SPACE COMPUTER CORPORATION
2800 OLYMPIC BLVD., SUITE 104 • SANTA MONICA, CALIFORNIA 90404-4119 • FAX: (213)829-1694
(213)829-7733

HYBRID WAFER-SCALE PROCESSOR

FINAL REPORT

Small Business Innovative Research Topic No. 87-3

MARCH 14, 1989

By

Space Computer Corporation
2800 Olympic Blvd., Suite 104
Santa Monica, Ca
90404-4119

Submitted to:

U.S. Army Research Office
Contract No. DAAL03-88-C-0007

Approved for Public Release;
Distribution Unlimited

SUMMARY

The basic goal of this project is to develop and demonstrate techniques for the reduction of power consumption of space-based processors for infrared surveillance systems. The primary technique is to minimize the capacitive loading encountered in off-chip communications for highly concurrent processing architectures. Both processing architecture and chip packaging are simultaneously considered to maximize MOPS per watt by increasing throughput while reducing system capacitance, signal delay, noise, voltage swing, and power consumption (the costs of system communications).

With conventional packaging technology, highly concurrent processing architectures result in hardware implementations that are extremely large, very heavy, and that consume excessive power. Monolithic wafer-scale integration is theoretically ideal but requires an extensive amount of redundant circuitry and provisions for circuit reconstructurability because of manufacturing yield problems. In the hybrid wafer-scale integration (HWSI) approach, individual pre-tested chips are bonded to a fine-line interconnect structure fabricated on the surface of a wafer-scale substrate. With this technique, high yields can be achieved without redundancy.

During Phase I we determined that it should be possible to achieve major reductions in size, weight, and power through the use of hybri.. wafer-scale integration techniques for interconnect and packaging. Alternative interconnect and packaging approaches were considered and a conceptual design (electrical and mechanical) has been established for a miniaturized processor. While alternative processing architectures were considered, fine grain architectures like the Associative String Processor were our initial desired approach. However, neither chips nor software support are currently available and there is no indication when they might become available. Consequently, a medium grain architecture, for which both chips and software support are immediately available, was selected for a proof of principle processor.

While reduction of power consumption is the primary goal, the packaging technique also leads to opportunities for miniaturization and significant weight reduction. The degree of miniaturization that can be achieved permits processor radiation shielding schemes that would otherwise be impractical from a weight consideration. With this additional shielding, requirements for radiation hardened chips could be relaxed. The reduction of weight is not only realized in the processor, but also in the equipment required to generate, regulate, distribute, and dissipate the electrical power consumed. Finally, hybrid wafer-scale integration as a packaging technique, allows new chips to be developed with greater processing density because of the disproportionate chip area currently being devoted to bonding pads and pad drivers.

TABLE OF CONTENTS

LIST OF FIGURES

1. INTRODUCTION

Signal and data processors for infrared surveillance systems must employ massively-parallel architectures with large numbers of processing elements in order to achieve the high throughputs required (up to billions of operations per second or more). With conventional technology, the hardware required to implement these architectures is extremely large, very heavy and consumes large amounts of power. For example, the signal processor for the Army's Airborne Optical Adjunct infrared sensor, which has a throughput on the order of ten billion operations per second, weighs thousands of pounds and consumes tens of kilowatts of power.

In order for space-based sensors with onboard processing to be practical, it will be necessary to miniaturize the computing hardware by a very substantial amount. The purpose of the proposed research effort is to develop and demonstrate concepts and techniques to accomplish this objective. Specifically, the goal is to reduce the size and weight of parallel processor hardware by factors approaching an order of magnitude or more from values achievable with conventional packaging and integrated circuit technology.

It is important to note that miniaturization of the processor hardware must be accompanied by a substantial reduction in the power consumed by the processor circuitry. There are two basic reasons for this requirement. The first reason is that the maximum power density allowable within the processor is limited by thermal considerations such as the maximum allowable temperature of semiconductor junctions and the limitations of available cooling techniques. The second reason involves consideration of the weight required to generate, regulate, distribute and dissipate the electrical power consumed by the processor in addition to the weight of the processor itself (since the overall cost driver for the system is the total throw weight which must be boosted into orbit). Even with advanced solar-cell technology, the weight of the power system plus the weight of heat sinks and heat radiators is approximately 0.5 or more pounds per watt of power consumed. This is comparable to or greater than the weight of the processor hardware itself. Current spacecraft computers, for example, have weight-to-power ratios in the range of 0.1 to 0.5 pounds per watt, and future miniaturized hardware will have even lower ratios.

From a fundamental point of view, the large size, weight and power required for a high-throughput parallel processing system result from the high cost of communication (relative to logic and storage), in terms of chip and board area as well as power consumption. As an example, the transistor switches on the chips rarely use more than five percent of the available silicon area, which in turn is only about one percent of the total board area. Furthermore, the power consumption is dominated by the power required to drive the huge parasitic capacitances of the chip packages and the chip-to-chip

connection lines. It follows that the basic approach to the reduction of size, weight and power should be:

a. Eliminate individual IC packages;

b. Shrink system dimensions;

c. Minimize system communications.

The net effect of this approach is to maximize performance in MOPS per watt.

2. PHASE I TECHICAL OBJECTIVES

The technical objectives for our original Phase I proposal (dated January, 1987), entitled "Passive Ranging with Electro-Optical Sensors", involved two basic tasks:

a. Development and evaluation of algorithms based upon the velocity filter approach to target detection and tracking;
b. Study of processor implementation for execution of these algorithms (including architecture and Phase II brassboard feasibility).

By the time we were awarded this Phase I contract, we had already essentially covered task (a). We therefore emphasized task (b) during this Phase I effort. This emphasis was noted in our first Phase I monthly progress report. Our revised Phase I technical objectives have been:

a. Identify parallel processing architectures (and their VLSI implementations) that are designed for extremely high throughput and determine their suitability for our processor concept;
b. Identify and compare alternative interconnect and packaging approaches for these architectures and IC implementations, including configurations, materials, and fabrication processes;
c. Establish a conceptual design (electrical and mechanical) for a miniaturized processor, targeted for space-based applications, that exhibits an extremely high MOPS/watt figure of merit;
d. Define a Phase II approach that would prove the feasibility of achieving a substantial MOPS/watt improvenment using hybrid wafer scale integration techniques.

3. TECHNICAL APPROACH

The fundamental limitation in a parallel processing architecture implemented with VLSI is the high cost of communication relative to logic and storage [1]. This high cost is manifested in terms of area, power, and performance. Most of the chip and board area required for the implementation of a parallel processor is in fact devoted to communication. In the case of the internal circuitry of the chips themselves, for example, the transistor switches rarely use more than 5 percent of the available silicon area. Communication is also expensive in sending signals between chips, where large areas are used for bonding pads, pad drivers and packages as well as for printed-circuit board wiring.

Dynamic power dissipated in the circuits that switch capacitive signal loads is typically dominated by the parasitic capacitance of the internal wires, bonding pads, and chip-to-chip connection lines rather than by the capacitance of the transistor gates. For VLSI technologies such as CMOS, in which the static power is negligible, communication thus accounts for most of the power consumed and dissipated by the chips.

Communication is also expensive in terms of delay, both internally within a chip as well as between chips. In MOS technologies, which exhibit the highest circuit densities but a poor relationship between transistor driving capabilities and the wiring parasitics, circuit speeds are dominated by parasitic wiring capacitance. In fact, the disparity between internal signal energies and the macroscopic world of bonding pads, package pins and interchip wiring is so large that the delay penalty in amplifying a signal so that it can run between chips can be comparable to a clock period.

The above area, power, and performance costs of communication translate directly into size, weight, and power costs for a parallel processor with a specified throughput. However, these costs can be reduced by (1) choosing architectures in which communication is localized as much as possible, and (2) by reducing system capacitances through the elimination of individual IC packages and the use of short, fine-line, high-density, chip-to-chip interconnects similar to those used on the ICs themselves.

There are two basic approaches to elimination of individual IC packages and down-sizing of the chip interconnects: monolithic wafer-scale integration (MWSI) and hybrid wafer-scale integration (HWSI). The monolithic approach is theoretically superior in many respects, but requires an extensive amount of redundant circuitry and provisions for circuit reconstructurability because of manufacturing yield problems. In the hybrid approach, where individual pretested chips are bonded to an interconnect structure fabricated on the surface of a wafer-scale substrate, high yields can be achieved without redundancy. Furthermore, levels of performance and circuit density can be

achieved which are comparable to or even greater than those obtainable with the monolithic approach.

Table 1 below shows a comparison of wafer-scale integration and conventional packaging approaches based upon recently published studies:

Table 1. Comparison of Interconnect and Packaging Approaches (Derived from References [4-6]).

APPROACH	Power X Delay (Normalized)	Size or Weight (Normalized)
Printed-Wiring Board	1.00	1.00
Thick-Film Hybrid	1.08	0.42
Co-fired Multilayer Ceramic	0.34	0.20
Wafer-Scale Integration		
Monolithic	0.10	0.09
3-D Hybrid	0.08	0.07

The use of HWSI leads directly to the reduction of volume and weight by approximately an order of magnitude. It also leads directly to a significant degree of power reduction in the case of CMOS circuitry (which is preferred for space applications because of its low static power dissipation as well as for other reason.) For such circuitry, the chip power consumption P is given by the following formula:

$$P = P_{INT} + N(1/2\ C_L\ V^2\ f)$$

where P_{INT} is the internal power dissipated in the chips, N is the number of output drivers on the chip, C_L is the average load capacitance per driver, V is the voltage swing and f is the clock frequency. As an example, consider a typical CMOS chip with N = 50, P_{INT}=0.1 watts, V = 5 volts, and f = 20 MHz. With conventional packaging, the value of C_L due to the package and (say) 5 inches of board wiring is about 150 pF, giving a power dissipation for the chip of 1.9 watts. If the package is eliminated and the length of the conductor line is reduced to (say) one inch, the value of C_L will be reduced to about 10 pF and the power dissipation will be reduced to 0.23 watts. This is almost an order of magnitude decrease.

The use of HWSI may permit additional reduction in power through the reduction of current switching (delta-I) spikes and other forms of electrical noise. Reduction of noise level should permit a reduction in the required voltage swing V, which in turn will reduce the value of P.

In summary, our basic approach is to exploit the advantages of hybrid wafer scale integration to shrink system dimensions through the elimination of individual IC packages, and to reduce power requirements through fine line interconnects thereby minimizing the capacitive loading on the inter-chip communication lines. To maximize these advantages in a high performance parallel processor, requires that careful consideration be given to the overall architecture especially as it pertains to communication.

4. ARCHITECTURE AND VLSI IMPLEMENTATION

Many schemes have been devised to interconnect processors to achieve increased performance through parallel or concurrent operations. With the advent of VLSI technology, parallel processing architectures have been devised on several levels. On one level, the processing architecture consists of a collection of chips that are interconnected in some manner to provide a processing node. In turn, multiple nodes may be interconnected to provide the parallel processing environment. On another level, a single chip itself might contain the parallel processing architecture consisting of numerous processing nodes that can be expanded further through the integration of additional, similar chips.

Single chip parallel processing architectures tend to be regular in nature; that is, each node is typically identical and the interconnection on chip, and between chips, tend to be highly symmetrical. In contrast, medium grain parallel processing architectures while normally exhibiting a high degree of symmetry between nodes, often have very little symmetry in the chip to chip communication lines within a node.

4.1 Communication Considerations

Nonsymmetrical or random communication schemes are more difficult to implement since conflicts in communication paths are often encountered. Implementations that use multilayer communication planes relieve this situation somewhat, but not all path lengths can be optimized to be as short as possible (and ideally, every communication path in the architecture should be). Hence, some path lengths have to be increased and thus processing delays and additional capacitive loading are introduced. However, with hybrid wafer scale integration, the additional capacitive loading introduced will be much less than if packaged chips and printed circuit boards were involved. The processing delay is still there, but this is a trade-off against (1) less complex nodes (fine grain) where there are less routing conflicts within the node and hence the potential for shorter path lengths, but (2) less capability at each node and hence the need for more nodes which can give rise to situations where an internode communication path length is increased. (The latter would not be true in architectures that only allowed communication with nearest neighboring nodes.)

4.2 Algorithm Considerations

Obviously, the choice of an architecture for a spaced-based surveillance processor is highly dependent on the nature of algorithms and overall throughput requirements. For example, the processing steps required in the

velocity filter approach to target detection and tracking are shown in Figure 1. An estimate of a processor's real-time throughput requirement (by processing step) is also shown in the figure. These estimates are based on the assumption that the sensor generates one million pixels per second. The estimate for the velocity filtering part is based on 1000 velocity filters. If 3000 filters were required, the estimate would increase to 6000 MOPS or 6 GOPS.

The types of computations associated with these processing steps are two dimensional FFT's and Inverse FFT's, complex array multiplication, shifting of arrays, and addition of arrays. While each individual computation is relatively simple, the requirement for high throughput arises because of the large number of pixels involved. Clearly, the stressing computations are associated with the velocity filter bank which could require billions of operations per second. But again, the computations only involve shifts and adds.

The algorithmic scheme shown in Figure 1 also uses frame-to-frame correlation techniques. Since millions of pixels are involved in each frame, the processing system would require a substantial amount of memory.

While the computational load for velocity filtering is demanding, its structure can be characterized as being very regular; that is, the algorithms involve little or no data dependent branching. Hence, this computational problem lends itself to pipelined and/or concurrent processing schemes.

Pipelining alone to achieve parallel operations is inappropriate for space-based processors. The reason is that if any node of the pipeline should fail, the entire system is inoperable (and space based systems have extremely long-term reliability requirements associated with them, such as 0.9 for periods of 10 years or more). Hence, parallel pipelines would be more appropriate for the example class of algorithm.

4.3 Reliability Considerations

Parallel processing architectures of interest during the Phase I effort included both medium grain and fine grain. A desirable characteristic of a parallel architecture for space-based applications is that it consist of a network of identical processing nodes (of either granularity) interconnected in a redundant, fault-tolerant configuration. If the architecture is based on a concept of identical nodes, spare nodes are more easily provided to support the long term reliability issues. (Typically spare nodes equal in number to about 20 percent of the total number of nodes are needed to meet the reliability requirements.) Figure 2 shows a schematic diagram of such a parallel, fault-tolerant processing network.

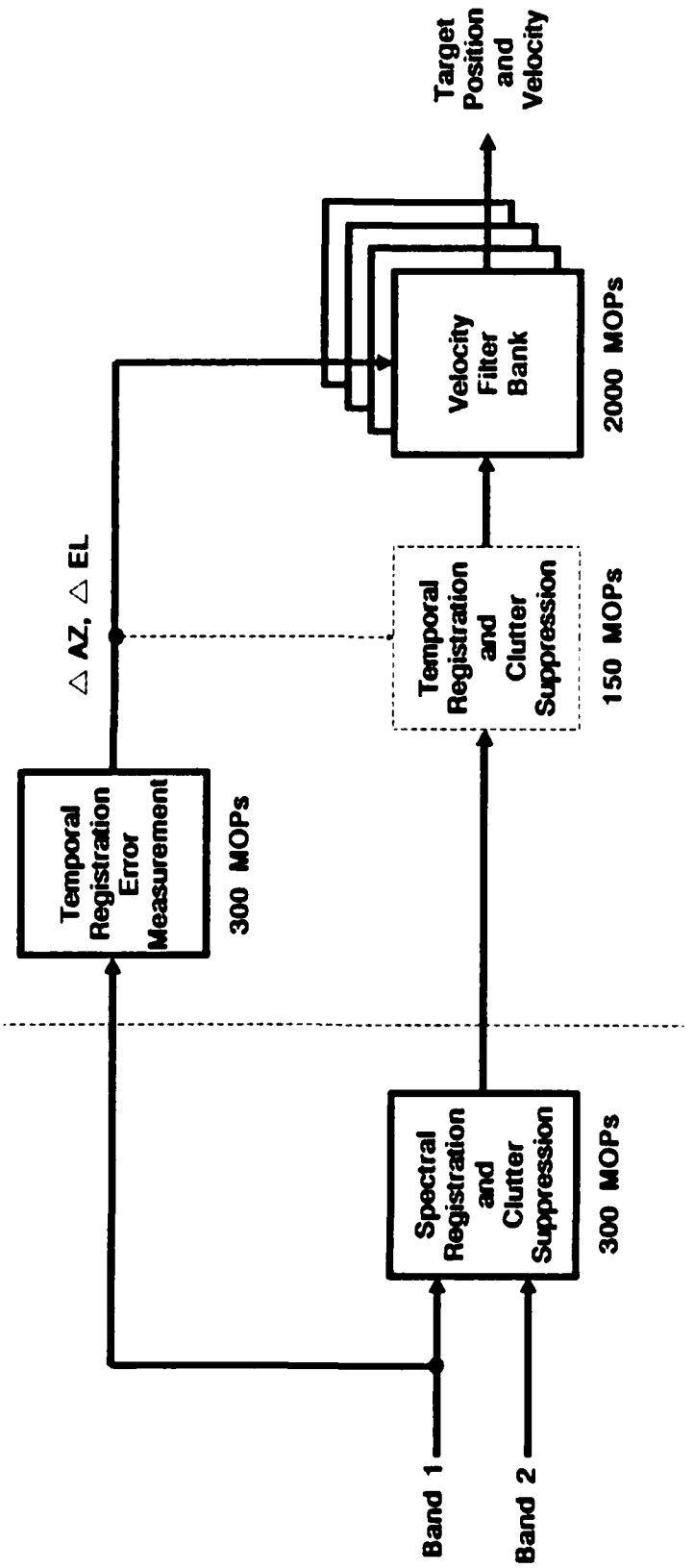


Figure 1. Velocity Filter Approach to Target Detection and Tracking.

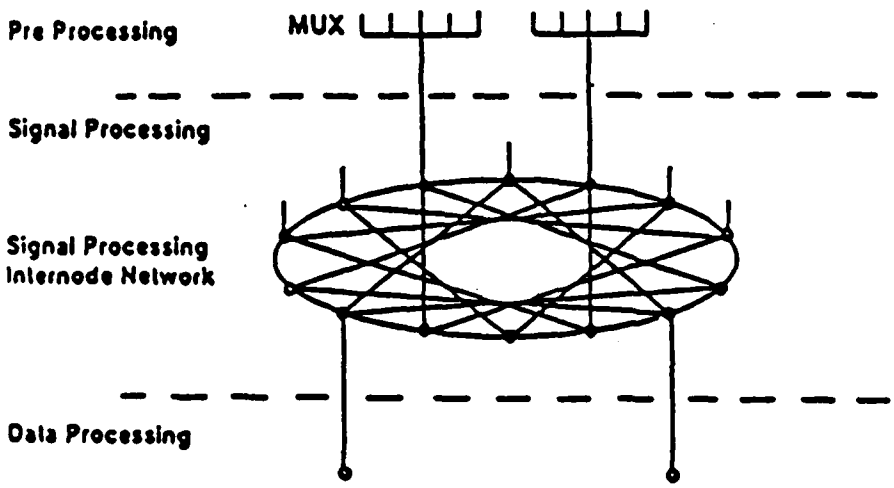


Figure 2. Parallel, Fault-Tolerant Processing Network.

4.4 VLSI Implementations

As part of this Phase I effort, a number of fine and medium grain parallel processing architectures were identified and evaluated as to their suitability for the velocity filter approach to target detection and tracking. Originally, emphasis was placed on the fine grain architectures since it appeared that the fine grain approach might be more suited to our objectives for low power, miniaturized processors. In particular, fine granularity supports the concept of minimizing chip to chip communications, assuming of course that a suitable number of processing elements could be implemented on a chip or wafer relative to the total number of processing elements required of the system. (Of course, this ratio would be unimportant if in the final implementation scheme, the architecture exhibited a loosely-coupled characteristic where the fine grain processing elements operated more or less autonomously and there was very little inter-processor communication required.)

Some of the architectures that were identified were specifically developed for image processing applications like the Hughes/University of Massachusetts Image Understanding Architecture(IUA), NASA's Massively Parallel Processor (MPP), and the MPP's refined version, the Blitzen Chip currently under development at the Microelectronics Center of North Carolina. Clearly, the functionality embodied in these architectures reflects insights into the requirements of image processing and it would appear that these architectures have direct relevance to our objectives. However, these architectures require significant adaptation for use in real-time applications, and in any case have not yet been implemented in a form available to us for our purposes and schedules.

Other fine grain architectures like the (Thinking Machines Inc) Connection Machine, the (Aspex Microsystems Ltd) Associative String Processor, and the (Active Memory Technology Inc) Distributed Array Processor, to name a few, were not necessarily developed for image processing but rather for a broader set of applications that also require high throughput. Yet, they might prove very useful to our specific application. For example, we identified a vendor of content addressable memory (Coherent Research Inc) who has initiated in-house research efforts to use associative devices in data compression techniques (where information is not lost) and in matrix operations where the matrix is sparse.

4.4.1 Blitzen (Microelectronics Center of North Carolina)

The BLITZEN chip is a fine grain, parallel processor with strong similarities to the Goodyear Aerospace Corporation's Massively Parallel Processor (MPP). It was specifically designed to provide high-throughput image processing for satellite applications. The custom VLSI chip contains 128 bit-serial Processing Elements (PE) in an 8 X 16 array with a 1024 X 1 Ram

associated with each PE. The PE's are interconnected on chip with an X-shaped grid that allows each PE to communicate with its eight nearest neighbors. This same interconnection scheme extends across chip boundaries so that an array of chips can be uniformly interconnected.

Unlike a pure SIMD machine where a single instruction is issued to all PE's (or if a memory operation is involved a single address is delivered to all PE's), the Blitzen chip has the ability to selectively turn off processing operations at individual PE's while allowing other PE's to perform. This masking feature allows processing to take place at a PE only if some condition is satisfied. This capability supports the high level IF - THEN conditional execution concept.

I/O operations are also not pure SIMD-like. Under program control, a specific column of PE's (one of 16 columns) can be selected for I/O; or through a broadcast command, all PE's on a given row can execute an input command. This structure is useful for array operations.

A prototype version of the chip has been fabricated and is in early test.

4.4.2 Connection Machine (Thinking Machines Corp)

The Connection Machine is a fine grain, parallel processor. It is implemented with a proprietary VLSI chip containing 16 bit-serial PE's. The CM integrates 4096 chips for a total of 65,536 PE's. Each PE has 65,636 X 1 RAM associated with it. Interprocessor network communication is flexible with support for 1-D to 16-D nearest neighbor links (order-1 to order-16 hypercube configurations). In its current configuration, the Connection Machine is extremely large and heavy, with high power consumption. Furthermore, its proprietary chips are not available to others.

We have recently learned that a desktop parallel processor similar in certain respects to the Connection Machine is under development by MasPar, Inc. in Santa Clara, California. This processor utilizes proprietary VLSI chips currently in the early stages of prototype fabrication.

4.4.3 Associative String Processor (Aspex Microsystems LTD.)

The ASP is a parallel processing computational structure consisting of a string of identical associative processing elements and an inter-processor communication network. The parallel processing elements of the ASP system are bit-serial devices, each with a small amount of associative memory. Since the circuitry of each processing element can be made extremely simple, a very large number of them can be fabricated on a single chip. (Current designs employ 64 processing elements per chip, while future designs may have as many as 1024.) The massive parallelism more than compensates for the loss of speed incurred by the use of serial arithmetic. The ASP is of particular interest because of its low overhead control and data communications techniques, combined with the elegant manner in which its architecture is matched to the

technological opportunites, as well as the constraints, of VLSI design and fabrication. However, neither chips nor software support are currently available.

4.4.4 Distributed Array Processor (Active Memory Technology Inc.)

The DAP is a fine grain, array processor based on a grid of proprietary, VLSI chips each containing 64 single-bit PE's in an 8 X 8 grid with a 32,768 X 1 RAM associated with each PE. Each PE has a dedicated connection to each of its 4 nearest neighbors and two bus interface connections to connect PE's by rows and columns. The proprietary chips are not available at this time.

4.4.5 Geometric Array Parallel Processor (NCR Corp.)

GAPP is a systolic array processor. Each GAPP chip contains 72 bit-serial PE's arranged in a 6 X 12 matrix with 128 X1 RAM associated with each PE. The PE's are connected in a nearest neighbor fashion to permit bidirectional, inter-processor communications in the north, south, east and west direction. The GAPP system integrates 32 chips for a total of 2,304 PE's arranged in a 48 X 48 array. GAPP chips are available, but software support is not extensive and equivalent throughput is limited for algorithms of interest here.

4.4.6 Image Understanding Architecture

The basic concept of this architecture is to provide processing arrays that are organized in hierarchical fashion. The top layer, called the General Purpose Processor Array (GPPA), is an 8 X 8 array of commercial, 32-bit microprpcessors (Motorola 68020's). Immediately below is a medium grain array of word-parallel, arithmetic oriented processors. This middle layer, called the Numeric Processor Array (NPA), contains 4096 Texas Instrument Digital Signal Processing chips arranged in a 64 X 64 array. The lowest level, called the Content Addressable Array Parallel Processor (CAAPP), is a 512 X 512 fine grain, content addressable array of bit serial processors for "processor per pixel" operations.

The fine grain (CAAPP) layer is based on a custom chip. Each chip contains 64 bit-serial PE's arranged in an 8 X 8 array with a 320 X 1 RAM associated with each PE. Routing circuitry allows communication with 4 nearest neighbors in the grid. A useful feature of the PE is an activity bit that controls whether or not a PE would actively respond to instructions. This would allow the capability to selectively process pixels. Each PE has two sources of associative feedback, a some/none response and a count response. This circuitry would be very useful for the rapid creation of histograms. For example, all (PE's) pixels having an intensity above some value would flag their some/none condition and the count responder circuitry would provide the number of pixels in this histogram bin. Although this feature has no critical value to velocity filtering, it could be useful in determining threshold

levels as part of the thresholding function. VLSI chips under development by Hughes Aircraft for the CAAPP layer are not yet available.

4.4.7 Microprocessors

There are many 32-bit microprocessors readily available, and relatively easy to program. Modern RISC like processors, including the forthcoming RH-32, exhibit excellent performance (on the order of 10 to 40 MIPS or more). Some of these are still in development, but there are also many commercially available like the R3000 from MIPS Computer Systems Inc., the Motorola 88000, and the Intel i860. However, one serious limitation that many of these microprocessors have, is that they were not designed, in general, to be interconnected in large arrays. The exception is the Inmos T800 transputer which was specifically designed for parallel processing since each transputer provides four, built-in, high speed I/O ports for interprocessor communication.

The role of the microprocessor in medium-grained node architectures is primarily to handle data routing and low-throughput computations. As shown in Figure 3, it will be used in conjunction with a co-processor or "slave" processor optimized for stressing computations such as those required for FFTs or velocity filters. A variety of such co-processors are currently available as so-called digital signal processors or vector signal processors. These devices utilize a form of fine-grained parallelism to achieve very high processing speeds. The Zoran 34161, vector signal processor, for example, has an equivalent throughput of about 100 MOPS for 2-D FFT computations. New devices currently under development will have even higher throughputs. For example, a 200-MOPS one-chip processor from Plessey is reported to calculate 1024-point FFTs in 97 microseconds. Other, more specialized, devices with throughputs up to 600 and 800 MOPS are also being reported.

5. INTERCONNECTION AND PACKAGING

It is generally recognized that a fundamental limitation in VLSI implementations is the high cost of communication (especially inter-chip communication) in terms of chip area devoted to output drivers, processing delays, and power consumption. A way around this is to implement entire processing systems on a single wafer so as to minimize the need for off-chip communications. This approach is known as monolithic wafer scale integration (MWSI). Unfortunately, MWSI technology has not matured to the point where the manufacturing yields are satisfactory. Until MWSI becomes practical, hybrid wafer scale integration (HWSI) is a viable alternative.

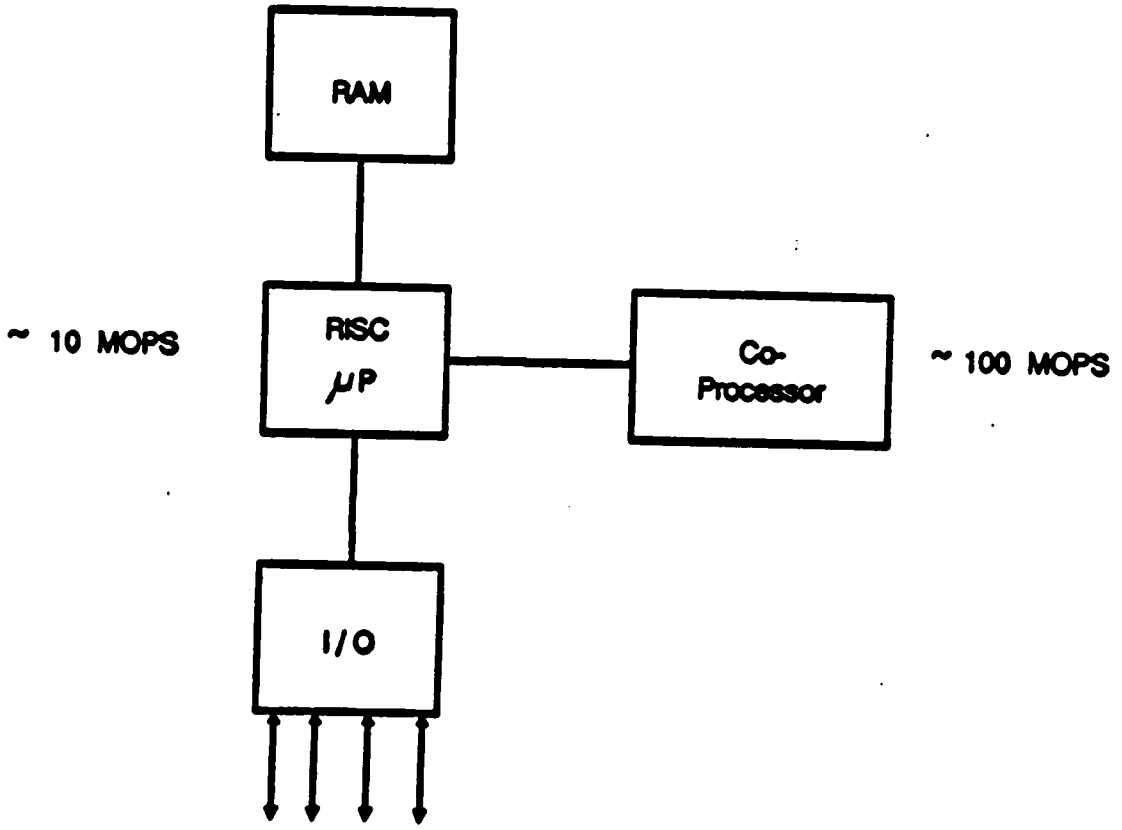


Figure 3. Medium-Grained Node Architecture.

5.1 General

Hybrid wafer scale integration refers to the concept of bonding pre-tested, unpackaged, integrated circuit chips to a substrate which also provides the interconnection scheme among the chips. The monolithic approach is theoretically superior in many respects, but requires an extensive amount of redundant circuitry and provisions for circuit reconstructurability because of manufacturing yield problems. In the hybrid approach, where individual pretested chips are bonded to an interconnect structure fabricated on the surface of a wafer-scale substrate, high yields can be achieved without redundancy. Furthermore, levels of performance and circuit density can be achieved which are comparable to or even greater than those obtainable with the monolithic approach.

5.2 Configurations

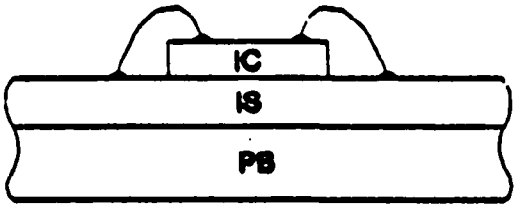
Figure 4 illustrates the major HWSI interconnect and packaging configurations currently under development. The basic HWSI elements shown are the integrated circuit chip (IC), the interconnect substrate (IS) upon which the multilayer interconnect structure is fabricated, the package base (PB) and the chip interconnect scheme (wire bond, solder bump or TAB).

Configuration (a) is the simplest and most commonly employed alternative. The ICs are die-bonded in a face-up position to a silicon or alumina interconnect substrate and electrically attached to the interconnect with traditional downhill wire bonding. A disadvantage of this configuration is low system assembly yield due to the difficulty of providing full functional and dynamic probe testing of bare IC chips prior to assembly. If such testing cannot be performed, then the "product of yields" effect dominates the overall yield of the entire asssembly process. To illustrate, assume that the probability of a die meeting its full performance specifications after successful completion of a probe test is 80 percent. If there are, say, 16 of these devices in the HWSI system, then the probability of all of them meeting their performance specifications is only 2 percent.

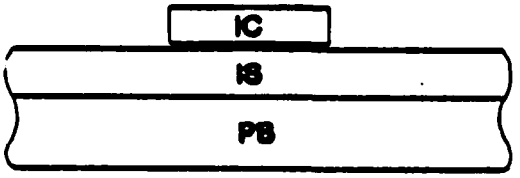
Configuration (b) is a flip-chip arrangement with face-down ICs and solder-bump chip attachment. Heat dissipation is a problem with this configuration, since the solder bumps must provide thermal as well as electrical conduction. Also, system yield is low for the same reason as with configuration (a).

Configuration (c) has a composite substrate structure consisting of:

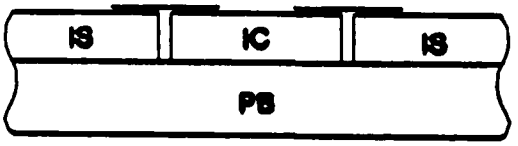
a. An interconnect substrate, which supports a thin-film, multilayer polyimide-metal interconnect structure deposited on its upper surface. This substrate has holes within which the chips are mounted.



(a) Chip Mounted on Interconnect Substrate with Downhill Wire Bonding



(b) Flip-Chip Arrangement with Solder Bump Bonding



(c) Chip Mounted within Die Cavity with Tape-Automated Bonding (TAB)

Figure 4. Alternative HWSI Configurations (IC = Integrated Circuit, IS = Interconnect Substrate, PB = Package Base).

b. A thermal substrate, which provides heat conduction as well as mechanical support for the chips. The material used for this substrate should have a very high thermal conductivity (as well as a thermal coefficient of expansion which closely matches that of silicon).

The use of this composite structure permits independent optimization of electrical, mechanical and thermal functions for both high performance and ease of manufacture. In particular, the use of planar TAB chip attachment provides exceptionally low-impedance lines as well as a capability for full functional and dynamic IC testing prior to assembly to achieve high fabrication yield.

The chips are located within die cavities formed by holes in the interconnect substrate and electrically connected to the interconnect pads by short, planar TAB leads. With the TAB process, the outer TAB leads are employed for IC testing after inner lead bonding of the chip. The TAB structure is then attached to the interconnect (after excising the surplus lead material) by outer lead bonding. In this manner, the chips can be tested to their full performance specifications to achieve yields as high as 99.5 percent prior to assembly. For a 16-chip system, this results in an overall HWSI yield of 92 percent.

The technology required for implementation of configuration (c) is currently under development by Space Computer Corporation.

5.2.1 Interconnect Substrate

The material used for the substrate upon which the multilayer interconnect structure is fabricated should have the following properties:

a. It should be capable of being lapped and polished to a high surface finish;
b. It should be chemically inert in order to remain stable after numerous etching and cleaning operations;
c. It should be capable of supporting the deposition of reliable, stable thin films with fine linewidths and a high degree of film-to-substrate adhesion.

For package configurations in which the chip is mounted within a die cavity, as in Figure 4 (c), an additional property is required: that of precision machineability. Such machineability is extremely difficult or impossible with ordinary ceramics such as alumina as well as with silicon (because of its anisotropic crystalline structure). However, it can readily be achieved with Fotoceram, a photosensitive glass-ceramic material manufactured by Corning Glass Works. This material can be machined with high precision to form die cavity holes, substrate boundaries and other features by an ultraviolet photolithography process. With this process, a sheet of the material is exposed to ultraviolet light through a mask, forming a 3-D latent

image through the entire thickness of the sheet. The image is then fixed by a controlled heat treatment, and the imaged material is removed by chemical etching. The process yields highly accurate parts which can be lapped and polished to a high surface finish. The Fotoceram surface provides a high degree of film-to-substrate adhesion for polyimide as well as aluminum, copper and other materials, and will support the reliable deposition of metal linewidths as narrow as 5 microns.

5.2.2 Multilayer Interconnect Structure

The multilayer interconnect structure is fabricated on the surface of the interconnect substrate. Typical materials employed are aluminum or copper for the conductors, and polyimide, benzocyclobutene (BCB) or polyphenylquinoxaline (PPQ) for the dielectric layers. Typical interconnect structure characteristics are listed below.

a. Conductor lines: width 10-25 microns, pitch 20-50 microns;
b. Number of conductor layers: 4 to 6;
c. Conductor thickness: 1 to 5 microns;
d. Dielectric layer thickness: 5 to 10 microns;
e. Dielectric constant: less than 3.
f. Conductor resistance (12 micron width): < 5 ohms/cm
g. Parasitic capacitance: < 2 pF/cm
h. Conductor inductance (12 micron width): < 4 nH/cm
i. Via resistance (12 micron width): < 0.01 ohms.

The cross section of a typical multilayer interconnect structure, including vias, is shown in Figure 5.

Fabrication of the multilayer interconnect structure requires clean-room facilities plus equipment for photolithography, polymer handling, wet and dry etching, sputtering, testing, inspection and rework/repair. Deposition of the dielectric and conductor layers is typically carried out by spin casting and sputtering, respectively. Excellent adhesion and environmental stability as well as high fabrication yields have been achieved through the use of stress reduction, bias sputtering and other thin-film design and processing techniques.

It is important to note that the multilayer interconnect can be designed in such a way that electrical noise effects are considerably reduced relative to conventional technology. Crosstalk between adjacent signal paths, for example, can be kept to a minimum through the use of metal planes under the signal lines. Furthermore, induced power and ground noise that results from many IC output drivers switching simultaneously can be reduced dramatically through the use of an integrated decoupling capacitor built into the interconnect structure. This noise is given by the formula

$$V_{noise} = N\ L\ di/dt$$

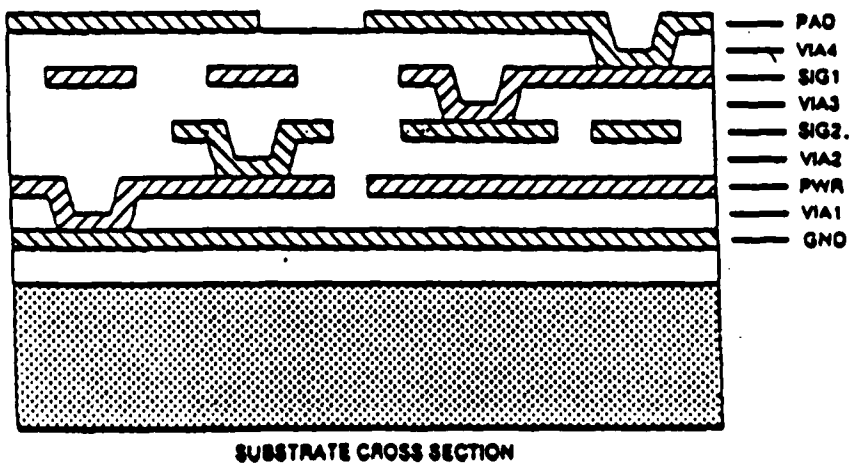


Figure 5. Cross Section of Typical Multilayer Interconnect Structure.

where N is the number of drivers, L is the inductance of the path from the IC pad to the voltage reference planes, and di/dt is the single driver switching current rate. With a silicon substrate, the decoupling capacitor structure can be fabricated directly on the silicon surface using a high-K material such as aluminum oxide, so that it is extremely close to the chip I/O pads.

Reduction of the noise voltage will permit a reduction in the voltage swing required for reliable operation of the ICs. This in turn will reduce the dynamic power, which varies as the square of the voltage swing.

5.2.3 Packaging

Aluminum nitride or silicon carbide are excellent materials for the thermal substrate or package base. These materials have thermal coefficients of expansion which closely match that of silicon over a considerable temperature range, as well as very high thermal conductivities (about an order of magnitude greater than that of alumina). Furthermore, they can be fabricated or integrated into hermetically-sealed packages incorporating large numbers of input-output leads.

5.2.4 Three-Dimensional Structures

Techiques for the fabrication of three-dimensional structures utilizing stacked substrates are under development by Space Computer Corporation. Figure 6 shows one concept utilizing a multilayer flex cable with "gold dot" connections. Another concept, shown in Figure 7 utilizes Fotoceram "button boards" and low-impedance wire "buttons" similar to those developed by TRW for VHSIC applications.

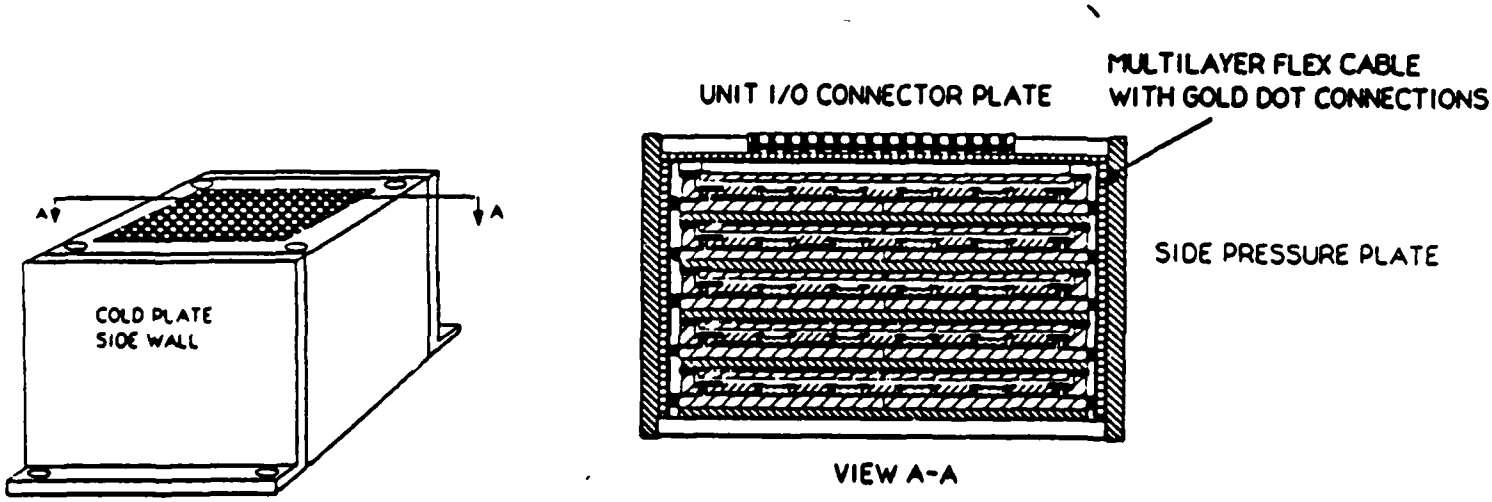


Figure 6. Three-Dimensional Packaging Concept with Gold Dot Flex Cable Connections.

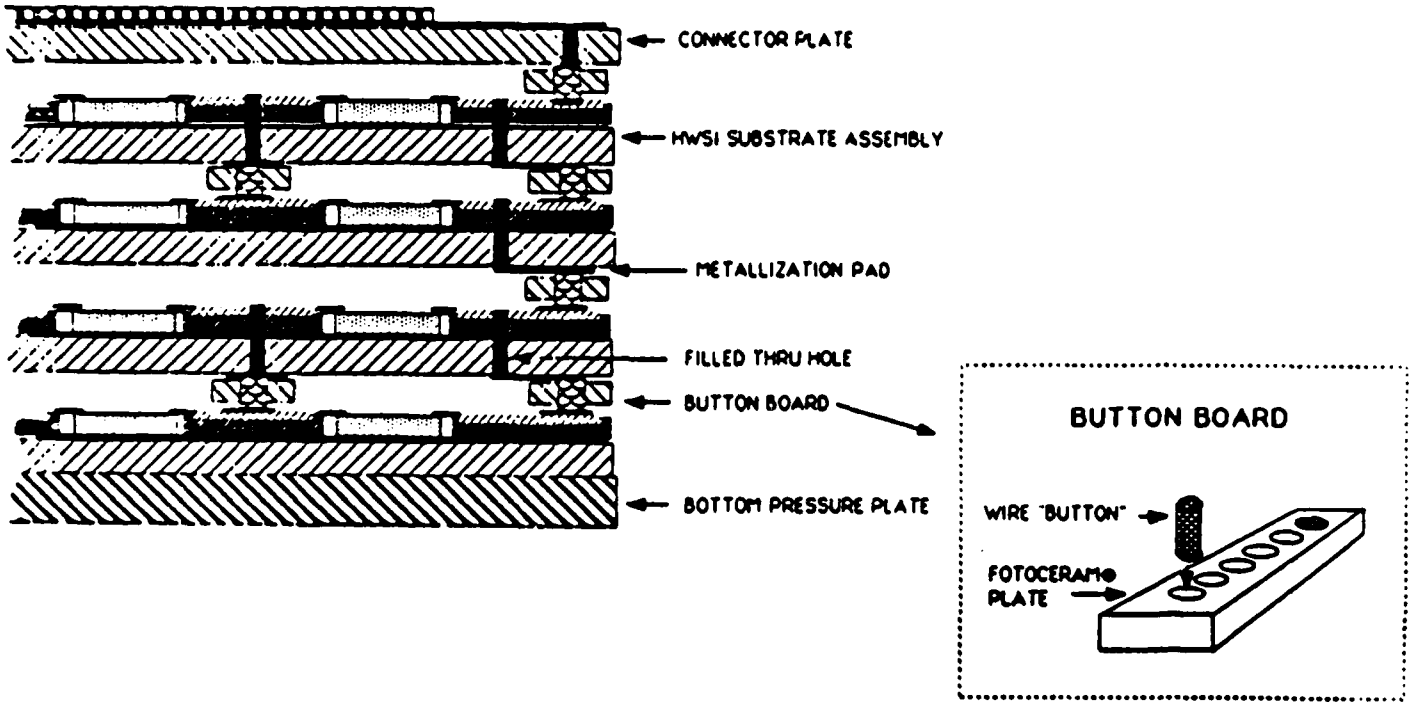


Figure 7. Three-Dimensional Packaging Concept with Button Boards.

6. DESIGN CONCEPT FOR 3-D HWSI PROCESSOR

We have developed a preliminary design concept for a parallel processor utilizing advanced CMOS VLSI chips, 3-D HWSI packaging and heat pipe cooling to illustrate possibilities for future low-power, miniaturized space equipment. The key features of the conceptual design for the 3-D HWSI Processor are (1) high throughput, low power, parallel processing modules that, (2) can be expandable in a modular fashion to match the throughput requirements of the particular application with, (3) integrated thermal management provisions for each module. Figure 8 illustrates the packaging concept for the processor.

As shown, each module consists of two substrates mounted back-to-back on a planar heat pipe. Thermal flow is from each module to the (secondary) side plate heat pipe and then ultimately down to the cold finger which connects to the space platform heat sink. The interconnected modules are stacked and enclosed within a hermetically-sealed package with external electrical connectors and cold-plate side walls. For processing modules implementing a medium grain parallel processing architecture, the characteristics of the assembly are estimated as follows:

a. Substrate size: 4 X 4 inches
b. No. of nodes per substrate: 8
c. No. of chips per substrate: 64
d. No. of substrates: 16
e. No. of modules: 8
f. Total no. of chips: 1024
g. Total power: 80 watts

Figure 9 shows a single 4 x 4 inch substrate with 64 chips. The dimensions of the assembly will be approximately 4.2 x 4.2 inches and the weight will be approximately 7 pounds. When enclosed in its hermetically-sealed package (with connector), the total dimensions and weight will be approximately 5 x 5 x 5 inches and 10 pounds, respectively.

The weight of a tantalum shield that would completely enclose a volume of 5 x 5 x 5 inches at a thickness of 0.25 inches to 0.35 inches is about 15 pounds to 21 pounds, respectively. By using a multi-layered, "shower curtain" shielding scheme of tantalum, aluminum, and gold foil, the internal environment of the processor would not exceed 10^4 rads Si. Since CMOS components have been observed to withstand this environment, adding the shielding to processors where the application can accomodate the additional weight penalty opens the door to using lower-cost standard parts. Whereas this shielding is not effective for the mitigation of single-event upsets (which in any case are better handled by error detection and correction

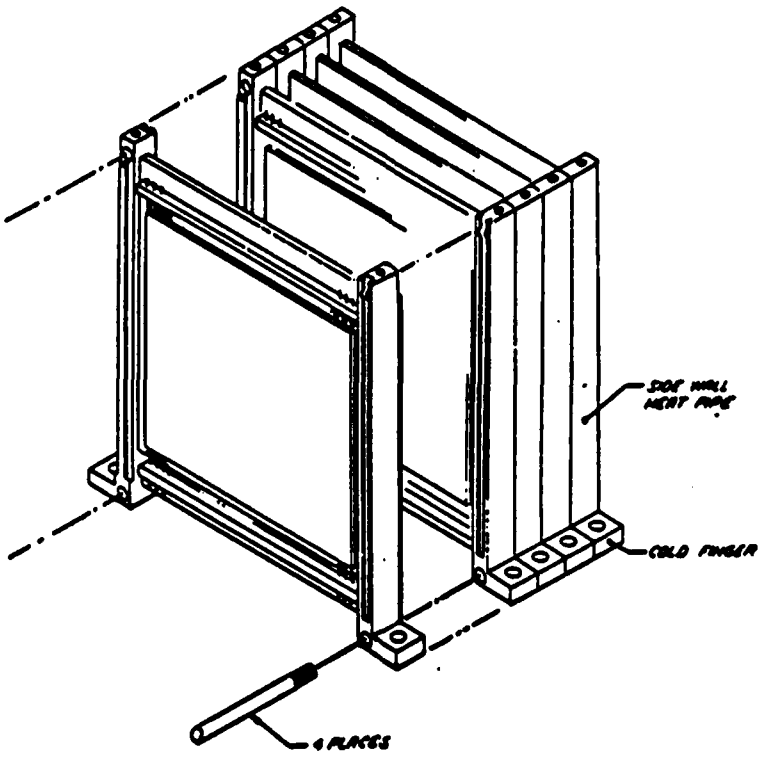


Figure 8. 3-D HWSI Packaging Concept.

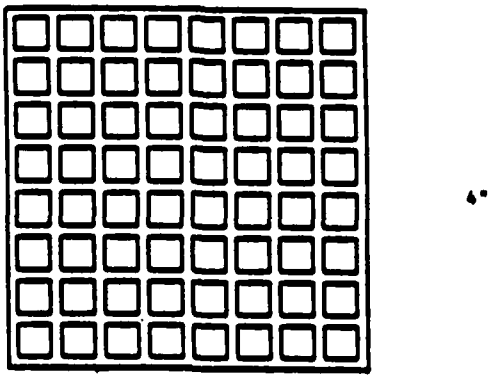


Figure 9. Substrate Layout.

circuitry), it can attenuate X-ray and other radiation components responsible for total dose and dose rate effects that cause latchup and damage in semiconductor devices. Such shielding is generally not practical at present because the large volumes to be enclosed require an excessive weight of material.

The power density within the 3-D assembly will be approximately 1.08 watts per cubic inch or 1,900 watts per cubic foot. The most efficient and practical method for cooling the assembly is the use of heat pipes. A heat pipe is a device that transfers heat by evaporation of liquid from hot areas with subsequent condensation onto cooler areas from which the liquid is wicked back to the heat source. The advantage of the heat pipe is that it has a very high effective thermal conductivity and can transport much more energy than solid conductors of the same area. Furthermore, it does not require an external source of energy as in the case of liquid cooling. Heat pipes are especially well suited for zero-g environments and have been used on a variety of spaceborne electronic equipments. In some of these applications, power densities of 55,000 watts per cubic foot or more have been successfully handled by heat pipes.

7. CONCLUSIONS

Our conclusions from the Phase I study are summarized below.

a. It is possible to achieve the high throughputs required for real-time execution of the velocity filter and other highly-structured algorithms for typical space-based infrared surveillance systems with both fine-grained and medium-grained parallel processor architectures.

b. The fine-grained approach appears to have a number of potential advantages with respect to minimum communication requirements and efficiency of VLSI implementation. In particular, it employs a repetitive, regularly-connected structure which is well matched to the technological opportunities as well as the technological constraints of VLSI design and fabrication. However, there are thus-unsolved problems with (1) real-time operation for high data rate input such as that from a high-resolution IR sensor as well as (2) programming. In addition, VLSI implementations of these architectures are not readily available at the present time.

c. The medium-grained approach appears to be more practical for near-term applications, since there is no significant real-time input problem with high data rates, existing VLSI chips are available for implementation, and programming problems are of a more familar nature. Communication requirements, while perhaps more severe than with the fine-grained approach, appear manageable for applications in which the

algorithms are highly structured, and where repetitive computations with little or no data-dependent branching are employed.

d. Reduction of power consumption (as well as size and weight) can be accomplished by the use of hybrid wafer-scale integration (HWSI) interconnect and packaging techniques. This reduction is obtained through elimination of individual IC packages, shrinkage of interconnect line dimensions, reduction of capacitance, and reduction of elelctrical noise effects.

e. A variety of HWSI configurations and materials suitable for application to parallel processor architectures can be employed. All of these employ multilayer, thin-film interconnect structures using aluminum or copper conductors and polyimide-like dielectric layers. Typical line widths are in the 10-25 micron range. Preference is given to configurations with composite substrates and TAB die attachment for production systems. For Phase II demonstration purposes, however, it is simpler and less expensive to use a simple silicon wafer substrate and ordinary wire bonding.

f. With advanced 3-D HWSI packaging technology it should be possible to obtain reductions in size, weight and power approaching an order of magnitude relative to conventional printed-circuit board packaging. A preliminary design concept has been developed for a space-based processor with 128 medium-grain nodes implemented with 1024 VLSI chips within a 5 x 5 x 5 inch enclosure. The total weight and power are 10 pounds and 80 watts, respectively (not including a possible radiation shield which could permit the use of non-radiation-hardened devices).

g. Critical technical issues have been identified, many of which can be resolved during a Phase II demonstration of power reduction. This demonstration can be carried out in an effective, relatively low-cost manner through the design, fabrication and test of a generic parallel processing element using low-cost commercial components packaged with simplified HWSI techiques.

8. KEY ISSUES

The primary issues in this project are those concerning the multilayer thin-film interconnect and its associated structures. These include the following:

a. Degree of downsizing of system dimensions possible;

b. Selection of dimensions and geometries of the conductor layers, the dielectric layers and the vias to minimize capacitance, inductance and resistance;

c. Resolution of complex tradeoffs between interconnect density, fanout, crosstalk, delay and attentuation;

d. Feasibility of fabrication of integral bypass capacitors within interconnect structure for reduction of current switching noise;
e. Feasibility of reduction of voltage swing with reduced levels of current switching and other types of noise;
f. Feasibility of performance improvement with reduced noise levels;

Another issue--in the area of chip design--is the ratio of internal power to driver power. Many existing chips have been designed to operate with conventional packages and with conventional printed-circuit board packaging technology. These chips utilize a large amount of silicon area for the drivers and require a large amount of power when operated in these conventional environments. As a result, there has been little incentive for the chip designer to minimize the internal power dissipation. Such chips could benefit from a redesign to reduce the internal power (as well as the silicon area devoted to drivers). Although such redesign is outside the scope of this program, it is possible to determine the extent to which redesign would be effective by measuring the power dissipation components under various controlled electrical constraints for different classes of devices such as microprocessors, static RAMs, etc.

9. REFERENCES

[1] C. E. Seitz, "Concurrent VLSI Architectures", *IEEE Transactions on Computers*, Vol. C-33, No. 12, December 1984.

[2] E. P. Verheiden, Jr. and R. Mauriello, "Concurrent Error Detection for a Signal Processing ASIC", *1988 Government Microcircuit Applications Conference Digest of Papers*, p. 219.

[3] R. M. Lea, "ASP: A Cost-effective Parallel Microcomputer", *IEEE MICRO*, October 1988.

[4] [2] R. O. Carlson and C. A. Neugebauer, "Future Trends in Wafer Scale Integration", *Proceedings of the IEEE*, Vol. 74, No. 12, December 1986.

[5] C. A. Neugebauer and R. O. Carlson, "Comparison of Wafer Scale Integration with VLSI Packaging Approaches", *IEEE Transactions on Components, Hybrids, and Manufacturing Technology*, Vol. CHMT-10, No. 2, June 1987.

[6] R. R. Tummala and E. J. Rymaszewski (Eds.), *Microelectronics Packaging Handbook*, Van Nostrand Reinhold, New York (1989), p. 693.

[7] G. Messner, "Cost-Density Analysis of Interconnections", *IEEE Transactions on Components, Hybrids, and Manufacturing Technology*, Vol. CHMT-10, No. 2, June 1987.

[8] K. D. Warren, J. H. Reche, W. J. Jacobi and R. M. Lea, "A 3D HDI ASP: A Cost-Effective Alternative to WSI Signal Processors", *1989 Wafer Scale Integration Conference*, San Francisco, January 1989.

[9] W. J. Jacobi and J. J. H. Reche, "3-D Wafer-Scale Interconnect and Packaging Using Photosensitive Glass-Ceramic Substrates", to be published in *1989 NAECON Conference Proceedings* (May 1989).

[10] J. E. Kohl, C. W. Eichelberger, R. J. Wojnarowski, R. O. Carlson and L. M. Levinson, "High Density Overlay for Bare Chip Interconnect", *1988 Government Microcircuit Applications Conference Digest of Papers*, p. 445.

[11] T. A. Lane, F. J. Belcourt and R. J. Jensen, "Electrical Characteristics of Copper/Polyimide Thin-Film Multilayer Interconnects", *IEEE Transactions on Components, Hybrids, and Manufacturing Technology*, Vol. CHMT-12, No. 4, December 1987.

[12] A. A. Evans and J. K. Hagge, "Advanced Packaging Concepts-- Microelectronics Multiple Chip Module Utilizing Silicon Substrates", *SAMPE Conference Proceedings*, June 24, 1987.

[13] C. J. Bartlett, J. M. Segelken and N. A. Teneketges, "Multichip Packaging Design for VLSI-Based Systems", *IEEE Transactions on Components, Hybrids, and Manufacturing Technology*, Vol. CHMT-12, No. 4, December 1987.

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION: Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT: Approved for public release; distribution unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S): SCC-R-104-F	5. MONITORING ORGANIZATION REPORT NUMBER(S): ARO 26019.1-EL-SBI
6a. NAME OF PERFORMING ORGANIZATION: SPACE COMPUTER CORPORATION; 6b. OFFICE SYMBOL (If applicable):	7a. NAME OF MONITORING ORGANIZATION: U. S. Army Research Office
6c. ADDRESS (City, State, and ZIP Code): 2800 Olympic Blvd. Suite 104 Santa Monica, CA 90404-4119	7b. ADDRESS (City, State, and ZIP Code): P. O. Box 12211 Research Triangle Park, NC 27709-2211
8a. NAME OF FUNDING/SPONSORING ORGANIZATION: U. S. Army Research Office; 8b. OFFICE SYMBOL (If applicable):	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER: DAAL03-88-C-0007
8c. ADDRESS (City, State, and ZIP Code): P. O. Box 12211 Research Triangle Park, NC 27709-2211	10. SOURCE OF FUNDING NUMBERS: PROGRAM ELEMENT NO. SBIR; PROJECT NO. SDIO; TASK NO. 87; WORK UNIT ACCESSION NO. -3
11. TITLE (Include Security Classification): ~~PASSIVE RANGING WITH ELECTRO OPTICAL SENSORS (U)~~ See Cover	
12. PERSONAL AUTHOR(S): WILLIAM J. JACOBI	
13a. TYPE OF REPORT: FINAL; 13b. TIME COVERED: FROM 7/15/88 TO 1/14/89; 14. DATE OF REPORT (Year, Month, Day): 1989, MARCH 1[illegible]; 15. PAGE COUNT	
16. SUPPLEMENTARY NOTATION: The view, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.	
17. COSATI CODES: FIELD / GROUP / SUB-GROUP	18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number): HYBRID WAFER-SCALE INTEGRATION, PARALLEL PROCESSING, PACKAGING, INTERCONNECT, MINIATURE ELECTRONICS, LOW POWER ELECTRONICS

19 ABSTRACT (Continue on reverse if necessary and identify by block number)

The basic goal of this project is to develop and demonstrate techniques for the reduction of power consumption of space-based processors for infrared surveillance systems. The primary technique is to minimize the capacitive loading encountered in off-chip communications for highly concurrent processing architectures. Both processing architecture and chip packaging are simultaneously considered to maximize MOPS per watt by increasing throughput while reducing system capacitance, signal delay, noise, voltage swing, and power consumption (the costs of system communications).

With conventional packaging technology, highly concurrent processing architectures result in hardware implementations that are extremely large, very

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT: ☑ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION: Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL: WILLIAM J. JACOBI	22b. TELEPHONE (Include Area Code): (213) 829-7733; 22c. OFFICE SYMBOL

DD FORM 1473, 84 MAR — 83 APR edition may be used until exhausted. All other editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED

heavy, and that consume excessive power. Monolithic wafer-scale integration is theoretically ideal but requires an extensive amount of redundant circuitry and provisions for circuit reconstructurability because of manufacturing yield problems. In the hybrid wafer-scale integration (HWSI) approach, individual pre-tested chips are bonded to a fine-line interconnect structure fabricated on the surface of a wafer-scale substrate. With this technique, high yields can be achieved without redundancy.

During Phase I we determined that it should be possible to achieve major reductions in size, weight, and power through the use of hybrid wafer-scale integration techniques for interconnect and packaging. Alternative interconnect and packaging approaches were considered and a conceptual design (electrical and mechanical) has been established for a miniaturized processor. While alternative processing architectures were considered, fine gain architectures like the Associative String Processor was our initial desired approach. However, neither chips nor software support are currently available and there is no indication when they might become available. Consequently, a medium grain architecture, for which both chips and software support are immediately available, was selected for a proof of principle processor.

While reduction of power consumption is the primary goal, the packaging technique also leads to opportunities for miniaturization and significant weight reduction. The degree of miniaturization that can be achieved permits processor radiation shielding schemes that would otherwise be impractical from a weight consideration. With this additional shielding, requirements for radiation hardened chips could be relaxed. The reduction of weight is not only realized in the processor, but also in the equipment required to generate, regulate, distribute, and dissipate the electrical power consumed. Finally, hybrid wafer-scale integration as a packaging technique, allows new chips to be developed with greater processing density because of the disproportionate chip area currently being devoted to bonding pads and pad drivers.