This makes simple, cost effective, and one step (gang) removal virtually impossible. Due to some of the inherent problems associated with tin tape, however, and the need for a fine pitch TAB assembly technology using gold tape, some companies are currently developing removal and repair processes for gold to gold bonds. Mechanical OLB bonds, a recent development, offer a great potential for rework and repair.

9.4.10 Manufacturing Issues and Costs

The manufacturing of TAB devices, regardless of the packaging scheme, depends on an industry infrastructure that provides the proper materials (TAB tape) and processing equipment (bonders, tooling etc.) to form a reliable and robust process. Historically, the capabilities of TAB equipment vendors have focused on low lead count devices for consumer electronic products. This focus has limited the application of the equipment in areas of high lead count, fine pitch TAB assembly, needed for multichip applications. As a result, the infrastructure for advanced TAB materials and assembly equipment needs further development, particularly in the United States. Equipment suppliers, tape manufacturers and contract assembly houses must cooperate to fill this technology gap. Consideration of the manufacturing processes, material choices and temperature exposures are issues needing attention in the successful implementation of TAB. This section will highlight a few issues and costs that should be given some attention.

TAB Tape

The major issues associated with TAB tape include consistent tape quality, the cost for advanced circuits and shelf life for tin tape.

The most crucial aspect of a robust TAB process rests on the material consistency of the TAB tape. Dimensional requirements must be met repeatably and depend upon the design of the tape as well as on the material selection and tape processing practices. Fine pitch tapes, for example, with long, unsupported cantilevers can cause dimensional problems during tape processing and handling. Consistent plating is required to enable the ILB and OLB processes to maintain controlled process windows.

The cost of the TAB leadframe influences the decision to pursue a TAB application and greatly depends on the design of the leadframe and the challenges it presents to the tape manufacturer. Tape costs range from $5.00 per frame to over $50.00 per frame (1992 dollars), depending on the complexity of the design (one metal layer tape versus multi-metal layer tape) and the volumes ordered.

The shelf life of the tape is a determination factor for the low volume user. These users must balance the low volume versus cost versus shelf life to justify the use of TAB. One possible option is the use of a JIT approach to receiving the tape material.

Bumping. The cost of wafer bumping can have a negative impact on the implementation of a TAB process, as this cost is concentrated only on the yielding die. As a result, large, low yielding die can bear the burden of wafer bumping costs totaling as much as $100 per wafer. Wafer bumping also requires expensive semiconductor processing equipment and cleanroom facilities.

Many users, buying die from merchant semiconductor companies, do not have the option of having them bumped. This, along with cost issues, has led to bumping alternatives requiring less capital intensive methods, such as transfer bump TAB (TBTAB), gold ball bumping etc. These methods, however, may not have the ultra-fine pitch and reliability capabilities of wafer bumping and are, therefore, limited in application.

Inner Lead Bonding. Inner lead bonding equipment costs can total anywhere from $200,000 to $400,000 (1992 dollars), depending on the level of automation and the type of bonder. As mentioned earlier, single point bond technology has yet to realize full automation and use of this equipment can result in slower production times. Automated gang bonders have limited application to larger, fine pitch die and require die specific tooling.

Outer Lead Bonding. The cost of setting up an OLB process can be as much as $500,000 (1992 dollars), depending on the desired bonding technology and level of TAB complexity. Fine pitch applications require very accurate placement systems which contribute the bulk of the cost. Single point bonding systems may require the purchase of separate placement and bonding systems, whereas gang bonders usually have the bonder and placement functions integrated. Custom tooling is also a cost concern, as the TAB standards do not take full advantage of the density attributes of TAB.

Package quality is an important consideration and is reflected in the choice of bonding technology. Critical among these considerations is metallurgy consistency and flatness (for gang bonding). Flux contamination may be a reliability issue. Cleaning the flux is also, currently, a major environmental issue.

Temperature Hierarchy. Encapsulation, test on tape/burn-in and repair are other steps in a TAB assembly process that should be looked at closely, especially from a temperature exposure perspective. Each of the TAB processes affect both previous and future processes. Burn-in can be detrimental to the solderability of tin plated tapes by causing accelerated growth in copper-tin intermetallics. The OLB process temperatures can have an effect on the encapsulation materials. These issues force the broad consideration of the entire process flow and the effects it may have on materials and other processes. A coordinated approach with design, development and manufacturing is necessary.

9.4.11 Comparison with Other Connection Technologies

The choice of which connection technology (wire bond, TAB or flip chip) to use on MCM applications is based on a balance between cost and performance. TAB is most cost effective in high volume, low product mix environments. At lower volumes and higher part mixes, TAB becomes an expensive alternative. This expense can be justified where performance (electrical, thermal, etc.) of TAB justifies the use of a more expensive alternative.

TAB inherently adds cost to the assembly process. The additional step of wafer bumping, the cost of each TAB frame and custom tooling all contribute to a higher assembly cost for TAB. These costs can make TAB unattractive unless the user has a high volume product.

Cost considerations have lead to the development of new materials and methods that may enable TAB to compete with the other chip connection technologies. Developments in single point and laser bonding techniques help eliminate ILB and OLB pattern specific gang bond tooling. TAB standards help minimize the number of excise and form tools needed, as well as the hard tooling for tape fabrication.

Electrical performance is improved through the use of TAB assembly. As the number of I/Os on the die increase, the corresponding pad pitches decrease. This leads to the use of finer wire in wire bond applications that eventually limit electrical performance. TAB leads, at these pitches, using fine pitch, peripheral leaded die can achieve the desired electrical performance. These designs, however, must be carefully tailored to the electrical environment and may necessitate flip TAB configurations with short TAB leads or two metal layer tapes to attain the desired electrical goals. The result can be a balancing of high performance needs versus cost requirements since the use of short TAB leads goes against the current TAB standards and would require expensive custom tape designs and tooling. Likewise, two metal layer TAB tape is more costly than single metal layer tape.

Thermal management issues also affect the connection decision. While conventional TAB configurations conduct heat through the substrate, as with wire bond, flip TAB designs require novel heat removal designs. Flip TAB designs also incorporate shorter lead lengths that help electrical performance and increase density. Again, the extra cost must be balanced against performance issues.

Reliability studies of TAB assemblies indicate failures unique to TAB, while other concerns are common to all connection methods. These unique failures include TAB lead and solder joint failure which is controlled or eliminated with proper TAB design and material choices.

Testability, and burn-in and repair are important attributes of TAB, especially for the MCM user, and constitute more advantages of TAB over the other connection methods. An inner lead bonded die can be tested or burned-in prior to committing that die to an expensive module, enabling the user to screen bad die out of the assembly flow. Die that do fail after attachment to the MCM can be removed and replaced at the OLB level, depending on the metallurgy. This process is difficult, if not impossible with wire bonded devices and is not a trivial process with flip chip.

Flip chip applications are natural and, probably, inevitable extensions of packaging connection assembly. While the flip chip eliminates the tape frame and custom tooling required by TAB, bumping still is required. Presently, the infrastructure and assembly know-how for this technology is immature to the world at large. This makes TAB, especially flip TAB, a natural alternative. When the flip chip infrastructure has matured sufficiently to handle the advanced assembly needs of the MCM user, knowledge gained from flip TAB may be applied to future flip chip applications.

9.4.12 Summary

TAB assembly for MCM applications offers the multichip designer an opportunity to utilize high lead count, fine pitch semiconductors and create densely packed modules with superior electrical performance. The inner lead bonded die also can be tested or burned-in prior to bonding to the package, enabling the user to screen out bad die prior to their commitment to an expensive MCM. Die also can be removed at the OLB level and replaced with new die, preventing the loss of an expensive assembly.

Materials and equipment for implementing a TAB process are available, allowing the use of TAB in design and manufacturing.

Tape material and assembly method options have presented the manufacturing engineer with a variety of process options, enabling a tailoring of designs for different uses and environments. Reliability data for high performance TAB applications, a relatively young practice, has begun to filter through the electronic packaging industry.

The choice of assembly methods for multichip applications, TAB, wire bond or flip chip, depends on the balance between cost and the desired performance. New methods and materials allow TAB to reduce its cost structure and compete, favorably, with the other assembly technologies.

9.5 FLIP CHIP CONNECTION TECHNOLOGY

Chee C. Wong

9.5.1 Introduction

Flip chip connection technology as a first level chip to package connection option traditionally is regarded as being synonymous with the Controlled Collapse Chip Connection (C4) process pioneered by IBM more than 20 years ago. The C4 process has set the highest record in I/O density, chip packing density and electrical performance, and establishes the industrial benchmark in field reliability. Details of the C4 process are presented in Section 9.6.

This section presents the flip chip connection technology as a generic technology with the C4 process as a subset example of one particular application. Emphasis is placed on the concepts behind the design of the flip chip connection configuration, the material options for the connection medium, the processing options in implementing flip chip and the cost and manufacturability issues in the context of inherent process limitations and existing infrastructure. The objective here is to introduce an overall perspective on flip chip connection technology beyond the C4 process to enable a judicious comparison to be made between the several flip chip variants and between flip chip connection and other chip connection options.

This is a section on concepts rather than details. It offers an organized format of questions to be asked and provides a framework for answering those questions on an individual basis. Hopefully, it will stimulate the reader to evaluate the applicability of flip chip technology for his or her product goals.

9.5.2 The Basics

Definitions

Flip chip is defined by the schematic in Figure 9-25 which shows a bare IC device flipped upside down with its active area or I/O side attached to a substrate via a connecting medium. In this generic description, the device may be a silicon microelectronic IC or any other monolithically integrated active functional block. The substrate in Figure 9-25 may be any of the MCM substrates providing an interconnection network between the flipped active device and other active, or even passive devices. The connecting medium may be any suitable interface serving the various needs of the matchmaking between the flipped device and the underlying structure. Each member of this flip chip ensemble is examined in later sections.

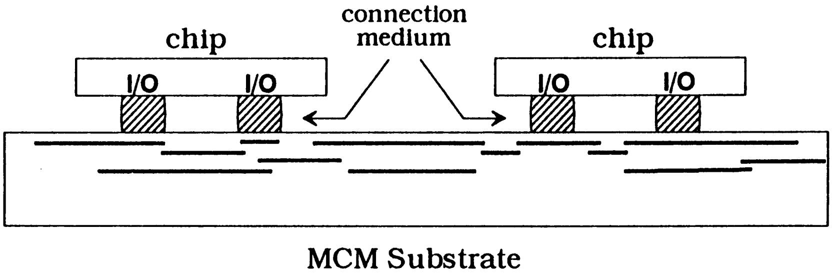


Figure 9-25 Flip chip configuration consisting of chip, connection medium, and substrate.

Why Flip a Chip?

What is the greatest distinction of the flip chip configuration? *Why a flip chip?* Flip chip is the only connection configuration that allows assembled active chips to approach the form in which they were originally created, namely, the form of a wafer. This is an advantage because it provides superior electrical and thermal performance.

To understand this point one must realize that the goal of any packaging scheme is to allow each chip to perform at its peak and to allow the system as a whole to take full advantage of the peak performance of each individual component. Circuit speed on the bare chip level is the highest speed achievable. As soon as the chip leaves its original wafer form and enters the first level of packaging, its performance begins to suffer. Why then don't we build entire systems or subsystems on a single wafer? This approach, called "wafer scale integration," has met with little success. The problem is poor yield. While most parts of the system may function as designed, functional failure of a single part can "doom" the entire system. Hence, the current approach in hybrid microelectronics follows the modular concept, namely, to break a big system into smaller systems, build many small systems on a wafer, isolate and package the functioning small systems (IC chips) and reassemble them back into a big system. Figure 9-26 shows a schematic comparison between monolithic wafer scale integration and two modular alternatives, an unpackaged flip chip version and a packaged surface mount version.

Since wafer scale integration has proven to be impractical because of yield issues, the next best thing in terms of performance is to build a separate interconnection structure of commensurate interconnection density (such as an MCM) and to assemble the unpackaged IC chips back onto the MCM substrate so as to resemble their original wafer form as closely as possible. This translates

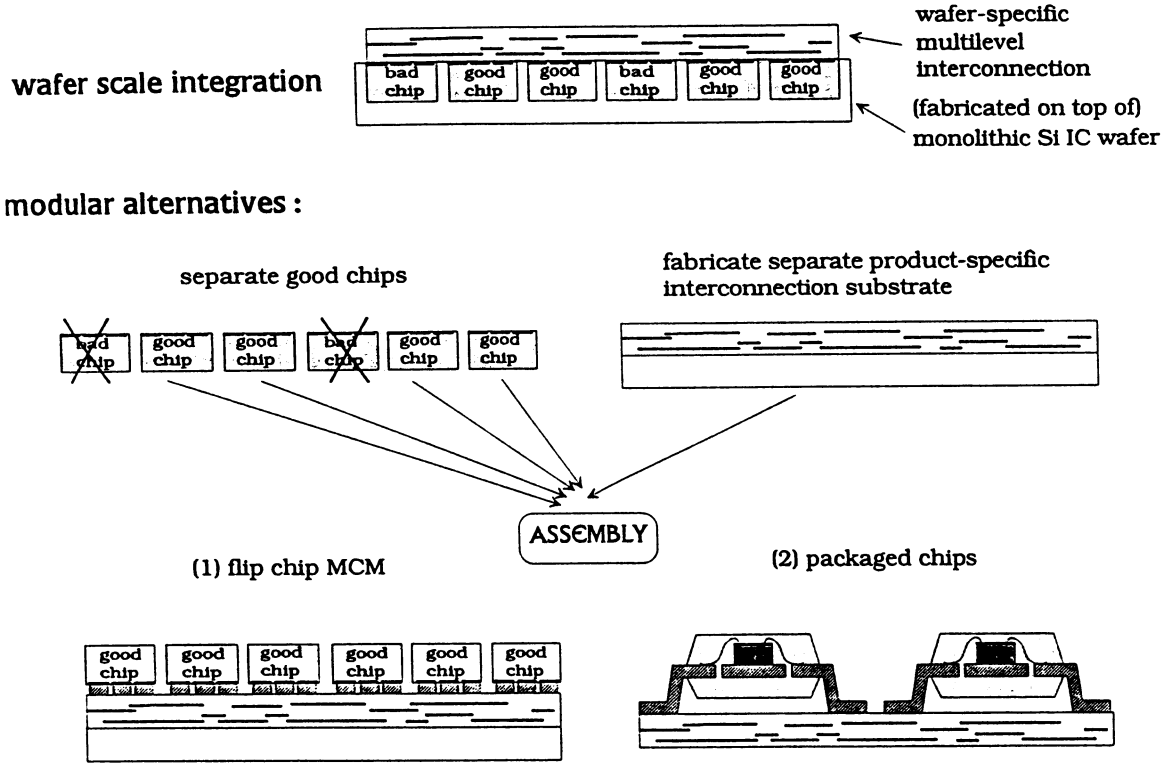


Figure 9-26 Comparison of wafer scale integration and modular approaches to multichip packaging, highlighting the packing density achievable by flip chip MCMs.

into a requirement for a connection technique which permits the closest possible chip proximity and a connection medium whose dimensions are contained within the area of the chip. Also, the connection medium should be amenable to short connection lengths to minimize electrical parasitics. By flipping a chip and directly attaching its I/Os via a connecting bump of controllable height onto the substrate as shown in Figure 9-25, the maximum footprint that the chip requires is that of its own. No fanout is required. This constitutes the distinct advantage of flipping a chip. As shown in Figure 9-26, the packing density of a flip chip MCM could, in principle, approach that achieved in wafer scale integration.

Members of the Ensemble

The characteristics of the individual members in a flip chip ensemble are presented as follows:

1. **IC Chips**. Most silicon IC chips presently are designed for perimeter wire bonding. The I/O pads, on the order of 4 mils, are finished with an Al metallization and surrounded by a passivating layer of dielectric. The degree of perfection of this passivation layer is inadequate in providing mechanical and environmental protection for the chip. The number of I/O pads on IC chips could range from several tens to several hundreds. There is no industrial standard in the spatial arrangement of I/O pads. Silicon chips generally are not readily available in bare wafer form. Instead, they are readily available only in die form.

2. **Substrate**. The substrates for MCMs could be in the form of cofired or thin film ceramics, thin film silicon, printed boards or flex circuits. These various structures are reviewed in Chapters 1, 5, 6 and 7. MCM substrate design, being considerably less mature compared to chip design, is more tolerant of the needs of chip connection. To make full use of any chosen MCM platform, design of the substrate and design of chip connection need to be carried out in parallel.

3. **Connection Medium**. The connecting medium couples the chips to the substrate to form a functional and reliable MCM. TAB and wire bonding techniques achieve connection using leads which fanout from the chip I/O to the corresponding pad on the substrate. Flip chip bonding achieves electrical and thermal connectivity using bumps (before joining), called joints after joining. These joints provide mechanical support for the flipped chip on the substrate. The C4 process uses bumps made of solder or solder-coated copper balls. Organic conductors are new candidates for the connection medium.

Note that out of the several functions of the IC chip package discussed in Chapter 1, flip chip connection has fulfilled only the functions of electrical and thermal connection. Mechanical protection generally is delayed until the module packaging level. If the module level package is not considered adequate for environmental protection, then chip encapsulation techniques are used to protect the chip from operating ambients. The issue of chip testability in its bare chip or wafer form prior to module level assembly is considered a major inadequacy of flip chip technology.

Mechanical support of the chip itself, which was not an issue in the case of unflipped IC chips die bonded onto the package, is now provided by the joints and introduces an important new variable in fatigue-related reliability. Hence, to complete the picture of using flip chip as a connection technique for MCMs, at least three new members have to be added to the ensemble depicted in Figure 9-25: a chip testing capability for the bare or bumped die prior to assembly, an encapsulation technology after assembly [55], and proper designs for minimizing susceptibility to thermal fatigue [56]-[57].

Why flip a chip revisited

As mentioned, flipping a chip onto an MCM could achieve wafer level packing density by eliminating fanout. There is another feature unique to having the active side of the chip face the top of the interconnecting substrate. Since the I/O pads on the chip also are fabricated on the active side, the layout of these pads easily can be expanded into an array covering the entire inner area of the chip, rather than being confined onto the perimeter. Area arrays offer a way of increasing I/O density without taxing other technologies for a finer I/O pitch. For example, for a chip size of 5 mm and a constant I/O pad spacing of 100 μm, a perimeter array could accommodate about 200 I/Os while an area array could accommodate about 2000 I/Os, a tenfold increase. Only the flip chip configuration provides the ability to achieve higher I/O density without decreasing I/O pitch.

By having the active side down, flip chip bonding also offers the shortest possible leads with the lowest inductance, maximizing the operating frequency. This consideration alone could justify the usage of flip chips in high performance systems. The actual length of the lead, or, in this case, the standoff height of the joint, could be controlled by any of the flip chip techniques in use, and is usually chosen to optimize other criteria rather than being predicated by the geometry of the unflipped chip. Joint height is usually designed for better fatigue endurance and chip underside cleaning.

Flip chip assembly is inherently a batch bonding process. This contrasts with wire bonding which proceeds serially. The throughput advantage of batch bonding is obvious at high I/O densities. Batch bonding also is offered in beam lead bonding and gang bonding versions of TAB. While TAB could sidestep the high tooling cost of TAB gang bonding techniques by temporarily employing single point bonding techniques, the flip chip technique has to confront the assembly issue of high speed batch bonding directly. Once the initial barriers of developing versatile and cost effective flip chip batch bonding machines are overcome, batch bonding could become the bonding method of choice for most current chip sizes. As maximum chip size increases, tooling for batch bonding becomes more difficult because of issues in planarity, heat distribution and wider disparity in chip sizes. At that stage, the benefits of batch bonding can be realized only by investing in more costly tooling.

The robustness of flip chip connections has set the reliability benchmark in the connection industry. The absence of leads makes the IC chip rugged and easier to handle. Issues of thermal fatigue have so far been adequately addressed by proper joint design (see Section 9.6). As chip sizes increase, issues of fatigue life again will dominate the question of joint mechanical integrity. This point argues in favor of using silicon MCM substrates for flipped silicon chips to circumvent the detrimental effects of coefficient of thermal expansion (CTE) mismatches and the resultant fatigue phenomenon.

9.5.3 Connection Medium (I): Solder Bumps

The connection medium between the flipped silicon chips and the underlying MCM substrate is the key in realizing any flip chip technology. This is where the issues of manufacturability and cost are defined. The material and fabrication method chosen for the connection medium influence chip and substrate reliability, yield and throughput in assembly.

The two connection media currently in use for flip chip MCMs are solder and organic conductors. This section presents an overview of solder bumping (bumping using organic conductors is deferred to another section), with emphasis on the comparative strengths of different techniques and the basic roles of various materials used in the formation of the bump. One of the techniques mentioned is the C4 process, treated in greater detail in Section 9.6.

Bump Location

The three possibilities of locating the solder bump on the substrate, on the chip, or both, are shown in Figure 9-27. The decision is based on the following:

- Which is the easiest to do
- Which side is likely to experience less impact going through a solder bumping process
- Which is better for handling and storage

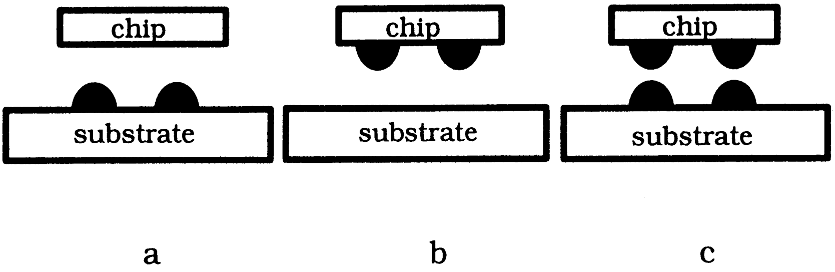


Figure 9-27 Three possibilities of locating bumps connecting chip to substrate.

- Which is the easier to assemble
- Which is the easier to repair and
- Which is the most cost effective in terms of overall yield.

To answer these questions, consider the members chosen for the flip chip ensemble. Silicon chips are produced in wafer form, regardless of type and function. MCM substrates could range from silicon wafers (making them equivalent to silicon chips in terms of processing) to flexible copper polyimide circuits. The equipment for handling and processing these various substrates are equally varied, each specializing in optimized processing for that particular substrate. If different substrates for different applications are envisioned, it would be practical to concentrate on placing bumps on silicon wafers only, since the wafer form of silicon chips remains invariant. The placement of solder bumps onto chip wafers turns out to be advantageous in areas such as testability and repair also. Details of the rework process for solder bumped chips are discussed in Section 9.6.

There is also an overall yield advantage for placing bumps on chip wafers. Any silicon chip on any MCM substrate is always smaller than the substrate. For a given unit area being processed, more chips are produced than substrates. Let's take an example of a five chip set MCM of 1 inch square. Here, in the same inch square area being solder bumped for one MCM substrate, five chips could be processed. Assume also that the defect level is such that one bad bump (short or missing) is produced per inch square on average. This one bad bump on the substrate would render the entire MCM useless, whereas it would only disable one of the five chips. A healthy chip from a neighboring group could be substituted in place of the bad chip to still produce a good MCM. The moral here is that, for a given defect level, smaller objects are more tolerant of defects than larger objects in terms of overall yield. Since chips are always smaller than their mating substrates, it is justified to place bumps on the chip wafer.

Placing the solder bump onto the substrate (Figure 9-27a) is justified in the case where chip wafers cannot survive the temperature or the physical and chemical environment of a solder bumping process. For example, an aggressive backsputtering step (for via cleaning) in a solder bumping process may damage CMOS chips sensitive to strong radiation. Bumping substrates also is justified if chips are not available in wafer form. Placing bumps on both the chips and the substrates adds more cost; however, this scheme becomes mandatory if the solder height limitation inherent in any single bump architecture is insufficient to meet joint height specifications.

Bump Shape

Schematics of a solder bump before and after reflow are shown in Figures 9-28a and 9-28b, respectively. Reflow is a heating process which takes the solder bump through a solid to liquid to solid transition, allowing the solder to consolidate its bonding with the connecting interface. While the geometry of the bump in Figure 9-28a could differ depending on the processing technique, the reflowed bump shape in Figure 9-28b is universal, governed solely by the forces of surface tension, gravity and the tendency of the liquid solder (during reflow) to assume a shape of minimum surface energy. For small bumps where the effect of gravity can be neglected, the equilibrium shape is a spherical segment.

The surface onto which the solder bump is fabricated consists of two distinct areas in the vicinity of the bump: a wettable and a non-wettable area. The wettable area is the bonding interface to the solder. Usually the chip's I/O pad with an Al finish is located directly below the wettable area for electrical

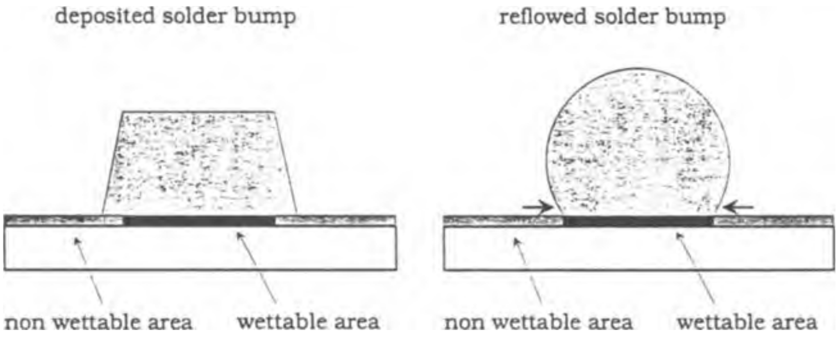


Figure 9-28 Schematic of a solder bump (a) before and (b) after reflow.

connection, although dummy bumps whose wettable areas are not connected to I/O pads also could be fabricated for reasons of mechanical robustness or improved thermal performance. The non-wettable area is necessary to confine the solder within its allowable area, thus controlling the final height of the bump for a given volume of solder.

The shape transformation from Figure 9-28a to Figure 9-28b is determined entirely by the volume of the deposited material and the area of the wettable region. The original area of the solder deposit may or may not correspond to the area of the wettable base, depending on the technique and the design. The final footprint of the reflowed bump, however, corresponds exactly to the wettable area, assuming complete wetting during reflow. In other words, if the solder deposit area is smaller than the wettable area, then solder spreads outward during reflow; conversely, the solder footprint shrinks back. This fact, coupled with the knowledge that the equilibrium shape of a reflowed solder bump is that of a spherical segment, allows us to predict the final shape and height of the reflowed bump.

The plot in Figure 9-29 shows the reflowed bump height as a function of the height of the solder deposit. The plot is delineated into three regions, corresponding to different shapes of the spherical segment. Figure 9-29b shows a perfect hemisphere, while Figures 9-29a and 9-29c show spherical segments which are smaller (sub-hemisphere) and larger (super-hemisphere) than the hemisphere, respectively. The line of unity slope in the plot is given as a yardstick to distinguish the region where the reflowed height is larger than the deposited height (above this line) and the region where the opposite is true (below this line). The reflowed height of the solder bump has an approximately cube root dependence on the height of the solder deposit. This curve crosses the straight line in the region where the shape of the bump is a super-hemisphere. This means that as long as the bump shape is that of a hemisphere or smaller, the deposited volume of solder is used efficiently in building the height of the bump. When the shape crosses into the regime of super-hemisphere, a further increase in the deposited volume makes little contribution to the height of the reflowed bump, most of the material going toward enlarging the waist of the bump instead. This can be seen in the shape of the super-hemisphere where the largest cross sectional area of the bump is no longer at the base; rather, it has migrated toward the middle. This has an effect of creating a reentrant corner at the point where the solder bump meets the wettable area. Similar considerations regarding solder joint geometry have been presented by Goldman [57], for which the same conclusions apply. Unfortunately, such reentrant corners would concentrate strain at the base of the joint, rather than distributing it throughout the solder volume. In other words, the solder joint would not be used efficiently as a mechanical support. The point of this discussion is that geometrical design

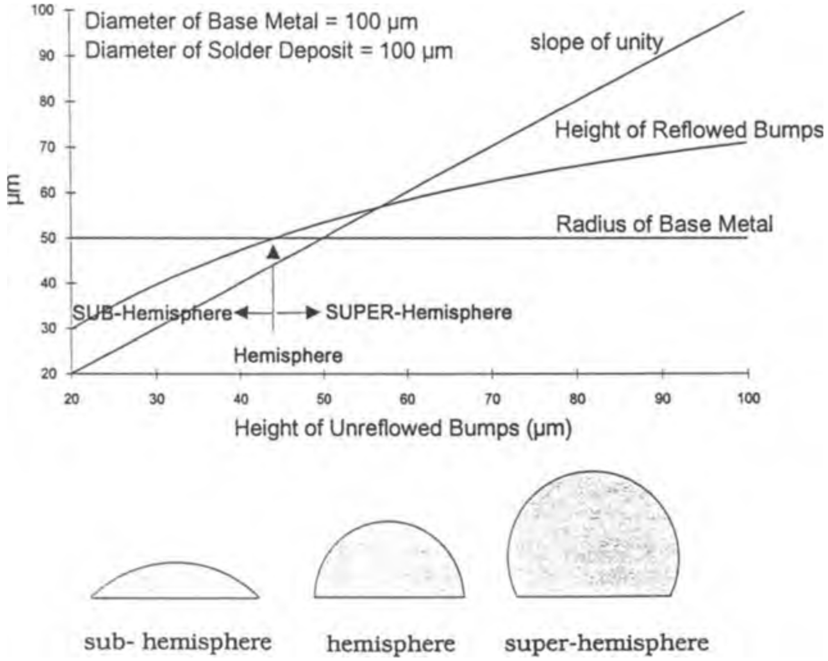


Figure 9-29 Plot of reflowed solder bump height as a function of unreflowed bump height, using the spherical segment approximation. Regions corresponding to a bump shape of sub-hemisphere, hemisphere and super-hemisphere are delineated.

is an important factor in optimizing processing efficiency as well as in mechanical integrity. The influence of joint shape in fatigue life is further discussed in Section 9.6.

Solder Bump Material Choices

Three classes of materials are of interest in making a flip chip solder bump: the solder itself, the wettable area (base metals) and the non-wettable area (solder dam). Most solders currently used are of the lead-tin (Pb-Sn) family, although the momentum toward limiting usage of Pb may necessitate consideration of other families of solders. Pb-Sn solder materials and their characteristics are reviewed by Wassink [58]. The exact composition chosen and any additives thereof depend on the desired reflow temperature, fatigue performance, corrosion susceptibility and ease of fabrication. The use of solders with 95% Pb and 5% Sn in the C4 process for many years has generated the most extensive research and development and field reliability database for this family of solders.

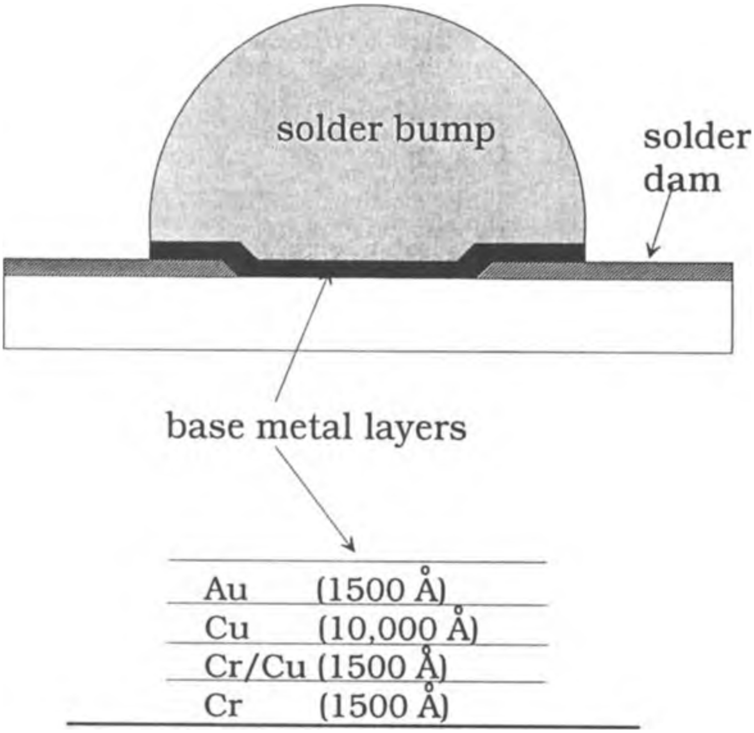


Figure 9-30 Cross section of a solder bump showing the multilevel base metal layers.

The base metallurgy for a 95Pb/5Sn solder bump fabricated using the C4 process is shown in Figure 9-30. Other metallurgies are reviewed elsewhere [59]. Invariably, the base metallurgy has to provide three functions: adhesion to the underlying surface, wettability by the solder and a barrier between the solder and whatever lies underneath. The first function is typically filled by one of the classical glue layers of titanium (Ti) or chromium (Cr). These metals stick well to other metals and most dielectrics. This point is important in chip wafer bumping because, as shown in Figure 9-30, the base metal footprint typically extends beyond the area of the I/O pad to provide a sealing function to the last remaining unpassivated area of the chip. To do this well, glue layers must have excellent adhesion to metals (I/O pads) as well as to dielectrics (solder dams). The Ti or Cr adhesion layers also can provide the barrier function; however, they are not solderable.

Solderable metals such as copper (Cu), nickel (Ni) and silver (Ag) have been studied and used in other applications involving 95Pb/5Sn solders. They become natural choices for wettable base metal layers in flip chips with one important distinction - the Cu in a base metal multilayer structure is much thinner than the Cu on a printed wiring board. Thin films of these materials may be completely consumed or exhausted during multiple reflows, resulting in dewetting of the solder from the base. Although a thin film of Cu is still being used as a solderable layer, an additional layer of phased, or codeposited, Cr/Cu is now interposed between the Cu and the underlying Cr. This mixed layer performs a dual function: part of it is a barrier metal (Cr) and part of it is a solderable metal (Cu). The result is that, at the proper compositional ratios, the mixed layer exhibits sufficient wettability to prevent non-wetting but limits chemical interactions to prevent material exhaustion and dewetting during multiple reflows.

The final top layer of the base metals shown in Figure 9-30 is a thin layer of gold (Au). The nonoxidative properties of Au and its excellent solderability preserve the wetting function of the base metals throughout material handling and storage.

The reflow process initiates aggressive chemical interactions between the liquid solder and the base metals. Au is dissolved almost instantaneously. The dissolution rate of the common solderable materials has been reported by Bader [60]. The chemical interactions add a new member to the multilayer structure - Sn-based intermetallics. The brittleness of these intermetallic compounds has been shown in some cases to introduce a weak link into the structure. Reviews on the formation and effects of intermetallics can be found elsewhere [61]-[63].

Lastly, the role of the solder dam can be filled by most dielectrics which do not interact with molten solder and which do not degrade at the reflow temperature (about 330°C for 95Pb/5Sn). The finishing passivation layer on the surface of a chip generally is adequate as a solder dam, although an additional protection layer such as polyimide can be used.

Although the above discussion makes use of material examples taken from chip wafer solder bumping, the roles of the different layers remain unchanged in the case of substrate solder bumping. If the MCM substrate were a silicon wafer, then solder bumping materials and procedure could be identical to that of the chip wafer. Otherwise, in the case of other substrates such as ceramics, wettable base layers and solder dams can be chosen from materials and techniques mature for the processing of that particular class of substrates. The reader is referred to Tummala's work [G1] for a review of various substrate pad structures.

Bump Formation

This section discusses several common techniques currently used in solder bumping. Understanding the functions of individual steps is critical in the design of a process sequence suitable to the needs and expectations of each application. In essence, process sequences are not unique; many different sequences may achieve the same goal.

Steps in making solder bumps consist of:

- Deposition of base metal layers,
- Deposition of solder
- Patterning of the base metals and solder layers.

The goal here is to create a structure similar to the one shown in Figure 9-29.

Deposition of Base Metal Layers. Such deposition is usually accomplished by sputtering, evaporation, electroplating or a combination of these methods. Deposition can be done under two different conditions: blanket deposition, which covers the entire surface uniformly with the deposit or patterned deposition which deposits the material through a masking layer containing the pattern to be transferred. The former option requires only that the object (chip wafer or substrate) be compatible with the deposition environment, while the latter option extends the requirement to the masking material. For sputtering and evaporation techniques, the environment is one of high vacuum and heat; for electroplating, the environment is reactive chemical solutions. In the case of electroplating, a conductive plating base is required to initiate and sustain the plating reaction. This plating base often is deposited using either sputtering or evaporation. Thus, one of the two vacuum techniques is a prerequisite to base metal deposition unless electroless plating methods could be developed for the materials discussed in the previous section.

Both sputtering and evaporation (electron beam or thermal) have been used extensively in the thin film industry. They are well characterized, well controlled and well supported by mature and sophisticated equipment. These processes also can be automated. The materials suitable as base metal layers for solder bumping fall within the range of capability of this equipment. Their major distinction lies in the end requirement for uniformity. The surface of uniformity for evaporation is that of a sphere. The size of the sphere determines the deposition rate, the number of wafers that can be processed together and the required size of the vacuum chamber. For sputtering, the surface of uniformity is roughly that of a plane of area to the target. Wafers receive deposition serially, supported on a rotating carousel. As wafer size increases, evaporation techniques require larger vacuum chambers whereas sputtering requires larger targets and larger sputtering guns. In both cases, substantial increases in cost would be incurred. Beyond this distinction, both techniques are well suited to the deposition of base metal layers, including the codeposited layer discussed in the previous section.

Electroplating of base metals has been used in the deposition of the solderable layer (Cu, Ni, for example) and the Au finish [64]. The ability to electroplate a mixed layer (Cr and Cu, for example) of a microstructure similar to that attained by evaporation or sputtering has yet to be demonstrated. The wide use of electroplating in industries, such as circuit board manufacturing, has resulted in a mature technology for coating large surfaces. In applications requiring large area processing with large feature sizes not requiring precise thickness control, electroplating can be the most cost effective.

Deposition of Solder for Flip Chips. Deposition could be carried out using thermal evaporation or electroplating. Sputtering is not suitable for low melting point materials such as solder. Screen printing of solder paste, although widely used in the surface mount industry, does not to meet the fine pitch requirements of flip chips.

Evaporation is the more common technique for solder deposition. It is the technique used in the C4 process. The primary difference between evaporation of base metals and evaporation of solder is the thickness of deposit. The evaporation of base metals could follow standard IC processing of thin films (1 μm), but the evaporation of much thicker solder deposits (tens of μm) requires special tooling. Pounds of material are needed for the charge, requiring special crucibles and special power supplies. The long periods of deposition needed to achieve large thicknesses tend to generate much higher temperatures on the wafer surface which, in an uncontrolled situation, could lead to melting of the solder coating. If a patterned deposition technique is used, care must be exercised to ensure that the masking material does not deform under these temperatures. These complexities notwithstanding, the evaporation process produces solder coatings of high purity, high uniformity and consistent composition at high throughput.

Solder can be deposited using an alloy or two elemental charges (for Pb-Sn). In an alloy charge, the different vapor pressures of the two elements cause the Pb layer to be deposited first, followed by that of Sn. Elemental charges could reverse the position of these two layers if desired. An alloy charge is preferable because it occupies the central position in the chamber, the optimized location for uniformity. Two elemental charges cannot occupy the same ideal location.

Electroplating of solder has been reported in [59] and [65]. There are two primary concerns in the plating of solders: bath chemistry and electrode design. Additional factors are: the need to ensure efficient mass transport inside the bath, the maintenance of bath composition and the control of electrical current density. These factors combined determine the purity and uniformity of the solder deposits. Since equipment for electroplating are not as automated as those used in vacuum deposition, the human error factor becomes more pronounced. Electroplating of solder often is used in a patterned deposition, rather than a blanket deposition scheme. In this case, the masking material has to be evaluated with regard to each different plating chemistry being considered.

Patterning. Patterning is the key to forming solder bumps of the desired geometry. In general, patterning techniques fall into two categories: additive (patterned deposition followed by liftoff) or subtractive (blanket deposition followed by patterned etching), as discussed in Chapter 2 and 7. Liftoff techniques have to contend with residues while etching techniques have to contend with unwanted chemical attack. In either case, a suitable masking layer has to be inserted at some point within the process sequence, and removed at some other point. The different ways to insert and remove one or more masking layer(s) among the two prerequisite deposition steps (base metal and solder) constitute the different process sequences unique to each flip chip solder bumping technology. Only two cases are presented as examples.

The C4 process uses a double liftoff technique for the base metal and the solder. The masking layer in this case is a metal mask (physically distinct from a glass mask used for patterning photoresists) with openings corresponding to the I/O patterns of the chip wafer. The mask is aligned with the wafer and both are mechanically held together in a fixture. After base metal evaporation, the mask is removed (first liftoff), a second mask with larger openings put in place, and solder is deposited through the second mask (second liftoff). This is a purely subtractive process; no etching is involved. The second liftoff step is not necessary if the design does not call for a solder footprint larger than the base metal footprint. A polymeric mask, such as dry film resist used in the circuit board industry can be substituted for the metal mask and the same sequence applied. Polymeric masks have the virtue of conformal coating to the wafer surface, avoiding the misregistration issues in mechanical fixturing of metal masks. On the other hand, metal masks do not introduce residues since they were never adhered onto the surface. In terms of resolution for fine pitch I/O, photosensitive polymeric masks are more broadly applicable since they conform to standard photolithographic practices in IC manufacturing.

The baseline electroplating process discussed in [65] uses liftoff for solder and etch patterning for base metals. The base metals are blanket deposited so that they would form the plating base for solder. A dry film resist is patterned onto the base metals before electroplating solder through the via openings. Following removal of the resist (liftoff for solder) the base metals are chemically etched using the solder bumps as a mask (etch patterning). In this case, the etching chemistry has to be chosen such that etching is preferentially done on the base metals, leaving the solder intact.

One other issue critical to patterning using liftoff is that of via cleanliness. This usually is accomplished by chemical etching or a physical means such as backsputtering or ion cleaning. Failure to clean vias could lead to nonplating of solder in the case of electroplating or non-adhesion of base layers in the case of the C4 process.

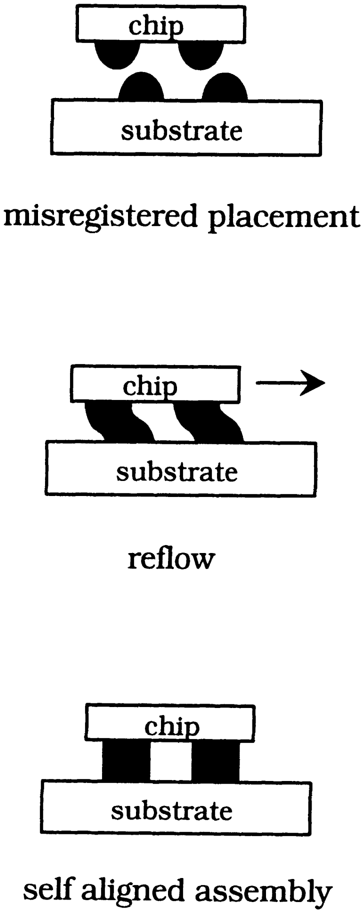
Attachment

The final step in creating solder connections for flip chip MCMs is reflow and assembly. As mentioned before, reflow allows the molten solder to form chemical bonds with the base metal layers of both the chip and the substrate. The heating is done either in a reducing atmosphere such as H_2 or in the presence of fluxes to remove oxides. A common practice is to reflow solder bumps prior to assembly to consolidate the solder into its equilibrium shape and to confirm its bonding with the underlying interface, whether it be a chip or an MCM substrate. After assembling the flip chip onto the substrate, a second reflow is carried out to turn solder bumps into solder joints. Following cleaning (if flux has been used), the flip chip solder MCM is ready for the next level of testing or packaging.

Common problems detected at the reflow stage are solder dewetting from the base and the formation of solder voids. Dewetting is a result of improper design or processing of base metals. A solder bump with this problem is useless. Voids, on the other hand, arise from a variety of sources, and come in a variety of sizes. Voids due to trapped fluxes or impurities can create huge hollow spaces within the bump, making it mechanically unsound. Shrinkage voids due to natural material freezing phenomena can be small compared to the dimension of the bump and have not been shown to cause deleterious effects, provided they do not coalesce to form a big void.

An important benefit derived during the second reflow stage due to the nature of molten solder is self alignment - the ability of the chip to center itself onto the mating footprint on the substrate regardless of placement misregistration. This is illustrated in Figure 9-31. The solder bump seeks its equilibrium shape during the first reflow, and the solder joint during the second reflow. The tendency of the joints to assume their equilibrium shapes provides the driving force to center the bump pads footprint form the chip onto the corresponding substrate. This eases the requirement for high assembly accuracy. An ideal assembly of flip chip MCMs approaches the standard set by surface mount assembly, since both processes involve batch bonding using a solder connection medium. To approach this ideal for flip chip MCMs, some considerations are:

1. Heat tacking (application of heat during placement in addition to pressure) is required to mounting flip chip solder bumps. Tacky flux (flux which glues the chip onto the substrate in preparation for reflow, discussed in Section 9.6) also may be used. Both procedures assist in the self alignment process. Pick-up tools currently used in pick and



misregistered placement

reflow

self aligned assembly

Figure 9-31 Self alignment of solder bumped components during reflow.

place machines incorporate heating and heat distribution elements to ensure uniform heating of the chip. Alternatively, heating may need to be applied to the entire substrate. Depending on the choice of base metal and solder materials, heat may have to be applied to both.

2. Solder pads on MCMs are at least a factor of two smaller than those on surface mount boards. Even though self alignment eases the tolerance on registration, placement accuracy must be sufficient to ensure that all

of the solder pads on one surface at least touch the appropriate features on the mating surface.

3. Unlike chip carriers which have a number of standard sizes, there is no emergent standard on IC chip sizes, MCM substrate types or MCM substrate sizes. A "standard" MCM assembly tool for flip chips needs to handle disparate IC sizes and substrate sizes and types. Although custom tooling always is available, standardized tooling is the key to lowering costs.

The infrastructure for assembling flip chip solder bumped MCMs currently lags that available for fabricating solder bumps. Solder bumping techniques, despite various methods, follow well established procedures in processing, while the assembly of MCMs is particular to each technology and its attendant chip connection method. Given sufficient demand, flip chip MCM assembly could approach the cost and performance standards of surface mount assembly.

9.5.4 Connection Medium (II): Conductive Polymers

Electrically conductive epoxies are used extensively as an alternative to eutectic die bonding, as discussed in Section 9.2. The conductivity in these polymers is isotropic. Such polymers are suited only for making a single connection such as ON the backside of a die to a package leadframe. An extension of this technology has led to the development of conductive polymers whose conductivity is anisotropic. These polymers conduct current preferentially in the z-direction (normal to the plane of the polymer film) while maintaining electrical isolation in the xy-plane of the film. This characteristic qualifies films such as a multi-I/O connection medium. They are referred to as anisotropic conductive adhesive films (ACAF) [66]-[69].

Most conductive polymers are formed by dispersing metallic particles into the polymer film so that current is conducted through the polymer via the bridging of the particles. ACAFs are prepared by controlling the dispersion of conductive particles, placing a sufficient concentration to enable conduction in the z-direction only. This is accomplished using the single particle bridging concept illustrated in Figure 9-32 which shows a schematic of a bare die flip bonded onto the substrate using ACAF. The metallic particles commonly used are made of Ag, Ni or Au. By controlling particle size, ACAFs successfully connect chips with 4 mil I/O pads [66].

An ACAF has one important difference from solder bumps - it does not need to be patterned and, thus, becomes more versatile. Only a blanket

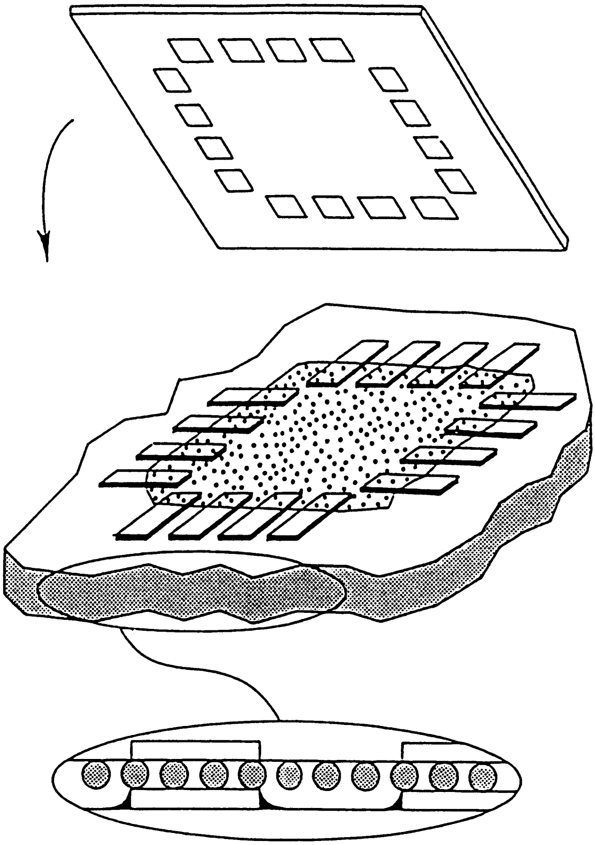


Figure 9-32 Schematic of chip connection using an ACAF material. Inset shows particle bridging between conductors. (Reprinted from [15], courtesy of A. M. Lyons)

application of this material onto the appropriate I/O surface (preferably Au) is needed. ACAFs can, in principle, be applied onto any chip wafer and any MCM substrate. Assembly consists of aligning and placing the chip onto the substrate with concurrent application of heat and pressure to ensure good adhesion and curing of the polymer. Unlike solder bumps assembly, a reflow step is not present. Because of the absence of reflow, the benefits of self alignment are not available; meaning that ACAF-type flip chip assembly has to rely on placement machinery with higher accuracy than those needed for flip chip solder bumps.

Electrically, ACAFs show negligible inductance since the connection length is on the order of the diameter of a conductive particle, which ranges from 0.5 - 30 μm [66]. The performance of this material at high frequencies is under investigation.

The usage of ACAFs in flip chip technology is relatively immature compared to that of solder bumps. Formulations of ACAF materials need to be standardized and individually tested. Performance, reliability, yield and throughput data have yet to be collected. It is likely that high end systems based on rigid substrates will continue to use solder bumping, while flex circuits and other compliant substrates, will utilize ACAFs. Also, the versatility of ACAFs could very well extend into other hierarchies of packaging beyond the MCM level.

9.5.5 The Whole Picture

Flip chips have a unique geometry, allowing them to be packed at a density higher than that of any other chip connection technology. This feature is best exploited when matched with MCM substrates with fine pitch features. Thin film MCM substrates are therefore the best candidates for utilizing this particular strength of flip chip technology. Thin film substrates, because of their high electrical performance, also are well matched to the low inductance, high performance connections of flip chips. In terms of reliability, the susceptibility of solder joints to fatigue due to CTE mismatch argues for the use of silicon MCM substrates. Other substrates must be CTE-matched to silicon before large IC chips can be used in high reliability flip chip MCM products.

In terms of cost, the expense of solder bumping has to be viewed on a cost per connection basis since solder bumping cost is based on wafers (for chip wafer bumping) with different size chips and variant numbers of I/O. The total number of I/Os per wafer is a product of the number of I/Os per chip and the number of chips per wafer. For 4" wafers, at small chip sizes (< 5 mm) and high chip I/O counts (> 100), the cost per connection for solder bumping becomes extremely competitive. The impending increase in I/O density certainly favors the use of solder bumped flip chips. This advantage is maximized for chips designed with area array pads since flip chips can accommodate higher I/O density without incurring the risks of a finer pitch technology.

A similar viewpoint on cost also can be adopted for the issue of assembly because of batch bonding which allows all connections to be made at once, regardless of number. The forgiving nature of self alignment enables us to project the future status of flip chip assembly as close to the cost and throughput standard of surface mount assembly. Thus, low cost flip chip assembly can be envisioned, though not currently realized.

Given the unquestioned performance and foreseeable low cost of flip chip attachment, the only issue that remains is that of availability. Besides IBMs C4, other companies are beginning to install manufacturing capabilities based on flip chip solder bumping, notably, Intel, Motorola and AT&T. As higher levels of IC integration demand higher performance levels in chip packaging, expect to see increasing activity in flip chip bumping manufacturing.

A final view of the whole picture extends beyond the role of flip chip in hybrid circuit technology to the role of hybrid circuits in electronic systems in general. Hybrid circuit functions can be replaced by advanced ICs. However, these advanced ICs will require advanced level hybrid packaging to connect them to other advanced ICs to build yet more powerful systems. The unique contributions of flip chip connection to high performance hybrid packaging schemes will ensure its role in the continuing evolution of electronic systems.

Acknowledgments

I wish to thank King Tai for introducing this subject to me, Alan Wachowicz for partnering with me to mass produce solder bumps, T. Dixon Dudderar for his patience and insights, Alan Lyons and Rich Bentson for providing me with useful references and Flora Tsai for lending me her Macintosh in a time of need.

9.6 FLIP CHIP SOLDER BUMP (FCSB) TECHNOLOGY: AN EXAMPLE

Karl J. Puttlitz, Sr.

9.6.1 Introduction

IBM first introduced flip chip solder bump (FCSB) interconnections in 1964 as an integral part of solid logic technology (SLT) hybrid modules utilized in the System/360. It still remains as the primary chip to substrate connection technique practiced by IBM. The connection technique consists of chips with solder bump terminals and a matching set of solder wettable pads on the substrate. The chip is placed upside down (flip chip), each solder bump aligned with its matching substrate pad, as discussed in Section 9.5. All the solder joints are processed simultaneously by reflowing the solder in a furnace. The surface tension force and the confinement of solder volume between solder wettable terminals pads, prevent chip collapse during reflow. Therefore, this type of chip connection often is referred to as a controlled collapse chip connection (C4). Early joints contained a copper ball standoff. Some physical changes have evolved over the several generations of module programs since its inception. FCSB technology is usable with several types of chip carriers: ceramic substrates with thick or thin film metallization, directly attached to polymer cards or flex circuits and on silicon.

FCSB technology is extendible to meet the requirements of future high performance chips. Since FCSB connections are arranged over an area, much larger numbers of I/O terminals can be accommodated in comparison to wire bonding or TAB, whose pads are arranged peripherally on a chip. The feasibility of fabricating dense 128 × 128 area arrays (25 μm bumps, 60 μm centers) has been demonstrated and further advances are anticipated.

Initially, the driving force behind FCSB was to develop a competitive practice to manual wire bonding in the area of cost reduction, increased reliability and productivity. This technology is characterized by self alignment during the chip joining process, high joint strength, ruggedness and the ability to make large numbers of bonds simultaneously. The technology also provides high yield, high chip connection density and high reliability. In recent years, a significantly increased level of activity in FCSB technology has been reported in the literature by major component manufacturers. The demand for increased I/O capacity and higher connection density imposed by VLSI and ULSI fuels this interest.

Various aspects of FCSB technology and its use in multichip module (MCM) applications are discussed in this section. Fabrication processes of the technology are described first, followed by topics including manufacturing viability, extendability to increased I/O density, reliability and chip replacement capability.

9.6.2 Fabrication, Process Flow, Tools and Hardware

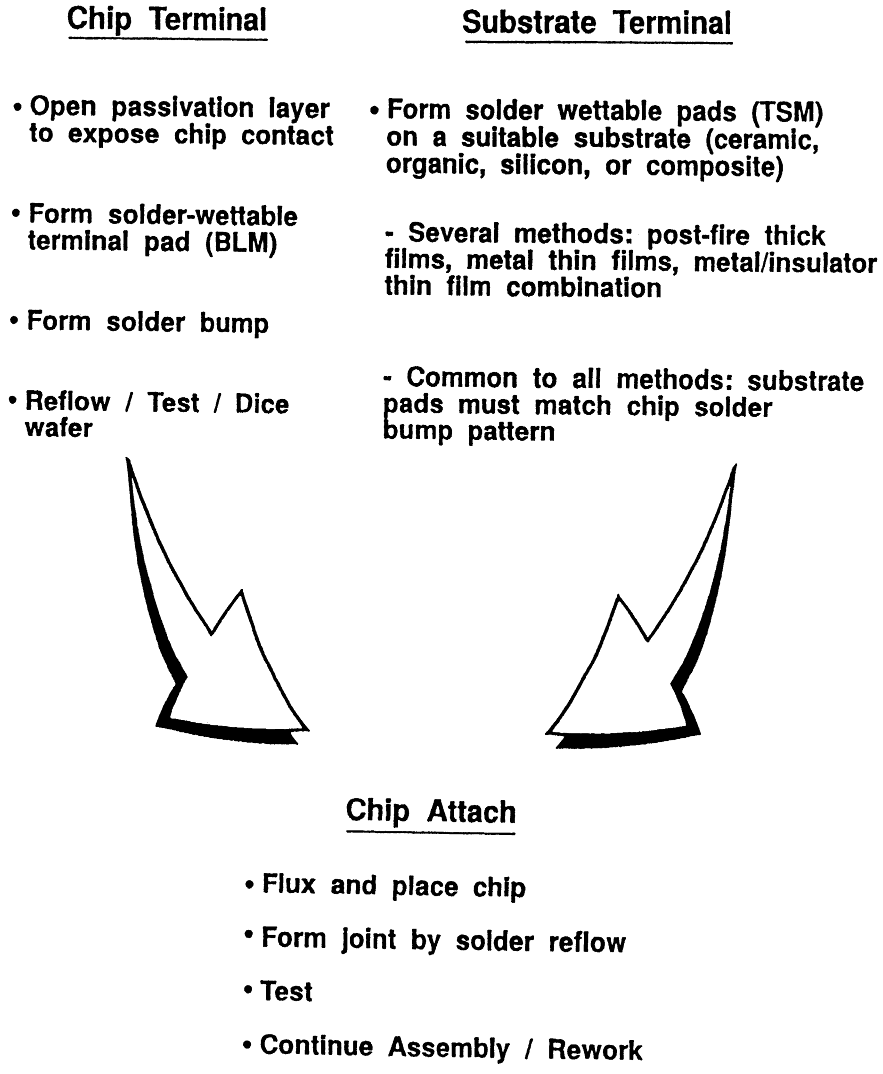
This section describes how the chip connection features unique to FCSB technology are accomplished (namely the solder bump terminals on chips), matching solder wettable areas on the substrate and method of joint formation. An overview of the general process flow is shown in Figure 9-33.

Chip Terminations

Chip terminals are defined during the final stage of chip fabrication. Thus, the majority of a chip's fabrication is the same as chips with terminals other than solder bumps. In the case of FCSB and chips bumped for TAB bonding, the similarity goes even further. Most chips are passivated with a suitable material, such as rf-sputtered quartz, polyimide etc., through which holes or vias are etched to provide an external communication path. Opened vias must be sealed hermetically by evaporating consecutive layers of chromium, copper and gold (Figure 9-34) through a molybdenum mask. Chromium and other materials promote adhesion to the passivation layer and serve as a reaction barrier between the aluminum-based chip metallization and the solder. Copper or some other solder wettable metal is required for subsequent solder reflow. The chromium and copper are phased (a few hundred angstroms of each metal is coevaporated to avoid separation at their interface since theses metals do not exhibit mutual solubility). A thin gold layer (< 0.1 µm) protects the solder wettable copper from oxidation when exposed to room ambient. The solder evaporation which follows is performed in a different evaporator to avoid contamination. To ensure a proper seal, the evaporated pad diameter is made larger than and concentric with the etched vias in the passivation layer. The thin film pads are the only solder wettable regions on the chip. These regions are referred to as the ball limiting metallurgy (BLM).

To complete the chip terminal, solder is evaporated through the same mask used for the BLM films or another mask. A solder alloy is selected whose melting point is sufficiently high to assure compatibility with subsequent processes and assembly.

High lead content Pb/Sn alloys (95 Pb/5 Sn) often are used [71]-[72]. Choice of alloy, however, may be driven by application or assembly-specific requirements. For example, the process temperatures of lead-indium alloys are



Chip Terminal

- Open passivation layer to expose chip contact
- Form solder-wettable terminal pad (BLM)
- Form solder bump
- Reflow / Test / Dice wafer

Substrate Terminal

- Form solder wettable pads (TSM) on a suitable substrate (ceramic, organic, silicon, or composite)

- Several methods: post-fire thick films, metal thin films, metal/insulator thin film combination

- Common to all methods: substrate pads must match chip solder bump pattern

Chip Attach

- Flux and place chip
- Form joint by solder reflow
- Test
- Continue Assembly / Rework

Figure 9-33 General FCSB (C4) technology process flow to create chip and substrate terminals and to achieve chip reflow.

about 100°C less than standard lead-tin FCSB alloys. This, coupled with their superior fatigue resistance, makes them candidates for high thermal mismatch applications, such as direct chip attach (DCAs) to organic chip carriers.

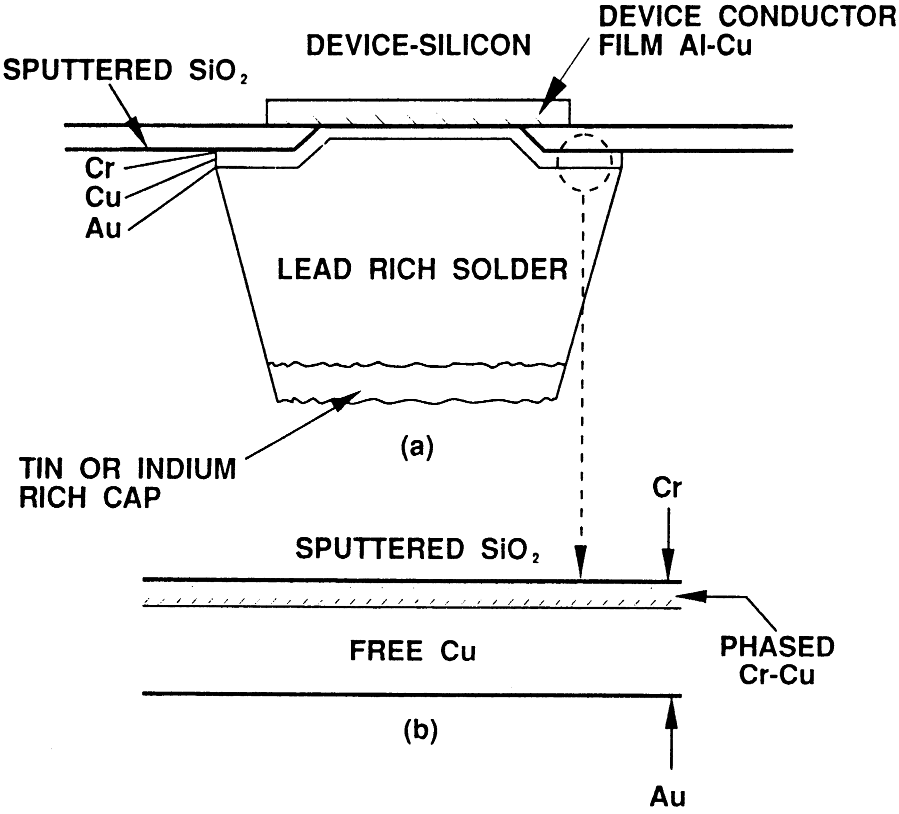


Figure 9-34 As evaporated FCSB solder pad and BLM structure. (a) Cross sectional view of the device terminal metallurgy, (b) BLM layered structure (taken from [70]).

Due to a considerable vapor pressure difference between the elemental constituents, as-evaporated solder deposits consist of two discrete layers (see Figure 9-34a). Evaporated lead-tin and lead-indium solder pads therefore are reflowed to achieve homogeneity. Each chip is electrically tested prior to dicing.

FCSB: Compatible Substrates

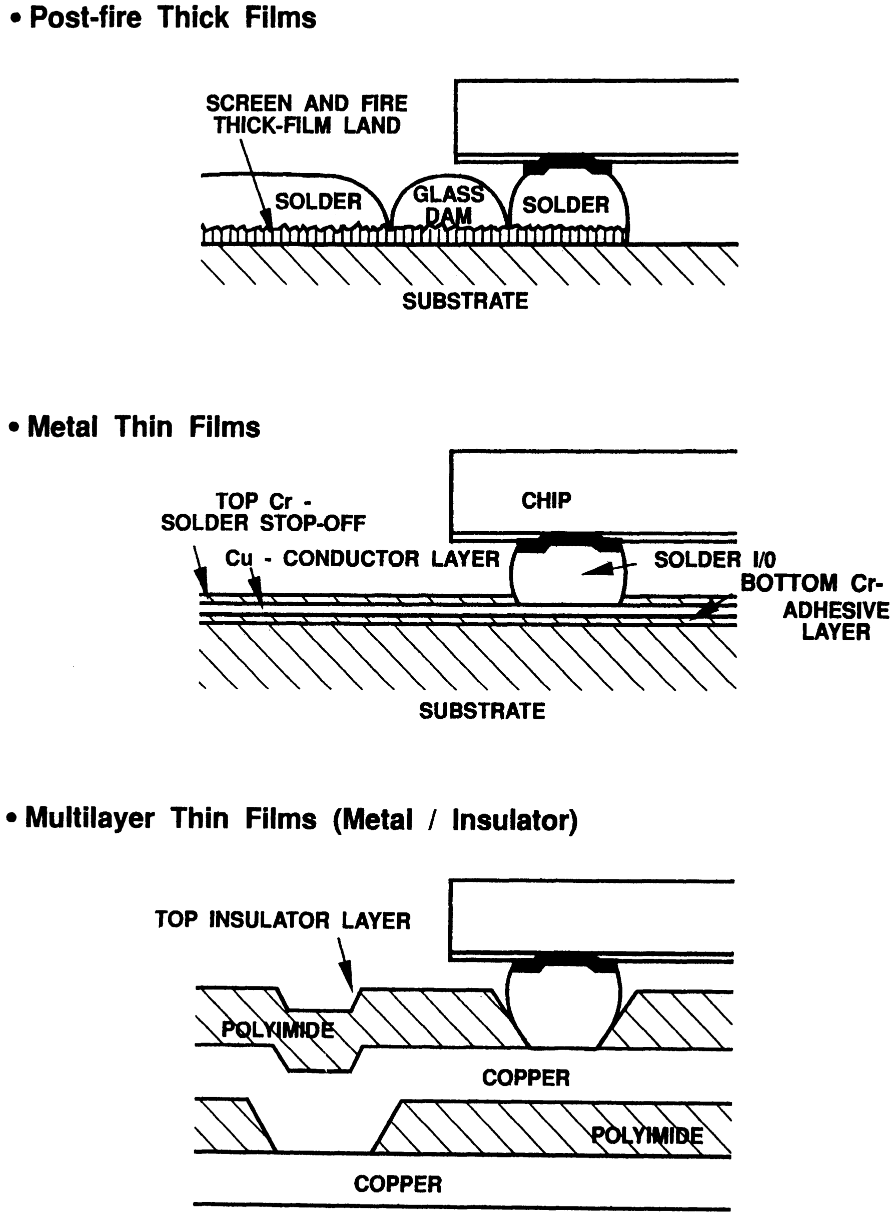
Materials. The choice of substrate material to attach FCSBs is limited by the need to withstand solder reflow temperatures exceeding at least 200°C and more typically in the range of about 340°C. Additionally, the material must possess sufficient strength and stiffness (high Young's modulus) to satisfy handling and fabrication requirements, have good thermal conductivity to avoid thermal gradient induced cracks and be chemically benign to process fluids. Ceramics ideally meet these requirements. Alumina (Al_2O_3) has been the material of choice. Other ceramic materials are used for applications requiring a better thermal expansion match with the chip to reduce joint stress, and improved thermal conductivity. However FCSB also can be joined to polymer-based chip carriers (polyimide) for DCA applications such as flex circuits and CTE-matched cards and boards (such as laminates with copper and Invar layers).

Terminals (TSM). The only substrate surface feature required for flip chip solder bump connections is an array of solder wettable pads which match the chip solder bump I/O pattern. These terminals to which FCSBs are attached (referred to as the top surface metallurgy—TSM), each must be isolated from other solder wettable areas to prevent chip collapse when joined by solder reflow. Depending on substrate type, these pads are formed in various manners.

Postfired Thick Films. Thick film technology most often is used to form circuits on staked pin, pressed ceramic substrates with materials such as Ag, Pd, Au Pt and Ag Pd Au. Ceramic substrates are required to withstand the thick film firing temperatures. TSM pads are formed by screening of thick film glass dams near the ends of the circuit lines (lands) as shown in Figure 9-35a. Being confined, molten solder is restrained from flowing along the lands during reflow, preventing chip collapse.

Metal Thin Films. A metal thin film structure can be used to form TSM pads. This method is applied in the IBM metallized ceramic (MC) technology, which consists of a three layer structure (Cr-Cu-Cr). The bottom Cr layer provides adhesion between the ceramic substrate and thick copper (50K Å) on conductor layer. As shown in Figure 9-35b TSM pads are defined by selectively etching the top Cr film to expose the underlying solder wettable copper. The chromium acts as a solder stop, preventing chip collapse during reflow. Since metal thin films (deposited by evaporation or sputtering) are applied at much lower temperatures compared to thick films, the method has application to substrate materials other than just ceramics. Thin films also have the advantage of permitting narrower lines and spaces, increasing the allowable number of lines per channel for inboard pad escapes.

Multilayer Ceramic Microsockets. The fabrication of multilayer ceramic (MLC) substrates is described in Chapter 6. Therefore, only the via structure, essential to FCSB technology is discussed here. Some of these vertical columns, formed by stacking metal-filled punched holes in an individual green sheet, intercept the substrate top surface and are referred to as microsockets. Substrate microsockets are patterned to mate with chip solder bumps (I/Os). A refractory metal, such as molybdenum or tungsten, is required to withstand the high sintering temperatures of cofired ceramic substrates. The metals also must exhibit good electrical conductivity. Refractory metals, are not solder wettable


• Post-fire Thick Films

• Metal Thin Films

• Multilayer Thin Films (Metal / Insulator)

Figure 9-35 Side view construction illustrating how FCSB connections are made to various substrate schemes. (a) Postfired thick films, (b) metal thin films, and (c) multilayer thin film structure (not to scale).

and thus, are plated with electroless nickel (about 2 μm), followed by immersion gold for corrosion protection [73].

Multilayer Thin Film Structure. Various combinations of metal and insulator thin films (usually polyimide) are used to define substrate TSM pads to which FCSB can be solder reflowed. As shown in Figure 9-35c, TSM pads are defined by selectively opening holes in the top insulator film by laser ablation or other suitable means. This exposes the underlying solder wettable metal to which the chip solder pads are reflowed.

Chip Connection

The FCSB connection technology is similar for any of the substrate types described in the previous section. That is, a nonactivated flux (containing no halide species) is applied to the chip site prior to chip place. Water-white rosin (mostly abietic acid) [74], dispensed in xylene and water soluble fluxes, have been found suitable for standard lead-rich alloys, and lower melting FCSB solders respectively. Flux promotes solder wetting to the TSM pads by cleansing the molten solder bumps of oxides or other surface contaminants. Using pattern recognition or split optics tooling, chips are placed face down so that their solder I/O pads align with mating substrate pads. Chip reflow is accomplished on an individual basis, using a local heat source or simultaneously using an oven or belt furnace to reflow all the chips placed on a substrate [74].

Tools and Hardware

The tools required to form solder bumps on chips are commercially available (wet and reactive ion etch tools, evaporation and sputter deposition tools etc). However, bare die are not available for bumping. In fact, completely processed chips with FCSB terminations generally are not available to the merchant market. Sources for flip chip solder bump will come about in response to demand. But, for the near term, flip chips solder bumps will generally be available only on a contract basis from chip manufacturers with this capability (IBM, Intel, Motorola). IBM, a leader in this technology, has recently made its device and electronic packaging technology available to the merchant market. This should aid the industry in moving along the learning curve more quickly.

Obtaining FCSB compatible substrates is much less difficult, since some infrastructure already is in place. Much of the technology is very mature and the tools readily available. Therefore, MCM manufacturers to place circuits on vendor substrates with metal thick or thin films. These technologies should be sufficient for most applications. Tool costs and process complexity increase dramatically with multilayer thin film technology and is justified only for special applications.

Manual chip placement tools are available for flip chip but automatic production tools are not. Belt furnaces for chip attach reflow also are available, but post join flux clean tools, designed to remove flux residue from beneath the chip, are not.

Alternative Approaches
Electroplated copper bumps capped with sufficient solder for reflow bonding is an alternative to solder bumps [75]. The copper bumps act as standoffs to prevent chip collapse. Several versions of solderless flip chips are described in the literature. The bump to substrate connection typically is made with a conductive epoxy or conductively filled polymer. The chip bumps consist of copper or conductive polymer formed by screen printing. The comparatively low process temperature for these materials, in comparison to solder reflow, allows for a diversity of substrate materials, including many plastic and low temperature composites. Conductive polymer joints are more compliant than their solder counterparts owing to their low elastic modules [76]. However, electric current and moisture-induced silver migration is a concern with this type of material.

A solderless flip chip connection system recently has been demonstrated where chip and substrate are not physically attached [77] , as illustrated in Figure 9-36. LSI chips with gold bumps (3 μm high, 10 μm pitch, 2330 I/O) are held in contact with a substrate by the contractile forces generated within an ultraviolet light setting resin. Chips are replaceable since the resin is soluble, and is therefore extendable to ULSI (10,000 I/O).

9.6.3 Manufacturability

The ease with which chip connections are achieved, with acceptable yields in a manufacturing environment, is a major consideration. This section discusses several features, unique to FCSBs technology, that enhance manufacturability.

Self Alignment
The self alignment feature of FCSBs allows coarsely or misplaced chips to be pulled into position (centered) by surface tension forces when the solder is molten. A misalignment of up to approximately one-quarter pad can be tolerated and still assure self alignment of the final joints. Self alignment can be used to address a variety of issues, such as aligning small pad sizes (24 μm) and low cost flip chip soldering in which large alignment pads are used to align smaller functional joints without requiring precision placement as described in Section 9.5.

Ruggedness/Strength
Unattached flip chips are sufficiently rugged to permit handling, including automated testing and chip placement. Unlike TAB and wire bonds, which are

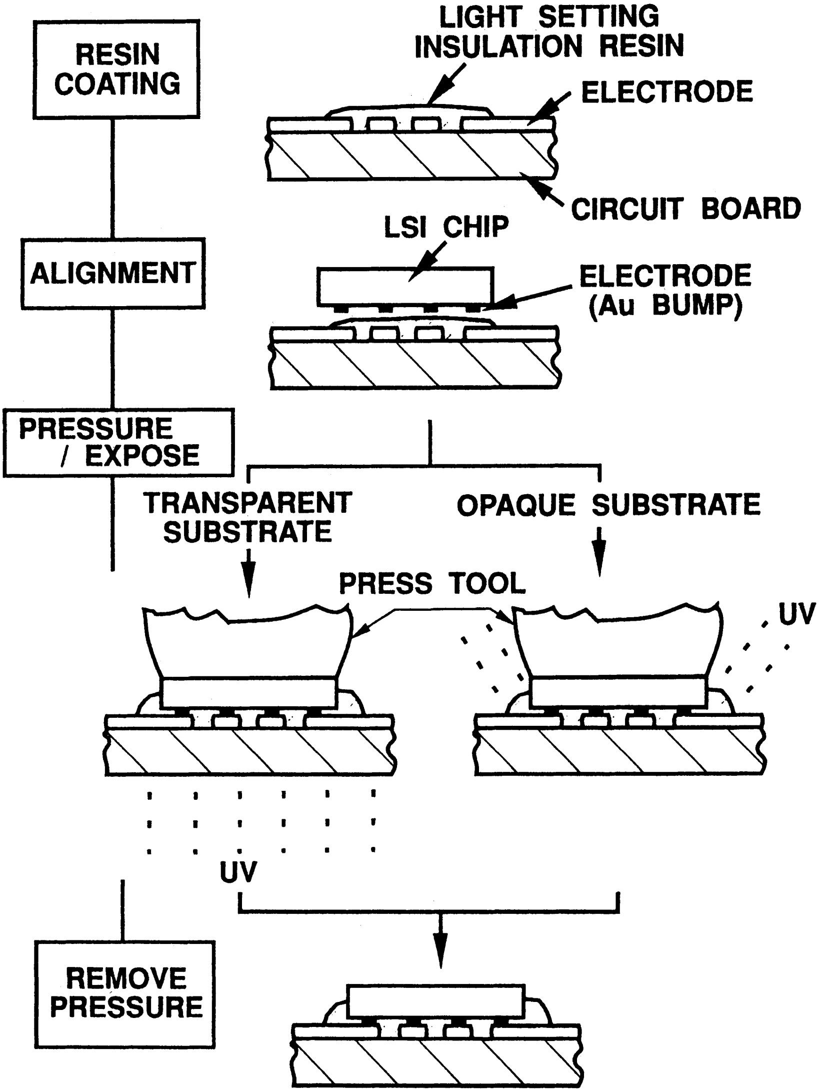


Figure 9-36 Process flow for achieving flip chip connections by resin-generated compressive forces (taken from [77]).

both rather delicate and exposed, FCSB joints are located beneath the chip and therefore are protected. Also, the joint tensile strength is high, compared with other chip connections, ranging from about one pound for the original three terminal SLT chip to over 80 pounds for some current VLSI chips. Electrical shorts sometimes occur in wire bonded devices when the encapsulating material is poured around them. This is not a problem with FCSB joints as they are quasi-rigid.

Yield/Throughput

It has been pointed out that the high cost of chips ($2 to $100 per IC), combined with the large number of total I/O from several attached chips (which may total several thousand), necessitates a very high yield chip to substrate connection technology. For example, an MCM package with 4000 chip I/O, with only a 90% overall yield, requires a 99.997 per I/O attachment yield [78].

Flip chip technology provides the highest bond yield and joint reliability. Also, being a batch process, throughput does not depend upon the total chip I/O on a module. Identifying defective chips prior to chip attach is becoming increasingly more important in meeting MCM yield objectives. Burn-in can be achieved with solder bumped flip chips, but not with the versatility of TAB. A rework capability with a high degree of control is fundamental to attaining yield levels necessary for economic viability. This is another key attribute which makes FCSB technology particularly well suited for MCM applications.

9.6.4 Rework

In the case of single chip modules (SCM), rework costs usually are not warranted; unsatisfactory parts are simply discarded. However, if one or more chips require replacement on multichip carriers, cost considerations prohibit discarding these carriers and their functional chips.

Among the more common reasons for replacing chips prior to shipping include: incorrect orientation (chip is rotated), wrong chip at a particular site, defective chip (electrical test escapes, lower solder volume causing an electrical open or partial wet), mechanical damage due to handling and changes necessary to satisfy customer requests. Also, performance upgrades are achieved by replacing chips on modules from the field with new state of the art chip sets.

In general, the replacement of components is considered a significant technical hurdle [78], particularly ultrasonically bonded TAB and wire bonded chip formats which do not easily lend themselves to replacement. Wire bonded chips usually are removed by manually severing individual wires at their bonds, a very time consuming operation, especially for high I/O count chips. A TAB assembly is removed by heating and remelting solder reflowed outer lead bonds (OLB). Heating also sufficiently weakens the die attach adhesive to permit chip removal. The following sections discuss several techniques for replacing flip chip with solder bumps.

Chip Removal

Mechanical methods offer the most direct approach to chip removal. However, sufficient space must be available to grasp the chip in a suitable manner. For example, a chip with less than 300 I/O may be removed by a few back and forth rotations (torque approximately 2 degrees in each direction), as depicted in Figure 9-37. Chips with larger numbers of FCSB joints, however, can, be safely removed from multichip carriers using ultrasound methods. A transducer coupled to a target chip in the 20 - 40 watt range is sufficient for most applications. Removal is rapid, requiring only a few seconds or less, depending on I/O count.

Melting the solder joints allows chips to be lifted from the surface with a vacuum pencil. Heating the entire module above the solder joint liquidus temperature is an obvious approach. Localized heating is preferred to prevent the formation of intermetallics, which can weaken the solder joint interfaces.

Multichip carriers are normally pre-heated (referred to as bias heating) from the I/O side to a temperature approximately midway between room ambient and

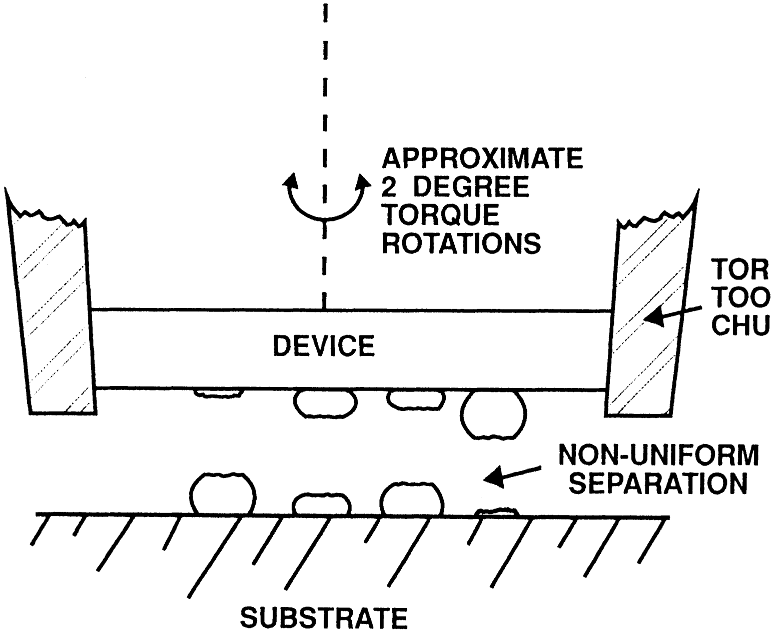


Figure 9-37 Cross sectional schematic of FCSB mounted device removed from a chip carrier by mechanical means (torque) (not to scale) (taken from [104]).

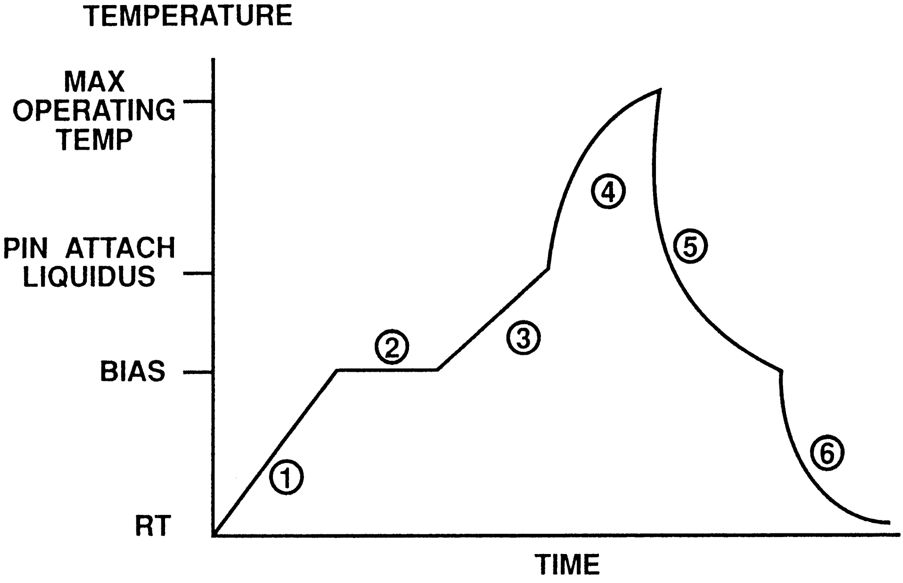


Figure 9-38 Generalized profile for thermally-enhanced, chip removal techniques at various stages of heating and cooling. Stages represented are: (1) heating from ambient to bias temperature, (2) temperature equilibration, (3) grading the temperature between a replacement site and surroundings to prevent sharp thermal gradients, (4) locally spiking the temperature at the replacement site, (5) slow convection cooling and (6) forced cooling. (Time and temperature not to scale.) (Taken from [103].)

solder solidification temperature. Additional energy from another source is directed to locally heat only the chip targeted for removal.

A general thermal profile for localized, thermally enhanced, chip replacement techniques is shown in Figure 9-38. Figure 9-38 depicts typical stages of heating and cooling.

When direct conduction techniques are utilized, they usually consist of a heated element in contact with the back side of a chip. Maintaining the required planarity and parallelism is difficult, due in large part to heat element warpage. Convective techniques, such as directing a flow of heated gas (nitrogen gas) to a target chip, are not sensitive to planarity. Typically, these methods require ample chip to chip spacing to avoid melting the joints of neighboring chips. Radiation methods, such as focused infrared or laser sources, also are suitable for removing closely spaced chips.

Solder Dress

Residual solder is left on substrate pads after the chips are removed. Site dress is the process of removing residual solder prior to replacing a chip. Site dress prevents electrical shorts due to solder volume buildup at sites which experience multiple chip replacement cycles. Since dressed sites are planarized, replacement chips are placed with the same accuracy as initially joined chips. Also, there is a strong relationship between fatigue life and solder volume of FCSB joints. Chip solder pads are evaporated to possess both the required uniformity and optimum solder volume when fabricated. Attaching these chips to undressed sites can have a significant adverse effect on reliability. Adjusting replacement chip solder volume would result in a substantial proliferation in chip part numbers, complicating manufacturing logistics. Solder dressing eliminates these concerns, rendering substrate pads to a near original condition.

Residual solder can be shaved mechanically, however, this procedure should only be used to remove debris such as chip fragments occasionally left after mechanical removal (tensile pull). Mechanical shaving potentially can damage the substrate.

The industry has long practiced ways of removing molten solder (solder suckers, solder wicks etc.). Similar techniques have been developed to remove molten residual solder from FCSB sites. As with other localized heating processes described earlier, the entire module is preheated to reduce the potential for introducing thermal gradient-induced stress cracks. Also, as before, external heat sources are utilized to raise the temperature at replacement chip sites. One such source is a semiautomatic hot gas site dress (HGD) tool. A tip with pedestals surrounds the target site and contacts the substrate as indicated in the sketch of Figure 9-39. A heated inert gas (nitrogen) directed at the substrate, heats and melts the residual solder. The velocity is sufficient to propel the molten residue into the gas stream removing it from the site [74].

Although HGD is not considered extendible to large footprints (2.5 mm square array is about maximum), there are techniques which are relatively footprint insensitive. One technique utilizes a solder wick chip [80]. The residual solder spreads over a grooved silicon surface prepared by an anisotropic etch process, and in followed by a solder wettable film deposit. Another technique utilizes a porus metal slug fabricated from any solder wettable metal by standard powder metallurgical practices. The molten residual solder is absorbed into slug by capillarity.

9.6.5 Reliability

Thermal Fatigue

Flip chip solder bump joints experience a displacement with each machine on and off cycle due to strains resulting from a (CTE) mismatch between silicon

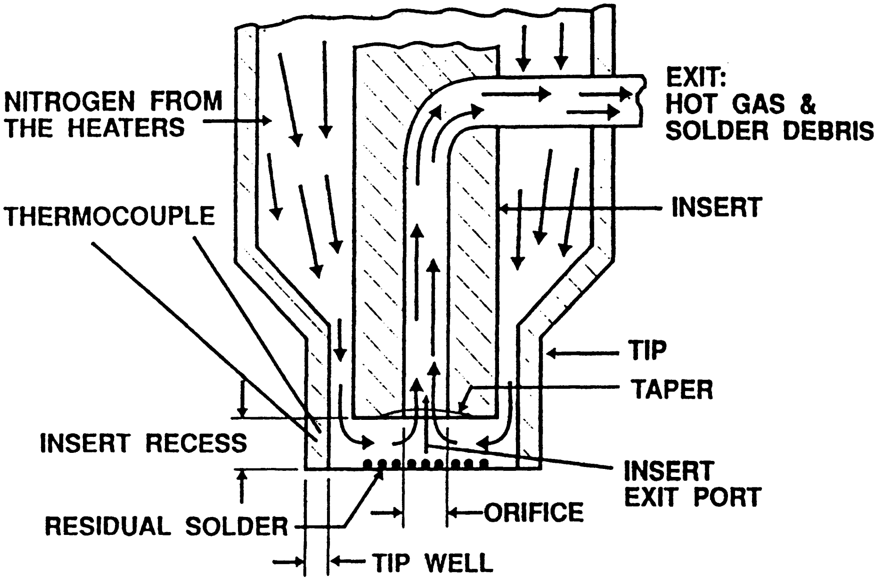


Figure 9-39 Side view construction indicating the direction of gas flow within the probe tip of the hot gas solder dress tool (taken from [74]).

chips (2.8 × 10^{-6}/°C) and the alumina substrate (6 × 10^{-6}/°C). The degree of strain experienced by a particular solder joint depends directly on its distance from the zero displacement point or neutral point (DNP). Normally the neutral point is located at or close to the geometric center of a chip. Thus, corner joints in square arrays experience the greatest strain, as illustrated in Figure 9-40. The concern for premature failure due to CTE mismatch often is cited as a key reason FCSB technology has not gained wide spread acceptance.

The joint is susceptible to thermal fatigue failure since it is the pliable member separating chip and substrate, two relatively rigid elements. Accordingly, a major function of FCSB joints is strain accommodation to avoid premature failure and loss of functionality. The accumulated plastic (permanent) deformation over a lifetime can be severe, exceeding a 1000% in some cases.

Design Considerations

The shape of FCSB joints, best approximated as a truncated sphere, depends on three geometric factors: the terminal radius between the joint and chip, r_c (BLM) terminal radius between the joint and substrate, r_c (TSM) and joint height, h

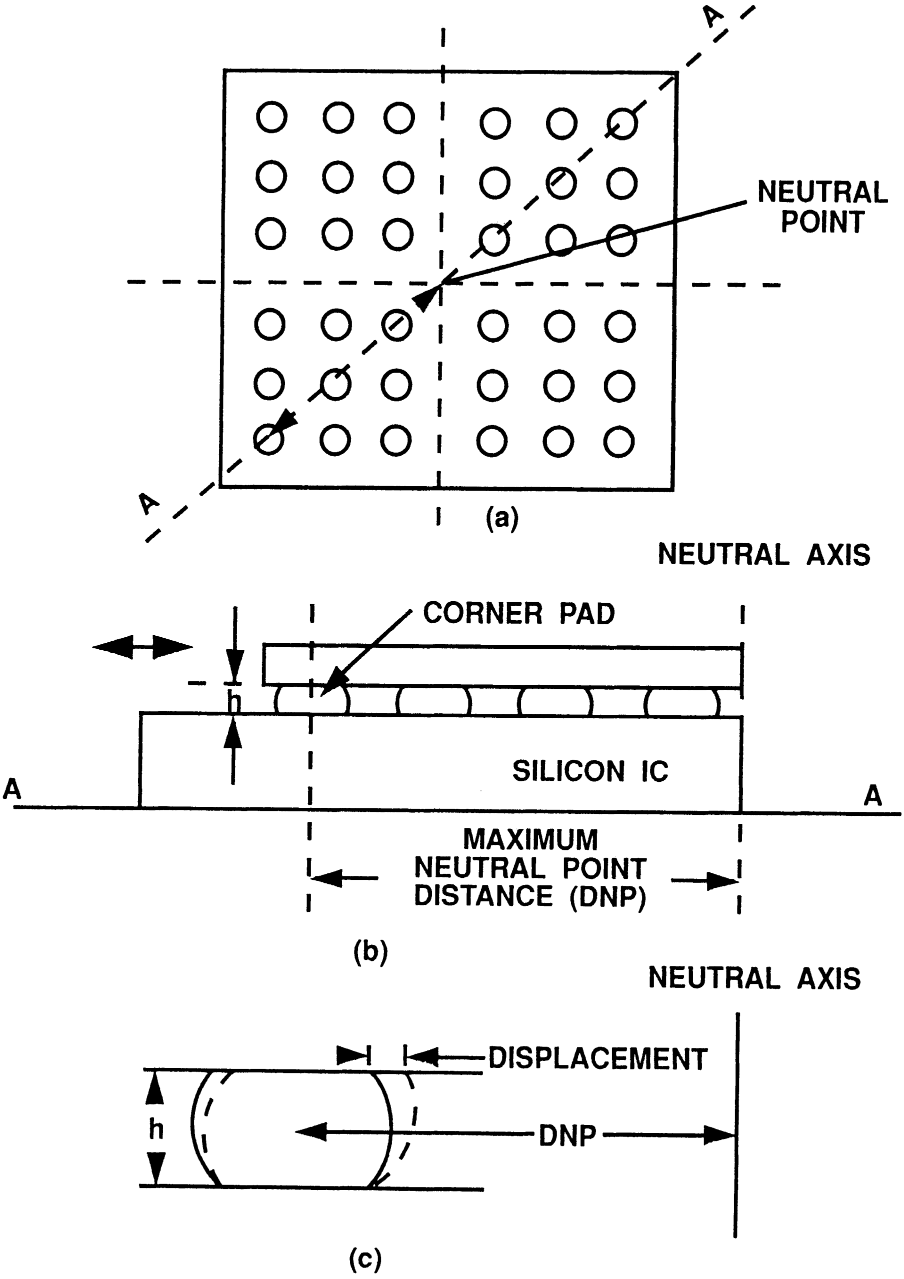


Figure 9-40 Schematic illustrating the displacements of FCSB joints from the neutral point due to the mismatch in thermal expansivity between chip and chip carrier.

(Figure 9-41). Chip weight has little effect on joint shape. The I/O count of current chips is more than sufficient to exert the surface tension forces necessary to resist collapse when the joints are molten.

The procedure normally followed is to optimize the substrate terminal dimensions and joint solder volume with respect to the chip terminal dimensions, usually fixed at the time of substrate design. The effect that several radius ratios, R = r_s (TSM) /r_c (BLM), have on the maximum local strain experienced by FCSB joints as a function of solder volume is shown in Figure 9-41. Actual strain values depend upon temperature, so the graph has been normalized to indicate relative differences. The solder volume corresponding to the lowest strain, has been assigned a value of unity. That condition occurs, as Figure 9-41 illustrates, when the terminals have the same dimensions (R = 1). Under this optimized condition, the strain is distributed symmetrically, and the maximum strain is relatively insensitive to solder volume variations which may occur during processing. However, for any set of terminal dimensions, there is an optimum solder volume for which the joint strain is a minimum. That minimum represents the most favorable balance between two competing conditions. Increasing joint solder volume increases joint height, which in turn reduces the shear strain. however, the joint horizontal cross section also increases, effecting strain distribution. Selecting matched terminals (R=1) only ensures optimum joint fatigue life if failure occurs within the bulk solder or that the solder-bond strength at both terminals is about equal. Otherwise the radius ratio must be shifted to appropriately compensate for a weak interface. Joint fatigue life is optimized by balancing these design parameters such that the predominant mode of failure is within the bulk solder, not at the terminal interfaces.

Models, Tests, Field Experience

The Coffin-Mansion relationship is widely utilized in predicting low cycle fatigue life (10K cycles) of metals. Elastic strain contributions, small compared to plastic strain, are neglected. But, since operating conditions are above the homologous temperature (typically one-half the absolute melting point) for solder materials, Norris and Landzberg [82] introduced frequency and temperature parameters to account for two plastic flow failure mechanisms in FCSB joints. One mechanism is characterized by intragranular deformation, while the other is characterized by relative movement of grains along their boundaries with attendant creation of intergranular voids. Both processes cause defects within the solder from which fatigue cracks can initiate and propagate. The cycle condition determines which mechanism predominates. Accelerated laboratory tests are used to predict the fatigue life of FCSB chip connections under field conditions

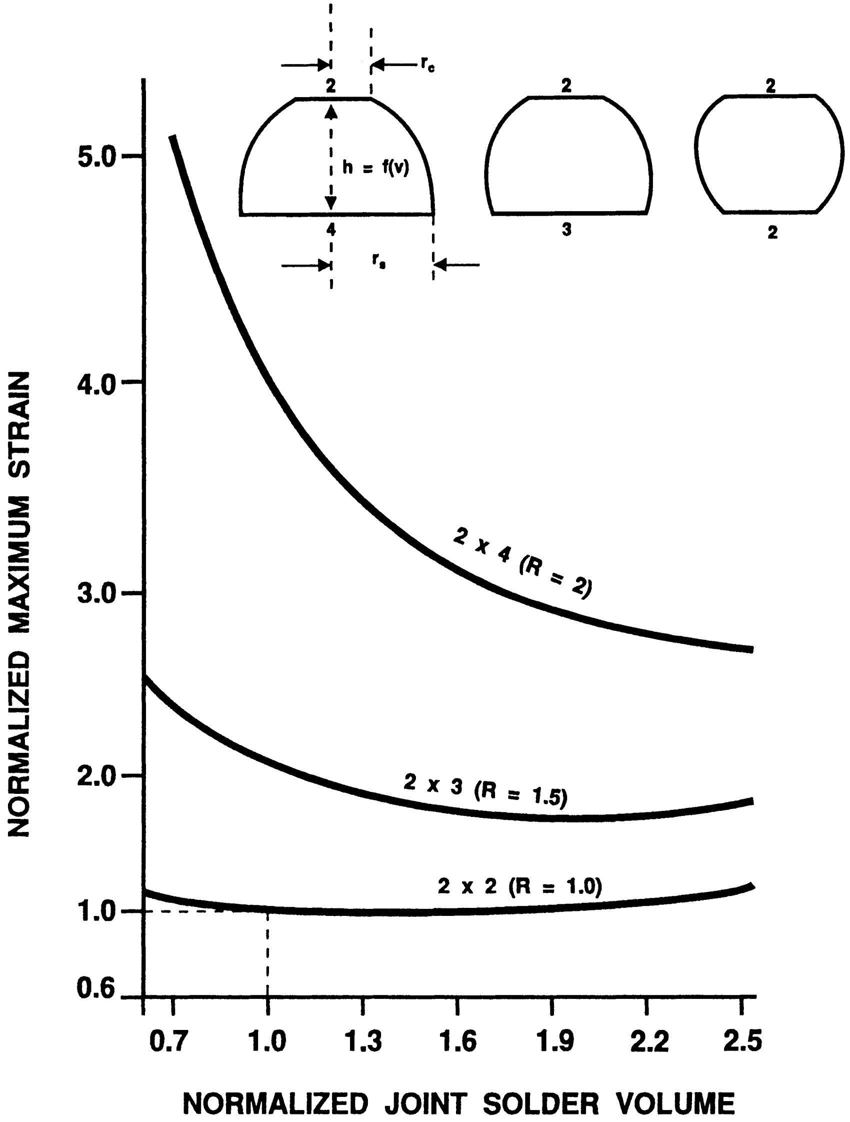


Figure 9-41 The effect of volume on the maximum strain on FCSB joints as a function of solder volume for various terminal ratio combinations [81].

on a per joint basis. Accordingly, fatigue life in the field (N_F) is expressed as:

$$N_{F(x)} = N_{T(X)} \left(\frac{\gamma_T}{\gamma_F} \right)^n \left(\frac{f_F}{f_T} \right)^{\frac{1}{3}} \exp \left\{ 1414 \left(\frac{1}{T_F} - \frac{1}{T_T} \right) \right\}$$

(9-1)

where: γ = maximum shear strain, in./in.
n = empirical constant, approximately 2.0
f = frequency, cycles per day
T = maximum temperature of the excursion
F = field, T = test indicates percentage of joints failed in $F_{(x)}$ and $T_{(x)}$.

The following example illustrates the application of Equation (9-1). Assume the FCSB joints of a chip mounted on an MCM are expected to experience an average of six thermal excursions between room ambient (25°C) and 60°C per day resulting in a 0.005 in./in. maximum plastic shear strain. We also assume the average fatigue life, $N_{T(50)}$ of FCSB joints mounted on similar parts subjected to a laboratory thermal cycle test to be 3,000 cycles. The temperature variations on this laboratory test are taken to be between 0°C and 100°C, 72 times per day, resulting in a maximum plastic strain of 0.02 in./in. Accordingly, the projected FCSB connection fatigue life for the field conditions is stated as:

$$\begin{aligned} N_{F(50)} &= 3000 \left(\frac{0.02}{0.005} \right)^{1.9} \left(\frac{8}{72} \right)^{\frac{1}{3}} \exp \left\{ 1414 \left(\frac{1}{333} - \frac{1}{373} \right) \right\} \\ &\approx 31{,}500\text{cycles} \end{aligned}$$

(9-2)

Stress relaxation, which takes place in plastically deformed FCSB joints eventually establishes a state of equilibrium. A minimum value of six is used for the frequency terms to account for this condition. Below this value there are no additional physical changes which take place. Also comparisons between field and test only can be made at the same failure percentage, (use $N_{T(50)}$ to determine $N_{F(50)}$) Utilizing their model, Norris and Landzberg predicted that the end of life (EOL) failure rate for logic chips utilized in the IBM System/370 would not exceed 10^{-7}% per 1000 power on hours (POH) on a per joint basis, a projection verified by field data [83]. Thus, FCSB connections also set the

Even so, there have been two lingering concerns. Manufacturers and users have been reluctant to depend upon process control to assure the reliability of FCSB array interior joints which cannot be inspected visually. Additionally chip dimensions have steadily increased with an accompanying increase in neutral point distances. That is, in the nearly two decades since the initial FCSB reliability projections were made, chip area has increased about 90 fold, accompanied by a more than 200 fold increase in I/O (based on a 6.5 mm logic chip, 27 × 27 solder pad array). However, over that same period, there has never been a single FCSB joint that failed in the field which represents many billions of power on hours (POH) - 30 billion POH for MCM-C packages alone [84] for which shipping began in 1978. These results should not be interpreted as suggesting that thermal fatigue of FCSB joints is not a concern. They do, however, indicate that within the design and use condition parameters exercised to date, FCSB joints continue to be the most reliable chip to next level of assembly connection in the industry. Additionally, it underscores the fact that a rather simple model, given the metallurgical complexities and dynamic nature of the joint, has been remarkably accurate in predicting field behavior and, therefore, is a very useful design guide. Model modifications may become necessary in the future as chip sizes and attendant DNP's continue to grow.

Fatigue Enhancements

The strain experienced by FCSB joints is directly proportional to the difference between chip and substrate CTE, ambient and maximum operating temperature and distance from the neutral point. But joint strain varies inversely with joint height, h as shown in Equation 9-3.

$$\gamma \propto \frac{(\Delta \mathrm{CTE})\,(\Delta \mathrm{T})\,(\mathrm{DNP})}{\mathrm{h}} \qquad (9\text{-}3)$$

The following steps can be taken to reduce the strain experienced by FCSB connections.

- **Eliminate High DNP Terminals.** Truncating the corners of large square arrays eliminates high DNP pads which are most prone to failure this results in a significant improvement with only a slight reduction in I/O density.

- **Reduce CTE Difference.** Various alternatives to alumina-based substrates have been reported in the literature. For example, Greer [85] using a polyimide Kevlar substrate achieved dramatic improvements in using a polyimide Kevlar substrate achieved dramatic improvements in FCSB thermal fatigue life by matching the substrate (CTE) to that of silicon, as shown in Figure 9-42. Matched card and board materials, such as copper-Invar-copper, are now commercially available, providing the opportunity for FCSB. CTE matched ceramic materials have also been pursued. Most notable among these are silicon carbide, aluminum nitride and glass-ceramics [86]. Also, silicon on silicon applications, which provide a direct CTE match, are becoming more prevalent [87].

- **Increase Joint Height**. Several means of increasing joint height to decrease joint stress (Equation 9-3) have been demonstrated. Among these methods are stacking solder pads to form a column or solidify elongated joints which are stretched while molten. Stretched joints have been fabricated by centrifugal methods and also using two solders whose melting points vary only slightly. Surface tension forces, and thus volume, of the higher melting point solder must be sufficient to stretch the functional pads. This method pertains only to applications wherein chip real estate availability is not an issue. Regardless of how achieved, stretching provides FCSB joints with an additional benefit - an enhanced shape factor. The as-reflowed barrel shape of FCSB joints. (Figure 9-43a) concentrates most of the strain deformation and accompanying fatigue deformation in planes close to the joint interfaces. However, stretched joints with slightly barrel or slightly hour glass shapes (Figures 9-43b and 9-43c distribute the deformation more uniformly throughout the joint volume. Improvements of 10 times and more are not unusual compared with standard barrel-shaped joints. Fatigue life is, however, adversely affected if the degree of stretch is excessive (Figure 9-43d), with early failure occurring within the middle portion of the joint.

- **Conformal Coating and Encapsulant Materials**. A wide range of materials are utilized to protect electronic devices from moisture, ionic contaminants, radiation and hostile environments. They also are installed in the gap between chip and substrate to enhance fatigue resistance. In one approach, the gap is only partially filled, conformally coating the substrate chip site, FCSB joint and chip bottom surfaces as illustrated in Figure 9-44. Low stiffness materials have been used in this application. Among these are dispensed liquids, such as amide-imide polyimide (AIP), which are subsequently cured [91] and chemical vapor-deposited Parylene [92]. Only peripheral (high DNP) pads need be coated to counter the ill effects due to chip bending, a factor whose importance increases with chip size [91]. Fatigue life enhancements of 1.5 - 3.0 times are typical.

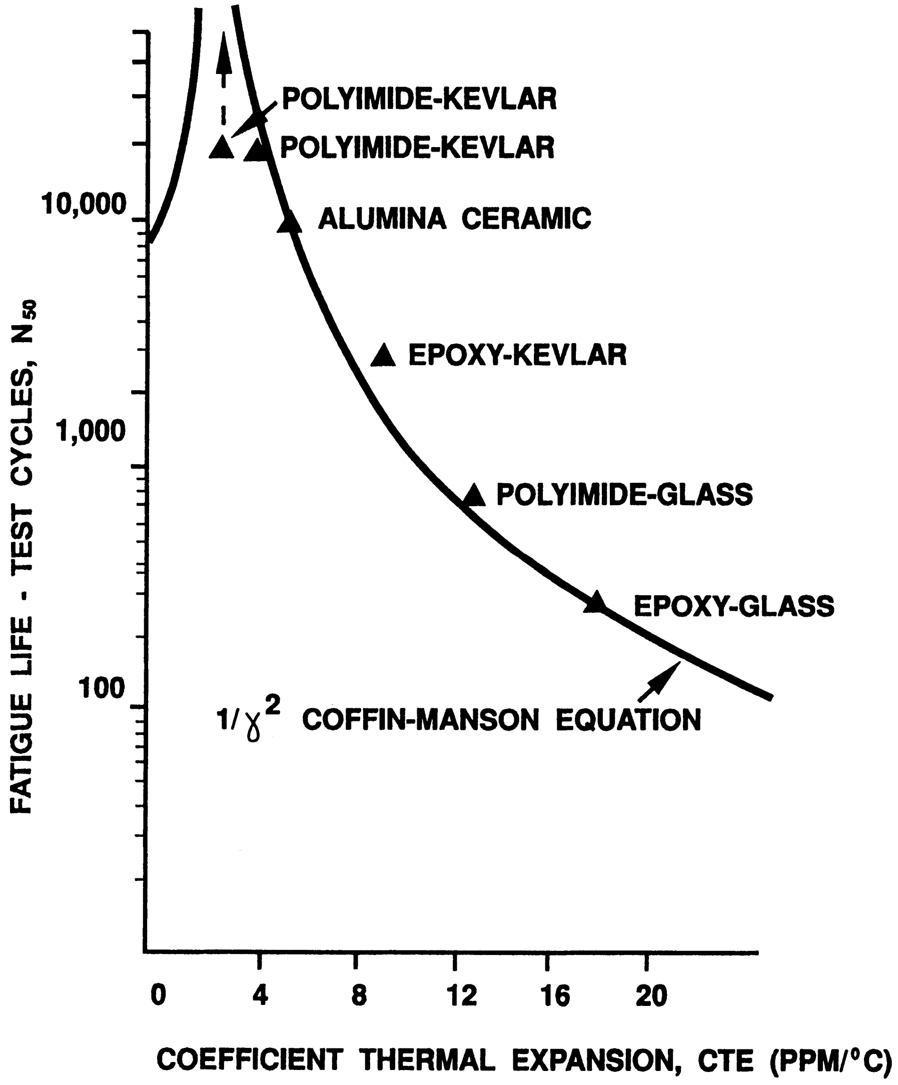


Figure 9-42 Effect of the coefficient of thermal expansion of a chip carrier on the fatigue life of FCSB mounted silicon chips (taken from [85]).

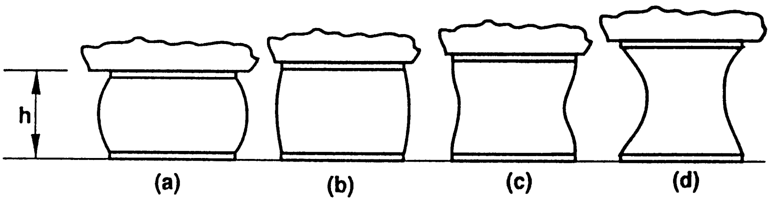


Figure 9-43 Cross sectional schematic of FCSB joint profiles with increasing degrees of stretch, (a) as-reflowed, barrel, (b) near column (c) hour glass, desirable for enhanced fatigue resistance, (d) highly concave, which, if excessive, has an adverse effect on fatigue life (not to scale).

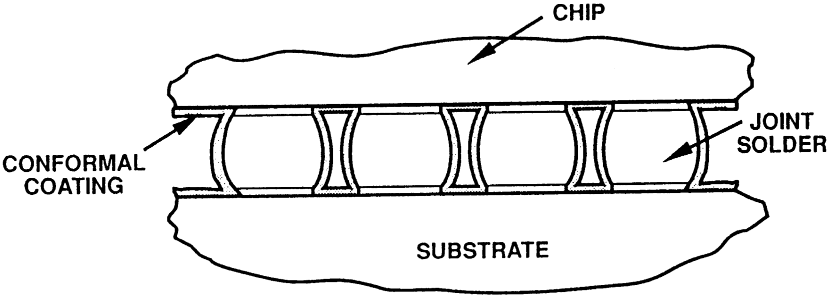


Figure 9-44 Cross sectional schematic of conformally coated FCSB joints (not to scale).

Yet another approach is to completely fill the chip-to-substrate gap with a filled epoxy resin (Figure 9-45). Dramatic fatigue life improvements (greater than 10 times are not uncommon, see Figure 9-46) can be achieved by this method. Improvements are optimum when encapsulant and joint solder CTEs are matched. This development greatly extends the allowable DNP and operating temperature ranges for FCSB mounted devices, freeing designers from the imposition of those constraints.

Currently, none of the available matched CTE materials is reworkable, thus limiting their use in MCM applications. However, the expectation is that reworkable materials will be available shortly.

- **Joint Solder Materials.** Lead indium alloys offer an opportunity for significant fatigue life enhancement over 95 Pb-Sn, a commonly used

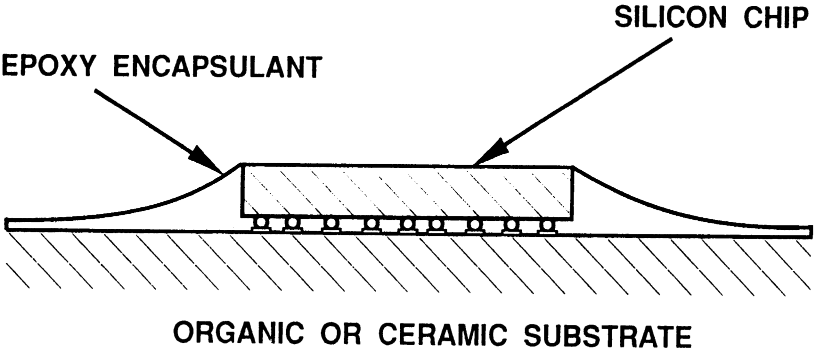


Figure 9-45 Cross sectional schematic representation of a FCSB mounted chip whose joints are encapsulated (taken from [93]).

FCSB solder. A parabolic-like dependence exists between indium concentration and fatigue life, with a minimum at 10 - 15% In. Improvement factors of 2, 3 and 20 times relative to 95 Pb-Sn correspond to indium concentrations of 5%, 50% and 100% respectively. Corrosion susceptibility increases with In content [94] and was recognized as a key concern for nonhermetic package applications [70]. Under these conditions the potential for corrosion fatigue exists as well. Additionally, 50 Pb-In FCSB joints and Pb-In solders, in general exhibit a much greater thermomigration rate in comparison to Pb-Sn alloys. Howard [94] has recommended Pb-low In (3% - 5%) alloys as candidates since their susceptibility to corrosion is virtually nonexistent.

Ambient Control. Recrystallization, creep and microstructural coarsening processes, which occur at high homologous temperatures, all favor fatigue damage. Even under normal conditions FCSB solder joints operate at over two-thirds of their absolute melting points [95]. Environmental effects are generally greatest under low cycle fatigue conditions [96]. An order of magnitude enhancement was observed in early fatigue studies of lead tested in a vacuum compared to air. Similar enhancements were reported for solder joints protected from oxidation by vacuum grease. In a comparative study, 95 Pb-Sn FCSB joints hermetically sealed in packages under a dry nitrogen atmosphere exhibited no failures, even when cycled to twice the fatigue life of similar joints not hermetically packaged [95]. Closely controlling the packaged environment is yet another factor that can be exploited to enhance fatigue life.

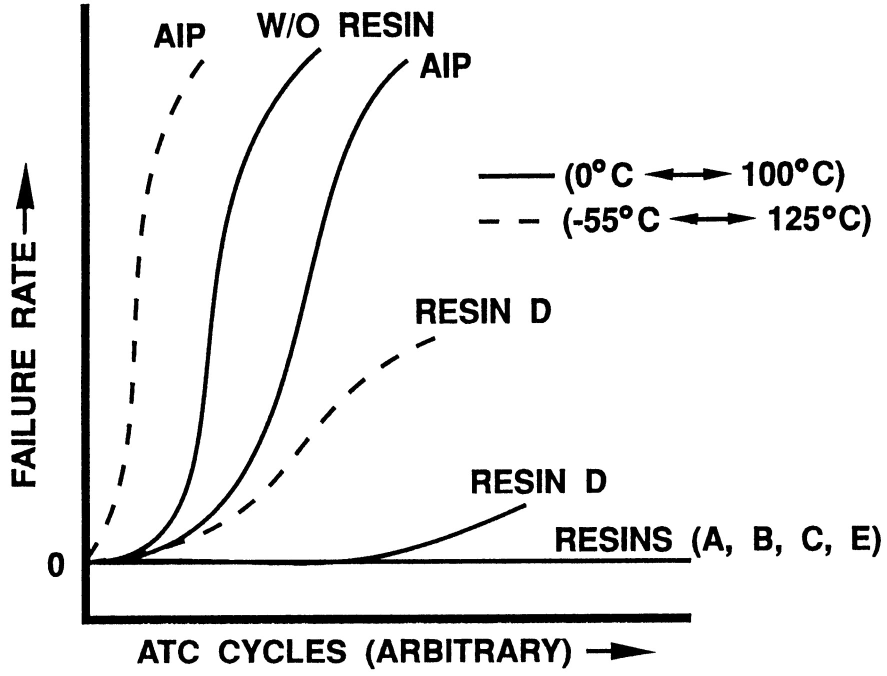


Figure 9-46 Effect of temperature cycle range and encapsulation on thermal fatigue life of FCSB mounted silicon chips (taken from [93]).

9.6.6 Performance

Cooling Capability

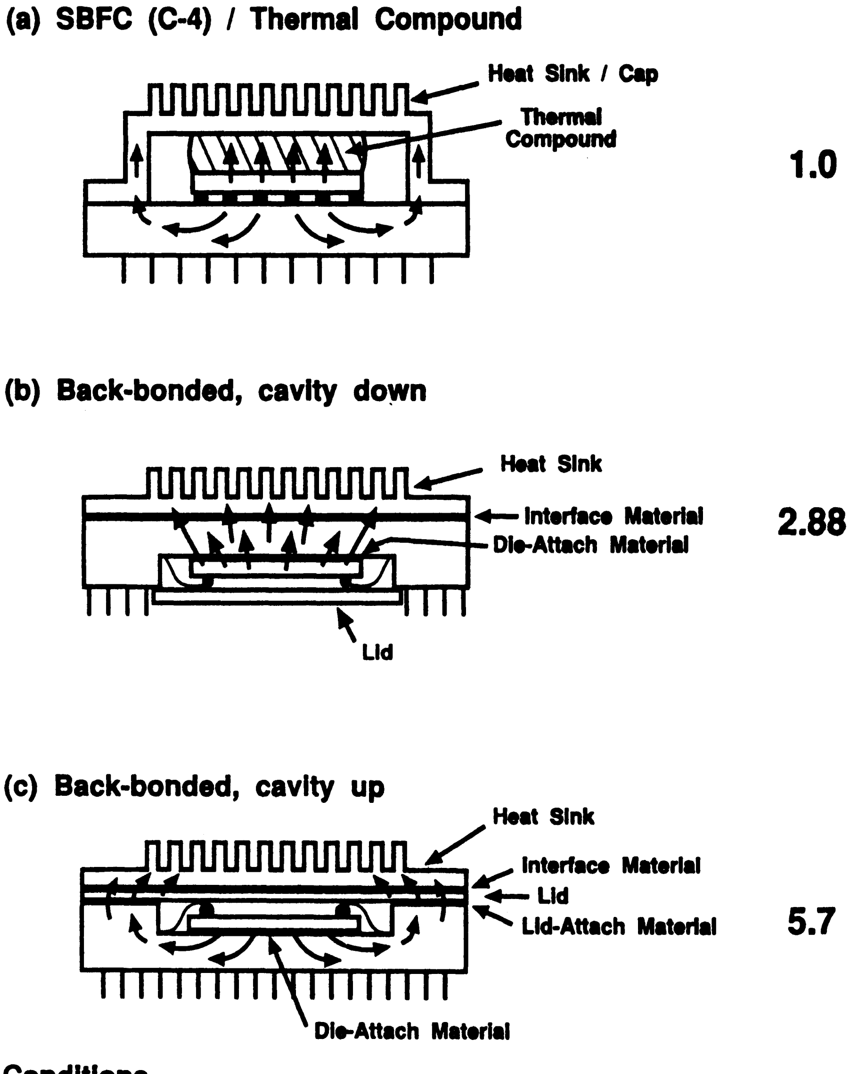
Initially FCSB technology's major heat dissipation path was through the chip solder joints. These early applications typically required less than 25 I/O (peripheral), offering limited capacity to dissipate heat in comparison to back bonded chips. At the time, however, total module dissipation requirements were also low, rarely exceeding 0.5 watt. Cooling capability has been significantly improved due to VLSI, which has driven the technology to large and dense (27 × 27, and 9 mil pitch respectively) area arrays, greatly increasing the area of contact with the substrate. This has made it possible to maintain chip junction temperatures of FCSB utilized in variety of IBM cost/performance products at or below 85°C using standard convection cooling. Typical of these is a 28 mm module with four chips (4.6 mm, 121 I/O) and a module heat dissipation rating of 5.4 watts utilized in the IBM System/38 computer.

Various techniques are employed to enhanced cooling efficiency which address both the back and front (I/O side) sides of flip chips. For example, a direct thermal path is established by introducing a thermal compound (TC) [97] [98] in the gap between the chip and cap (Figure 9-47) or die bonding the chip to the cap [99]. Thermal compounds, sometimes referred to as thermal greases, are commercially available with conductivities ranging from about 0.5 - 1.1 W/m-°C. They consist of thermally conductive but chemically benign material such as ZnO or BN dispersed in a suitable carrier (such as silicone oil). This method of cooling is possible with FCSB technology because the solder joints are located and protected beneath the chip, compared with delicate wire and TAB bonds which are exposed on the surface. The cooling advantage provided by FCSB/TC package with two standard back bond configurations (wire bond connections) is shown in Figure 9-47. The conditions are the same for all cases: chip and substrate size, chip thermal dissipation and external convection cooling. The external package resistance (heatsink resistance) also is the same for all cases compared. Figure 9-47 show the relative differences in internal thermal resistance for the conditions noted. This difference is responsible for the ability of the FCSB/TC configuration to maintain the chip junction temperature (T_j) at 69°C when compared to 81°C and 99°C for the back bond cavity down and cavity up configurations respectively. Although the values may change with other examples, the trends are the same because FCSB technology provides the lowest internal thermal resistance, taking into account the total path length and bulk thermal conductivities of materials in the heat path between the chip and heatsink. FCSB/TC have a short chip to heatsink path, typically about 5 mils. As noted earlier, heat also is dissipated through the solder joints into the substrate. The ability to dissipate heat from both sides of the chip provides FCSB technology with a distinct advantage.

The cooling requirements of some demanding applications have been achieved by directly contacting the chip backside with a spring loaded piston [86], [99]-[100]. Pistons have been used on modules containing up to 133 chips [81]. Chip front side (I/O side) thermal enhancement also is achieved with high thermal conductivity ceramics such as silicon carbide and aluminum nitride [87], [101] whose conductivity is 6 to 10 times better than alumina (Al_2O_3).

Electrical

Chip connections have a significant impact on the electrical performance of MCM packages. It is desirable to minimize the chip to substrate lead length and distances between chips. FCSB technology allows chips to be placed in close proximity on MCMs, with a greater than 90% packaging density possible [102].



Conditions

- Alumina (Al_2O_3) substrate, 32mm
- Chip, 10mm dissipating 5 watts
- Convection cooling, 150 W/m^2 - C°

Figure 9-47 Cooling comparison between FCSB mounted devices with thermal compound and back bonded chips. (Courtesy of R. Sherif [104].)

Trends of increased device performance coupled with smaller chip and substrate feature dimensions is forcing a movement away from wire bonded, single chip modules to MCM utilizing TAB connections and ultimately to flip chip solder bump systems to minimize inductance. A key factor is the optimal geometry of FCSB joints (3 mil high, 5 mil diameter typically) in comparison to the smaller cross sections and much longer length characteristic of TAB and wired bond connections. The inductance of TAB and wire bond connections is typically in the range of 0.5 - 2.0 nH compared with FCSB joints, which are several orders of magnitude less. Lead inductance affects the number of simultaneous switch operations a chip driver may execute. The lead inductance also influences the maximum operating frequency. Under the best conditions this frequency is judged to be several hundred MHz for TAB. Although better than wire bonds, it does not compare favorably with FCSB joints, which operate in the GHz frequency range. Also, since FCSB are area array connections, power is brought close to the point of application when compared with either TAB or wire bonded chips where power is brought to the chip periphery. This also results in flip chips having shorter chip lines. These factors all serve to enhance the electrical performance of FCSB mounted devices.

9.6.7 Summary

This section notes that FCSB technology is capable of I/O extendibility, greatest connection density, pre-attach burn-in, test and rework. Thus, progress achieved in moving from a VLSI to an ULSI era makes it apparent FCSB technology is well suited to meet both I/O and performance (electrical, thermal etc.) requirements for current and future MCM applications. However, wider practice will require significant effort on the part of the electronics packaging industry and device manufacturers since the infrastructure currently is not in place.

9.7 SUMMARY

Wire bond, TAB and flip chip (FCSB) will be the dominant chip connection technologies in the foreseeable future. Each of these chip connection methods has advantages and disadvantages. The technologies are maturing partly to meet industry requirements and also because of their competition with one another. Tables 9-21, 9-22 and 9-23 provide a current comparison of the relative strengths and weaknesses for each of these technologies. The categories and ratings in these tables are summarized below and are based on previously discussed material.

9.7.1 Cost

Many factors contribute to the cost of an MCM. The choice of a chip connection technology can have a significant influence on the cost of an MCM. Some of the considerations involved that impact on the total cost are discussed in this section and in the following section where achievable performance also is related to cost.

Direct Costs

Each chip connection technology has costs directly related to its implementation and use. Equipment and tooling costs are examples of direct costs and are common to all three methods. Wire bond equipment is relatively mature in comparison to TAB and FCSB, resulting in an infrastructure that provides equipment at more competitive costs (refer to Table 9-21). TAB equipment can be very expensive due to the fine geometries, the degree of automation and the number of qualified vendors. FCSB assembly equipment is available, but is automated to a lesser degree than that of wire bond or TAB. Die specific tooling costs are a major issue with TAB, especially for custom, non-standard designs. Material and device preparation costs also are important. TAB and FCSB may require expensive wafer bumping processes. TAB also requires tape frames. These costs are reflected in the chip connection and device preparation cost comparisons in Table 9-23.

Indirect costs

The choice of chip connection technology also can affect other design and manufacturing considerations. Methods of approaching the cooling of MCMs are affected by the choice of flip TAB or flip chip over wire bond. Densities of chips on a module, or the ratio of silicon area to package area also has an effect on cost. FCSB has the potential for the densest assembly while conventional TAB and wire bond designs consume the most area. Repair and rework

Table 9-21 Comparison of Chip to Package Connection Technologies.

CONNECTION TECHNOLOGY	INDUSTRY USE	MATURITY	INFRA-STRUCTURE	DESIGN COMPLEXITY
Wire Bond	H	H	H	L
TAB	M	M	M	M
FCSB	L	L	L	M

L = Low M = Medium H = High

requirements, especially in MCMs, are critical factors when making a choice between these technologies. While TAB and FCSB are amenable to repair in certain configurations, wire bond does not lend itself easily or inexpensively to repair processes, as compared in Table 9-22.

9.7.2 Performance

The choice of a chip connection technology can have a significant influence on the performance of an MCM. Thermal performance, electrical performance and reliability are components of the general performance of the module and are related closely to the cost the MCM manufacturer is willing to pay to achieve specific performance objectives.

Thermal performance is influenced by the type of chip connection technology selected. Wire bond and conventional TAB configurations allow for the die to be cooled by conducting heat through the package. Flip TAB and FCSB, on the other hand, may require novel and potentially more expensive cooling methods. The desired thermal performance probably can be achieved if the MCM manufacturer is willing incur the expense.

Electrical performance also is a function of the chip connection choice. High performance applications may rule out wire bond because of the limitations of the fine wire. Conventional TAB also may be limited, unless a two metal layer tape is used. Only flip TAB with its short leads and FCSB with its leadless connections offer the greatest potential for optimizing electrical performance.

Wire bond is the most understood of the chip connection technologies. The documented reliability of wire bonded packages is plentiful. TAB, and especially FCSB, are not as mature or as well documented. Special considerations need to be made for the various metallurgical systems and materials allowed by TAB.

Table 9-22 Comparison of Chip to Package Connection Technology Design Issues.

CONNECTION TECHNOLOGY	DEVICE AVAIL-ABILITY	SILICON DENSITY	I/O DENSITY	ELECTRICAL PERFORMANCE	REWORK POTENTIAL
Wire Bond	H	L/M	L/M	L	L
TAB	L	L/M	L/M	M	M
FCSB	L	H	H	H	M

L = Low M = Medium H = High

Table 9-23 Comparison of Chip to Package Connection Technology Costs Assuming Low Volume MCM Production.

CONNECTION TECHNOLOGY	CAPITAL INVESTMENT	TOOLING COST	CONNECTION COST	DEVICE PREPARATION COST
Wire Bond	L	L	L	L
TAB	H	H	H	H
FCSB	M	L	M	H

L = Low M = Medium H = High

FCSB must be concerned with material thermal mismatches that contribute to thermally induced mechanical stress.

9.7.3 Tradeoffs: Making a Choice

Choosing a chip connection technology involves compromise. There is no one perfect solution for all products. The ultimate choice over time can be determined only by matching requirements against the available solutions. Careful evaluation of the cost tradeoffs, together with the performance requirements, should be made during the design phase of the MCM so that an optimum choice can be made. High performance applications may be able to absorb the higher cost associated with TAB and flip chip for the added performance. Low performance or cost driven products, on the other hand, may look solely at the cost issue.

References

1 T. L. Hodson, "Bonding Alternatives for Multichip Modules," *Electr. Packaging and Prod.*, vol. 32, no 4, pp. 38-42, April 1992.

2 R. K. Shukla, N. P. Mencinger, "A Critical Review of VLSI Die-Attachment in High Reliability Applications,' *Solid State Technology*, vol. 28, no. 7, pp. 67-74, July 1985.

3 T. A. Bishop, "A Review of Repairable Die Attach for CuPI MCM," *Proc. Internat. Electr. Packaging Conf.*, (Marlborough MA), pp. 77-89, 1990.

4 P. Burggraaf, "Polyimides in Miroelectronics," *Semi. Internat.*, vol. 11, no. 4, pp. 15-17, March 1988.

5 H. K. Charles, Jr., "Applications of Adhesives and Sealants in Electronic Packaging," *Proc. Internat. Soc. Hybrid Microelectr. Conf.*, (Orlando FL) pp. 139-146, 1991.

6 M. N. Nguyen, "Low Stress Silver-Glass Die Attach Material," *IEEE Trans. Components, Hybrids and Manufact. Techn.*, vol. 13, no. 3, pp. 478-483, Sept. 1990.

7 T. L. Herrington, G. G. Ferrier, H. L. Smith, "Low Temperature Silver Glass Die Attach Material," *Johnson Matthey Corp.*, San Diego CA.

8 S. M. Dershem, C. E. Hoge, "Effects of Materials and Processing on the Reliability of Silver-Glass Die Attach Pastes," *Quantum Materials, Inc.* 9988 Via Pasar, San Diego, CA 92126, 1992.

9 E. Razon, Y. Tal, "Silver Glass Die Attach," *Kulicke and Soffa Industries, Inc.*, Israel Div.

10 J. E. Ireland, "Epoxy Bleedout in Ceramic Chip Carriers," *Medtronic Corporation*, Tempe, AZ.

11 M. L. White, "The Removal of Die Bond Epoxy Bleed Material by Oxygen Plasma," *Bell Laboratories*, 555 Union Boulevard, Allentown, PA 18103.

12 R. C. Benson, T. E. Phillips, N. DeHaas, "Volatile Species from Conductive Die Attach Adhesives," *IEEE Trans. Components, Hybrids and Manufact. Techn.*, vol. 12, no. 4, pp. 571-577, Nov. 1989.

13 T. E. Phillips, *et al.*, "Silver-Induced Volatile Species Generation from Conductive Die Attach Adhesives," *Proc. 42nd ECTC*, (San Diego CA), pp. 225-233, May 1992.

14 R. H. Pater, *et al.*, "LaRC-RP41: A Tough, High Composite Matrix," *NASA Tech. Briefs*, vol. 15, no. 9, pp. 82, Sept. 1991.

15 R. K. Lowry, K. L. Hanley, "Volatility Behavior of Organic Die Adhesive Materials," *Harris Semiconductor*, P. O. Box 883, Melbourne, FL 32901.

16 C. J. Lee, J. Chang, "Reworkable Die Attachment Adhesives for Multichip Modules," *Proc. ISHM*, (Orlando FL), pp. 110-114, 1991.

17 MMT Staff, "Selection Guide: Die Bonding," *Microelectr. Manuf. and Testing*, pp. 16-20, July 1990.

18 K. Chung, *et al.*, "MCM Die Attachment Using Low-Stress, Thermally Conductive Epoxies, *Proc., IEPS*, (San Diego CA), pp. 167-175, Sept. 1991.

19 M. Akhavain, "Cost Effective Multi-Chip Modules Utilizing High Temperature Cofired Multilayer Ceramic Technology," *NEPCON West Proc.*, (Anaheim CA), pp. 1422-1430, Feb. 1992.

20 "Monitoring of Ultrasonic Wire Bonding Machines," *Hewlett-Packard Application Note 393*, (1990).

21 P. H. Singer, "Hybrid Wire Bonding Advances," *Semiconductor Internat.*, vol. 7, no. 7, pp. 22-25, 1984.

22 D. Herrell, H. Hashemi, R. Smith, W. Weigler, "Electrical Performance of Tape Automated Bonding," *Handbook of Tape Automated Bonding*, J. H. Lau ed., New York: Van Nostrand Reinhold, 1992, pp. 44-84.

23 E. Philofsky, "Design Limits when Using Gold-Aluminum Bonds," *9th Annual Proc. Reliability Physics*, (Las Vegas NV), pp. 11-16 (1971).

24 H. A. Schafft, "Testing and Fabrication of Wire-Bond Electrical Connections—A Comprehensive Survey," *NBS Tech. Note 726*, 1972.

25 G. G. Harman, "Reliability and Yield Problems of Wire Bonding in Microelectronics—The Application of Materials and Interface Science," Reston VA: ISHM, 1989.

26 H. B. Eisenberg, I. Jensen, "When Control Charting is Not Enough: A Wirebond Process Improvement Experience," *Proc. Internat. Soc. for Hybrid Microelectr. Conf.*, (Chicago IL), pp. 61-66, 1990.

27 K. Fukuta, T. Tsuda, T. Maeda, "Optimization of Tape Carrier Materials," *Proc. 4th Internat. TAB Symp.*, (San Jose CA), pp. 283-312, 1992.

28 A comprehensive handbook on TAB technology is, J. H. Lau, ed., *Handbook of Tape Automated Bonding*, New York: Van Nostrand Reinhold, 1992.

29 A. S. Rose, F. E. Scheline, T. V. Sikina, "Metallurgical Considerations for Beam Tape Assembly," *Proc. IEEE 27th Electr. Computer Conf.*, (Arlington VA), pp. 130-134, 1977.

30 K. Hatada, *et al.*, "New Film Carrier Assembly Technology - Transfer Bump TAB," *Proc. IEEE Internat. Electrical Manuf. Techn. Symp.*, pp. 122-127, 1986.

31 C. Montgomery, "A Low Cost High Performance Interconnection Technique for GaAs Devices," *1991 VLSI Packaging Workshop*, (Tucson AZ), pp. 9-12, Sept. 1991.

32 R. D. Pendse, *et al.*, "Demountable TAB - A New Path for TAB Technology," *Proc. 4th Internat. TAB Symp.*, (San Jose CA), pp. 9-24, 1992.

33 A. Reubin, B. Bohrn, R. Smith, "Transfer Molded TAB Package," *Proc. NEPCON West*, (Anaheim CA), pp. 894-905, 1990.

34 J. H. Lau, S. J. Erasmus, D. W. Rice, "Overview of Tape Automated Bonding," *Circuit World*, vol.1, no. 2, pp. 5-24, 1990.

35 J. M. Smith, S. M. Stuhlbarg, "Hybrid Microcircuit Tape Chip Carrier Materials/Processing Tradeoffs," *IEEE Trans. Parts Hybrid Packaging*, vol. PHP-13, no. 3, pp. 257-268, 1977.

36 V. Iyer, R. Pendse, "A Novel Inner Lead Bonding Technique for TAB," *Proc. 40th Electrical Computer and Techn. Conf.*, (Las Vegas NV), pp. 754-756, 1990.

37 D. Walshak, "The Effects of Bonder Parameters on Au-Au TAB Inner Lead Bonding," *Proc. NEPCON West '90*, (Anaheim CA), pp. 906-913, 1990.

38 C. J. Speerschneider, J. M. Lee, "Solder Bump Reflow Tape Automated Bonding," *Proc. ASM Internat. Electr. Material and Processing Cong.*, (Philadelphia PA), pp. 7-12, 1989.

39 D. A. Field, "A Tin Based TAB Assembly Process," *Proc. 1991 IEEE Internat. Manuf. Techn. Symp.*, (San Francisco CA), pp. 31-35, 1991.

40 P. J. Spletter, "Au/Au Inner Lead Bonding with a Laser," *Proc. 4th ITAB Symposium*, (San Jose CA), pp. 58-71, 1992.

41 E. Zakel, G. Azdasht, H. Reichl, "Investigations of Laser Soldered TAB Inner Lead Bonds," *Proc. 41st Electr. and Computer Techn. Conf.*, (Atlanta GA), pp. 497-506, 1991.

42 A. Emamjomeh, *et al.*, "TAB Inner Lead Process Characterization for Single Point Laser Bonding," *Proc. 1991 IEEE Internat. Manuf. Techn. Symp.* (San Francisco CA), pp. 21-26, 1991.

43 E. Zakel, R. Leutenbauer, H. Reichl, "Reliability of Thermally Aged Au and Sn Plated Copper Leads for TAB Inner Lead Bonding," *Proc. 41st Electr. and Computer Techn. Conf.*, (Atlanta GA), pp. 866-876, 1991.

44 W. T. Chen, *et al*,. "A Fundamental Study of the Tape Automated Bonding Process," *J. of Electr. Packaging*, vol. 113, no. 3, p. 216, 1991.

45 M. Wong, "TAB Outer Lead Bonding and SMT," *Proc. NEPCON West*, (Anaheim CA), pp. 785-789, 1988.

46 J. M. Altendorf, "SMT-Compatible, High Yield TAB Outer Lead Bonding Process," *Proc. NEPCON West*, (Anaheim CA). 233-249, 1989.

47 D. J. Arnone, J. J. Tong, "Determination of Placement Accuracy Requirements for Fine Pitch and Very Fine Pitch Component Assembly," *Proc. NEPCON West*, (Anaheim CA), p. 2154, 1991.

48 M. Y. F. Kou, "Assessing Fine Pitch Placement Machine From a System Standpoint," *Proc. NEPCON West*, (Anaheim CA), p. 2125, 1991.

49 L. Fox, "High Performance TAB Package," *1991 VLSI Packaging Workshop*, (Scottsdale AZ), pp. 21-22, Sept. 1991.

50 J. Deeny, D. Halbert, L. Nobi, "TAB as a High Lead Count PGA Replacement," *Proc. Internat. Electr. Packaging Conf.*, (Marlborough MA), pp. 660-669, 1990.

51 M. Mahalingam, J. A. Andrews, "TAB vs. Wire Bond - Relative Thermal Performance," *Trans. IEEE-CHMT*, vol. CHMT-8, no. 4, pp. 490-499, 1985.

52 D. Herrell, D. Carey, "High Frequency Performance of TAB," *Trans. IEEE CHMT*, vol 10, no.2, pp. 199-203, 1987.

53 J. H. Lau, D. W. Rice, G. Harkins, "Thermal Stress Analysis of TAB Packages and Interconnections," *Trans. IEEE CHMT*, vol. CHMT-13, no.1, pp. 152-187, 1990.

54 M. J. Bertram, "Repair Method for Solder Reflow of TAB OLB," *Proc. 2nd ITAB*, (San Jose CA), pp. 147-164, 1990.

55 D. S. Soane, Z. Martynenko, "Encapsulation and Packaging of Integrated Circuits," *Polymers in Microelectronics*, New York: Elsevier, 1989, p. 213.

56 K. C. Norris, A. H. Landzberg, "Reliability of Controlled Collapse Interconnections," *IBM J. Res. Develop.* vol. 13, no. 3, p. 266, 1969.

57 L.S. Goldman, "Geometric Optimization of Controlled Collapse Interconnections," *IBM J. Res. Develop.*, vol. 13, no. 3, p. 251, 1969.

58 R. J. Wassink, "Solder Alloys," *Soldering in Electronics, 2nd ed.*, Scotland: Electrochemical Publications Limited, 1989, p. 135.

59 T. Kawanobe, *et al.*, "Solder Bump Fabrication by Electrochemical Method for Flip Chip Interconnection," *Proc. 31st Electr. Computer Conf.*, (Atlanta GA), p. 149, May 1981.

60 W. G. Bader, "Dissolution of Au, Ag, Pd, Pt, Cu and Ni in a Molten Tin-Lead Solder," *Welding J.*, vol. 48, no. 12, p. 551, 1969.

61 P. J. Kay, C. A. MacKay, "Barrier Layers Against Diffusion," *Trans. Inst. Metal Finishing*, vol. 54, part 4, p. 169, 1979.

62 K. N. Tu, R. D. Thompson, "Kinetics of Interfacial Reaction in Bimetallic Cu-Sn Thin Films," *Acta Met.*, vol.30, p. 947, 1982.

63 D. Olsen, R. Wright, H. Berg, "Effects of Intermetallics on the Reliability of Tin Coated Cu, Ag and Ni Parts," *13th Annual Proc. of the Reliability Physics Symp.*, (Las Vegas NV), p. 80, April 1980.

64 A. vanderDrift, W. G. Gelling, A. Rademakers, "Integrated Circuits with Leads in Flexible Tape," *Phillips Technical Review,* vol. 34, no. 4, p. 85, 1974.

65 E. Yung, I. Turlik, "Electroplated Solder Joints for Flip Chip Applications," *IEEE Trans. on CHMT*, vol. 14, no. 3, p. 549, 1991.

66 D. Chang, *et al.*, "Design Considerations for the Implementation of Anisotropic Conductive Adhesive Interconnection," *Proc. NEPCON West*, (Anaheim CA), 1992.

67 H. Yoshigahara, *et al.*, "Conductive Epoxies for Attachment of Surface Mount Devices," *4th Internat. SAMPE Electr. Conf.*, (Albuquerque NM), p. 255, 1990.

68 B. Sun, *Connection Technology*, vol. 20, no. 8, p. 31, Aug. 1988.

69 I. Tsukagoshi, *et al.*, *Hitachi Technical Report*, No. 16. p. 23, 1991.

70 K. J. Puttlitz, "Corrosion of Pb-50 In Flip Chip Interconnections Exposed to Harsh Environment," *IEEE Trans. CHMT.*, vol. 13, no. 1, pp. 183-193, March 1990.

71 V. C. Marcotte, N. G. Koopman, P. A. Totta, "Review of Flip Chip Bonding," *Proc. 2nd ASM Internat. Elec. Mat. Proc. Congress*, pp. 73-81, April 1989.

72 C. Dostal, M. Woods eds., *Electronic Materials Handbook*, Materials Park,6 OH: ASM International, 1989, vol. 1, p. 231.

73 A. H. Kumar, "Corrosion of the Joining Metallurgy in Multilayer Ceramic Substrates During Processing," *Proc. 40th Elec. Comp. Techn. Conf.*, vol. 1, pp. 89-93, May 1990.

74 K. J. Puttlitz, "Flip Chip Replacement within the Constraints Imposed by Multilayer Ceramic (MLC) Modules," *J. of Elec. Mat.*, vol. 13, no. 1, pp. 29-46, Jan. 1984.

75 K. G. Heinen, W. H. Schroen, O. R. Edwards, *et al.*, "Multichip Assembly with Flipped Integrated Circuits," *IEEE Trans. CHMT.*, vol. 12, no. 4, pp. 650-657, Dec. 1989.

76 F. W. Kuleszsa, R.H. Estes, "Solderless Flip Chip Technology," *Hybrid Circuit Techn.*, pp. 24-27, Feb. 1992.

77 K. Hatada, H. Fujimoto, "A New LSI Bonding Technology, Micron Bump Bonding Technology," *Proc. IEEE 39th Elec. Comp. Conf.*, pp. 45-49, May 1989.

78 M. Bartschat, "An Automated Flip-Chip Assembly Technology for Advanced VLSI Packaging," *Proc. 38th Elec. Comp. Conf.*, pp. 335-341, May 1988.

79 B. Inpyn, "Practial Considerations for Tape Automated Bonding," *Circuits Manufacturing*, vol. 6, pp. 42-43, June 1989.

80 N. Basavanhally, S. Gahr, J. Liu, H. Nguyen, "Flip Chip Repair Process," *Proc. 41st Elec. Comp. Techn. Conf.*, pp. 779-782, May 1991.

81 P. Lin, J. Lee, S. Im, "Design Considerations for a Flip Chip Joining Technique," *Solid State Techn.*, pp. 48-54, July 1970.

82 K. C. Norris, A. H. Landzberg, "Reliability of Controlled Collapse Interconnections," *IBM J. Res. & Devel.*, vol. 13, no. 3, pp. 266-271, May 1969.

83 P. A. Tobias, N. A. Sinclair, A. S. Van, "The Reliability of Controlled-Collapse Solder LSI Interconnections," *Proc. 1976 Internat. Microelectr. Symp.*, ISHM, pp. 360-363, Oct. 1976.

84 S. Ahmed, R. Tummala, "Packaging Technology for IBM's Latest Mainframe Computers," *Adv. Techn. Workshop '91 on Multichip Modules*, (Oqunquist ME), June 1991.

85 S. E. Greer, "Low-Expansivity Organic Substrate for Flip-Chip Bonding," *Proc. 28th Elec. Comp. Conf.*, pp. 166-171, May 1978.

86 R. R. Tummala, H. Potts, S. Ahmed, "Packaging Technology for IBM System 390/ES9000, Models 820 and 900 Mainframe Computers," *Proc. 41st Elec. Comp. Techn. Conf.*, pp. 682-688, May 1991.

87 K. K. Hagge, "Ultra-Reliable Packaging for Silicon-on-Silicon WSI," *IEEE Trans. CHMT.*, vol. CHMT-12, no. 2, pp. 170-179, June 1989.

88 M. Matsui, S. Sasaki, T. Ohsaki, "VLSI Chip Interconnection Technology Using Stacked Solder Bumps," *Proc. IEEE 37th Elec. Comp. Conf.*, pp. 573-578, May 1987.

89 K. Puttlitz, T. Reiley, "Centrifugal Stretching of C-4 Joints," *IBM Tech. Disc. Bulletin*, vol. 27, no. 10B, pp. 6198-6200, March 1985.

90 R. Satoh, *et al.*, "Development of a New Micro-Solder Bonding Method for VLSI," *Proc. 3rd Internat. Elec. Pkg. Conf.*, pp. 455-461, Oct. 1983.

91 K. Beckham, *et al.*, "Solder Interconnection Structure for Joining Semiconductor Devices to Substrates that have Improved Fatigue Life, and Process for Making," U.S. Patent No. 4,604,644, Aug. 5, 1986.

92 H. M. Tong, L. Mok, K. Grebe, H. Yeh, "Parylene Encapsulation of Ceramic Packages for Liquid Nitrogen Application," *Proc. 40th Elec. Comp. Techn. Conf.*, vol. 1, pp. 345-350, May 1990.

93 D. Suryanarayana, R. Hsiao, T. P. Gall, J. M. McCreary, "Flip Chip Solder Bump Fatigue Life Enhanced by Polymer Encapsulation," *Proc. 40th Elec. Comp. Techn. Conf.*, vol. 1, pp. 338-344.

94 R. T. Howard, "Packaging Reliability and How to Define and Measure It," *Proc. 32nd Elec. Comp. Conf.*, pp. 376-384, May 1982.

95 R. T. Howard, "Optimization of Indium-Lead Alloys for Controlled Collapse Chip Connection Application," *IBM J. Res. and Develop.*, vol. 13, no. 1, pp. 29-46, Jan. 1984.

96 K. J. Lodge, D. J. Pedder, "The Impact of Packaging on the Reliability of Flip Chip Solder Bonded Devices," *Proc. 40th Elec. Comp. Techn. Conf.*, vol. 1, pp. 470-476, May 1990.

97 H. Ewalds, R. Wanhill, *Fracture Mechanics*, London: Edward Arnold Publishers Ltd., 1985, p. 180.

98 A. Bar-Cohen, "Thermal Management of Air and Liquid-Cooled Multichip Modules," *Proc. 23rd ASME Nat. Heat Transfer Conf.*, (Denver CO),, 1985.

99 S. Oktay, R. J. Hanneman, A. Bar-Cohen, "High Heat From a Small Package," *Mech. Engr.*, vol. 108, no. 3, pp. 36-42, March 1986.

100 T. Hatsuda, H. Doi, T. Hayasida, "Thermal Strains in Flip-Chip Joints of Die-Bonded Packages," *Proc. Internat. Electr. Pkg. Conf.*, IEPS, vol. 2, pp. 826-832, Sept. 1991.

101 T. Watari, H. Murano, "Packaging Technology for the NEC SX Supercomputer," *Proc. 35th Elec. Comp. Conf.*, pp. 192-198, May 1985.

102 T. Horton, "MCM Driving Forces, Applications and Future Directions," *Proc. NEPCON West*, (Anaheim CA), pp. 487-494, 1991

103 K. Puttlitz, "An Overview of Flip Chip Replacement Technology on MLC Multichip Modules," *Proc Internat. Electr. Packaging Conf.*, vol. 2, pp. 909-928, Sept. 1991.

104 R. Sherif, Private communication, 1992.

10

MCM-TO-PRINTED WIRING BOARD (SECOND LEVEL) CONNECTION TECHNOLOGY OPTIONS

Alan D. Knight

10.1 INTRODUCTION

Electronic packaging involves many levels of connections for components, subsystems and systems. The connections between the individual die and the multichip module (MCM) substrate forms a level 1 connection. A level 2 connection joins the MCM to the printed wiring board (PWB), either directly or indirectly. This should not be confused with level 1 (MCM) and level 2 (PWB) packages, as defined in Chapters 1 and 3.

There are two basic purposes of the level 2 connection. The first is to provide the electrical paths between the MCM and PWB to transfer electrical signals. The second is to provide the electrical power required for the chip to function [1]-[3].

A connector is often used to provide a level 2 connection from the MCM to the PWB. As well as providing the electrical path, a connector incorporates structures that help to provide a solid contact between the conductors on the MCM and PWB and the conductors in the connector. A connector is manufactured separately from the PWB and MCM and used when all three are assembled together.

In some cases, the second level connection is relied upon for physical attachment and for retention of the MCM to the PWB. This eliminates the cost of attachment hardware and assembly labor. However, this approach should be considered only for small, lightweight MCMs used in shock and vibration free environments and is not recommended for most applications.

Because level 2 connections are the interface for the MCM to the rest of the system, the method of connection requires careful consideration. The choice of connection and connector technology directly influences electrical performance, manufacturability, reliability and cost. As with the MCM itself, the technology chosen must address appropriately any mechanical and electrical performance issues that affect signal transmission quality.

The ideal connection is cost effective, easily processed and compatible with the thermal performance and test requirements of the MCM. Just as important, the connection should be as electrically transparent as possible - signals through the connection should not be distorted, attenuated, reflected or delayed. The ideal connection also would be capable of providing for a very large number of connections. The number of I/O pins of a large MCM can be very high, even greater than 10,000.

10.1.1 Second Level Connection Alternatives

The basic second level connection choices are summarized in Figure 10-1. The choices are differentiated by the geometry, the lead type and the connection method. There typically are two geometrical patterns used to distribute the connections: area array and edge array.

An area array geometry places connections on a square grid over the bottom, or most of the bottom, surface of the MCM. An identical pattern is placed on the PWB. Since the entire bottom surface may be used for the connections, this pattern provides either the greatest quantity of connections or the maximum spacing between connections. However, all connections are concealed beneath the MCM and, as such, may not be visually inspectable. An exception may be an MCM with pins soldered to through-holes in the PWB.

Edge arrays have all connections distributed on one, two, three or four edges of the MCM in one or more rows per edge. The same or expanded pattern is provided on the PWB. This configuration offers the largest variety of the connection means, most of which are inspectable. However, as the pin count increases, edge arrays generally require closer spacing of the connections than an area array.

For either area or edge geometries, the MCM package can include leads or be leadless. If leads are used, they are either formed flat leads or pins. Formed leads are used only with a single row edge geometry. (Forming means that the leads are bent.) Leaded MCMs are similar to surface mount packages used for

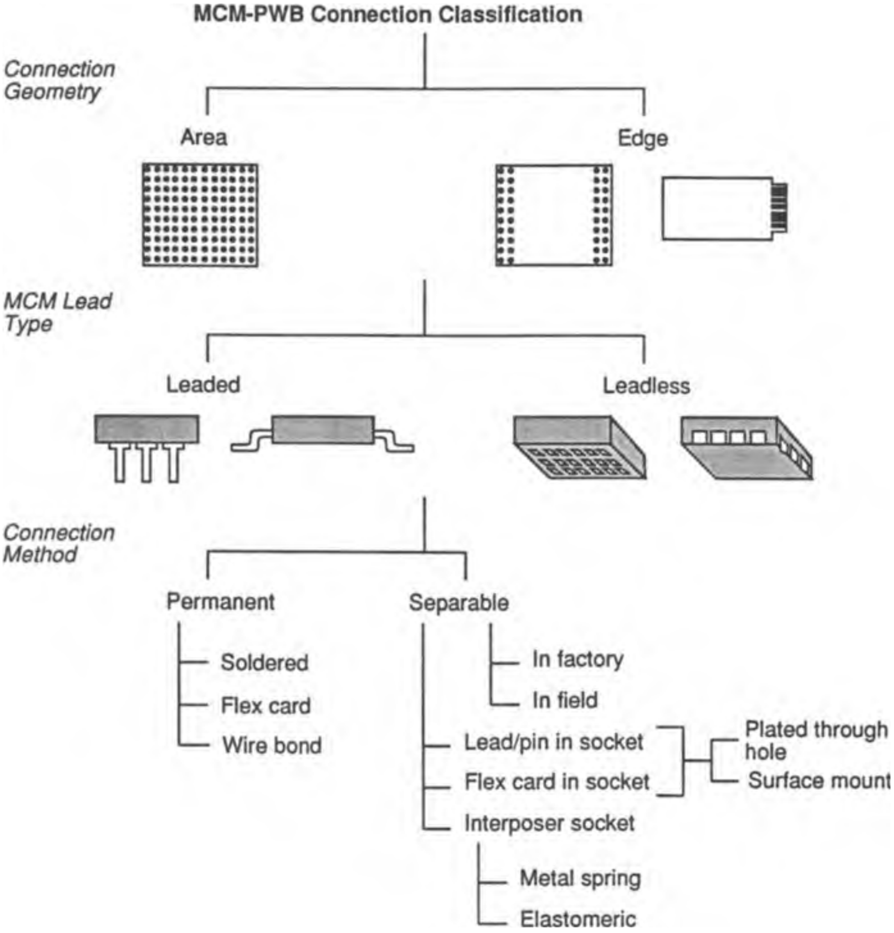


Figure 10-1 Classification of MCM-PWB (second level) connection geometries and methods.

single chip packages such as quad flat packs (QFP). Pinned packages are often found in the pin grid array (PGA) format. A common form of leadless array is the land grid array (LGA) in which pads are arrayed over the bottom of the package. Leadless packages also may be configured in the edge array format.

The methods by which MCMs are attached to PWBs fall into two basic categories: permanent and separable connections [4]-[5].

Permanent connections are typified by the use of a metallic bond, such as solder, to create the electrical connection between the MCM and PWB. As such, the MCM becomes an integral part of the PWB and requires special factory operations for removal and replacement. Thus replacing the MCM in the field usually requires replacement of the PWB as well. There are a number of additional options. In addition to direct solder attachment, the MCM also can be connected by a flex card (a flexible PWB). In this case, the wires at each end of the flex card are attached permanently to the MCM and PWB. Another alternative is to wire bond the MCM directly to the PWB. If the MCM has pins, these pins are inserted into matching holes on the PWB and soldered into place. A new method of attachment known as a pressure contact connection, uses gold bumps on the MCM contact pads and matching gold pads on the PWB. The MCM is glued to the PWB using a special adhesive which contracts as it cures and maintains a compressive force on the MCM-PWB interface. In many cases, leaded and leadless MCMs are soldered to the surface of the board (surface mount). This allows MCMs to be placed on both sides of the PWB. Out of the above methods, only the flex card is an example of a connector; the others consist of components (leads and pads) manufactured together with the MCM and/or the PWB.

Separable connections use a metallic element of some sort creating an electrical connection through mechanical means. Accordingly, the MCM can be installed and separated from the PWB at anytime during assembly and test or after shipment and installation. All separable connection methods require the use of a connector, which, together with any retention mechanism, is referred to as a socket.

Separable connectors are classified in terms of where the connection is made and by their type. With an in-factory separable connector, the connection is made, broken and reformed only in clean conditions. With a field-separable connector, the connection can be broken and reformed anywhere. The connector must have the ability to displace particulates and penetrate deposited films to ensure a good and reliable connection.

There are three types of separable connectors. In the first type, the MCM leads or pins are fitted into corresponding gaps or holes in the socket. In the second type, flex cards which have been connected permanently to the MCM are used to make a separable connection. Instead of making a flex card connection by soldering the leads to the PWB, the leads are force fit into a socket on the board. Both of these types of sockets are classified further by how they attach to the PWB. They either are pinned and soldered into holes in the board, much like a permanently mounted PGA is soldered, or they are soldered into place using surface mount techniques, just like a QFP or LGA is soldered into place. One large advantage of the surface mount alternative is that it allows MCMs to be placed on both sides of a PWB. Another alternative to mounting the socket to the PWB is to make the pins larger than the holes in the PWB and force fit the socket body onto the PWB without a soldering step.

The third type of separable contact or socket involves use of an interposer [6]. In an interposer, the connections are provided by a sheet of contacts placed (interposed) between matching pads on the MCM and PWB. This is used with leadless MCMs that have I/Os on the bottom such as in a LGA. The MCM and interposer are clamped to the PWB. Since there are no pins to fit into the socket, interposers allow MCMs to be used on both sides of the PWB. There are two types of interposers: those based on metal springs and those based on wires supported in a rubber material (elastomer).

Another connection scheme, not shown in Figure 10-1, is to bypass the MCM directly and combine the first level and second level connections by attaching the bare die directly to the PWB. This is referred to as chip on board (COB).

This chapter discusses the basic mechanical and electrical issues that affect connection choice. The different connection methods then are discussed in further detail.

10.2 BASIC ISSUES AFFECTING CONNECTION CHOICE

10.2.1 Mechanical and Materials Issues with Permanent Soldered Connections

There are a number of considerations that must be taken into account when soldering leaded and leadless MCMs onto PWBs.

The choice between peripheral or area array connection usually is based on the nature of the MCM package. Since peripheral pads are placed along the edges of the MCM, there is less space for connections than in array area configurations which make use of the entire bottom side of the package. In MCM applications where there are often a large number of connections, the peripheral approach usually requires tight center line spacing. Currently it is difficult to space solder connections closer than about 0.3 mm (12 mils). Area arrays allow more connections in the same area with wider pad-to-pad separation.

Area array contacts are not without their disadvantages, however. Traditionally, there has been some concern about the wisdom of soldering leadless MCMs where the pads are underneath the MCM [7]. However, the technology of reflow soldering LGA modules has improved dramatically in recent years. The techniques used to reflow solder LGAs are similar to the techniques used to reflow solder bump arrays described in Sections 9.5 and 9.6.

The soldered joint must be able to absorb the stresses that arise due to different levels of thermally induced expansions of the MCM and PWB. A coefficient of thermal expansion that closely matches that of the board and the MCM is highly desirable. The larger the package and substrate, the higher the levels of stress. Stress levels are reduced by using a leaded MCM with compliant (bendable) leads. This is why the leads in leaded packages have bends (gull wing or J-lead). Straight leads have little compliancy. In leadless packages the only compliancy provided is that of the solder itself. Again, the amount of compliancy there is small. This is discussed in Chapter 9 with respect to flip chip solder bump arrays.

Most of the thermal stress arises during the actual reflow step when the solder is heated and flows to wet (cover) the metal surfaces being connected. The temperature must be well below the glass transition temperature of the PWB. If the dielectric materials in the PWB start turning into glass during the reflow step, the thermal stresses will be much higher.

10.2.2 Separable Connector Interface Physics

The physics of electrical contacts in separable connections is a broad, complex subject [8]. The following discussion highlights some of the main issues dealing with MCM connections. For a more complete discussion, consult reference [1]. There are two areas of concern in the quality of an electrical connection: the DC performance and the AC performance. The DC performance relates to the ability of the connection to carry the required current with a minimum voltage drop. The AC performance relates to the distortion or delay experienced by high frequency signals as they cross the connection. DC issues are discussed first in terms of bulk properties of the metal and the surface properties of the interface.

Contact Conductivity

The contact must have low bulk resistance, meaning a substantial cross section of high conductivity metal. A high conductivity (low resistivity) material minimizes the voltage drop across the contact and minimizes the possibility of incorrect signal levels. Generally, materials with good spring properties do not have good conductivity. Beryllium-copper or beryllium-nickel, for example, often chosen for their spring properties and their ability to withstand high

Table 10-1 Resistivities of Typical Contact Metals.

CONTACT METAL	BULK RESISTIVITY (μΩ-cm)
Silver	1.6
Copper	1.7
Gold	2.2
Aluminum	2.8
Tungsten	5.3
Nickel	6.8
Titanium	43
Beryllium-copper	5.6 to 8.5*
Lead	22
Tin	11
Tin/lead solder (63/37)	15
Tin/lead solder (5/95)	21.5

* Depending on composition.

temperatures over the long term, have low conductivity.

The need for low resistivity contacts is increasing due to shrinking semiconductor geometries, bringing about a reduction in IC working voltage - from 5 V to 3.3 V and even as low as 1.6 V - while leaving the current the same. As the working voltage decreases, the voltage drop across the contacts becomes a greater percentage of this voltage and thus is a greater concern. Table 10-1 lists the resistivities of metals typically used in both separable and permanent connections [8].

Contact Reliability: Force and Wipe

In addition to good bulk properties, a good contact must have a low contact resistance. Whenever two dissimilar surfaces are placed into contact, there will be some resistance to the flow of current from one material to the other. A reliable contact interface must have a very low end of life electrical resistance - in the range of 10 - 20 mΩ at 60°C after 15 years. To obtain a low and stable contact resistance, the contact must penetrate and fracture any oxide or nonconductive film which has been deposited, chemisorbed (chemically bound adsorbed layer) or adsorbed on contact surfaces or the substrate contact area. This surface layer is removed by a combination of wipe and normal force or normal force alone. Typically, the two contacts are pressed together and motion occurs in the plane of the contact surface. This combination of motion and normal force cleans the surface just as one cleans a dirty dish with a brush. If there is no lateral motion, cleaning is completed through normal force alone (a force perpendicular to the plane of the surface). This force cleans the surface through a compressing action similar to squeezing dirty water from a household sponge. As a normal force requirement alone does not account properly for the effects of wipe, connector requirements often are expressed in terms of the stress in the contact surface. A contact surface stress in excess of 150,000 psi is required to clean the surface. Insertion force is the force required to insert the part into the socket and is created by a combination of normal force and wipe [9].

The required force or stress is highly dependent on the metallurgy of the contact surfaces. Metallurgies are classified as either reactive metallurgies, such as copper, brass, beryllium-copper, tin or tin-lead, or as noble metallurgies, such as gold, palladium or palladium-nickel.

Reactive metallurgies are low cost and easy to obtain. However, they react with atmospheric gases producing non-conducting compounds such as oxides. When these compounds are present, not only is it difficult to make a good metal to metal contact during insertion, but they add to the thickness of the base material. This increased volume increases further if airborne moisture is absorbed. If this volumetric expansion occurs after the material is mated, it may eventually overcome the normal force of contact, producing intermittent or open circuit contact. This corrosion of the contact is one example of a time dependent failure [9].

A second potential problem can arise if the connector is plugged and unplugged many times. This repeated action grinds the non-conductive material into the base metal, increasing the resistance. Also, channels, through which gas and moisture reach the reactive material under the final contact point, may be created. This also leads to intermittent contact or non-contact, excessive heating and total connection failure. This is another time dependent failure mechanism.

Noble metals, while more costly to obtain, have the advantage of providing a non-reactive, stable surface for the connection. However, when placed directly onto copper or alloys of copper, commonly used for the PWB printed wiring patterns and contact pads, the noble metal is absorbed by the copper, producing a reactive contact surface thus negating the rationale for using noble metals. Therefore, diffusion barriers, such as nickel, should be present between the base contact material and the final metal surface. A further caution must be exercised when using noble metals. The presence of the noble metal does not guarantee high performance reliability. The metal coating must be thick enough to provide a continuous surface in the wear track as well as at the final contact area. Wear-through due to multiple insertions or minute holes in the plating, called "pores," may result in corrosive products of the base or diffusion barrier metal to appear on the surface.

Contact surfaces often are formed by adding another complication. Microscopic hydrogen bubbles may cling to the metal being plated, with the result that the plating is deposited around the bubbles. As the plating thickness increases, the bubble migrates away from the substrate. A fistula or pin hole (Figure 10-2) results, frequently containing plating solution, wetting agent or organic complex or residues. Osmotic pressure often holds these deposits temporarily inside the bubbles, even at high temperatures.

When the pin holes are small, the contacts may pass accelerated environmental tests such as humidity and temperature cycling, porosity tests and fuming nitric acid tests. Two or more years later, when the moisture inside the bubbles has dissipated, gases enter the void, initiating corrosion. The accumulation of corrosion products separates the contacts, even those with high contact forces. The combined action of wipe and normal force displaces surface layers of the metal and burnishes it to seal the pores in the mating area and eliminate the problem. During subsequent mating cycles, wiping cleans and seals the mating areas [10].

The second reason for wipe is the need, particularly in dirty environments, for dislodging, removing and displacing dust particles. IBM, for example, has

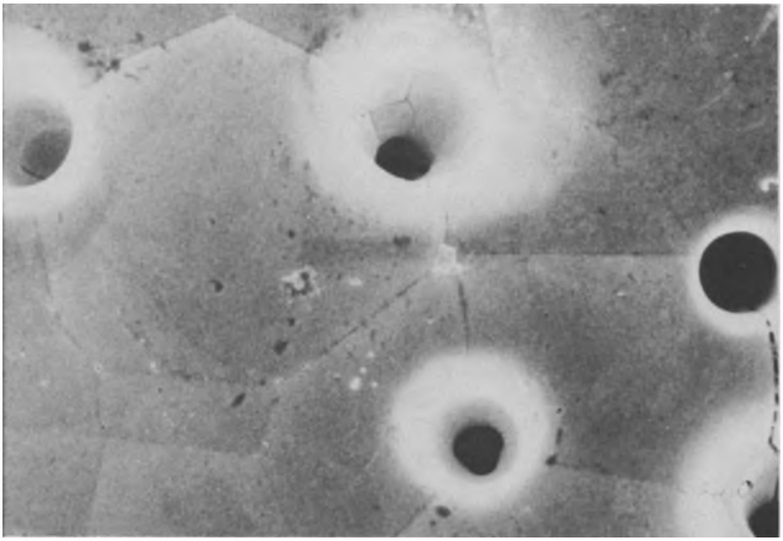


Figure 10-2 Pinholes in plating form a failure mechanism that can be controlled by contact wipe. (Courtesy of AMP.)

recognized the threat of such particles to system performance and has created artificial dust to help judge contact reliability. To pass the test, contacts must have a local high stress point and substantial wipe in a very small area.

Wipe is provided in different ways. For example, when a socket is used where pins are plugged into holes, the wipe and normal force often are provided by having the pins slightly larger than the holes. In this case, if all the pins are inserted at once, considerable insertion force might be required to press fit all of the pins simultaneously. Conversely, considerable withdrawal forces might be required to separate the parts. If these forces are large, manual insertion and withdrawal might be difficult and mechanical aids such as cams, wedges and screws must be provided. In most cases, wipe is provided by the displacement of either the MCM lead or a spring loaded contact inside the socket in such a way as to provide a scrubbing action between the two surfaces.

When little wipe occurs, the required normal force depends on the configuration of the contact forces. Two clean, flat contact surfaces (pads) require considerably more force for a reliable metal-to-metal contact than a half sphere shape on a flat pad. A configuration with half spheres provided on both surfaces requires the least force. In the last two configurations, geometry acts to concentrate the force onto a small area, creating a high surface stress. Typical force values for a well designed contact with geometric force concentrators are 100 grams per contact when noble metals are used and 200 grams per contact for non-noble metals, though success is achieved with smaller forces.

Failure to address the need for normal force and wipe leads to reliability problems. Consider, for example, a proposal consisting of a gold plated flexible circuit in direct contact through an elastomer with a gold plated PWB surface. This approach provides low forces and no wipe, based on the theory that gold to gold interfaces require only low contact force and no wipe. Such an approach, however, results in low reliability and high field failures, since contaminants and oxidation products migrate into micro pinholes formed during the plating process, particularly if products sit in inventory for any length of time. A classic, and painful, example occurred in a program which incorporated low force connection of gold plated flat pack leads to a gold plated multilayer PWB. The system passed all accelerated tests, but failed after less than two years in the field. Other approaches, such as anisotropic and conductive gel connectors, generally have poor reliability because of the absence of sufficient force and wipe.

Contact Compliance in Interposers

The contact must have a high degree of mechanical compliance, with an interposer separable connector, to compensate for any lack of coplanarity and flatness in the MCM and the PWB. The contact also must be compressed easily. A working range of 0.02" (0.5 mm) is desirable, although difficult to achieve.

In a metal spring interposer, this is achieved through bending the spring. With an elastomer interposer, this is achieved through compression of the rubber-like material that supports the metal wires. Interposers are discussed in detail later in the chapter.

10.2.3 Electrical Performance Issues

AC Performance

A driving force in electronics is higher packaging densities at all levels, from semiconductors to systems. This increased density is required to reduce the overall size of the end product and to decrease the electrical path lengths in the circuit. As the clock speed of digital electronics increases, there is less time for signal propagation between components. Clock speed in a digital circuit is akin to frequency in an analog circuit and, as the clock speed increases, a digital circuit takes on radio frequency circuit attributes requiring consideration of electrical impedance and coupling of circuit elements. What constitutes high speed depends on the application, with the electrical length of the circuit playing a part in the determination.

Impedance and crosstalk control are of special concern at high speeds; in many cases, both are treated through similar design approaches. Impedance mismatches and disturbances cause reflections, distorting signals. To ensure maximum transfer of a signal, all the components in the transmission path should have the same characteristic impedance. Crosstalk, which couples energy from one signal line to another through either capacitive or inductive coupling, must also be considered. The AC performance issues in a second level connection are similar to those on the MCM itself (see Chapter 11).

Transmission Line Performance

The performance of an connection is related to its resistance, capacitance and inductance. The capacitance and inductance of the lines affect the high speed performance. Interactions occur both between signals and from signals to power and ground lines. As the length of a line increases, both its capacitance and inductance also increase resulting in signal delays and distortions. At high speeds, circuit designers must treat a circuit as a transmission line, using the distributed rather than the lumped properties of all components in the design [11].

Transmission line rules generally become applicable when the length of a conductor approaches one-quarter of the signal wavelength, $\lambda/4$, or the length of a conductor approaches 1/100 of the rise length of a digital signal. (Rise length is the distance a signal travels during its rise time.)

The length of a circuit element, whether a wire, circuit trace or socket contact, is described in electrical terms. In high speed applications, electrical length is the major consideration - how the length of the circuit element compares to the signal wavelength. If the electrical length of the circuit element is long, impedance control becomes important. For a very short length, conventional lumped element circuit analysis is sufficient. In between those lengths is a gray area where the decision depends on other system-related factors, such as acceptable noise margins and cost and performance tradeoffs.

Rise time is the "turn on" time of a pulse, measured between the 10% and 90% (or 20% and 80%) points of amplitude. The rise time is used to calculate an equivalent frequency for the signal based on Fourier time-to-frequency analysis. This frequency then is used in the modeling of the performance of the electrical system.

In a high speed system, the designer has two options: either to use a controlled impedance connection or to keep the circuit element short enough to eliminate the need for impedance matching. If the circuit element is considerably smaller than one-quarter of a wavelength, impedance matching may not be necessary. This is the approach usually taken for second level connections as it is less expensive. However, the circuit element still presents a discontinuity in the transmission lines passing through the connection. This discontinuity disturbs the high speed signals propagating through it and reflects part of the signal. This is measured through time domain reflectometry (TDR). With this measurement technique, a high speed signal is sent down a transmission line with a second level connection placed somewhere near the middle and the resulting signals are observed at both ends of the line using an oscilloscope.

Crosstalk

Crosstalk results when signals from one line on a circuit couple onto another line. Crosstalk causes signal loss on one line, as well as contaminating the signal on the other line. Energy transferred from one line to the other produces a signal on the second line which might create an unexpected change of state. Crosstalk increases with the frequency of the signal and decreases as the separation between the two lines grows. Crosstalk is minimized by ensuring that lines operating at high frequency are placed sufficiently far apart or by using shielding to isolate the lines. Power and ground lines are used to shield the signal lines by placing each signal line close to a ground or power line. The coupling then occurs between the signal line and ground, rather than between adjacent signal lines.

Propagation Delay

Propagation delay is the time a signal takes to pass through a section of circuit. Differences in propagation delay between lines cause signal skewing in parallel transmissions typically found in bus configurations. A good connection delays all lines in a bus by the same amount to minimize skew. Since the path through a second level connection tends to be short, propagation delay is not usually a problem. As the frequency of systems increases, this becomes a greater issue and drives the size of connectors smaller.

Electromagnetic Modeling

The complex electromagnetic environments of a high speed, high density system means that verifying logic design is not enough. Component electromagnetic properties must be modeled early in the design process and tested to verify the modeling. Modeling must include the transmission line effects of the connection: reflections, crosstalk and any other effects that may disrupt a signal passing through the connection.

The connection does not exist electromagnetically alone in space. The influence of surroundings must be included, such as the capacitance of a plated through-hole solder connection, and the effect of the signals themselves. Signal rise times, for example, directly influence the number and placement of ground lines. As the signal rise times decrease, a larger number of power and ground connections usually are needed to minimize ground bounce during switching.

Properly modeling a connection is a difficult task, particularly with a socket, because of its complex structure. The model should evaluate the distributed values of capacitance, inductance and resistance, and should allow multicontact analysis through a matrix model [12]. (Contact capacitive and inductive coupling effects are not limited to adjacent contacts. A matrix model also studies the effects of simultaneous switching on several lines.)

A frequent and significant mistake is to ignore or to simplify the connection when using a circuit analysis program such as SPICE. Because the connection can have a dramatic impact on system performance, it must be included.

10.3 BASIC APPROACHES TO MCM LEVEL TWO CONNECTIONS

As described earlier an MCM can be permanently and directly attached to the PWB or attached using a socket, so that they are easily separated again. Cost evaluations should consider component costs and total system costs. System costs include such factors as field replaceablity, upgradeability, troubleshooting and inventory costs in both the factory and repair depot. The cost of the lowest field replaceable unit (FRU) varies significantly depending on whether it is an entire printed circuit assembly or an easily swapped MCM. The reliability of a socketed MCM depends on a number of factors, including the quality of the mating connections and the protection provided from atmospheric elements.

Permanent, direct attachment generally simplifies reliability considerations since there are fewer contact interfaces and the interfaces are in more intimate contact. It is this same feature that makes field replacement of a directly attached MCM very difficult or impossible. Directly attached MCMs still face harsh environments during assembly on the soldering line and in the field. In some cases, the MCM may be encapsulated after mounting on the PWB to protect the unit and its connections.

10.3.1 Direct Attachment Through Soldering

Soldering is used with PGAs, and peripheral leaded and leadless MCM packages. PGA pins are soldered to plated through-holes, while peripheral devices are surface mounted. The pins on a PGA often are spaced at 0.05" or 0.1" intervals while the leads on a surface mount device often are spaced at intervals as small as 0.012".

Because this permanent, direct connection avoids the cost of a socket, it presents obvious component part cost savings. Any decrease in the number of connections tends to improve the reliability. Direct attachment through solder presents only one connection, while a socket presents two interfaces.

Another attractive feature of soldering is that it allows processing of MCM connections simultaneously with other components. The growing popularity of surface mount components and processing requires that the MCM withstand the high temperatures of reflow soldering. In a typical surface mount process, a solder paste is applied to copper lands on the board. Components are mounted to the board with the pads or leads on the paste. When the board is then passed through the soldering equipment, the solder melts (reflows) around the contacts and cools to form a solder joint. An alternative method of soldering is wave soldering which passes a wave of molten solder over the areas to be connected. Wave soldering does not subject the MCM or PWB to the full temperature of the molten solder and may be preferable in some cases. In all soldering, if a flux is used to promote surface wetting, it must be carefully cleaned off to prevent corrosion of the connection.

A device does not have to be a surface mount component to be processed by surface mount assembly techniques. Even a pinned area array MCM, which uses through-hole mounting, can be processed in a reflow soldering line. To be compatible with generic surface mount processing, a component must have the following characteristics:

- Ability to be mounted to the PWB by pick and place equipment
- Ability to be mounted on EIA-481 compliant tape
- Ability to be reflow soldered by infrared or forced air convection
- Ability to withstand aqueous solvent cleaning
- Allow visual inspection of solder joints

A component with characteristics different from these usually requires special equipment or techniques. Leadless area array MCMs obviously do not allow visual inspection of solder joints, although inspection can be performed with x-rays. Mass production soldering of such packages has been difficult. On the other hand, MCMs often are used in applications where standard high volume processing techniques are not required because of the specialized nature of the application. This lessens the need for compatibility with standard surface mount assembly equipment.

The decision to solder affects the choice of plating for the MCM leads. The most common form of solder is tin-lead which easily joins to gold and other noble metals. The most preferable arrangement is to have the leads of the MCM pre-tinned (coated) with tin-lead solder to simplify the joining process. In some configurations, it may be necessary to solder more than one connection between the MCM and PWB. For example, a lead frame may be soldered first to the MCM and then to the PWB. In this case, a soldering temperature hierarchy is required. The solder used to join the lead frame to the MCM is chosen to melt at a higher temperature than that used for the leads to PWB connection so that soldering the leads to the PWB do not melt the connections to the MCM.

Wire Bonded Modules

This form of permanent connection is an extension of chip wire bonding techniques (see Section 9.3). The MCM typically is bonded to the PWB with an adhesive. Wire bonding equipment is used to connect module edge mounted pads with corresponding pads on the PWB. The number of connections is limited by the pitch of the pads on the PWB. Visual inspection and rework of the resulting connection is performed readily and usually no post process cleaning is required. However, wire bonds from the MCM top surface to the PWB can be fairly long. This results in extra inductance in the connection and adversely affects the electrical performance as described earlier [14].

Flex Connected Modules

In flex connected modules, the discrete wires used in the wire bonded example are replaced with a series of flex ribbon cables. These ribbons act as jumper cables and usually consist of parallel copper lines placed on a thin dielectric film, with or without a fanout pattern at the PWB end. The circuit lines generally terminate in pads at each end of the flex which reflect the PWB and MCM pad patterns.

Typically assembly starts with attaching the pads on one edge of the flex to the MCM. The MCM and flex assembly then is precisely aligned and bonded to the PWB with adhesive. The pads on the free edge of the flex circuit are attached to the PWB, usually accomplished by soldering or compression bonding techniques (see Section 9.3). When solder attachment is used, attention must be given to the cleaning process to assure complete removal of all flux residues between the flex and the MCM or PWB for long term joint reliability.

Inspection and rework are performed readily with flex connected modules. Reliability is generally high, since thermal expansion differences are absorbed by flexure of the flex circuit. Electrical performance is enhanced through the use of alternating signal and reference (power or ground) lines on the flex as well as through the addition of a ground plane.

10.3.2 Chip-on-Board Connections

The COB technology merges level 1 and level 2 connections by attaching chips directly and permanently to the PWB [15]. This approach has been used for years in applications requiring maximum space conservation. Intel Corp., for example, used COB technology to create an IBM XT-compatible personal computer on a board slightly larger than a credit card. The microprocessor was mounted directly to the board.

The COB approach is accomplished in one of three ways, each described in Chapter 9. These are: wire bonding, flip chip solder bump and tape automated bonding.

COB applications typically encapsulate the chip in a protective silicone or epoxy material to prevent moisture and gases from degrading bonding wires and joints. While encapsulants do not provide a true hermetic seal, they provide very good moisture sealing in high reliability applications. As thermal vias are rarely provided in COB PWBs, the thermal performance generally is poor. Both the materials above and below the chip are poor conductors of heat. Due to limitations on the minimum pad size available on a PWB, COB generally is not used for chips with a large number of I/Os since long wires would be needed to connect from the chip to the much larger pattern on the PWB.

10.3.3 Separable Connections (Sockets)

Sockets permit removal and reinstallation of the MCM to test the device, isolate parts of a circuit or to swap modules. They also permit PWB integrity to be checked prior to installing expensive components. Sockets allow fast, easy field replacement of faulty MCMs and eliminate the need for special desoldering equipment.

If new MCMs are not available in full quantity, sockets allow the manufacturing process to proceed through testing by using a small quantity of on-hand modules to test the boards. When late deliveries arrive, they are tested and plugged into the board. Since MCM cost is relatively high, inventory costs are minimized by installing the MCM at the last moment rather than carrying it in inventory for months. Additionally, portions of a system are assembled at lower cost offshore and imported at a low tariff. Assembly of the MCM and its socket can then be completed domestically.

Sockets for MCMs include those designed for peripheral and area substrates, both leaded and leadless. The selection of a socket depends largely on the MCM packaging approach. Some sockets are essentially standard IC sockets, while others are designed especially for MCMs and similar high density, high pin count applications.

Peripheral Versus Area Array

The heart of the socket is the electrical contact, which serves as the electrical interface between the MCM and PWB. While several contact geometries have evolved to meet packaging needs, a given contact design can be adapted to many different socket configurations. Typical contact types include leaf springs and pogo pins. A leaf spring contact consists of a conductive member configured as a cantilever beam which applies force through the displacement of the beam. A pogo contact consists of a pin assembly containing an integral coiled spring which applies the required force.

The choice between peripheral and area array MCM and socket combinations presents several considerations. For a given number of MCM I/Os, the peripheral package presents the least area for contacts. Consequently, one either enlarges the package for large lead counts or places the contacts on close center lines.

Tight center lines are more rigorous in their requirements for alignment and maintenance of exacting tolerances. But peripheral packages have an attractive advantage: they permit easy access for visual inspection and for probes during testing. The area array package, with its greater area for I/O, relaxes the demand for tight center lines at contact points but at the expense of access for inspection or testing.

10.3.4 Sockets for Leadless MCM Substrates

Spring Contact Sockets for Peripheral and Area Array MCMs

The INTERPOSER™ contact (Figure 10-3) is an example of a basic contact structure that forms the basis for a variety of sockets [16]. It is a beryllium

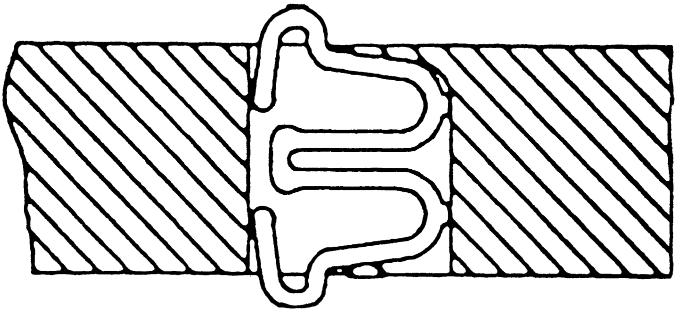


Figure 10-3 Basic INTERPOSER™ contact. (Courtesy of AMP.)

copper contact, featuring a double spring design with two internal and two external wiping surfaces that provide wiping action on the module, the board and the contact. When the contacts are interposed between the two mating surfaces and compressed, they exert an outward force to achieve and maintain a gas tight interconnection. The contacts are assembled into a rugged liquid crystal polymer housing which forms the basic module from which a variety of sockets are made. This solderless approach relies on a top and bottom plate on either side of the board held together by screws. A strong cover plate is required for a socket with a large number of contacts.

A smaller alternative of this socket for high speed, high density applications features a 0.090" contact suited to both peripheral and area applications [4]. Compression forces are reduced. The contacts are beryllium-copper with plated gold over nickel for mating with gold-plated interfaces, or made using palladium alloy (PAL-7). An example is an area array connector made by AMP using an 86 × 86 array with contacts on a true 1.2 mm grid. Each individual contact resides in its own plastic housing, and these housings are packed into a metal housing. At 120 grams per contact, the connector requires over 2,000 pounds for compression, presenting two problems. First, sufficiently rigid hardware is required with screws or clamps to compress the contacts. Second, the MCM substrate may be deformed or flexed by the clamping forces sufficiently to lower the reliability of the chip connections.

Contact Array Sockets for Area Array MCMs

The socket seen in Figure 10-4 shows an approach that further reduces contact height and required compression forces, complementing the drive toward higher

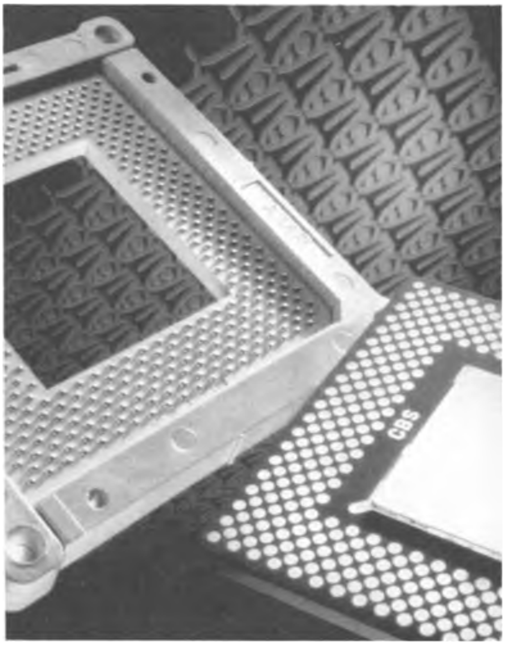


Figure 10-4 Contact area array socket. (Courtesy of AMP.)

density systems packaging by reducing the size of the socket. The compressed contacts are less than 0.010" high, and compression forces are low at 50 grams. The low compression force makes the contact suited only to noble metal applications.

At the heart of the socket is the contact array (Figure 10-5). A sheet of beryllium copper is chemically etched to produce the contact pattern. The contacts are then formed and plated with gold over nickel. Plastic insulating sheets are laminated to the array; this laminated structure is punched to isolate the contacts electrically. This provides an array of contacts on an 0.050" grid. The contact array is assembled into the socket [16].

A new type of interposer manufactured by Cinch consists of an array of conductive "fuzz" buttons which bridge the gap between pads on the MCM and matching pads on the PWB. The fuzz button (called a button board connector) accommodates variations in planarity and insures a good contact. The button board connector is made by twisting a small diameter wire of a known length and having it collapse on itself to form a small cylindrical shape of dense wire (Figure 10-6), which is then inserted into the pre-drilled holes in the polymer material to form the connector (Figure 10-7) [17]-[18].

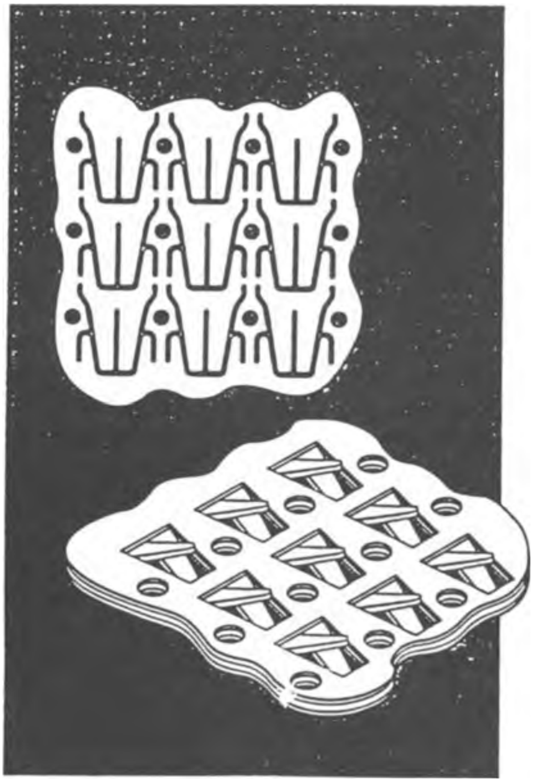
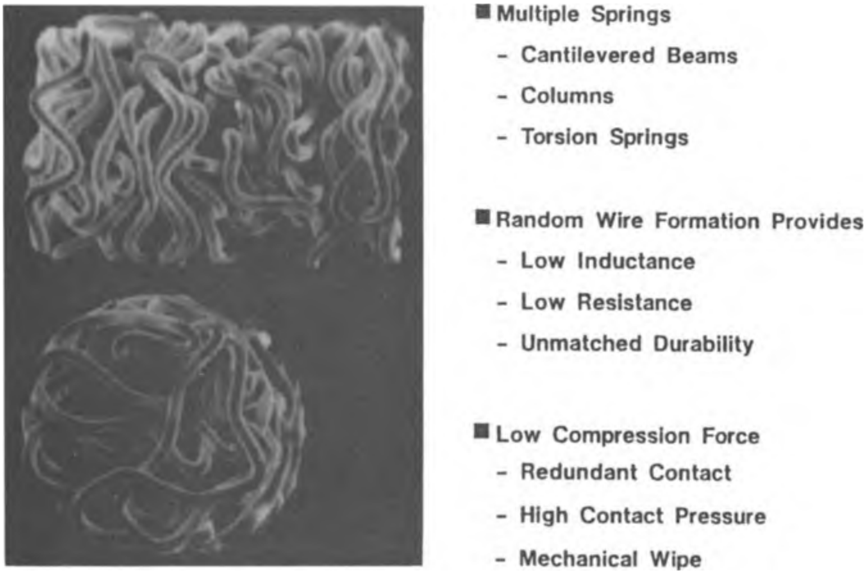


Figure 10-5 AMPFLAT™ socket contact array. (Courtesy of AMP.)

SIMM Sockets for Peripheral MCMs

The socket used for single in-line memory modules (SIMM) also may be used in laminate and some cofired MCM applications. Figure 10-8 shows a representative SIMM socket. SIMM contacts must furnish enough force (over 200 grams) to provide a reliable, gas tight connection for tin-to-tin interfaces. Because of the popularity of SIMMs, SIMM sockets are manufactured in very high volumes. As a result, they offer the benefits of low cost and configuration and option flexibility such as vertical, angled or right angle SIMM insertion; one or two rows; 22 to 84 positions; gold or tin-lead plating; and 0.100" or 0.050" center lines.



- Multiple Springs
  - Cantilevered Beams
  - Columns
  - Torsion Springs
- Random Wire Formation Provides
  - Low Inductance
  - Low Resistance
  - Unmatched Durability
- Low Compression Force
  - Redundant Contact
  - High Contact Pressure
  - Mechanical Wipe

Figure 10-6 Button board connector showing connections to random wire configuration. (Courtesy of Cinch Connectors, a division of Labinal Components and Systems.)

usually feature a high temperature plastic that allows processing in surface mount soldering lines. SIMM sockets eliminate secondary operations such as wave soldering.

The main disadvantage of the SIMM socket in MCM applications is the low pin count of standard sockets (84 maximum) which may not suffice for many MCMs. Because only one side of the MCM can plug into the socket, routing on the module also is more complex. In addition, there are practical limits to extending the socket for larger sizes; the assembly becomes too long. These sockets also present a longer electrical path, making them less suited to high speed applications.

Elastomeric Connectors for Peripheral and Area Array MCMs

Elastomeric connectors consist of two components. First is the elastomeric element, a rubber material in which unidirectional conductors are formed. The elastomeric element is a form of interposer. Second is the rest of the connector

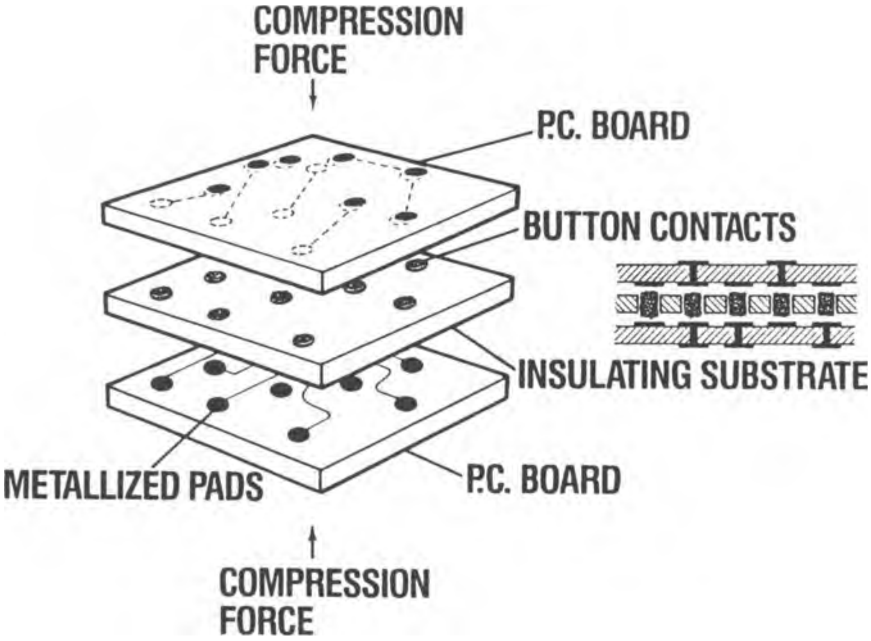


Figure 10-7 CIN:: Apse™ button board connections to printed wiring board. (Courtesy of Cinch Connectors, a division of Labinal Components and Systems.)

elastomeric element is a form of interposer. Second is the rest of the connector consisting of a holder to retain the elastomeric element. The hardware used to mount the element, attach it to the MCM and PWB and provide the necessary force is also part of the holder [19].

The basic advantages of elastomeric connectors include their high contact density, low profile, very low resistance, tolerance of wide temperature ranges and hostile environments, shock and vibration resistance, ease of installation and replacement and packaging versatility. Disadvantages arise from the fact that most elastomeric element conductors provide for minimal or no wiping action (through a compressing action) and that the elasticity of the rubber reduces with age. Normal force also is reduced as the elastomer ages. This is referred to as permanent set. As a result, the initial deflection (compression) and the initial force provided to do this must be increased in compensation for reduction of normal force [20]-[21].

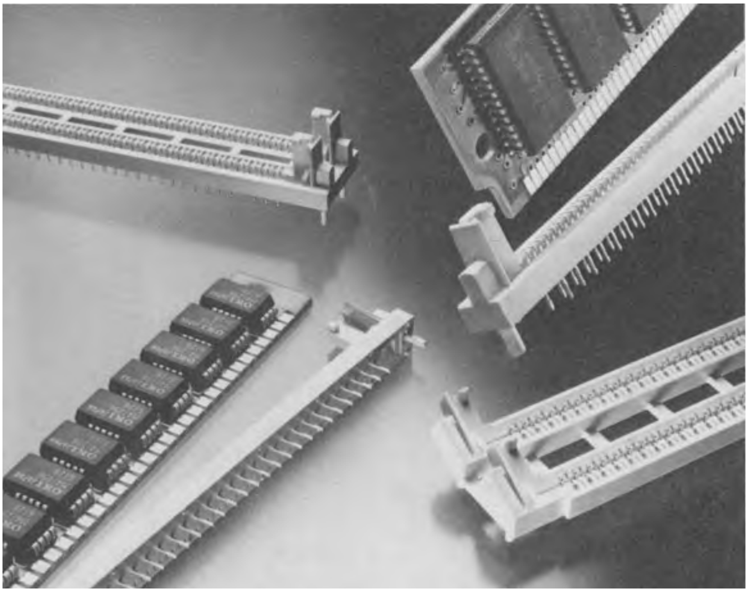


Figure 10-8 SIMM (single in-line memory module) socket. (Courtesy of AMP).

Resistance and current carrying ability in these connectors vary with the geometry of the elastomeric elements and the substrate pads. The choice of element type depends on the maximum resistance that can be tolerated, the contact area, the possible clamping pressure and the cost [22].

The elastomer interposer connector group encompasses numerous configurations and compositions. Almost all use a form of silicone rubber, either as a solid or a foam of open or closed cell construction. This is used not only as the contact carrier but also to provide the energy storage that springs provide in more traditional contact systems. To create conductivity, a host of materials are used in either particle or continuous conductor form [23].

Conductive particles dispersed in the silicone rubber are one form of elastomeric interposer connectors. Typically the particles are copper or nickel and may be plated with silver or gold.

Compression of a thin silicone sheet (20 mils or less) with the correct density of particles displaces the silicone between the particles until adjacent particles touch. This creates conductive paths only in the compressed axis direction, making an electrical connection from the top to the bottom of the elastomeric sheet. Because of the thinness of the silicone sheet, this embodiment has an operating range of only a few mils, requiring very flat connecting surfaces.

Subjecting nickel-based particles to a magnetic field of varying intensity during solidification of the silicone produces a somewhat ordered, columnar structure of particles in the area of high magnetic flux. Generally, these columns can be placed on the desired grid creating conductive paths. These silicone sheets tend to be relatively thicker to accommodate the column of particles and, therefore, have a larger operating range.

Similar fabrication techniques have been used to produce elastomer connectors with solid wires embedded in the silicone rubber. These wires are made of copper, silver or gold and are straight or buckled. Upon compression, the film of silicone rubber, which may be over the ends of the wires, must be displaced to provide conductivity. These systems have the advantage of eliminating the multiple contact points between particles and, therefore, exhibit lower bulk resistance per contact path.

Layered elastomeric interposers are constructed by laminating alternating sheets of silicone and conductive layers together, creating rows of conductive paths. The conductive layer may be a solid conductor or a series of parallel line segments on a typical fine pitch of 5 mils. Thus, many parallel paths are used per contact providing redundancy. Constructions made by Elastomeric Technologies, illustrated in Figures 10-9a and 10-9b, typify this connector technique.

Continuous, parallel conductive strips, placed on a this sheet of silicone or polyimide and wrapped around a silicone rubber core, producing another version of elastomeric interposer connectors. These are normally produced on a fine line pitch to produce multiple conductive paths per contact. Typical values are 5 mil lines on 10 mil centers.

The metal on elastomer (MOE) is a robust elastomeric connector. The element consists of multiple rows of gold ribbon conductors across a silicone rubber medium along the z-axis. The components which make up the connector are silicone rubber and solid gold. These materials are resistant to harsh environments which make the second level connection ideal for space and military applications. The Matrix MOE connector which relies almost entirely on silicone rubber to apply the contact force useful when required to make a connection between nonplanar surfaces [17].

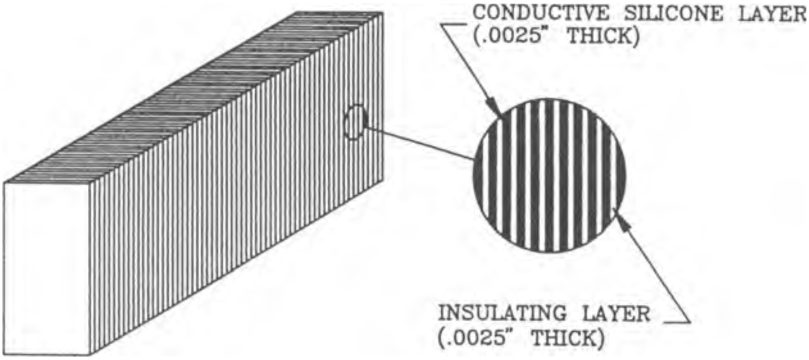


Figure 10-9a Layered elastomeric connector formed by alternating layers of conductive and nonconductive silicone rubber.

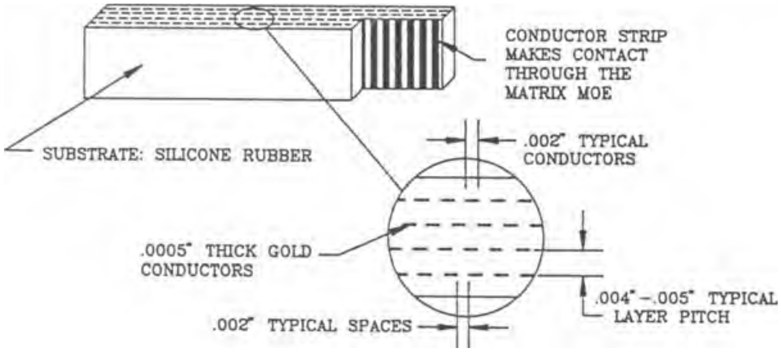


Figure 10-9b Details of a matrix metal on elastomer (MOE) construction. (Courtesy of Elastomeric Technologies.)

Another variety of elastomeric connector, produced by the Rogers Corp., uses rigid metal contacts formed in a S-type shape. These individual contacts are inserted into slits in a sheet of elastomer on the desired grid. During compression, the S-shape contact rotates, depressing the elastomer beneath the top and bottom portions of the S, producing both wipe and normal force on both top and bottom interfaces.

Most elastomer interposer connectors require high compressive loading to produce a limited operating range. They also have poor debris and film penetration ability. Normal force is lost with time as the result of stress relaxation (from 15 - 50% at end of life) and the particle filled elastomers have greater internal resistance than the solid conductor versions.

Connector holders generally are used to position mating substrates before clamping pressure is applied to the elastomer element. Such holders also provide a deflection stop; the slot in the holder must be wide enough to accept the deflection in width and length of the element. Parallel, coplanar, perpendicular and offset configurations are available. The height of the holder is equivalent to the desired substrate separation. Mechanical features built into the holder fit into matching features on substrates to ensure alignment to pads.

Most commonly, deflection force is applied with machine screws or eyelets. Plastic and metal edge clips or molded plastic snaps provide clamping along the entire side of the assembly. Pressure is the predominant method for establishing and maintaining contact, so, a stable rigid backing is needed to ensure reliable, long term operation. Thus a stable rigid back up plate is always essential. The PWB is not rigid enough to provide this function.

In many designs, connections are made as the package is assembled. The clamping force is maintained by latching with interference fits of the mating package sections. If the connector is unlikely to be removed, plastic posts can be molded onto the package or adhesives can be used to hold parts in place for surface mount applications.

The contact area is pliable and conforms to irregular surfaces. The resiliency of the connector core also helps to create a sealed area around the contacts providing environmental protection at the contact interface. A wide variety of cross sectional configurations and lengths are available. The circuit is assembled into a housing and offered as a completed connector assembly.

The basic elastomeric strip easily is adapted to a variety of connector styles for both area and peripheral devices. Figure 10-10 shows an example of a socket designed for area array applications, while Figure 10-11 shows a socket for peripheral applications.

Because the elastomeric member can have a free height diameter as small as 0.040", it presents a short electrical path causing minimum signal disturbance. The main concern with such connectors is contact resistance. A wide variety of elastomeric member or conductor combinations have been devised with resistances ranging from low tens of mΩ to several hundred mΩ. Low resistance designs are available, for example, offering a resistance of 20 mΩ, inductance of 1 nH and a capacitance of 0.2 pF. A typical propagation delay is 12 ps [23].

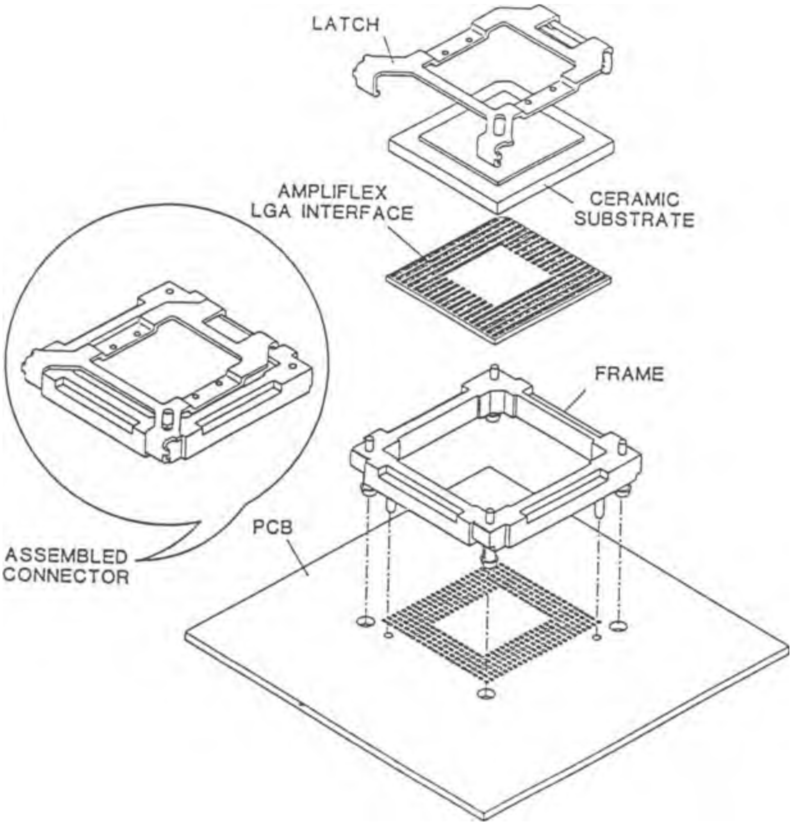


Figure 10-10 Elastomeric socket for leadless area array applications. (Courtesy of AMP)

10.3.5 Sockets for Leaded MCM Substrates

Leaded MCM Sockets for Peripheral MCMs

A leaded MCM socket, developed for leaded single and MCM packages, is shown in Figure 10-12. This socket doesn't need the clamping forces associated with other sockets discussed above. Instead, it uses flexible circuit leads that extend from the perimeter of the MCM over tuning fork-shaped contacts. The wedge-shaped protrusion of the socket cover forces the MCM leads into the

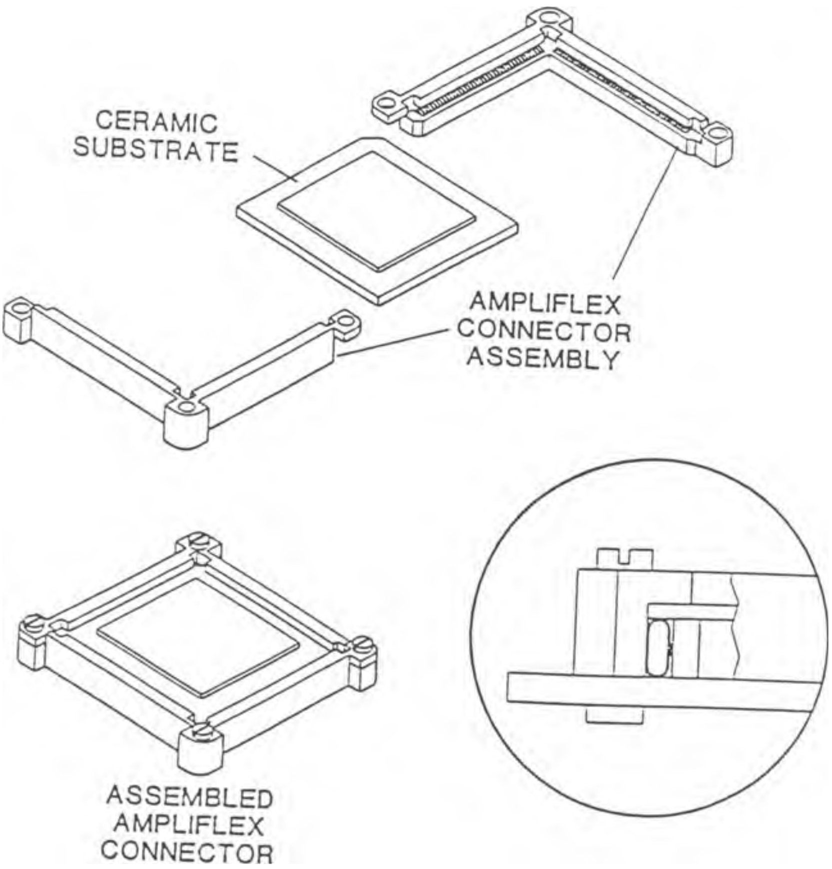


Figure 10-11 Elastomeric socket for leaded peripheral applications. (Courtesy of AMP.)

contacts. The configuration allows the MCM substrate to expand and contract without transferring micromotion to the contact interface. Motion is absorbed elastically at the base of the contact. While this approach seems simple, it requires particular attention to the alignment and registration of the flexible circuits on all four sides of the MCM. Tolerance analysis of the leads, connector components and PWB reveals the precise positioning and alignment required for a reliable connection.

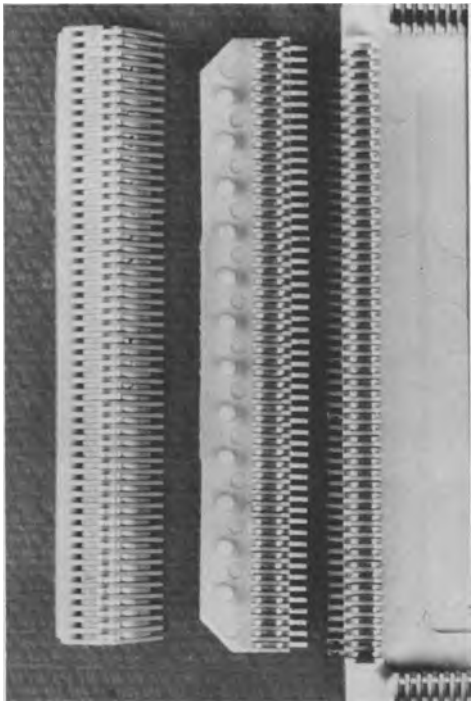


Figure 10-12 Leaded socket for peripheral MCMs. (Courtesy of AMP.)

Prototype implementations of this socket use several short 54-position strips with contacts on 0.5 mm center lines. An alignment plate positions the strips on the PWB, the strips are reflow soldered and the alignment plate is replaced with a nest. The nest serves the function of a bottom housing, a seat for securing the MCM. The MCM is placed in the nest and the cover is installed and pushed down.

Socketed Flex Connected MCMs

A socket similar to the SIMM socket is used with a flex card connector to make a separable connector. In this case, the flex leads are attached permanently to the MCM and also attached to the PWB through a variation of the SIMM socket with narrower openings. The flex leads are wedged into these contact openings.

Quad Flat Pack Sockets for Peripheral MCMs

QFP sockets for peripheral MCMs were originally designed for JEDEC QFPs with gull-wing shaped leads. The sockets (Figure 10-13) use a two piece arrangement. Normal force is created by the cover pressing the MCM leads into the contacts. These sockets generally offer very high packaging density and a low profile. Their main attraction is that they are production tooled for compatibility with high volume commercial semiconductors. Inexpensive and readily available, they also are compatible with reflow soldering.

Current versions use through-hole mounting to the PWB with contact legs arranged on a 0.100" × 0.75" grid. Through-hole mounting makes trace routing on the board more difficult. On a PWB, plated through-holes used for component mounting also may serve as vias. (Vias interconnect different layers.) The difference is that normal interlayer vias are small (about 0.013") and plated through-holes are large (0.035"), making routing of conductors on all layers of the board more difficult. An additional disadvantage of through-hole mounting is the effect on performance driven applications. The contacts typically offer higher inductance and resistance as well as a longer electrical path.

Surface mount versions of QFP sockets are available, with a footprint identical to the device. Through-holes are eliminated to make routing easier. The QFP sockets make it easier to mount MCMs on both sides of the PWB.

Pin Grid Array Sockets for Pinned Area Array MCMs

PGA sockets for pinned area array MCM, like the QFP and SIMM sockets, are production sockets originally designed for commercial, high volume application.

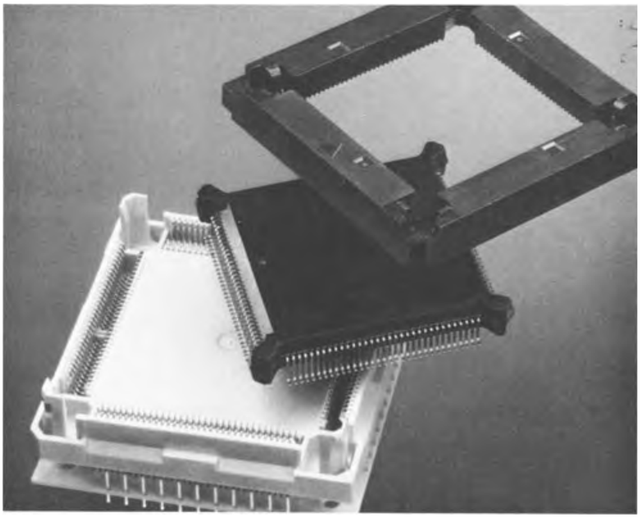


Figure 10-13 Quad flat pack (QFP) sockets for peripheral MCMs. (Courtesy of AMP.)

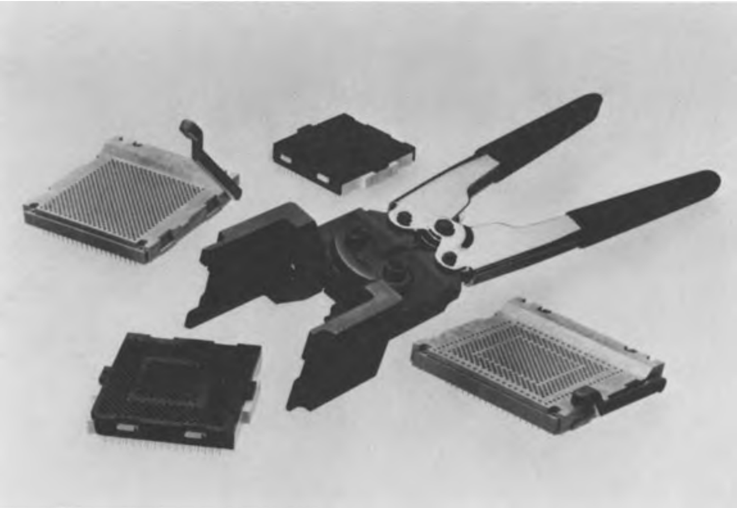


Figure 10-14 Zero insertion force (ZIF) sockets for pin grid array (PGA) packages (Courtesy of AMP.)

These sockets are available in both low insertion force (LIF) and zero insertion force (ZIF) versions. LIF sockets use contacts staggered on two heights to lower the engagement forces, while ZIF sockets (Figure 10-14) employ a camming or spring mechanism to hold the contacts open during MCM insertion and then close the contact around the pin. ZIF sockets are further divided by actuation method - handle or tool.

PGA sockets, with grids up to 20 x 20 (400 positions), are suited to high pin count MCMs. Versions of ZIF sockets with high temperature materials particularly are well suited to burn-in applications since the absence of insertion force accommodates high cycle life.

The main disadvantages of the PGA socket are the same as for the QFP socket: through-hole mounting on the PWB makes routing more difficult and can present undesirable high inductance, resistance and propagation delays in some applications.

A surface mount zero insertion force pin grid array socket is shown in Figure 10-15 and 10-16. This socket was used in the NEC ACOS 3900 [24] to connect an MCM containing 9440 contacts and, being surface mount, it allows two sided mounting. The contact shown in Figure 10-16 is 4.3 mm high and has a 0.4 mm solder terminal. The contacts are soldered to the PWB pads which are arranged on a 2.54 mm staggered grid (0.1"). The contacts are separated from

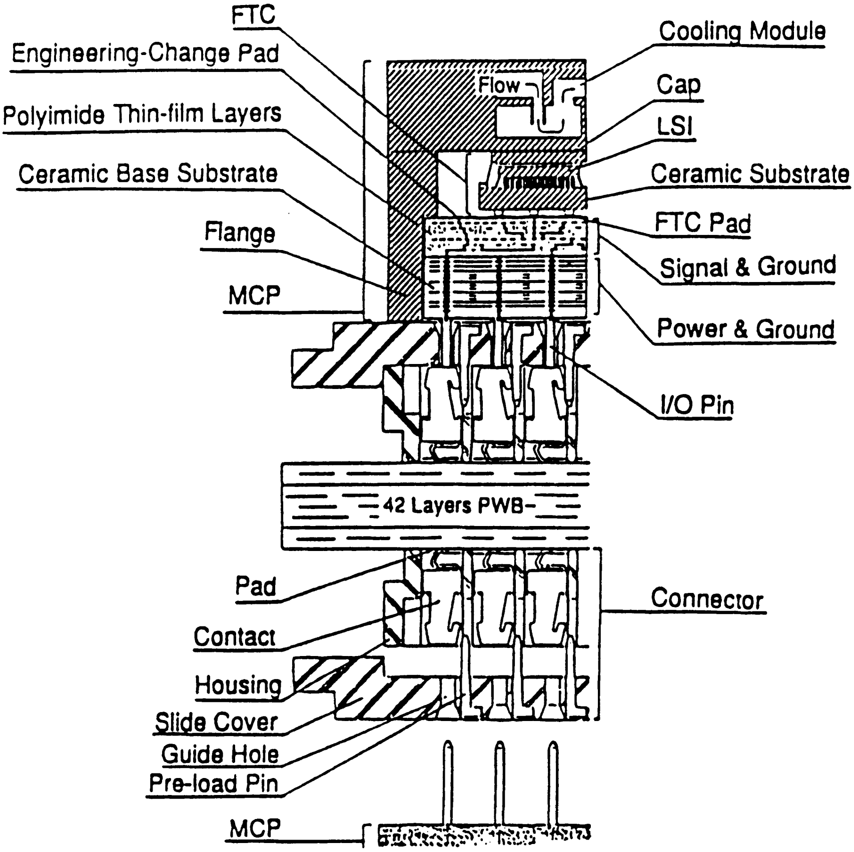


Figure 10-15 Packaging cross section of NEC ACOS 3900 showing the complete connector [24].

each other by a plastic molded housing. During assembly, the MCM pins are inserted into the slide cover (Figure 10-15) and the slide cover (with MCM pins) is inserted into the housing in one motion. As the MCM pin enters the contact, the pre-load pin compresses the contact around the MCM pin.

10.4 STANDARDS ACTIVITIES

Standards provide stability to industry, hasten acceptance and implementation of new technologies, reduce development costs and break down trade barriers [13].

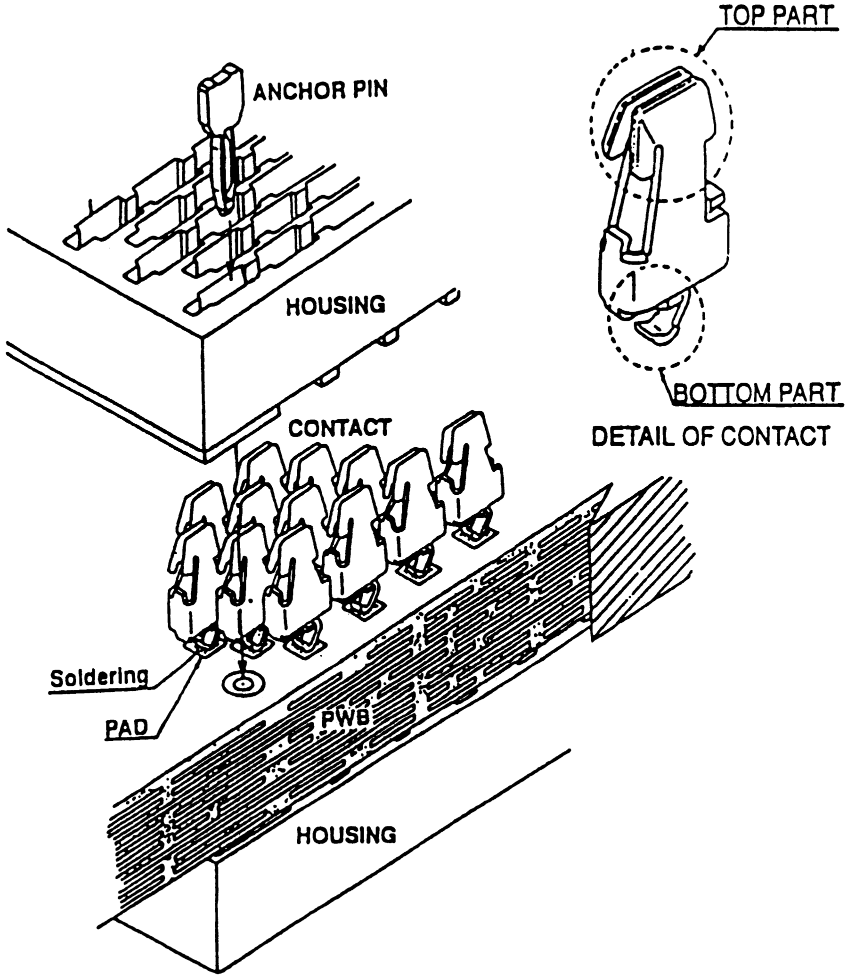


Figure 10-16 Details of the surface mount contact and housing for the NEC ACOS 3900 [24].

As with any standard, those affecting MCMs (and consequently their level 2 connections) are evolving at a rapid pace. With the worldwide emphasis on microminiaturization and surface mount technology, the job of standards groups has been made even more critical. The need for faster, denser and more cost effective designs brought to market in a short time complicates the challenge of developing and issuing standards. Several groups are involved in standards

Table 10-2 IEEE Task Force Recommendations for MCM Package Sizes.

SUBSTRATE SIZE (mm)	PACKAGE SIZE (mm)	
	Peripheral	Area Array
40	55	52
48	65	62
62	80	77
81	100	97
97	115	112

activities as discussed below. Some recommendations from these groups are given in Table 10-2.

10.4.1 Supporting Groups for Standards

The IEEE Computer Society Technical Committee on Packaging

This IEEE (Institute of Electronic and Electrical Engineers) committee appointed a special task force in the fall of 1990 to "seek early consensus on the need to standardize MCM sizes, and to propose some possible sizes." Its efforts promoted the development of standards limiting the number of MCM package sizes, around which an infrastructure could be built economically.

The task force recommended five sizes of peripheral packages with leads spaced on 0.5 mm center lines and five sizes of pin and pad area array packages using established PGA center lines. These are summarized in Table 10-2.

The EIA JEDEC Committee 11 (JC-11)

The Electronic Industries Association (EIA) Joint Electronic Device Engineering Council (JEDEC) committee establishes voluntary standards for the mechanical outlines of solid state and related products. Early in 1992, a task force developed MCM packaging standards, also in an attempt to stem a proliferation of many different package sizes. Its efforts include PGA and ceramic QFPs.

JEDEC Publication 95 contains published outlines prepared by the JC-11 Committee.

The EIA Committee on Sockets (CE-3.0) has established voluntary standards for sockets [13].

The Institute for Interconnecting and Packaging Electronic Circuits (IPC) The IPC has established standards for printed circuit land patterns for electronic products.

10.5 SUMMARY

10.5.1 Recommendations

MCMs do not necessarily represent a new need in second level connector technology. The leadless area MCM has its counterpart in an LGA package used for packaging microprocessors. Likewise, tight center line spacing is as common with high volume ICs as with MCMs. Challenges unique to MCM-to-board connections include very large I/O counts, large component size and high speed performance. These require careful analysis for a satisfactory connection.

Speed, in particular, must be carefully analyzed. In any high speed system, the characteristics of the level 2 connection must be included in any performance model since these connections can significantly affect signal transmission quality. The connection is more than hardware or simple mechanical connections. It is an electrical connection and part of the transmission path. The often irregular geometry and relatively long electrical length of any connection can be a significant portion of the path. The adverse influence of the connection is best minimized either through some degree of impedance matching or by keeping the connection length short enough so that it is insignificant to signal transmission.

The choice of connections covers a wide range of options, including direct connection through soldering, standard high volume sockets and specialized sockets. Any choice involves a tradeoff between cost and performance. But failure to calculate this tradeoff rigorously runs the risk of degrading the rest of the system.

10.5.2 Future Trends

The growing preference for surface mounting is contributing to a trend away from PGA packages to LGAs and to various styles of QFPs. Socket makers are accommodating these changes. To address the continuing popularity of PGAs, designers of surface mount boards can avoid making through-holes for the PGAs by placing the packages in surface mount sockets. However, socketed PGAs present a very high profile. The vulnerability of the pins to damage is another disadvantage of socketed PGAs. In contrast, LGAs are a practical alternative. With no pins, they are less vulnerable to damage and provide very short electrical paths. The trend is toward smaller and smaller packaging to meet the needs of hand-held computers and communications devices.

As the speed of components continues to increase, new forms of packaging are required. Optical connects can be expected to play an important role as clock speeds approach the GHz level. This will create a whole new range of problems including the need for very precise alignments and a quick and cost effective separable connector containing many fibers. Future connectors may even include lasers to communicate with remote components in the system.

Another trend in socketing follows the transition in the United States toward metric dimensioning of flat packs. New flat packs generally have contacts on 1 - 0.3 mm pitches. New package outlines submitted for the JEDEC registration process must be hard metric, although extensions to existing families based on inch measurements may be acceptable to JEDEC.

In response to customer demands, package designers have created a proliferation of outlines for flat packs, more than 100 in Japan, and about 60 registered with JEDEC. This diversity represents a challenge to socket manufacturers to their product offerings, and both surface mount and through-hole versions of their products. JEDEC and EIA Japan are working together to develop standards.

Acknowledgments

The author gratefully acknowledges the significant technical and organizational help of the editors. He would also like to thank AMP for allowing him to use a number of their photographs and diagrams, and for giving him access to technical descriptions of many of their connectors. Sincere thanks are given to Dr. Leonard Buchoff of Elastomeric Technologies, Inc. for similar contribution. He also thanks Dr. Thomas C. Russell and Zeev Lipkes of Alcoa Electronic Packaging for their technical assistance in preparing the final version of this chapter.

References

1 *Connectors and Interconnections Handbook, Vol. 1* , G. Derman, ed. Deerfield MA: Int. Inst. Connect. & Connect. Tech., 1990.

A. J. Bilotta, *Connections in Electronic Assemblies*, New York: Marcel Dekker, 1985.

2 W. E. Gilmour, "Connector and Interconnection Devices," *Electronic Packaging and Interconnections Handbook*, C. A. Harper ed., New York: McGraw Hill, 1991, Ch. 3.

3 J. B. Gillett, B. D. Washo, "Connector and Cable Packaging," *Microelectronics Packaging Handbook*, R. R. Tummala, E. J. Rymaszewski, eds., New York: Van Nostrand Reinhold, 1989, Ch. 14.

4 M. Freedman, "MCM Interconnection Options," *Proc. NEPCON West,* (Anaheim CA), p. 1527, 1991.

J. H. Whitley, "The Mechanics of Pressure Connections," AMP Inc., Harrisburg PA, Dec. 1964.

5 W. H. Knausenberger, "Multichip Module Connections," *Thin Film Multichip Modules*, G. Messner, I. Turlik, J. W. Balde, P. E. Garrou, eds, Reston VA: ISHM Press, 1992, Ch. 10.

6 D. Grabbe, "ModuPak: A New, Low Cost, High Speed MCM Socketing System," *Proc. NEPCON West*, (Anaheim CA), p. 1537, 1991

7 H. Kent, "Multi-chip Module Connectors and Sockets: High Density and High Speed," *Proc IEPS Conf.*, (Marlborough MA), p. 726, Sept. 1990.

8 R. Holm, *Electronic Contacts*, New York: Springer-Verlag, 1967.

9 S. S. Simpson, M. E. St. Lawrence, "A High Density Land Grid Array (LGA) Connector with Wipe and High Compliance," *Proc. IEPS Conf.*, (Marlborough MA), pp. 951-955, Sept. 1990.

J. H. Whitley, Anatomy of a Contact: A Complex Metal System," *Insulation/Circuits*, vol. 8, p. 39, Fall 1981.

10 R. S. Mroczkowski, "Materials Considerations in Connector Design," Technical Paper, AMP, Inc., pp/ 31011-1388, 1989.

11 J. A. Defalco, "Reflection and Crosstalk in Logic Circuit Interconnections," *IEEE Spectrum*, vol. 11, pp. 44-50, July 1970.

12 D. Royle, "Rules Tell Whether Interconnections Act Like Transmission Lines," *EDN*, vol. 23, pp. 131-160, June 1988.

13 *Microelectronics Standards MS002-MS008*, JEDEC Publication no. 95, Washington DC: EIA Eng Dept., 1989.

14 G. G. Harman, *Wire Bonding in Microelectronics*, Reston VA: ISHM Press, 1989.

15 "Guidelines for Chip-on-Board Technology Implementations, ANSI/IPC-SM-784, Lincolnwood, IL: IPC, 1990.

16 D. G. Grabbe, H. Merkelo, "High Density Electronic Connector for High Speed Digital Application," *AMP J. of Techn.*, pp. 800-90, Nov. 1991.

17 C. W. Pike, R. Hassan, "Wire Button Contacts as a Connection Alternative: Design Opportunities and Challenges," *Proc. IEPS Conf.*, (Marlborough MA), p/ 944, Sept. 1990.

18 CINCH Connectors, 1500 Morse Avenue, Elk Grove Village, IL 60007.

19 H. W. Markstein, "Applications Widen for Elastomeric Connector," *Electr. Packaging and Production*, vol. 32, no. 5, pp. 29-32, May 1992.

20 W. R. Lambert, W. H. Knausenberger, "Elastomeric Connections-Attributes, Comparisons and Potential," *Proc NEPCON West*, (Anaheim CA), pp. 1512-1516, 1991.

21 C. A. Haque, ""Characterization of the Metal-in-Elastomer Contact Material," *Proc. 35th Holm Conf. Elec. Contacts*, Chicago, IL, pp. 117-120, 1989.

22 A. Strange, L. S. Buchoff, S. Ross, "Elastomeric Connectors Meet Critical Aerospace Requirements," *Proc. NEPCON West*, (Anaheim CA), pp. 651-656, Feb. 1992.

23 L. S. Buchoff, "Elastomeric Sockets for Chip Carriers and MCMs," *Proc. 42nd Electronic Components and Techn. Conf.*, (San Diego CA) pp. 316-320, May 1992.

24 M. Yamada, *et al.*, "Packaging Technology for the NEC ACOS System 3900," *Proc. 42nd Electr. Components and Techn. Conf.* (San Diego CA), pp. 745-751, May 1992.

11

ELECTRICAL DESIGN OF DIGITAL MULTICHIP MODULES

Paul D. Franzon

11.1 INTRODUCTION

The aims in package electrical design are to maximize the performance of the system, as limited by the interconnect (here "interconnect" refers to interconnections and connections) delay, while minimizing the possibility of false operation in the field, due to electrical noise, and minimizing the cost. To this end, electrical design of digital systems consists of the following:

1. Selecting the appropriate packaging and semiconductor technology mixture, and the appropriate partitioning of functions between chips and packages, so that the system design is likely to meet its cost and performance aims (see Chapter 3).

2. Generating the logical design (gate level design), and determining the logic families to be used.

3. Generating the timing design (when signal events will take place) and the noise budget (the required signal quality or signal integrity).

4. Determining the appropriate models and selecting the appropriate

simulation tools that allow interconnect signals and timing to be accurately predicted from the physical design.

5. Generating the physical design. This includes a placement, which describes where the chips and other components are located, and a layout, which describes where conductors run.

Electrical design is very important because a large fraction of the clock cycle time of a computer or any other digital system can be attributed to the delay involved in getting a signal off the chip and transmitting it between chips. A good electrical design maximizes performance by controlling these delays. In today's high performance system designs, the speed of ICs already strains the ability of conventional PWB technologies to provide comparably fast interconnections. The use of MCM technology allows clock speed increases of between 30% and 100%. A good electrical design also minimizes the possibility of noise induced errors occurring during operation. Good design practices are particularly relevant to MCM design because it is much more difficult and costly to diagnose and fix an electrical design problem on a fine pitch MCM prototype than it is on a PWB prototype. Nevertheless, in both technologies it is highly desirable not to have to iterate the design once the prototype is constructed.

This chapter starts by defining delay and noise and how they relate to digital interconnection design. It then describes the primary delay and noise phenomena important to MCM interconnect design (many of which are common to conventional package design). Finally the activities that take place during electrical design are discussed.

11.2 DELAY AND NOISE IN DIGITAL DESIGN

In any digital design there is some critical path whose delay limits the maximum possible speed of the system. For example, imagine that the performance of a system is limited by the delay, t_{total}, between the two flip-flops or latches shown in Figure 11-1. (A flip-flop or latch samples the voltage at the input, pin D, such as D1 in Figure 11-1, and transfers the same logical level, 0 or 1, to the output, pin Q, such as Q1, whenever a 0 to 1 edge occurs at the clock pin, Ck, such as Ck1.) A buffer is placed at the output of the first flip-flop for the purpose of driving the package interconnect structures. The buffer might also be called a driver or just an output. The gate input at the end of the interconnection is referred to as an input, a receiver or just a load.

The total delay, t_{total}, is the sum of the following components:

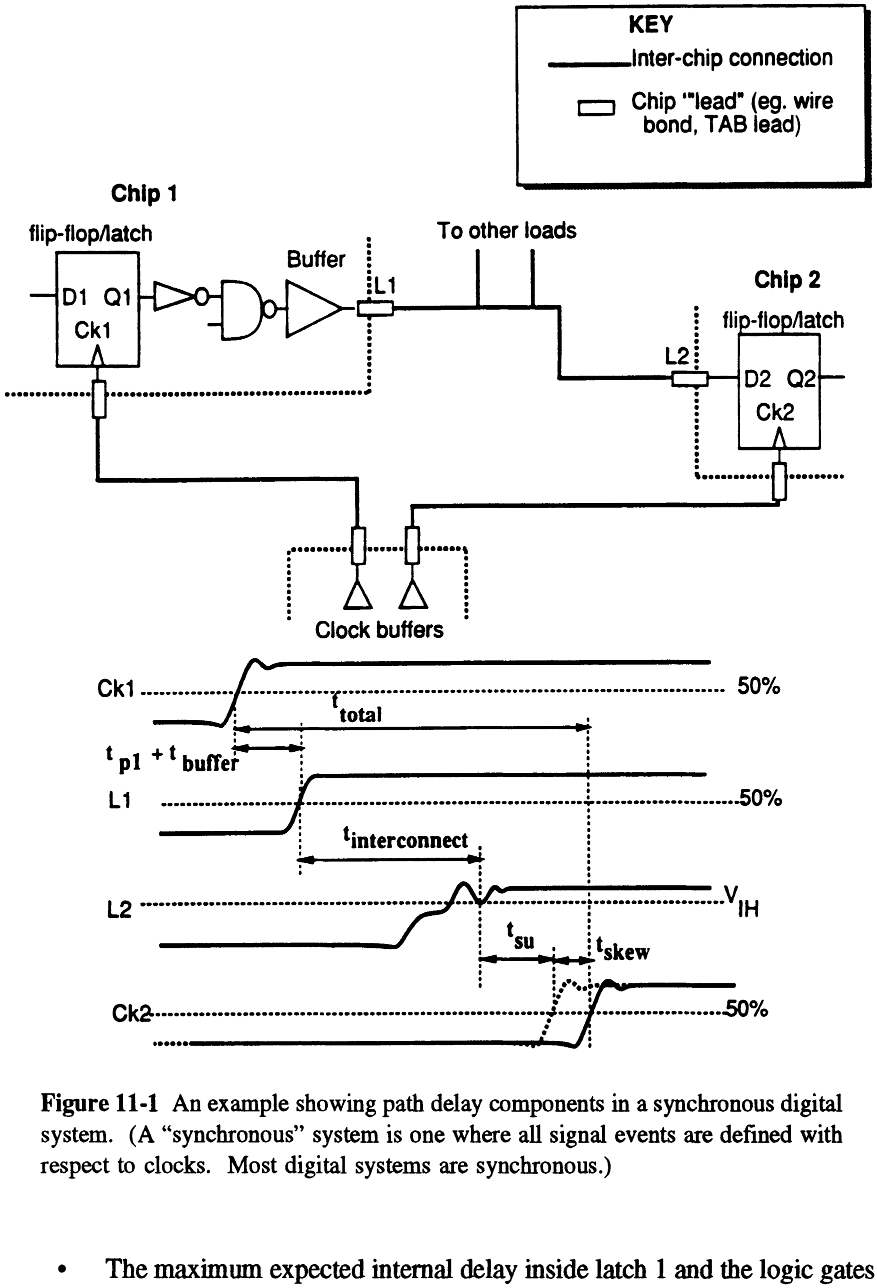


Figure 11-1 An example showing path delay components in a synchronous digital system. (A "synchronous" system is one where all signal events are defined with respect to clocks. Most digital systems are synchronous.)

- The maximum expected internal delay inside latch 1 and the logic gates that come after it, t_{p1}.
- The internal delay within the buffer, t_{buffer}.
- The delay introduced by the interconnect, $t_{interconnect}$.

- The setup time required by the receiver latch t_{su}. (This is the minimum time for the signal at D2 to be stable before being sampled by a clock edge at Ck2).
- The maximum possible uncertainty in the exact position of the edge on Ck2 with respect to the edge on Ck1. This is referred to as clock skew, t_{skew} when the uncertainty is constant. Each individual clock also experiences a small amount of jitter, t_{jitter} (uncertainty when each edge will occur as compared with when its arrival is expected), which should also be added in.

The maximum expected result for the delay, t_{total}, determines the minimum timing between the clock events Ck1 and Ck2. If Ck1 and Ck2 come from the same clock source, as is usually the case, then t_{total} would be the clock period, T_{clock}, of the system and its inverse would be the clock frequency, $f_{clock} = 1/T_{clock}$. The words “maximum expected result” refer to the fact that manufacturing and process variations in the circuits and packages result in different actual delays in different produced systems. The clock period must be selected so that it is greater than the delay expected in more than 999 out of 1000 of the manufactured systems. If nominal values of delay are used to select the clock period, then only 50% of the manufactured systems will work.

Of main interest here, is the package interconnect delay, $t_{interconnect}$. This is shown in Figure 11-2 as consisting of two delays. First there is the time, t_{prop} between when the output L1 reaches the 50% point of its logic swing and when the input L2 reaches a similar point. The 50% point is used as a reference as it

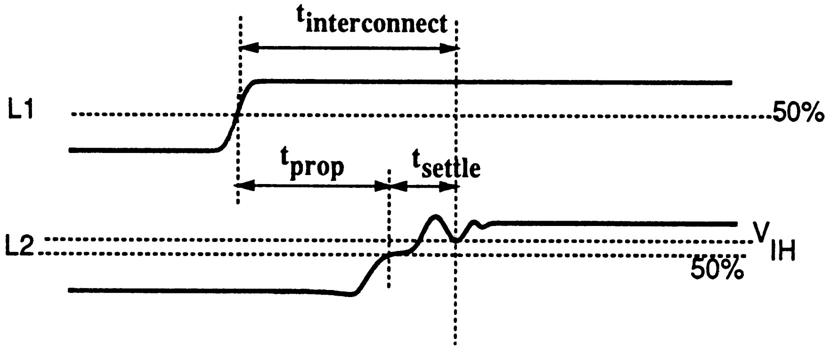


Figure 11-2 Data signal delay, defined as the delay between the 50% points in the signal with noise settling delay added in.

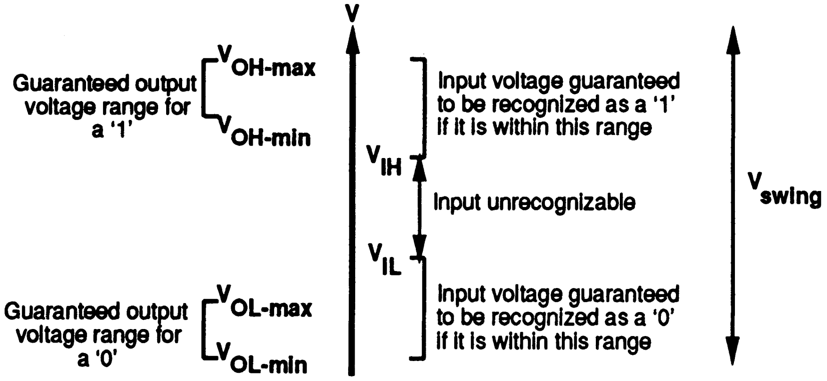

V
Guaranteed output voltage range for a '1'
$V_{OH\text{-}max}$
$V_{OH\text{-}min}$
V_{IH}
V_{IL}
Guaranteed output voltage range for a '0'
$V_{OL\text{-}max}$
$V_{OL\text{-}min}$
Input voltage guaranteed to be recognized as a '1' if it is within this range
Input unrecognizable
Input voltage guaranteed to be recognized as a '0' if it is within this range
V_{swing}

Figure 11-3 Definitions of DC voltage properties for a digital gate.

is the standard point used to define delay in digital circuits. Second there is the time, t_{settle}, that is introduced by the requirement that any electrical noise on the signal must settle before the signal can be safely sampled by the flip flop and determined to be a logic-0 or logic-1. In Figure 11-2, the signal must settle to a value greater than V_{IH} in order to be considered a valid logic-1. The reason for this relates to the DC properties of the circuit (Figure 11-3):

- Maximum and minimum output low voltage, $V_{OL\text{-}max}$ and $V_{OL\text{-}min}$, represent respectively the highest and lowest possible output voltage which corresponds to the logic-0 state. $V_{OL\text{-}min}$ is also the nominal output low voltage (such as 0 V in CMOS and TTL circuits).

- Maximum input low voltage, V_{IL}, is the highest possible input voltage that the circuit recognizes as a logic-0.

- Maximum and minimum output high voltage, $V_{OH\text{-}max}$ and $V_{OH\text{-}min}$, represent respectively the highest and lowest possible output voltage which corresponds to the logic-1 state. $V_{OH\text{-}max}$ is the nominal output high voltage (such as 5 V in a CMOS circuit).

- Minimum input high voltage, V_{IH}, is the lowest possible input voltage that the circuit recognizes as a logic-1. (This is V_{IH} as used above.)

- The nominal voltage swing V_{swing}.

Any voltage between V_{IL} and V_{IH} is not classifiable as a logic-0 or logic-1 and if sampled could possibly be interpreted as either, resulting in a logic error.

Thus, if the output of a gate is at a logic-0 level, the DC properties of the gate guarantee that as long as the magnitude of any noise voltage added to this signal is less than V_{IL} - $V_{OL\text{-}max}$, then the signal is always recognized as a logic-0 at the input of any similar gate. This difference is thus defined as the low DC noise margin,

$$NM_L = V_{IL} - V_{OL-max} \tag{11-1}$$

of the logic gate. The high DC noise margin is similarly defined as

$$NM_H = V_{OH-min} - V_{IH}. \tag{11-2}$$

Logic gates are usually designed so that these properties remain constant for most of the products within one circuit family for any one manufacturer. Typical DC voltage properties for different families are given in Table 11-1. (Here, and from now on the -max and -min suffixes will be dropped, with $V_{OL} = V_{OL\text{-}max}$ and $V_{OH} = V_{OH\text{-}min}$.)

Table 11-1 Typical DC Parameters and Noise Margins for Different Logic Families.

FAMILY	V_{swing}	V_{OL}	V_{IL}	NM_L	V_{OH}	V_{IH}	NM_H
Advanced CMOS	5	0.1	1.65	1.55	4.9	3.85	1.05
Advanced LS TTL	3.4	0.5	0.8	0.3	2.7	2.0	0.7
10K ECL	0.8	-1.63	-1.48	0.16	-0.96	-1.11	0.15

It has been recognized that digital gates can safely withstand short AC noise pulses of magnitude greater than the DC noise margin. A plot of the maximum safe noise voltage versus its 50% width is referred to as the AC noise immunity curve for a particular logic family, an example of which is given in Figure 11-4. Virtually no manufacturers guarantee values for AC noise immunity as they guarantee values for DC noise margins. It must be characterized by the designer.

The noise control requirements on digital signals are referred to collectively as "signal integrity" requirements. Note that signal integrity requirements are

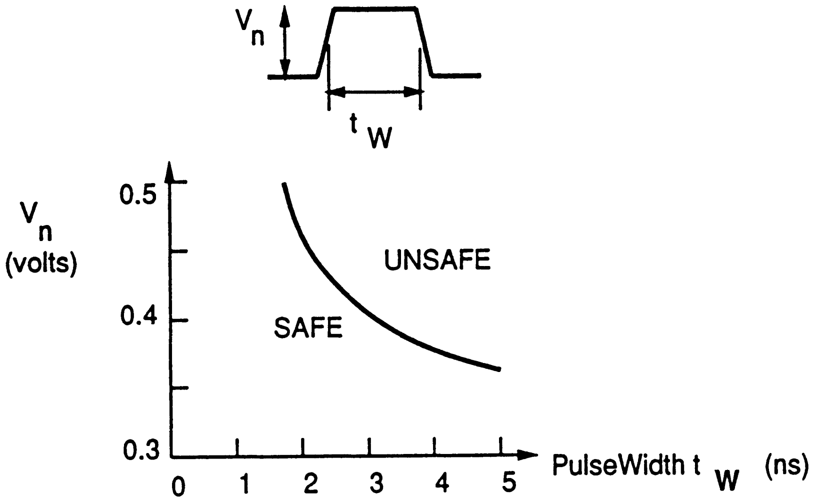


Figure 11-4 A typical AC noise immunity curve for ECL logic. If the gate input pulse height and width fall in the "safe" region, it will not result in a potential logic error [3].

more severe on clock signals than on data signals. On the latter, extra delay time can sometimes be included to allow noise to settle. On the former, excessive noise might be incorrectly recognized as a clock edge, leading to the latching of an incorrect data value. The techniques used to establish the noise margin and noise immunity of digital circuits are discussed in a number of sources: [1] - [5].

The remainder of this chapter discusses the basics of delay and noise control for MCM interconnections, including how they affect decision-making in the technology selection process and during the design process. It starts by discussing the main phenomena of interest: delay, reflection noise, crosstalk noise and simultaneous switching noise and then describes the design process.

11.3 PROPAGATION DELAY AND REFLECTION NOISE

In order to understand propagation delay and reflection noise it is first necessary to determine if an interconnecting segment (signal line or signal connector) should be treated as a transmission line or as a lumped circuit modeled by discrete capacitors, inductors and resistors. This determination is related to the frequency spectrum of the signal. A digital signal is actually communicated over

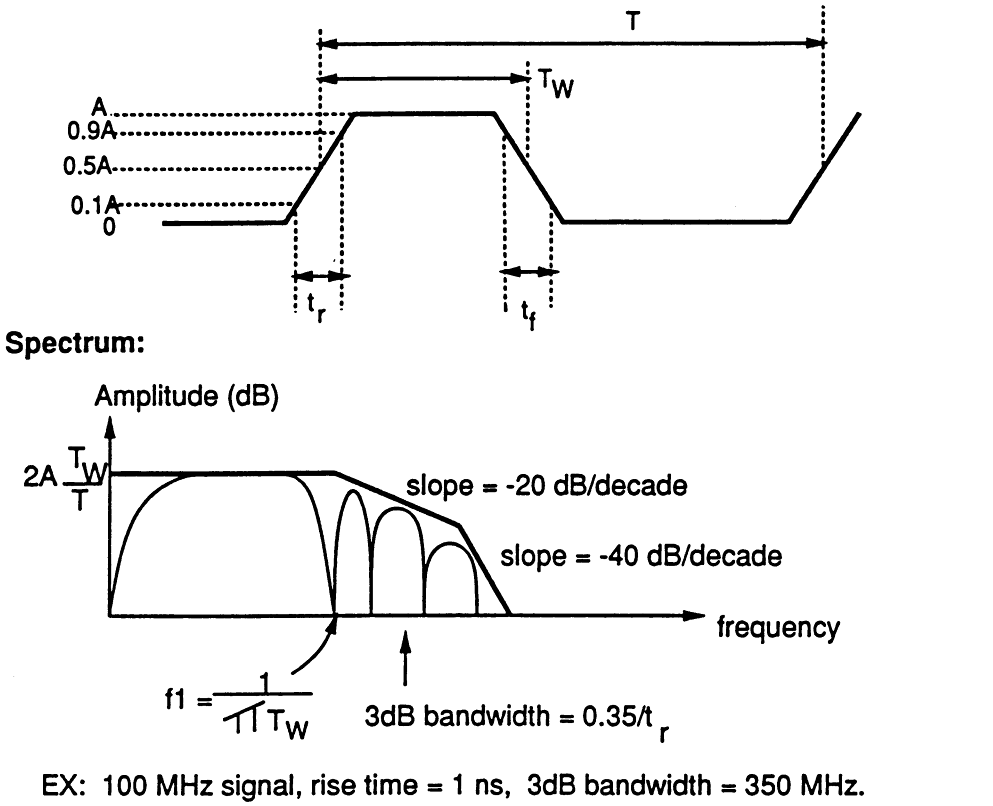


Figure 11-5 Frequency spectrum of a digital signal.

a wide range of electromagnetic frequencies as shown in Figure 11-5. The bandwidth, BW, of this signal is given by

$$BW = 0.365 / t_r \tag{11-3}$$

where t_r is the rise time of the signal or the time taken for the signal to transit between 10% and 90% of the voltage swing. For example, the bandwidth of a signal with a 1 ns rise time is 350 MHz. For this signal, we define a "pulse design wavelength" λ,

$$\lambda = \frac{c / \sqrt{\varepsilon_r \, \mu_r}}{N \times BW} \tag{11-4}$$

where c = 3.0 × 10^8 m/s is the speed of light in a vacuum, ε_r is the relative dielectric constant, μ_r is the relative permeability constant, normally 1, (and thus $c/\sqrt{\varepsilon_r}$ is the speed of light in that medium) and N is an integer that determines the quality of the signal. If the length of a structure exceeds $\lambda/8$ then a transmission line treatment is in order. Otherwise a lumped circuit treatment suffices. With a polyimide dielectric constant, $\varepsilon_r = 3.5$, and a rise time, $t_r = 1$ ns, then with N = 4, $\lambda / 8 = 1.4$ cm suggests that even 2 cm long interconnect structures should be treated as transmission lines.

If the length is shorter than $\lambda / 8$ then a lumped circuit, as shown in Figure 11-6, is used to model a point-to-point connection. A lower bound estimate on the propagation delay of this line is the RC delay given by

$$t_{prop} = t_{50\%}$$
$$t_{50\%} \approx 0.7 \left(R_{out} (C_1 + C_{line} + C_2 + C_{in}) + \frac{1}{2} R_{line} C_{line} + R_{line} (C_2 + C_{in}) \right) \tag{11-5}$$

where R_{out} is the equivalent resistance of the driver, R_{line} is the total line resistance, C_{line} is the total line capacitance, C_1 and C_2 are the lead (chip connection) capacitances, and C_{in} is the input capacitance of the die (typically about 2 pF for CMOS). The effect of any line inductance, L_{line}, is to increase this delay by 10 - 30% or more.

Figure 11-7 shows a point-to-point connection and simple equivalent circuit model for a line that must be treated as a transmission line. Note that because the chip connection leads are shorter than $\lambda / 8$, they are treated as lumped circuit elements. If a short via or connector were placed in the center of this line, it would be modeled by a lumped circuit equivalent and referred to as a discontinuity. Typical values for lead inductance and capacitance are given in Table 11-2.

Figure 11-7 also shows what happens to a signal produced by the output buffer. In Figure 11-7, the buffer is modeled by its Thévenin equivalent circuit, that is a voltage source and an equivalent output resistor. The input or receiver is replaced by its equivalent capacitance. The combined effect of the output circuits and the load presented to the signal at the output is that the signal will have a certain rise time, t_{rise}, which contributes to the 50% delay. The rise time is determined in part by these lumped circuit parasitics. The faster the circuit family, the faster the rise time. For example, ECL rise times are often well under a nanosecond. The buffer design determines t_{buffer} and t_{rise}, both of which decrease with reduced load capacitance. For example, a CMOS buffer that might

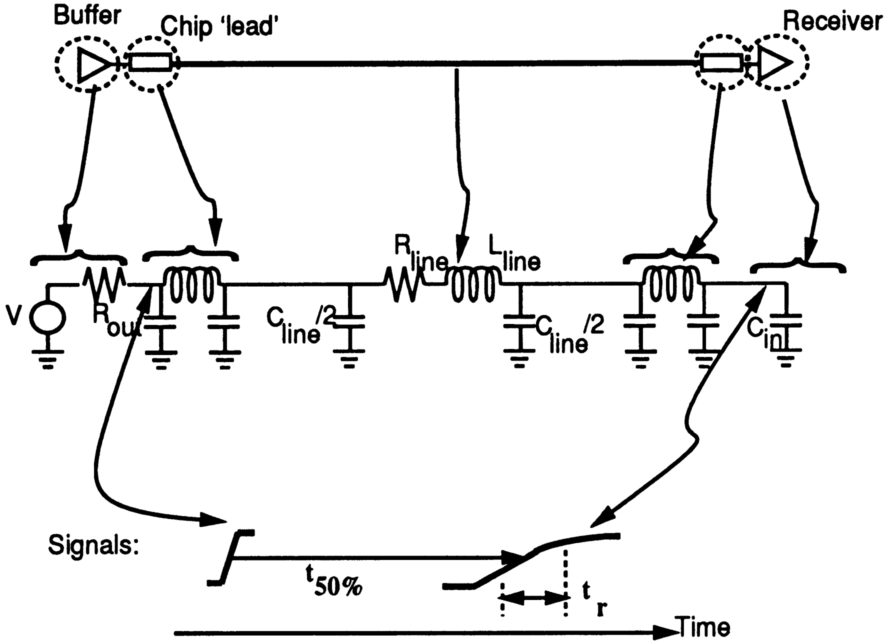


Figure 11-6 A lumped equivalent model for an electrically short point-to-point interconnect line.

have a delay $t_{buffer} = 10$ ns and a rise time, $t_{rise} \approx 1.5$ ns when mounted in a single chip package, finds these times reduced to $t_{buffer} = 7$ ns [6] and has a rise time, $t_{rise} \approx 800$ ps when mounted on a MCM. This results from the lower capacitance of the MCM interconnection and represents significant advantages for MCM technology.

When the signal leaves the output buffer it travels down the transmission line at the speed of light in that medium. At the receiver, the rise time of the signal is increased further by the inductive and capacitive parasitics of the receiver chip lead. All these effects combine to give the total delay of the signal as shown at the bottom of Figure 11-7. If first incidence switching is achieved (see Section 11.3.3), the total propagation delay is given by

$$t_{prop} = \frac{l}{c/\sqrt{\varepsilon_r}} + t_{rise-time-degradation} \qquad (11\text{-}6)$$

Table 11-2 Typical Values for Lead Inductance and Capacitance.

Lead Type	Capacitance (pF)	Inductance (nH)
SMT Package	1	1 - 12
PGA	1	2
Wire Bond	0.5	1 - 2
TAB	0.6	1 - 6
Solder Bump	0.1	0.01

Note:

- Inductance increases linearly with lead length unless a ground plane is provided (hence the reduced inductance of a PGA when compared with an SMT package).
- Lead capacitance is split evenly between the two capacitors shown in the chip lead model.

where l is the length of the interconnection, c is the velocity of light in a vacuum, ε_r is the relative dielectric constant of the dielectric and $t_{\text{rise-time-degradation}}$ is the delay effect of the increase in rise time between the end and start of the line (no simple equation is available for estimating the size of this delay contribution). The propagation velocity is

$$v_{prop} = c / \sqrt{\varepsilon_r} \qquad (11\text{-}7)$$

the speed of light in that medium and $l / \left(c / \sqrt{\varepsilon_r}\right)$ is referred to as the time-of-flight, t_{flight}. Note that the propagation delay decreases with smaller values for dielectric constant.

A transmission line allows the electromagnetic waves to propagate uniformly, an example of which is given in Figure 11-8. These electromagnetic waves are associated with a voltage wave and current wave as shown in Figure 11-9. The return current flowing in the reference plane in Figure 11-9 has the same magnitude as the current on the signal line. The characteristic impedance Z_0 of the transmission line is defined as the ratio of the voltage and current waves traveling down the line and thus has the units of ohms [7] - [9].

The signal line shown in Figure 11-8 must maintain a constant characteristic impedance over any length greater than $\lambda / 8$. When it does so it is referred to as a controlled impedance line. It can be shown that for a lossless line, the

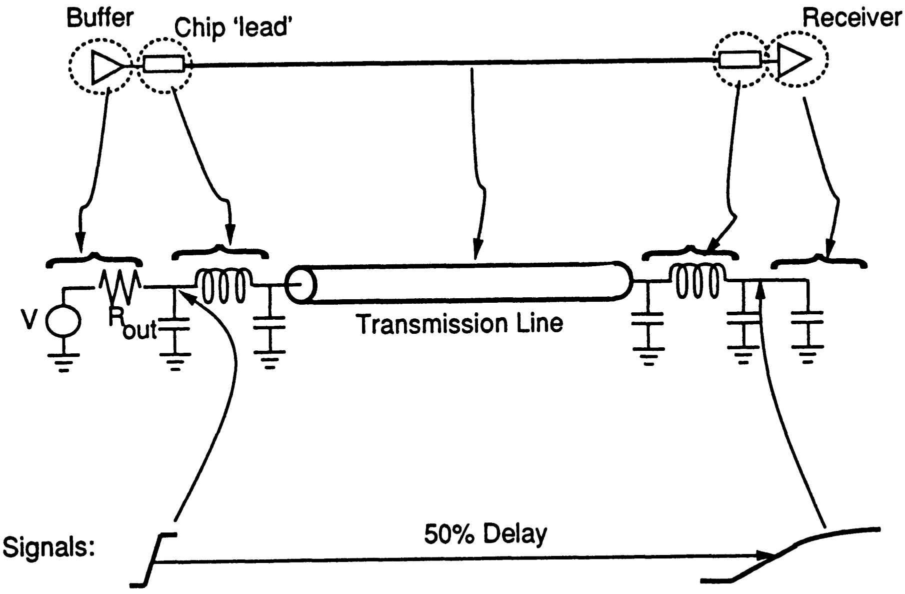


Figure 11-7 A transmission line model for an interconnection showing how propagation time and increased rise time due to discrete capacitive and inductive loads all contribute to delay.

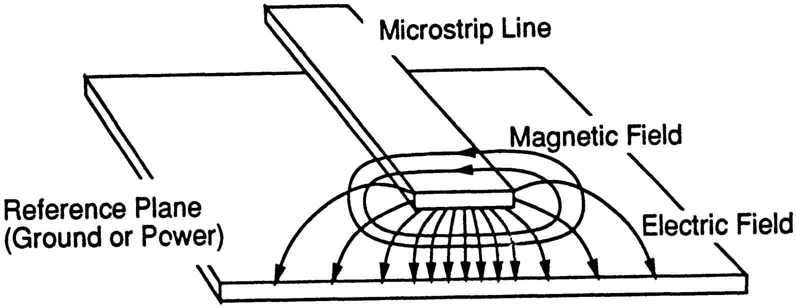


Figure 11-8 In a transmission line an electromagnetic wave is propagated along the line at the speed of light in that dielectric.

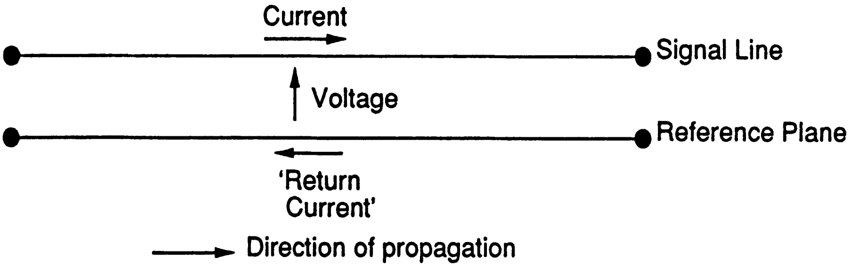


Figure 11-9 Voltage and current associated with the electromagnetic wave being propagated along the line.

characteristic impedance is expressed in terms of the inductance per unit length, L_o, and the capacitance per unit length, C_o.

$$Z_o = \sqrt{\frac{L_o}{C_o}} \ \Omega. \tag{11-8}$$

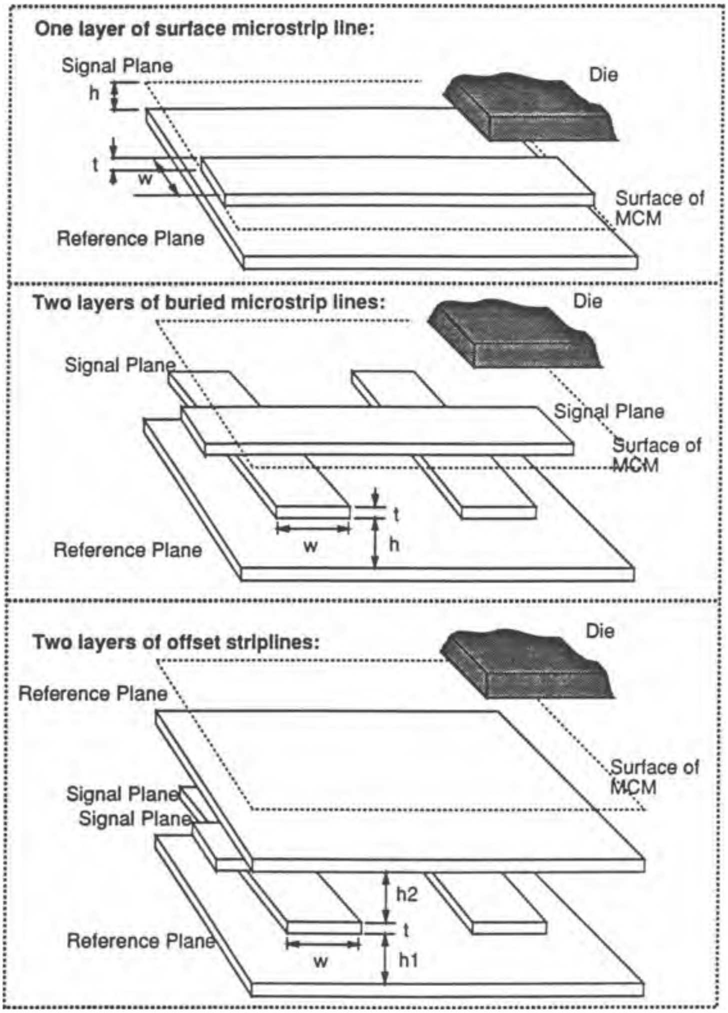
One way to maintain a constant impedance is to maintain a constant geometrical relationship between the signal conductor and the reference plane. If the line narrows, for example, the capacitance decreases and the impedance increases.

The three arrangements that are most commonly used in MCMs and PWBs to provide controlled impedance interconnections are shown in Figure 11-10. While all three arrangements are common in PWBs and laminate MCMs, only the last two are generally used in thin film and cofired ceramic MCMs (though a multilayer ceramic might use surface microstrips).

An empirical (and approximate) equation for Z_0 for a surface microstrip transmission line is

$$Z_o = \frac{87}{\sqrt{\varepsilon_r + 1.41}} \ln \frac{5.98h}{0.8w + t}, \tag{11-9}$$

ε_r is the dielectric constant of the dielectric and the dimensions h, w and t are described in Figure 11-10.



A 'Reference' Plane can either be a power or ground plane.

Figure 11-10 The three most common techniques for layering signal and reference (ground/power) planes in a multichip module.

Note that in order to obtain a higher value for Z_0 you either have to reduce the width, w, reduce the thickness, t, increase the height, h, and/or reduce ε_r. (All of these reduce the capacitance of the line and the first three also increase the inductance of the line.) The time-of-flight delay for a signal traveling down a surface microstrip (in nanoseconds per meter) is

$$t_{\text{flight}} = 3.337\sqrt{0.475\varepsilon_r + 0.67}. \tag{11-10}$$

Similar equations for a buried microstrip are

$$Z_o = \frac{60}{\sqrt{\varepsilon_r + 1.41}} \ln \frac{5.98h}{0.8w + t}, \tag{11-11}$$

and

$$t_{\text{flight}} = 3.337\sqrt{\varepsilon_r}. \tag{11-12}$$

Note that signals propagate fastest on surface microstrips. Also note that a buried microstrip line must be narrower or thinner than a surface microstrip to achieve the same value of Z_o if h and ε_r are the same. The reason is that some of the signal is propagating in the air layer ($\varepsilon_r = 1$) above the surface microstrip, reducing the capacitance. While there is no simple empirical relationship for the characteristic impedance of offset striplines, the time-of-flight delay is given in Equation 11-12. Note that as an offset stripline has two reference planes, its capacitance is higher (inductance is lower), and thus, its impedance is lower than the same sized microstrip line with the same dielectric thickness.

The reference planes (either power or ground) tend to be meshed in thin film and many cofired ceramic MCM structures to promote adhesion between layers. (Each plane is actually a solid sheet with square holes in it.) By breaking up the reference planes, the meshing can affect the characteristic impedance and propagation delay of the signal lines. It also increases the DC resistance of the power and ground circuits. If the holes are large compared with the line pitch, then considerable care must be taken to ensure that the signal lines are routed only over the solid parts of the ground plane. Unrestricted routing might result in characteristic impedance variations of 30% or more. On the other hand, if the mesh pitch is comparable to the conductor pitch, the effect on conductor characteristics is small [10]. If the holes cannot be made this small, then this effect can be reduced partially by running the lines diagonally across the grid.

11.3.1 Reflections

Whenever a change in characteristic impedance occurs, part of the incident electromagnetic wave is reflected, just like part of a light beam is reflected upon striking a sheet of glass.

Characteristic impedance changes occur whenever the line branches (the waves "see" two lines in parallel) or the line ends. The portion of the incident electromagnetic wave voltage that is reflected at a change of characteristic impedance is given as the reflection coefficient, Γ

$$\Gamma = \frac{Z_{load} - Z_o}{Z_{load} + Z_o}. \qquad (11\text{-}13)$$

Z_o is the characteristic impedance of the line that the incident wave is traveling on and Z_{load} is the impedance of the load being seen by the line. If the load is an open circuit, such as a gate input, then $Z_{load} = 0$ and $\Gamma = 1$ and the reflected wave has the same voltage as the incident wave. (This reflected wave also causes the voltage at the end of the line to be doubled, the reflected traveling wave voltage to be added to the voltage provide by the previous wave.) If the load is a short circuit, then $Z_{load} = 0$ and $\Gamma = -1$ and the reflected wave is inverted with respect to the incident wave. On the other hand, if the load is matched to the line impedance, $Z_{load} = Z_o$, then no reflection occurs.

The successive partial reflections of the wave from each end creates a damped ringing signal as shown in Figure 11-11. The ringing signal can be

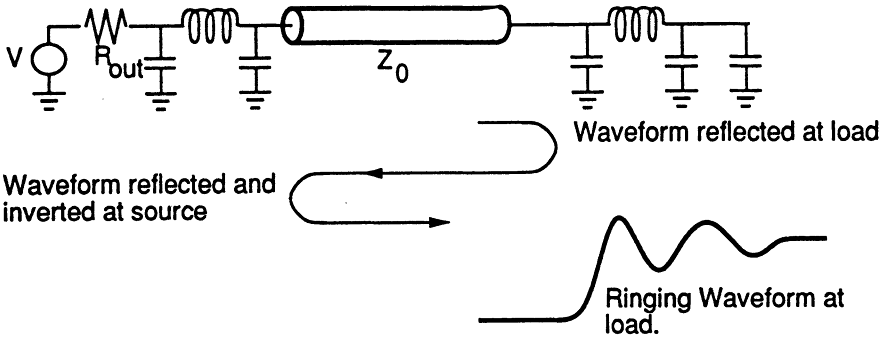

V
Rout
Z0
Waveform reflected at load
Waveform reflected and inverted at source
Ringing Waveform at load.

Figure 11-11 Multiple reflections act to cause ringing.

analyzed quantitatively using a lattice diagram [11] and [7] or a Bergeron diagram [12].

Ringing is a potential problem if the time it takes for the wave to travel down the line is longer than one-quarter of the rise time,

$$t_{prop} > t_{rise}/4. \tag{11-14}$$

Waiting for the ringing to settle down can take up to an additional 4 × t_{prop}. The reason why a short line controls ringing noise is that the ringing settles down before the end of the rise time. Equations 11-14 and 11-6 indicate that for a 30 cm long PWB line ($\varepsilon_r = 4.5$), ringing could be a problem for signals with rise times less than 8 ns. For a 10 cm long cofired MCM line ($\varepsilon_r = 10$), on the other hand, ringing becomes a potential problem with rise times less than 4 ns. Rise times less than 1 ns - 2 ns are not uncommon in high speed circuits.

If ringing is a potential problem, then one or sometimes both ends of the line must be terminated with a matched resistive load. The four most common techniques for providing a matched impedance load are shown in Figure 11-12. (The series termination shown in Figure 11-12 prevents noise by not allowing the wave reflected from the load to be re-reflected.) Note that circuits operating at CMOS voltage levels do not typically use the Thévenin equivalent style of termination because the DC voltage at the join of the two resistors usually falls between the V_{IL} and V_{IH} when there are no other signals on the line. The AC termination style is used to prevent a DC current flowing through the termination, reducing power consumption [31].

Reflections also occur whenever the line branches, for example whenever a short section of line called a stub is used to attach the main line to an intermediate load. In this case, a signal traveling up the line sees a load of Z_0 in parallel with itself, $Z_o / 2$, and a reflection occurs. The effect of the reflection is small, however, if the propagation delay along the length of the stub length is kept small when compared with the signal rise time.

Small reflections also occur when the uniform transmission line structure is changed even for a distance of less than $\lambda / 8$ such as at vias, bends and input gates placed midway along the line.

11.3.2 Line Losses

There are two main sources of losses in transmission lines:

1. Resistive losses experienced by currents traveling in the signal conductor and reference plane return path.

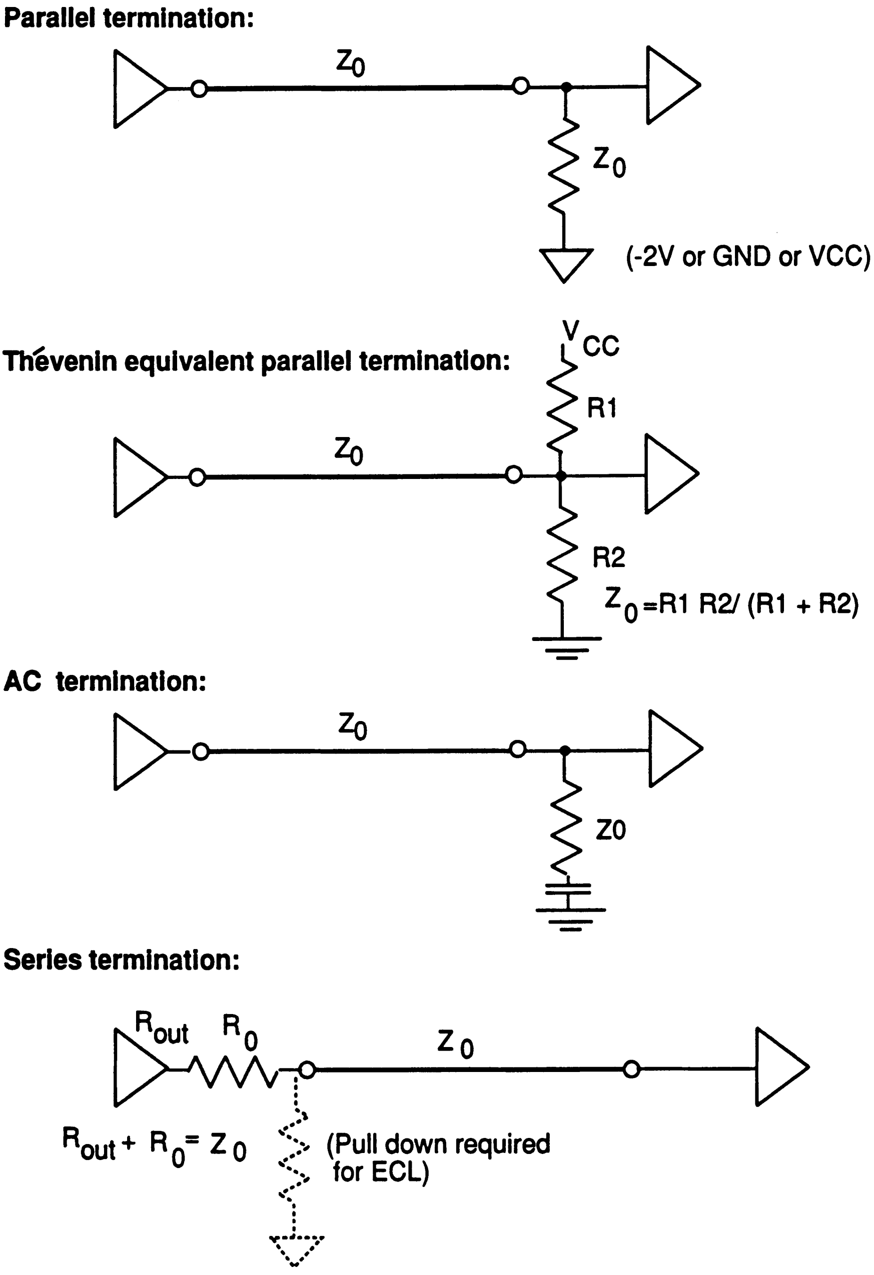


Figure 11-12 The four most common matching termination styles.

2. Dielectric losses due to some of the electromagnetic wave being absorbed in the dielectric material.

If the dielectric materials are chosen correctly, dielectric losses are negligible at most frequencies of interest. This is discussed in Chapters 5 through 8. However resistive losses can be significant, particularly for the very thin conductors used in thin film MCMs. The net effect of line losses is to attenuate the signal voltage and to increase its rise time as shown in Figure 11-13.

The signal voltage, V, placed initially on the line is attenuated to a voltage,

$$V_{\text{end-of-line}} = Ve^{-Rl/2Z_o} \qquad (11\text{-}15)$$

as it travels down the line. Here, R is the resistance per unit length and *l* is the line length. As shown in Figure 11-13, this attenuated signal has a slower rise time because of the line losses, and is followed by a slowly rising signal that eventually brings the voltage at the end of the line to the output voltage of the driver.

The line losses are dominated by the resistance of the signal line, the resistance of the reference return path usually being small (but not negligible at high speeds). This resistance depends on the material being used, the cross sectional geometry, and the pulse design bandwidth.

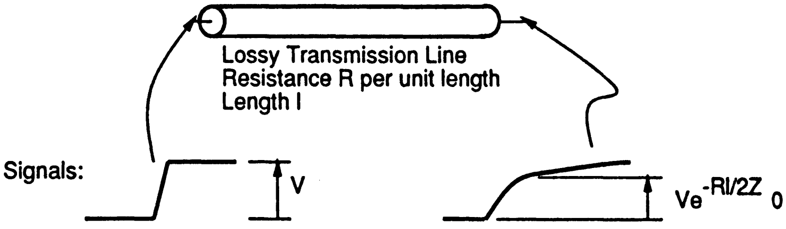


Figure 11-13 Effect of line losses on signal propagation.

Table 11-3 Resistivities of Common Conductor Materials.

MATERIAL	RESISTIVITY (10^{-8} Ω-m)	TYPICAL APPLICATIONS
Molybdenum	5.7	Cofired MCMs
Tungsten	5.7	Cofired MCMs
Copper	1.67	Thin film MCMs LTCCs PWBs
Aluminum	2.8	Thin film MCMs

Typical metal resistivities are given in Table 11-3. At DC, this translates into a resistance per unit length of

$$R_{DC} = \frac{\rho}{W \times T} \qquad (11\text{-}16)$$

where ρ is the resistivity, W is the line width and T is the thickness. For conductors typically found on PWBs, laminate MCMs and cofired ceramic MCMs (typically W = 100 μm and T = 30 μm) this translates into a resistance of less than 10 - 20 Ω/m. For a 30 cm long 50 Ω characteristic impedance line, less than 5 - 10% of the signal would be lost due to this resistance. On the other hand, the conductors in thin film MCMs are very small, typically 2.5 - 10 μm in thickness and 10 - 25 μm wide. A typical 2.5 μm × 15 μm aluminum conductor has a resistance of 747 Ω/m while a 10 μm × 15 μm copper conductor has a resistance of 110 Ω/m. For a 10 cm long, 50 Ω line, these resistances result in losses equal to 67% and 10% respectively. This is a major disadvantage in using aluminum conductors.

At frequencies above DC, however, the current concentrates in the skin of the conductor, as shown in Figure 11-14, with the current density decreasing with distance from the conductor edge. The current density distribution is characterized by a skin depth, δ_S, which is the depth at which the current density is the fraction 1 / e of the density at the conductor surface. This depth is given by

$$\delta_s = \sqrt{\frac{\rho}{\pi \mu f}} \qquad (11\text{-}17)$$

where ρ is the resistivity of the metal (see Table 11-3), μ is the magnetic permeability $\mu \approx \mu_0 = 4\pi \times 10^{-7}$ H/m and f is the frequency. The skin effect causes a significant increase in resistance at frequencies above those for which the skin depth becomes half the thickness of the conductor. This occurs at a frequency of 64 MHz for a 30 μm thick cofired ceramic MCM molybdenum conductor, for example, and above 670 MHz for a 5 μm thick thin film MCM copper conductor. If a significant portion of the pulse design bandwidth is much greater than this frequency then the skin effect increases signal attenuation. Furthermore, in thin film MCMs this causes the current to be concentrated in the higher resistivity nickel or chromium barrier metal [14]. Above these frequencies, the only way to decrease the resistance of the conductor is to make

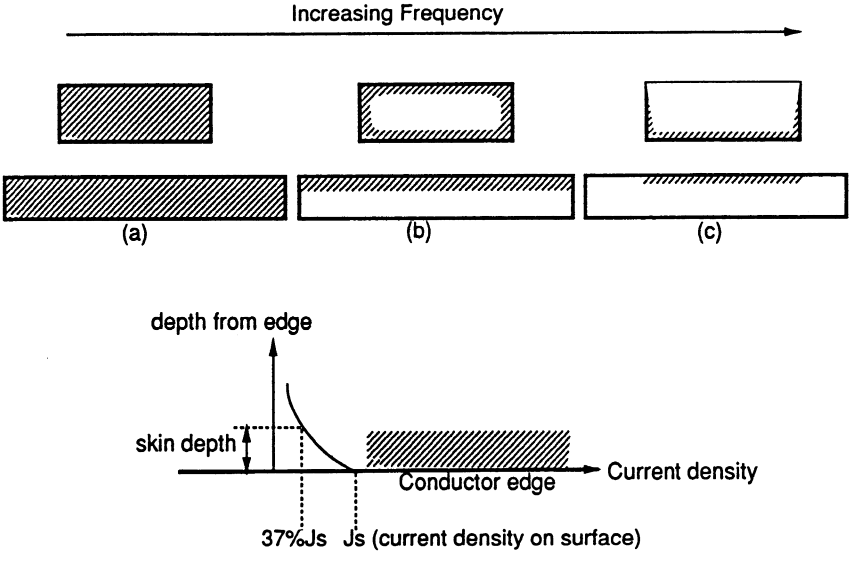


Figure 11-14 The skin effect. As the frequency increases, the current becomes more concentrated a → b → c.

it wider, the resistance being independent of thickness. Skin effect is only of concern for very fast digital systems operating at clock frequencies above several hundred MHz [15].

Note that as line losses attenuate reflected signals, the reflection noise in a lossy line is less than the noise in a lossless line. Sometimes it is possible to rely on line losses, rather than matching terminations, to control reflection noise, at the expense of increased propagation delay.

11.3.3 First Incidence Switching

To get the fastest propagation delay on a controlled impedance line, the first signal to arrive at the end of the line must have sufficient voltage to switch the receiver. (It should exceed V_{IH} on a 0 → 1 transition and be below V_{IL} on a 1 → 0 transition.) When this is achieved, the situation is called "first incidence switching." First incidence switching also requires that reflection noise be controlled well enough for the noise settling delay t_{settle} to be close to zero. If first incident switching is not achieved, then the propagation delay might be as much as five times the time-of-flight delay.

The first incidence voltage at the end of a matched terminated line is calculated by

$$V_{first} = V_{swing} \frac{Z_o}{R_{out} + Z_o} e^{-Rl/2Z_o} \qquad (11\text{-}18)$$

where V_{swing} is the open circuit output voltage swing of the gate, l is the length of the line, R is the resistance of the line per unit length and R_{out} is the output impedance of the buffer circuit. In Equation 11-18, the $Z_o/(R_{out} + Z_o)$ factor arises from the voltage division between the gate output and the line impedance. The Thévenin equivalent output resistance, R_{out}, is nonlinear and may have a value anywhere from 5 - 7 Ω for ECL, fast TTL and GaAs outputs, to 10 - 20 Ω or more for high speed CMOS and regular TTL outputs and to over 100 Ω for low speed outputs. Thus, in order to achieve first incidence switching, line losses must be low and the characteristic impedance of the line must be significantly greater than the Thévenin equivalent output resistance of the driver. For this reason, in high speed systems, the line characteristic impedance tends to fall in the range between 40 Ω and 75 Ω, higher values being preferred, particularly for CMOS drivers with their high values for R_{out}. The tradeoff is, however, that a higher value for Z_O requires either a narrower line or a thicker dielectric, both of which increase the wiring pitch. (See the crosstalk discussion below.)

Because the wire cross sections can be so thin in MCM-D technology, controlling line losses to achieve first incidence switching sometimes requires careful design. For example, consider a CMOS circuit. As $NM_H < NM_L$ (see Table 11-1), the $0 \rightarrow 1$ transition requires the largest V_{first} magnitude. Assuming $R_{out} = 15\ \Omega$ in the logic-1 state, $Z_0 = 65\ \Omega$, $V_{swing} = 5$ V, and $V_{first} = V_{IH} = 3.85$ V, then using Equation 11-18, first incidence switching requires that $R \times l < 7\ \Omega$. For a thin film 2.5 μm × 15 μm aluminum line $R_{DC} = 747\ \Omega/m$ and the longest line would be 9 mm, which is too short to be useful. For a 5 μm × 20 μm thin film copper conductor $R_{DC} = 165\ \Omega/m$ and the longest line would be 4.2 cm. For a cofired tungsten line (dimensions 30 μm × 100 μm, $R_{DC} = 19\ \Omega/m$), lines losses are low enough so that first incidence switching is not a problem for any practical line length. If no terminations are used then V_{first} is increased by a factor of $(1 + \Gamma)$ and as long as the line resistance sufficiently attenuates the reflection signal, then first incident switching can be achieved with longer lossy lines. However, the resistance increases with frequency. This changes the shape of the signal [15], increases the rise time and makes first incidence switching more difficult for long lossy lines for design bandwidths above 200 - 500 MHz. The smaller output impedances of TTL and ECL circuits make first incidence switching easier to achieve than in CMOS circuits.

If first incidence switching is achieved, then delay primarily depends on the values of dielectric constant and the capacitive load, as shown in Figure 11-15. Reducing delay is the main reason to reduce dielectric constant. Delay is also controlled by limiting the number of loads on a line, preferably to three or four. Propagation delay is affected, to some extent, by the chip connection choice because the chip connection technology is included in the load capacitance. Flip chip solder bump connections add the smallest excess capacitance and single chip packaging the most capacitance.

If first incidence switching is not achieved, then the delay is increased. The value of the extra delay depends on the reason for the non-first incident switching. If excessive line losses are the cause, then the extra delay is only an increase in $t_{rise\text{-}time\text{-}degradation}$, which might be small or large, depending on the magnitude of the losses. No simple analytic formula exists to predict this. If the cause is a high output impedance, $R_{out} > Z_0$, then the extra delay is large, several time-of-flight delays, t_{flight}. (The reflection coefficient at both ends is positive and the successive reflections build the voltage up to the required value.) A similarly sized extra delay is expected if excessive ringing noise is the cause.

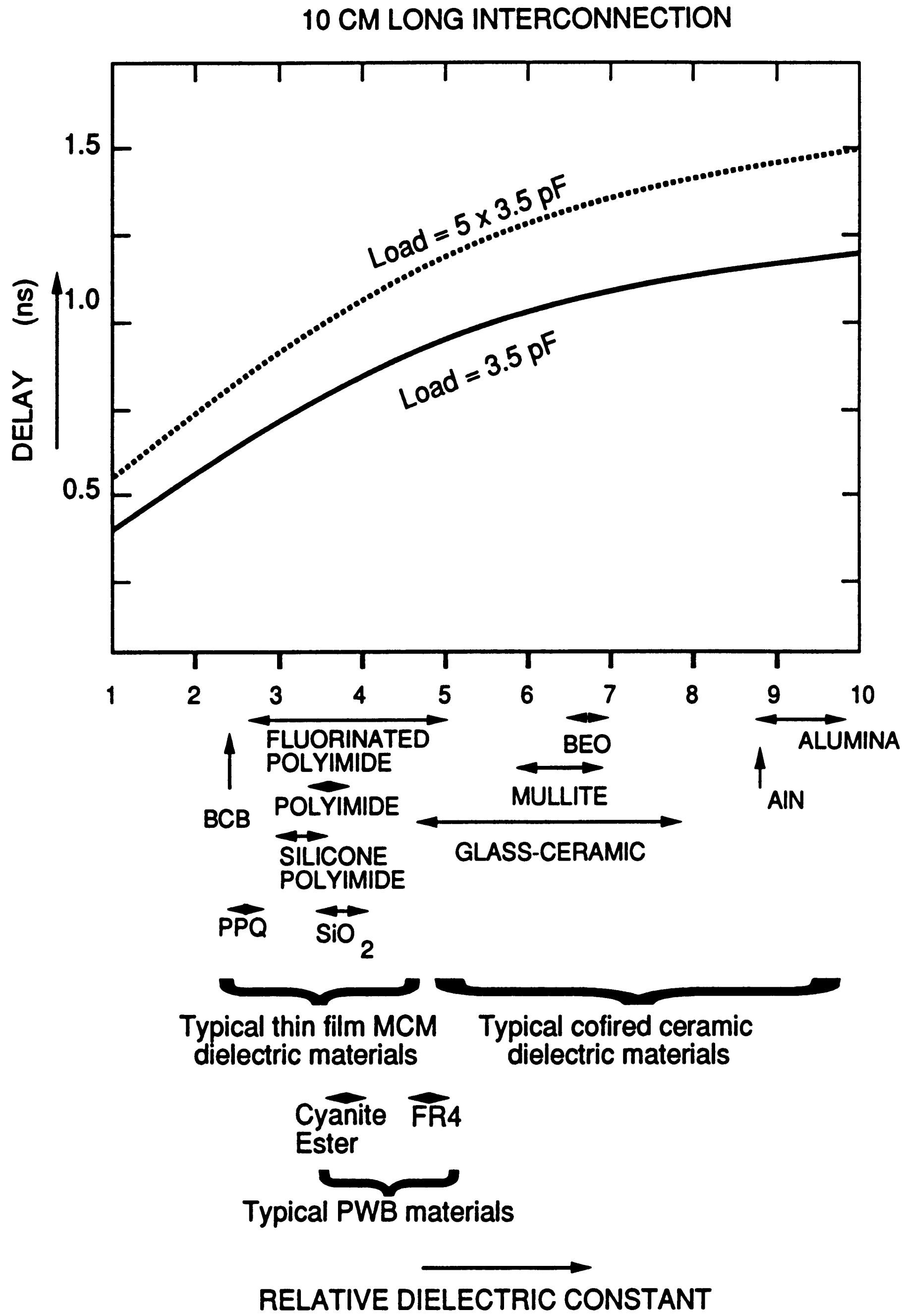


Figure 11-15 Effect of dielectric constant and capacitive loading on delay [16].

11.3.4 Net Topology

A net refers to the network of wires that join a set of transistor circuits. In this case, digital drivers and receivers. The topology refers to the net shape, as viewed from above.

If the timing design requires that ringing noise be controlled, and the net length is long ($t_{prop} > t_{rise} / 4$), then a matching termination, either series or parallel, is needed and the stub lengths must be limited. Controlling the stub lengths means controlling the net topology. Acceptable net topologies for controlling reflection noise are shown in Figure 11-16. Note that a ring topology

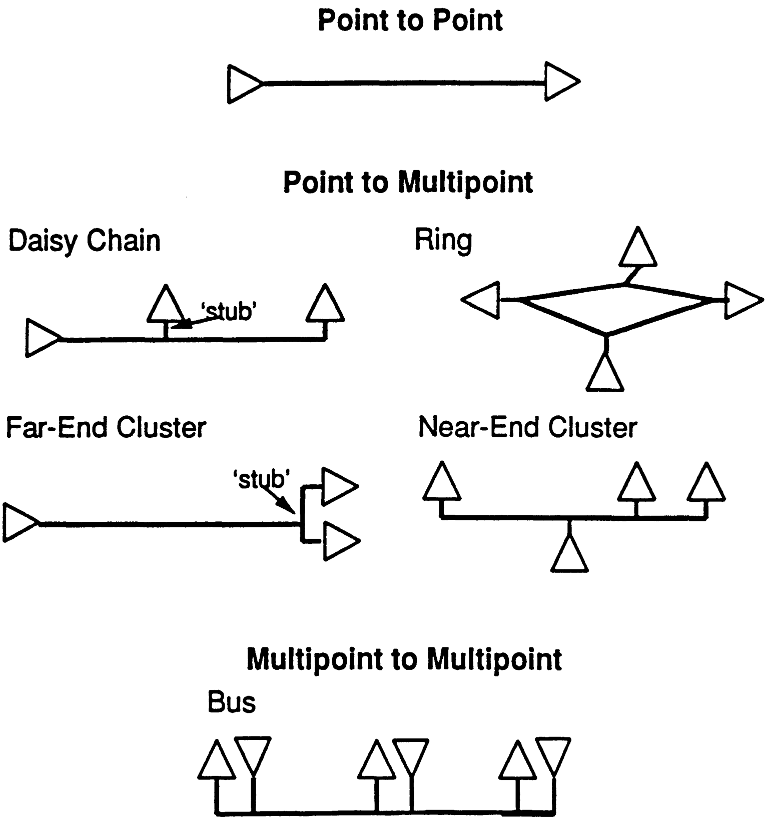


Figure 11-16 Different net topologies to handle reflection noise.

does not require a matching termination, at the expense of increased wiring, and the bus topology can be treated as either a daisy chain or near end cluster, depending on which driver is in use. The most common point to multipoint topology is the daisy chain. The acceptable stub length depends on how much ringing noise is acceptable and how fast the rise time is. The faster the rise time, the shorter the acceptable stub length.

11.3.5 Effect of Loading

If the loads (receivers) are closely spaced along a net, then their capacitance effectively adds to the self-capacitance of the line. This is referred to as distributed loading along the line. By closely spaced, a spacing, l_s, between the loads such that $l_s / v_{prop} < t_r/2$ is meant. The capacitive loading might be in the form of the capacitance of the attached drivers and receivers, mutual capacitance from crossing lines or mutual capacitance with surface pads. If the total distributed load capacitance is C_L, then the effect of this loading is to slow the propagation velocity to

$$v_{prop} = \frac{c / \sqrt{\varepsilon_r}}{\sqrt{1 + C_L / C_0 l}} \qquad (11\text{-}19)$$

where C_0 is the per unit length self capacitance of the line, and l is the length of line over which the load C_L can be considered to be distributed. As the load capacitance has a large effect on delay, critical nets are limited to a maximum of three to four loads if possible.

The characteristic impedance Z_0 is reduced to

$$Z_L = \frac{Z_0}{\sqrt{1 + C_L/C_0 l}} . \qquad (11\text{-}20)$$

With heavily loaded lines, this reduction might be 50% or more. This can be compensated by increasing Z_0 (this is often done in backplanes, where Z_0 might be as high as 120 Ω). If Z_L becomes too low then first incidence switching cannot be achieved and the increase in delay is even more than the decrease in v_{prop} would indicate.

The effects of loading also are compensated by reducing R_{out}. This is done by inserting a special driver chip after the outputs of the ASIC. Manufacturers produce drivers that guarantee first incidence switching for loaded characteristic impedances as low as 30 Ω. The internal delay of the driver must be accounted for in the total delay.

11.4 CROSSTALK NOISE

Mutual inductance and capacitance between different electrical signal paths contribute to unwanted electrical coupling known as crosstalk noise. Whenever a signal edge travels down a signal wire, chip attach lead or connector lead, both forward and backward crosstalk noise pulses are induced in the neighboring two wires, as shown in Figure 11-17. Capacitive coupling, $K_C = C_m / C_0$, and inductive coupling, $K_L = L_m/L_0$, between adjacent lines add at the near end of the quiet line and subtract at the far end. The maximum noise voltage at the near end can be approximated by:

$$V_n \approx K_B \left(\frac{2}{v_{prop}}\right)\left(\frac{V_S}{T_1}\right) l \text{ if } l < \frac{v_{prop} T_1}{2} \qquad (11\text{-}21)$$

$$V_n \approx K_B V_S \text{ if } l > \frac{v_{prop} T_1}{2} \qquad (11\text{-}22)$$

where $K_B = (K_C + K_L) / 4$ is the coupling coefficient, C_m is the mutual capacitance between the lines per unit length, L_m is the mutual inductance per unit length, C_0 is the self capacitance per unit length of each line, L_0 is the self inductance per unit length, V_s is the voltage swing on the active line, $v_{prop} = c/\sqrt{\varepsilon_r}$ is the propagation velocity of electromagnetic waves in the dielectric, T_1 is 0% - 100% rise time of the signal, and l is the coupled line length. When the crosstalk stops increasing with coupled length (see Equation 11-22) is referred to as saturated crosstalk and the length $l = v_{prop}\ T_1 / 2$ is referred to as the saturated line length. More accurate, but more complex, analytic expressions for crosstalk can be found in reference [17].

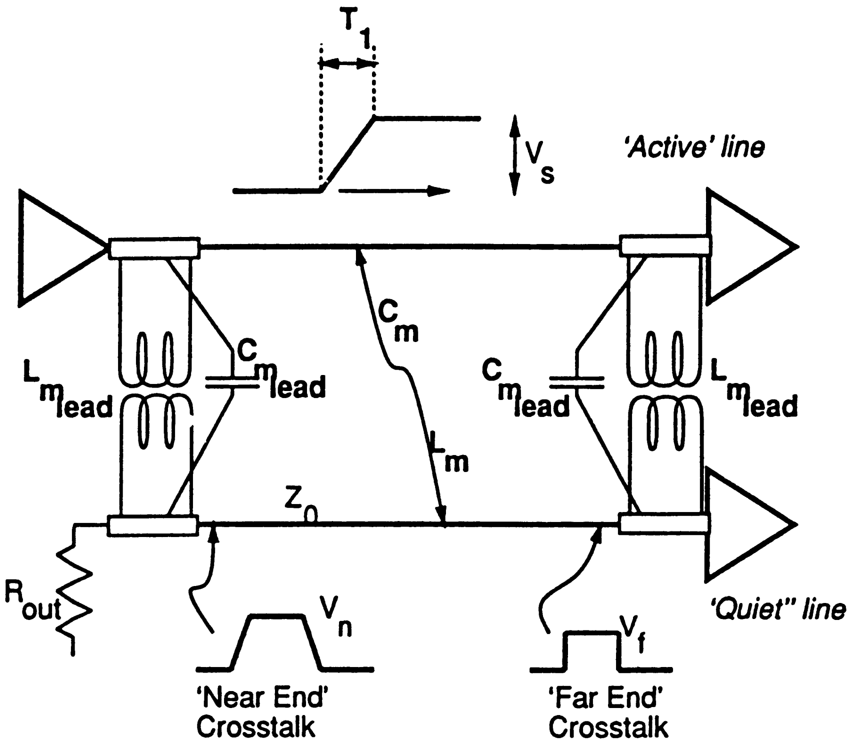


Figure 11-17 Crosstalk noise arises from the effects of mutual inductance and capacitance. This diagram show how a signal travelling down an "active" line causes crosstalk signals at both ends of a quiet line.

The maximum noise voltage at the far end is given approximately as,

$$V_f \approx K_F \left(\frac{2}{v_{prop}} \right) \left(\frac{V_s}{T_1} \right) l \qquad (11\text{-}23)$$

where $K_F = (K_C - K_L) / 4$ is the coupling coefficient. If the medium in which the line is buried contains no other materials beside the dielectric then the far end crosstalk is zero because $C_m / C_0 = L_m / L_0$. This is only perfectly achieved within homogeneous-dielectric, stripline conductors, and if no other conductors are present. Even with other conductors present, and for buried microstrips, however, K_F is usually small.

If the neighboring wires do not have matching terminating resistors then the noise pulses are reflected from each end. As a result, part of the near end noise shown in Figure 11-17 will still arrive at the receiver and matching terminations can be used to help reduce the effects of crosstalk noise.

Plots showing near end crosstalk versus length for an alumina cofired MCM with a 7.5 mil line separation, taken from Sons *et al*, [18], are shown in Figure 11-18. For the plot labeled $T_1 = 1$ ns it can be seen that the saturated line length is about 2.5" (6.35 cm).

Crosstalk requirements can determine the required spacing between the lines and thus the signal line pitch (line width + line spacing). Often the required spacing is at least twice the minimum spacing that the technology would allow. As the spacing increases, C_m and L_m decrease and crosstalk noise decreases. The required line pitch to control crosstalk also must increase with characteristic impedance Z_o. When Z_o is increased by making the lines narrower, the mutual capacitance and inductance increases if the spacing remains the same. To keep the crosstalk noise voltage constant, the line spacing must be increased. As the increase in spacing is greater than the decrease in width, the line pitch increases. The characteristic impedance value chosen might actually be a compromise between the desire for first incidence switching, particularly when the line is loaded, and the desire to maximize interconnect density [1].

Choosing a material with a lower dielectric constant has two positive effects on crosstalk. First, by allowing the lines to be brought closer to the reference planes for the same impedance Z_o the interline spacing can be reduced with the coupling coefficients K_B and K_F remaining unchanged. Second, it reduces v_{prop} and increases the saturated length, effectively moving the plots in Figure 11-18 to the right.

Choosing a stripline over a microstrip configuration also allows a reduced spacing for the same impedance Z_o, as the presence of two reference planes in the former reduces the ratio C_m / C_0 (therefore, K_B and K_F). In either case, the greater the distance from the ground plane, the larger both the line width and spacing must be for the same Z_o and crosstalk.

The placement of a ground line between adjacent microstrip or stripline signal lines can be used to reduce crosstalk by almost 50%. This is only beneficial if the spacing already is large enough to consider adding a ground line. The ground line cannot be placed too close to the signal lines or it changes the characteristic impedance of the signal line [19].

Sometimes, crosstalk also may be reduced by rerouting parallel lines so that the length of the parallelism is minimized. In thin film MCMs this can be done with little penalty because of the MCM's large routing capacity.

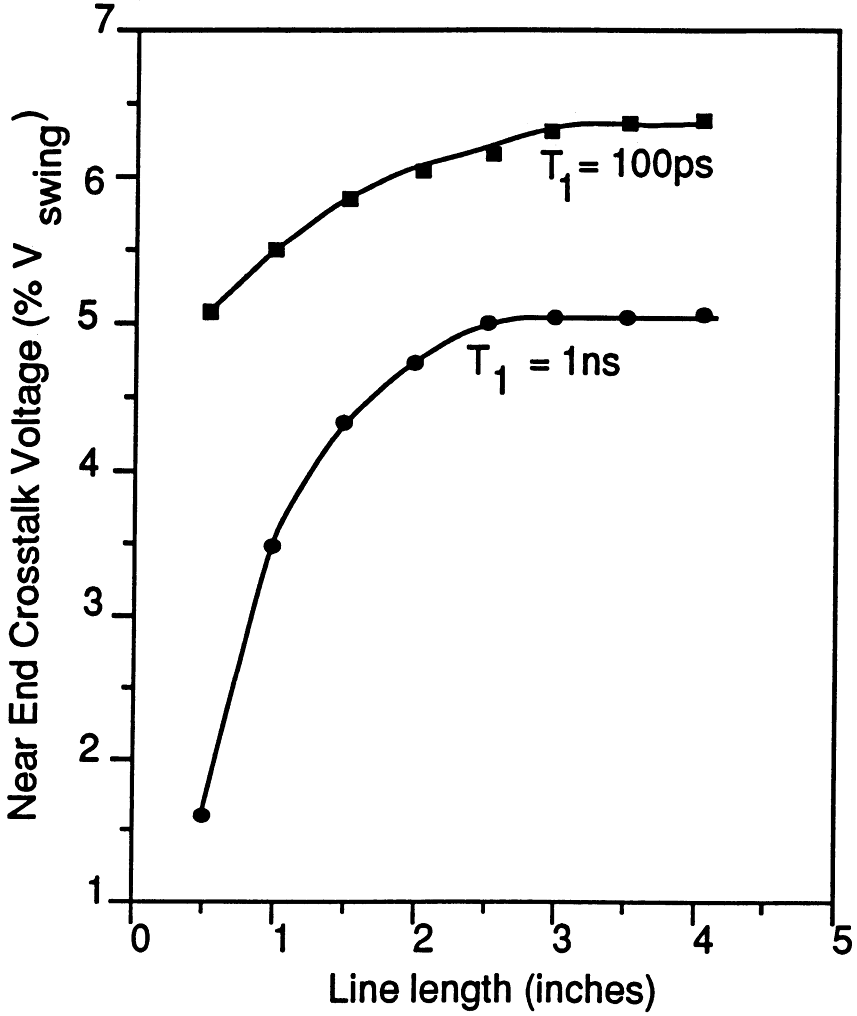


Figure 11-18 Simulated crosstalk versus length for an alumina cofired ceramic substrate [18].

Crosstalk is greatly affected by the capacitive loading represented by the crossing of lines in the adjacent signal layer. The reason why the lines are crossing in the first place is to prevent the high amounts of crosstalk that occurs when lines in different layers run parallel on top of each other for any significant length. This orthogonality is enforced by designating alternate layers as being mainly X running or Y running layers. The effect of multiple orthogonally crossing lines, however, is to increase C_0 and to decrease C_m. Deutsch *et al*, [20] report an example, where the capacitive coupling K_C reduces by about 40% when the crossing lines are pitched at minimum pitch and are continuously distributed along the signal line. As a result, the near end noise decreases and the far end noise increases. This effect is greater for two crossing microstrip layers than for two crossing offset stripline layers. The effect disappears when symmetric stripline layers are used, that is one signal line only centered between every pair of reference planes.

Another source of crosstalk is between chip connection bonding leads. Inter-lead crosstalk gets worse as the leads become longer or become more tightly spaced, and also as the rise time of the signal becomes shorter. For wire bonds and TAB leads of equal length and pitch, the crosstalk contribution on the bonding leads is nearly the same (with $C_m \approx 1$ pF and $L_m \approx 1$ nH). With high speed systems ($f_{clock} > 50$ - 75 MHz), wire bond and TAB leads must be kept short and/or consideration be given to incorporating a ground plane beneath them (multimetal TAB [20]). The crosstalk between flip chip bonds is very small due to the short lead lengths. Similar considerations apply in single chip packages. For example, most PGAs contain extensive ground planes, while most surface mount packages do not. As a result, lead crosstalk usually is smaller in a PGA package.

Crosstalk between MCM-to-board connector leads must also be considered for signal lines that leave the MCM. Controlling this noise at high speeds might require that every other pin be assigned to power or ground.

11.5 SIMULTANEOUS SWITCHING NOISE

When a number of off-chip loads are switched simultaneously in a digital system, a current change is produced in the power and ground supply network. For example, consider the CMOS circuits at the top of Figure 11-19. Whenever a 1 → 0 occurs at the outputs of a buffer (driver), the capacitive loads connected to that driver are discharged through the ground, producing a current spike through the ground. For example, consider a 5 V, 32-bit driver chip with a rise time of 2 ns driving a load of 320 pF (10 pF/gate). This corresponds to a $di/dt = C\Delta V / \Delta t = 0.8$ A/s. Switching noise is also a problem in TTL and ECL logic families though it tends to be worse in CMOS logic. For example, in an ECL logic state change, an 0.8 V swing into a load with an impedance of 50 Ω might occur in 700 ps. This corresponds to a rate of change of current di/dt of 16 mA/700 ps = 0.02 A/s. In either case, when this transient current passes through the inductive power distribution network, a noise voltage is produced ($V = Ldi/dt$), which is referred to as simultaneous switching or delta-I noise and is shown in Figure 11-19.

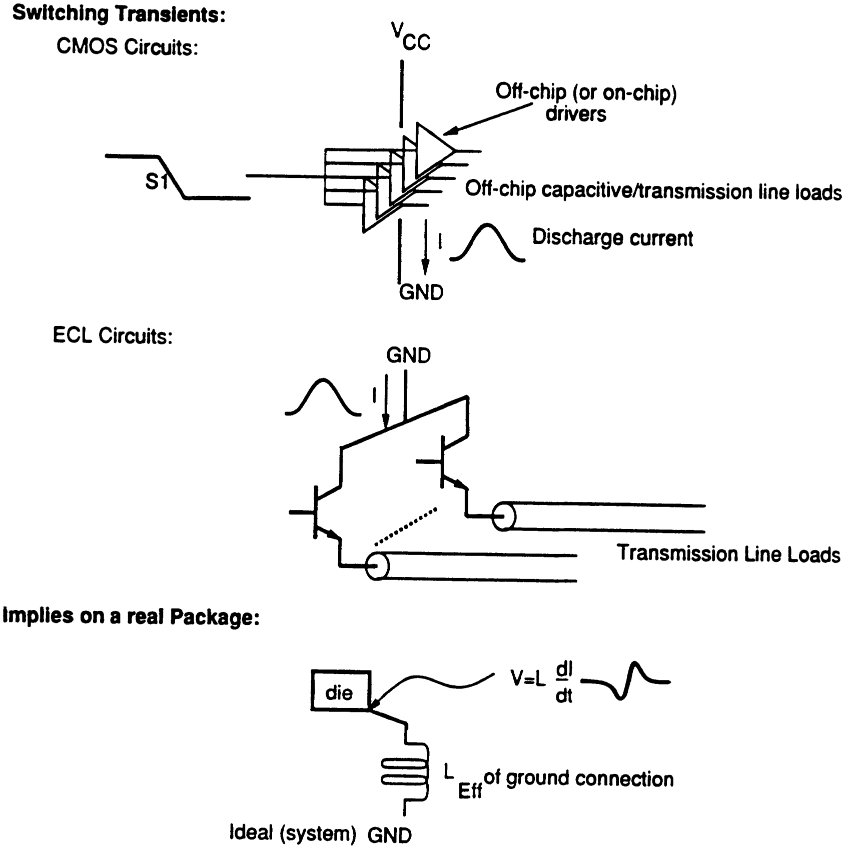


Figure 11-19 Simultaneous switching noise in CMOS and ECL circuits.

Note that noise can occur in a similar fashion on power rails, such as V_{CC} and V_{EE}. However, this noise is less than the noise on the ground connections mainly due to the properties of the circuits. For this reason, simultaneous switching noise is sometimes referred to solely as "ground bounce." (The noise is likened to the ground voltage bouncing.) The transmission line return current transients (see Figure 11-9) flowing on the ground and power reference planes make a small contribution to the ground and power supply noise at each chip.

Switching noise can result in a number of problems if not handled correctly [21]. From top to bottom in Figure 11-20, the following effects occur:

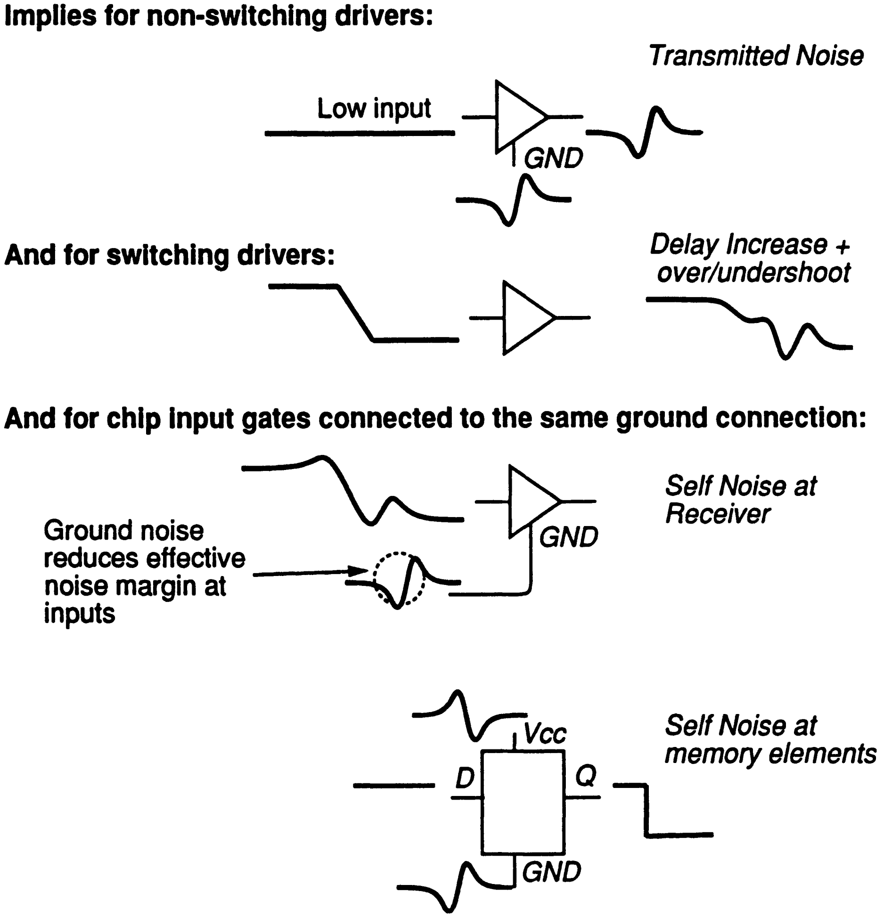


Figure 11-20 Effects of simultaneous switching noise.

1. Noise appears at the output of what were intended to be quiet off-chip drivers. This noise appears at the inputs of connected receivers.

2. The changes in internal chip supply voltage make the circuits operate more slowly and, thus, increase the delay in the switching drivers. The delay increase might be up to 2 - 3 ns or more for CMOS and TTL circuits, depending on the circuit details and the number of switching drivers. Overshoots and undershoots might also appear in these drivers.

3. For on-chip circuits acting as input gates, simultaneous switching noise acts to reduce the effective noise margin at the inputs.

4. For on-chip memory devices, such as latches, large amounts of ground-rail and power-rail noise might cause false changes in state (a logic-1 becoming a logic-0).

To a first order, the noise generated by the simultaneous switching of *N* output drivers can be estimated as

$$\Delta V = N L_{eff} \frac{di}{dt} \tag{11-24}$$

where L_{eff} is the effective inductance of the power and ground connections and di/dt is the peak rate of change of current. This rate is often approximated as $\Delta i/\Delta t$ where Δi is the current demand of each driver during the switching event, and Δt is the rise (fall) time of the signal. It is only a first order equation because it makes the assumption that di/dt is independent of N and L_{eff}. In fact ΔV does not increase linearly with L_{eff} or N because of a feedback effect between ΔV and di/dt. Any increase in ΔV due to an increase in L_{eff} or N tends to result in a decrease in di/dt as the reduced voltage slows down the circuits. Thus use of Equation 11-24 often leads to an overestimate of expected noise. Improved expressions accounting for this effect in CMOS drivers are presented in references [23] and [24]. Even these expressions will tend to overestimate ΔV, however.

The effective inductance, L_{eff}, is primarily a function of the package design. In contrast, N is dependent on the logic design and di/dt on the circuit design. Reducing L_{eff} requires minimizing the inductances of the power and ground distribution networks and also making use of bypass capacitors. Bypass capacitors placed between the power and ground pins of each chip can act as a local source of charge during switching events so that not all of the switching current has to be supplied from the system ground, minimizing the local change in voltage. The equivalent circuit model for a CMOS output driving a capacitance is shown in Figure 11-21. In this model, $L_{ground\text{-}lead}$ is the inductance of the ground lead in the chip attach, L_{gnd}, is the inductance of the ground plane or wiring between the chip attach and the bypass capacitor, R_0 and L_0 are the resistance and inductance between the bypass capacitor and the power supply. Also shown is the parasitic inductance and capacitance associated with the bypass capacitor. Thus, the factors through which the package contributes to L_{eff} are as follows:

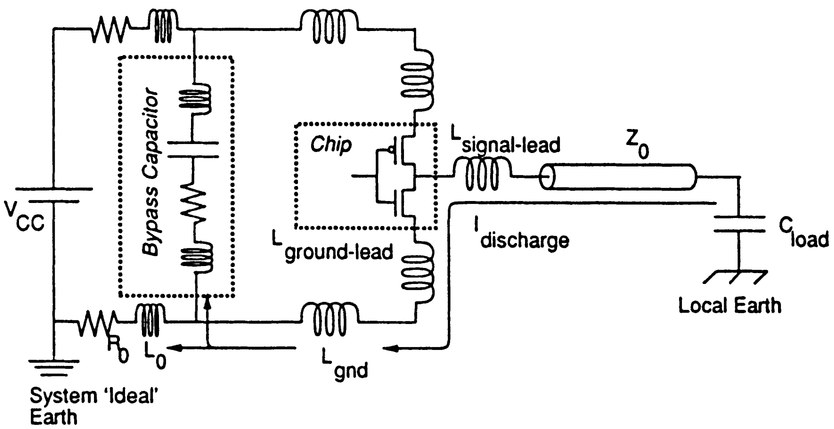


Figure 11-21 CMOS circuit model for simultaneous switching noise. With correct choice of bypass capacitor, most of the discharge current is supplied by the charge stored on this capacitor.

- **Choice of chip connection technology** - typical lead self inductances for different technologies are given in Table 11-2. Basically, inductance increases with the length of any round or rectangular lead. Of the MCM connection techniques, solder bump technology has the least inductance while fanout TAB and wire bonds have the most. The ground lead inductance, however, is less than the self inductance due to the effect of mutual inductances [25]:

$$L_{eff} = \frac{L_{self}\, I_{ground} - \Sigma\, L_{mutual-i}\, I_{signal-i}}{I_{ground}} \qquad (11\text{-}25)$$

where L_{self} is the self inductance of a ground lead, $L_{mutual\text{-}i}$ is the mutual inductance of the ground lead with signal lead i, $I_{signal\text{-}i}$ is the current flowing in signal lead i and I_{ground} is the current flowing in the ground lead. Note that the ratio of $I_{signal\text{-}i}$ to I_{ground} is the same as the signal to ground lead count ratio. The more ground leads provided for the same number of signal lines, the less the switching noise (actually this is mainly due to the decrease in N in Equation 11-24). The effective inductance is also substantially reduced if ground and power planes are used instead of leads, as in multimetal TAB and PGAs.

- **Choice of bypass capacitor** - at high frequencies it is important to minimize the parasitic inductance of the bypass capacitance. One technique to do this in a thin film MCM is to closely space the power and ground planes and use this as a low inductance bypass capacitor as done in the DEC VAX-9000 MCM (see Chapter 17).

- **Location of bypass capacitor** - on PWBs, bypass capacitors are located right next to the chips. On MCMs this might be undesirable due to their size (about 1 mm x 3 mm). One solution is to use the power and ground planes in a thin film MCM. Another is to build the capacitors into the MCM in a cofired or multilayer thick film ceramic MCM. Alternatively, the capacitors could be placed at the edge of the MCM, or off the MCM entirely. Both of the last two approaches lead to an increase in L_{gnd} (shown in Figure 11-21) and L_{eff}. If the capacitor is placed off the MCM, the MCM-PWB connector inductance must be included in L_{eff} and N must be evaluated MCM-wide rather than for each chip [26] and [22] - [23].

- **Choice of ground and power network components** - to minimize L_{gnd} and L_0, ground and power planes should be used, the connectors should be selected for minimum inductance and multiple ground and power pins should be distributed evenly through all connectors.

Techniques to minimize switching noise and its effects are discussed in a number of sources [11], [12], [22] and [27] - [28].

11.6 OTHER SOURCES OF NOISE

There are several minor sources of noise that must be considered. First, the resistive voltage drop caused by the DC supply currents in the power and ground planes must be kept to less than 1% of the power supply voltage. This is a potential problem in high power thin film MCMs. If so, solutions include using multiple power supply planes (such as the VAX-9000, see Chapter 17) or using a ceramic substrate base with thick film ground and power planes within it.

Second, if a parallel termination style is used, the resistance of the signal lines prevents the DC line voltage from settling at the nominal voltages for logic-0 or logic-1, thus consuming part of the noise margin. The driver output resistance, the line resistance and the parallel termination form a voltage divider circuit that reduces the voltage across the latter and across the chip input.

Third, there might be small amounts of electromagnetic interference noise (EMI) produced in the circuit from outside sources of electromagnetic radiation, such as electric motors. This effect is small in MCMs because of their small size (the noise depends on the size of the antennas formed by the circuit interconnections) and can be neglected.

Fourth, some of the internal chip noise might appear at the outputs of the chip.

Finally, there is so called thermal noise caused by different operating temperatures for connected chips. The operating temperature of the chip has an effect on DC parameters, such as V_{IH} and V_{OH}, particularly for ECL circuits. Thus, the expected temperature difference between chips reduces the noise margin by a small amount.

11.7 THE ELECTRICAL DESIGN PROCESS

11.7.1 Technology Selection and System Planning

This process is summarized in Chapter 3. From an electrical design point of view, a packaging technology must be selected so that, with a suitable choice for partitioning and floorplan, delay and noise aims are met.

Noise control has the greatest number of requirements. As the circuits in the system get faster, rise times become faster, and more stringent requirements on the technology arise. In particular, controlled impedance interconnection is required, and chip connection and connector styles with lower inductances and less mutual coupling are preferred. For example, at signal frequencies above 50 MHz, long leaded wire bonds and TAB frames become undesirable. Either short lead connection methods must be used or reference planes are needed beneath the leads. (One way to provide this reference plane for wire bonds is to use multitier chip connection structures, examples of which are given in Chapter 6. The lower tier forms a ground plane over which the signal wire bonds run.) As the rise time increases, striplines are preferred over microstrips so that the mutual coupling between crossing lines in adjacent signal layers is reduced. Also the design of bypass capacitors becomes more critical. One advantage of MCM technology over single chip technology is that it makes it easier to meet these requirements.

The desire to produce fast first incidence switching creates requirements on line impedance, dielectric constant and line losses. The latter become particularly difficult to handle with thin film interconnect if aluminum lines are

Table 11-4 Noise Budget for the ECL VAX-8600 [29].

SOURCE	BUDGET (mV)
Load Reflections	100
Interconnect Impedance Mismatch	100
Crosstalk	100
Simultaneous Switching Noise	150
-2.0 V AC Noise	25
Signal IR Drop	25
V_{CC} IR Drop	14
Internal Chip Noise	50
Temperature	6
Sum	570
RSS	237

used or if the design bandwidth is so high that the skin effect becomes important. Guidelines for thin film technology selection accounting for this and other effects, are given by Gilbert and Walters [27].

The desire for a high line impedance must be balanced against the desire to minimize the line pitch, as determined by crosstalk requirements. If the line pitch is large, then extra layers might have to be added to create more interconnect capacity. An impedance of 50 Ω is commonly used but higher impedances are sometimes used for CMOS systems and heavily loaded lines. In any case, as the rise time gets faster, the line spacing has to increase. For longer lines, the minimum manufacturable spacing is unlikely to be acceptable. Tighter spacings can be used in stripline than in microstrip layers. This consideration is particularly important in laminate and ceramic technologies as one wishes to avoid the extra cost of adding layers. The wiring capacity provided by two layers of interconnect in a thin film MCM is sufficient for most systems today.

Noise Budgeting

The relative effort that goes into controlling each noise source, reflection noise, crosstalk noise and simultaneous switching noise, depends on the noise budgeted for each. Determining the noise budget is part of the system planning process. An example of a noise budget, used in the design of the VAX-8600 [29], an ECL system, is given in Table 11-4. This noise budget accounts separately for two different types of reflection noise, reflections from loads and reflections due to mismatches between different transmission lines. The crosstalk noise is given a budget of 100 mV. The simultaneous switching noise refers to the noise placed at the outputs of quiet drivers when grounded. Simultaneous switching noise on the -2.0 V power rail, resistive voltage drop noise, internal chip noise and thermal noise also are accounted for. Another good example of a noise budget is given in reference [30].

The total DC noise margin for ECL circuits is given in Table 11-1 as 150 mV. A typical ECL AC noise immunity curve is given in Figure 11-4. It can be seen that the sum of the different noise sources exceeds both the DC noise margin and the highest point on the AC noise immunity curve. This worst case, when pulses produced by the various noise sources arrive simultaneously at the receiver, is unlikely. If this assumption is used to determine the noise budget, the allowed noises would be very small and the system very difficult to design. Instead it is assumed that the noise sources arrive at the receiver at random times. This is far more reasonable and is sufficiently safe if a margin of safety is included. With this assumption, it can be shown that the total noise does exceed the Root Sum of Squares (RSS) of the different noise voltages for over 99.7% of the time. In this case, the RSS Noise voltage is

$$V_{RSS} = \Big(V^2_{\text{load-reflection}} + V^2_{\text{mismatch-reflection}} + V^2_{\text{crosstalk}} + V^2_{\text{SSN}} + V^2_{\text{AC}} + V^2_{\text{IR-sig}} + V^2_{\text{IR-}V_{CC}} + V^2_{\text{chip-noise}} + V^2_{\text{thermal}}\Big)^{\frac{1}{2}} . \tag{11-26}$$

In the case of the VAX-8600, the designer chose a V_{RSS} that was significantly less than the noise immunity for a suitably long pulse. (It appears to be about 150 mV less.)

The relative weight given to the different noise sources depends on how difficult it is to control each source. In the noise budget in Table 11-4, it is recognized that the simultaneous switching noise is the most difficult to reduce further and, therefore, is given the largest weighting. For example, if the simultaneous switching noise could be reduced, by improved technology, it might be possible to increase the crosstalk budget and thus allow narrower line spacings.

The noise budget for Table 11-4 is for a data connection. As discussed earlier, noise control on clock connections is more important. Noise budgets for clock signals tend to be far more conservative.

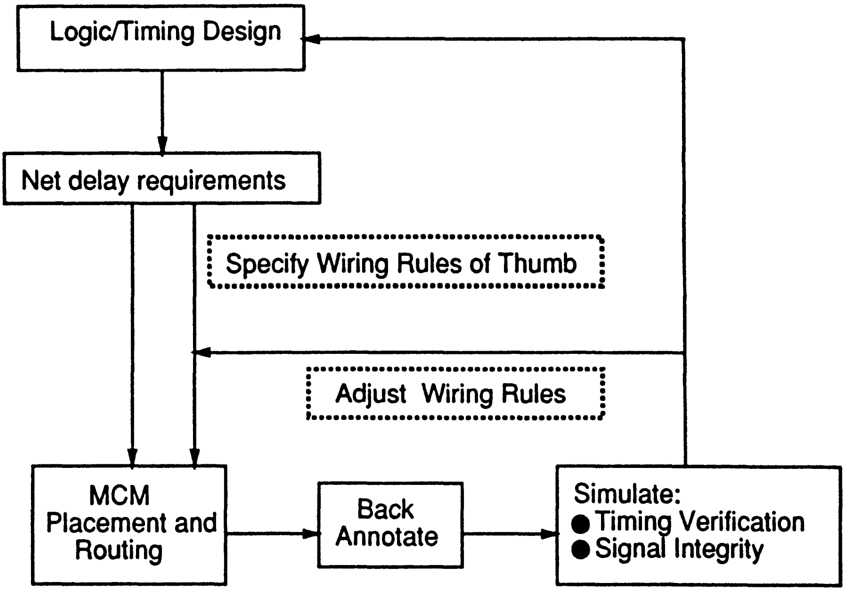


Figure 11-22 The steps in producing an MCM layout from an electrical design point of view.

11.7.2 Modeling, Simulation, and MCM Layout

The final objective of the electrical design of the MCM is to produce a layout which is a description of the artwork used to make the masks to be applied in MCM production.

The process to produce this layout, as supported by today's computer aided engineering (CAE) tools is described in Figure 11-22. From the timing design, delay requirements for each net are produced by subtracting the worst case delays of the active components between each pair of latches (see Figure 11-1) from the clock period. (Note: Any effects of simultaneous switching noise on this delay must be included. Data books rarely include this in their delay specifications. Typically for CMOS drivers, an extra 250 ps of extra delay is required for each simultaneously switching driver.) From these delays an estimate of the wiring rules is produced. The wiring rules specify which nets need to use controlled topologies and matching terminations, the limits on the lengths to each receiver and the stub length limit for those nets, as well as the spacing requirements between nets. Not all nets require that the conditions for first incident switching be satisfied. The rules for these nets do not have to be as stringent. See references [12] and [1] for more details. Care must be taken in using Equation 11-19 to estimate delay because the propagation delay is only a part of the total delay. This is particularly true in thin film MCMs where line losses can significantly increase $t_{\text{rise-time-degradation}}$ [31]-[32].

The wiring rules are passed to the placement and routing CAD tools which produce the layout. It is necessary to verify that delay requirements and noise budget requirements (signal integrity requirements) are met. This requires that electrical models be obtained and the simulations be undertaken and studied. The results of these simulations are compared with the requirements. If they predict that the layout will not produce a working design, then the wiring rules are adjusted and the process iterated. For example, if the simulated delay is too long, the maximum allowed length is reduced. Unfortunately, many iterations might be required. The need for iterations might be avoided by taking more care in producing the initial rules [1]. Computer aided techniques to automate and improve the determination of initial wiring rules are a current area of investigation [31].

Models must be obtained for the drivers, receivers, transmission lines, and line discontinuities including vias, chip connection leads and connectors. The driver and receiver models are either Thévenin equivalents or, for more accuracy, full circuit models (typically use of full circuit models improve delay prediction accuracy by at least 300 ps). Take care, however, as many of the circuit models provided by parts vendors are approximate models, not the real ones they use in design. They might need to be verified and qualified through measurements. There are a variety of modeling approaches for the packaging structures including using empirical equations (for example, Equation 11-11 for Z_0 of a microstrip), various numerical solution techniques implemented in a range of CAE tools and measurement-based models.

A variety of simulation techniques have been implemented in a number of computer-based simulators. It is important to note that some interconnection parameters have frequency dependent behavior, particularly the skin effect resistance. If this frequency dependence is strong within the design bandwidth then specific modeling and simulation techniques must be used [33].

11.8 SUMMARY

Electrical design is concerned with delay and noise control, the control limits being determined by the timing design and the noise budget. To some degree, over-the-budget noise on data signal lines can be tolerated by allowing some time for the noise to settle. Excess noise on clock lines cannot be tolerated. The delay of the data paths depends on the output buffer characteristics, the characteristic impedance of the line, the length of the line, the dielectric constant of the dielectric, the line losses and the load capacitance of the line and the amount of noise that must settle out. Thus, delay is improved by ensuring that the paths are short, the dielectric constant is low, the line resistance is low, the chip leads have low capacitance and inductance and that noise is controlled (high levels of signal integrity). Simultaneous switching noise is controlled through careful system design including the use of bypass capacitors and other techniques. The control of reflection noise and crosstalk noise involves careful design of the interconnections and connections themselves. Many of the design issues for CMOS and TTL systems are covered further by Buchanan [12], for ECL systems by Blood [3] and for GaAs systems by Long and Butner [5].

The packaging technology choice has a very strong effect on delay and noise. By using MCMs, particularly thin film MCMs, delay and noise are substantially reduced. Most of this improvement comes from the removal of the single chip package whose leads introduce considerable capacitive and inductive parasitics and whose size often increases path lengths substantially. Even after the basic technology is chosen, however, there are still many decisions to make. For example, the interconnect material affects line resistance and delay. The chip connection technology affects both crosstalk noise and simultaneous switching noise. Fanout TAB, with its long leads, is the worst in both regards unless a ground plane is added. Short lead TAB and short wire bonds usually are acceptable except for the fastest systems. Flip chip connection provides the most superior electrical connection. Further technology factor guidelines for thin film MCMs, categorized by desired clock frequency of operation, have been formulated by Gilbert and Walters [27].

The use of MCM technology has one disadvantage, however. The MCM has to be carefully designed as prototyping is expensive, debugging difficult and redesign undesirable. The appropriate CAE tools must be used in high speed designs. Nevertheless, the resulting performance advantage is a significant factor in driving many system users to consider MCM technology.

Acknowledgments

The author sincerely thanks Hassam Hashemi and Peter Sandborn for their extensive suggestions and feedback on earlier portions of this draft. He would also like to thank Steve Lipa, Sharad Mehrotra, Slobodan Simovich, Michael Steer, Douglas Thomae, and the anonymous reviewers for their comments and suggestions.

References

1 E. E. Davidson, G. A. Katopis, "Package Electrical Design," R. R. Tummala, E. J. Rymaszewski, eds., *Microelectronics Packaing Handbook*, New York: Van Nostrand Reinhold, 1989, Chapter 3.

2 C. F. Hill, "Noise Margins and Noise Immunity in Logic Circuits," *Microelectronics*, vol. 1, pp. 16-21, April 1968.

3 W. R. Blood Jr., *MECL System Design Handbook*, Phoenix AZ: Motorola Semiconductor Products, Inc., 1983.

4 J. Lohstroh, "Static and Dynamic Noise Margins in Logic Circuits," *IEEE J. of Solid State Circuits*, vol. 14, no. 3, pp. 591-598, June 1979.

5 S. I. Long, S. E. Butner, *Gallium Arsenide Digital Integrated Circuit Design*, New York: McGraw Hill, 1990.

6 J. Shiao, D. Nguyen, "Performance Modeling of a Cache System with Three Interconnect Technologies: Cyanate Ester PCB, Chip-on-Board and CU/PI MCM," *Proc. IEEE MCM Conf.* (Santa Cruz CA), pp. 134-137, March 1992.

7 R. E. Matick, *Transmission Lines for Digital and Communcations Networks*, New York: McGraw Hill, 1969.

8 E. H. Fooks, R. A. Zakarevicius, *Microwave Engineering Using Microstrip Circuits*, New York: Prentice Hall, 1990.

9 T. Itoh, *Planar Transmission Line Structures*, New York: IEEE Press, 1987.

10 L. Smith, "High Density Copper/Polyimide Interconnect," *3rd Internat. SAMPE Electr. Conf.*, pp. 939-947, 1989.

11 H. B. Bakoglu, *Circuits, Interconnections and Packaging for VLSI*, New York: Addision Wesley, 1990.

12 J. E. Buchanan, *BiCMOS/CMOS Systems Design*, New York: McGraw Hill, 1990.

13 E. Burton, "Transmission Line Methods Aid Memory-Board Design," *Electr. Design*, vol. 28, pp. 87-91, Dec. 1988.

14 M. S. Lin, A. H. Engvik, J. S. Loos, "Measurements of Transient Response on Lossy Microstrips with Small Dimensions," *IEEE Trans. Circuits and Systems*, vol. 37, no. 11, pp. 1383-1393, Nov. 1990.

15 L. T. Hwang, *et al.*, "Thin Film Pulse Propagation Analysis Using Frequency Techniques," *IEEE Trans. CHMT*, vol. 14, no. 1, pp. 192-198, March 1991

16 M. W. Hartnett, P. D. Franzon, M. B. Steer, E. J. Vardaman, "Worldwide Status and Trends in Multichip Module Packaging," Austin TX: Techsearch International, 1990.

17 H. You, M Soma, "Crosstalk Analysis of Interconnection Lines and Packages in High Speed Integrated Circuits," *IEEE Trans. Circuits and Systems*, vol. 37, no. 8, pp. 1019-1026, Aug. 1990.

18 T. Sons, Y. Wen, A Agrawal, "Electrical Considerations for Multichip Module Design," *Proc. ISHM Conf.*, (Orlando FL), pp. 287-291, 1991.

19 C. S. Chang, "Electrical Design of Signal Lines for Multilayer Printed Circuit Boards," *IBM J. of Res. and Devel.*, vol. 32, no. 5, pp. 647-657, Sept. 1988.

20 A. Deutsch, *et al.*, "Electrical Charactristrics of Lossy Transmission Lines for High Performance Computer Applications," *Internat. Symp. on Advances in*

Interconnections and Packaging, SPIE (Boston MA), vol. 1389, pp. 161-186, Nov. 1990.

21 R. T. Smith, H. Hashemi, "Electrical Analysis of TAB Tape in Bonding High I/O Chips to High Density Boards," *Proc. Internat. Electr. Packaging Conf.*, (San Diego CA), Nov. 1988.

22 G. A. Katopis, "Delta-I Noise Specification for a High Performance Computing Machine," *Proc. IEEE*, vol. 73, no. 9, pp. 1405-1415, Sept. 1985.

23 P. A. Sandborn, H. Hashemi, B. Weigler, "Switching Noise in a Medium Film Copper/Polyimide Multichip Module," *Internat. Symp. on Advances in Interconnection and Packaging*, SPIE, vol. 1389, pp. 177-186, Nov. 1990.

24 R. Senthinathan, J. L. Prince, "Simultaneous Switching Ground Noise Calculation for Packaged CMOS Devices," *IEEE JSSC*, vol. 26, no. 11, pp. 1724-1728, 1989.

25 R. Kaw, R. Liu, R. Crawford, "Effective Inductance for Switching Noise in Single Chip Packages," *Proc. Internat. Electr. Packaging Conf.*, (Marlborough MA) pp. 756-761, Sept. 1990.

26 H. Hashemi, *et al.*, "Analytical and Simulation Study of Switching Noise in CMOS Circuits," *Proc Internat. Electr. Packaging Conf.*, (Marlborough MA) pp. 762-774, Sept. 1990.

27 B. K. Gilbert, W. L. Walters, "Design Options for Digital Multichip Modules Operating at High System Clock Rates," *Proc Internat. Conf. on Multichip Modules*, ISHM, (Denver CO), pp. 167-173, 1992.

28 A. J. Rainal, "Computing Inductive Noise of Chip Packages," *AT&T Bell Laboratories Technical J.*, vol. 63, no. 1, pp. 177-195, Jan. 1984.

29 H. Hackenburg, "Signal Integrity in the VAX-8600 System," *Digital Technical Review*, vol. 1, no. 1, pp. 43-65, Aug. 1985.

30 National Semiconductor, *FAST Applications Handbook*, South Portland, MN: National Semiconductor Corporation, 1987.

31 P. D. Franzon, *et al.*, "Tools to Aid in Wiring Rule Generation for High Speed Interconnects," *Proc Design Automation Conf.*, IEEE and ACM, (Anaheim CA), pp. 466-471, June 1992.

32 T. Mikazuki, N. Matsui, "Statistical Design Techniques for High-Speed Circuit Boards," *Proc. IEEE CHMT '90 IEMT Symp.* (Baltimore MD), pp. 185-191,

33 M. S. Basel, M. B. Steer, P. D. Franzon, "Simulation of High Speed Digital Interconnection with Nonlinear Terminations," submitted for publication, available from author.

12

THERMAL DESIGN CONSIDERATIONS FOR MULTICHIP MODULE APPLICATIONS

Kaveh Azar

12.1 INTRODUCTION

Operational integrity and longevity of electronic components are directly affected by their operating temperature. Thus the object in thermal design is to control the temperature sufficiently to meet the reliability requirements of the system. Achieving this goal requires the cooperation of many engineers on a design team. It also requires close attention of the thermal engineer. However, the relative ease or difficulty of achieving thermal aims depends a lot on design decisions made by the systems engineer, the logic and circuit engineers and the layout engineer. They also need to understand thermal design principles and processes.

The thermal design of multichip modules (MCMs) is both more critical and more complex than for single chip modules (SCMs). It is more critical because a single MCM generates a lot more heat in a space comparable to a large SCM. Removing this additional heat requires larger cooling capacity. It is more complex for a number of reasons. First, an MCM contains several components whose temperatures must be controlled. Second, it contains several heat sources, all of which might dissipate different powers. Third, an MCM contains multiple materials and materials interfaces and is asymmetric. In comparison, a SCM usually contains only two materials (a metal and plastic or ceramic) and has several degrees of symmetry, both of which simplify modeling.

This chapter emphasizes a systems approach to module cooling and also modeling and evaluation techniques used in the design process. It starts with a description of the objectives in thermal design and a brief overview of the thermal design process. Following that, the thermal phenomena that a reader must understand in order to conduct design are discussed. This leads into a description of thermal management alternatives for MCMs. Finally, the analytical, computational and experimental methods used to conduct thermal design are discussed.

12.2 THERMAL MANAGEMENT

12.2.1 Objectives in Thermal Management

Temperature is the key player in reliable operations of MCMs and other electronic components. Though it is difficult to define the failure properties of a composite structure with one or a few parameters, transistor junction temperature reduction is viewed as the primary goal toward reliability enhancement. Having such a single goal simplifies the entire design process. Thermal management then embraces efforts to reduce or maintain junction temperature, T_j, within the design specifications. Although there is no industry set standard, T_j design limits vary from 80°C - 180°C. The industry norm for most components appears to be 125°C.

The reason for using a single temperature metric becomes more evident if we look at the equation used for reliability calculation of electronic components. Consider the ratio A_T of the time to failure at a temperature T_2 relative to that at T_1 [1]:

$$A_T = \exp\left[\frac{E_a}{k_B}\left(\frac{1}{T_1} - \frac{1}{T_2}\right)\right]. \qquad (12\text{-}1)$$

Here, E_a is a characteristic activation energy, and k_B is the Boltzmann constant. By reducing the operating temperature, T_2, the failure rate is reduced exponentially.

Also, as the temperature increases in the MCM the materials within it expand. Unfortunately, the rates of expansion of the different materials are often different and thermal stresses arise. If the thermally induced stresses exceed the elastic limit, the materials fail by coming apart. This might happen during manufacturing with a poorly designed process. If the composite structure experiences many temperature variations, such as power on, power off cycles, then fatigue field failure might eventually result.

12.2.2 Thermal Paths

Thermal design of MCMs, similarly to SCMs, can be divided into two areas: internal and external. For the sake of discussion, we refer to these as paths. The internal path is the one that directly deals with the structure of the component (module). Hence, it varies from component to component and differs significantly from the SCMs. In the external path, heat that has come to the component surface from the internal path is taken away either by gases or liquids. This part of the design does not have a strong dependency on the internal design of the MCM; it is more a function of the circuit board and system configuration.

Thermal design of MCMs, specifically with respect to their internal paths, typically takes two forms. These are experimental or computational simulations. The latter is the first chosen form for two reasons. First, it is much quicker to develop a model based on finite element or finite difference methods. Second, the computational models allow parametric simulations. This means that key parameters influencing design are varied to gauge their effects on junction temperature. The experimental simulation often follows the computational modeling. Some computational simulations of SCMs are described in references [2] - [4]; an example of the more involved modeling required for MCMs is given in [5].

12.3 THERMAL PHENOMENA IN ELECTRONIC ENCLOSURES

Thermal phenomena govern the removal of heat from components. A thermal process is defined as the merger of heat transfer and fluid flow to transport energy.

In this section, we begin by defining the basic principles of heat transfer. Then, heat transfer in electronic components is discussed. The concept of thermal resistance is presented with its uses. We also discuss why it should not be used to uniquely characterize SCM and MCM thermal properties. Since circuit boards contain MCMs and SCMs and play an important role in the thermal response of the MCM, thermal transports in circuit boards are discussed. The last two parts of this section discuss the thermal coupling (communication) between elements that form an electronic enclosure (system).

12.3.1 Heat Transfer Mechanisms

There are three modes of heat transfer: conduction, convection and radiation. Conduction heat transfer is when the heat is transferred by molecular vibration - solids or stagnant fluids. An example of a solid is the molding compound or the substrate in an MCM. An example of a stagnant fluid is the air trapped between the MCM and circuit board. The conduction heat transfer through a block of material is governed by the Fourier cooling law defined as:

$$Q = \frac{kA}{L}(T_h - T_c) \qquad (12\text{-}2)$$

where Q is the heat flow in units of power (watts) and k is the thermal conductivity of the material. The other terms are defined in Figure 12-1. T_h and T_c are the temperatures on opposite sides of the block. L is the length of the heat path and A is the cross sectional area. Table 12-1 shows the thermal conductivity of typical materials used in MCMs.

Convection heat transfer is when the transport of heat takes place by fluid motion. Three types of convection heat transfer are recognized: natural (free), forced and mixed. Natural convection occurs as a result of fluid (air) being in contact with a heated surface. The density of the fluid decreases causing it to rise, thus creating a natural circulation.

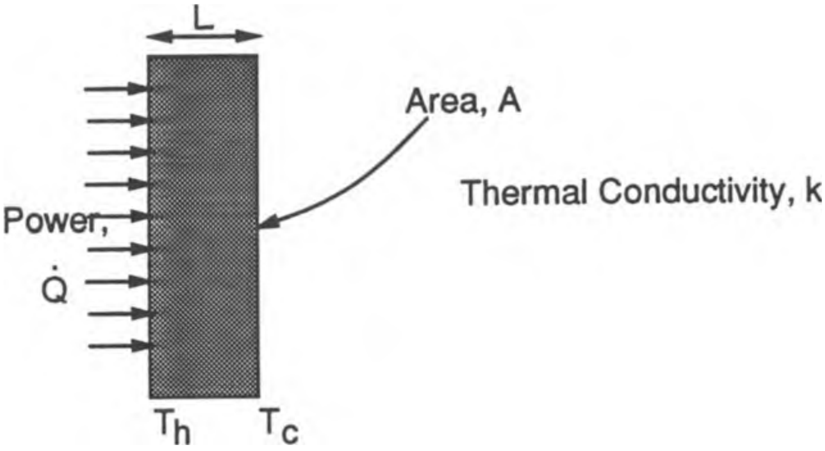


Figure 12-1 Conduction heat transfer in a solid.

Forced convection occurs when the fluid motion is induced by external sources. These include fans, pumps and blowers. Mixed convection occurs when natural and forced convection both are present. This is typically observed in low velocity flows at the presence of high powered components.

Convection heat transfer is governed by Newton's cooling law:

$$Q = hA(T_s - T_f), \tag{12-3}$$

where h is the heat transfer coefficient, A is the surface area exposed to the fluid, T_s is the temperature of that surface and T_f is the temperature of the fluid. The value of a specially constructed heatsink in thermal design is that it greatly increases the surface area.

Regardless of the type of convective heat transfer, Equation 12-3 is used for its solution. What sets the three types of convective heat transfer apart is h (heat transfer coefficient), which is obtained from empirical data. Figure 12-2 gives ranges of h encountered in the cooling of electronic components [6] - [10].

A word of caution seems merited at this point with respect to h and its correlations with fluid parameters. The data is typically reported in a form of correlation that relates the Nusselt number (Nu) to the Reynolds number (Re). When using a given correlation, its constraints must match the specific problem. Otherwise, that particular correlation is unsuitable for your analysis.

Radiative heat transfer occurs when heat is transported by photons or electromagnetic waves. What sets radiation heat transfer apart from conduction and convection is the medium for transport. Radiation heat transfer does not require a medium to transport energy. Radiation heat transfer is always present and its magnitude, similar to other modes of heat transfer, is a function of temperature difference. The belief that radiation heat transfer can be ignored is in error. Especially in natural convection problems, an excess of 20% of heat transfer is attributed to radiation.

Radiation heat transfer is governed by Equation 12-4,

$$Q = \sigma \varepsilon F_{hc} A (T_h^4 - T_c^4). \tag{12-4}$$

Here σ is the Stefan-Boltzmann constant, and ε is the emissivity of the radiating surface of area A. The temperatures of the (hot) radiating surface and of (cold) neighboring bodies are T_h and T_c, respectively, in degrees Kelvin. Values for F_{hc} (view or shape factor) are tabulated in heat transfer texts [11]-[12].

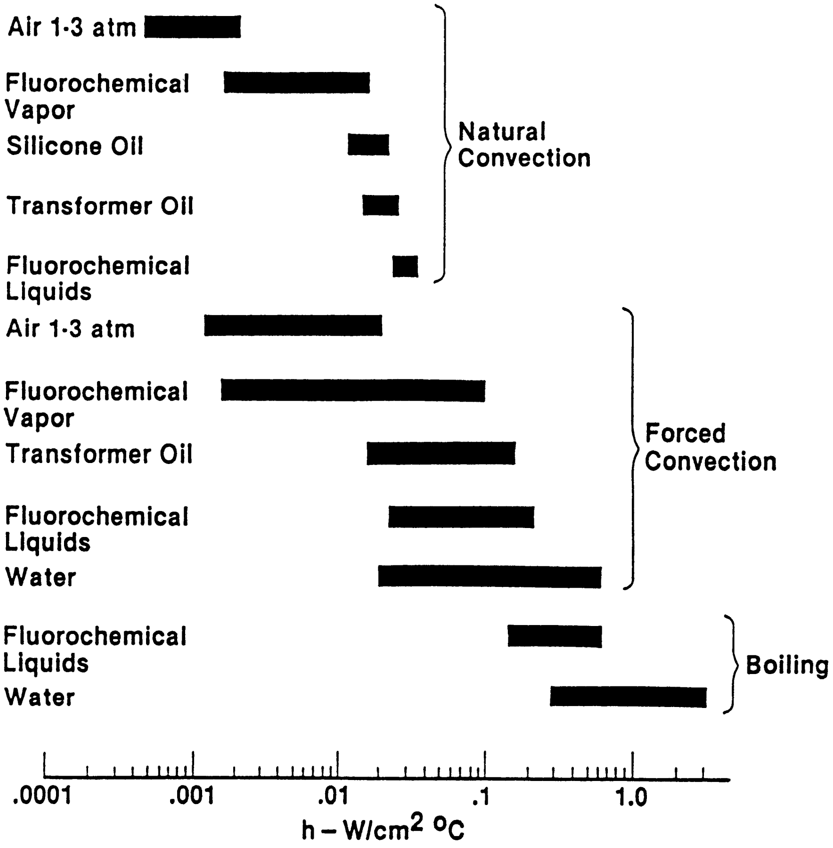


Figure 12-2 Range of heat transfer coefficients for various coolants.

12.3.2 Heat Transfer in Electronic Components (Modules)

This section discusses generation, spreading and eventual departure of heat from a component, in accordance with the three heat transfer mechanisms described above. Electronics components (modules) are made of an aggregate of materials with different physical geometries.

Consider a typical low performance plastic packaged MCM residing on circuit board in an air cooled system as shown in Figure 12-3. The electrical signals are brought to MCMs via the leads and then to the chips (typically) through the wire bonds. The leads are either press fitted, brazed or soldered to the substrate. The component is connected to the board through its leads either by surface mounting or through hole methods. This creates a package with