Chapter 14

Physical Design Automation of MCMs

MultiChip Modules (MCMs) have been introduced as an alternative packaging approach to complement the advances taking place in the IC technology. Even though the steps in the physical design cycle of MCMs are similar to those in PCB and IC design cycle, the design tools for PCB and IC cannot be used for MCM directly. This is mainly due to the fact that MCM layout problems are different from both IC layout and PCB layout problems. The existing PCB design tools cannot handle the dense and complex wiring structure of MCMs. On the other hand, IC layout tools are inadequate to decipher the complex electrical, thermal and geometrical constraints of the MCM problems. As a result, the lack of CAD tools for MCMs is impeding further development in this area. Most of the commercial CAD tools available are the adapted versions of existing PCB tools and do not address the real problems associated with the MCM designs. Let us just consider the problem of routing in MCM. The signal effects of long lines in terms of crosstalk, noise, and reflections must be taken into account during routing. In addition, as high speeds are explored, the transmission line behavior of the interconnect must be modeled accurately to optimize the layout. All of these conditions have to be met, subject to the main goal of the interconnect, which is to route the signals between the chips. In designing CAD tools for MCM, many effects have to be taken into consideration such as clock skew, power noise disturbance, assembly effects of thermal mechanical nature that are caused by close positioning of chips, and limitations of assembly equipment. As a result, the design of multichip modules involves several disciplines such as electrical, chemical, material and mechanical engineering.

As MCMs are used for high performance system packaging, all steps in their physical design are performance driven. This makes the existing delay models for IC and PCBs inappropriate for MCMs. Therefore, new delay models will have to be developed for designing MCMs more accurately in order to comply with the stringent performance requirements.

The rest of the chapter has been organized as follows: In order to understand the issues and problems related to the physical design automation of MCMs different types of MCM technologies will be briefly described in Section 14.1. Section 14.2 will outline the different steps involved in the physical design of an MCM. Partitioning, the first phase of the MCM physical design cycle will be discussed in Section 14.3. MCM placement is discussed in Section 14.4. MCM routing problems will be described in Section 14.5.

14.1 MCM Technologies

MCMs, or more precisely, non programmable MCMs, are generally categorized into the following three, MCM-L,MCM-C, and MCM-D. MCM-L describes high density, laminated printed circuit boards. MCM-C refers to the ceramic substrates icluding both cofired and low-dielectric constant ceramics. MCM-D covers modules with deposited metallic wiring on silicon or ceramic support substrates. Yet another approach for fast turnaround is *Programmable MCM* (PMCM). In this section, we present a brief review of both programmable and non programmable MCM technologies. The methods for attaching chips to MCMs will also discussed in this section.

MCM-L (Laminates) is the oldest technology available. MCM-L is essentially an advanced PCB on which bare IC chips are mounted using Chip-On-Board (COB) technology. The well established PCB infrastructure can be used to produce MCM-L modules at a low cost. This makes them an attractive electronic packaging alternative for many low-end MCM applications with low interconnect densities. MCM-L becomes less cost-effective at higher densities where many additional layers are required. For cost- effectiveness, MCM technology must increase the functionality of each layer instead of adding more layers. MCM-L is considered a suitable technology for applications which require low risk packaging approach and most of the steps have already been automated.

MCM-C (ceramic) refers to MCMs with substrates fabricated with cofired ceramic or glass-ceramic techniques. These have been in use for many years and MCM-C has been the primary packaging choice in many advanced applications requiring both performance and reliability. Due to excellent thermal conductivity and low thermal expansion, ceramic substrates have also been used to serve as the package. Although interconnect densities are in the range of $200\text{-}400\ \mathrm{cm/cm^2}$, the same are not enough for high-end applications.

MCM-D (deposited) technology is closest to IC technology. It consists of substrates which have alternating deposited layers of high density thin-film metals, and low dielectric materials such as poly or silicon dioxide. MCM-D technology is an extension of conventional IC technology. It is developed specifically for high performance applications demanding a superior electrical performance and a high interconnect density. Since, this technology is relatively recent, it does not offer either a cost-effective manufacturing infrastructure, or a high volume application. Therefore, no significant commercial driving force

Characteristics	MCM-L	MCM-C	MCM-D
Line density (cm/cm^2)	250-400	200-400	> 400
Line width/separation (μms)	750/2250	125/125-375	10/10-30
Turnaround time	9-13 weeks	1 month	10-25 days
Years of availability	50	> 10	> 5

Table 14.1: Multichip Modules Classifications

exists. Table 14.1 compares the MCM families in terms of line widths, line density, line separation, turnaround time and the number of years for which these technologies have been available.

A full-custom design of an MCM requires significant engineering efforts. The lack of a mature infrastructure further magnifies the problem, since high density and high performance multichip modules are still expensive to fabricate and the cost increases with the number of mask layers. In order to side-step these difficulties, PMCMs have been introduced to minimize both the engineering delays and the cost. Programmable MCM approach is somewhat similar to Field Programmable Gate Arrays (FPGAs) technique. Just like gate arrays, PMCM wafers are manufactured in large quantities. A PMCM wafer has sites for chips and several layers of programmable interconnect. The customization process is carried out by setting programmable switches to establish the connectivity needed by the user. This is done after the chips have been placed on the chip sites. Thus, the customization consists of only placement of chips and programming the fuses (just like FPGA) to complete the routing.

Irrespective of the types of the MCM technology used, bare chips have to be attached to the substrates. Bare chips are attached to the MCM substrates in three ways, viz., wire bonding, Tape Automated Bonding (TAB) and flip-chip bonding. In wire bonding (illustrated in Figure 14.1 (a)), the back side of a chip (nondevice side) is attached to the substrate and the electrical connections are made by attaching very small wires from the I/O pads on the device side of the chip to the appropriate points on the substrate. The wires are attached to the chip by thermal compression. TAB is a relatively new method of attaching chips to a substrate. It uses a thin polymer tape containing metallic circuitry. The connection pattern is simply etched on a polymer tape. As shown in Figure 14.1(b), the actual path is simply a set of connections from inner leads to outer leads. The inner leads are positioned on the I/O pads of the chips, while the outer leads are positioned on the connection points on the substrate. The tape is placed on top of the chip and the substrate and pressed. The metallic material on the tape is deposited on the chip and the substrate to make the desired connections. Flip-chip bonding uses small solder balls on the I/O pads of the chip to both physically attach the chip and make required electrical connections (see Figure 14.1(c)). This is also called face down bonding, or Controlled-Collapse Chip Connections (C4).

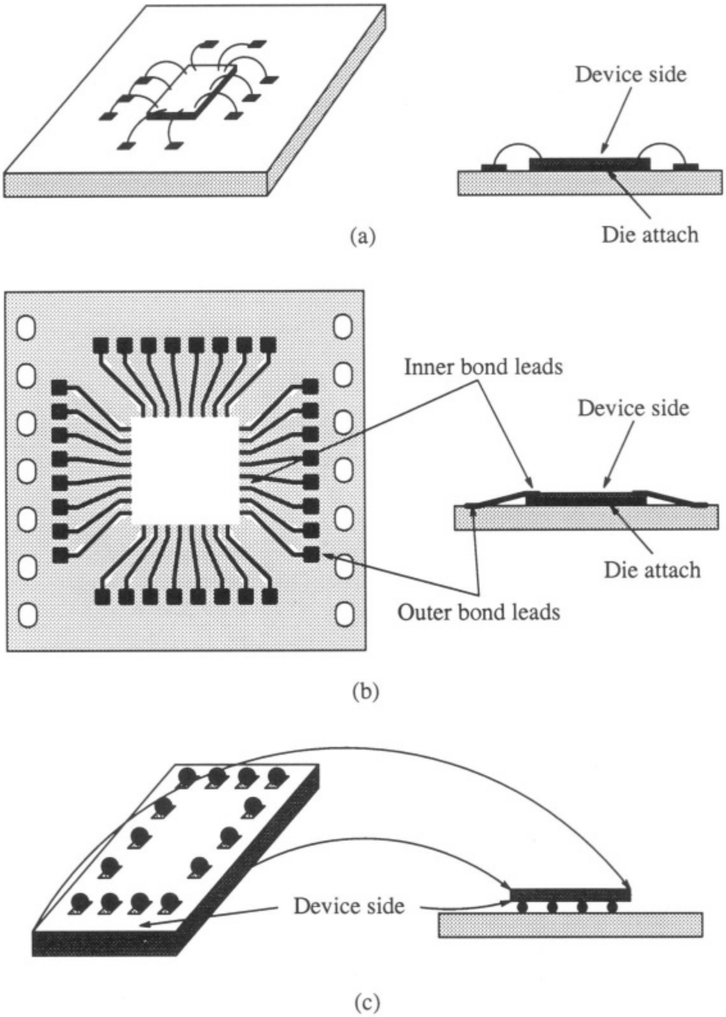


Figure 14.1: Die attachment techniques (a) Wire bonding (b) Tape automated bonding (c) Flip-chip bonding

14.2 MCM Physical Design Cycle

The physical design considerations of an MCM differ significantly from their counterparts for an IC. The input to the MCM physical design cycle is the circuit design of the entire system. The output is the MCM layout.The physical design cycle of an MCM pursues the following steps (also illustrated in Figure 14.2):

1. **Partitioning:** An MCM may contain as many as 100 chips. In turn, each chip can accommodate a certain number of transistors. The first assignment herein is to partition the given circuit into subcircuits.The partitioning should warrant fabrication of each subcircuits on a single chip. Simultaneously, the number of subcircuits should be equivalent to or less than the number of chips that the MCM can sustain. Please note that the MCM designs require performance driven approach. This requirement necessitates consideration of the power and timing constraints in the partitioning step.These requiremnts shall be in addition to the traditional I/O constraints and area constraints for chip sites.

2. **Placement:** The placement step is concerned with mapping the chips to the chip sites on the MCM substrate. Placement, of course affects not only the thermal characteristics of an MCM but also routing efficiency, which translates directly into manufacturability and cost. The number of components involved with the chip placement is much less as compared to the IC placement phase. However, timing and power constraints in MCM placement problem makes it a significantly different problem compared to the IC placement. Thermal considerations in MCM placement are important because bare chips are placed closer together and generate significant amount of heat. When the chip sites are prefabricated, the MCM placement problem lends itself to gate array based approach. Another variation of the MCM placement arises when the chips manufactured in different technologies need to be placed on an MCM. A critical difference between IC placement and MCM placement is allocation of routing regions. Unlike IC placement, no routing region needs to be allocated in MCMs since routing is done in routing layers and not between chips.

3. **Routing:** After the chips have been placed on the chip sites, the next phase of the MCM physical design is to connect these chips specified by the net list. The objective of minimizing routing area in IC design is no longer valid in MCM routing environment. Instead, the objective of the MCM routing is to minimize the number of layers, as the cooling requirement and therefore the cost of an MCM depends on the number of layers used. Because of the long interconnect wires involved in MCM design, crosstalk and skin effect become important considerations which are not of much concern in IC layout. In particular, in MCM-D, skin effect of the interconnect becomes more severe. The parasitic effects also degrade the performance if not accounted for in routing of MCMs.

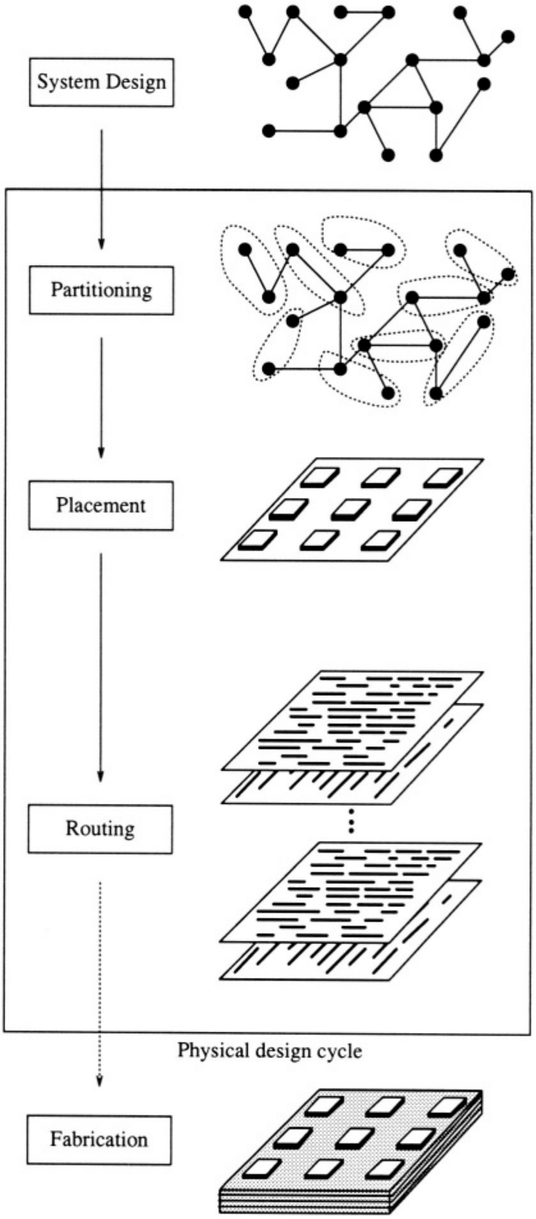


Figure 14.2: Physical design cycle of MCM

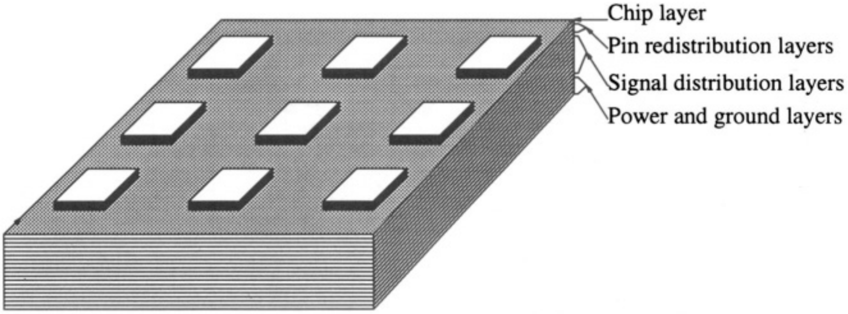


Figure 14.3: MCM routing environment

Power and ground signals do not complicate global routing because these signals are distributed on separate layers, and taps to the power supply layers are easy to make. However, overall dimensions must be tightly controlled (to fit within the package) and the packaging delay must be carefully controlled. The routing environment of an MCM can be viewed as a 3 dimensional space as shown in Figure 14.3.

14.3 Partitioning

As discussed in Chapter 5, the design of a complex system such as computer system consisting of tens of millions of components necessitates breaking the system into subsystems using a divide and conquer strategy. This process of decomposing the system into subsystems is called partitioning. Traditionally, partitioning has been applied at three levels, system level, board level and chip level. System level partitioning breaks the system design into sub-circuits which can fit on a PCB. Board level partitioning partitions each sub-circuit into a set of chips. The last step in the hierarchy of partitioning, chip level partitioning decomposes a chip circuit into smaller sub-circuits in order to ease the task of the chip designer.

With the introduction of multichip modules, the intermediate board level partitioning is replaced by *module level partitioning*. We refer to module level partitioning as MCM partitioning. The module level partitioning is characterized by high performance and high density design. Thus the module level partitioning is performance driven. The module level partitioning is becoming an important ingredient for complex design with the rapid increase of the device density. The device density has, on an average, doubled annually for almost two decades. It is anticipated that such advances will continue to be made well into 1990s. This growth in the devices per unit area makes the problem of MCM partitioning challenging.

The MCM partitioning depends on design style. If an gate array type

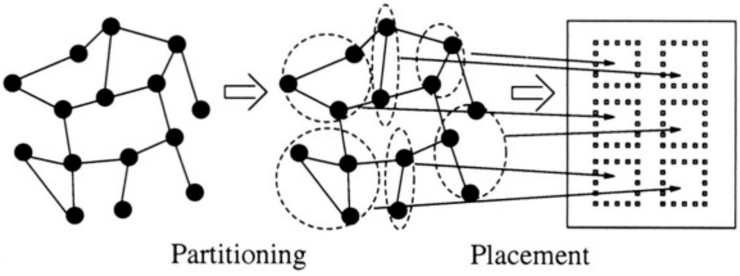


Figure 14.4: MCM partitioning and placement

design style is used, the MCM partitioning problem is similar to the gate array partitioning problem except each 'gate' corresponds to a chip. We refer to this approach as a *chip array* approach. If a full custom type approach is used, then each chip can have a different size and the MCM partitioning is analogous to the full custom partitioning problem. In the following, we restrict our discussion to chip array approach.

MCM partitioning is defined as an optimum mapping of the design to a set of chips (see Figure 14.4). However, as the performance considerations enter the design, the MCM partitioning process must consider other constraints as well. So for high performance system designs, MCM partitioning can be defined as a partition of the design to a set of chips that minimizes the inter-chip wire crossings subject to timing constraints, area constraints, thermal constraints and I/O pin count constraints.

An MCM package can be considered to contain a set ($\mathcal{C}$) of equal sized chips, each chip placed in a chip slot. Each chip $c \in \mathcal{C}$ has constraints on area A_c, thermal capacity $\mathcal{H}_c$, and maximum number of terminals (I/O pins) $\mathcal{T}_c$. A synchronous digital system consists of registers and blocks of combinational logic. For simplicity, all clock generation and distribution circuits are ignored. The system can be represented by an edge weighted graph called *system graph* $G = (V, E)$ [SKT94], where $V = R \cup B$, R is the set of nodes representing registers, B is the set of nodes representing combinational blocks, and E is the set of all *directed edges,* which correspond to signal flow in the system. Associated to each edge $e_{ij} \in E$, there is a weight w_{ij} representing the total number of wires between nodes v_i and v_j in V. Associated with each node $v_i \in V$, we have three parameters, area a_i, power consumption h_i, and internal delay d_i. Figure 14.5 shows a system graph.

The MCM partitioning problem is to find an optimum mapping $\varphi : V \rightarrow C$ such that the number of total inter-chip connections

$$W = \sum_{\forall i,j,\varphi(v_i)\neq\varphi(v_j)} w_{ij}$$

is minimized while satisfying the timing constraints, area constraints, terminal

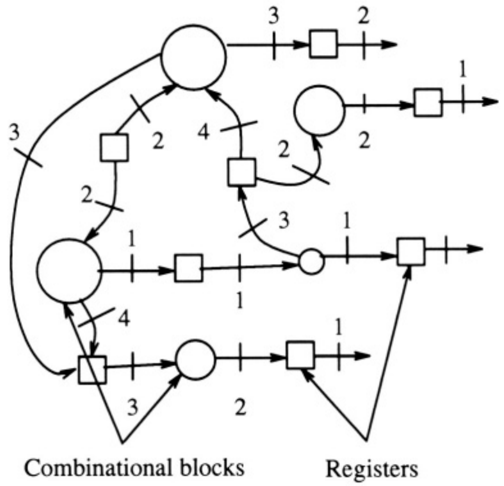


Figure 14.5: System graph

constraints, and thermal constraints. The area and the terminal constraints is the same as the area and the terminal constraints of IC design partitioning. However, for the timing constraints, we need to consider the internal delay of each circuit. The timing and the thermal constraints can be stated as:

1. **Timing constraints:** A register-to-register delay through some combinational logic blocks must be less than or equal to the cycle time.

$$d_j + D(\varphi(r_i), \varphi(c_j)) + D(\varphi(c_j), \varphi(r_k)) \leq T_{cycle}$$

   for all $c_j \in C$, where $r_i, r_k \in R$, $e_{ij}, e_{jk} \in E$, D is the time delay between objects and T_{cycle} is the given cycle time.

2. **Thermal constraints:** The total heat generated by a partition must be less than or equal to the thermal capacity of the corresponding mapped slot.

$$\sum_{\forall i, \varphi(v_i)=s} h_i \leq \mathcal{H}_s$$

Thermal constraints can be treated in a similar fashion as area constraints. Therefore, any performance driven partitioning algorithm may be applied by taking into consideration the thermal constraints. For the gate-array based approach, where each chip slot is of equal size, any gate-array partitioning algorithm may be applied with appropriate modifications. Similarly, any high-performance full custom partitioning algorithm may be applied for the generalized full custom based MCMs. In [SKT94], Shin, Kuh, and Tsay presented a performance driven integrated partitioning and placement technique for MCMs.

They only considered timing and area constraints. Two different delay models have been considered: 1) constant delay model, and 2) linear delay model. For constant delay model, their approach is essentially a partitioning algorithm which will be briefly described below.

In [SKT94], given a system graph, assuming that delay time for the signal traveling between a combinational block and a register that are grouped into the same partition is negligible. In addition, the delay time for the signal traveling between a combinational block and a register that are partitioned into different groups is a constant. Each group is called a *super node* and corresponds to a chip.

For each combinational block c_i, the algorithm finds the two registers r_j and r_k that are adjacent to the block in the system graph. The procedure of constructing super nodes is shown by an example in Figure 14.6. In this example, we assume the system cycle time is $T_{cycle} = 6$ which requires that the maximum delay time between any two registers should be less than or equal to 6. The delay time between a combinational block and a register is assumed as $D = 2$. The super nodes are constructed according to the following three cases.

1. Both registers must be combined: In this case, the condition $d_i + D > T_{cycle}$ must be satisfied. Consider the example shown in Figure 14.6(a) with $d_i = 5$. If one of r_j and r_k is assigned to a different partition than c_i, the time delay will be at least 2 + 5 = 7, thus violating the timing constraint. So, all these vertices have to be included in the same super node.

2. At least one of the registers must be combined: In this case, the conditions $d_i + D < T_{cycle}$ and $d_i + 2 \times D > T_{cycle}$ must be satisfied. Consider the example shown in Figure 14.6(b) with $d_i = 3$. If both registers are assigned to different partitions than c_i, the time delay will be 7, thus violating the timing constraint. In this situation, the super node consists of combinational block and either one of the registers.

3. No registers need to be combined: In this case, the condition $d_i + 2 \times D < T_{cycle}$ must be satisfied. Consider the example shown in Figure 14.6(c) with $d_i = 1$, the registers can be assigned to any partitions without violating the timing requirement. Thus, each super node consists of only one vertex.

The algorithm repeats until no nodes can be combined. At this stage, the number of super nodes is equal to the number of chips required in the MCM.

14.4 Placement

The thermal and timing considerations in the MCM placement problem make it significantly different than the IC placement. With the increase in the density of the individual chip, the thermal requirements have also gone up. High

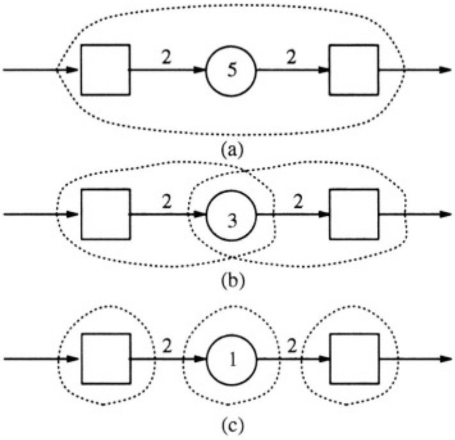


Figure 14.6: Super node construction

speed VLSI chips may generate heat from 40 to 100 watts. In order to ensure proper operation of the design, such a large amount of heat must be dissipated efficiently. The heat dissipation of an MCM depends directly on how the chips are placed. In addition to this, the timing constraints for the design must also be satisfied. These timing constraints are responsible for the proper operation of the module at high frequencies. The placement problem in MCM is to assign chips to the chip sites on the substrates subject to some constraints. If the placement is not satisfactory, then the subsequent steps of routing will be inefficient.

Chip level placement determines the relative positions of a large number of blocks on an IC as well as organizes the routing area into channels. As opposed to IC placement problem, MCM placement involves fewer components (100-150 ICs per MCM compared to 10-1000 general cells per IC) and the sizes and shapes of ICs on an MCM are less variable than the general cells within the IC. MCM placement is more complex because many interrelated factors determine layout quality. Wide buses are very prevalent, propagation delays and uniform power dissipation are much more important. As opposed to IC placement problem, the main objective of MCM placement is to assign the chips to the chip sites such that the number of routing layers is minimized. In addition, other constraints such as timing constraints and thermal constraints make the MCM placement problem more difficult. The MCM placement problem can be formally stated as follows: given a set of chips C, and a set of chip sites on the substrate $\mathcal{S}$, assign $\mathcal{C}$ to $\mathcal{S}$, e.g., find a mapping $\phi : C \rightarrow S$ subject to timing constraints and thermal constraints and to minimize the number of layers. The typical values for $|\mathcal{C}|$ and $|\mathcal{S}|$ range between 4-100.

There are mainly two types of placement related to MCMs, namely, chip array and full custom.

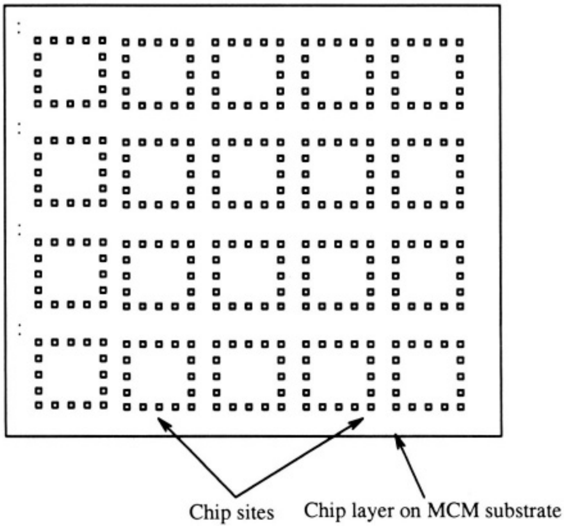


Figure 14.7: MCM placement problem

14.4.1 Chip Array Based Approach

The MCM placement approach when the chip sites are symmetric, becomes very similar to the conventional gate array approach. In this case, the MCM placement problem is the assignment of the chips to predefined chip sites. However, the key difference between the IC placement and the MCM placement problem is the type of constraints involved. Figure 14.7 shows a chip array MCM substrate. The two approaches to MCM placement problem have been discussed by LaPotin [LaP91] as part of the early design analysis, packaging and technology tradeoffs.

14.4.2 Full Custom Approach

One of the important features of the MCMs is that it allows the integration of mix of technologies. This means, each individual chip can be optimally fabricated using the technology best suited for that chip. Figure 14.8 shows an arrangement depicting a concept of *2.5-D integration* scheme derived from ideas postulated by McDonald [McD84] and Tewksbury [Tew89]. This concept can be viewed as an advanced version of the existing MCMs. It is envisioned that this hypothetical system will respond directly to the cost limitations of VLSI technologies. The system could be assembled on a large-area active substrate. The technology of such a substrate could be optimized for yield, power, and speed of the interconnect. This substrate could dissipate a large percentage of the total power and could be cost-effective if fabricated with relaxed design rules in stepper-free, interconnect-oriented technology. The performance-

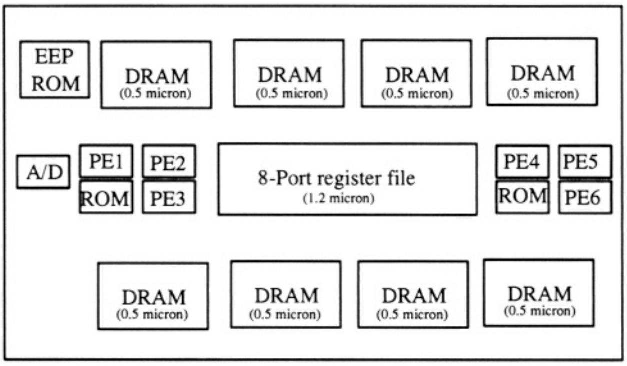


Figure 14.8: 2.5-D integration placement in MCM

critical system components could be fabricated separately on fabrication lines oriented toward high volume and high performance. They could be attached to the active substrate with rapidly maturing flip-chip technology. This way only those system elements that really require ULSI technology (for example, data path) would be fabricated with the most expensive technologies. It is obvious that placement problem in 2.5-D integration scheme is that of full-custom approach. In addition to the usual area constraints, the placer of this type must be able to complete the task of placement subject to the thermal and timing constraints.

14.5 Routing

After the chips have been placed on the chip sites, the next phase of the MCM physical design is to connect these chips specified by the net list. As mentioned earlier that unlike IC design, performance is the main objective in MCM design. Therefore, the main objective of routing is to satisfy timing constraints imposed by the circuit design. Also, the cost of an MCM is directly proportional to the number of layers used in the design. Thus minimizing the total number of layers used is also an objective of MCM routing. In particular, in MCM-D, cross talk, skin and parasitic effect of the interconnect become more critical. Crosstalk is a parasitic coupling between neighboring lines due to the mutual capacitances and inductances. In the design of high speed systems, crosstalk is a primary concern. Excessive crosstalk compromises noise margins, possibly resulting in false receiver switching. The crosstalk between the lines can be minimized by making sure that no two lines are laid out in parallel or next to each other for longer than a maximum length.

In addition to crosstalk, the skin effect is also a major consideration in

MCM routing. Skin effect is defined as characteristic of current distribution in a conductor at high frequencies by virtue of which the current density is greater near the surface of the conductor than its interior. As the rise time of digital pulses is reduced to the sub-nanosecond range, the skin effect becomes an important issue in high speed digital systems. As the frequency, the conductivity, and permeability of the conductor are increased, the current concentration is also increased. This results in increasing resistance and decreasing internal inductance at frequencies for which this effect is significant. These effects must be taken into account while routing long lines.

14.5.1 Classification of MCM Routing Algorithms

The routing of an MCM is a three-dimensional general area routing problem where routing can be carried out almost everywhere in the entire multilayer substrate. However, the pitch spacing in MCM is much smaller and the routing is much denser as compared to conventional PCB routing. Thus traditional PCB routing algorithms are often inadequate in dealing with MCM designs.

There are four distinguished approaches for general (non-programmable) MCM routing problems:

1. Maze Routing
2. Multiple Stage Routing
3. Topological Routing
4. Integrated Pin Distribution and Routing

The routing of programmable MCMs is very similar to that of FPGAs. In this section, we discuss routing of both MCMs and PMCMs.

14.5.2 Maze Routing

The most commonly used routing method is three dimensional maze routing. Although this method is conceptually simple to implement, it suffers from several problems. First, the quality of the maze routing solution is very much sensitive to the ordering of the nets being routed, and there is no effective algorithm for determining a good net ordering in general. Moreover, since the nets are routed independently, global optimization is difficult and the final routing solution often uses a large number of vias despite the fact that there is a large number of signal layers. This is due to the fact that maze router routes the first few nets in planar fashion (using shorter distances), the next few nets use a few vias each as more and more layers are utilized. The nets routed towards the end tend to use a very large number of vias since the routing extends over many different layers. Finally, three dimensional maze routing requires long computational time and large memory space.

14.5.3 Multiple Stage Routing

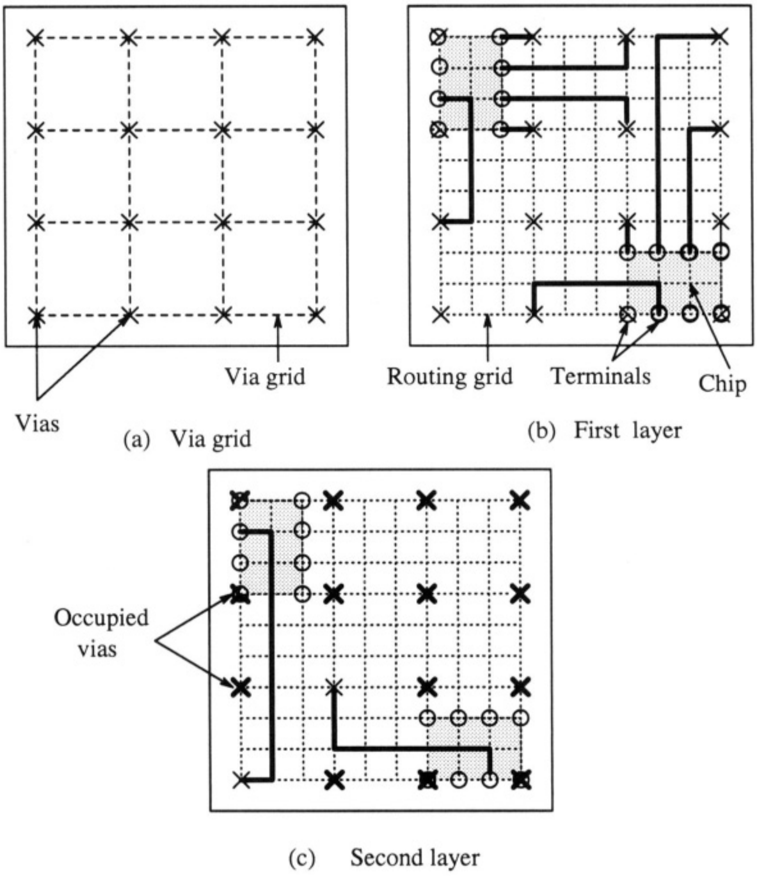
In this approach, the MCM routing problem is decomposed into several subproblems. The close positioning of chips and high pin congestion around the chips require separation of pins before routing can be attempted. Pins on the chip layer are first redistributed evenly with sufficient spacing between them so that the connections between the pins of the nets can be made without violating the design rules. This redistribution of pins is done using few layers beneath the chip layer. This problem of redistributing pins to make the routing task possible, is called *pin redistribution.* After the pins are distributed uniformly over the layout area using pin redistribution layers, the nets are assigned to layers on which the assigned nets will be routed. This problem of assigning nets to layers is known as *layer assignment problem.* The layer assignment problem resembles the global routing of the IC design cycle. Similar to the global routing, nets are assigned to layers in a way such that the routability in layer or in a group of layers is guaranteed and at the same time the total number of layers used is minimized. The layers on which the nets are distributed are called *signal distribution layers.* The detailed routing follows the layer assignment. The detailed routing may or may not be reserved layer model. The horizontal and vertical routing may be done in same layer or different layers. Typically, nets as distributed in such a way that each pair of layers is used for a set of nets. This pair is called $x - y$ *plane pair* since one layer is used for horizontal segments while the other one is used for vertical segments. Another approach is to decompose the net list such that each layer is assigned a planar set of nets. Thus MCM routing problem become a set of *single layer* problem. Yet another routing approach may combine the x–y plane pair and single layer approaches. In particular, the performance critical nets are routed in top layers using single layer routing because xy-plane pair routing introduces vias and bends which degrade performance.

We now discuss each of these problems in greater detail in the following subsections.

14.5.3.1 Pin Redistribution Problem

Pins in chip layer need to be redistributed to help in the routing process. This is accomplished in pin distribution layers. The pin redistribution problem can be stated as: Given the placement of chips on an MCM substrate, redistribute the pins using the pin redistribution layers such that one or more of the following objectives are satisfied (depending the the design requirements):

1. minimize the total number of pin redistribution layers.
2. minimize the total number of signal distribution layers.
3. minimize the cross-talks.
4. minimize the maximum signal delay.
5. maximize the number of nets that can routed in planar fashion.



(a) Via grid

(b) First layer

(c) Second layer

Figure 14.9: Pin redistribution example

It is to be noted that the separation between the adjacent via-grid points may affect the number of layers required [CS91]. The pin redistribution problem can be illustrated by the example shown in Figure 14.9. The terminals of chips need to be connected to the vias shown in Figure 14.9(a). Usually, it is impossible to complete all the connections. In this case, we should route as many terminals as possible (shown in Figure 14.9(b)). The unrouted terminals are brought to the next layer and routed in that layer as shown in Figure 14.9(c). This procedure is repeated until each terminal is connected to some via. In [CS91], various approaches to pin redistribution problem have been proposed.

14.5.3.2 Layer Assignment

The main objective of layer assignment for MCMs is to assign each net in x-y pair of layers subject to the feasibility of routing the nets on a global routing grid on each plane-pair. This step determines the number of plane pairs required for a feasible routing of nets and is therefore important step in the design of the MCM. The cost of fabricating an MCM, as well as the cooling of the MCM when it is operation, are directly related to the number of plane-pairs in the MCM, and thus it is important to minimize the number of plane-pairs. There are two approaches known to the problem of layer assignments [HSVW90b, SK92]. The problem of layer assignment has been shown to be NP-complete [HSVW90b].
An approximation algorithm, for minimizing the number of layers, has been presented by Ho, Sarrafzadeh, Vijayan and Wong [HSVW90b].

14.5.3.3 Detailed Routing

After the nets have been assigned to layers, the next step is to route the nets using the signal distribution layers. Depending on the layer assignment approach, the detailed routing may differ. Routing process may be single-layer routing or x-y-plane-pair routing. Usually a mixed approach is taken in which the single-layer routing is first performed for more critical nets, followed by x-y-plane-pair routing for less critical nets. Two models can be employed for x-y-plane-pair routing, namely xy-reserved model and xy*-free* model. One advantage in xy*-free* model is that bends in nets do not necessarily introduce vias where bends in nets introduce vias in xy-reserved model. The detailed routing was presented in [LSW94].

14.5.4 Topological Routing

In [DDS91], Dai, Dayan and Staepelaere developed a multilayer router based on rubber-band sketch routing. This router uses hierarchical top-down partitioning to perform global routing for all nets simultaneously. It combines this with successive refinement to help correct mistakes made before more detailed information is discovered. Layer assignment is performed during the partitioning process to generate routing that has fewer vias and is not restricted to one-layer one-direction. The detailed router uses a region connectivity graph to generate shortest-path rubber-band routing.

The router has been designed primarily for routing MCM substrates, which consist of multiple layers of free (channelless) wiring space. Since MCM substrate designs have potentially large number of terminals and nets, the router of this nature must be able to handle large designs efficiently in both time and space. In addition, the router should be flexible and permit incremental design process. That is, when small changes are made to the design, it should be able to be updated incrementally and not recreated from scratch. This allows faster convergence to a final design. In order to produce designs with fewer vias, the router should be able to relax the one-layer one-direction restriction. This is an

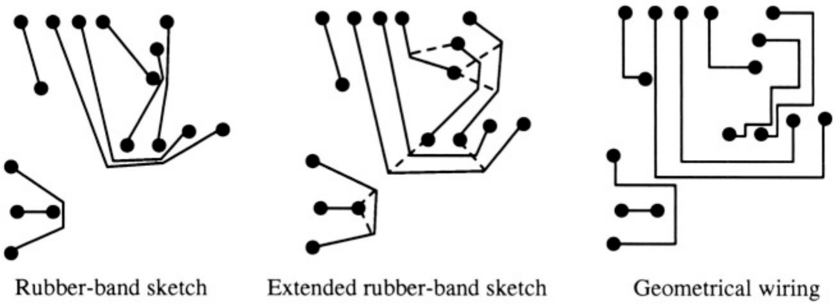


Figure 14.10: Rubber-band representations

important consideration in high speed designs since the discontinuities in the wiring caused by bends and vias are a limiting factor for system clock speed.

In order to support the flexibility described above, the router must have an underlying data representation that models planar wiring in a way that can be updated locally and incrementally. For this reason, SURF models wiring as rubber-bands [CS84, LM85]. Rubber-band provides canonical representation for planar topological wiring. Because rubber-bands can be stretched or bent around objects, this representation permits incremental changes to be made that only affect a local portion of the design. A discussion of this representation has been described in [DKJ90].

Once the topology of the wiring is known, the rubber-band sketch can be augmented with spokes to express spatial design constraints such as wire width, wire spacing, via size, etc. [DKS91]. Since successful creation of the spoke sketch guarantees the existence of a geometrical, wiring (Manhattan or octilinear), the final transformation to fixed geometry wiring can be delayed until later in the design process. This allows most of the manipulation to take place in more flexible rubber-band format. Figure 14.10 shows different views of the same wiring topology. These represent various states of the rubber-band representation.

In this context, a topological router has been developed that produces multilayer rubber-band sketches. The input to this router is a set of terminals, a set of nets, a set of obstacles, and a set of wiring rules. These rules include geometrical design rules and constraints on the wiring topology. The topological constraints may include valid topologies (daisy chain, star, etc.) as well as absolute and relative bounds on segment lengths. The output of the router is a multilayer rubber-band sketch in which all the points of a given net are connected by wiring. Although the routability of a sketch is not guaranteed until the successful creation of spokes. At each stage, the router uses the increasingly detailed information available to generate a sketch without overflow regions. This increases the chance that the sketch can be successfully transformed into a representation (the spoke sketch) that satisfies all of the spatial constraints. In addition the router tries to reduce overall wire length and the number of vias. A more detailed analysis of routability of a rubber-band sketch is described in [DKS91].

14.5.5 Integrated Pin Distribution and Routing

In [KC92], Khoo and Cong presented an integrated algorithm SLICE for routing in MCM. The basic idea is to redistribute pins simultaneously with routing in each layer, instead of the pins distribution prior to routing. SLICE performs planar routing on a layer by layer basis. Subsequent to routing on one layer, the terminals of the unrouted nets are propagated to the next layer. The routing process is then continued until all the nets are routed.

An important feature of SLICE is computation of planar set of nets for each layer. The algorithm strives to connect maximum number of nets in each layer. The algorithm attempts partial routing of nets that cannot be routed completely in a layer. This facilitates completion of nets in the subsequent layer with shorter wires. The routing region is scanned from left to right. A topological planar set of nets is computed for each adjacent column-pair using maximum weighted non-crossing matching. The matching is comprised of a set of non-crossing edges that extend from the left column to the right column. Thereafter, the physical routing between the column-pair is generated based on the selected edges in the matching. This process is carried out for each column from left to right. The completion of the planar routing in a layer is followed by distribution of the terminals of the unrouted nets so that they can be propagated to the next layer without causing local congestions. The left to right scanning operation in the planar routing culminates in predominantly horizontal wires in the solution. A restricted two-layer maze routing technique is adopted for completion of the routing in vertical direction. Unnecessary jogs and wires are eliminated after each layer is routed. The terminals of the unrouted nets are propagated to the next layer. Finally, the routing region is rotated by 90^0 so that the scanning direction is orthogonal to the one used in the previous layer. The process is iterated until all the nets have been routed. Details of the planar routing, pin redistribution, and maze routing are available in [KC92].

14.5.6 Routing in Programmable Multichip Modules

Like gate arrays, routability is a key concept in the design of programmable MCMs. In a programmable MCM design, most if not all, of the masking or phototooling steps are defined prior to commencement of the system designing. Initially, a substrate is manufactured in a generic fashion. Subsequently, it is customized for fulfilling the specific needs of the user. The capability for routing complex and dense multichip designs requires early designing of a highly routable wiring topology. An important component for achieving efficient programmable designs is the design tool that can sustain the dual responsibility of: one, deciphering the programmable wiring structure; and two, performing the actual routing (customization) needed to realize an application specific MCM. It is noteworthy that the routing efficiency is a factor of both the base wiring density and the resource utilization. The base wire density is typically measured in inches of wire per square inch of the substrate area. The resource utilization refers to the fraction of available wiring that can be utilized in routing a design. The total wire length used, relative to the minimum theoretical routing length, must be accounted for.

Electrical performance is a key ingredient to any programmable custom MCM design. If the programmable approach fails to meet the performance goals, then its application objectives will not be accomplished. In many circumstances, electrical performance of the signal interconnect will be relatively good even without rigorous design for characteristic impedance, low loss etc. This can be ascribed to the electrical length of the signal wiring. In most MCM environments, the same is short as compared to the wavelength/rise times of the IC signals. The crux of the issue in nearly all cases is capacitive loading reduction for CMOS systems in order to minimize delay caused by RC time constraints. In other words, a large fraction of system designs will be needed to address signal delay more than high bandwidth signal fidelity. This may not be the case in a more conventional single chip packaging/PC board implementation where physical/electrical lengths of interconnect are longer and more significant. Perhaps a more compelling issue associated with signal fidelity is power distribution. Many signal noise problems develop due absence of clean power and ground supplies. Due to these, noise is fed forward through output drivers, which diminishes noise margins at the receivers. This imposes an additional demand on the design of a programmable MCM. It decrees that the power distribution scheme must be supportive of high performance, in addition to being flexible. The power distribution network of the MCM design is usually predefined and accommodates a myriad of supply voltages, variable supply potentials, and a variety of both AC/DC current requirements.

Figure 14.11 illustrates a simplified cut-away view of a programmable multichip module with a substrate wherein antifuses have been incorporated. The substrate is comprised of four metal layers separated by dielectric layers. The lower two layers are used for power distribution. On the other hand, the upper two layers are used for an orthogonal wiring grid with permanent vias or antifuses in selected grid interconnections. The uppermost layer also houses the bonding pads. The bare chips will be electrically connected to these pads upon completion of the programming. A signal path can be programmed through the substrate by linking previously uncommitted line elements together via the antifuses. The interconnection line architecture of actual designs is much more convoluted than the one presented in the above mentioned simplified example. However, the principle of programming remains unaltered in either case. Since, all line elements are accessible from a bonding pad, a programming pulse can be applied. A programming pulse with a voltage amplitude larger than the threshold voltage is applied using a wafer prober to a pair of wiring elements in order to connect them to each other.

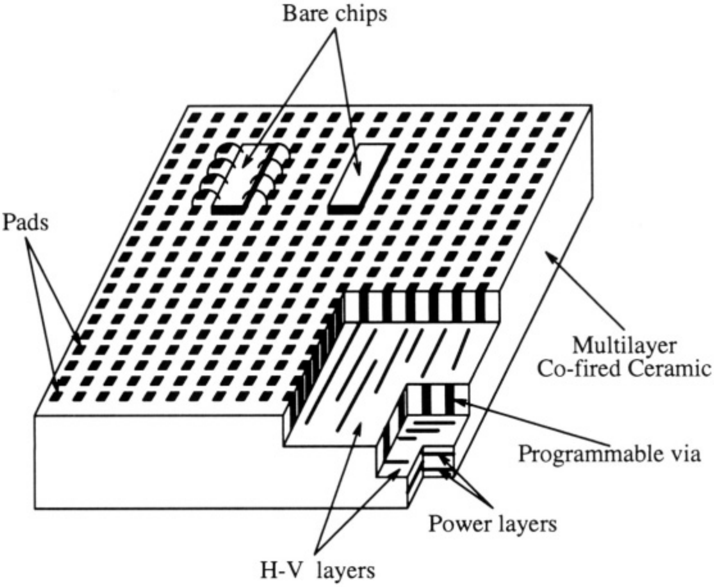


Figure 14.11: Routing in programmable multichip modules

14.6 Summary

The MCM approach to microelectronic packaging has significantly improved the system performance. Such improvement has been acheived by bridging the gap between the existing PCB packaging approach and the progressing VLSI IC technology. The physical design of MCMs is an important ingredient of the overall MCM design cycle. The density and complexity of contemporary VLSI/ULSI chips require automation of the physical design of MCMs. Further developments of MCMs face stuff challenges due to limited research in the area of development of algorithms requisite for MCM physical design. This is primarily attributable to the fact that MCMs pose an entirely new set of problems which cannot be solved by existing PCB or IC layout tools. Therefore, considerable research efforts need to be steered towards development of algorithms for MCM physical design automation.

14.7 Exercises

1. A Multi-Chip Module (MCM) consists of many interconnected bare chips. Consider a hypothetical MCM with four chip slots. Each slot has five terminals. Does there exist a 4-way partition of the graph in Figure 14.12,

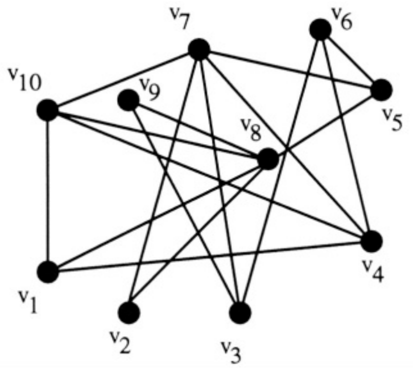


Figure 14.12: A graph partitioning problem.

each having no more than three vertices and the number of terminals for each partition is no more than five?

† 2. Consider the thermal-driven placement problem in which the chips are to be placed onto chip sites such that the heat distribution across the multichip module is uniform. Develop an algorithm for such a placement.

† 3. In MCM placement problem, the heat distribution should be uniform over the MCM. Modify the simulated annealing algorithm described in Chapter 5 to take the heat effect into account so that it can be used in MCM placement.

4. The routing problem for MCMs is three dimensional. Extend maze routing algorithm for routing a two-terminal net in three-dimensions.

† 5. Extend line probe algorithm for global routing a two-terminal net in three-dimensions.

† 6. Let L be the longest possible length of a net that does not cause undue skin effects. Develop a global router that guarantees the length constraints imposed by the skin effects.

‡ 7. Formulate global routing in three-dimensions as a Hierarchical Integer Program with an objective of optimizing overall wire length.

‡ 8. Develop a heuristic algorithm for pin redistribution such that it minimizes the net lengths and the number of of layers needed.

‡ 9. Develop a *crosstalk-driven router* for MCM, which routes all the nets and also minimizes the crosstalk between the neighboring lines. Assume that the system is to be assembled on a multichip module using silicon substrate and silicon dioxide as dielectric layer.

† 10. Consider the following channel routing problem motivated by the crosstalk. Let L be the longest distance two nets can run parallel to each other without causing undue crosstalk problems. Modify Yoshimura-Kuh channel routing algorithm to minimize the crosstalk.

Bibliographic Notes

A comprehensive introduction to the technology of MCM-based electronic packaging, covering all aspects of MCM, including classification, design, and CAD tools, and explaining methods and materials used in the design of MCM-based systems is given in the book *Introduction to MCMs* [SYB95]. A textbook by Tummala and Rymaszewski [TR89] covers the fundamental concepts of microelectronic packaging. A survey of electronic packaging technology appears in [Tum91]. A mathematical analysis of different system packaging parameters can be found in [Mor90]. An excellent discussion on die attachment techniques can be found in the book by Bakoglu [Bak90]. The discussion on early design analysis can be found in [CL89, LaP91]. Discussion on testing and diagnosis of multichip modules can be found in [KT91]. A detailed analysis about the skin effect in thin-film interconnections for ULSI/VLSI packages has been described in [HT91]. An electrical design methodology for multichip modules is described by Davidson [Dav84]. An excellent discussion about thermal issues in MCMs has been presented by Buschbom [Bus90]. In [RP96]an adaptive genetic algorithm for Performance driven MCM Partitioning is presented. In [CL96], a multilayer, MCM router called MCG, is introduced for x-y routing. The book [Lic95] is a guide to using multichip modules (MCMs) in the design, testing, and manufacture of electronic systems and equipment, for students and professionals in electronics, computer, and materials engineering. [CF96] presents current and future techniques and algorithms of high performance multichip modules (MCMs) and other packaging methodologies. A genetic algorithm for building-block placement of MCMs and ICs is presented which simultaneously minimizes layout area and an Elmore-based estimate of the maximum path delay while trying to meet a target aspect ratio is presented in [EK96b]. In [LGKM96], chip pad migration is shown as a key component to high performance MCM design. The book [SK84] collects together a large body of important research work that has been conducted in recent years in the area of Multichip Module (MCM) design. All major aspects of MCM physical design are discussed, including interconnect analysis and modeling, system partitioning and placement, and multilayer routing. IMAPS-International Microelectronics And Packaging Society plays a key role in advancing the state of the art in MCM technology by organizing workshops, conferences and educational tutorials. The www site for IMAPS is (www.imaps.org).