Early Analysis of Cost/Performance Trade-Offs in MCM Systems†

Vivek Garg, David E. Schimmel, Darrell Stogner, Craig Ulmer, D. Scott Wills, and Sudhakar Yalamanchili

Packaging Research Center
School of Electrical and Computer Engineering,
Georgia Institute of Technology, Atlanta, Georgia 30332-0250
Phone: (404) 894-2940, Fax: (404) 894-9959
e-mail: sudhakar.yalamanchili@ece.gatech.edu

Abstract

This paper explores early analysis of the complex relationships between system architectures and the active and packaging materials from which they are implemented. The goals of this analysis are to enable the designer to specify cost effective technologies for a particular system and to uncover resources which may be exploited to increase performance of such a system, early in the design process. We describe a prototype tool called IMPACT, which will predict cost, performance, power, and reliability, and demonstrate its use on several problems.

1.0 Introduction

The resource requirements of current generation integrated circuit technology are close to exceeding capabilities of traditional packaging techniques. Critical packaging resources include I/O bandwidth, off-chip signal transmission time, system footprint and mass. Multi-chip modules (MCM) utilize chip scale packaging (CSP) techniques to eliminate the intermediate level package, enabling direct placement of dice on a substrate, which contains the interconnections to realize circuit connectivity.

This paper explores the interaction between the resources provided by the packaging technology and the system architecture. It has been reported that decisions made very early in the design cycle have a significant impact on implementation and expenses incurred in the development of a product [10]. We introduce the concept of *early analysis* which allows a designer to assess the effects of technology-based decisions on the cost, performance, reliability, and power metrics of a system. Early analysis is essential to provide designers with the ability to very rapidly evaluate architectural alternatives without actually implementing the system. The architecture may be specified structurally at a very abstract level, and coupled with the desired technology to compute the various metrics which evaluate the design. The goal of early analysis is to help designers develop a *better* product in less time than required by design and evaluate methodology. Shorter design cycles expedite time to market, and improve competitiveness.

†. This research was supported by the National Science Foundation under grant EEC-9402723.

0-7803-3639-9/96 $4.00

2.0 Impact of Multi-Chip Modules (MCM) on System Design

The design of computing systems is fundamentally a process of optimizing system parameters given physical constraints. The advent of MCM technology is producing a large change in these physical constraints with consequent redefinition of the appropriate package and die boundaries. Our goal is the early analysis of the impact of these constraints so that the system may be designed to make the most effective use of MCM technology. The main effects of using MCM technology in the design of system architectures are the following.

2.1 Performance

The elimination of intermediate level packaging implies that the chip is directly attached to the MCM substrate using either wirebond or controlled collapsible chip connection (C4). Modern architectures are limited by off-chip delays. The parasitics associated with an off-chip interconnect in a traditional package typically include the resistance, capacitance, and inductance of the wirebond used to attach the die to the package, and the resistance, capacitance, and inductance of the brazed pins of the through hole or surface mount package. These parasitics are several times larger than those encountered in direct chip attach technologies. C4 parasitics are an order of magnitude less than those associated with through hole and surface mount packages [14]. This reduction in parasitics encountered by the signals going off-chip improves the signal transmission speeds. Also the conductors on the substrate behave like transmission lines resulting in faster signal transmission times than on-chip interconnects, further contributing to the performance of MCM based systems.

Incorporating the components of a printed circuit board onto an MCM substrate typically results in a reduction in footprint. This helps reduce the interconnect lengths between components, reducing the transmission time between them. Furthermore, for some of the MCM technologies the dielectric constant of the substrate is much smaller than FR4 or other materials used in the printed circuit board technology, resulting in faster propagation speeds. As monolithic designs become large, the on-chip aluminum interconnect delays become larger than off-chip interconnect delays on the substrate, providing further reasons to move to MCM based designs. Figure 1 provides an example comparison of off-chip vs. on-chip delay for an MCM-D (MMS D-500) process. Details can be found in [6].

2.2 Cost

While MCM packaging has proved to be useful, currently it is limited to high performance applications, where cost is not the primary issue. Since this technology is used only in "niche" applications, it is not manufactured in high volume, and as a result is not economically competitive with PCB technology. The use of MCMs in automotive applications has proven that when manufactured in high volume, this technology can indeed be cost-effective. Other factors adding to the cost of MCMs are testing of unpackaged (bare) dice to ensure correct functionality, also known as the known good die (KGD) issue, and signal redistribution on existing dice designed for peripheral I/O to enable area array bonding. Most of these costs are related to the fact that MCMs are not widely used in applications, and as a result the process and equipment involved in its manufacturing is not cost competitive with other more mature technologies. Active research involving both industry and academia is focusing on the goal of making MCM technologies price competitive with printed circuit board technology. Several low-cost processes have been introduced and others are being proposed which may lead to a reduction of substrate costs by a factor of 5 and 10 respectively [5,11,13,15].

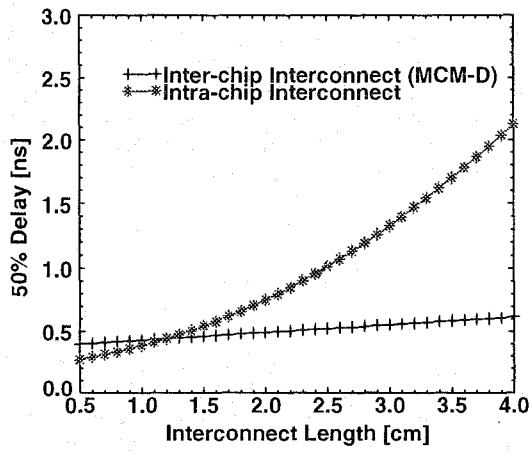

Figure 1. Comparison of intra-chip and inter-chip interconnect delays.

Lower off-chip delays incurred in MCM technology encourages the partitioning of large monolithic dice into smaller ones. Smaller die sizes result in higher yields, and as a result decrease the cost of dice. This decrease in cost can be used to offset some of the costs incurred in the use of MCM packaging. Ultimately we expect the MCM process to mature to a point where it is price-competitive with the PCB technology. At that time partitioning of the dice will reduce the cost of the system, possibly without appreciably affecting the performance.

2.3 Reliability

System reliability depends directly upon the reliability of its individual components. As a result, the die manufacturers must provide better test coverage and delivered die yield for bare dice than for packaged die, since system reliability is now the product of the probabilities of all the chips on the MCM being good. It has been reported that the reliability of C4 die attach mechanism is 0.5 ppm which is six times more reliable than wirebond (3 ppm) [4]. Consequently, use of C4 type die attach will increase the reliability of a given system. Furthermore, elimination of the intermediate packaging removes reliability concerns related to these components. However, this elimination leads to other problems involving thermal (CTE) mismatches between the silicon dice and the MCM substrate. Epoxy encapsulants are often used along with C4 connections to minimize thermal mismatch problems.

2.4 Power

The lower parasitics of the C4 or wirebond connections, as compared to brazed pins, imply that to maintain the same performance the signal drivers can be smaller hence reducing the power consumption. Concentration of components closer together on the MCM substrate may give rise to more challenging thermal management problems, but depending on the substrate technology and cooling strategy chosen, it may not be an issue. For example, MCM-C substrates are typically better conductors of heat than printed circuit boards, and may compensate for the increased heat flow.

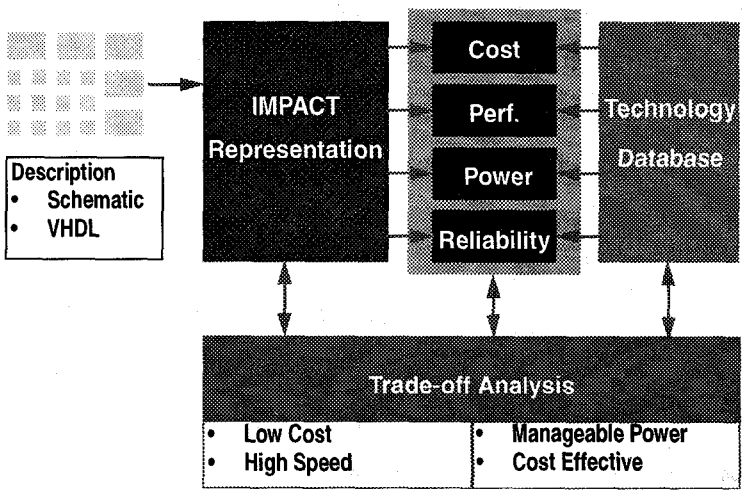

Figure 2. Organization of IMPACT tools.

2.5 Ergonomics

The largest driving force for MCMs next to performance is perhaps volume and mass. With the recent explosion in portable consumer electronics such as mobile telephones, personal digital assistants, laptop computers, etc., the industry is constantly striving for smaller and lighter products. An MCM package involves bare dice, discrete components, and a substrate, which houses the wiring required to connect all the circuits to be placed on the package. As a result, the MCM package can incorporate the functionality of a printed circuit board (PCB) by replacing the board with the substrate, and by attaching bare dice and discrete components directly on the substrate, hence reducing the overall size and weight of the system. Due to the elimination of the intermediate level packages, the effective usage of the MCM substrate increases, since a die typically comprises of only 20% of the package area. This implies that larger printed circuit boards can be reduced to small MCM substrates, resulting in significant reduction in system footprints.

3.0 IMPACT Methodology

There is a major research effort in progress at the Packaging Research Center (PRC), an NSF sponsored Engineering Research Center, at Georgia Tech. The center is multi-disciplinary with the common goal of developing economically viable MCM technology for consumer applications. As part of this effort, we are developing the IMPACT tools to help designers perform early analysis of the impact of MCM packaging on system architectures. These tools are based on a set of core models that take technology parameters for substrate, die, packaging, and assembly, and system specifications as inputs. The architecture is evaluated based on cost, performance, reliability, and power metrics computed using IMPACT. The accuracy of these models directly affects the predictions of the tool. At present the models that are being used are from the published literature. In the future, we will incorporate models from the Packaging Research Center's

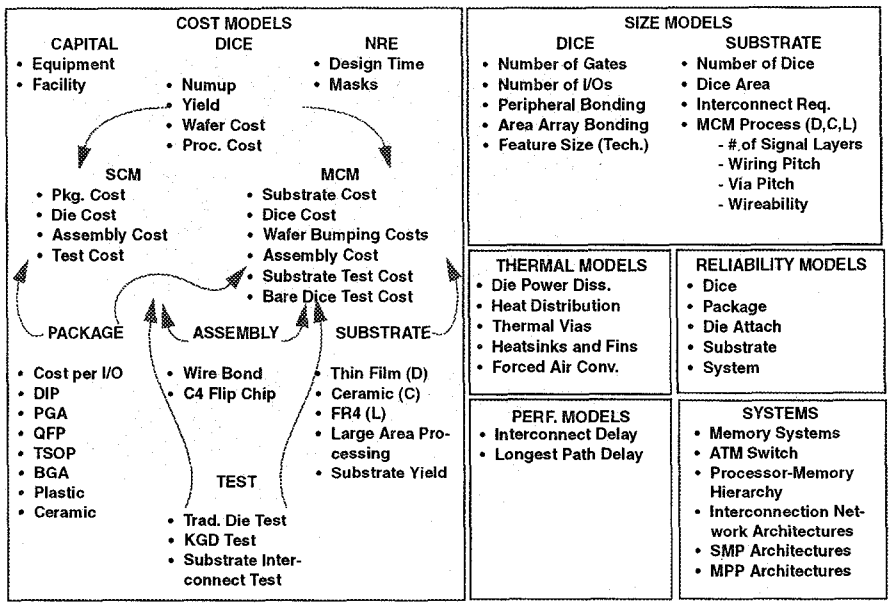

Figure 3. Model Flow in IMPACT Tools.

manufacturing process as they become available. The models are fully interchangeable, so that designers may use technologies and processes of choice to evaluate their designs.

The core set of models include cost models for die, substrate, assembly, and test. Models for on-chip and off-chip interconnects provide signal transmission delay estimates, and in turn performance metrics. Reliability and power dissipation models are under development. Figure 2 shows the organization of the IMPACT tools. The design may be entered as a schematic or in VHDL. In either case a structural description of the design is extracted to perform the analysis. It should be noted that the design is specified at a very abstract level in terms of computational units, memory units, and information channels. Each of these components have some basic properties associated with them to describe their physical features and functional specifications. Currently the designer is expected to provide the total number of functional units in the system, identify them as memory or logic die, and assign them to partitions, where each partition represents one MCM. The connectivity between these functional blocks is represented via information channels, which are simply the rate of information flow between the corresponding blocks. For example, consider a parallel computing system consisting of 16 processors, and corresponding memory. A partitioning strategy is to assign 4 processor-memory pairs to each partition yielding a total of 4 MCMs. Once the partitioning has been identified, the cost and performance metrics for the system as defined can be calculated. Other configurations of the system may be evaluated to achieve the desired specifications. An iterative process may be used to identify the appropriate technology - MCM-L,C, or D, C4 or wirebond, stacked die, etc. - to satisfy design specifications. The trade-offs to be considered are cost-effectiveness, performance, price/performance, reliability, and thermal management. The flow of models implemented in IMPACT is illustrated in Figure 3.

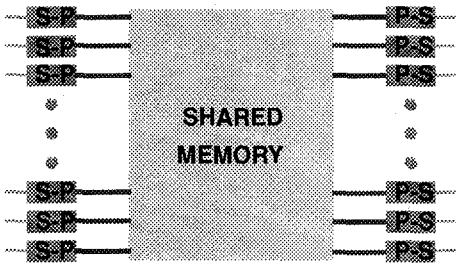

Figure 4. Shared Memory ATM switch architecture.

4.0 Example Systems and Analyses

There are several useful design issues that can be addressed at an early stage. To answer these questions effectively we need information about the IC and MCM technologies, and specifications of the system architecture. An iterative method may be used to arrive at a solution that is technologically and architecturally viable to meet the specified performance of the system. The approach is illustrated with the following examples.

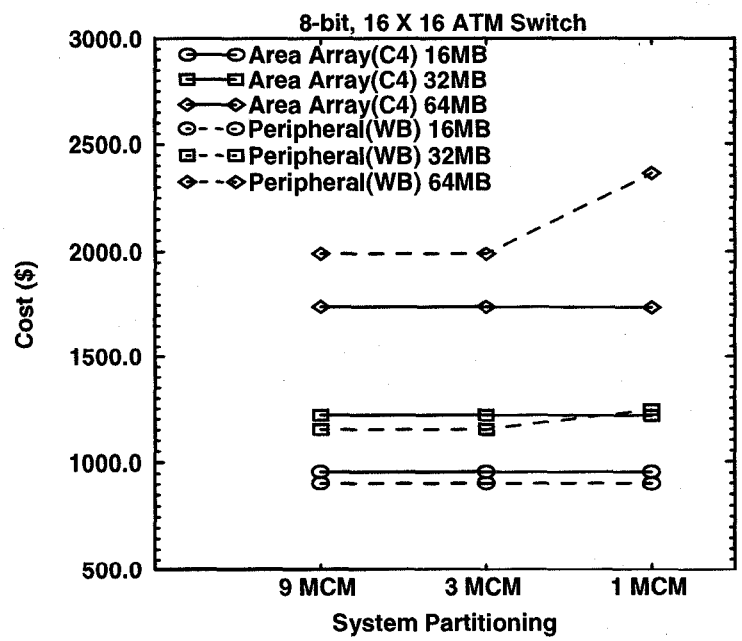

Figure 5. Cost analysis for 8-bit ATM switch for various memory configurations and system partitionings.

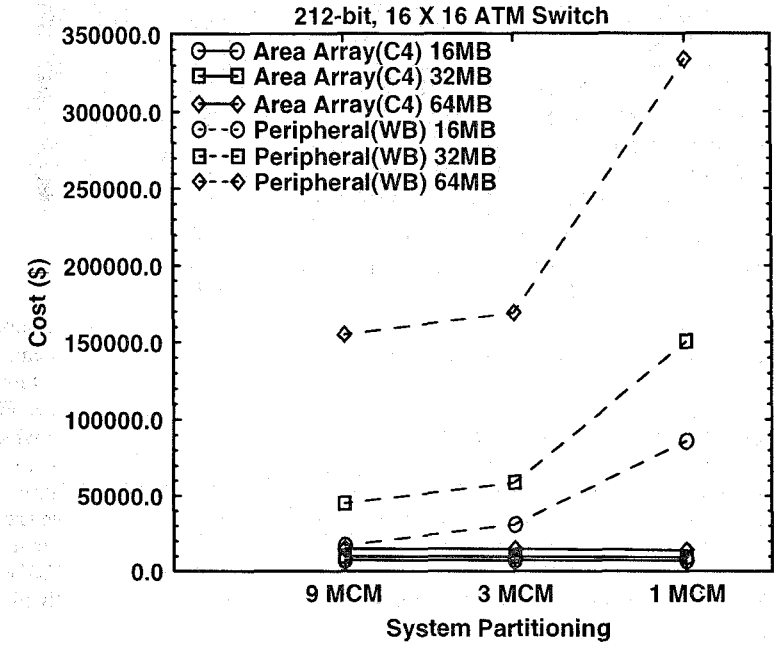

Figure 6. Cost analysis for 212-bit ATM switch for various memory configurations and system partitionings.

4.1 ATM Switch Architecture

Consider an analysis of the effect of packaging on a shared memory ATM switch design shown in Figure 4. The analysis performed here covers three different partitioning strategies. We can package all the serial to parallel (S-P) and parallel to serial (P-S) converters and memory on one MCM, or we can package the S-P converters on one MCM, memory on another MCM and finally P-S converters on a separate MCM, yielding a 3 MCM system. Another alternative is to use 4 MCMs each to package the S-P and P-S converters, and to use one MCM for the memory, yielding a 9 MCM design. The cost analysis for the use of 8-bit and 212-bit internal buses between the converters and shared memory is shown in Figure 5 and Figure 6 respectively. For the 8-bit case it is clear that switches with all three memory configurations are feasible, but that area array bonding may provide a more cost-effective solution in some cases. For the 212-bit case it is apparent that peripheral bonding is simply infeasible to use to cost-effectively realize these switches. We also note that for both the 8-bit and 212-bit buses, the system cost changes for the different partitionings for the peripheral bonding, while it stays almost constant for the area array bonding case. This leads us to the conclusion that the peripheral bonded designs are I/O limited, and as a result larger dice, and hence larger MCM substrates are required to implement the ATM switches.

4.2 SIMPil: A SIMD Pixel Array Processor

SIMPil is a single instruction stream, multiple data stream (SIMD) array processor designed to be used for embedded image processing and computer vision applications [2,3]. The architecture consists of several thousand processing elements (PEs) interconnected in a 2-dimen-

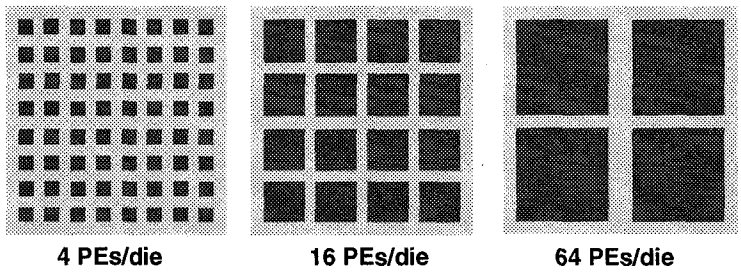

Figure 7. Various die partitionings for a single MCM 256-node system.

sional array topology. We present an example analysis to answer the question, where should the die boundaries be placed for a single MCM design. This is a trade-off between the number of I/Os and the chip size, and the MCM cost is used for evaluation. We also explore the cost impact of using distinct MCM fabrication technologies, and various semiconductor technologies. Figure 7 illustrates the 3 configurations that are considered. At one extreme we have 4 PEs/die which provides high yield due to smaller die area, but also restricts the number of I/Os, and since a large number of dice need to be assembled onto the MCM, the overall system reliability may be adversely affected. At the other extreme, fabricating 64 PEs on a die increases the die size rather drastically which in turn increases the MCM substrate area as well. As a result the system cost for either of these cases is quite large. An economically superior strategy uses 16 PEs/die which yields better system cost than the other two partitionings. These results are graphically illustrated in Figure 8.

Figure 9 shows the cost of the same system taking into account projection of semiconductor evolution. It indicates that in the years 1995-1998, when the integration levels are relatively mod-

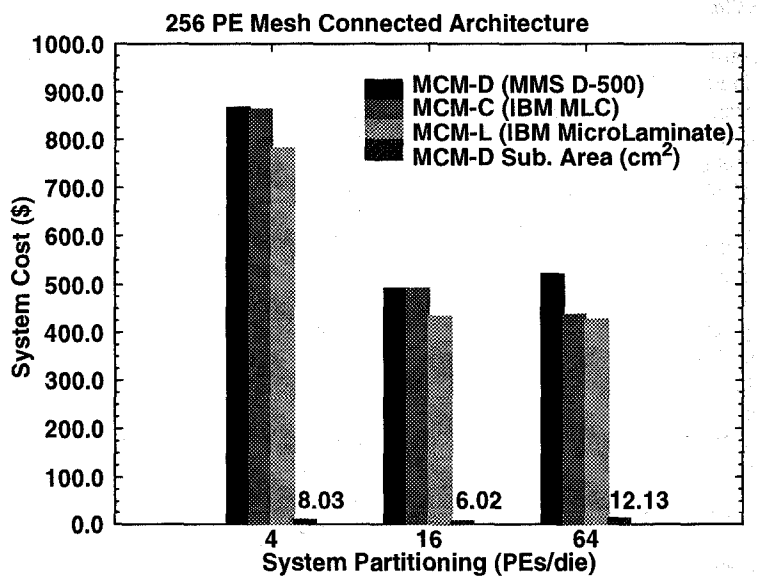

Figure 8. MCM-D,C,L costs for various system partitionings.

erate, the strategy using 16 PEs/die is economically better than the other two. But as the levels of integration improve over the years, i.e. as the defect density decreases, the effect of die area on the yield and hence cost of the die is minimized, and as a result systems with a higher degree of integration become economical once again.

This cost analysis we have performed does not take into account the NRE costs associated with the design and fabrication process. The system cost is the sum of the cost of bare dice, calculated based on die yield models and number of dice fabricated on a wafer; cost of the substrate, cost of C4 die attach, and cost of testing dice and substrate. The substrate area calculations were obtained from the maximum of the surface area required to accommodate the dice, and wireability analysis using Bakoglu's extension of the Donath model [1]. All die process related parameters were obtained from [12], and the MCM process related parameters were obtained from the available commercial process specifications. The models and parameters used to analyze the design can be found in [7].

5.0 Conclusions

The goal of this work is an understanding of the impact of MCM packaging technology on system design. Preliminary results suggest that MCM technology can be exploited to realize a new class of cost effective system designs. We have developed a suite of tools called IMPACT to help designers assess the effects of packaging on system architecture and design. The goal of IMPACT is to provide decision support for designers very early in the design cycle. Having the ability to predict the effects of packaging on system design early in the cycle can help shorten design cycles leading to higher profitability. Early decision support mechanisms also promote cost-effective use of packaging technologies, and provides the designers with a venue to evaluate alternate architectures. As packaging technologies advance, the traditional limitations such as

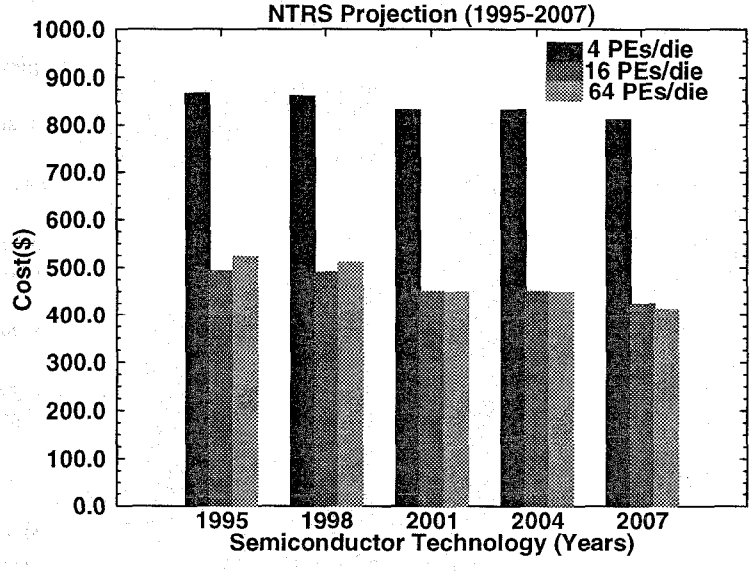

Figure 9. MCM-D cost for various system partitionings, assuming fixed size die.

limited I/O pads and slower off-chip bandwidth, etc. are no longer applicable. As a result, traditional architectural styles may be altered to realize more cost-effective designs which provide better or comparable performance.

As MCM technologies advance and mature, they will become an increasingly viable option for a wide range of applications. Our objective is to facilitate this process with early analysis tools that can reliably predict the impact of packaging options on system level metrics.

6.0 References

[1] Bakoglu, H.B., *Circuits, Interconnections, and Packaging for VLSI*, Addison Wesley Publishing Company, 1990.

[2] H. H. Cat., J. C. Eble, D. S. Wills, V. K. De, M. Brooke, N. M. Jokerst, "Low Power Opportunities for a SIMD VLSI Architecture Incorporating Integrated Optoelectronic Devices," *GOMAC'96 Digest of Papers*, pp. 59-62, Orlando, FL, March 1996.

[3] H. H. Cat, M. Lee, B. Buchanan, D. S. Wills, M. A. Brooke, N. M. Jokerst, "Silicon VLSI Processing Architectures Incorporating Integrated Optoelectronic Devices," *Proceedings of the 16th Conference on Advanced Research in VLSI*, pp. 17-27, Chapel Hill, NC, March 1995.

[4] Elenius, P., "Why Flip Chips," Seminar Presentation, Kulicke & Soffa Industries, Jan. 31, 1996.

[5] Frye, D.C., Skinner, M.P., and Heistand II, R.H., "Cost Implications of Large Area MCM Processing," *Proceedings of the 1994 International Conference on Multichip Modules*, pp. 69-80, Denver, Colorado, April 13-15, 1994.

[6] Garg, V., et al., "Incorporating Packaging Constraints into System Design," *Proceedings of the European Design and Test Conference*, pp. 508-13, Paris, France, March 11-14, 1996.

[7] Garg, V., et al., "IMPACT Models and Parameters," Technical Report: Work in Progress, School of ECE, Georgia Tech, 1996.

[8] Ginsberg, G.L. and Schnorr, D.P., *Multichip Modules and Related Technologies: MCM, TAB, and COB Design*, McGraw-Hill, Inc., New York, 1994.

[9] Lau, J.H., Ed., *Chip On Board: Technologies for Multichip Modules*, Van Nostrand Reinhold, New York, 1994.

[10] Sandborn, P.A. and Moreno, H., *Conceptual Design of Multichip Modules and Systems*, Kluwer Academic Publishers, 1994.

[11] Schreiber, C.M., "MCM Substrates Fabricated Using Flexible Polyimide Film," *Proceedings of 1995 International Conference on Multichip Modules*, pp. 567-72, Denver, Colorado, April 19-21, 1995.

[12] Semiconductor Industry Association, *The National Technology Roadmap for Semiconductors*, 1994.

[13] Trask, P.A. and Pillai, V.A., "Large Format MCM-D Processing," *Proceedings of the 1994 International Conference on Multichip Modules*, pp. 94-99, Denver, Colorado, April 13-15, 1994.

[14] Tummala, R.R. and Rymaszewski, E.J., Eds., *Microelectronics Packaging Handbook*, Van Nostrand Reinhold, New York, 1989.

[15] White, G.E., Perfecto, E., DeMercurio, T., McHerron, D., Redmond, T., and Norcott, M., "Large Format Fabrication - A Practical Approach to Low Cost MCM-D," *Proceedings of the 1994 International Conference on Multichip Modules*, pp. 86-93, Denver, Colorado, April 13-15, 1994.