MIPS R5000 Microprocessor

Technical Backgrounder

Performance:	SPECint95	5.5
	SPECfp95	5.5
Instruction Set	MIPS-IV	
ISA Compatibility	MIPS-I, MIPS-II, AND MIPS-III	
Pipeline Clock	200 MHz	
System Interface clock	Up to 100 MHz	
Caches	32 kB I-cache and 32 kB D-cache, each 2-way set associative	
TLB	48 dual entries; Variable Page size (4 kB to 16 MB in 4x increments)	
Power dissipation:	10 watts (peak). at maximum operating frequency	
Supply voltage	min. 3.0 Vtyp. 3.3 Vmax. 3.6 V	
Packaging:	272-pin cavity-down Ball Grid Array (BGA)	
	223-pin ceramic Pin Grid Array (PGA)	
Fabrication Technology:	Vendor specific process including 0.35 micron	
Die Size:	80-90 mm^2 (Vendor Dependent)	
Number of Transistors:	3.6 million (4 Transistor SRAM cell), 5.0 million (6 Transistor SRAM cell)	
	(Of these totals, logic transistors number 800,000).	

Overview

This backgrounder introduces the R5000 microprocessor from MIPS Technologies, Inc. The information presented in this paper discusses new features in the R5000, i.e. how the R5000 differs from previous microprocessors from MIPS.

This section provides general information on the R5000, including:

- Introduction
- The R5000 microprocessor
- Packaging
- Future upgrades
- Block Diagram

Introduction to RISC

Reduced instruction-set computing (RISC) architectures differ from older complex instruction-set computing (CISC) architectures by streamlining instruction execution. The MIPS architecture, developed by MIPS Technologies, is firmly established as the leading RISC architecture today.

On introduction, RISC microprocessors were used for high performance computing applications. Lately, these processors have found their way into the consumer electronics and embedded systems markets as well. Today, MIPS RISC processors are used in a wide variety of applications, from embedded controllers to supercomputers.

Between 1985 and 1994, three generations of the MIPS architecture have been introduced and widely adopted. The first commercial MIPS processor, the R2000, ran at 8-MHz and implemented the MIPS I 32-bit architecture. The R3000 family raised the operating frequency to 40 MHz. The R4x00 family extends the MIPS architecture to 64 bits to increase the address space and data and instruction bandwidth. The first R4000 also used a fast super-pipeline to reach high operating frequencies. Some members of the R4x00 family also include support for multilevel caches and multiprocessing. The R4x00 family currently operates at pipeline speeds up to 250 MHz and implements the MIPS III instruction set. The R8000 was the first MIPS product to implement the MIPS IV instruction set. This was followed a year later by the introduction of the R10000. The R5000 processor is the third MIPS processor to implement the MIPS IV instruction set.

The MIPS semiconductor partners have manufactured and shipped MIPS standard processors in a variety of process technologies and have designed numerous derivative products for ancillary markets such as embedded applications. MIPS semiconductor partners include Integrated Device Technology, Inc., LSI Logic Corp., NEC Corporation, NKK Corporation, Philips, Siemens and Toshiba Corporation. Users of the MIPS architecture include AT&T, Cisco, Network Computing Devices, NEC, NeTpower, Pyramid Technology, QMS, Siemens-Nixdorf, Silicon Graphics, Sony, Tektronix, Tandem, and many more.

The R5000 Microprocessor

The R5000 is designed to offer excellent price/performance. It achieves high performance levels while restricting the die area to less than 90 sq. mm. The R5000 provides an upgrade path to users of the R4600/R4700.

The R5000 was optimized to run software in a modern desktop computing environment. This includes applications that make use of interactive 3D graphics to provide a more sophisticated user interface. Recent examples of this type of application are data visualization and VRML (Virtual Reality Modeling Language) extensions to the World-Wide Web. The R5000 accelerates geometry-processing for 3D graphics by providing an extremely high throughput rate for single-precision floating-point calculations.

The R5000 uses a variety of implementation techniques to provide high performance at low cost. These techniques include:

- dual-issue instruction mechanism
- large on-chip caches,

- 48 dual-entry TLB
- fully-pipelined floating-point unit with separate ALU and DIV/SQRT units,
- multiple paths for integer and floating-point ALU operations,
- implementation of low-latency multiply-add (MADD) and conditional move instructions that enhance floating-point performance.
- relatively small die size (80-90 mm^2) and pin count (272 pin) helps to reduce package cost.
- BGA package implementation allows for higher bus speeds by reducing package capacitance and inductance.
- Hooks for cache coherency and multi-processor support.

Packaging

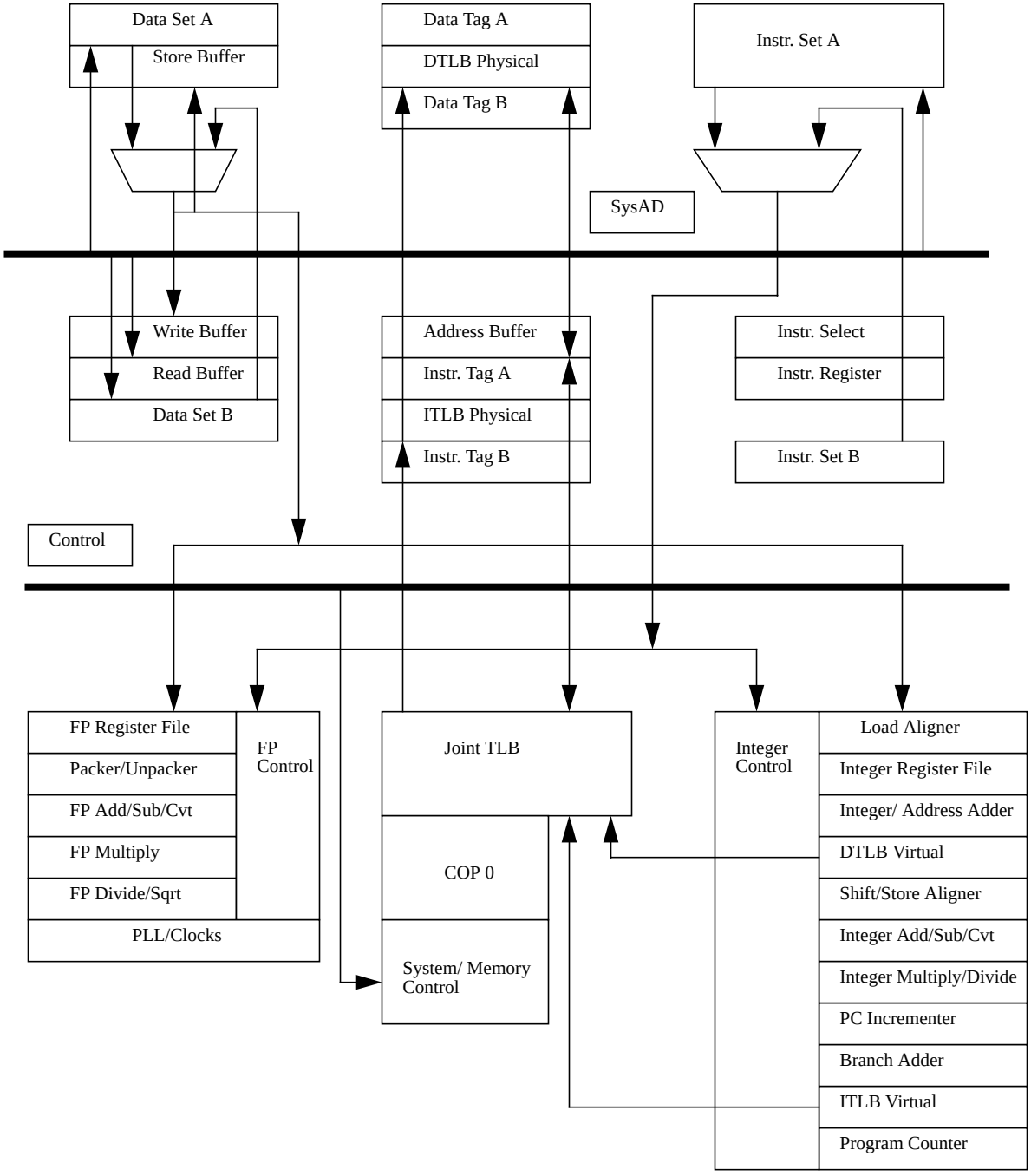
The R5000 will be made available in both a single 272-pin BGA package and a 223-pin PGA package. The relatively low pin count reduces package cost, while the BGA package increases I/O bandwidth by facilitating a 100 MHz system interface.

Future Upgrades

A family of processors based on the R5000 core will provide higher clock speeds and additional features. These will provide an extra performance boost at the same low price points as the current R5000. The first R5000 achieves 200Mhz core speed early in 1996. Later in 1996, this speed increases to 250Mhz. Early in 1997, an improved core, the R5000A will boost performance further with additional functionality.

Block Diagram

Figure below shows a block diagram of the R5000 processor.



Implementation

The R5000 fits naturally as an upgrade to the R4600 microprocessor. Some of the features which differentiate the R5000 processor from the R4600 are:

- Increased floating-point performance for 3-D graphics geometry processing
- Dual-issue mechanism
- Multi-processing support
- Secondary cache support
- Flexible clocking mechanism
- Cost reduction

These features are discussed in more detail in the following sections.

Optimized for 3D Graphics Performance

Geometry-processing performance is dramatically increased using the following techniques:

- Inclusion of Multiply-add (MADD) instruction
- Pipelined floating-point ALU allows for high throughput and single-cycle repeat rates for most single-precision floating-point arithmetic instructions
- Large on-chip caches
- High clock speed
- Write buffer
- Pipelined writes

MADD Instruction

The MIPS IV instruction set supports four floating-point multiply-add/subtract instructions which allow two separate floating-point computations to be performed with one instruction. The four instructions are;

1. Multiply-add (MADD)
2. Multiply-subtract (MSUB)
3. Negative Multiply-add (NMADD)
4. Negative Multiply-Subtract (NMSUB)

The product of two operands is either added to or subtracted from a third operand to produce one result. The result is then rounded according to the rounding mode specified in the instruction. The final result is then placed in another floating-point register whose location is also defined in the instruction.

3D graphics (geometry-processing intensive) applications are typically floating-point based, and a multiply followed by an add or subtract operation is a common way to calculate geometric transforms. Typically in a multiply-add operation, the multiply is performed and the result written to a register. On the following cycle the register is read, the add operation is performed, and the result written back to the register. With the MADD instruction the result of the multiply is immediately used by the add operation, eliminating the intermediate step normally required when the multiply result is written to a register. Having a single instruction which performs both the multiply and the add in one cycle enhances graphics performance by reducing the cycle time required for this operation to complete.

Pipelined ALU's

The R5000 processor boosts its floating-point performance by reducing the repeat rate of many instructions from 2 cycles to 1. This allows those instructions critical to 3-D applications such as the MADD instruction to be issued on every cycle, as opposed to every other cycle.

Separate integer and floating-point ALU's allows these operations to occur simultaneously. Integer instructions no longer have to wait for long-latency floating-point operations to finish before being fetched, and vice-versa. Load/store operations may also be issued simultaneously with FP ALU instructions to reduce load latencies and bandwidth.

Large On-chip Caches

The R5000 processor contains separate 32 kB data and instruction caches, double the size of the R4600. Each cache is 2-way set associative, which helps to increase the hit rate over a direct-mapped implementation, and has a 32-byte fixed line size. Cache lines may be classified as write-through or write-back on a per-page basis.

Both caches are virtually indexed and physically tagged. A virtually indexed cache allows the cache access to begin as soon as the virtual address is generated, as opposed to waiting for the virtual to physical translation. The cache is accessed at the same time as the address translation is performed. The physical address is then compared against the corresponding instruction or data cache tag. If the compare is valid, the data which has been retrieved from the cache is used. If the compare is not valid, meaning that the address requested does not reside in the cache, the data is not used and a cache miss is generated. Having a physically tagged data cache also allows for coherency between the primary and secondary caches in a system.

Having large primary caches allows more of the application to be executed on-chip, reducing accesses to slower secondary cache and main memory. This in turn reduces bus utilization and allows the application to run faster since fewer off-chip accesses are required.

Dual Entry TLB

The TLB of the R5000 processor contains 48 dual entries. This implementation is equivalent to a 96-entry TLB but requires less die area and logic to implement. Each virtual page number (VPN) entry equates to two physical frame numbers (PFN), one even and one odd. Since addresses are often sequential, back-to-back TLB accesses often occur on adjacent virtual pages. The lower bit of the VPN is used to determine whether the even or odd PFN will be used. In addition, the TLB is fully-associative and maintains R4x00 family compatibility. This allows all existing R4x00 code to run unmodified and also supports the Microsoft Windows NT operating system.

Write Buffer

Writes to external memory, whether cache miss writebacks or writes to uncached or write-through addresses, can use the on-chip write buffer. The write buffer holds up to four 64-bit address and data pairs, or one cache line to be written out. The write buffer allows the CPU to write data into the buffer without accessing the system bus. Since data cache writebacks are typically performed on a line basis, an entire line can be written to the buffer, allowing the CPU to resume normal execution. The contents in the write buffer are then written out at a later time. Without a write buffer, the CPU would have to write a single 64-bit doubleword, then wait until the memory operation completes, before writing another. The writing of a data cache line requires four 64-bit writes (32 bytes). Typically data in the primary cache is written out only when the line is replaced, or becomes dirty.

Graphics applications such as a screen update can be done performing uncached write cycles. For uncached write cycles, the write buffer can significantly increase performance by allowing the pipelining of multiple writes. As with cacheable write cycles, the buffer allows the CPU to write data to the buffer and immediately begin processing the next write data. Without the buffer, the CPU would output the write data, then be forced to wait until the uncached write operation has completed before processing the next write.

Pipelined Writes

The pipelined write protocol also uses the write buffer to allow pipelining of write cycles. In the original R4000 architecture there is a two clock delay between the generation of back-to-back addresses due to the fact that certain control signals must be asserted two clocks before each address can be output. This results in two dead clocks between back-to-back cycles. In the R5000 processor, performance is significantly increased due to the fact that the two null cycles between each write cycle are eliminated. Write cycles can be performed back-to-back without any dead clocks between cycles.

Dual-Issue Mechanism

The dual-issue mechanism implemented in the R5000 processor allows a floating-point ALU instruction to be issued simultaneously with any other instruction type. Whenever a floating-point ALU instruction is fetched with any non-FP-ALU instruction, both instructions can be issued in the same cycle. In most geometry-processing applications, a floating-point load or store is frequently issued together with a MADD instruction. Load and store instructions in one pipeline usually provide enough data bandwidth to permit a new MADD instruction to be issued every cycle for a sustained period. Well structured code (e.g. OpenGL libraries) can take full advantage of this pipeline structure.

Dedicated Integer and FP ALU's

Separate Integer and FP ALU's allow instructions of both types to be performed simultaneously. This is useful when running CAD-type applications as both fixed-point and floating-point math calculations are being performed. Integer instructions are not stalled while long latency floating-point operations are being executed.

Separate FP Execution Units

In addition to the dual-issue mechanism, the R5000 processor also contains separate acceleration hardware for most floating-point ALU instructions than for DIV/SQRT instructions. This allows long-latency operations such as divide and square-root to be performed in a dedicated unit, thereby allowing other shorter-latency operations such as MADD and subtract to be overlapped while the divide or square-root operation is in progress.

Scaleable for Multiple Processors

The R5000 processor incorporates 8 external signals dedicated to multiprocessor support. These signals allow for arbitration and data coherency between processors. Symmetric multiprocessing (SMP) systems implementing the full MESI (Modified Exclusive Shared Invalid) cache consistency protocol in both primary and secondary caches, as well as other styles of multiprocessing will be supported by the R5000 family. Multiprocessor support is not available in the first production version of the R5000. The performance of SMP systems based on the R5000 is expected to scale well up to 4 processors.

Secondary Cache Support

The R5000 processor contains a dedicated secondary cache interface. These signals provide an efficient interface between the processor, the secondary cache, and the secondary cache tag RAM. All tag RAM interface signals (such as data and chip enables, output enable, address match, cache valid, line index, and word index) are provided by the R5000 processor. The secondary cache supports multiple cache sizes and both the write-through and write-back data transfer protocols. Data transfers to the secondary cache share the 64-bit system bus.

Multiple Cache Sizes

The secondary cache can be configured as 512 kB, 1Mbyte, or 2 Mbyte, allowing large applications to run within the secondary cache, reducing the number of accesses to slower main memory. The secondary cache is accessed through the system bus. Uncached bus cycles are not evaluated by the secondary cache control logic as they travel to the external agent. Uncached operations such as video screen updates can be passed directly to the system logic responsible for routing the data to the screen without any delays from the secondary cache logic.

Simultaneous Access

To maximize data throughput, the main memory accesses can be initiated while the secondary cache tag is being compared. If the requested address is found to be in the secondary cache, the memory access is aborted. However, if the address is not found in the secondary cache, the main memory access has already begun and the data can be retrieved more quickly.

Flexible Clocking Mechanism

The clocking mechanism in the R5000 processor has been simplified from previous generations and offers a number of pipeline frequencies based on the frequency of the input clock.

Single External Clock Signal

The R5000 processor greatly simplifies the clocking mechanism currently used in both the R4000 and the R4600. A single clock signal (SysClock) is used for the system interface, as opposed to three. The R5000 eliminates the Rclock, Tclock, and MasterOut clock signals that existed in the R4000. Having only one clock simplifies system design, as well as reducing the circuit complexity of the internal clock mechanism.

On-chip Clock Multiplication Circuitry

The R5000 includes on-chip clock frequency multiplication circuitry to support 200-MHz internal operation from an external 50-MHz clock. The R5000 has the option of operating internally at 2, 3, or 4 times the frequency of the external clock. Maximum bus speed of the system interface (SysAD) is 100 MHz.

Cost Reduction Features

The R5000 is designed for low cost. The main areas contributing to microprocessor cost are packaging, test and assembly, and die costs.

Packaging cost reduction

The R5000 is available in a 272-pin plastic Ball Gird Array (and also Pin Grid Array). The plastic material and relatively low ball count helps to keep package cost down, and low inductance characteristics of a BGA allows the R5000 processor bus to maintain high speed.

Test & assembly cost reduction

The R5000 implements column redundancy in instruction and data primary caches. Column redundancy reduces the chip's sensitivity to defects in the caches. The test process first tests the integrity of each bit column in the cache; polysilicon fuses in the chip are then blown to swap good redundant bit columns for defective bit columns. Column redundancy increases the die yield at the test stage, which in turn decreases testing costs per good die.

Die cost reduction

The die area was reduced by the following techniques:

- High-density design rules (including 0.35 micron)
- Use of 4-transistor RAM cells in the caches
- Optimized cache and TLB size

Comparison Chart

Table 1 shows a comparison between the R5000, R4600, and R4400 processors.

Table 1

Parameter	R5000-200	R4600-133	R4400-250
Primary Data Cache Size	32 KBytes	16 KBytes	16 KBytes
Primary Instruction Cache Size	32 KBytes	16 KBytes	16 KBytes
TLB Size	96 entries	96 entries	96 entries
ISA Support	MIPS-IV	MIPS-III	MIPS-III
Pipeline	5-stage	5-stage	8-stage
Secondary Cache Support	Yes	No	Yes
Multiprocessor Support	Yes[a]	No	Yes
Instruction Dual Issue	Yes	No	No
Special Integer and FP ALUs	Yes	No	Yes
Die Size	84 mm^2	70 mm^2	100mm^2
Number of Transistors	3.6 million	1.8 million	2.3 million

a. Not available in first version

Table 2 shows a comparison between the R5000, Pentium, and PentiumPro processors.

Table 2

Parameter	R5000-200	Pentium-133	Pentium Pro-200
Primary Data Cache Size	32 KBytes	8 KBytes (unified)	8 KBytes
Primary Instruction Cache Size	32 KBytes	8 KBytes (unified)	8 KBytes
ISA Support	MIPS-IV	x86	x86

Table 2

Parameter	R5000-200	Pentium-133	Pentium Pro-200
Secondary Cache Support	Yes	No	Yes
Multiprocessor Support	Yes[a]	No	Yes
Multiple Instruction Issue	Yes (2-way)	Yes (2-way)	Yes (3-way)
Separate Integer and FP ALUs	Yes	Yes	Yes
Die Size	84 mm^2	93 mm^2	196 mm^2
Number of Transistors	3.6 million	3.3 million	5.5 million
SpecInt95	5.5	4.1	8.1
SpecFP95	5.5	2.5	6.0

a. Not available in first version

Summary

The R5000 processor obtains high performance at low cost. New features include the following:

- The R5000 processor runs the MIPS IV instruction set which contains additional floating-point instructions such as the multiply-add (MADD) which accelerates geometry-processing in 3-D graphics applications.
- A dual-issue instruction mechanism allows floating-point load (or store) and floating-point ALU instructions to be dispatched in the same cycle. Highly pipelined floating point ALU instructions greatly enhance 3D graphics performance.
- An on-chip write buffer enhances bus performance by facilitating pipelined write operations.
- Separate integer and floating-point ALU's enhance performance in mixed integer/fp applications by allowing long latency instructions such as floating-point divide and square-root operations to be performed at the same time as integer ALU operations.
- Large on-chip instruction and data caches, each 32 Kbytes in size, allows many common applications to run within the primary cache, enhancing performance by reducing the number of accesses to both secondary cache and main memory.
- The multi-processing capability of the R5000 processor enables multiprocessor servers with additional processors added at low-cost. This system configuration is especially valuable for the Windows NT mid-range server market.
- On-chip clock multiplication circuitry allows for a flexible clocking scheme which can meet the requirements of almost any system. The reduction in the number of external clocks simplifies system design and reduces on-chip clock logic complexity.

Glossary

BGA (Ball Grid Array): A semiconductor packaging technology where each pin is represented by a solder ball on the package. The BGA package eliminates the actual pin, reducing the amount of capacitive and inductive loading, allowing for higher performance and higher bus speeds by reducing the overall time for the signal to reach its destination.

Cache: An on-chip temporary storage area containing a copy of main memory fragments. Cache access is much faster than main memory access.

CPU (Central Processing Unit): The part of a microprocessor where the majority of the instructions are executed.

Die: The silicon chip after it has been cut from a wafer and before it has been packaged.

FPU (Floating-Point Unit): Dedicated logic to accelerate calculations using floating-point numbers.

IU (Integer Unit): The part of a CPU that performs calculations using integer arithmetic.

LVCMOS (Low-voltage CMOS): An IEEE standard for low-voltage (3.3V) logic design.

MMU (Memory Management Unit): That part of a microprocessor which implements virtual-to-physical address translation and the memory system hierarchy including cache memory.

MTI (MIPS Technologies, Inc.): The developer of the MIPS RISC architecture, the leading RISC architecture worldwide.

Page Table: An area of main memory containing sets of virtual addresses with their corresponding physical addresses and protection data.

RISC (Reduced Instruction Set Computing): A design philosophy that avoids implementing complex functions in silicon but realizes large performance increases through executing simpler, standardized instructions at faster, more efficient rates.

Pipeline: A mechanism to allow multiple instructions to overlap during execution for greater throughput. A five-stage pipeline offers peak performance five times that of a non-pipelined processor.

TLB (Translation Lookaside Buffer): An on-chip "page table" cache containing copies of the page tables used by the MMU for virtual-to-physical address translation.

Write Buffer: The flush buffer is a temporary storage location for data that is being written from the pipeline or cache to main memory. The flush buffer allows the processor to continue executing instructions while data is being written to main memory.

64-bit Processor: A processor in which all address and data paths are 64 bits wide. Leading-edge applications require 64-bit processors today. 32-bit capability in a 64-bit processor is important to manage the smooth transition of software from 32 to 64 bits. The MIPS R4000, R4400 and R5000 MPUs can run in either 32-bit or 64-bit mode.

Contact MIPS Technologies, Inc. for a more complete list of publications available on RISC technology and the MIPS architecture.

MIPS Technologies, Inc. Information Service:

1-800- I GO MIPS or 1-800-446-6477Inside the US

415-688-4321Outside the US

World Wide Web: URL to http://www.mips.com