Appears in the proceedings of the. Tech. Digest of the GaAsIC Symp. Nov. 1996, pp. 127-132.

Complementary GaAs Technology for a GHz Microprocessor

Richard B. Brown, Todd D. Basso, Phiroze N. Parakh, Spencer M. Gold,
Claude R. Gauthier, Ronald J. Lomax, and Trevor N. Mudge

Electrical Engineering and Computer Science Dept.
University of Michigan
1301 Beal Avenue
Ann Arbor, MI 48109-2122, USA

Abstract—A DARPA-funded project at the University of Michigan has as a goal the development of technologies and tools needed to implement microprocessors that can be clocked at GHz speeds. A Complementary GaAs HIGFET technology from the Motorola CS-1 facility (CGaAs) is the target semiconductor process. While this technology is immature, it is years ahead of CMOS in terms of fast gate delay at low power supply voltages. A major focus of this work is advanced packaging, which supports partitioning of the design into multiple integrated circuits, each having an integration level that should be achievable in CGaAs. This paper touches on the major aspects of the project, process technology, circuit design, packaging, architecture, CAD tools and software, with an emphasis on application of the CGaAs technology.

I. Introduction

As the 25th anniversary of the development of the microprocessor is observed in 1996, there is no denying the impact microprocessors have had on both the scientific world and on our personal lives. The remarkable decrease in cost vs. performance for microprocessors has made computing ubiquitous in our society. Through their astonishing advances in performance, microprocessors have replaced traditional mainframes and supercomputers with workstations and servers [1].

Current processor trends can be seen by surveying the microprocessors presented at ISSCC over the past ten years: gate delays have decreased at 12% per year; clock frequencies have increased at 40% per year; and transistor counts have grown at 40% per year [2]. The performance of systems made from these processors, as measured by the integer SPEC benchmarks, has been growing at 59% per year, resulting in an increase in computing power of more than 100 fold over the decade.

The disparity between gate delay and clock frequency improvement is accounted for by noting that some of the additional transistors have been spent to pipeline the processors, reducing the number of gates between latches, and to add on-chip cache, reducing the latency to memory. Additional pipeline depth provides diminishing performance returns, though, and cannot be expected to maintain the steep increase in clock frequency.

To keep new processors on the performance curve, architects next used their growing transistor budget to increase parallelism in the form of multiple functional units and concurrent instruction execution (superscalar architectures). Unfortunately, the benefits of parallelism also diminish with scale in general purpose machines. For example, the designers of the Intel Pentium Pro reduced the initial instruction-issue width of four to three in order to reduce chip size and complexity and enable an increase in the initial clock frequency from 100 to 150 MHz [3]. A general industry trend at the 1995 Hot Chips conference toward simpler CPUs running with higher clock frequencies was noted in [4].

To summarize the state of high-end microprocessors, much of the performance gain in the past decade has derived from increased pipelining and multiple instruction issue, but staying on the performance curve will be increasingly more difficult, as existing processors have already taken pipelining and superscalar to their points of diminishing return. Thus, technology improvements must provide more of the clock speed improvement which will lead to increased performance in future computers. Scaling of CMOS to channel lengths of 0.35 µm, together with planarization techniques which can produce five levels of fine-pitch metal interconnect have pushed the clock speed of commercial CMOS processors to as high as 500 MHz [5]. A number of leading semiconductor manufacturers will offer 0.25-µm CMOS processes in the first half of 1997. The importance of semiconductor technology to future high-end computer performance warrants evaluation of exotic processes such as complementary GaAs.

Another justification for exploring digital CGaAs is the growing portable (battery-powered) computer market, which could benefit from complementary GaAs because of its speed at low power supply voltages. The state-of-the art in CMOS low-power processors is exemplified by the IBM 401GF, which consumes just 40 mW (typical) from a 2.5-V supply at 25 MHz; 100-MHz parts are expected before the end of 1996 [6]. CGaAs, operating with a 1-V supply, would have a dynamic-power advantage of 1:6.25 over this 2.5-V CMOS process if parasitic capacitances were equivalent.

II. Complementary GaAs

CGaAs overcomes many of the problems of E/D MESFET DCFL technology, in which we have designed several prototype microprocessors [7, 8, 9]. Digital circuits are less efficient in DCFL than CMOS because the MESFET source resistance prohibits the stacking of enhancement devices, resulting in a NOR-only logic family. Furthermore, a large off-state leakage current limits the fan-in of high-speed gates to four. The Schottky diode at the MESFET gate becomes forward biased when the input is a logic ONE, clamping the input voltage (which would otherwise be at the power-supply voltage of 2.0 V) to about 0.7 V. This characteristic of DCFL reduces noise margins, causes IR drops on signal lines, makes pass-gate logic difficult to design, and requires superbuffer drivers for many signals. Two positive effects of the gate diodes are that the circuits have a current-steering nature which keeps internally-generated noise low compared to that in complementary circuits, and the small logic swings enhance switching speed. Because the transistors must be ratioed, layout area is greater than would be expected from the simple unipolar logic forms of DCFL. The most difficult challenge in implementing an E/D MESFET microprocessor is on-chip memory design. The current-leakage problem and lack of a complementary pull-up device make these circuits large and power-hungry compared to CMOS memories.

A complementary heterostructure-insulated-gate FET process, on the other hand [10], has many of the characteristics of CMOS. Though it, too, has a Schottky-diode gate, the forward voltage of the heterostructure diode is large enough to make gate current negligible under static operating conditions. Like CMOS, CGaAs has the flexibility to implement unratioed complementary gates, pass-gates, and domino gates, as well as the ratioed *p*-loaded DCFL, including stacked-transistor NAND structures. These logic families can be mixed on an integrated circuit, giving the designer a range of speed and power from complementary to DCFL. The power and speed can also be adjusted by varying the power supply voltage from below 0.9 V to 1.5 V. One can also realize source-coupled logic in the CGaAs process, but we have not used it because of its higher power dissipation and the greater interconnect requirements of differential logic.

Memories implemented in CGaAs are similar to CMOS memories. The layout density is limited by transistor and metal design rules, rather than by an area/power trade-off as in DCFL. The low power supply voltage, however, does make the sense amplifier design challenging. Because the transconductance ratio of *p*- to *n*-devices is 1:4 in CGaAs, superbuffers are needed to drive large loads.

There are several interesting differences between CGaAs and CMOS. In CGaAs, there is no need to contact the well or substrate, and latchup is not a possibility. Drains of *p*- and *n*-transistors can be abutted and can share an ohmic contact.

The major challenges in CGaAs seem to be developmental, rather than fundamental. The process is not yet mature. Both the process and device models are still being tuned. Yield, integration levels, metallization, and scaling all need attention. The University of Michigan project provides feedback to Motorola regarding changes that are needed to better suit CGaAs to digital VLSI applications.

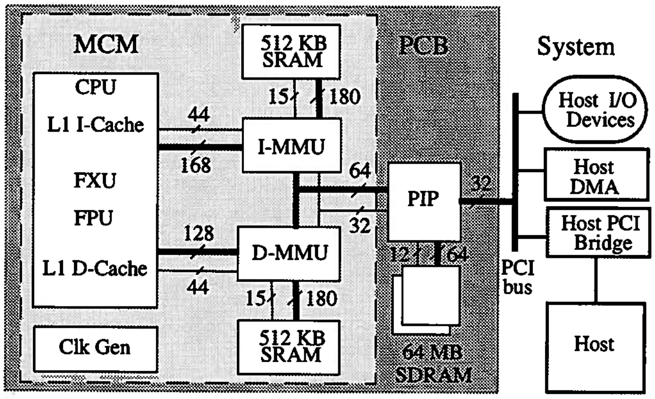

Fig. 1. Diagram of proposed prototype computer, showing microprocessor on MCM, mounted on PC Board, which plugs into host chassis.

III. Microprocessor Project

In addition to semiconductor process technology, which has been described, the UM GaAs microprocessor project, called PUMA, includes architecture, operating system, compiler, packaging, CAD tool and circuit design aspects.

A. Architecture

The general architecture of the PUMA processor follows that of the PowerPC [11], but we view the fabrication process as the performance driver, so the instruction set and microarchitecture are being tailored to fit the CGaAs technology. As seen in Fig. 1, the limited integration levels of CGaAs force a partitioning of the processor onto three VLSI chips and a small clock generator chip. The CPU chip is to include both integer (FXU) and floating-point (FPU) processors, as well as first level instruction and data caches; this circuit will require approximately one million transistors. Separate memory management chips for the instruction and data streams will interface the processor to BiCMOS level-two caches, and through a PCI interface chip (PIP), to synchronous silicon DRAM and to the host computer's memory and I/O.

Architectural work in this project emphasizes high bandwidth interconnect to memory, which is supported by the advanced packaging scheme described below. The characteristics of CGaAs (short gate delay and comparatively low integration levels) dictate that the datapaths and control be simple. Since the integration level will not support much parallelism, the processor's performance must come from high clock frequency, which argues for a deep pipeline (currently at 9 stages). Deep pipelines require good branch prediction because of the high performance cost of flushing the pipe on mispredicted branches.

All microarchitectural decisions are made based on results from a trace-driven simulator which takes as inputs, streams of instruction- and data-references generated from benchmark programs, and produces plots of performance loss in terms of cycles per instruction (CPI) attributable to various components of the architecture (e.g., cache misses, branch mispre-

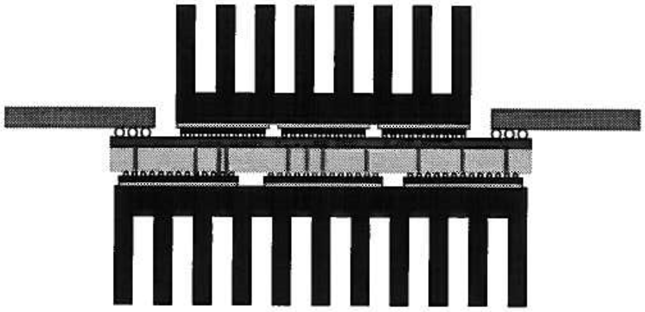

Fig. 2. Cross-section of packaging scheme, showing flip-chip mounted CGaAs ICs on top of MCM and vias through the substrate to silicon memory chips mounted on the back side of the MCM.

diction, data-dependency pipeline stalls, etc.). At this point in time, the simulator has been used to model four major configurations of the FXU pipeline and three versions of the FPU, together with a baseline MMU model. As microarchitectural features such as the size of the branch target buffer or number of forwarding paths on the pipeline are varied, one can see the effect on overall CPI, and determine whether spending the hardware resources to implement a given feature is justified.

B. Software

Changes in the instruction set or microarchitectural features require modifications to the software compiler and operating system needed to build a working prototype. An example of the interdependence of software and architecture is the software prefetch capability being added to the PUMA processor and compiler in order to reduce the performance loss due to data- or instruction-cache misses.

C. Packaging

The primary objectives of the fine-pitch, flip-chip, array-pad packaging approach proposed in this project are to increase bandwidth to memory (as mentioned above) through providing high I/O counts, and to minimize interconnect length throughout the processor. ICs having peripheral pads can be flip-chip mounted, but like wire-bonded or TAB-mounted parts, they require wide on-chip power rails to assure stable voltages in the core, and they usually have long interconnect routes from the core to bonding pads. Conventional flip-chip array packaging methods redistribute signals from the periphery to an array of bumps covering the chip surface; while there are some advantages to these approaches, the power distribution situation has not been helped, and interconnect is made even longer.

Fig. 2 is a high-level cross-sectional view of our array-pad packaging approach, showing CGaAs circuits flip-chip mounted on the top side of the MCM, with a heat sink attached to the back of these chips. Signal I/O will be distributed across the chip on fine-pitch gold bumps, which will connect to receivers, drivers, and ESD protection modules which are sized for MCM interconnect loads, and distributed with the I/O bumps to minimize interconnect length between the points of signal origination or destination and the pads. Power pads will provide Vdd and Ground for each module, reducing the size of power rails. Additional thermal vias can be added as needed. One distinguishing feature of this approach is the ability to make vias through the MCM substrate to minimize interconnect length to the second-level cache, mounted on the back of the MCM; again, a heat sink will be attached to the back sides of the ICs. The MCM could be mounted to the printed circuit board with solder bumps, as shown, to maintain a high I/O count at the next level of packaging.

Expected benefits of this packaging approach are smaller integrated circuits for a given functionality, faster logic paths due to less total interconnect, lower power dissipation because of reduced parasitic capacitance, and a higher percentage of the interconnect on the MCM where signal propagation characteristics are better. CAD tool developments are also vital to the packaging strategy.

D. CAD Tools

The proposed packaging approach requires CAD developments in order for the tools to directly place and route the array pads. The integrated circuits are being designed with the Cascade Design Automation EPOCH [12] circuit compiler. This tool suite is being extended to enable it to distribute pads, pad-drivers and receivers in an array. In addition, a number of MCM-specific capabilities are being added to the tools. In a future version, EPOCH will be able to run static timing analysis across multiple chips on the MCM. A major packaging advance will come through a new ability to concurrently place integrated circuits on the MCM and modules within the integrated circuits, so as to minimize the total interconnect length, especially on critical nets.

Our IC design flow begins with a circuit description written in the Verilog hardware description language. As pioneers in VLSI digital CGaAs, we have designed and modeled our own cell library, including complementary, Domino and *p*-loaded DCFL circuits. The initial layout was done in Mentor's IC Station, after which the cells were transferred to the EPOCH database using MasterPort, a Cascade tool which translates physical layout to new design rules. Because the CGaAs design rules are still changing, the ability to efficiently regenerate the cell library has been extremely valuable. EPOCH compiles, synthesizes, places and routes the circuits automatically from the Verilog description, but its floorplanner allows user interaction, which we employ extensively. Electrical and design-rule checks are then run on the full physical design, and a gate-level Verilog model generated by EPOCH is verified to have behavior equivalent to the original behavioral Verilog chip specification.

E. Circuit Design

Using the design methodology described, together with preliminary device models and architectural specifications, our

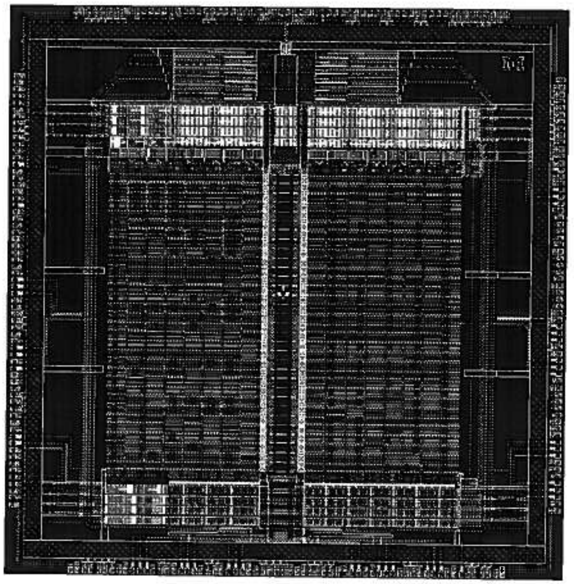

Fig. 3. CGaAs SRAM configured as 4KB data cache with tags.

circuit design group has been able to evaluate both the technology and the proposed architecture by designing and simulating major microprocessor circuit blocks. This work has included a comparison of complementary, unipolar, and dynamic logic styles, and optimization of both leaf cells and complex modules such as adders and shifters in CGaAs.

Fig. 3 is the layout of a CGaAs SRAM test chip designed full-custom with projected design rules and device models. Our SRAM design experience was a major influence in our recommending that for digital CGaAs, the threshold voltages be reduced. This circuit was designed in a 0.5 µm process, assuming +/- 0.3 V thresholds. Based on the projected model, which has not yet been realized, the embedable core of this 4KB (plus 1KB tags) RAM would have 0.97 ns access time, 1.14 ns cycle time, and about 1.2 W power dissipation.

Fig. 4 is the layout of a PowerPC ALU test chip realized in CGaAs, and based on the same projected design rules and device models. A major objective of this design was to explore Domino CGaAs logic. The ALU design provided important feedback to the architecture group. The critical path of 1.8 ns was found in the rotate-and-mask unit. The next longest path (adder) had a critical path of 750 ps. Since the rotate-and-mask unit is required by only three instructions, the processor performance can be improved dramatically by employing an architectural- or compiler-based approach for executing those instructions, so that all of the other instructions can run much faster. With buffering delays and setup and hold times, the maximum simulated frequency for the ALU based on the projected model would be 617 MHz, and the dynamic execution unit would dissipate about 1 W.

IV Summary

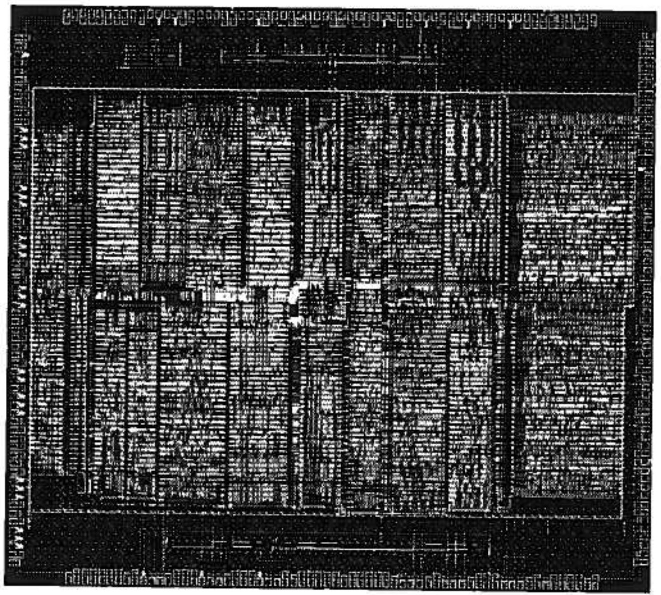

Fig. 4. CGaAs PowerPC ALU in Domino logic.

The University of Michigan PUMA project is helping to evaluate and explore CGaAs for VLSI digital applications. CGaAs requires further scaling, reduced thresholds, higher integration levels, and at least one additional metal layer to meet the requirements of high-performance digital circuits.

Advanced packaging and CAD capabilities, under development in this project, will also contribute to the technology base needed to realize microprocessors operating at GHz clock frequencies.

Acknowledgment

This work is funded by the U.S. Defense Advanced Research Projects Agency, under ARO DAAH04-94-G-0327.

References

[1] Linley Gwennap, "The Death of the Superprocessor," *Microprocessor Report*, vol. 9, no. 13, p. 3, October 2, 1995.

[2] Michael D. Upton, *Architectural Trade-offs in a Latency Tolerant Gallium Arsenide Microprocessor,* Ph.D. Dissertation, University of Michigan, Ann Arbor, 1996.

[3] David B. Papworth, "Tuning the Pentium Pro Microarchitecture," *IEEE Micro*, vol. 16, no. 2, pp. 8–15, April 1996.

[4] Ron Wilson, "Superscalar sunset seen at Hot Chips?" *Electronic Engineering Times*, issue 862, pp. 1 and 108, August 21, 1995.

[5] Linley Gwennap, "Digital's 21164 Reaches 500 MHz," *Microprocessor Report*, vol. 10, no. 9, pp. 1 and 12, July 8, 1996.

[6] Jim Turley, "401GF is Coolest, Cheapest PowerPC," *Microprocessor Report*, vol. 10, no. 8, pp. 9–10, June 17, 1996.

[7] R. Brown, et al., "Gallium-Arsenide Process Evaluation Based on a RISC Microprocessor Example," *IEEE Journal of Solid-State Circuits*, vol. 28, no. 10, pp. 1030–1037, October 1993.

[8] M. Upton, et al., "A 160,000 transistor GaAs microprocessor," ISSCC Dig. Tech. Papers, vol. 36, Feb. 1993, pp. 92–93.

[9] M. Upton, T. Huff, T. Mudge, R. Brown, "Resource Allocation in a High Clock Rate Microprocessor," ASPLOS, San Jose, CA, Oct. 4–7, 1994, pp. 98–109.

[10] B. Bernhardt, et al., "Complementary GaAs (CGaAs^{TM}): A High Performance BiCMOS Alternative," GaAs IC Symposium, San Diego, CA, Oct. 29–Nov. 1, 1995, pp. 18–21.

[11] Shlomo Weiss and James E. Smith, Power and PowerPC: Principles, Architecture, Implementation, Morgan Kaufmann, San Francisco, 1994.

[12] Cascade Design Automation, *EPOCH User's Manual, ver. 3.2*, Bellevue, WA 98006, 1995.