Annual Progress Report for the Year ending 28 Feb 1996
and
Request for Continued Funding

M. J. Flynn, P.I.

Grant NSF MIP-9313701

1. Brief Summary of Progress

Our studies continue in several areas, described below.

Algorithms (Professor M. Flynn)

In this past year, several aspects of floating-point division have been investigated. First, two techniques were proposed for reducing the average latency of division operations. It has been determined that many scientific applications contain recurring or redundant computation, where the same division computation is performed on multiple occasions. This behavior can be exploited using division and reciprocal caches which store frequently used division results. For multiplication-based division implementations, the reciprocal can be reused rather than the quotient, increasing the likelihood of the computation being redundant. Additionally, due to the similarity between division and square root computation, square-root operations can share a reciprocal cache. By checking first in the appropriate cache at the beginning of such an operation and using an existing result when available, the overall division performance can be increased.

Another study in floating-point division investigated techniques to minimize the complexity of SRT division tables. The analysis derived the allowable divisor and partial remainder truncations for radix 2 through radix 32, and it quantified the relationship between table parameters and the number of product terms in the logic equations defining the tables. By mapping the tables to a library of standard-cells, true delay and area values were measured. The results showed that Gray-coding of the quotient-digits allows for the automatic minimization of the quotient-digit selection logic equations. Second, using a short carry-assimilating adder with a few more input bits than output bits can reduce table complexity. Third, reducing the number of bits in the partial remainder estimate and increasing the length of the divisor estimate increases the size and delay of the table, offsetting any performance gain due to the shorter external adder. Finally, while delay increases nearly linearly with radix, area increases quadratically, limiting practical SRT table implementations to radix 2 and radix 4.

With the added integration offered by technology scaling, microprocessor designers have gone from software emulation of floating operations, to dedicated FP chips (e.g. 80387), to on-chip FPUs, and finally to multiple FPUs on a chip. Meanwhile, the latencies of most FP operations have gone from hundreds of cycles to two or three cycles. The basis of these design efforts is the fundamental need for fast FP executions. Despite these efforts, allocation of die area to FPUs remains an art based on engineering intuition and past experience. We develop the Floating Point Unit Cost Performance Analysis Metric (FUPA) to allow quantitative tradeoffs between performance and cost. The FUPA metric incorporates five key aspects of VLSI systems design: latency, die area, power, minimum feature size and profile of applications. FUPA utilizes technology projections based on scalable device models in order to identify the design/technology compatibility and allows FPU designers to make high level tradeoffs in optimizing designs.

Publications

Steve T. Fu, Nhon Quach and Michael J. Flynn. Architecture Evaluator's Workbench and its Application to Microprocessor Floating Point Units. Technical Report No. CSL-TR-95-668, Computer Science Laboratory, Stanford University, June 1995.

Steve T. Fu, Nhon Quach and Michael J. Flynn,. FUPA: Floating Point Unit Cost Performance Metric and its Application to Microprocessors. Submitted to *IEEE Transactions on VLSI Systems*, 1995.

Steve T. Fu and Michael J. Flynn. CacheOpt–A High Level Synthesis Tool For On-Chip Cache Hierarchy Synthesis. Submitted to the *33rd Design Automation Conference*, 1996.

Steve T. Fu and Michael J. Flynn. Optimal On-Chip Cache Hierarchy Synthesis with Scaling of Technology. To be presented at the *15th Annual IEEE International Phoenix Conference on Computers and Communications*, March 1996.

Stuart F. Oberman and Michael J. Flynn. "Design Issues in Division and Other Floating-Point Operations." To appear in *IEEE Transactions on Computers*.

Stuart F. Oberman and Michael J. Flynn. "Measuring the Complexity of SRT Tables." Technical Report No. CSL-TR-95-679, Computer Systems Laboratory, Stanford University, November 1995.

Stuart F. Oberman and Michael J. Flynn. "Minimizing the Complexity of SRT Tables." Submitted to *IEEE Transactions on VLSI Systems*, 1995.

Stuart F. Oberman and Michael J. Flynn. "Implementing Division and Other Floating-Point Operations: A System Perspective." In Proceedings of *SCAN-95* (International Symposium on Scientific Computing, Computer Arithmetic, and Validated Numerics), Wuppertal, Germany, September 1995.

Stuart F. Oberman and Michael J. Flynn. "Reducing Division Latency with Reciprocal Caches." In *Proceedings of SCAN-95* (International Symposium on Scientific Computing, Computer Arithmetic, and Validated Numerics), Wuppertal, Germany, September 1995.

Stuart F. Oberman and Michael J. Flynn. "An Analysis of Division Algorithms and Implementations." Technical Report No. CSL-TR-95-675, Computer Systems Laboratory, Stanford University, July 1995.

Stuart F. Oberman and Michael J. Flynn. "On Division and Reciprocal Caches." Technical Report No. CSL-TR-95-666, Computer Systems Laboratory, Stanford University, April 1995.

CAD tools (Professor: G. De Micheli)

A) Wave pipeline with CMOS (with F. Klass – Delft)

We have investigated the applicability of wave pipeline techniques to CMOS technology. Klass has shown that wave pipelining can be used in CMOS technology where logic gates have pattern-dependent delays, provided that such delays are well-characterized and that path-delays are computed while taking into account possible variations of gate delays. Techniques for wave pipelining in CMOS are described in Klass' thesis where he reports also on the successful design of a demonstration chip.

B) Retiming an pipelining with uncertain delays (with I. Karkowski – Delft)

Retiming synchronous circuits means moving latch boundaries to reduce the cycle time. Retiming can be used to achieve effective pipelined implementations. We consider here retiming for arithmetic circuits, where delays of basic constituents are uncertain, i.e. they vary within a range. Whereas retiming with fixed, known delays has been solved by Leiserson et al., retiming with uncertainb delays is a novel and challenging problems. We investigated the use of possibilistic programming techniques, that use a triangular delay distribution, and an integer linear programming formulation that allows us to restructure circuits while maximizing the probability of achieving maximum performance for a given circuit technology characterized by delay parameters.

C) Power-performance tradeoff

Reducing power dissipation, while preserving performance levels, is required by most design of portable communication and computation devices. Power Dissipation can be reduced by avoiding glitches, that are often cause Of wasted power in arithmetic circuits. Whereas hazard (glitch) removal is essential in asynchronous circuit design, it is desirable also for synchronous design. We have investigated analysis methods to detect hazards by enhanced logic simulation, that yields power measurements comparable to SPICE in accuracy with limited computation effort. We are currently investigating glitch removal algorithms, that preserve Both performance and testability features of electronic circuits.

Circuits and Packaging (Professor R. F. Pease)

Reversible Pressure Contacts for High Speed Testing to Assure Known Good Die (Student Ming Zhang)

One of the key barriers to implementing system level packaging is the challenge of assuring the quality of the I.C. chips prior to committing them to a system; this is well known and is referred to as the 'known good die' issue. The problem arises from the difficulty of testing the die at speed under realistic conditions prior to packaging. One approach is the active array membrane probe card in which probes are mounted on a membrane and the test circuitry is integrated into the silicon frame of the membrane. This approach, still in the research phase, has difficulties with assuring good contact between the probe and the contact pad of the I.C. under test and in building the test circuitry.

An alternative approach arises from the recent observation that solder balls, similar to those used in making permanent contacts between die and substrate (the IBM 'C4' process) can make surprisingly good pressure contact to aluminum contact pads and the contact can be re-used by simple separation and thermally reforming the solder into the ball shape (1). This approach may well overcome the difficulties of the membrane probe card approach.

We are researching both the scientific and technological aspect of this phenomenon. On the scientific side the fact that the native aluminum oxide can be so easily penetrated is remarkable and may be associated with the nature of the oxides of the solder material. A variety of novel microscopic analyses is being applied to understand this phenomenon. On the technological side the condition for achieving reliable contact and re-use are to be the subject of the research following the scientific phase. The high frequency electrical performance of the contacts will be the subject of the third phase of the contacts.

Publications

D. B. Tuckerman, B. Jarvis, Chang-Ming Lin, P. Patel, and M. Hunt. "A cost-effective wafer-level burn-in technology." In *Proceedings of 3rd International Conference on Multichip Modules* (SPIE Vol. 2256). Denver, CO, USA, 13-15 April 1994. (USA: ISHM-Microelectron. Soc. 1994, p. 34–40).

J. B. P. Williamson. "The Microworld of the Contact Spot." *Electrical Contacts* 1981, IIT (1981), p. 1.

Membrane MCM Technology in Silicon (Professor S. Wong)

A membrane MCM [1] is made by attaching chips to a membrane which is fabricated on a silicon substrate with conventional IC processing techniques. This allows the membrane MCM to take advantage of a large, existing technology base. The chip to chip connections are performed by conducting wires that run on the substrate and the polyimide membrane. Multiple levels of wires can be fabricated to increase the flexibility of routing and to reduce wiring parasitics. The use of silicon as the module substrate allows active components to be built in the substrate if desired.

Process Flow

The key steps for fabricating a membrane multi-chip module are described [1]. Silicon nitride is first deposited on both sides of the wafer. The silicon nitride is patterned on the backside and serves as a mask for a timed KOH etch. About 30 μm of silicon is left after the etch and acts as a mechanical support for subsequent process steps. 2.5 μm low temperature oxide (LTO) is deposited on the front side and patterned. A reactive ion etch is performed on the LTO to form a grid pattern and about 1 μm layer of LTO is left as an etch stop for later step. Polyimide is spun and cured on the front side. Then aluminum is sputter deposited, patterned, and etched. Multiple layers of polyimide and aluminum are deposited to fabricate multiple levels of interconnections. After the final polyimide is cured, LTO is deposited on the top to serve as a protection layer. On the backside of the wafer, the silicon layer is removed by a highly selective dry etch, which stops on the LTO layer. The membrane is finally released by immersing the wafer in HF to remove the exposed LTO. The grid pattern built into the LTO layer is replicated at the bottom of the released membrane.

To attach a chip to the membrane, a layer of polyimide is spun on the chip as an adhesive. After a 5 minute bake at 90°C to evaporate most of the solvent, the chip is then visually aligned to the membrane, attached and bonded at 120 °C. To avoid the formation of air bubbles and ensure perfect contact between the chips and the membrane, the attachment is performed under vacuum, with differential air pressure applied to the membrane in a customized apparatus [1]. After all the chips are attached, the entire MCM is cured at 350°C. The grid pattern in the polyimide membrane provides escaping channels for the gases generated during the curing process and avoids the formation of air bubbles.

In order to electrically connect the conducting wires on the chips and the membrane, contact vias are patterned and formed by an oxygen RIE through polyimide down to the aluminum pads on the chips. Holes in the wires in the membrane allow the etch to proceed down to the pads on the chips. Aluminum/TiW layers are then sputter deposited, patterned and etched to cover the vias, and provide electrical connections.

Contact Chain Experimental Results

Contact chains have been used to evaluate the electrical connections between the chip and multiple layers of wiring in the MCM substrate. The chips are 0.9 X 0.9 cm^2 and the membrane openings are 1 X 1 cm^2. The contact chains are distributed throughout the chip. Continuity of chains with 400 contacts has been confirmed. Two different sizes of vias, 10 X 10 μm^2 and 20 X 20 μm^2, have been studied. The contact resistance is 0.060 to 0.080, and 0.024 to 0.040 =/contact, respectively. The contact resistance could be further decreased by filling the vias with a CVD metal, which offers better step coverage than sputtered metal. However, the high stress of CVD metal may cause other problems.

Interface Circuits

MCM provides additional flexibility in system partitioning and circuit design. It enables the incorporation of devices/circuits from different technologies and materials [2-4]. For example, low voltage swing interconnections can be implemented by buffering CMOS chips with bipolar driver and receiver circuits. These bipolar circuits can be built either in the silicon module substrate or in another chip. In order to accurately determine the signal delay, various experimental interface circuit chains have been implemented to account for the delay caused by the measurement probes and cables. The circuits were fabricated with a 2 μm BiCMOS technology. Only MOS devices are used in one of the chips, and bipolar transistors are fabricated in the other. The chips are connected together by membrane MCM Technology. From delay measurements performed on the composite chains, the corresponding delay of each circuit is calculated. The fastest chip-to-chip delay achieved is 5.6 nsec, which is over 40% faster than the delay associated with wire bonded systems [4]. The improvement of the signal delay in membrane MCM is due to the reduction in size of the contact pads and the associated parasitic capacitances. Further reduction in delay is expected with contemporary sub-micron devices.

Power Supply Noise

As the number of input and output buffers increases, the noise associated with simultaneously switching these buffers becomes more severe. The system performance and reliability will be compromised. Integrated decoupling capacitors could be used to reduce the switching noise in the power and ground lines. For Si-based MCMs, bipolar transistors can also be built into the substrate to isolate the supply voltages to each individual chip, and hence reduce the coupling of the switching noise between chips.

To evaluate the effectiveness of the integrated decoupling capacitors and the regulating bipolar transistors in reducing switching noise, the on-chip voltages of a noisy chip and a quiet chip are measured. The noisy chip contains four 41-stage NMOS ring oscillators to generate a large amount of noise. The quiet chip only has four 3-stage inverter buffers. The use of 1.2 nF decoupling capacitors decreases the noise on the noisy chip by a factor of 1.5, but does not improve that on the quiet chip, because the coupling remains between these two chips. With bipolar transistors as series regulators, the peak-to-peak noise voltages of the quiet and the noisy chips are decreased by a factor of 1.8 and 1.5, respectively. The regulating bipolar transistors are connected with a emitter follower connection to isolate the chip supply voltage V_{DD} from the system supply voltage V_{CC}. A reference voltage, $V_{REG} = V_{DD} + V_{BE}$, is distributed to the base of the regulating bipolar transistors. With both the capacitors and bipolar transistors, the peak-to-peak noise voltages are reduced by a factor of about 3.2 for both chips.

Thermal Management

Thermal management is an important consideration in MCM. Since high power is to be dissipated over the small area of a module, the heat must be conducted away efficiently. For membrane MCM technology, both the silicon substrate and the attached chips can be placed on a heat sink. To further increase the efficiency of the heat removal, water can be sealed between the substrate and the chips. The water provides an additional heat removal path to the heat sink, through either the heat sink or the module substrate. For switching operations, which generate heat pulses, the water can increase the thermal capacity and minimize the fluctuation of temperature.

In order to evaluate the heat removal capability of the membrane MCM, a chip which contains ECL circuits as a heat source and diodes as a temperature-sensing device is designed. The heat source consists of 60 identical circuits. By varying the supply voltages, power dissipation up to 4.4 W can be produced. Three diodes are located at different positions on the chip to detect the temperature variation. Two are in the proximity of the heat source, while the third one is 0.2 cm away from it. The I-V characteristic of a diode allows the temperature to be determined.

In our experiment, modules with and without water have been compared. By adding water between the chips and the substrate, the temperature near the heat source is decreased from 135 °C to 100 °C with a power dissipation of 4.4W. The thermal resistance of the system is determined to be 25 °C/W without sealed water, and 17 °C/W with sealed water.

Summary

By utilizing conventional IC processing technology, Membrane MCM systems have been fabricated to study the electrical design and thermal management of advanced systems. Bipolar interface circuits and bipolar series regulators can be built in the module substrate or in a separate chip. The bipolar interface circuits, when connected to CMOS chips, have been shown to reduce the chip-to-chip signal delay. Bipolar series regulating transistors combined with integrated decoupling capacitors have reduced power supply noise by a significant factor. Due to the configuration of the Membrane MCM systems, water can be sealed between the chips and the substrate to decrease the thermal resistance and provide additional thermal capacity.

References

[1] W. Cheng, M. Beiley, and S. Wong, Membrane Multi-Chip Module Technology on Silicon. Proceedings of the IEEE Multi-Chip Module Conference, pp. 69-73, 1993.

[2] B. Wooley, M. Horowitz, R. Pease, and T. Yang, Active Substrate System Integration. Proceedings of 1987 IEEE International Conference on Computer Design, pp. 468-471, 1987.

[3] R. Day, C. Hruska, L. Tai, R. Frye, M. Lau, and P. Sullivan, A Silicon-on-Silicon Multichip Module Technology with Integrated Bipolar Components in the Substrate. Proceedings of the IEEE Multi-Chip Module Conference, pp. 64-67, 1994.

[4] C. Chao, K. Miyamoto, K. Sakui, W. Cheng, and B. Wooley, An Active Substrate MCM System. Symposium on VLSI Circuits Digest of Technical Papers, pp.47-48, 1994.

4. Change in Current Research Support of Senior Personnel

Michael J. Flynn, P.I.: "Smart Photonic Networks and Computer Security for Image Data," Grant DAAH04-95-1-0123, funded $500,000 from 6/1/95–5/31/96.