A Bare-chip Probe for High I/O, High Speed Testing

Alan Barber, Ken Lee, and Hanns Obermaier

HPL–94–18
March, 1994

test, wafer probe,
thin film, chip

Abstract —We will describe a bare chip probing fixture for temporary interconnection of a VLSI tester to a die. It is capable of connecting to an area array of die pads, can operate beyond 1 GHz, and is extensible to 1000 signal I/O's. This probe has been adapted to an existing VLSI tester by attaching it to a custom DUT board and has been used to test operational silicon.

The fixture consists of a four metal layer membrane probe which is an enhancement to a previously described burn-in fixture with a novel alignment scheme and no-wipe contacting buttons. The probe is electrically connected to the DUT pcb with an array of button connections, and board I/O is through coaxial cables to the tester. A mechanical structure provides alignment of the pcb, button connector, and membrane probe while providing controlled pressure between the membrane and die, and at the same time cooling the die.

We will describe the electrical performance of the interconnect and the results of testing a circuit toggling at up to 1 GHz, compare them with another probing solution and describe future improvements contemplated. In addition, we will briefly describe the potential for use as a very fast bare chip burn-in fixture.

1 Background

The search for high performance, low cost computing has led many to consider mounting multiple large dice in multichip modules (MCM's). The penalty for mounting bad chips in MCM's is high, and well documented [1]. VLSI testers capable of operation at very high speeds and high I/O counts are available, but when the device under test is unpackaged, available probing solutions cannot match the performance of the tester. We sought a probing technology to match the bandwidth and accuracy of today's best VLSI testers [2].

The objectives of high bandwidth and accuracy led to these specific goals:

1. **1000 I/O's**. Power and ground connections are additional for a total pad count of 1500-2000. Such a large number of device connections require an area array of pads. Solder bumped pads were tested, since that is probably the most common area array chip attachment method in use for VLSI today.
2. **660 Mbps**.
3. **Low distortion**. When parametric measurements will be made at high speed, signal interconnections must have low loss, reflection, and crosstalk. And even more demanding, the power supplies and grounds must be able to supply many amps of current at very short rise times, with very little deviation in voltage. How much distortion is tolerable is dependant upon the application, of course, but here we will relate the distortion to the resulting degradation in tester accuracy specifications.

Potential wafer probing solutions were explored. The most aggressively designed cantilever probes still offer several nH of inductance and cannot accommodate the large I/O count and area pad array capability [3]. Arrays of spring loaded pins or buckling beams have been described that can meet the I/O count requirement [4, 5], but they have substantial inductance for the frequency or distortion requirement. The simulated performance of one such probe was examined and will be shown. Membrane probes or overlay fixtures have excellent high frequency properties, but the commercially available ones do not realize their full high frequency performance capability either because they do not have the multiple layer capability required to bring power and ground planes and microstrip connections to a chip area array or they are combined with sockets that have large uncontrolled impedances [6, 7, 8, 9].

We have built and tested a prototype bare chip probe system which meets the second and third goals, and has sufficient wiring density to meet the first. This design is a modification of a previously reported bare chip burn-in fixture [10]. The enhancements include additional layers to accommodate full microstriplines all the way to the chip area array pads as well as independent power and ground planes over the full area of the chip. Also the signal I/O design was modified for high frequency performance and the mechanical support structures were redesigned to accommodate a high speed VLSI tester.

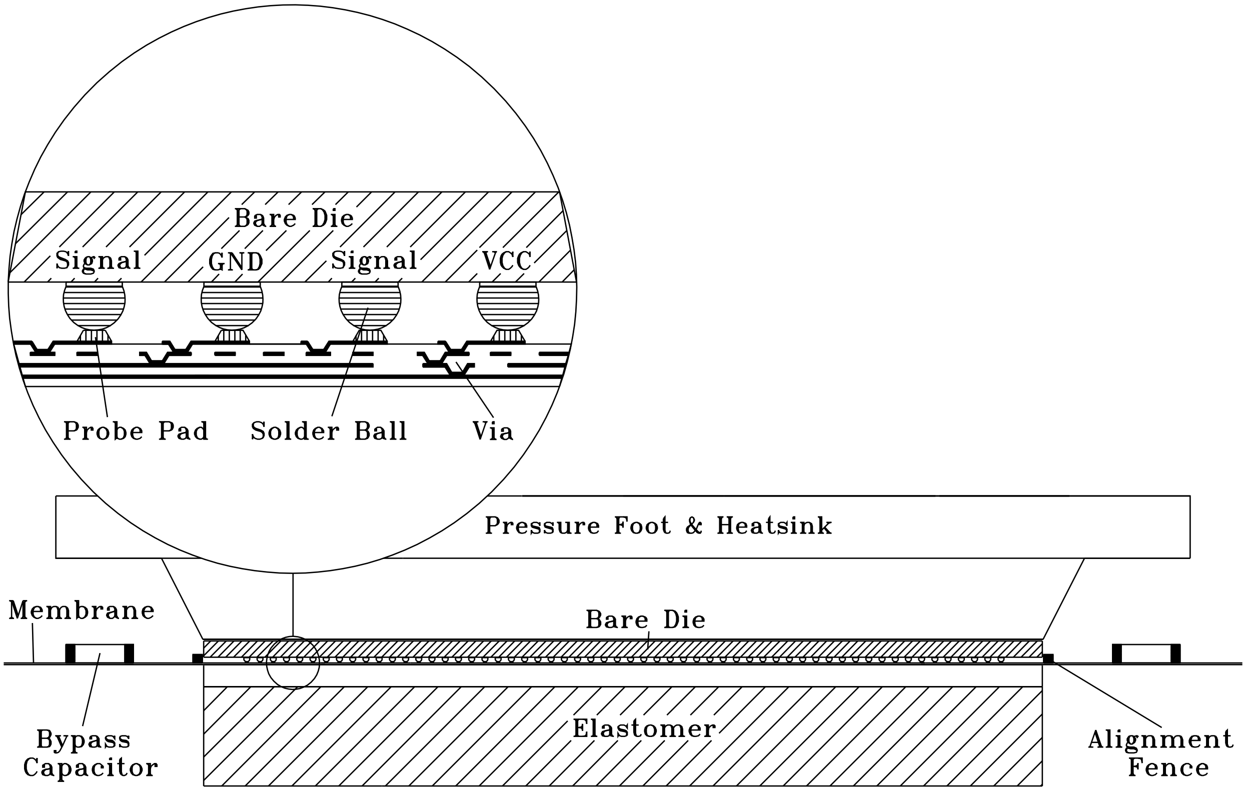

Fig. 1 Cross Section Detail of Membrane and Chip

2 Physical Configuration and Operation

Figure 1 shows a cross section of the chip to probe interconnection. The probe itself is a four metal layer polyimide membrane supported by an aluminum frame. The membrane holds bypass capacitors and an integrated alignment fence which holds the die on the membrane. In this case, the chip has an area array of pads at 250 µm pitch with solder bumps applied. The chip and membrane are compressed between a lower support and an upper piston designed to apply a controlled force between membrane probe pads and chip pads. The four metal layers of the membrane include the probe pad layer with the oxide penetration features, one signal layer, one power distribution plane, and one ground plane. For our prototype, a membrane was designed to accommodate an existing test chip containing a variety of high speed receivers, data latches, drivers, and other test circuitry. Figure 2 shows a plan view of the signal layer. The alignment fence can be seen. Of the 122 chip signal I/O's, 114 are clustered in one corner of the chip, occupying about 5% of the chip area. With an additional signal layer this probe could accommodate 1000 I/O's. Chip I/O's are routed with embedded microstrip lines to pads over the aluminum support frame on the periphery of the membrane.

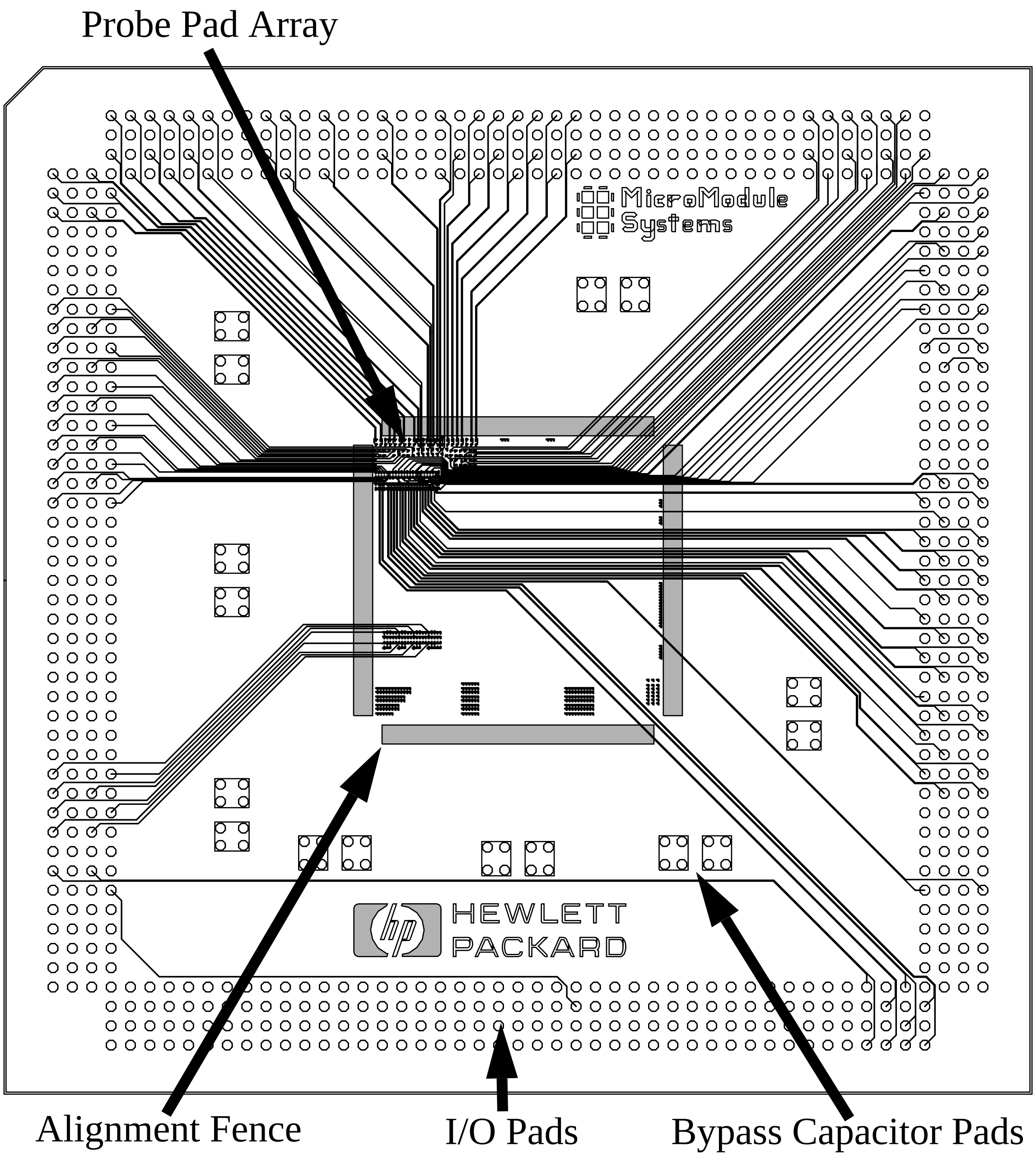

Fig. 2 Membrane Signal Layer

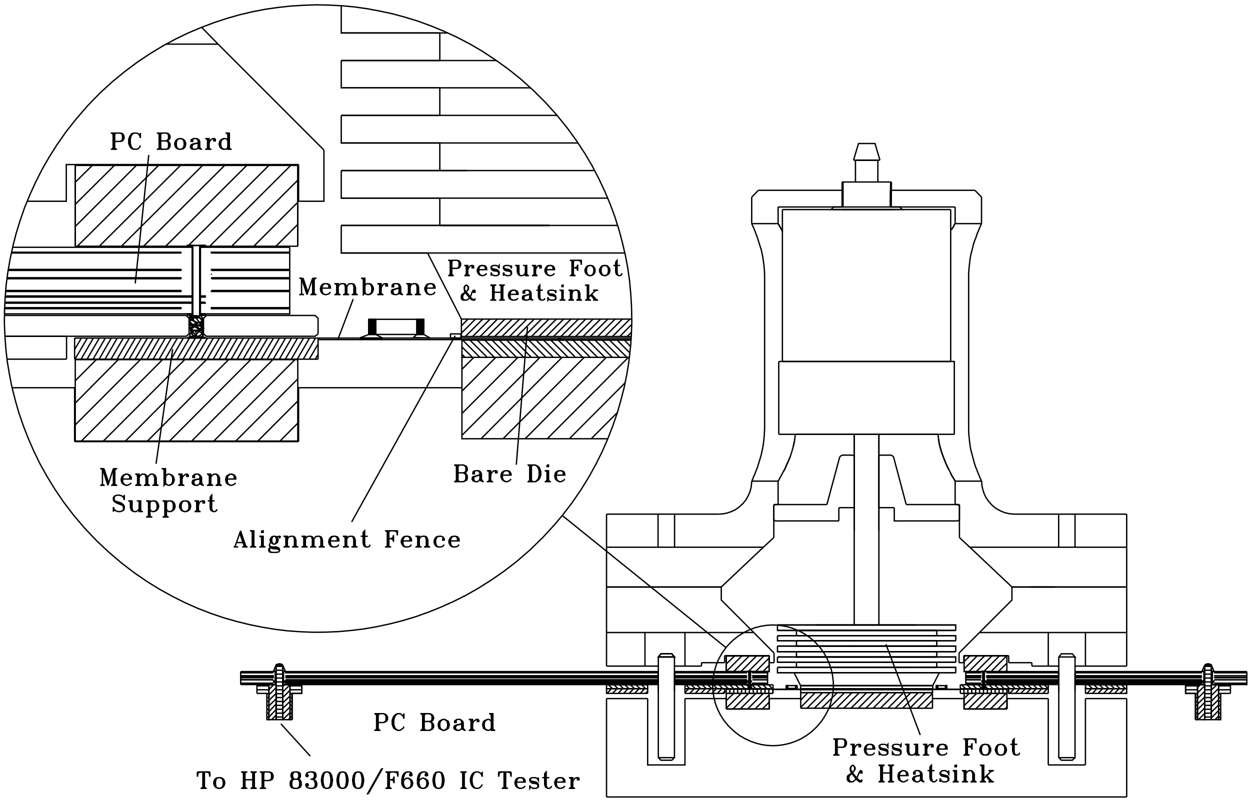

Fig. 3 DUT Board and Test Fixture Details

Figure 3 shows the assembled probe with the mechanical support and cooling structure. The connection between the membrane I/O pads and the DUT board is through an array of button contacts. These are crumpled gold plated molybdenum wire formed into a cylindrical shape and stuffed into holes in a thin polyetherimide interposer. Above the interposer is a spring alignment plate that centers the membrane over the interposer. The button contacts are compressed between membrane I/O pads and DUT board I/O pads. The interposer is aligned to the DUT board with four pins in slots that are designed to minimize misalignment during thermal excursions by maintaining a true center reference between board and membrane. Within the DUT board, striplines connect the button contact pads to SMA coaxial connectors. In operation, the membrane, holding the chip, is pretensioned from the bottom by a pressure plate. The membrane, interposer, and DUT board are all clamped between this bottom pressure plate and the upper support structure shown in Figure 3. The upper structure holds a pneumatically actuated pressure foot which applies controlled force against the chip. The pressure foot also serves as a heat sink and the upper support structure contains an air duct through which heated or chilled air can be forced to control chip temperature. The die under test is removed and replaced by raising the pressure foot, swinging the upper support assembly aside and moving the chip with a vacuum pickup. The whole DUT board assembly mounts into the DUT interface frame of the VLSI tester mainframe and coaxial cables connect the SMA connectors to the pin electronics within the mainframe. Alternately, either a pulse generator and oscilloscope, or a microwave vector network analyzer can be connected directly to the SMA connectors for system characterization or actual chip testing.

3 Making Reliable Contact

When probing large, dense arrays of I/O pads on a die there are several important requirements for making reliable electrical contact:

1. Alignment of probe and die pads in X and Y.
2. Compliance between probe and die pads in Z.
3. Penetration of non-conducting surface films.
4. Repeatability within a range of expected operating conditions.

In the prototype test fixture single dice are placed manually with a vacuum pencil into the integrated alignment fence. The fence opening is typically oversized by 0.024 mm to allow for the small variations in die size. The required dicing accuracy for this alignment scheme of .012 mm in overall size and pad-to-edge dimensions is well within the capability of today's dicing equipment.

A pressure foot gimbal and an elastically suspended membrane support assure probe-to-die coplanarity. Plastic deformation of the solder bump compensates for most variations in bump height. Elasticity and flexing of the membrane allows probing of aluminum pads. Probably the most difficult task in making reliable contacts on conventional aluminum die pads is the penetration of the hard, non-conductive oxide film. Conventional tungsten wire probes on aluminum pads rely on a controlled wipe to break through. Spring loaded pin probes (pogo pins or needles) typically rely on stress concentration to displace oxides by cold flowing underlying metals. Oxide films on softer metals such as tin/lead used for solder bumps on flip-chips are generally easier to break and require a relative small amount of force to displace. Buckling beam probes used on solder bumps are normally designed to have a small wipe as well as sufficient force to displace oxides. In all cases probing must not contaminate or destroy the die pad or solder bumps so that subsequent processes can be performed successfully. Wire probe marks on aluminum pads, for example, could interfere with the solder bump processes.

Wear and contamination of probe tips or pads is an issue for long term reliability. Buckling beam probe users have reported a build-up of tin/lead and its oxides on probes used for solder bumps. A cleaning procedure at regular intervals is required for reliability. Only a relative small number of mating cycles with the membrane probe were tested at this time; more extensive tests to demonstrate reliability are under way.

Figure 4 shows the membrane probe pads employed in this test. They rely on stress concentration produced by a unique surface structure (Figure 5) to achieve redundant

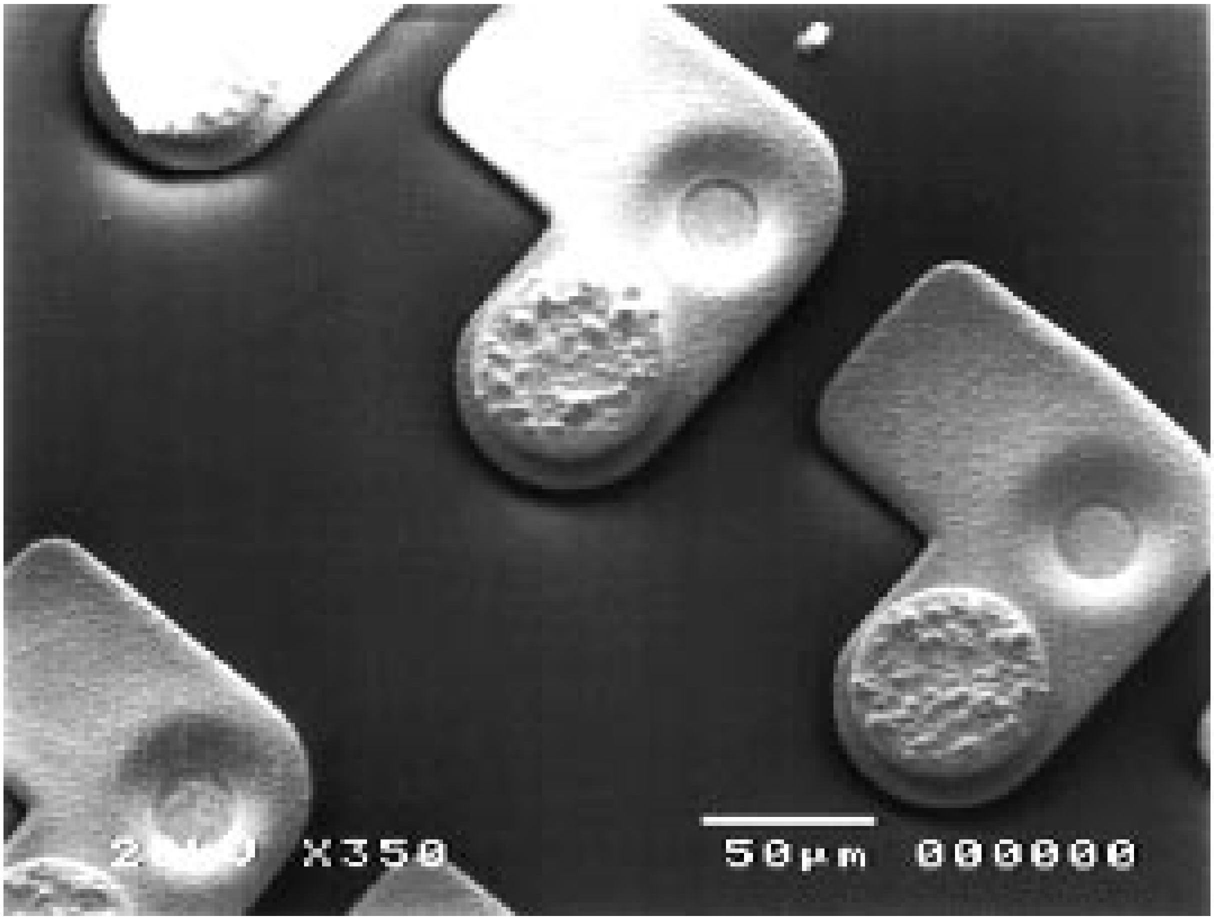

Fig. 4 Membrane Probe Pads

contacts on each pad or solder bump. Previous tests of burn-in sockets using identical pad structures on a membrane and contacting aluminum pads on a die indicated predictable contact resistance over a number of mating cycles and temperature ranges [10]. Our tests were focussed on demonstrating feasibility of the described technology for a test and burn-in application of solder bumped flip chips. Figure 6 shows the result of early probe tests. Stable contact resistance is reached typically at about 10 gram/bump with solder bumps and 20 gram/pad with aluminum pads. A force greater than 20 gram/pad results in severe deformation of the solder bumps requiring an additional reflow step prior to assembly. More extensive tests to validate probing of large arrays are under way using test chips with a number of daisy chains and four wire test pads.

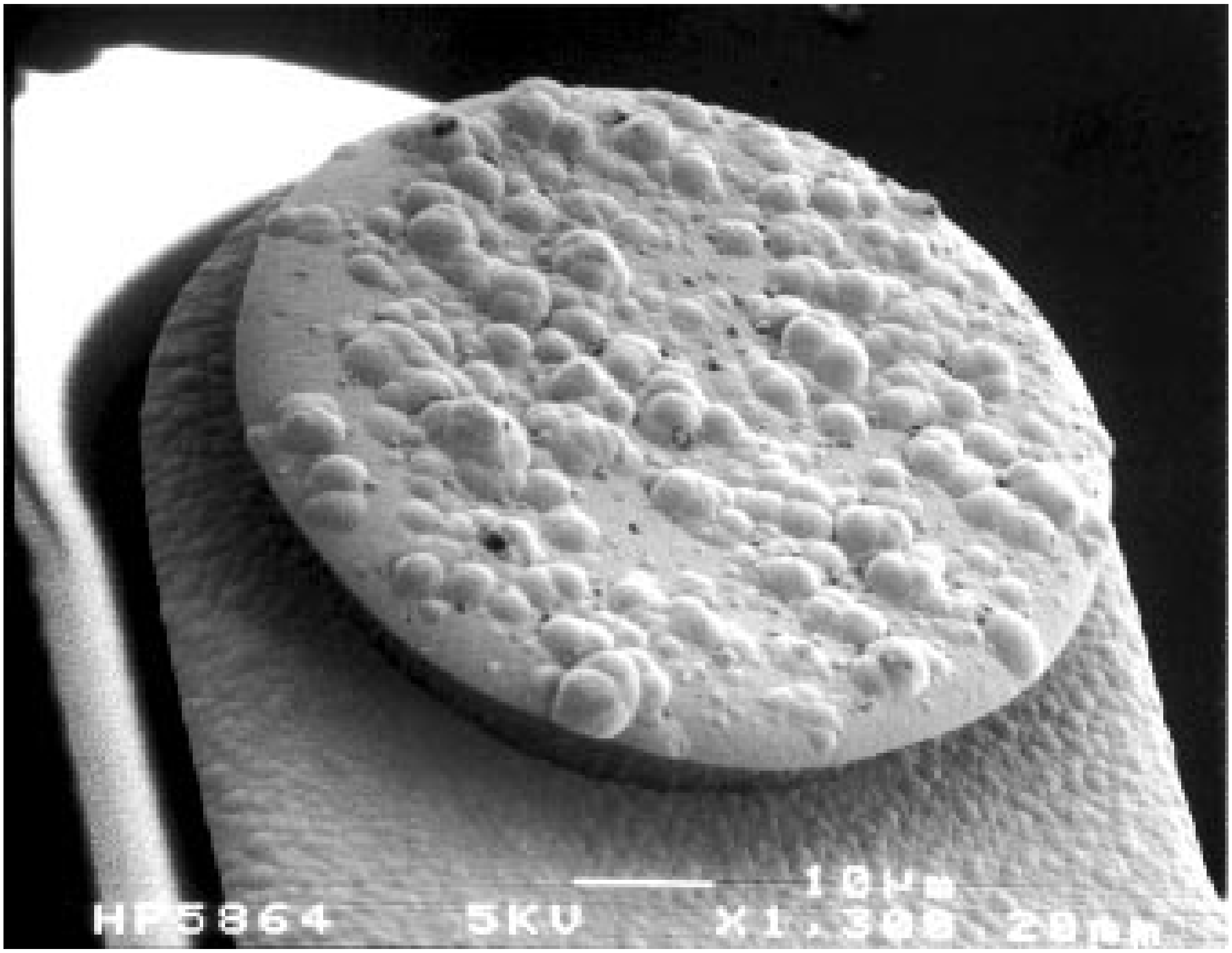

Fig. 5 Pad Surface Structure

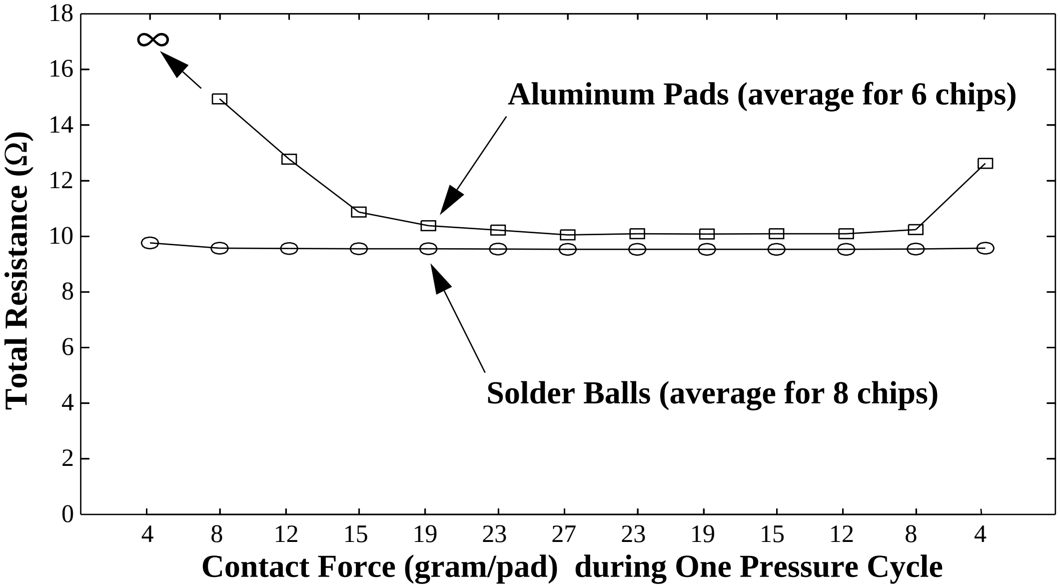

Fig. 6 Contact Force vs. Resistance

4 Electrical Performance

The probe was measured with a microwave vector network analyzer, tested with a high speed pulse generator, and operated installed in a VLSI tester. An electrical model was built and compared to the measured data in the frequency domain. That electrical model was then used in the time domain to explore the effects of the probe on timing measurements made on a high speed CMOS circuit. Finally, these effects were compared to those predicted for another high performance probe using buckling beams.

Figure 7 shows the round trip frequency response from the connector on the DUT board through the pc board, through a membrane line, through a very short line on the chip, and back out through a similar path to another DUT board connector. This is the magnitude of the S_{21} scattering parameter as measured with a 50 Ω microwave network analyzer. Also shown is the frequency response as modeled with microwave design software. Figure 8 shows the model used to simulate the network with commercial analysis software. The coaxial connectors, PCB traces, and membrane probe traces are described by their physical dimensions and static electrical characteristics (i.e. dielectric constant and conductivity). The capacitance of the button connector pads and chip I/O pads are the calculated parallel plate capacitances. The inductance and resistance of the probe contacts themselves were calculated in a way described below. Good agreement is seen between measured and modeled responses.

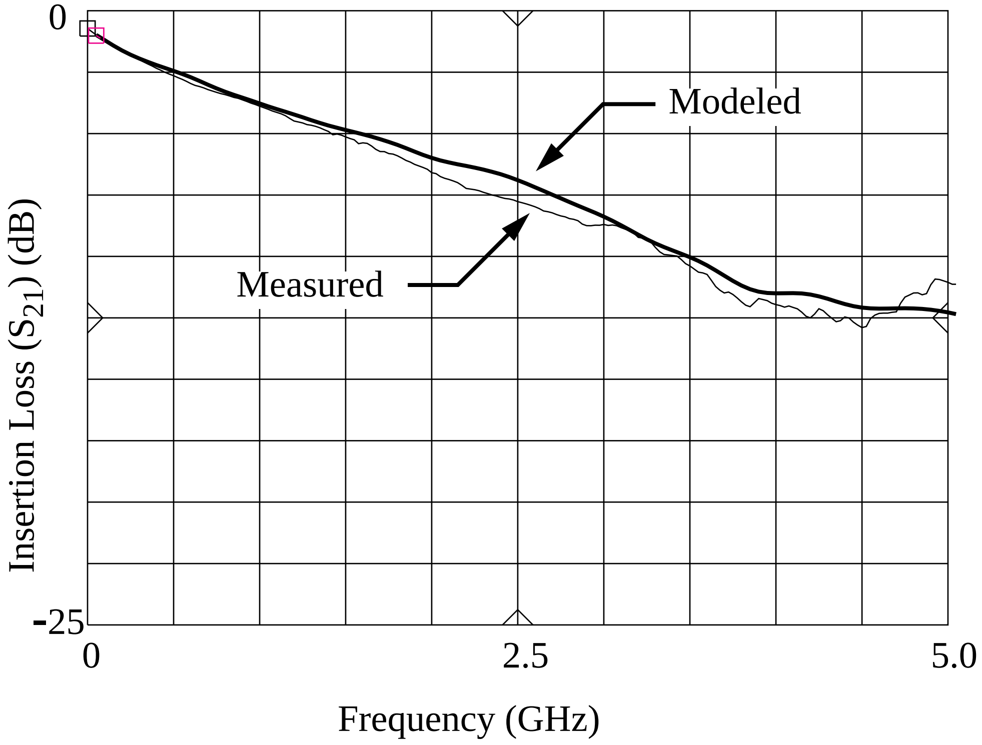

Fig. 7 Probe and Interface Board Insertion Loss vs Frequency

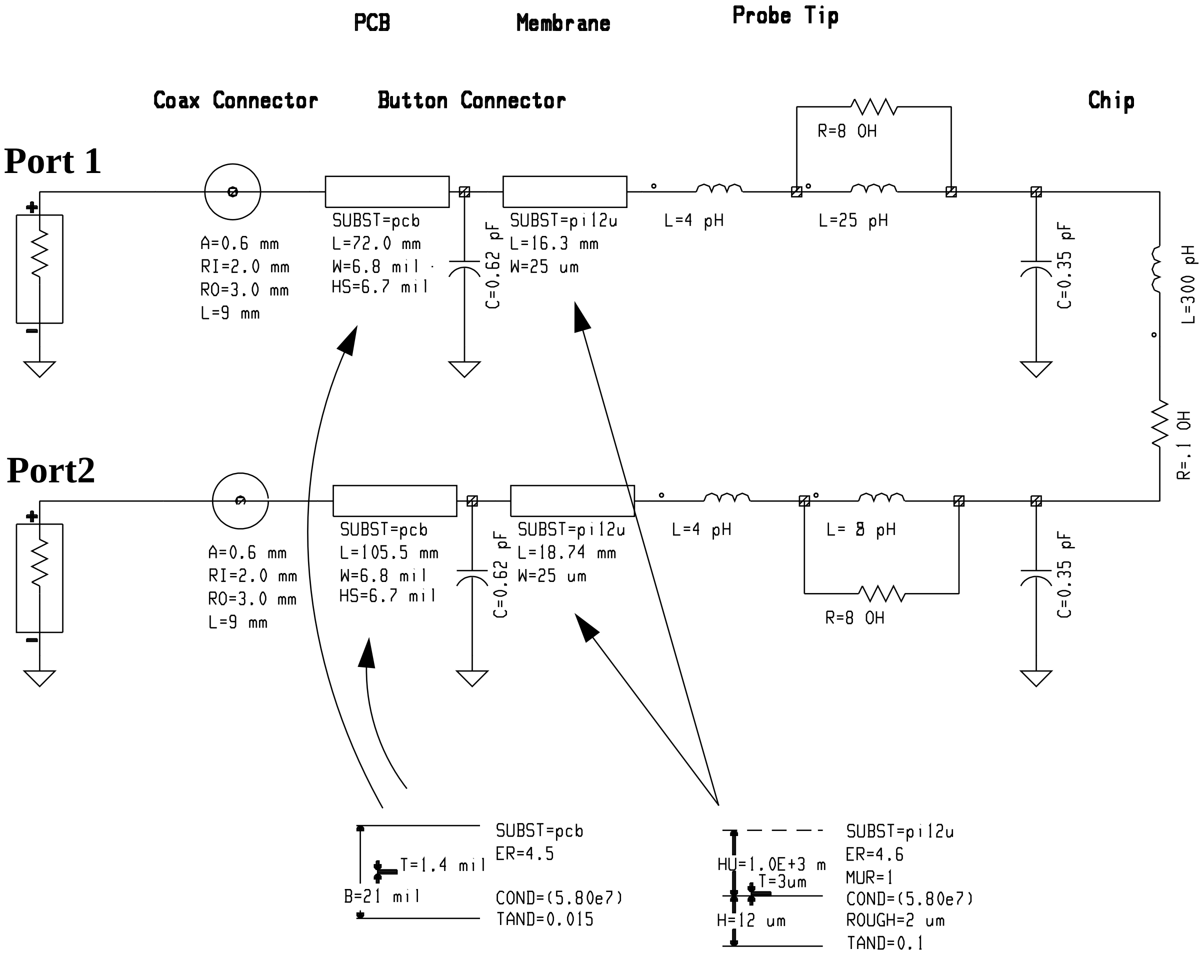

Fig. 8 Model of Membrane Probe and PCB

In addition, crosstalk between adjacent lines was measured as follows. 50 Ω microwave probes [11] were used to contact two adjacent active pads of the membrane probe. Each active pad has an adjacent ground pad and is connected via the line on the membrane to a button contact at its periphery. In this measurement the probe was not mounted in the DUT board so the button contacts were unterminated. The microwave probes, in turn, were connected to the microwave network analyzer with coaxial cables and the analyzer was calibrated at the microwave probe tips, thus establishing a reference plane at the membrane probe contact point. An S_{21} measurement with this arrangement yields the crosstalk between adjacent lines and includes the effect of coupling between the probe tips, the lines on the membrane, and the button connectors. It also combines forward and backward crosstalk components since backward crosstalk is measured directly and forward crosstalk is reflected from the unterminated button connector and measured. The crosstalk was measured in the frequency domain and converted to time domain to give the peak voltage induced. For 200 psec risetime edges the p-p crosstalk was under 0.5% and for 75 psec risetimes it was approximately 2%. The value is low because the membrane lines are close together for only a very short distance near the chip pads.

The signal distortion contributed by the membrane line and probe tip was simulated through use of the model of Figure 8. Reference [12] contains a good discussion of the signal distortion of fast edges caused by lossy lines such as this. The lower (and longer) path from port 2 to the chip interface was simulated and the frequency domain data converted by Fourier transform to the time domain to examine the response to a unit step input at port 2. The result is shown in Figure 9 with the bandwidth limited to produce a 200 psec step rise time. The input step has been translated in time to coincide with the output step and adjusted by 5% in scale to match the d.c. levels of the input step to allow a direct comparison of input and output edge shapes, as described in [12]. The transition from fast-rise region to slow-rise region caused by line resistance is seen to occur very high on the waveform, contributing little to waveform distortion and delay in the middle portion of the rise. This transition will occur at:

$$\frac{V_{tt}}{V_{ot}} = e^{(-Rl)/(2R_o)}$$

where R is the line resistance per unit length, l is line length, and R_o is the real part of the line characteristic impedance. In this case, 93% of the resistance is in the membrane portion of the line (R=2.3 Ω/cm, l=1.874 cm), the rest in the PCB portion

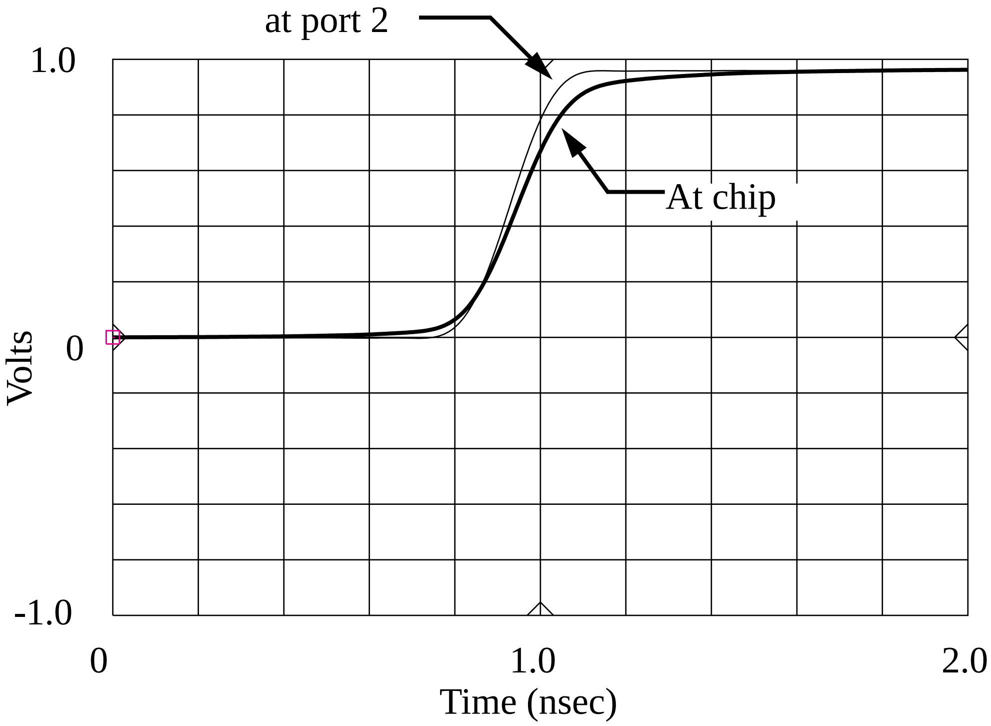

Fig. 9 Distortion of Fast Rise Edge Caused by Membrane Probe and PCB

(R=28.5 mΩ/cm, l=10.55 cm). Combining these values into the above expression yields a transition point at 95% of the output step. Keeping this transition point high on the waveform reduces its impact on the device being tested, whose switching point will typically be near the midpoint of the waveform.

Perhaps more difficult and important than a good quality signal path is a good quality power and ground connection to the DUT. The ideal power supply would deliver amps of current at very fast rise times with no voltage deviation at the chip connections. This requires that a source of charge be available to the chip terminals with very little inductive reactance, a burden which falls heavily on probe design. Another way to look at this requirement is to say that the impedance presented to the chip by the probe power supply terminals be very low at all frequencies. Failure to meet this requirement at high frequencies will cause power supply voltage to change when fast steps of current are required. Failure to meet it at low frequencies will cause power supply voltage to change when sustained surges are required. The high frequency requirement is typically met by placing bypass capacitors as close to the chip as possible. Any resonance or high impedance will result in a ringing or glitching power supply voltage with large numbers of simultaneously switching drivers on the chip. This membrane structure provides a nearly ideal implementation of this high frequency requirement by devoting a whole metal plane to power supply and another plane to ground and by providing large value bypass capacitors close to the chip. To evaluate the contribution of the probe to meeting this high frequency requirement a measurement of the impedance presented to the chip by the probe was made. This measurement was made directly at the probe pads by probing with the same microwave probes and network analyzer configuration described above. The probe was not installed in a DUT board. Figure 10 shows the result plotted on a Smith

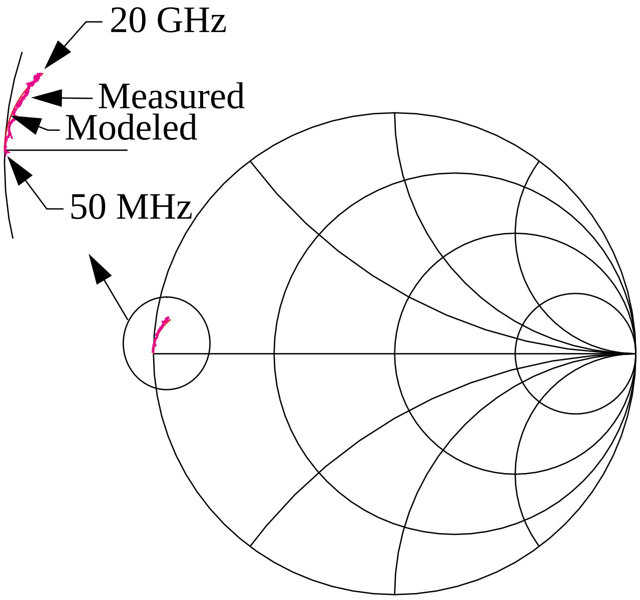

Fig. 10 Impedance of Probe Power Supply Presented to Chip Pads

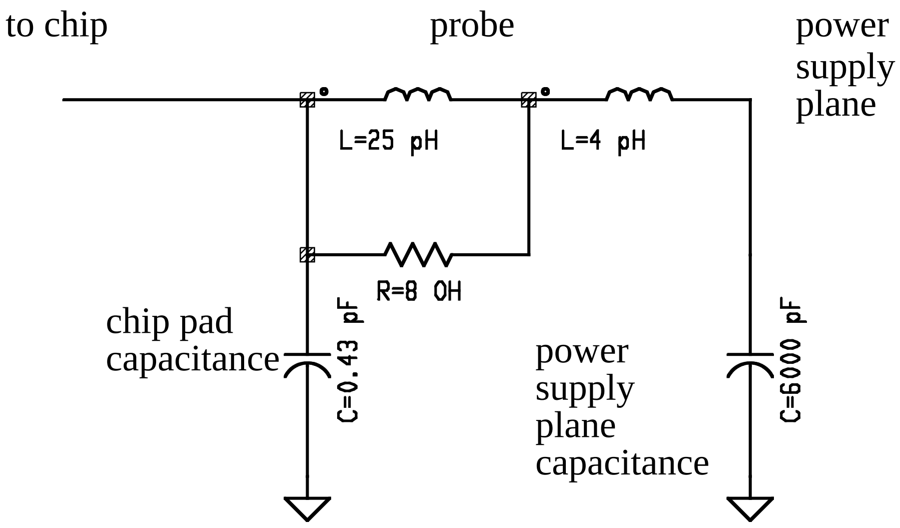

Fig. 11 Model of Power Supply Plane Impedance Presented to Chip

Chart. It certainly approaches the high frequency ideal of a short circuit from 50 MHz to 20 GHz. In addition, a simple lumped equivalent circuit, Figure 11, was generated and fitted to the measured response. This equivalent circuit was also used for the signal line to chip transition described above since this transition is identical for signal connections and power connections.

D.C. power supply performance may also suffer with a membrane probe due to the high resistance of the thin film metallization, only 2 µm thick with sheet resistance of about 10 mΩ per square. In this case, however, with a full plane devoted to power supply and ground each and with power delivered from all four sides of the membrane, there is only about 1/8 square, or about 1.25 mΩ of power supply and ground resistance.

Each of the measurements and simulations above was carried out in a 50 Ω environment. We simulated performance in a measurement environment by using the signal and power supply models described above combined with models for a CMOS inverter and line driver as the device under test (DUT) in a SPICE simulation. The CMOS circuits used 0.6 µm gate length devices with a line driver capable of a 200 psec risetime into a 50 Ω transmission line. Three circuits were driven simultaneously in order to study the waveform and delay distortion created by interference from neighbor circuits. In addition, we created an electrical model for another high performance area array probe and substituted this model for the membrane probe model to compare the performance of the two. The three devices under test are contacted by a small array of ten probe points, three inputs, three outputs, and four power supplies. The power supply probe points were positioned between DUT's to reduce interference among them. Figure 12 shows the arrangement for the comparison probe case. The model for this comparison probe was an array of

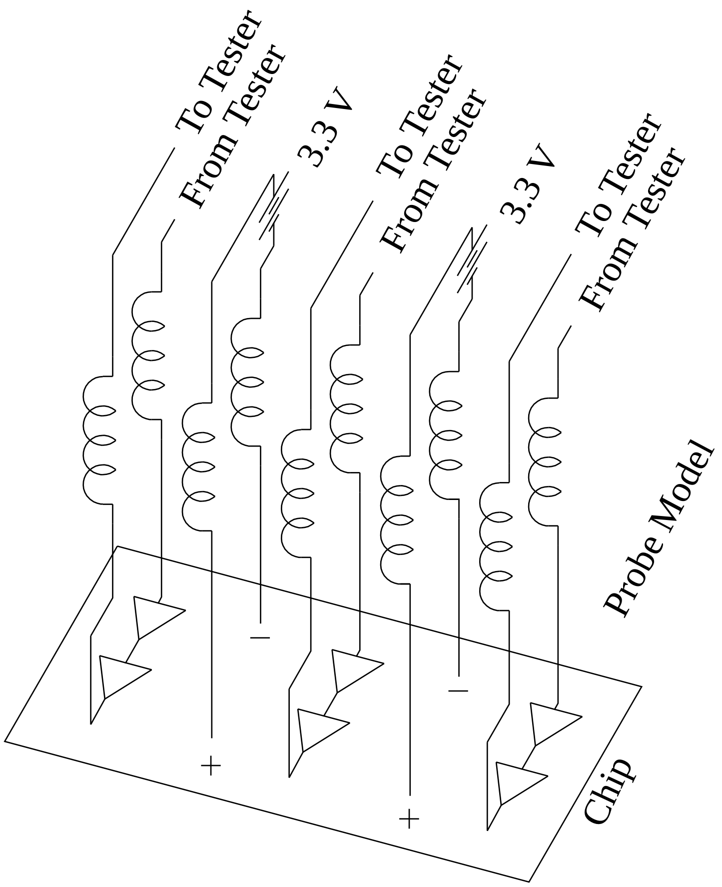

Fig. 12 Model of Comparison Probe Used to Evaluate Distortion

ten mutually coupling inductors with inductance values computed from published dimensions [4]. The computed self inductance of a single probe was 4.5 nH and the mutual inductance varied from 2.9 nH for nearest neighbors to 1.58 nH for diagonally opposite neighbors. For the case of the membrane probe, the probe tip model of Figure 8 was substituted for each of the inductors in Figure 12. To study the effects of just the probe, a perfect power supply was assumed on the tester side of the probes. The SPICE connection to the probes (simulating the tester) was through 50 Ω transmission lines and the input signals were single 200 psec risetime edges simultaneously applied to all three circuits. Resulting waveforms for the center of the three circuits are shown in Figures 13a and 13b. Distortion of the step output waveform as well as power and ground voltage deviations are evident in the comparison probe. In many digital test applications these distortions are quite acceptable as long as data can be successfully toggled into and out of the DUT. However, if one wishes to make an ac parametric measurement, such as setup time, delay, or maximum frequency of operation, they can do damage. For example, if the edge timing of two devices is varied while examining the delay through a third device,

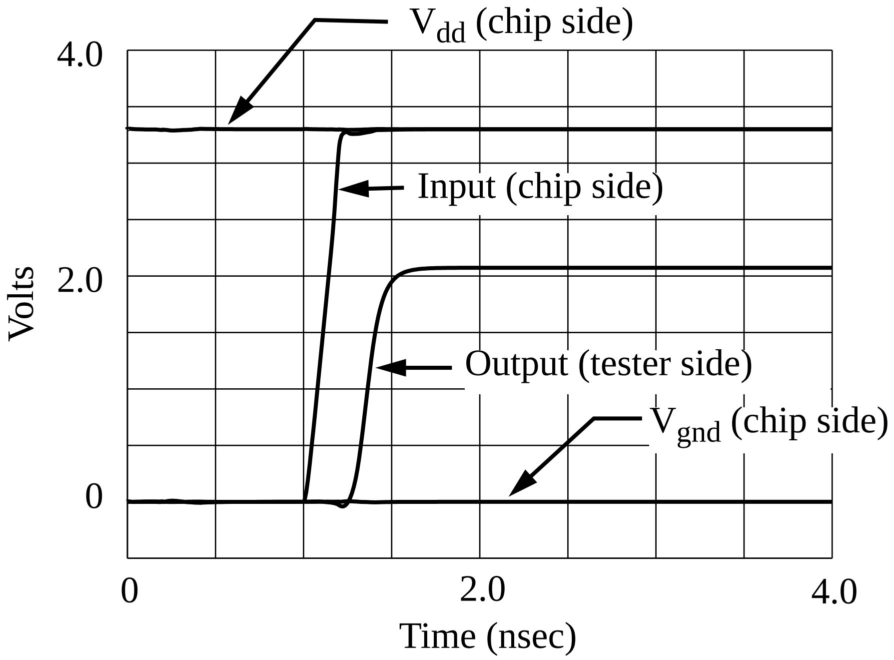

Fig. 13a Output and Power Supply Waveforms for Membrane Probe

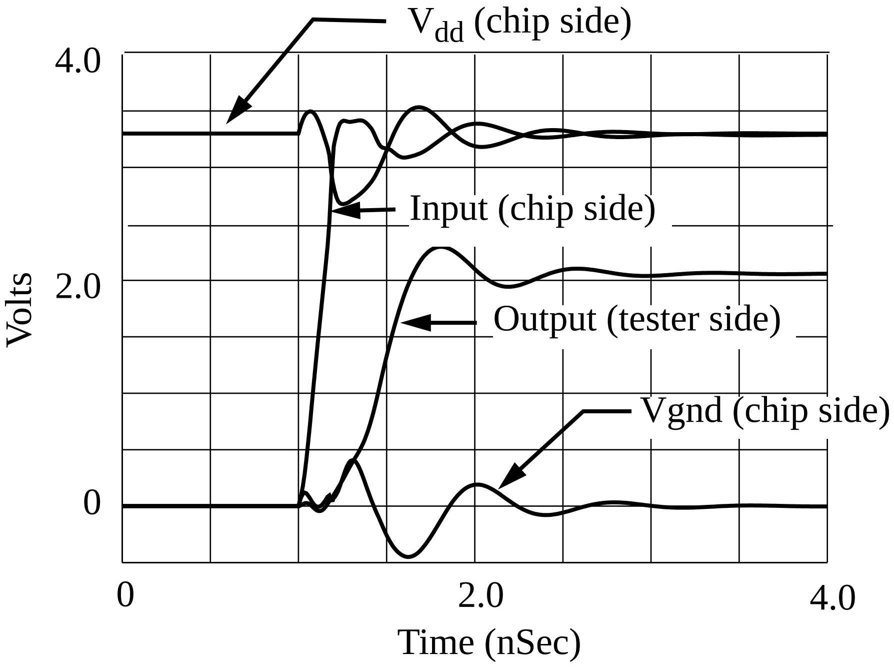

Fig. 13b Output and Power Supply Waveforms for Comparison Probe

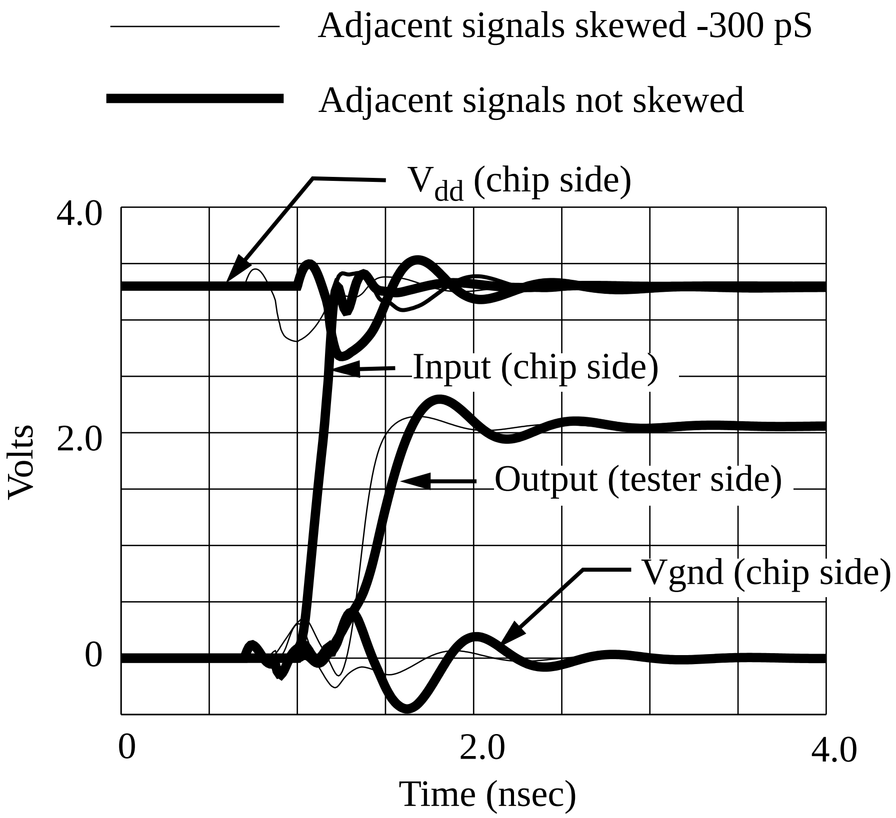

Fig. 14 The Effect of Skew of Adjacent Interfering Signals on Output Waveforms

the power supply noise caused by the two will be coupled into the third. This is illustrated in Figure 14. To the waveforms of Figure 13b have been added a set of waveforms,again for the center circuit,but in this case the two nearest neighbor signals are skewed 300 psec earlier. The power supply coupling has moved the output edge under observation by about 100 psec. Figure 15 shows how the measured delay through the center of the three DUT's in this example varies as the skew of the two adjacent edges is varied for both the membrane and the comparison probe. Such delay variations would be very difficult to calibrate out in a practical measurement situation and would contribute directly to measurement inaccuracy. In other words, this 100 psec delay variation adds directly to the tester's specifications for edge placement accuracy and resolution, which in some testers is at least 50 and 10 psec respectively.

The probe was also evaluated with a test chip containing a variety of high speed bipolar receivers, latches, and drivers. Figure 16 shows output data after being latched into and out of the chip at 1 Gbps using a high speed data generator and oscilloscope.

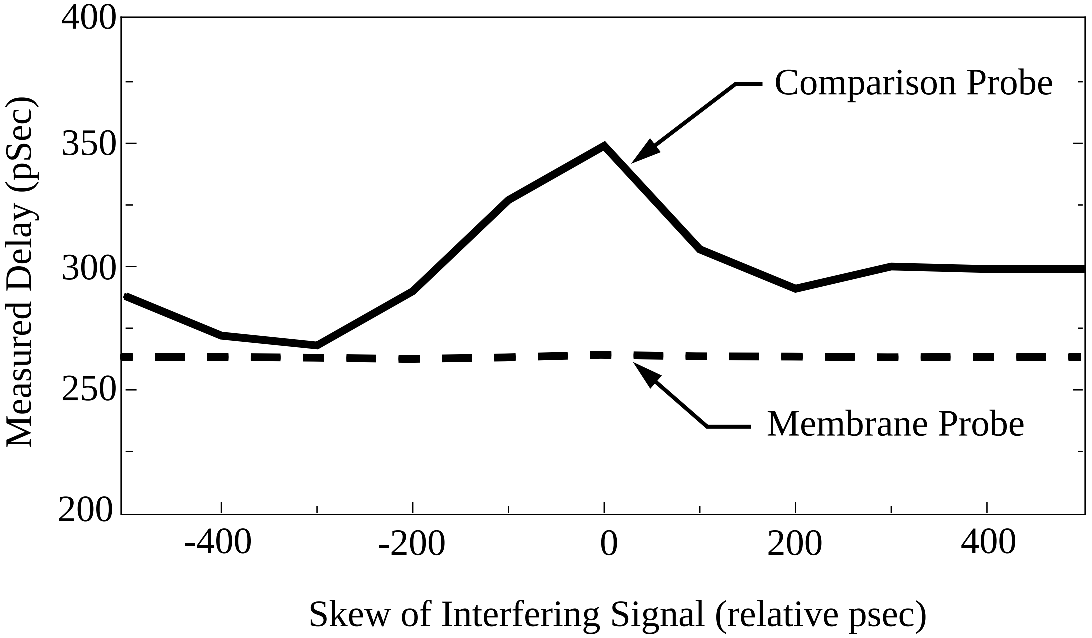

Fig. 15 Measured Input to Output Delay Variation with Skew of Adjacent Interfering Signals

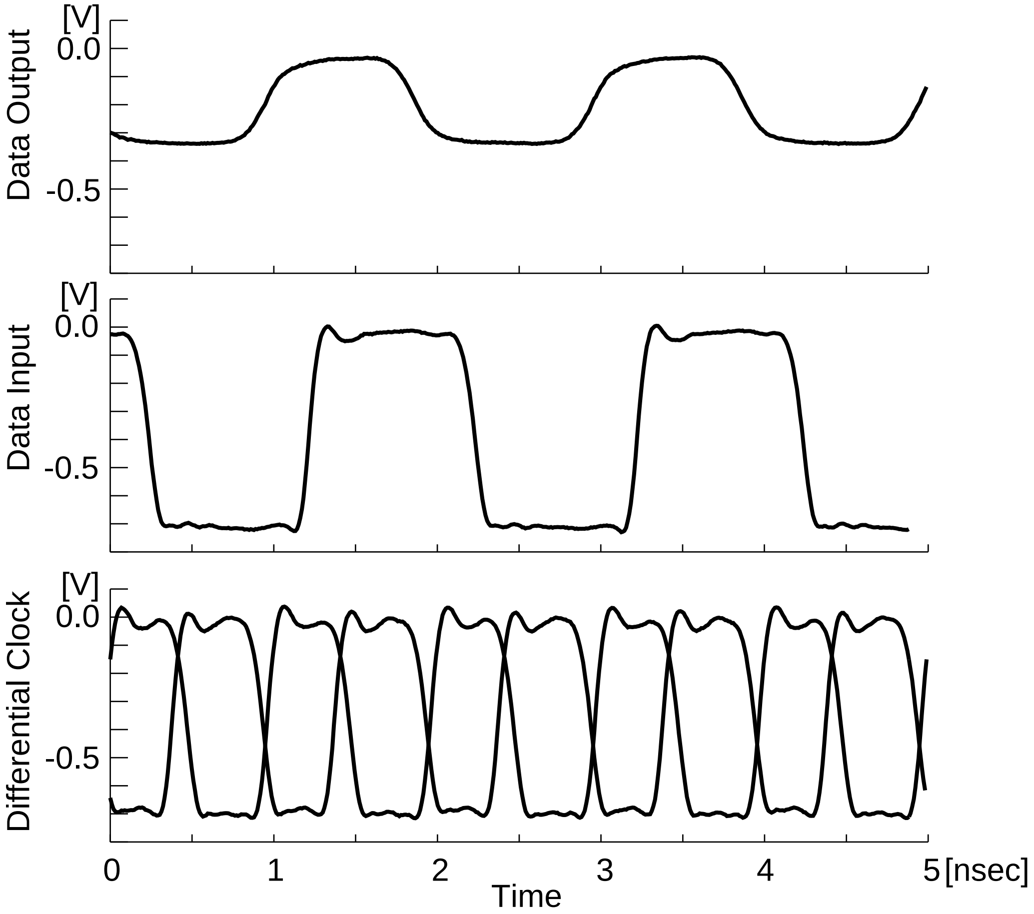

Fig. 16 Data Output: VLSI Tester

5 Future Work

This first prototype has suggested to us additional analysis and product improvements that are now planned:

1. **Reliability Testing**. Tests with large numbers of mating cycles with solder bump metallization.
2. **Lower loss lines**. The loss in the coaxial cables can be reduced by using larger cables. The use of teflon dielectric in the DUT board results in lower loss lines and the membrane itself can be designed smaller, and the lines shorter, if an additional metal layer is used for signal distribution.
3. **Eliminate button contacts**. The existing pressure connection between the probe pads and solder chip metallization seems to be very workable. We believe we can extend this technique to the I/O pads and eliminate the whole button and interposer structure, making a direct pressure connection between the I/O pads and DUT board. This will simplify the structure and reduce the size of the largest reflection producer, the membrane pads.
4. **Use for burn-in**. A method has been developed [13] to screen for infant mortality reliability problems by making ac parametric and functional tests while rapidly ramping device temperature. It is more sensitive than traditional High Temperature Operating Life burn-in and can be done in about a minute on a high speed VLSI tester with proper chip handling and temperature control. This probing method would allow such screening to be done in the bare chip form, saving the cost of packaging devices which would ultimately fail HTOL burn-in.

6 Summary

We have described a bare chip probing technology that can contribute to solving the MCM Known Good Die problem. It is easily adaptable to a VLSI tester and allows chip testing with large numbers of high data rate signals with either peripheral or area array connections. The electrical properties allow highly accurate delivery of signals and power to the chip under test. Future developments include higher performance, simpler structure, and use simultaneously for test and burn in.

Acknowledgements

Thanks go to Chung Ho and Fariborz Agahdel of MicroModule Systems for the design and fabrication of the membrane, and to Steve Dielman and Matthias Werner of Hewlett Packard's Semiconductor Systems Center for training on and use of their HP83000 VLSI tester.

References

[1] D. A. Doane and P. D. Franzon, Multichip Module Technologies and Alternatives: The Basics, Chap. 13, Van Nostrand Reinhold, New York, 1993.

[2] H. Engelhart and U. Schottmer, "High Performance Pin Electronics with GaAs--A Contradiction in Terms", International Test Conference 1992 Proceedings, pp. 538-545, 1992.

[3] E. Subramanian and R. Nelson, "Discover the New World of Test and Design", International Test Conference 1992 Proceedings, pp. 936-939, 1992

[4] D. J. Genin and M. M. Wurster, "Probing Considerations in C-4 Testing of IC Wafers", The International Journal of Microcircuits and Electronic Packaging, Volume 15, Number 4, pp. 229-238, Fourth Quarter 1992.

[5] S. Kasukabe, S. Harada, T. Maruyama, and R. Takagi, "Contact Properties of the Spring Probe for Probing on a Solder Bump" Proceedings of the Thirty-Eighth IEEE Holm Conference on Electrical Contacts, pp. 187-190.

[6] T. Fisher and J Kister, "Reducing Test Costs for High-Speed and High Pin-Count Devices", pp. 96-98, EE-Evaluation Engineering, February 1992.

[7] B. Leslie and F. Matta, "Membrane Probe Card Technology", International Test Conference 1988 Proceedings, pp. 601-607, 1988.

[8] R. Parker, "Bare Die Test", Proceedings of the 1992 IEEE Multi-Chip Module Conference", pp. 24-27, 1992.

[9] M. Andrews et al, "ARPA/WL ASEM Consortia for Known Good Die (KGD)", Microelectronics and Computer Technology Corporation, 1994.

[10] F. Agahdel, C. Ho, R. Roebuck, "Known Good Die: A Practical Solution", Proceedings of the 1993 International Conference and Exhibition on Multichip Modules, pp. 177-182, 1993.

[11] D. Carlton, K. Gleason, E. Strid, "Microwave Wafer Probing", Microwave Journal, vol. 26, pp. 747-748, 1990.

[12] M. Lin, A. Engvik, J. Loos, "Measurements of Transient Response on Lossy Microstrips with Small Dimensions", IEEE Transactions on Circuits and Systems, vol 37, number 11, pp. 1383-1393, November 1990.

[13] T. E. Figal, "Below a Minute Burn-in", United States Patent 5,030,905, July 9, 1991.