R18000™

The Latest SGI™ Superscalar Microprocessor

Tim Fu, Farshid Iravani, Mahdi Seddighnezhad, Kenneth Yeager, David Zhang

Silicon Graphics, Inc.
www.sgi.com
Outline

- R18000 Processor MicroArchitecture
- R18000 Memory Hierarchy
- Verification & Test
- Technology
- Summary
R18000 Chip Block Diagram

R18000 Chip

Superscalar Processor

Inst Cache

Data Cache

Cache Controller

System Interface

1 MB Secondary Cache

Tertiary Cache Tags

2 to 64 MB Tertiary Cache

System Controller

Local Main Memory

Second R18000

Network Links

Node
Scalable System Architecture

SGI’s ccNUMA system architecture

- Local memories are shared in a cache-coherent Non-Uniform Memory Architecture.
- Single-system image (SSI).
  - Single shared address space.
  - Single copy of operating system.
- Scales to very large processor counts.
  - High-bandwidth network.
  - Low-latency in hubs and routers.
  - A dozen 512-p systems have been installed.

Design Challenge

- Keep processors and/or network busy even when streaming data from remote nodes.
  - Remote latency can be hundreds of cycles.
  - Time exceeds a dozen block transfers.
• **4-way super-scalar microprocessor**
  - Mips-4 Instruction Set Architecture
  - Out of Order execution
  - Two Floating-point Execution Units
    - Each issues one Add, Multiply or Multiply-Add instruction per cycle

• **Large virtual and physical address spaces**
  - 52-bit virtual address (data)
  - 48 bits physical address
  - Larger TLB page sizes 64M, 256M and 1G page sizes
R18000 Processor MicroArchitecture
Processor Core Block Diagram

IC
32KB Instr. Cache


Instr Seq.

Active List

Free List

Branch

Mips-4 ISA
Pre-decode

32-bit Inst ⇒ 36-bit

Graduate

8 Entry FQ
8 Entry

64 Entry AQ
16 Entry

64 Entry IQ
16 Entry

64 Int. Reg.

64 FIt.Pt. Reg.

Multiply-Add

Multiply-Add

St
Ld
control
St
Ld

WrBack

Refill

Adr

ALU

ALU

TLB

64-bit Data Paths

32KB Data Cache

Hot Chips 13, August 19-21, 2001
• Two floating-point execution units
  • Second unit doubles peak FLOPS rate
  • Each unit is controlled by an 8-entry Flt.Pt. queue
  • Each queue can issue one instruction per cycle:
    • Add
    • Multiply
    • Multiply-add
  • Multiplier and adder are pipelined
    • 1-cycle repeat rate
  • Latency increased to 3 cycles for faster clock rate
  • Divide and square-root units use iterative SRT algorithms
R18000 Processor MicroArchitecture
New System Bus

• High bandwidth “SysTF” bus
  • Uni-directional wiring with source-synchronous DDR clocking
  • 64-bit data path to hub chip, for addresses and write data
  • 128-bit data path from hub chip, for read data and interventions
  • Two processor chips share bus
  • Programmable clock divisors

• Split transaction
  • Up to 28 outstanding operations (14 per processor).

• ECC protection

• SysAD mode for backward compatibility
Memory Hierarchy

Cache Hierarchy

- 3-level Cache hierarchy
  - L1: On-chip 32KB Instruction and 32KB Data cache
    - 2-way set associative
  - L2: On-chip 1M 4-way set associative on chip secondary cache
  - L3: Off-chip tertiary cache, up to 64 MB
    - On-chip tag
    - External cache is optional. Chip can operate with only internal caches

- Subset property
  - L1 and L2 lines must be subsets of L3
  - Altered L1 lines must also be subsets of L2, to simplify write-back
  - Refill from memory is loaded into on-chip L2, not into L3
  - ECC protection on L2 and L3 caches, data and tags. (Parity on L1)
  - Non-blocking operation. Write-back. LRU replacement algorithm
  - Cache locking supported for L2 and L3
• On-chip secondary cache
  • 1 MB, 4-way set-associative, 128B line size
  • Two banks of multiple small arrays
    • Each bank has a write-back buffer for 8 cache lines
    • Copy entire line in one cycle
  • Operates at processor pipeline clock rate
  • Latency is less than half external cache latency. (5 to 6 cycles versus 12 to 14)
  • Bandwidth is 6 times greater than external cache
  • Simultaneous read and write at full internal bus bandwidths
    • Internal DDR clocking reduces wiring and noise
    • Read 4 quadwords per cycle and write 2 quadwords per cycle
    • All transfers are 64B packets, sent in two cycles
  • Reduces interference between instruction and data cache refills
  • Eliminates interference between refills and write-back
• Tertiary Cache - Data Arrays
  • 4-way set-associative, 128B line size
  • Off-chip bandwidth is physically constrained by pin-out and wire length
    • 144 pins for 16 bytes data plus 8-bit ECC
    • Programmable clock rate. (Bit rate up to processor pipeline clock rate)
  • SDR or DDR synchronous SRAM, or
  • Fast DDR synchronous DRAM
    • Allows very large caches up to 64MB by 8-way sectoring
    • Same transfer rate as SRAM
    • Slightly longer latency is acceptable because:
      • Tertiary cache is only accessed if miss in secondary cache
      • Tertiary cache transfers 128-byte blocks
      • Latency is much less than main memory
• Tertiary Cache Tags
  • 400-KB memory array contains:
    • 16,536 sets of four 52-bit address tags
    • ECC error check and correction on each cache set
    • Similar circuit design as on-chip secondary cache
  • Tag check is completed before accessing external data RAM
    • 4-way set-associative organization
    • L3 miss does not use any external RAM bandwidth
  • Tag check and update take no external RAM bandwidth
    • System intervention to “snoop” cache cause minimal interference
  • Minimize latency for refill from main memory
• Size of on-chip cache tag array limits maximum size of off-chip cache. Sectoring raises limit from 8 MB to 64 MB.

• A single address tag can be used for 1, 2, 4, or 8 adjacent cache blocks (“sectors”)

• Each additional sector requires only an extra 2-bit state

• Allows 8 times larger cache with only 50% larger memory array
  • An 8-sector tag represents a 1024-byte “macro-line”
  • Sectors are filled only when referenced
  • Victim tag may require up to 8 lines to be written back to main memory
    • “Lazy” write-back logic allows sectors to be refilled first
    • Only one X-queue entry is kept busy
• Use parallel operation to achieve high performance
• Use high internal bandwidths to simplify logic design
• On-chip L2 cache is interleaved in 2 parallel banks
  • Each bank has separate read bus to processor
  • Data cache write-back bus is shared
  • Each bank is controlled by a separate X-Queue
    • Operations to different cache sets are independent
  • Cache bandwidth higher than read plus write buses combined
    • No interference between reads and writes
    • Easier arbitration
• Bandwidth of each internal bus higher than system bus plus external cache combined
  • Secondary cache can be refilled directly, with only minimum buffering
Memory Hierarchy

High-Bandwidth Buses

bytes / cycle

16  8

Fetch  Load/Store

32  32  32

Read  Write  Read

Write-back buffers

Processor

I-Cache  D-Cache

Primary Caches

System Bus

8

16

System clock clock

Processor

I-Cache

L2-Cache Bank 1

L2-Cache Bank 0

Secondary Cache

L3 External RAM

Tertiary Cache

Gen  Chk

Buffer  Buffer

External cache clock

16

128 bit

16

128 bit

8/31/01  16
2 X-queues control external operations (L2 and L3 caches, system interface).
- Bank selected by Address[7], because L2 lines are 128 bytes.
- Each XQ contains 8 entries.
- Each entry contains 6 independent control fields for:
  - Primary caches.
  - Secondary cache read.
  - Tertiary cache read and write.
  - System bus read and write.
Cache Control
Split operations into simple actions

Example:
Data cache miss is split into four (almost) independent actions:

- Write Dc victim into Sc.
- Write Sc victim into Tc.
- Write Tc victim into memory.
- Refill Dc and Sc from memory.

Bypass data to processor
Verification & Test

• **Tools and environment:**
  - In house HDL and simulator with backup and replay.
  - Graphical user interface for simulation and regression.
  - Instruction level simulator as a reference machine.
  - Programmable random code generators for UP and MP.
  - C-based system model supports.
    - Cache arrays, memory controller and arrays, bus controller.
    - Bus protocol checking.
  - 1 to 4 processor configuration for MP verification.

• **Diagnostics**
  - Architecture Verification Programs (AVP).
  - Micro-architecture Verification Programs (MVP).
  - Random diagnostics from programmable random code generators.
  - Diagnostics are self-checked and/or compared with a reference machine.
Verification & Test
Trace Recorder, On-Chip debugging & profiling device

Debugging Challenges
Activity hidden on chip (L1 and L2 caches).
Deep sub-micron features are difficult to observe in testers, impossible in a system.
Large systems have many processors.
Failure point may be difficult to find.
Exact failure condition must be recreated.

On-chip Trace Recorder
512 x 72-bit trace memory (4 KB).
Records signals at processor clock rate.
Multiple trigger and recording options.
Configured by program or by JTAG.
Helps identify dynamic state of processor:
  cache refill, branch mispredict, ordering.

- Trigger, Mask
- CPU State to Be Captured
- Processor Control
- Trace Recorder
  512 x 72-bit memory
- Register File
- JTAG
  JTAG Control
**NEC’s UX5 Advanced CMOS Technology**

- **Technology Generation**: 130 nm
- **Voltage**: $V_{cc} = 1.2$ V
- **Gate length**: $L_{poly} = 95$ nm
- **Core oxide thickness**: $T_{ox} < 19$ Å
- **Transistor current**: $I_n > 800$, $I_p > 300$ μA/μm
- **Leakage current**: $I_{offN}$ or $I_{offP} < 10$ nA/μm worst case
- **6T Memory Cell Area**: < 3 μm²
- **Interconnect wiring**: 9 Layers Full Copper Damascene
- **IMD Dielectric Constant**: < 3.0
Technology

Scanning and Transmission Electron Microscope Images

**Figure (a):** 3D SEM image of the metal stack on the core memory area

**Figure (b):** TEM image of R18000 device cross section. Achieving a 95nm gate length required a size reduction etch in combination with KrF lithography. A gate ReOx before extension formation reduces the Cov to less than 0.29 ff/μm.
Technology

- **Figure (c):** TEM image of the metal stack for the R18000 core chip. Low K dielectric (Ladder Oxide) is used to reduce the lateral metal capacitance. Ladder Oxide offers comparable K value to SiLK with better thermal properties.
• R18000 is a high-performance superscalar microprocessor
  • Designed for use in large scalable ccNUMA systems
• Supported by a high-bandwidth system bus and a 3-level cache hierarchy
  • 1 MB on-chip secondary cache
  • On-chip tags support a 64 MB external cache, with sectoring
• Multiple banks and non-blocking operation improve cache performance
  • Each operation is split into multiple independent actions
• ECC protection
• Trace recorder supports system debugging