Annual Progress Report

for the Year ending 28 Feb 1997

and

Request for Continued Funding

M. J. Flynn, P.I.

Grant NSF MIP-9313701

1. Brief Summary of Progress

Algorithms (Professor M. Flynn)

Over the past year we have completed a major study on improving the performance of arithmetic operations using variable latency techniques.

The performance and area of a functional unit depend upon circuit style, logic implementation, and choice of algorithms. The space of current circuit styles ranges from fully-static CMOS designs to hand-optimized self-timed dynamic circuits. Logic design styles range from automatically-synthesized random logic to custom, hand-selected gates. We investigate performance and area tradeoffs at all levels.

The three primary parameters in FP functional unit design are latency, cycle time, and area. The functional unit latency is the time required to complete a computation, typically measured in machine cycles. Designs can be either *Fixed Latency (FL)* or *Variable Latency (VL)* [S1]. In a FL design, each step of the computation completes in lock-step with a system clock. Further, any given operation completes after a fixed quantity of cycles. The cycle time in a FL design is the maximum time between the input of operands from registers and the latching of new results into the next set of registers. In contrast, VL designs complete after a variable quantity of cycles. This allows a result to be returned possibly sooner than the maximum latency, reducing the average latency. They achieve their variability through either the choice of algorithm (VLA) or choice of circuit design (VLC). VLA designs operate in synchronization with a system clock. However, the total number of cycles required to complete the operation varies depending upon other factors, such as the actual values of the input operands. VLC designs need not have any internal synchronization with the rest of the system.

Instead, such a design accepts new inputs at one time, and it produces results sometime later, independent of the system clock.

FL designs can be fully combinational or pipelined. Pipelining is a commonly used technique for increasing the throughput of functional units. A functional unit can be divided into smaller components by introducing explicit registers in-between the components. In this way, the cycle time of the unit becomes the maximum time for any of the components to complete. By increasing the number of components, the cycle time of the unit is decreased at the expense of increasing the latency. The primary motivation for pipelining is to allow for one operation to be initiated and another to be completed in each machine cycle, with more than one operation in progress at any time. As a result, the total latency for a sequence of operations is reduced by exploiting the component-level parallelism. These tradeoffs must be understood for optimal FL and VLA functional unit design.

Power–performance tradeoff and analysis (Professor G. De Micheli)

Reducing power dissipation, while preserving performance levels, is required by most design of portable communication and computation devices. Evaluation of power dissipation is important, when making trade-offs at the system architecture level.

We have addressed the problem of estimating the power consumed by systems that incorporate arithmetic circuits, such as floating-point adders and multipliers, We are modeling such units by using Verilog HDL models, that express completely the behavior and the decomposition into sub-units. We have considered power estimation models based on the I/O switching activity of the arithmetic module. Linear regression is used to estimate the average power.

Regression coefficients are determined by a training period, where the unit is exercised in situ. We have measured a larger accuracy with in situ measurements as compared to off-line measurements. In particular, we have studied the behavior of an IEEE-standard floating-point adder, and we could measure an error in power estimate of at most 15%, as compared to detailed electrical-level simulation. At the same time, we achieved almost a speed-up in the estimation of almost three orders of magnitude. The significance of this result is that the algorithms and tools being developed allow designers to quickly trade-off power/performance parameters in system using FP units, in the search for an architectures that best suit the implementation technology.

Membrane MCM Technology in Silicon (Professor S. Wong)

A problem associated with the propagation of high speed signal on a silicon substrate is the penetration of electric field into the non-insulating substrate, and hence the associated energy loss. This loss becomes significant for above GHz frequencies. Although the problem is well recognized, in-depth understanding and accurate predictive models

are lacking. We are developing a physical model for signal propagation along interconnection that includes the substrate effect. The initial version of the model has been applied to predict the behavior of on-chip spiral inductor, which has attracted growing interest due to the desire for Si-based RF communication circuits.

The scaleable physical model has been shown to accurately predict the behavior of on-chip inductors with different structural parameters over a broad range of frequencies. Each element of the model is consistent with the physical phenomena occurring in the part of the structure it represents. The inductance is determined with Greenhouse algorithm [1]. The model properly accounts for eddy current effect in the conductor, and the capacitance between the conductor and the substrate. The ohmic loss and charge storage in the substrate are represented with a parallel connection of resistor and capacitor whose values are experimentally determined. The model and experimental results confirm that losses in the substrate become significant for above GHz frequencies. We are currently extending the model for general on-chip interconnection.

[1] H. Greenhouse, "Design of planar rectangular microelectronic inductors," IEEE Trans. PHP 10(2):101–109, June 1974.

BiCMOS Active Substrate Probe Card Technology (Professor B. Wooley)

The ability to fully test integrated circuits at the wafer level is becoming increasingly important as a consequence of escalating packaging costs and the emergence of multichip module technologies. In typical test systems, the pin electronics module is located 50-100 cm from the device under test (DUT). During dynamic testing, transmission lines of this length can introduce substantial ringing and waveform distortion. Moreover, impedance mismatches between the DUT and the transmission lines can cause severe errors in delay and timing measurements.

This work introduces an active substrate silicon probe card employing a polyimide membrane formed on a silicon substrate. The probe card combines tungsten probe tips and aluminum interconnects in the polyimide membrane with active test circuitry, integrated in the silicon substrate. The card is potentially capable of providing more than 1000 probe tips in an array format.

The first part of the research focused on the design of a monolithic active substrate probe card in which linear buffers are used to isolate the device under test outputs from the transmission lines in the test system and drive those lines from matched impedances. Each buffer has a gain of 0.25 and employs an active pull down output driver in order to improve the falling slew rate of its output signals.

In the second part, an integrated timing measurement unit combining two test channels and a time digitizer will be described. Designed and integrated in a 0.6um BiCMOS technology, the chip employs phase interpolation to achieve a 90ps timing resolution. A 700MHz phase-locked loop (PLL) time reference is also fully integrated on the prototype

chip. A typical timing error of 38ps RMS has been achieved while dissipating 45mW per test channel from a 3.3-V supply. The PLL has 15.4ps RMS jitter relative to the reference clock.

List of Publications

1. G. McFarland and M. Flynn, "Limits of scaling MOSFETs", Technical Report: CSL-TR-95-662 Revised, Stanford University, Nov. 1995.
2. S. F. Oberman and M. J. Flynn, "Reducing division latency with reciprocal caches," *Reliable Computing*, vol. 2, no. 2, pp. 147–153, April 1996.
3. H. Al-Twaijry and M. J. Flynn, "Technology scaling effects on multipliers," Technical Report: CSL-TR-96-698, Stanford University, July 1996.
4. S. F. Oberman and M. J. Flynn, "A variable latency pipelined floating-point adder," in *Proc. Euro-Par'96, Springer LNCS vol. 1124*, pp. 183–192, Aug. 1996.
5. H. Al-Twaijry and M. J. Flynn, "Optimum placement and routing of multiplier partial product trees," Technical Report: CSL-TR-96-706, Stanford University, Sept. 1996.
6. S. F. Oberman, *Design Issues in High Performance Floating Point Arithmetic Units*, Ph.D. thesis, Stanford University, Nov. 1996.
7. Masoud Zargari and Bruce A. Wooley, "A BiCMOS Active Pull-Down ECL Output Driver for Low Power Applications," 1995 IEEE Symposium on Low-Power Electronics, Dig. of Tech. Papers, pp. 50–51.
8. Masoud Zargari, Justin Leung, S. Simon Wong, Bruce A. Wooley, "A BiCMOS Active Substrate Probe Card Technology for Digital Testing," 1996 IEEE International Solid-State Circuit Conference (ISSCC), Dig. of Tech. Papers, pp. 308–309
9. Justin Leung, Masoud Zargari, Bruce A. Wooley, S. Simon Wong, "Active Substrate Membrane Probe Card," 1995 IEEE International Electron Device Meeting (IEDM), pp. 709–712
10. C. P. Yue, C. Ryu, J. Lau, T. H. Lee, and S. S. Wong. "A physical model for planar spiral inductors on silicon." *International Electron Devices Meeting Technical Digest*, pp. 155–158, Dec. 1996.

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.