"...perhaps the most comprehensive book on the subject we've ever seen. It's incredibly well researched, well written, and jam packed with useful information."
—**Josh Norem, Reviews Editor, Maximum PC**

DVD-ROM
Includes:
2 hours of new how-to video with Scott Mueller

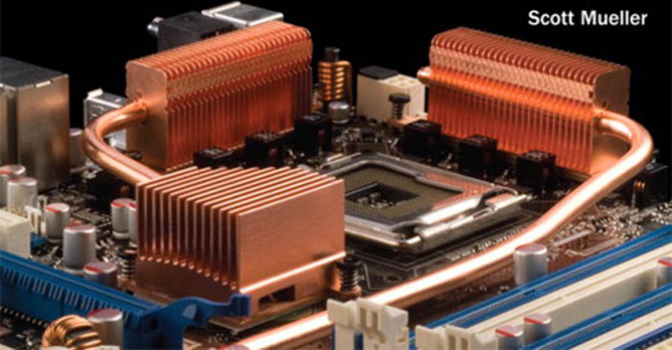
Scott Mueller's

UPGRADING AND REPAIRING PCs

19th EDITION

20th Anniversary

Scott Mueller



Upgrading and Repairing PCs, 19th Edition

ISBN-13: 978-0-7897-3954-4
ISBN-10: 0-7897-3954-2

Library of Congress Cataloging-in-Publication Data:

Mueller, Scott.
Upgrading and repairing PCs / Scott Mueller. — 19th ed.
p. cm.
ISBN 978-0-7897-3954-4
1. Microcomputers—Maintenance and repair. 2. Microcomputers—Equipment and supplies. 3. Microcomputers—Upgrading. I. Title.
TK7887.M84 2010
004.16—dc22
2009042724

Printed in the United States of America

First Printing: December 2009

Trademarks

All terms mentioned in this book that are known to be trademarks or service marks have been appropriately capitalized. Que Publishing cannot attest to the accuracy of this information. Use of a term in this book should not be regarded as affecting the validity of any trademark or service mark.

Warning and Disclaimer

Every effort has been made to make this book as complete and as accurate as possible, but no warranty or fitness is implied. The information provided is on an "as is" basis. The author and the publisher shall have neither liability nor responsibility to any person or entity with respect to any loss or damages arising from the information contained in this book or from the use of the DVD or programs accompanying it.

Bulk Sales

Que Publishing offers excellent discounts on this book when ordered in quantity for bulk purchases or special sales. For more information, please contact

U.S. Corporate and Government Sales
1-800-382-3419
corpsales@pearsontechgroup.com

For sales outside of the U.S., please contact

International Sales
international@pearson.com

Associate Publisher
Greg Wiegand

Acquisitions Editor
Rick Kughen

Development Editor
Todd Brakke

Managing Editor
Kristy Hart

Project Editor
Jovana San Nicolas-Shirley

Copy Editor
Bart Reed

Indexer
Erika Millen

Proofreader
Linda Seifert

Technical Editor
Mark Reddin

Publishing Coordinator
Cindy Teeters

Multimedia Developer
Dan Scherf

Book Designer
Ann Jones

Compositor
Nonie Ratcliff

Introduction

Welcome to *Upgrading and Repairing PCs, 19th Edition*. Since debuting as the first book of its kind on the market in 1988, no other book on PC hardware has matched the depth and quality of the information found in this tome. This edition continues *Upgrading and Repairing PCs*' role as not only the best-selling book of its type, but also the most comprehensive and complete PC hardware reference available. This book examines PCs in depth, outlines the differences among them, and presents options for configuring each system.

More than just a minor revision, the 19th edition of *Upgrading and Repairing PCs* contains much new, revised, and reworked content. The PC industry is moving faster than ever, and this book is the most accurate, complete, and in-depth book of its kind on the market today.

I wrote this book for all PC enthusiasts who want to know everything about their PCs: how they originated; how they've evolved; how to upgrade, troubleshoot, and repair them; and everything in between. This book covers the full gamut of PC-compatible systems, from the oldest 8-bit machines to the latest high-end 64-bit multi-core processors and systems. If you need to know everything about PC hardware from the original to the latest technology on the market today, this book and the accompanying information-packed disc is definitely for you.

Upgrading and Repairing PCs also doesn't ignore the less glamorous PC components. Every part of your PC plays a critical role in its stability and performance. Over the course of this book, you'll find out exactly why your motherboard's chipset might just be the most important part of your PC and what can go wrong when you settle for a run-of-the-mill power supply that can't get enough juice to that monster processor you just bought. You'll also find in-depth coverage of technologies such as new processors, chipsets, graphics, audio cards, PCI Express 2.x, Blu-ray drives, Serial ATA, USB and FireWire, and more—it's all in here, right down to the guts-level analysis of your mouse and keyboard.

Book Objectives

Upgrading and Repairing PCs focuses on several objectives. The primary objective is to help you learn how to maintain, upgrade, and repair your PC system. To that end, *Upgrading and Repairing PCs* helps you fully understand the family of computers that has grown from the original IBM PC, including all PC-compatible systems. This book discusses all areas of system improvement, such as motherboards, processors, memory, and even case and power-supply improvements. The book discusses proper system and component care, specifies the most failure-prone items in various PC systems, and tells you how to locate and identify a failing component. You'll learn about powerful diagnostics hardware and software that enable a system to help you determine the cause of a problem and how to repair it.

PCs are moving forward rapidly in power and capabilities. Processor performance increases with every new chip design. *Upgrading and Repairing PCs* helps you gain an understanding of all the processors used in PC-compatible computer systems.

This book covers the important differences between major system architectures, from the original Industry Standard Architecture (ISA) to the latest PCI Express interface standards. *Upgrading and Repairing PCs* covers each of these system architectures and their adapter boards to help you make decisions about which type of system you want to buy in the future and to help you upgrade and troubleshoot such systems.

The amount of storage space available to modern PCs is increasing geometrically. *Upgrading and Repairing PCs* covers storage options ranging from larger, faster hard drives to state-of-the-art storage devices.

When you finish reading this book, you should have the knowledge to upgrade, troubleshoot, and repair almost any system and component.

The 19th Edition DVD-ROM

The 19th edition of *Upgrading and Repairing PC*s includes a DVD containing valuable content that greatly enhances this book!

First, there's the all-new DVD video with new segments covering important PC components like processors, motherboards, chipsets, memory, hard disk drives, chassis, and power supplies. This includes tips that will help you select the best components when building or purchasing new systems and when upgrading or repairing existing ones. You'll also find a complete step-by-step guide to building a new PC from scratch, including tips that will make the build process run smoothly and help you build a durable, high-performance system that will be easy to upgrade or repair in the future.

The DVD-ROM content includes my venerable Technical Reference material, a repository of reference information that has appeared in previous editions of *Upgrading and Repairing PCs* but has been moved to the disc to make room for coverage of newer technologies. The DVD-ROM also includes the complete 16th Edition of this book, a comprehensive PC glossary, a detailed list of acronyms, and much more available in printable PDF format. There's more PC hardware content and knowledge here than you're likely to find from any other single source.

My Website: informit.com/upgrading

Don't forget about the InformIT Upgrading website! Here, you'll find a cache of helpful material to go along with the book you're holding. I've loaded this site with tons of material—mine as well as from other authors—ranging from video clips to book content and technology updates.

If you find that the video on this book's disc isn't enough, you'll find even more of my previously recorded videos on the website. Not to mention that it is the best place to look for information on all of Que's *Upgrading and Repairing* titles.

I also have a private forum (www.forum.scottmueller.com) designed exclusively to support those who have purchased my recent books and DVDs. I use the forum to answer questions and otherwise help my loyal readers. If you own one of my current books or DVDs, feel free to join in and post questions. I endeavor to answer each and every question personally, but I also encourage knowledgeable members to respond as well. Anybody can view the forum without registering, but to post a question of your own you will need to join. Even if you don't join in, the forum is a tremendous resource because you can still benefit from all the reader questions I have answered over the years.

Be sure to check the informit.com/upgrading website for more information on all my latest books, videos, articles, FAQs, and more!

A Personal Note

When asked which was his favorite Corvette, Dave McLellan, former manager of the Corvette platform at General Motors, always said, "Next year's model." Now with the new 19th edition, next year's model has just become this year's model, until *next* year that is....

I believe this book is absolutely the best book of its kind on the market, and that is due in large part to the extensive feedback I have received from both my seminar attendees and book readers. I am so grateful to everyone who has helped me with this book through each edition, as well as all the loyal readers who have been using this book, many of you since the first edition was published. I have had personal contact with many thousands of you in the seminars I have been teaching since 1982, and I enjoy your comments and even your criticisms tremendously. Using this book in a teaching environment has been a major factor in its development. Some of you might be interested to know that I originally began writing this book in early 1985; back then it was self-published and used exclusively in my PC hardware seminars before being professionally published by Que in 1988.

In one way or another, I have been writing and rewriting this book for more than 20 years! In that time, *Upgrading and Repairing PCs* has proven to be not only the first but also the most comprehensive and yet approachable and easy-to-understand book of its kind. With this new edition, it is even better than ever. Your comments, suggestions, and support have helped this book to become the best PC hardware book on the market. I look forward to hearing your comments after you see this exciting new edition.

Scott

Chapter

6

Memory

Memory Basics

This chapter discusses memory from both a physical and logical point of view. First, we'll examine what memory is, where it fits into the PC architecture, and how it works. Then we'll look at the various types of memory, speeds, and packaging of the chips and memory modules you can buy and install.

This chapter also covers the logical layout of memory, defining the various areas of memory and their uses from the system's point of view. Because the logical layout and uses are within the "mind" of the processor, memory mapping and logical layout remain perhaps the most difficult subjects to grasp in the PC universe. This chapter contains useful information that removes the mysteries associated with memory and enables you to get the most out of your system.

Memory is the workspace for the processor. It is a temporary storage area where the programs and data being operated on by the processor must reside. Memory storage is considered temporary because the data and programs remain there only as long as the computer has electrical power or is not reset. Before the computer is shut down or reset, any data that has been changed should be saved to a more permanent storage device (usually a hard disk) so it can be reloaded into memory in the future.

Memory often is called *RAM*, for *random access memory*. Main memory is called RAM because you can randomly (as opposed to sequentially) access any location in memory. This designation is somewhat misleading and often misinterpreted. Read-only memory (ROM), for example, is also randomly accessible, yet is usually differentiated from the system RAM because it maintains data without power and can't normally be written to. Although a hard disk can be used as virtual random access memory, we don't consider that RAM either.

Over the years, the definition of RAM has changed from a simple acronym to become something that means the primary memory workspace the processor uses to run programs, which usually is constructed of a type of chip called dynamic RAM (DRAM). One of the characteristics of DRAM chips (and therefore most types of RAM in general) is that they store data dynamically, which really has two meanings. One meaning is that the information can be written to RAM repeatedly at any time. The other has to do with the fact that DRAM requires the data to be refreshed (essentially rewritten) every few milliseconds or so; faster RAM requires refreshing more often than slower RAM. A type of RAM called static RAM (SRAM) does not require the periodic refreshing. An important characteristic of RAM in general is that data is stored only as long as the memory has electrical power.

Note

Both DRAM and SRAM memory maintain their contents only as long as power is present. However, a different type of memory known as flash memory does not. Flash memory can retain its contents without power, and it is most commonly used today in digital camera and player media and USB flash drives. As far as the PC is concerned, a flash memory device emulates a disk drive (not RAM) and is accessed by a drive letter, just as with any other disk or optical drive.

When we talk about a computer's memory, we usually mean the RAM or physical memory in the system, which is mainly the memory chips or modules the processor uses to store primary active programs and data. This often is confused with the term *storage*, which should be used when referring to things such as disk and tape drives (although they can be used as a form of RAM called virtual memory).

RAM can refer to both the physical chips that make up the memory in the system and the logical mapping and layout of that memory. *Logical mapping* and *layout* refer to how the memory addresses are mapped to actual chips and what address locations contain which types of system information.

People new to computers often confuse main memory (RAM) with disk storage because both have capacities that are expressed in similar megabyte or gigabyte terms. The best analogy to explain the relationship between memory and disk storage I've found is to think of an office with a desk and a file cabinet.

In this popular analogy, the file cabinet represents the system's hard disk, where both programs and data are stored for long-term safekeeping. The desk represents the system's main memory, which allows the person working at the desk (acting as the processor) direct access to any files placed on it. Files represent the programs and documents you can "load" into the memory. For you to work on a particular file, it must first be retrieved from the cabinet and placed on the desk. If the desk is large enough, you might be able to have several files open on it at one time; likewise, if your system has more memory, you can run more or larger programs and work on more or larger documents.

Adding hard disk space to a system is similar to putting a bigger file cabinet in the office—more files can be permanently stored. And adding more memory to a system is like getting a bigger desk—you can work on more programs and data at the same time.

One difference between this analogy and the way things really work in a computer is that when a file is loaded into memory, it is a copy of the file that is actually loaded; the original still resides on the hard disk. Because of the temporary nature of memory, any files that have been changed after being loaded into memory must then be saved back to the hard disk before the system is powered off (which erases the memory). If the changed file in memory is not saved, the original copy of the file on the hard disk remains unaltered. This is like saying that any changes made to files left on the desktop are discarded when the office is closed, although the original files are still preserved in the cabinet.

Memory temporarily stores programs when they are running, along with the data being used by those programs. RAM chips are sometimes termed *volatile storage* because when you turn off your computer or an electrical outage occurs, whatever is stored in RAM is lost unless you saved it to your hard drive. Because of the volatile nature of RAM, many computer users make it a habit to save their work frequently—a habit I recommend. Many software applications perform periodic saves automatically in order to minimize the potential for data loss.

Physically, the *main memory* in a system is a collection of chips or modules containing chips that are usually plugged into the motherboard. These chips or modules vary in their electrical and physical designs and must be compatible with the system into which they are being installed to function properly. This chapter discusses the various types of chips and modules that can be installed in different systems.

To better understand physical memory in a system, you should understand what types of memory are found in a typical PC and what the role of each type is. Three main types of physical memory are used in modern PCs. (Remember, I'm talking about the type of memory chip, not the type of module that memory is stored on.)

- **ROM**—Read-only memory
- **DRAM**—Dynamic random access memory
- **SRAM**—Static RAM

The only type of memory you normally need to purchase and install in a system is DRAM. The other types are built in to the motherboard (ROM), processor (SRAM), and other components such as the video card, hard drives, and so on.

ROM

Read-only memory, or ROM, is a type of memory that can permanently or semipermanently store data. It is called read-only because it is either impossible or difficult to write to. ROM also is often referred to as *nonvolatile memory* because any data stored in ROM remains there, even if the power is turned off. As such, ROM is an ideal place to put the PC's startup instructions—that is, the software that boots the system.

Note that ROM and RAM are not opposites, as some people seem to believe. Both are simply types of memory. In fact, ROM could be classified as technically a subset of the system's RAM. In other words, a portion of the system's random access memory address space is mapped into one or more ROM chips. This is necessary to contain the software that enables the PC to boot up; otherwise, the processor would have no program in memory to execute when it was powered on.

The main ROM BIOS is contained in a ROM chip on the motherboard, but there are also adapter cards with ROMs on them as well. ROMs on adapter cards contain auxiliary BIOS routines and drivers needed by the particular card, especially for those cards that must be active early in the boot process, such as video cards. Cards that don't need drivers active at boot time typically don't have a ROM because those drivers can be loaded from the hard disk later in the boot process.

Most systems today use a type of ROM called *electrically erasable programmable ROM (EEPROM)*, which is a form of flash memory. Flash is a truly nonvolatile memory that is rewritable, enabling users to easily update the ROM or firmware in their motherboards or any other components (video cards, SCSI cards, peripherals, and so on).

◀◀ For more information on BIOS upgrades, see "Upgrading the BIOS," p. 328 (Chapter 5, "BIOS").

DRAM

Dynamic RAM (DRAM) is the type of memory chip used for most of the main memory in a modern PC. The main advantages of DRAM are that it is very dense, meaning you can pack a lot of bits into a very small chip, and it is inexpensive, which makes purchasing large amounts of memory affordable.

The memory cells in a DRAM chip are tiny capacitors that retain a charge to indicate a bit. The problem with DRAM is that it is dynamic—that is, its contents can be changed. With every keystroke or every mouse swipe, the contents of RAM change. And the entire contents of RAM can be wiped out by a system crash. Also, because of the design, it must be constantly refreshed; otherwise, the electrical charges in the individual memory capacitors will drain and the data will be lost. Refresh occurs when the system memory controller takes a tiny break and accesses all the rows of data in the memory chips. The standard refresh time is 15ms (milliseconds), which means that every 15ms, all the rows in the memory are automatically read to refresh the data.

◀◀ See "Chipsets," p. 198 (Chapter 4).

Refreshing the memory unfortunately takes processor time away from other tasks because each refresh cycle takes several CPU cycles to complete. In older systems, the refresh cycling could take up to 10% or more of the total CPU time, but with modern systems running in the multigigahertz range, refresh overhead is now on the order of a fraction of a percent or less of the total CPU time. Some systems allow you to alter the refresh timing parameters via the CMOS Setup. The time between refresh cycles is known as *tREF* and is expressed not in milliseconds, but in clock cycles (see Figure 6.1).

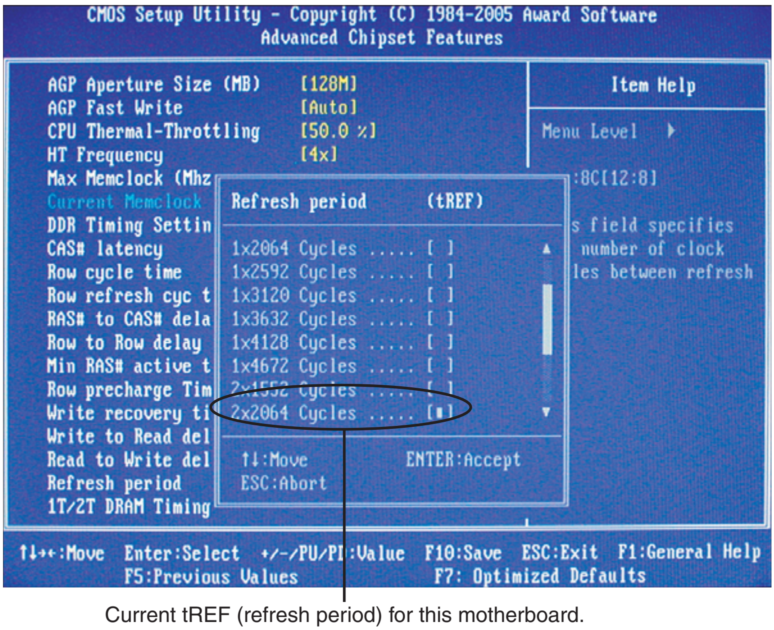


Figure 6.1 The refresh period dialog box and other advanced memory timings can be adjusted manually through the BIOS Setup program.

It's important to be aware that increasing the time between refresh cycles (tREF) to speed up your system can allow some of the memory cells to begin draining prematurely, which can cause random soft memory errors to appear.

A *soft error* is a data error that is not caused by a defective chip. To avoid soft errors, it is usually safer to stick with the recommended or default refresh timing. Because refresh consumes less than 1% of modern system overall bandwidth, altering the refresh rate has little effect on performance. It is almost always best to use default or automatic settings for any memory timings in the BIOS Setup.

Many modern systems don't allow changes to memory timings and are permanently set to automatic settings. On an automatic setting, the motherboard reads the timing parameters out of the serial presence detect (SPD) ROM found on the memory module and sets the cycling speeds to match.

DRAMs use only one transistor and capacitor pair per bit, which makes them very dense, offering more memory capacity per chip than other types of memory. Currently, DRAM chips are being prepared for production with densities up to 4Gb (512MB) per chip, which at one transistor per bit requires at least 4 billion transistors. The transistor count in memory chips is much higher than in processors, because in a memory chip the transistors and capacitors are all consistently arranged in a (normally square) grid of simple repetitive structures, unlike processors, which are much more complex circuits of different structures and elements interconnected in a highly irregular fashion.

The transistor for each DRAM bit cell reads the charge state of the adjacent capacitor. If the capacitor is charged, the cell is read to contain a 1; no charge indicates a 0. The charge in the tiny capacitors is constantly draining, which is why the memory must be refreshed constantly. Even a momentary power interruption, or anything that interferes with the refresh cycles, can cause a DRAM memory cell to lose the charge and thus the data. If this happens in a running system, it can lead to blue screens, global protection faults, corrupted files, and any number of system crashes.

DRAM is used in PC systems because it is inexpensive and the chips can be densely packed, so a lot of memory capacity can fit in a small space. Unfortunately, DRAM is also relatively slow, typically much slower than the processor. For this reason, many types of DRAM architectures have been developed to improve performance. These architectures are covered later in the chapter.

Cache Memory: SRAM

Another distinctly different type of memory exists that is significantly faster than most types of DRAM. SRAM stands for *static RAM*, which is so named because it does not need the periodic refresh rates like DRAM. Because of how SRAMs are designed, not only are refresh rates unnecessary, but SRAM is much faster than DRAM and much more capable of keeping pace with modern processors.

SRAM memory is available in access times of 0.45ns or less, so it can keep pace with processors running 2.2GHz or faster. This is because of the SRAM design, which calls for a cluster of six transistors for each bit of storage. The use of transistors but no capacitors means that refresh rates are not necessary because there are no capacitors to lose their charges over time. As long as there is power, SRAM remembers what is stored. With these attributes, why don't we use SRAM for all system memory? The answers are simple.

Compared to DRAM, SRAM is much faster but also much lower in density and much more expensive (see Table 6.1). The lower density means that SRAM chips are physically larger and store fewer bits overall. The high number of transistors and the clustered design mean that SRAM chips are both physically larger and much more expensive to produce than DRAM chips. For example, a high-density DRAM chip might store up to 4Gb (512MB) of RAM, whereas similar sized SRAM chips can only store up to 72Mb (9MB). The high cost and physical constraints have prevented SRAM from being used as the main memory for PC systems.

Table 6.1 Comparing DRAM and SRAM

Type	Speed	Density	Cost
DRAM	Slow	High	Low
SRAM	Fast	Low	High

Even though SRAM is impractical for PC use as main memory, PC designers have found a way to use SRAM to dramatically improve PC performance. Rather than spend the money for all RAM to be SRAM memory, they design in a small amount of high-speed SRAM memory, used as *cache memory*, which is much more cost-effective. The SRAM cache runs at speeds close to or even equal to the processor and is the memory from which the processor usually directly reads from and writes to. During read operations, the data in the high-speed cache memory is resupplied from the lower-speed main memory or DRAM in advance. To convert access time in nanoseconds to MHz, use the following formula:

$$1 / \text{nanoseconds} \times 1000 = \text{MHz}$$

Likewise, to convert from MHz to nanoseconds, use the following inverse formula:

$$1 / \text{MHz} \times 1000 = \text{nanoseconds}$$

Today we have memory that runs faster than 1GHz (1 nanosecond), but up until the late 1990s, DRAM was limited to about 60ns (16MHz) in speed. Up until processors were running at speeds of 16MHz, the available DRAM could fully keep pace with the processor and motherboard, meaning that there was no need for cache. However, as soon as processors crossed the 16MHz barrier, the available DRAM could no longer keep pace, and SRAM cache began to enter PC system designs. This occurred way back in 1986 and 1987 with the debut of systems with the 386 processor running at speeds of 16MHz to 20MHz or faster. These were among the first PC systems to employ what's called *cache memory*, a high-speed buffer made up of SRAM that directly feeds the processor. Because the cache can run at the speed of the processor, it acts as a buffer between the processor and the slower DRAM in the system. The cache controller anticipates the processor's memory needs and preloads the high-speed cache memory with data. Then, as the processor calls for a memory address, the data can be retrieved from the high-speed cache rather than the much lower-speed main memory.

Cache effectiveness can be expressed by a hit ratio. This is the ratio of cache hits to total memory accesses. A *hit* occurs when the data the processor needs has been preloaded into the cache from the main memory, meaning the processor can read it from the cache. A cache *miss* is when the cache controller did not anticipate the need for a specific address and the desired data was not preloaded into the cache. In that case the processor must retrieve the data from the slower main memory, instead of the faster cache. Any time the processor reads data from main memory, the processor must wait longer because the main memory cycles at a much slower rate than the processor. As an example, if the processor with integral on-die cache is running at 3.6GHz (3,600MHz) on a 1,333MHz bus, both the processor and the integral cache would be cycling at 0.28ns, while the main memory would most likely be cycling almost five times more slowly at 1,333MHz (0.75ns). So, every time the 3.6GHz processor reads from main memory, it would effectively slow down to only 1,333MHz. The slowdown is accomplished by having the processor execute what are called *wait states*, which are cycles in which nothing is done; the processor essentially cools its heels while waiting for the slower main memory to return the desired data. Obviously, you don't want your processors slowing down, so cache function and design become more important as system speeds increase.

To minimize the processor being forced to read data from the slow main memory, two or three stages of cache usually exist in a modern system, called Level 1 (L1), Level 2 (L2), and Level 3 (L3). The L1 cache is also called *integral* or *internal cache* because it has always been built directly into the processor as part of the processor die (the raw chip). Because of this, L1 cache always runs at the full speed of the processor core and is the fastest cache in any system. All 486 and higher processors incorporate integral L1 cache, making them significantly faster than their predecessors. L2 cache was originally called *external cache* because it was external to the processor chip when it first appeared. Originally, this meant it was installed on the motherboard, as was the case with all 386, 486, and first-generation Pentium systems. In those systems, the L2 cache runs at motherboard and CPU bus speed because it is installed on the motherboard and is connected to the CPU bus. You typically find the L2 cache physically adjacent to the processor socket in Pentium and earlier systems.

◀◀ See "Cache Memory," p. 64 (Chapter 3, "Processor Types and Specifications").

In the interest of improved performance, later processor designs from Intel and AMD included the L2 cache as a part of the processor. In all processors since late 1999 (and some earlier models), the L2 cache is directly incorporated as a part of the processor die, just like the L1 cache. In chips with on-die L2, the cache runs at the full core speed of the processor and is much more efficient. By contrast, most processors from 1999 and earlier with integrated L2 had the L2 cache in separate chips that were external to the main processor core. The L2 cache in many of these older processors ran at only half or one-third the processor core speed. Cache speed is very important, so systems having L2 cache on the motherboard were the slowest. Including L2 inside the processor made it faster, and including it directly on the processor die (rather than as chips external to the die) is the fastest yet. Any chip that has on-die full core speed L2 cache has a distinct performance advantage over any chip that doesn't.

A third-level or L3 cache has been present in some processors since 2001. The first desktop PC processor with L3 cache was the Pentium 4 Extreme Edition, a high-end chip introduced in late 2003 with 2MB of on-die L3 cache. Although it seemed at the time that this would be a forerunner of widespread L3 cache in desktop processors, later versions of the Pentium 4 Extreme Edition (as well as its successor, the Pentium Extreme Edition) dropped the L3 cache, instead using larger L2 cache sizes to improve performance. L3 cache made a return to PC processors in 2007 with the AMD Phenom and in 2008 with the Intel Core i7, both of which have four cores on a single die. L3 is especially suited to processors with four or more cores because it provides an on-die cache that all the cores can share. I expect L3 cache to be a staple in future multicore processors.

The key to understanding both cache and main memory is to see where they fit in the overall system architecture. See Chapter 4 for diagrams showing recent systems with different types of cache memory.

RAM Types and Performance

The speed and performance issue with memory is confusing to some because memory speed is sometimes expressed in nanoseconds (ns) and processor speed has always been expressed in megahertz (MHz) or gigahertz (GHz). Newer and faster types of memory usually have speeds expressed in MHz, thus adding to the confusion. Fortunately, you can easily translate MHz/GHz to ns, and vice versa.

A *nanosecond* is defined as one billionth of a second—a very short time indeed. To put some perspective on that, the speed of light is 186,282 miles (299,792 kilometers) per second in a vacuum. In one billionth of a second, a beam of light travels a mere 11.80 inches or 29.98 centimeters—less than the length of a typical ruler!

Chip and system speeds have often been expressed in megahertz (MHz), which is millions of cycles per second, or gigahertz (GHz), which is billions of cycles per second. Today's processors run in the 2GHz–4GHz range with most performance improvements coming from changes in CPU design (such as multiple cores) rather than pure clock speed increases.

Because it is confusing to speak in these different terms for speeds, I thought it would be interesting to see how they compare. Earlier in this chapter I listed formulas you could use to mathematically convert these values. Table 6.2 shows the relationship between common nanosecond (ns) and megahertz (MHz) speeds associated with PCs from yesterday to today and tomorrow.

Table 6.2 The Relationship Between Megahertz (MHz) and Cycle Times in Nanoseconds (ns)

Clock Speed	Cycle Time	Clock Speed	Cycle Time	Clock Speed	Cycle Time	Clock Speed	Cycle Time
4.77MHz	210ns	250MHz	4.0ns	850MHz	1.18ns	2,700MHz	0.37ns
6MHz	167ns	266MHz	3.8ns	866MHz	1.15ns	2,800MHz	0.36ns
8MHz	125ns	300MHz	3.3ns	900MHz	1.11ns	2,900MHz	0.34ns
10MHz	100ns	333MHz	3.0ns	933MHz	1.07ns	3,000MHz	0.333ns
12MHz	83ns	350MHz	2.9ns	950MHz	1.05ns	3,100MHz	0.323ns
16MHz	63ns	366MHz	2.7ns	966MHz	1.04ns	3,200MHz	0.313ns
20MHz	50ns	400MHz	2.5ns	1,000MHz	1.00ns	3,300MHz	0.303ns
25MHz	40ns	433MHz	2.3ns	1,100MHz	0.91ns	3,400MHz	0.294ns
33MHz	30ns	450MHz	2.2ns	1,133MHz	0.88ns	3,500MHz	0.286ns
40MHz	25ns	466MHz	2.1ns	1,200MHz	0.83ns	3,600MHz	0.278ns
50MHz	20ns	500MHz	2.0ns	1,300MHz	0.77ns	3,700MHz	0.270ns
60MHz	17ns	533MHz	1.88ns	1,400MHz	0.71ns	3,800MHz	0.263ns
66MHz	15ns	550MHz	1.82ns	1,500MHz	0.67ns	3,900MHz	0.256ns
75MHz	13ns	566MHz	1.77ns	1,600MHz	0.63ns	4,000MHz	0.250ns
80MHz	13ns	600MHz	1.67ns	1,700MHz	0.59ns	4,100MHz	0.244ns
100MHz	10ns	633MHz	1.58ns	1,800MHz	0.56ns	4,200MHz	0.238ns
120MHz	8.3ns	650MHz	1.54ns	1,900MHz	0.53ns	4,300MHz	0.233ns
133MHz	7.5ns	666MHz	1.50ns	2,000MHz	0.50ns	4,400MHz	0.227ns
150MHz	6.7ns	700MHz	1.43ns	2,100MHz	0.48ns	4,500MHz	0.222ns
166MHz	6.0ns	733MHz	1.36ns	2,200MHz	0.45ns	4,600MHz	0.217ns
180MHz	5.6ns	750MHz	1.33ns	2,300MHz	0.43ns	4,700MHz	0.213ns
200MHz	5.0ns	766MHz	1.31ns	2,400MHz	0.42ns	4,800MHz	0.208ns
225MHz	4.4ns	800MHz	1.25ns	2,500MHz	0.40ns	4,900MHz	0.204ns
233MHz	4.3ns	833MHz	1.20ns	2,600MHz	0.38ns	5,000MHz	0.200ns

As you can see from Table 6.2, as clock speeds increase, cycle time decreases proportionately.

Over the development life of the PC, memory has had a difficult time keeping up with the processor, requiring several levels of high-speed cache memory to intercept processor requests for the slower main memory. More recently, however, systems using DDR, DDR2, and DDR3 SDRAM have memory bus performance equaling that of the processor bus. When the speed of the memory bus equals the speed of the processor bus, main memory performance is optimum for that system.

For example, using the information in Table 6.2, you can see that the 60ns DRAM memory used in the original Pentium and Pentium II PCs up until 1998 works out to be an extremely slow 16.7MHz! This slow 16.7MHz memory was installed in systems running processors up to 300MHz or faster on a processor bus speed of 66MHz, resulting in a large mismatch between processor bus and main memory performance. However, starting in 1998 the industry shifted to faster SDRAM memory, which was able to match the 66MHz speed of the processor bus at the time. From that point forward, memory has largely evolved in step with the processor bus, with newer and faster types coming out to match any increases in processor bus speeds.

By the year 2000, the dominant processor bus and memory speeds had increased to 100MHz and even 133MHz (called PC100 and PC133 SDRAM, respectively). Starting in early 2001, double data rate (DDR) SDRAM memory of 200MHz and 266MHz become popular. In 2002, DDR memory increased to 333MHz; in 2003, the speeds increased further to 400MHz. During 2004, we saw the introduction of DDR2, first at 400MHz and then at 533MHz. DDR2 memory continued to match processor bus speed increases in PCs during 2005 and 2006, rising to 667MHz and 800MHz, respectively, during that time. By 2007, DDR2 memory was available at speeds of up to 1,066MHz, and DDR3 came on the market at 1,066MHz and faster. In 2009, DDR3 memory became the most popular memory type in new systems, with standard speeds of up to 1,600MHz. Table 6.3 lists the primary types and performance levels of PC memory.

Table 6.3 PC Memory Types and Performance Levels

Memory Type	Years Popular	Module Type	Voltage	Max. Clock Speed	Max. Throughput Single-Channel	Max. Throughput Dual-Channel	Max. Throughput Tri-Channel
Fast Page Mode (FPM) DRAM	1987–1995	30/72-pin SIMM	5V	22MHz	177MBps	N/A	N/A
Extended Data Out (EDO) DRAM	1995–1998	72-pin SIMM	5V	33MHz	266MBps	N/A	N/A
Single Data Rate (SDR) SDRAM	1998–2002	168-pin DIMM	3.3V	133MHz	1,066MBps	N/A	N/A
Rambus DRAM (RDRAM)	2000–2002	184-pin RIMM	2.5V	1,066MTps	2,133MBps	4,266MBps	N/A
Double Data Rate (DDR) SDRAM	2002–2005	184-pin DIMM	2.5V	400MTps	3,200MBps	6,400MBps	N/A
DDR2 SDRAM	2005–2008	240-pin DDR2 DIMM	1.8V	1,066MTps	8,533MBps	17,066MBps	N/A
DDR3 SDRAM	2008+	240-pin DDR3 DIMM	1.5V	1,600MTps	12,800MBps	25,600MBps	38,400MBps

MHz = Megacycles per second

MTps = Megatransfers per second

MBps = Megabytes per second

SIMM = Single inline memory module

DIMM = Dual inline memory module

The following sections look at these memory types in more detail.

Fast Page Mode DRAM

Standard DRAM is accessed through a technique called *paging*. Normal memory access requires that a row and column address be selected, which takes time. Paging enables faster access to all the data within a given row of memory by keeping the row address the same and changing only the column. Memory that uses this technique is called *Page Mode* or *Fast Page Mode* memory. Other variations on Page Mode were called *Static Column* or *Nibble Mode* memory.

Paged memory is a simple scheme for improving memory performance that divides memory into pages ranging from 512 bytes to a few kilobytes long. The paging circuitry then enables memory locations in a page to be accessed with fewer wait states. If the desired memory location is outside the current page, one or more wait states are added while the system selects the new page.

To improve further on memory access speeds, systems have evolved to enable faster access to DRAM. One important change was the implementation of burst mode access in the 486 and later processors. Burst mode cycling takes advantage of the consecutive nature of most memory accesses. After setting up the row and column addresses for a given access, using burst mode, you can then access the next three adjacent addresses with no additional latency or wait states. A burst access usually is limited to four total accesses. To describe this, we often refer to the timing in the number of cycles for each access. A typical burst mode access of standard DRAM is expressed as x-y-y-y; x is the time for the first access (latency plus cycle time), and y represents the number of cycles required for each consecutive access.

Standard 60ns-rated DRAM normally runs 5-3-3-3 burst mode timing. This means the first access takes a total of five cycles (on a 66MHz system bus, this is about 75ns total, or 5×15ns cycles), and the consecutive cycles take three cycles each (3×15ns = 45ns). As you can see, the actual system timing is somewhat less than the memory is technically rated for. Without the bursting technique, memory access would be 5-5-5-5 because the full latency is necessary for each memory transfer. The 45ns cycle time during burst transfers equals about a 22.2MHz effective clock rate; on a system with a 64-bit (8-byte) wide memory bus, this would result in a maximum throughput of 177MBps (22.2MHz× 8 bytes = 177MBps).

DRAM memory that supports paging and this bursting technique is called *Fast Page Mode (FPM)* memory. The term comes from the capability of memory accesses to data on the same page to be done with less latency. Most 386, 486, and Pentium systems from 1987 through 1995 used FPM memory, which came in either 30-pin or 72-pin SIMM form.

Another technique for speeding up FPM memory is called *interleaving*. In this design, two separate banks of memory are used together, alternating access from one to the other as even and odd bytes. While one is being accessed, the other is being precharged, when the row and column addresses are being selected. Then, by the time the first bank in the pair is finished returning data, the second bank in the pair is finished with the latency part of the cycle and is now ready to return data. While the second bank is returning data, the first bank is being precharged, selecting the row and column address of the next access. This overlapping of accesses in two banks reduces the effect of the latency or precharge cycles and allows for faster overall data retrieval. The only problem is that to use interleaving, you must install identical pairs of banks together, doubling the number of modules required.

Extended Data Out RAM (EDO)

In 1995, a newer type of DRAM called *extended data out (EDO)* RAM became available for Pentium systems. EDO, a modified form of FPM memory, is sometimes referred to as *Hyper Page mode*. EDO was invented and patented by Micron Technology, although Micron licensed production to many other memory manufacturers.

EDO memory consists of specially manufactured chips that allow a timing overlap between successive accesses. The name *extended data out* refers specifically to the fact that unlike FPM, the data output drivers on the chip are not turned off when the memory controller removes the column address to begin the next cycle. This enables the next cycle to overlap the previous one, saving approximately 10ns per cycle.

The effect of EDO is that cycle times are improved by enabling the memory controller to begin a new column address instruction while it is reading data at the current address. This is almost identical to what was achieved in older systems by interleaving banks of memory, but unlike interleaving, with EDO you didn't need to install two identical banks of memory in the system at a time.

EDO RAM allows for burst mode cycling of 5-2-2-2, compared to the 5-3-3-3 of standard fast page mode memory. To do four memory transfers, then, EDO would require 11 total system cycles, compared to 14 total cycles for FPM. This is a 22% improvement in overall cycling time. The resulting two-cycle (30ns) cycle time during burst transfers equals a 33.3MHz effective clock rate, compared to 45ns/22MHz for FPM. On a system with a 64-bit (8-byte) wide memory bus, this would result in a maximum throughput of 266MBps (33.3MHz×8 bytes = 266MBps). Due to the processor cache, EDO typically increased overall system benchmark speed by only 5% or less. Even though the overall system improvement was small, the important thing about EDO was that it used the same basic DRAM chip design as FPM, meaning that there was practically no additional cost over FPM. In fact, in its heyday EDO cost less than FPM and yet offered higher performance.

EDO RAM generally came in 72-pin SIMM form. Figure 6.4 (later in this chapter) shows the physical characteristics of these SIMMs.

To actually use EDO memory, your motherboard chipset had to support it. Most motherboard chipsets introduced on the market from 1995 (Intel 430FX) through 1997 (Intel 430TX) offered support for EDO, making EDO the most popular form of memory in PCs from 1995 through 1998. Because EDO memory chips cost the same to manufacture as standard chips, combined with Intel's support of EDO in motherboard chipsets, the PC market jumped on the EDO bandwagon full force.

◀◀ See "Fifth-Generation (P5 Pentium Class) Chipsets," p. 207 and "Sixth-Generation (P6 Pentium Pro/II/III Class) Chipsets," p. 209 (Chapter 4).

EDO RAM was used in systems with CPU bus speeds of up to 66MHz, which fit perfectly with the PC market up through 1998. However, starting in 1998, with the advent of 100MHz and faster system bus speeds, the market for EDO rapidly declined, and faster SDRAM architecture became the standard.

One variation of EDO that never caught on was called *burst EDO (BEDO)*. BEDO added burst capabilities for even speedier data transfers than standard EDO. Unfortunately, the technology was owned by Micron and not a free industry standard, so only one chipset (Intel 440FX Natoma) ever supported it. BEDO was quickly overshadowed by industry-standard SDRAM, which came into favor among PC system chipset and system designers over proprietary designs. As such, BEDO never really saw the light of production, and to my knowledge no systems ever used it.

SDRAM

SDRAM is short for *synchronous DRAM*, a type of DRAM that runs in synchronization with the memory bus. SDRAM delivers information in very high-speed bursts using a high-speed clocked interface. SDRAM removes most of the latency involved in asynchronous DRAM because the signals are already in synchronization with the motherboard clock.

As with any type of memory on the market, motherboard chipset support is required before it can be usable in systems. Starting in 1996 with the 430VX and 430TX, most of Intel's chipsets began to support industry-standard SDRAM, and in 1998 the introduction of the 440BX chipset caused SDRAM to eclipse EDO as the most popular type on the market.

SDRAM performance is dramatically improved over that of FPM or EDO RAM. However, because SDRAM is still a type of DRAM, the initial latency is the same, but burst mode cycle times are much faster than with FPM or EDO. SDRAM timing for a burst access would be 5-1-1-1, meaning that four memory reads would complete in only eight system bus cycles, compared to 11 cycles for EDO and 14 cycles for FPM. This makes SDRAM almost 20% faster than EDO.

Besides being capable of working in fewer cycles, SDRAM is also capable of supporting up to 133MHz (7.5ns) system bus cycling. Most PC systems sold from 1998 through 2002 included SDRAM memory.

SDRAM is sold in DIMM form and is normally rated by clock speed (MHz) rather than cycling time (ns), which was confusing during the initial change from FPM and EDO DRAM. Figure 6.5 (later in this chapter) shows the physical characteristics of DIMMs.

To meet the stringent timing demands of its chipsets, Intel created specifications for SDRAM called PC66, PC100, and PC133. For example, you would think 10ns would be considered the proper rating for 100MHz operation, but the PC100 specification promoted by Intel calls for faster 8ns memory to ensure all timing parameters could be met with sufficient margin for error.

In May 1999, the Joint Electron Device Engineering Council (JEDEC) created a specification called PC133. It achieved this 33MHz speed increase by taking the PC100 specification and tightening up the timing and capacitance parameters. The faster PC133 quickly caught on for any systems running a 133MHz processor bus. The original chips used in PC133 modules were rated for exactly 7.5ns or 133MHz; later ones were rated at 7.0ns, which is technically 143MHz. These faster chips were still used on PC133 modules, but they allowed for improvements in column address strobe latency (abbreviated as CAS or CL), which somewhat improves overall memory cycling time.

Note

JEDEC is the semiconductor engineering standardization body of the Electronic Industries Alliance (EIA), a trade association that represents all areas of the electronics industry. JEDEC was originally created in 1960 and governs the standardization of all types of semiconductor devices, integrated circuits, and modules. JEDEC has about 300 member companies, including memory, chipset, and processor manufacturers as well as practically any company involved in manufacturing computer equipment using industry-standard components.

The idea behind JEDEC is simple: to create open standards that can be freely adopted throughout the industry. For example, if one company were to create a proprietary memory technology, other companies who wanted to manufacture components compliant with that memory would have to pay license fees, assuming the company that owned the technology was interested in licensing at all! Parts would be more proprietary in nature, causing problems with interchangeability or sourcing reasonably priced replacements. In addition, those companies licensing the technology would have no control over the evolution of the technology or any future changes made by the owner company.

JEDEC prevents this type of scenario for things such as memory by getting all the memory manufacturers to work together to create shared industry standards covering memory chips and modules. JEDEC-approved standards for memory can then be freely shared by all the member companies, and no one single company has control over a given standard, or any of the companies producing compliant components. FPM, SDRAM, DDR, DDR2, and DDR3 are all examples of JEDEC memory standards used in PCs, whereas memory such as EDO and RDRAM are proprietary examples. You can find out more about JEDEC standards for memory and other semiconductor technology at www.jedec.org.

Table 6.4 shows the timing, rated chip speeds, and standard module speeds for various SDRAM DIMMs.

Table 6.4 SDRAM Timing, Actual Speed, and Rated Speed

Timing	Rated Chip Speed	Standard Module Speed
15ns	66MHz	PC66
10ns	100MHz	PC66
8ns	125MHz	PC100
7.5ns	133MHz	PC133
7.0ns	143MHz	PC133

SDRAM normally came in 168-pin DIMMs, running at several different speeds. Table 6.5 shows the standard single data rate SDRAM module speeds and resulting throughputs.

Table 6.5 JEDEC Standard SDRAM Module (168-pin DIMM) Speeds and Transfer Rates

Module Standard	Chip Type	Clock Speed (MHz)	Cycles per Clock	Bus Speed (MTps)	Bus Width (Bytes)	Transfer Rate (MBps)
PC66	10ns	66	1	66	8	533
PC100	8ns	100	1	100	8	800
PC133	7ns	133	1	133	8	1,066

MTps = Megatransfers per second

MBps = Megabytes per second

ns = Nanoseconds (billionths of a second)

DIMM = Dual inline memory module

▶▶ See "SIMMs, DIMMs, and RIMMS," p. 397 (this chapter).

Some module manufacturers sold modules they claimed were "PC150" or "PC166," even though those speeds did not exist as official JEDEC or Intel standards, and no chipsets or processors officially supported those speeds. These modules actually used hand-picked 133MHz rated chips that could run overclocked at 150MHz or 166MHz speeds. In essence, PC150 or PC166 memory was PC133 memory that was tested to run at overclocked speeds not supported by the original chip manufacturer. This overclockable memory was sold at a premium to enthusiasts who wanted to overclock their motherboard chipsets, thereby increasing the speed of the processor and memory bus.

Caution

In general, PC133 memory is considered to be backward compatible with PC100 memory. However, some chipsets or motherboards had more specific requirements for specific types of 100MHz or 133MHz chips and module designs. If you need to upgrade an older system that requires PC100 memory, you should not purchase PC133 memory unless the memory is specifically identified by the memory vendor as being compatible with the system. You can use the online memory-configuration tools provided by most major memory vendors to ensure that you get the right memory for your system.

DDR SDRAM

Double data rate (DDR) SDRAM memory is a JEDEC standard that is an evolutionary upgrade in which data is transferred twice as quickly as standard SDRAM. Instead of doubling the actual clock rate, DDR memory achieves the doubling in performance by transferring twice per transfer cycle: once at the leading (falling) edge and once at the trailing (rising) edge of the cycle (see Figure 6.2). This effectively doubles the transfer rate, even though the same overall clock and timing signals are used.

DDR SDRAM first came to market in the year 2000 and was initially used on high-end graphics cards because there weren't any motherboard chipsets to support it at the time. DDR finally became popular in 2002 with the advent of mainstream supporting motherboards and chipsets. From 2002 through 2005, DDR was the most popular type of memory in mainstream PCs. DDR SDRAM uses a DIMM module design with 184 pins. Figure 6.6 (later in this chapter) shows the 184-pin DDR DIMM.

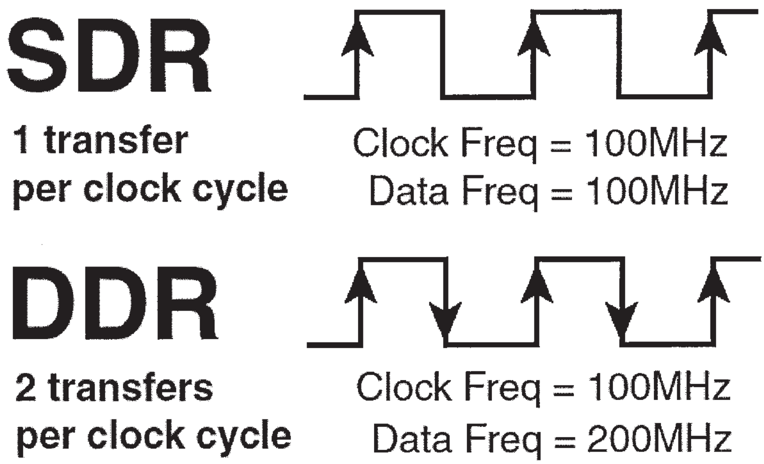


Figure 6.2 SDR (single data rate) versus DDR (double data rate) cycling.

DDR DIMMs come in a variety of speed or throughput ratings and normally run on 2.5 volts. They are basically an extension of the standard SDRAM DIMMs redesigned to support double clocking, where data is sent on each clock transition (twice per cycle) rather than once per cycle as with standard SDRAM. To eliminate confusion with DDR, regular SDRAM is often called *single data rate (SDR)*. Table 6.6 compares the various types of industry-standard DDR SDRAM modules. As you can see, the raw chips are designated by their speed in megatransfers per second, whereas the modules are designated by their approximate throughput in megabytes per second.

Table 6.6 JEDEC Standard DDR Module (184-pin DIMM) Speeds and Transfer Rates

Module Standard	Chip Type	Clock Speed (MHz)	Cycles per Clock	Bus Speed (MTps)	Bus Width (Bytes)	Transfer Rate (MBps)	Dual-Channel Transfer Rate (MBps)
PC1600	DDR200	100	2	200	8	1,600	3,200
PC2100	DDR266	133	2	266	8	2,133	4,266

Table 6.6 Continued

Module Standard	Chip Type	Clock Speed (MHz)	Cycles per Clock	Bus Speed (MTps)	Bus Width (Bytes)	Transfer Rate (MBps)	Dual-Channel Transfer Rate (MBps)
PC2700	DDR333	166	2	333	8	2,667	5,333
PC3200	DDR400	200	2	400	8	3,200	6,400

MTps = Megatransfers per second

MBps = Megabytes per second

DIMM = Dual inline memory module

DDR = Double data rate

The major memory chip and module manufacturers normally produce parts that conform to the official JEDEC standard speed ratings. However, to support overclocking, several memory module manufacturers purchase unmarked and untested chips from the memory chip manufacturers, then independently test and sort them by how fast they run. These are then packaged into modules with unofficial designations and performance figures that exceed the standard ratings. Table 6.7 shows the popular unofficial speed ratings I've seen on the market. Note that because the speeds of these modules are beyond the standard default motherboard and chipset speeds, you won't see any advantage to using them unless you are overclocking your system to match.

Table 6.7 Overclocked (non-JEDEC) DDR Module (184-pin DIMM) Speeds and Transfer Rates

Module Standard	Chip Type	Clock Speed (MHz)	Cycles per Clock	Bus Speed (MTps)	Bus Width (Bytes)	Transfer Rate (MBps)	Dual-Channel Transfer Rate (MBps)
PC3500	DDR433	216	2	433	8	3,466	6,933
PC3700	DDR466	233	2	466	8	3,733	7,466
PC4000	DDR500	250	2	500	8	4,000	8,000
PC4200	DDR533	266	2	533	8	4,266	8,533
PC4400	DDR550	275	2	550	8	4,400	8,800
PC4800	DDR600	300	2	600	8	4,800	9,600

MTps = Megatransfers per second

MBps = Megabytes per second

DIMM = Dual inline memory module

DDR = Double data rate

The bandwidths listed in these tables are per module. Most chipsets that support DDR also support dual-channel operation—a technique in which two matching DIMMs are installed to function as a single bank, with double the bandwidth of a single module. For example, if a chipset supports standard PC3200 modules, the bandwidth for a single module would be 3,200MBps. However, in dual-channel mode, the total bandwidth would double to 6,400MBps. Dual-channel operation optimizes PC design by ensuring that the CPU bus and memory bus both run at exactly the same speeds (meaning throughput, not MHz) so that data can move synchronously between the buses without delays.

DDR2 SDRAM

DDR2 is simply a faster version of DDR memory: It achieves higher throughput by using differential pairs of signal wires to allow faster signaling without noise and interference problems. DDR2 is still double data rate, just as with DDR, but the modified signaling method enables higher clock speeds to be achieved with more immunity to noise and crosstalk between the signals. The additional signals required for differential pairs add to the pin count—DDR2 DIMMs have 240 pins, which is more than the 184 pins of DDR. The original DDR specification officially topped out at 400MHz (although faster unofficial overclocked modules were produced), whereas DDR2 starts at 400MHz and goes up to an official maximum of 1,066MHz. Table 6.8 shows the various official JEDEC-approved DDR2 module types and bandwidth specifications.

Table 6.8 JEDEC Standard DDR2 Module (240-pin DIMM) Speeds and Transfer Rates

Module Standard	Chip Type	Clock Speed (MHz)	Cycles per Clock	Bus Speed (MTps)	Bus Width (Bytes)	Transfer Rate (MBps)	Dual-Channel Transfer Rate (MBps)
PC2-3200	DDR2-400	200	2	400	8	3,200	6,400
PC2-4200	DDR2-533	266	2	533	8	4,266	8,533
PC2-5300	DDR2-667	333	2	667	8	5,333	10,667
PC2-6400	DDR2-800	400	2	800	8	6,400	12,800
PC2-8500	DDR2-1066	533	2	1066	8	8,533	17,066

MTps = Megatransfers per second

MBps = Megabytes per second

DIMM = Dual inline memory module

DDR = Double data rate

The fastest official JEDEC-approved standard is DDR2-1066, which is composed of chips that run at an effective speed of 1,066MHz (really megatransfers per second), resulting in modules designated PC2-8500 having a bandwidth of 8,533MBps. However, just as with DDR, many of the module manufacturers produce even faster modules designed for overclocked systems. These are sold as modules with unofficial designations and performance figures that exceed the standard ratings. Table 6.9 shows the popular unofficial speed ratings I've seen on the market. Note that because the speeds of these modules are beyond the standard default motherboard and chipset speeds, you won't see any advantage to using these unless you are overclocking your system to match.

Table 6.9 Overclocked (non-JEDEC) DDR2 Module (240-pin DIMM) Speeds and Transfer Rates

Module Standard	Chip Type	Clock Speed (MHz)	Cycles per Clock	Bus Speed (MTps)	Bus Width (Bytes)	Transfer Rate (MBps)	Dual-Channel Transfer Rate (MBps)
PC2-6000	DDR2-750	375	2	750	8	6,000	12,000
PC2-7200	DDR2-900	450	2	900	8	7,200	14,400
PC2-8000	DDR2-1000	500	2	1000	8	8,000	16,000
PC2-8800	DDR2-1100	550	2	1100	8	8,800	17,600

Table 6.9 Continued

Module Standard	Chip Type	Clock Speed (MHz)	Cycles per Clock	Bus Speed (MTps)	Bus Width (Bytes)	Transfer Rate (MBps)	Dual-Channel Transfer Rate (MBps)
PC2-8888	DDR2-1111	556	2	1111	8	8,888	17,777
PC2-9136	DDR2-1142	571	2	1142	8	9,136	18,272
PC2-9200	DDR2-1150	575	2	1150	8	9,200	18,400
PC2-9600	DDR2-1200	600	2	1200	8	9,600	19,200
PC2-10000	DDR2-1250	625	2	1250	8	10,000	20,000

MTps = Megatransfers per second
MBps = Megabytes per second
DIMM = Dual inline memory module
DDR = Double data rate

In addition to providing greater speeds and bandwidth, DDR2 has other advantages. It uses lower voltage than conventional DDR (1.8V versus 2.5V), so power consumption and heat generation are reduced. Because of the greater number of pins required on DDR2 chips, the chips typically use fine-pitch ball grid array (FBGA) packaging rather than the thin small outline package (TSOP) chip packaging used by most DDR and conventional SDRAM chips. FBGA chips are connected to the substrate (meaning the memory module in most cases) via tightly spaced solder balls on the base of the chip is.

DDR2 DIMMs resemble conventional DDR DIMMs but have more pins and slightly different notches to prevent confusion or improper application. For example, the different physical notches prevent you from plugging a DDR2 module into a conventional DDR (or SDR) socket. DDR2 memory module designs incorporate 240 pins, significantly more than conventional DDR or standard SDRAM DIMMs.

JEDEC began working on the DDR2 specification in April 1998, and published the standard in September 2003. DDR2 chip and module production actually began in mid-2003 (mainly samples and prototypes), and the first chipsets, motherboards, and systems supporting DDR2 appeared for Intel processor–based systems in mid-2004. At that time variations of DDR2 such as G-DDR2 (Graphics DDR2) began appearing in graphics cards as well. Mainstream motherboard chipset support for DDR2 on Intel processor–based systems appeared in 2005. Notable for its lack of DDR2 support through 2005 was AMD, whose Athlon 64 and Opteron processor families included integrated DDR memory controllers. AMD processor–based systems first supported DDR2 in mid-2006, with the release of socket AM2 motherboards and processors to match. (AMD's Socket F, also known as 1207 FX, also supports DDR2 memory.)

It is interesting to note that AMD was almost 2 years behind Intel in the transition from DDR to DDR2. This is because AMD included the memory controller in its Athlon 64 and all subsequent processors, rather than incorporating the memory controller in the chipset North Bridge, as with the more traditional Intel designs. Although there are advantages to integrating the memory controller in the CPU, one disadvantage is the inability to quickly adopt new memory architectures, because doing so requires that both the processor and processor socket be redesigned. However, with the release of the Core i7 processors in 2008, Intel also moved the memory controller from the chipset into the processor, thus putting Intel and AMD in the same situation as far as memory architectures are concerned.

DDR3

DDR3 is the latest JEDEC memory standard. It enables higher levels of performance along with lower power consumption and higher reliability than DDR2. JEDEC began working on the DDR3 specification in June of 2002, and the first DDR3 memory modules and supporting chipsets (versions of the Intel 3xx series) were released for Intel-based systems in mid-2007. Due to initial high cost and limited support, DDR3 didn't start to become popular until late 2008 when Intel released the Core i7 processor, which included an integrated tri-channel DDR3 memory controller. In early 2009, popularity increased when AMD released Socket AM3 versions of the Phenom II, the first from AMD to support DDR3. During 2009, with full support from both Intel and AMD, DDR3 finally began to achieve price parity with DDR2, causing DDR3 to begin to eclipse DDR2 in sales.

DDR3 modules use advanced signal designs, including self-driver calibration and data synchronization, along with an optional onboard thermal sensor. DDR3 memory runs on only 1.5V, which is nearly 20% less than the 1.8V used by DDR2 memory. The lower voltage combined with higher efficiency reduces overall power consumption by up to 30% compared to DDR2.

DDR3 is most suited to systems where the processor and/or memory bus runs at 1,333MHz or higher, which is faster than the 1,066MHz maximum supported by DDR2. For higher-speed memory in standard (non-overclocked) systems, DDR3 modules rated PC3-10600 and PC3-12800 allow for throughputs of 10,667MBps and 12,800MBps, respectively. When combined in dual-channel operation, a pair of PC3-12800 modules result in a total throughput of an incredible 25,600MBps. Processors with tri-channel support, such as the Core i7, have memory bandwidths of 32,000MBps and 38,400MBps using DDR3-1333 and DDR3-1600, respectively. Table 6.10 shows the various official JEDEC-approved DDR3 module types and bandwidth specifications.

Table 6.10 JEDEC Standard DDR3 Module (240-pin DIMM) Speeds and Transfer Rates

Module Standard	Chip Type	Clock Speed (MHz)	Cycles per Clock	Bus Speed (MTps)	Bus Width (Bytes)	Transfer Rate (MBps)	Dual-Channel Transfer Rate (MBps)	Tri-Channel Transfer Rate (MBps)
PC3-6400	DDR3-800	400	2	800	8	6,400	12,800	19,200
PC3-8500	DDR3-1066	533	2	1066	8	8,533	17,066	25,600
PC3-10600	DDR3-1333	667	2	1333	8	10,667	21,333	32,000
PC3-12800	DDR3-1600	800	2	1600	8	12,800	25,600	38,400

MTps = Megatransfers per second

MBps = Megabytes per second

DIMM = Dual inline memory module

DDR = Double data rate

The fastest official JEDEC-approved standard is DDR3-1600, which is composed of chips that run at an effective speed of 1,600MHz (really megatransfers per second), resulting in modules designated PC3-12800 and having a bandwidth of 12,800MBps. However, just as with DDR and DDR2, many manufacturers produce nonstandard modules designed for overclocked systems. These are sold as modules with unofficial designations, clock speeds, and performance figures that exceed the standard ratings.

Table 6.11 shows the popular unofficial DDR3 speed ratings I've seen on the market. Note that because the speeds of these modules are beyond the standard default motherboard and chipset speeds, you won't see any advantage to using them unless you are overclocking your system and your motherboard supports the corresponding overclocked processor and memory settings that these modules require. In addition, because these modules use standard-speed chips that are running overclocked, they almost always require custom voltage settings that are higher than the 1.5V used by standard DDR3 memory. For system stability, I generally don't recommend using overclocked (higher voltage) memory, instead preferring to use only that which runs on the DDR3 standard 1.5V.

Table 6.11 Overclocked (non-JEDEC) DDR3 Module (240-pin DIMM) Speeds and Transfer Rates

Module Standard	Chip Type	Clock Speed (MHz)	Cycles per Clock	Bus Speed (MTps)	Bus Width (Bytes)	Transfer Rate (MBps)	Dual-Channel Transfer Rate (MBps)	Tri-Channel Transfer Rate (MBps)
PC3-11000	DDR3-1375	688	2	1375	8	11,000	22,000	33,000
PC3-13000	DDR3-1625	813	2	1625	8	13,000	26,000	39,000
PC3-14400	DDR3-1800	900	2	1800	8	14,400	28,800	43,200
PC3-14900	DDR3-1866	933	2	1866	8	14,933	29,866	44,800
PC3-15000	DDR3-1866	933	2	1866	8	14,933	29,866	44,800
PC3-16000	DDR3-2000	1000	2	2000	8	16,000	32,000	48,000

MTps = Megatransfers per second

MBps = Megabytes per second

DIMM = Dual inline memory module

DDR = Double data rate

The 240-pin DDR3 modules are similar in pin count, size, and shape to the DDR2 modules; however, the DDR3 modules are incompatible with the DDR2 circuits and are designed with different keying to make them physically noninterchangeable.

RDRAM

Rambus DRAM (RDRAM) was a proprietary (non-JEDEC) memory technology found mainly in certain Intel-based Pentium III and 4 systems from 2000 through 2002. Intel had signed a contract with Rambus in 1996 ensuring it would both adopt and support RDRAM memory into 2001. Believing that any memory it endorsed would automatically become the most popular in the industry, Intel also invested heavily in Rambus at the time. Because RDRAM was a proprietary standard owned by Rambus, using or producing it would require licensing from Rambus, something that was not very popular with other memory and chipset manufacturers. Still, the technology was licensed and Intel originally promised that supporting chipsets and motherboards would be available in 1998.

Unfortunately there were problems in getting the supporting chipsets to market, with delays of many months resulting in memory manufacturers stockpiling RDRAM chips with no systems to support them, while conventional SDRAM and DDR meanwhile came into short supply. The delays resulted in an industrywide debacle that caused Intel to rethink and eventually abandon its investment in the technology. After 2001, Intel continued to support RDRAM in existing systems; however, new chipsets and motherboards rapidly shifted to DDR SDRAM. AMD wisely never invested in the RDRAM technology, and as a result no AMD-based systems were ever designed to use RDRAM.

Without Intel's commitment to future chipset development and support, very few RDRAM-based systems were sold after 2002. Due to the lack of industry support from chipset and motherboard manufacturers, RDRAM was only used in PCs for a short time, and will most likely not play a big part in any future PCs.

With RDRAM, Rambus developed what is essentially a chip-to-chip memory bus, with specialized devices that communicate at very high rates of speed. What might be interesting to some is that this technology was first developed for game systems and first made popular by the Nintendo 64 game system, and it subsequently was used in the Sony Playstation 2.

Conventional memory systems that use SDRAM are known as *wide-channel systems*. They have memory channels as wide as the processor's data or memory bus, which for the Pentium and up is 64 bits, or even wider in dual-channel or tri-channel modes. The dual inline memory module (DIMM) is a 64-bit wide device, meaning data can be transferred to it 64 bits (or 8 bytes) at a time.

RDRAM modules, on the other hand, are narrow-channel devices. They transfer data only 16 bits (2 bytes) at a time (plus 2 optional parity bits), but at faster speeds. This was a shift away from a more parallel to a more serial design for memory and is similar to what has been happening with other evolving buses in the PC.

Each individual chip is serially connected to the next on a package called a *Rambus inline memory module (RIMM)*, which looks similar to a DIMM module but which is not interchangeable. All memory transfers are done between the memory controller and a single device, not between devices. A single Rambus channel typically has three RIMM sockets and can support up to 32 individual RDRAM devices (the RDRAM chips) and more if buffers are used. However, most motherboards implement only two modules per channel (four sockets in a dual-channel design) to avoid problems with signal noise.

The RDRAM memory bus is a continuous path through each device and module on the bus, with each module having input and output pins on opposite ends. Therefore, any RIMM sockets not containing a RIMM must then be filled with a continuity module to ensure that the path is completed. The signals that reach the end of the bus are terminated on the motherboard.

The 16-bit single-channel RIMMs originally ran at 800MHz, so the overall throughput is 800×2, or 1.6GB per second for a single channel—the same as PC1600 DDR SDRAM. Pentium 4 systems typically used two banks simultaneously, creating a dual-channel design capable of 3.2GBps, which matched the bus speed of the original Pentium 4 processors. The RDRAM design features less latency between transfers because they all run synchronously in a looped system and in only one direction.

Newer RIMM versions ran at 1,066MHz in addition to the original 800MHz rate, but very few chipsets or motherboards were released to support the higher speed.

Each RDRAM chip on a RIMM1600 essentially operates as a standalone device sitting on the 16-bit data channel. Internally, each RDRAM chip has a core that operates on a 128-bit wide bus split into eight 16-bit banks running at 100MHz. In other words, every 10ns (100MHz), each RDRAM chip can transfer 16 bytes to and from the core. This internally wide yet externally narrow high-speed interface is the key to RDRAM.

Other improvements to the design include separating control and data signals on the bus. Independent control and address buses are split into two groups of pins for row and column commands, while data is transferred across the 2-byte wide data bus. The actual memory bus clock runs at 400MHz; however, data is transferred on both the falling and rising edges of the clock signal, or twice per clock pulse. The falling edge is called an *even cycle*, and the rising edge is called an *odd cycle*. Complete memory bus synchronization is achieved by sending packets of data beginning on an even cycle interval. The overall wait before a memory transfer can begin (latency) is only one cycle, or 2.5ns maximum.

Figure 6.2 (shown earlier) depicts the relationship between clock and data cycles; you can see the DDR clock and data cycles used by RDRAM and DDR SDRAM. An RDRAM data packet always begins on an even (falling) transition for synchronization purposes. The architecture also supports multiple, simultaneous interleaved transactions in multiple separate time domains. Therefore, before a transfer has even completed, another can begin.

Another important feature of RDRAM is that it is designed for low power consumption. The RIMMs themselves as well as the RDRAM devices run on only 2.5 volts and use low-voltage signal swings from 1.0V to 1.8V, a swing of only 0.8V total. RDRAMs also have four power-down modes and can automatically transition into standby mode at the end of a transaction, which offers further power savings.

A RIMM is similar in size and physical form to a DIMM, but they are not interchangeable. RIMMs are available in module sizes up to 1GB or more and can be added to a system one at a time because each individual RIMM technically represents multiple banks to a system. Note, however, that they have to be added in pairs if your motherboard implements dual-channel RDRAM and you are using 16-bit wide RIMMs.

RIMMs are available in four primary speed grades and usually run in a dual-channel environment, so they have to be installed in pairs, with each one of the pairs in a different set of sockets. Each set of RIMM sockets on such boards is a channel. The 32-bit version incorporates multiple channels within a single device and, as such, is designed to be installed individually, eliminating the requirement for matched pairs. Table 6.12 compares the various types of RDRAM modules. Note that the once-common names for RIMM modules, such as PC800, have been replaced by names that reflect the actual bandwidth of the modules to avoid confusion with DDR memory.

Table 6.12 RDRAM Module Types and Bandwidth

Module Standard	Chip Type	Clock Speed (MHz)	Cycles per Clock	Bus Speed (MTps)	Bus Width (Bytes)	Transfer Rate (MBps)
RIMM1200	PC600	300	2	600	2	1,200
RIMM1400	PC700	350	2	700	2	1,400
RIMM1600	PC800	400	2	800	2	1,600
RIMM2100	PC1066	533	2	1,066	2	2,133

MTps = Megatransfers per second

MBps = Megabytes per second

RIMM = Rambus inline memory module

When Intel initially threw its weight behind the Rambus memory, it seemed destined to be a sure thing for success. Unfortunately, technical delays in the chipsets caused the supporting motherboards to be significantly delayed, and with few systems to support the RIMMs, most memory manufacturers went back to making SDRAM or shifted to DDR SDRAM instead. This caused the remaining available RIMMs being manufactured to be originally priced three or more times that of comparatively sized DIMMs.

Note

Rambus claimed it had patents that covered both standard and DDR SDRAM designs, contending that regardless of whether a company manufactured SDRAM, DDR, or RDRAM, it must pay royalties. Most of the cases that have gone to trial have so far ruled against Rambus, essentially invalidating its patents and claims on DDR and SDRAM. Many appeals are pending, and it will likely be a long time before the patent issues are resolved.

With support for RDRAM memory essentially gone by 2003, RDRAM quickly disappeared from the PC marketplace. Because RDRAM is in such limited supply, if you have existing systems with RDRAM memory, it is generally not cost effective to upgrade them by adding more memory.

Memory Modules

The CPU and motherboard architecture (chipset) dictates a particular computer's physical memory capacity and the types and forms of memory that can be installed. Over the years, three primary changes have occurred in computer memory—it has gradually become faster, wider, and larger in capacity. The CPU and the memory controller circuitry dictate the speed, width, and maximum amount supported. The memory controller in a modern PC resides in either the processor or the motherboard chipset. Even though a system might physically support a given maximum amount of memory, the type of software you run may dictate how much memory can actually be used.

We've already discussed memory types, speeds, and widths. Modern memory modules are 64 bits wide, and depending on the memory controller design, they are accessed in single-, dual-, or tri-channel mode. In single-channel mode, the memory is read and written 64 bits at a time, whereas in dual- or tri-channel mode, the memory bus width increases to 128 bits or 192 bits, respectively. With the exception of the ill-fated RDRAM memory type, memory is one of the few components in the PC to remain massively parallel. Most other parts of the PC have transitioned to serial interface designs.

Maximum physical memory capacity is dictated by several factors. The first is the amount addressable by the processor itself, which is based on the number of physical address lines in the chip. The original PC processors (8086/8088) had 20 address lines, which resulted in those chips being able to recognize up to 1MB (2 to the 20th power bytes) of RAM. The 286/386SX increased memory addressing capability to 24 lines, making them capable of addressing 16MB (2 to the 24th power bytes). Modern x86 processors have from 32 to 36 address lines, resulting in from 4GB to 64GB of addressable RAM. Modern x86-64 (64-bit) processors have 40 address lines, resulting in a maximum of 1TB (1 terabyte) of supported physical RAM.

◀◀ See "Processor Specifications," p. 37 (Chapter 3).

The operating mode of the processor may place further limits on memory addressability. For example, when the processor is operating in backward-compatible real mode, only 1MB of memory is supported.

◀◀ See "Processor Modes," p. 45 (Chapter 3).

Note that even though modern 64-bit processors can address up to 1TB, modern motherboards and/or chipsets generally limit the maximum amount of RAM to 8GB, 16GB, or 24GB. The type of software also has an effect. The 32-bit versions of Windows XP, Vista, and Windows 7 limit memory support to 4GB, whereas the 64-bit versions limit support to 8GB, 16GB, or 192GB, depending on the edition.

Note

See the "Chipsets" section in Chapter 4 for the memory limits on motherboard chipsets.

SIMMs, DIMMs, and RIMMs

Originally, PCs had memory installed via individual chips. They are often referred to as *dual inline package (DIP)* chips because of their physical designs. The original IBM XT and AT systems had 36 sockets on the motherboard for these individual chips—and more sockets could often be found on memory cards plugged into the bus slots. I remember spending hours populating boards with these chips, which was a tedious job.

Besides being a time-consuming and labor-intensive way to deal with memory, DIP chips had one notorious problem—they crept out of their sockets over time as the system went through thermal cycles. Every day, when you powered the system on and off, the system heated and cooled, and the chips gradually walked their way out of the sockets—a phenomenon called *chip creep*. Eventually, good contact was lost and memory errors resulted. Fortunately, reseating all the chips back in their sockets usually rectified the problem, but that method was labor intensive if you had a lot of systems to support.

The alternative to this at the time was to have the memory soldered into either the motherboard or an expansion card. This prevented the chips from creeping and made the connections more permanent, but it caused another problem. If a chip did go bad, you had to attempt desoldering the old one and resoldering a new one or resort to scrapping the motherboard or memory card on which the chip was installed. This was expensive and made memory troubleshooting difficult.

A chip was needed that was both soldered and removable, which was made possible by using memory modules instead of individual chips. Early modules had one row of electrical contacts and were called SIMMs (single inline memory modules), whereas later modules had two rows and were called DIMMs (dual inline memory modules) or RIMMs (Rambus inline memory modules). These small boards plug into special connectors on a motherboard or memory card. The individual memory chips are soldered to the module, so removing and replacing them is impossible. Instead, you must replace the entire module if any part of it fails. The module is treated as though it were one large memory chip.

Several different types of SIMMs, DIMMs, and RIMMs have been commonly used in desktop systems. The various types are often described by their pin count, memory row width, or memory type.

SIMMs, for example, are available in two main physical types—30-pin (8 bits plus an option for 1 additional parity bit) and 72-pin (32 bits plus an option for 4 additional parity bits)—with various capacities and other specifications. The 30-pin SIMMs are physically smaller than the 72-pin versions, and either version can have chips on one or both sides. SIMMs were widely used from the late 1980s to the late 1990s but have become obsolete.

DIMMs are available in four main types. SDR (single data rate) DIMMs have 168 pins, one notch on either side, and two notches along the contact area. DDR (double data rate) DIMMs, on the other hand, have 184 pins, two notches on each side, and only one offset notch along the contact area. DDR2 and DDR3 DIMMs have 240 pins, two notches on each side, and one near the center of the contact area. All DIMMs are either 64 bits (non-ECC/parity) or 72 bits (data plus parity or error-correcting code [ECC]) wide. The main physical difference between SIMMs and DIMMs is that DIMMs have different signal pins on each side of the module, resulting in two rows of electrical contacts. That is why they are called dual inline memory modules, and why with only 1" of additional length, they have many more pins than a SIMM.

Note

There is confusion among users and even in the industry regarding the terms *single-sided* and *double-sided* with respect to memory modules. In truth, the single- or double-sided designation actually has nothing to do with whether chips are physically located on one or both sides of the module, and it has nothing to do with whether the module is a SIMM or DIMM (meaning whether the connection pins are single- or double-inline). Instead the terms single-sided and double-sided are used to indicate whether the module has one or two internal banks (called ranks) of memory chips installed. A dual-rank DIMM module has two complete 64-bit wide banks of chips logically stacked so that the module is twice as deep (has twice as many 64-bit rows). In most (but not all) cases, this requires chips to be on both sides of the module; therefore, the term *double-sided* has often been used to indicate that a module has two ranks, even though the term is technically incorrect. Single-rank modules (incorrectly referred to as single-sided) can also have chips physically mounted on both sides of the module, and dual-rank modules can have chips physically mounted on only one side. I recommend using the terms *single rank* or *dual rank* instead because they are much more accurate and easily understood.

RIMMs also have different signal pins on each side. Three different physical types of RIMMs are available: a 16/18-bit version with 184 pins, a 32/36-bit version with 232 pins, and a 64/72-bit version with 326 pins. Each of these plugs into the same sized connector, but the notches in the connectors and RIMMs are different to prevent a mismatch. A given board will accept only one type. By far the most common type is the 16/18-bit version. The 32-bit version was introduced in late 2002, and the 64-bit version was introduced in 2004.

The standard 16/18-bit RIMM has 184 pins, one notch on either side, and two notches centrally located in the contact area. The 16-bit versions are used for non-ECC applications, whereas the 18-bit versions incorporate the additional bits necessary for ECC.

Figures 6.3 through 6.9 show a typical 30-pin (8-bit) SIMM, 72-pin (32-bit) SIMM, 168-pin SDRAM DIMM, 184-pin DDR SDRAM (64-bit) DIMM, 240-pin DDR2 DIMM, 240-pin DDR3 DIMM, and 184-pin RIMM, respectively. The pins are numbered from left to right and are connected through to both sides of the module on the SIMMs. The pins on the DIMM are different on each side, but on a SIMM, each side is the same as the other and the connections carry through. Note that all dimensions are in both inches and millimeters (in parentheses), and modules are generally available in error-correcting code (ECC) versions with 1 extra ECC (or parity) bit for every 8 data bits (multiples of 9 in data width) or versions that do not include ECC support (multiples of 8 in data width).

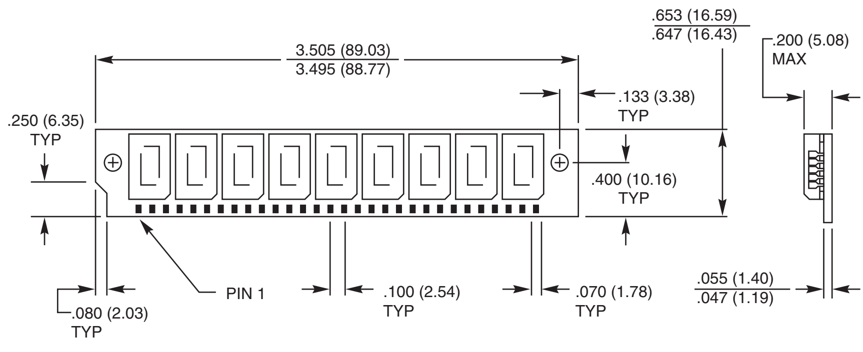


Figure 6.3 A typical 30-pin SIMM.

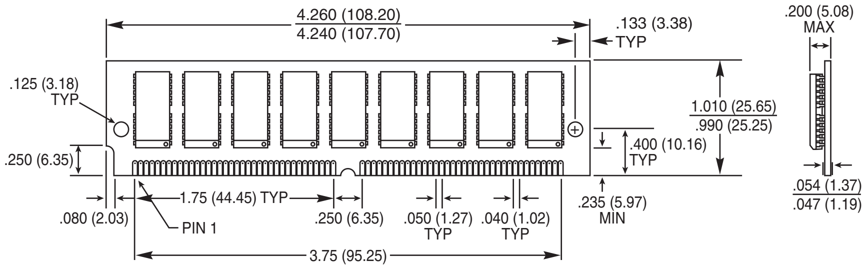


Figure 6.4 A typical 72-pin SIMM.

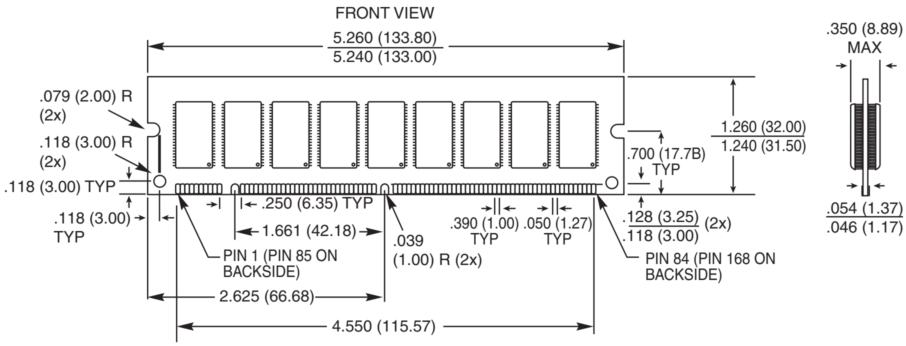


Figure 6.5 A typical 168-pin SDRAM DIMM.

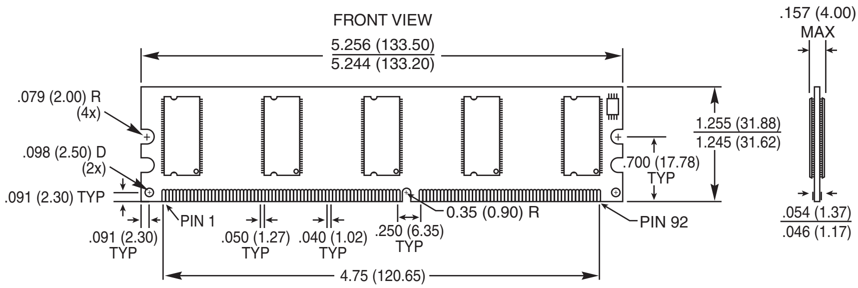


Figure 6.6 A typical 184-pin DDR DIMM.

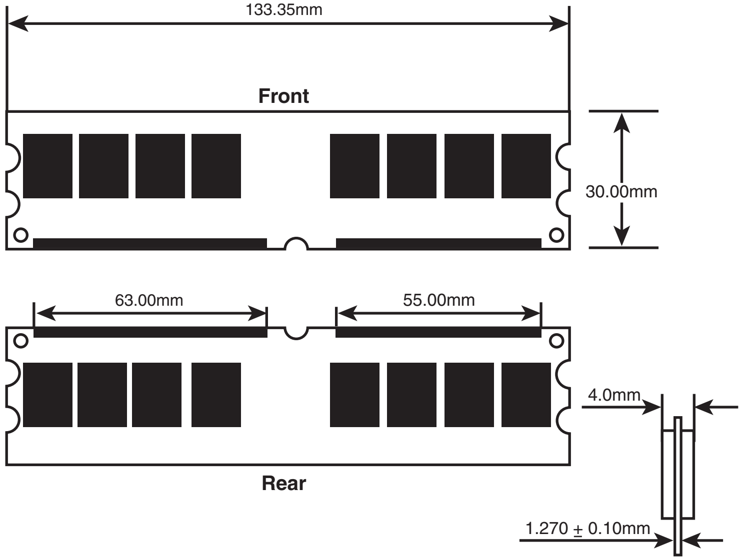


Figure 6.7 A typical 240-pin DDR2 DIMM.

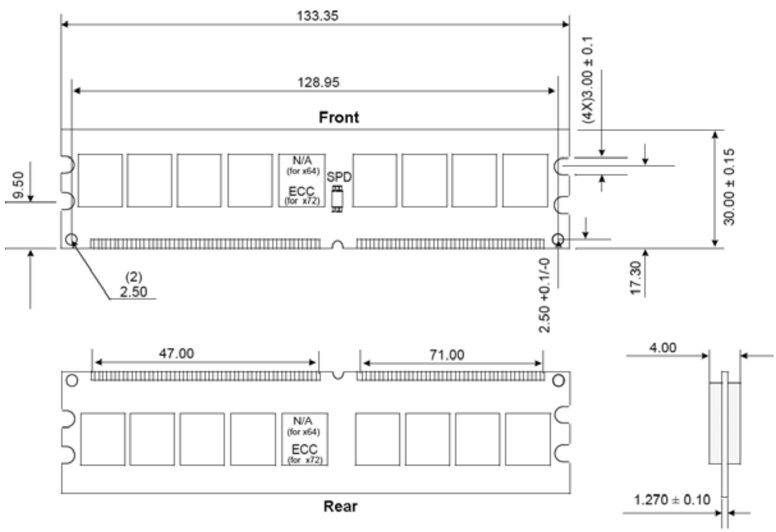


Figure 6.8 A typical 240-pin DDR3 DIMM.

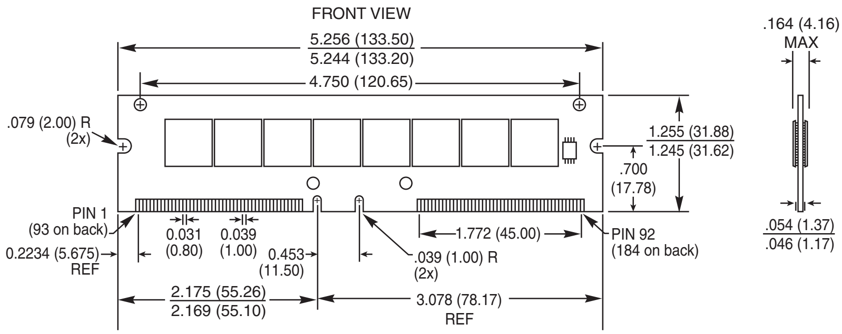


Figure 6.9 A typical 184-pin RIMM.

All these memory modules are fairly compact considering the amount of memory they hold and are available in several capacities and speeds. Table 6.13 lists the various capacities available for SIMMs, DIMMs, and RIMMs.

Table 6.13 SIMM, DIMM, and RIMM Capacities

Capacity	Standard Depth×Width	Parity/ECC Depth×Width
30-Pin SIMM		
256KB	256K×8	256K×9
1MB	1M×8	1M×9
4MB	4M×8	4M×9
16MB	16M×8	16M×9
72-Pin SIMM		
1MB	256K×32	256K×36
2MB	512K×32	512K×36
4MB	1M×32	1M×36
8MB	2M×32	2M×36
16MB	4M×32	4M×36
32MB	8M×32	8M×36
64MB	16M×32	16M×36
128MB	32M×32	32M×36
168/184-Pin DIMM/DDR DIMM		
8MB	1M×64	1M×72
16MB	2M×64	2M×72
32MB	4M×64	4M×72
64MB	8M×64	8M×72
128MB	16M×64	16M×72
256MB	32M×64	32M×72
512MB	64M×64	64M×72
1,024MB	128M×64	128M×72
2,048MB	256M×64	256M×72

Table 6.13 Continued

Capacity	Standard Depth×Width	Parity/ECC Depth×Width
240-Pin DDR2/DDR3 DIMM		
256MB	32M×64	32M×72
512MB	64M×64	64M×72
1,024MB	128M×64	128M×72
2,048MB	256M×64	256M×72
4,096MB	512M×64	512M×72
184-Pin RIMM		
64MB	32M×16	32M×18
128MB	64M×16	64M×18
256MB	128M×16	128M×18
512MB	256M×16	256M×18
1,024MB	512M×16	512M×18

Memory modules of each type and capacity are available in various speed ratings. Consult your motherboard documentation for the correct memory speed and type for your system. If a system requires a specific speed memory module, you can almost always substitute faster speeds if the one specified is not available. Generally, no problems occur in mixing module speeds, as long as you use modules equal to or faster than what the system requires. Because there's little price difference between the various speed versions, I often buy faster modules than are necessary for a particular application, especially if they are the same cost as slower modules. This might make them more usable in a future system that could require the faster speed.

Because SDRAM and newer modules have an onboard serial presence detect (SPD) ROM that reports their speed and timing parameters to the system, most systems run the memory controller and memory bus at the speed matching the slowest module installed.

Note

A *bank* is the smallest amount of memory needed to form a single row of memory addressable by the processor. It is the minimum amount of physical memory that is read or written by the processor at one time and usually corresponds to the data bus width of the processor. If a processor has a 64-bit data bus, a bank of memory also is 64 bits wide. If the memory runs dual- or tri-channel, a virtual bank is formed that is two or three times the absolute data bus width of the processor.

You can't always replace a module with a higher-capacity unit and expect it to work. Systems might have specific design limitations for the maximum capacity of module they can take. A larger-capacity module works only if the motherboard is designed to accept it in the first place. Consult your system documentation to determine the correct capacity and speed to use.

Registered Modules

SDRAM through DDR3 modules are available in unbuffered and registered versions. Most PC motherboards are designed to use unbuffered modules, which allow the memory controller signals to pass directly to the memory chips on the module with no interference. This is not only the cheapest design, but also the fastest and most efficient. The only drawback is that the motherboard designer must place limits on how many modules (meaning module sockets) can be installed on the board, and possibly also limit how many chips can be on a module. So-called double-sided modules that really have multiple banks of chips onboard might be restricted on some systems in certain combinations.

Systems designed to accept extremely large amounts of RAM (such as servers) often require registered modules. A registered module uses an architecture that has register chips on the module that act as an interface between the actual RAM chips and the chipset. The registers temporarily hold data passing to and from the memory chips and enable many more RAM chips to be driven or otherwise placed on the module than the chipset could normally support. This allows for motherboard designs that can support many modules and enables each module to have a larger number of chips. In general, registered modules are required by server or workstation motherboards designed to support more than four sockets. One anomaly is the initial version of the AMD Athlon 64 FX processor, which also uses registered memory because its Socket 940 design was based on the AMD Opteron workstation and server processor. Subsequent Socket 939, AM2, and Socket F versions of the Athlon FX no longer require registered memory.

To provide the space needed for the buffer chips, a registered DIMM is often taller than a standard DIMM. Figure 6.10 compares a typical registered DIMM to a typical unbuffered DIMM.

Tip

If you are installing registered DIMMs in a slimline case, clearance between the top of the DIMM and the case might be a problem. Some vendors sell low-profile registered DIMMs that are about the same height as an unbuffered DIMM. Use this type of DIMM if your system does not have enough head room for standard registered DIMMs. Some vendors sell only this type of DIMM for particular systems.

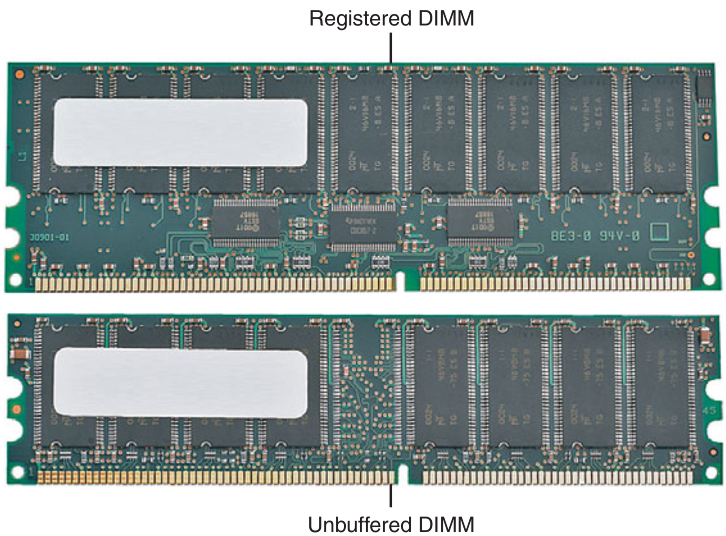


Figure 6.10 A typical registered DIMM is taller than a typical unbuffered DIMM to provide room for buffer chips.

The important thing to note is that you can use only the type of module your motherboard (or chipset) is designed to support. For most, that is standard unbuffered modules or, in some cases, registered modules.

SIMM Details

The 72-pin SIMMs use a set of four or five pins to indicate the type of SIMM to the motherboard. These presence detect pins are either grounded or not connected to indicate the type of SIMM to the motherboard. Presence detect outputs must be tied to the ground through a 0-ohm resistor or jumper on the SIMM—to generate a high logic level when the pin is open or a low logic level when the motherboard grounds the pin. This produces signals the memory interface logic can decode. If the motherboard uses presence detect signals, a power-on self test (POST) procedure can determine the size and speed of the installed SIMMs and adjust control and addressing signals automatically. This enables autodetection of the memory size and speed.

Note

In many ways, the presence detect pin function is similar to the industry-standard DX coding used on modern 35mm film rolls to indicate the ASA (speed) rating of the film to the camera. When you drop the film into the camera, electrical contacts can read the film's speed rating via an industry-standard configuration.

Presence detect performs the same function for 72-pin SIMMs that the serial presence detect (SPD) chip does for DIMMs.

Table 6.14 shows the Joint Electronic Devices Engineering Council (JEDEC) industry-standard presence detect configuration listing for the 72-pin SIMM family. JEDEC is an organization of U.S. semiconductor manufacturers and users that sets semiconductor standards.

Table 6.14 Presence Detect Pin Configurations for 72-Pin SIMMs

Size	Speed	Pin 67	Pin 68	Pin 69	Pin 70	Pin 11
1MB	100ns	Gnd	—	Gnd	Gnd	—
1MB	80ns	Gnd	—	—	Gnd	—
1MB	70ns	Gnd	—	Gnd	—	—
1MB	60ns	Gnd	—	—	—	—
2MB	100ns	—	Gnd	Gnd	Gnd	—
2MB	80ns	—	Gnd	—	Gnd	—
2MB	70ns	—	Gnd	Gnd	—	—
2MB	60ns	—	Gnd	—	—	—
4MB	100ns	Gnd	Gnd	Gnd	Gnd	—
4MB	80ns	Gnd	Gnd	—	Gnd	—
4MB	70ns	Gnd	Gnd	Gnd	—	—
4MB	60ns	Gnd	Gnd	—	—	—
8MB	100ns	—	—	Gnd	Gnd	—
8MB	80ns	—	—	—	Gnd	—
8MB	70ns	—	—	Gnd	—	—
8MB	60ns	—	—	—	—	—

Table 6.14 Continued

Size	Speed	Pin 67	Pin 68	Pin 69	Pin 70	Pin 11
16MB	80ns	Gnd	—	—	Gnd	Gnd
16MB	70ns	Gnd	—	Gnd	—	Gnd
16MB	60ns	Gnd	—	—	—	Gnd
16MB	50ns	Gnd	—	Gnd	Gnd	Gnd
32MB	80ns	—	Gnd	—	Gnd	Gnd
32MB	70ns	—	Gnd	Gnd	—	Gnd
32MB	60ns	—	Gnd	—	—	Gnd
32MB	50ns	—	Gnd	Gnd	Gnd	Gnd

— = No connection (open)
Gnd = Ground
Pin 67 = Presence detect 1
Pin 68 = Presence detect 2
Pin 69 = Presence detect 3
Pin 70 = Presence detect 4
Pin 11 = Presence detect 5

Unfortunately, unlike the film industry, not everybody in the computer industry follows established standards. As such, presence detect signaling is not a standard throughout the PC industry. Different system manufacturers sometimes use different configurations for what is expected on these four pins. Compaq, IBM (mainly PS/2 systems), and Hewlett-Packard are notorious for this type of behavior. Many of the systems from these vendors require special SIMMs that are basically the same as standard 72-pin SIMMs, except for special presence detect requirements. Table 6.15 shows how IBM defines these pins.

Table 6.15 Presence Detect Pins for IBM 72-Pin SIMMs

67	68	69	70	SIMM Type	IBM Part Number
—	—	—	—	Not a valid SIMM	n/a
Gnd	—	—	—	1MB 120ns	n/a
—	Gnd	—	—	2MB 120ns	n/a
Gnd	Gnd	—	—	2MB 70ns	92F0102
—	—	Gnd	—	8MB 70ns	64F3606
Gnd	—	Gnd	—	Reserved	n/a
—	Gnd	Gnd	—	2MB 80ns	92F0103
Gnd	Gnd	Gnd	—	8MB 80ns	64F3607
—	—	—	Gnd	Reserved	n/a
Gnd	—	—	Gnd	1MB 85ns	90X8624
—	Gnd	—	Gnd	2MB 85ns	92F0104
Gnd	Gnd	—	Gnd	4MB 70ns	92F0105
—	—	Gnd	Gnd	4MB 85ns	79F1003 (square notch) L40-SX
Gnd	—	Gnd	Gnd	1MB 100ns	n/a

Table 6.15 Continued

67	68	69	70	SIMM Type	IBM Part Number
Gnd	—	Gnd	Gnd	8MB 80ns	79F1004 (square notch) L40-SX
—	Gnd	Gnd	Gnd	2MB 100ns	n/a
Gnd	Gnd	Gnd	Gnd	4MB 80ns	87F9980
Gnd	Gnd	Gnd	Gnd	2MB 85ns	79F1003 (square notch) L40SX

— = No connection (open)

Gnd = Ground

Pin 67 = Presence detect 1

Pin 68 = Presence detect 2

Pin 69 = Presence detect 3

Pin 70 = Presence detect 4

Because these pins can have custom variations, you often must specify IBM, Compaq, HP, or generic SIMMs when you order memory for systems using 72-pin SIMMs. Although very few (if any) of these systems are still in service, keep this information in mind if you are moving 72-pin modules from one system to another or are installing salvaged memory into a system. Also, be sure you match the metal used on the module connectors and sockets. SIMM pins can be tin or gold plated, and the plating on the module pins must match that on the socket pins; otherwise, corrosion will result.

Caution

To have the most reliable system when using SIMM modules, you must install modules with gold-plated contacts into gold-plated sockets and modules with tin-plated contacts into tin-plated sockets only. If you mix gold contacts with tin sockets, or vice versa, you are likely to experience memory failures from 6 months to 1 year after initial installation because a type of corrosion know as *fretting* will take place. This has been a major problem with 72-pin SIMM-based systems because some memory and motherboard vendors opted for tin sockets and connectors while others opted for gold. According to connector manufacturer AMP's "Golden Rules: Guidelines for the Use of Gold on Connector Contacts" (available at www.tycoelectronics.com/documentation/whitepapers/pdf/aurulrep.pdf) and "The Tin Commandments: Guidelines for the Use of Tin on Connector Contacts" (available at www.tycoelectronics.com/documentation/whitepapers/pdf/sncomrep.pdf), you should match connector metals.

If you are maintaining systems with mixed tin/gold contacts in which fretting has already occurred, use a wet contact cleaner. After cleaning, to improve electrical contacts and help prevent corrosion, you should use a liquid contact enhancer and lubricant called Stabilant 22 from D.W. Electrochemicals when installing SIMMs or DIMMs. The company's website (www.stabilant.com) has detailed application notes on this subject that provide more technical details.

SDR DIMM Details

SDR (single data rate) DIMMs use a completely different type of presence detect than SIMMs, called *serial presence detect (SPD)*. It consists of a small EEPROM or flash memory chip on the DIMM that contains specially formatted data indicating the DIMM's features. This serial data can be read via the serial data pins on the DIMM, and it enables the motherboard to autoconfigure to the exact type of DIMM installed.

DIMMs can come in several varieties, including unbuffered and buffered as well as 3.3V and 5V. Buffered DIMMs have additional buffer chips on them to interface to the motherboard.

Unfortunately, these buffer chips slow down the DIMM and are not effective at higher speeds. For this reason, most PC systems (those that do not use registered DIMMs) use unbuffered DIMMs. The voltage is simple—DIMM designs for PCs are almost universally 3.3V. If you install a 5V DIMM in a 3.3V socket, it would be damaged, but fortunately keying in the socket and on the DIMM prevents that.

Modern PC systems use only unbuffered 3.3V DIMMs. Apple and other non-PC systems can use the buffered 5V versions. Fortunately, the key notches along the connector edge of a DIMM are spaced differently for buffered/unbuffered and 3.3V/5V DIMMs, as shown in Figure 6.11. This prevents inserting a DIMM of the wrong type into a given socket.

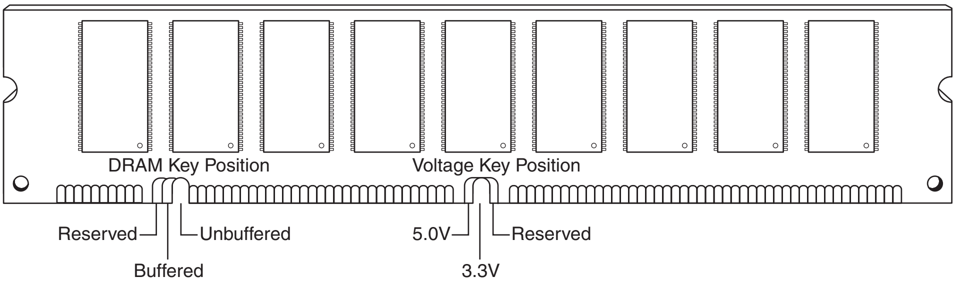


Figure 6.11 The 168-pin DRAM DIMM notch key definitions.

DDR DIMM Details

The 184-pin DDR DIMMs use a single key notch to indicate voltage, as shown in Figure 6.12.

DDR DIMMs also use two notches on each side to enable compatibility with both low- and high-profile latched sockets. Note that the key position is offset with respect to the center of the DIMM to prevent inserting it backward in the socket. The key notch is positioned to the left, centered, or to the right of the area between pins 52 and 53. This is used to indicate the I/O voltage for the DDR DIMM and to prevent installing the wrong type into a socket that might damage the DIMM.

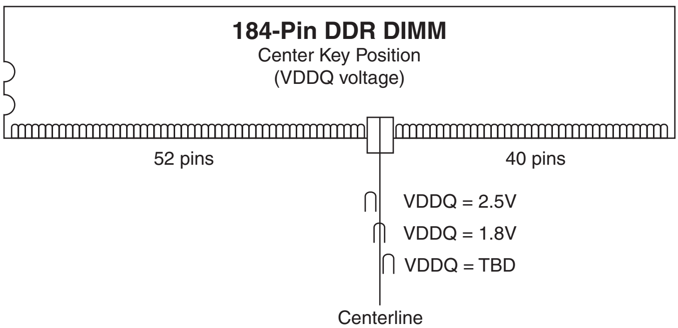


Figure 6.12 The 184-pin DDR SDRAM DIMM keying.

DDR2 DIMM Details

The 240-pin DDR2 DIMMs use two notches on each side to enable compatibility with both low- and high-profile latched sockets. The connector key is offset with respect to the center of the DIMM to prevent inserting it backward in the socket. The key notch is positioned in the center of the area between pins 64 and 65 on the front (184/185 on the back), and there is no voltage keying because all DDR2 DIMMs run on 1.8V.

DDR3 DIMM Details

The 240-pin DDR3 DIMMs use two notches on each side to enable compatibility with both low- and high-profile latched sockets. The connector key is offset with respect to the center of the DIMM to prevent inserting it backward in the socket. The key notch is positioned in the center of the area between pins 48 and 49 on the front (168/169 on the back), and there is no voltage keying because all DDR3 DIMMs run on 1.5V.

RIMM Details

The 16/18-bit RIMMs are keyed with two notches in the center. This prevents a backward insertion and prevents the wrong type (voltage) RIMM from being used in a system. Currently, all RIMMs run on 2.5V, but proposed 64-bit versions will run on only 1.8V. To allow for changes in the RIMMs, three keying options are possible in the design (see Figure 6.13). The left key (indicated as “DATUM A” in Figure 6.13) is fixed in position, but the center key can be in three different positions spaced 1mm or 2mm to the right, indicating different types of RIMMs. The current default is option A, as shown in Figure 6.13 and Table 6.16, which corresponds to 2.5V operation.

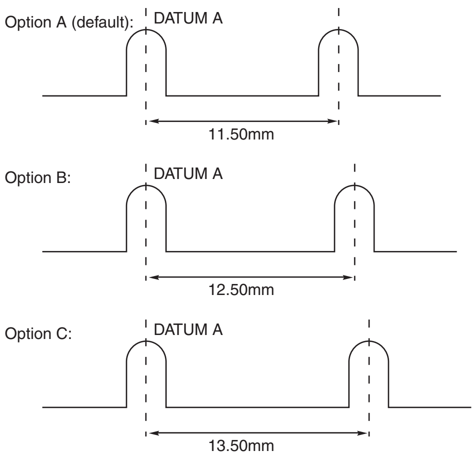


Figure 6.13 RIMM keying options.

Table 6.16 Possible Keying Options for RIMMs

Option	Notch Separation	Description
A	11.5mm	2.5V RIMM
B	12.5mm	Reserved
C	13.5mm	Reserved

RIMMs incorporate an SPD device, which is essentially a flash ROM onboard. This ROM contains information about the RIMM’s size and type, including detailed timing information for the memory controller. The memory controller automatically reads the data from the SPD ROM to configure the system to match the RIMMs installed.

Figure 6.14 shows a typical PC RIMM installation. The RDRAM controller and clock generator are typically in the motherboard chipset North Bridge component. As you can see, the Rambus memory channel flows from the memory controller through each of up to three RIMM modules in series. Each module contains 4, 8, 16, or more RDRAM devices (chips), also wired in series, with an onboard SPD ROM for system configuration. Any RIMM sockets without a RIMM installed must have a continuity module, shown in the last socket in Figure 6.13. This enables the memory bus to remain continuous from the controller through each module (and, therefore, each RDRAM device on the module) until the bus finally terminates on the motherboard. Note how the bus loops from one module to another. For timing purposes, the first RIMM socket must be 6" or less from the memory controller, and the entire length of the bus must not be more than it would take for a signal to go from one end to another in four data clocks, or about 5ns.

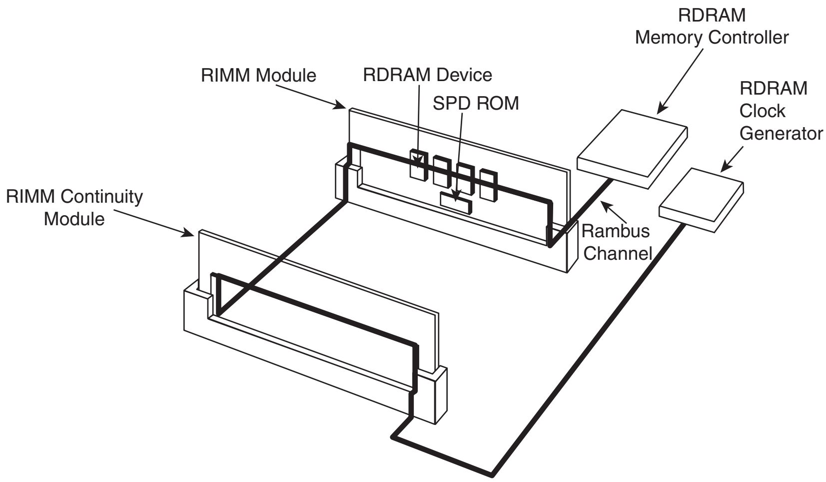


Figure 6.14 Typical RDRAM bus layout showing a RIMM and one continuity module.

Interestingly, Rambus does not manufacture the RDRAM devices (the chips) or the RIMMs; that is left to other companies. Rambus is merely a design company, and it has no chip fabs or manufacturing facilities of its own. It licenses its technology to other companies who then manufacture the devices and modules.

Determining a Memory Module's Size and Features

Most memory modules are labeled with a sticker indicating the module's type, speed rating, and manufacturer. If you are attempting to determine whether existing memory can be used in a new computer, or if you need to replace memory in an existing computer, this information can be essential. Figure 6.15 illustrates the markings on typical 512MB and 1GB DDR memory modules from Crucial Technologies.

1. Module size
2. Module type and speed
3. CAS Latency
4. Crucial Technology part number

Figure 6.15 Markings on 512MB (top) and 1GB (bottom) DDR memory modules from Crucial Technology.

However, if you have memory modules that are not labeled, you can still determine the module type, speed, and capacity if the memory chips on the module are clearly labeled. For example, assume you have a memory module with chips labeled as follows:

MT46V64M8TG-75

By using an Internet search engine such as Google and entering the number from one of the memory chips, you can usually find the data sheet for the memory chips. Consider the following example: Say you have a registered memory module and want to look up the part number for the memory chips (usually eight or more chips) rather than the buffer chips on the module (usually from one to three, depending on the module design). In this example, the part number turns out to be a Micron memory chip that decodes like this:

MT = Micron Technologies (the memory chip maker)

46 = DDR SDRAM

V = 2.5V DC

64M8 = 8 million rows × 8 (equals 64) × 8 banks (often written as 64 Meg × 8)

TG = 66-pin TSOP chip package

–75 = 7.5ns @ CL2 latency (DDR 266)

The full datasheet for this example is located at http://download.micron.com/pdf/datasheets/dram/ddr/512MBDDRx4x8x16.pdf.

From this information, you can determine that the module has the following characteristics:

- The module runs at DDR266 speeds using standard 2.5V DC voltage.
- The module has a latency of CL2, so it can be used on any system that requires CL2 or slower latency (such as CL2.5 or CL3).
- Each chip has a capacity of 512Mb (64 × 8 = 512).
- Each chip contains 8 bits. Because it takes 8 bits to make 1 byte, the capacity of the module can be calculated by grouping the memory chips on the module into groups of eight. If each chip contains 512Mb, a group of eight means that the module has a size of 512MB (512Mb × 8 = 512MB). A dual-bank module has two groups of eight chips for a capacity of 1GB (512Mb × 8 = 1024MB, or 1GB).

If the module has nine instead of eight memory chips (or 18 instead of 16), the additional chips are used for parity checking and support ECC error correction on servers with this feature.

To determine the size of the module in MB or GB and to determine whether the module supports ECC, count the memory chips on the module and compare them to Table 6.17. Note that the size of each memory chip in Mb is the same as the size in MB if the memory chips use an 8-bit design.

Table 6.17 Module Capacity Using 512Mb (64Mbit × 8) Chips

Number of Chips	Number of Bits in Each Bank	Module Size	Supports ECC?	Single or Dual Bank
8	64	512MB	No	Single
9	72	512MB	Yes	Single
16	64	1GB	No	Dual
18	72	1GB	Yes	Dual

The additional chip used by each group of eight chips provides parity checking, which is used by the ECC function on most server motherboards to correct single-bit errors.

A registered module contains 9 or 18 memory chips for ECC plus additional memory buffer chips. These chips are usually smaller in size and located near the center of the module, as shown previously in Figure 6.10.

Note

Some modules use 16-bit wide memory chips. In such cases, only four chips are needed for single-bank memory (five with parity/ECC support) and eight are needed for double-bank memory (10 with parity/ECC support). These memory chips use a design listed as capacity × 16, like this: 256Mb × 16.

You can also see this information if you look up the manufacturer and the memory type in a search engine. For example, a web search for *Micron Unbuffered DIMM Design* locates a table showing various DIMM organization, SDRAM density, and other information for listed modules.

As you can see, with a little detective work, you can determine the size, speed, and type of a memory module—even if the module isn't marked, as long as the markings on the memory chips themselves are legible.

Tip

If you are unable to decipher a chip part number, you can use a program such as HWiNFO or SiSoftware Sandra to identify your memory module, as well as many other facts about your computer, including chipset, processor, empty memory sockets, and much more. You can download shareware versions of HWiNFO from www.hwinfo.com and SiSoftware Sandra from www.sisoftware.net.

Memory Banks

Memory chips (DIPs, SIMMs, SIPPs, and DIMMs) are organized in banks on motherboards and memory cards. You should know the memory bank layout and position on the motherboard and memory cards.

You need to know the bank layout when adding memory to the system. In addition, memory diagnostics report error locations by byte and bit addresses, and you must use these numbers to locate which bank in your system contains the problem.

The banks usually correspond to the data bus capacity of the system's microprocessor. Table 6.18 shows the widths of individual banks based on the type of PC.

Table 6.18 Memory Bank Widths on Various Systems

Processor	Data Bus	Memory Bank Width	Memory Bank Width (Parity/ECC)	30-pin SIMMs per Bank	72-pin SIMMs per Bank	DIMMs per Bank
8088	8-bit	8 bits	9 bits	1	—	—
8086	16-bit	16 bits	18 bits	2	—	—
286	16-bit	16 bits	18 bits	2	—	—
386SX, SL, SLC	16-bit	16 bits	18 bits	2	—	—
486SLC, SLC2	16-bit	16 bits	18 bits	2	—	—
386DX	32-bit	32 bits	36 bits	4	1	—
486SX, DX, DX2, DX4, 5x86	32-bit	32 bits	36 bits	4	1	—
x86 and x86-64 running single-channel mode	64-bit	64 bits	72 bits	—	—	1
x86 and x86-64 running dual-channel mode	64-bit	128 bits	144 bits	—	—	2
x86 and x86-64 running tri-channel mode	64-bit	192 bits	216 bits	—	—	3

DIMMs are ideal for Pentium (and higher) systems because the 64-bit width of the DIMM exactly matches the 64-bit width of the Pentium processor data bus. Therefore, each DIMM represents an individual bank, and these can be added or removed one at a time. Many recent systems have been designed to use matched pairs or triples of memory modules for faster performance. So-called "dual-channel" and "tri-channel" designs treat two or three matched modules as a single bank of memory.

The physical orientation and numbering of the SIMMs or DIMMs used on a motherboard are arbitrary and determined by the board's designers, so documentation covering your system or card comes in handy. You can determine the layout of a motherboard or an adapter card through testing, but that takes time and might be difficult, particularly after you have a problem with a system.

Caution

If your system supports dual- or tri-channel memory, be sure you use the correct memory sockets to enable multichannel operation. Check the documentation to ensure that you use the correct sockets. Most multichannel systems will still run in single-channel mode if the memory is not installed in a way that permits full multichannel operation, but performance is lower than if the memory were installed properly. Some systems provide dual-channel support if an odd number of modules are installed, as long as the total capacity of two modules installed in one channel equals the size of the single module in the other channel and all modules are the same speed and latency. Again, check your documentation for details.

Memory Module Speed

When you replace a failed memory module or install a new module as an upgrade, you typically must install a module of the same type and speed as the others in the system. You can substitute a module with a different (faster) speed but only if the replacement module's speed is equal to or faster than that of the other modules in the system.

Some people have had problems when "mixing" modules of different speeds. With the wide variety of motherboards, chipsets, and memory types, few ironclad rules exist. When in doubt as to which speed module to install in your system, consult the motherboard documentation for more information.

Substituting faster memory of the same type doesn't result in improved performance if the system still operates the memory at the same speed. Systems that use DIMMs or RIMMs can read the speed and timing features of the module from a special SPD ROM installed on the module and then set chipset (memory controller) timing accordingly. In these systems, you might see an increase in performance by installing faster modules, to the limit of what the chipset will support.

To place more emphasis on timing and reliability, there are Intel and JEDEC standards governing memory types that require certain levels of performance. These standards certify that memory modules perform within Intel's timing and performance guidelines.

The same common symptoms result when the system memory has failed or is simply not fast enough for the system's timing. The usual symptoms are frequent parity check errors or a system that does not operate at all. The POST might report errors, too. If you're unsure of which chips to buy for your system, contact the system manufacturer or a reputable chip supplier.

▶▶ See "Parity Checking," p. 415 (this chapter).

Parity and ECC

Part of the nature of memory is that it inevitably fails. These failures are usually classified as two basic types: hard fails and soft errors.

The best understood are hard fails, in which the chip is working and then, because of some flaw, physical damage, or other event, becomes damaged and experiences a permanent failure. Fixing this type of failure normally requires replacing some part of the memory hardware, such as the chip, SIMM, or DIMM. Hard error rates are known as HERs.

The other, more insidious type of failure is the soft error, which is a nonpermanent failure that might never recur or could occur only at infrequent intervals. Soft error rates are known as SERs.

More than 20 years ago, Intel made a discovery about soft errors that shook the memory industry. It found that alpha particles were causing an unacceptably high rate of soft errors or single event upsets (SEUs, as they are sometimes called) in the 16KB DRAMs that were available at the time. Because alpha particles are low-energy particles that can be stopped by something as thin and light as a sheet of paper, it became clear that for alpha particles to cause a DRAM soft error, they would have to be coming from within the semiconductor material. Testing showed trace elements of thorium and uranium in the plastic and ceramic chip packaging materials used at the time. This discovery forced all the memory manufacturers to evaluate their manufacturing processes to produce materials free from contamination.

Today, memory manufacturers have all but totally eliminated the alpha-particle source of soft errors and more recent discoveries prove that alpha particles are now only a small fraction of the cause of DRAM soft errors.

As it turns out, the biggest cause of soft errors today is cosmic rays. IBM researchers began investigating the potential of terrestrial cosmic rays in causing soft errors similar to alpha particles. The difference is that cosmic rays are very high-energy particles and can't be stopped by sheets of paper or other more powerful types of shielding. The leader in this line of investigation was Dr. J.F. Ziegler of the IBM Watson Research Center in Yorktown Heights, New York. He has produced landmark research into understanding cosmic rays and their influence on soft errors in memory. One interesting set of experiments found that cosmic ray–induced soft errors were eliminated when the DRAMs were moved to an underground vault shielded by more than 50 feet of rock.

Cosmic ray–induced errors are even more of a problem in SRAMs than DRAMS because the amount of charge required to flip a bit in an SRAM cell is less than is required to flip a DRAM cell capacitor. Cosmic rays are also more of a problem for higher-density memory. As chip density increases, it becomes easier for a stray particle to flip a bit. It has been predicted by some that the soft error rate of a 64MB DRAM is double that of a 16MB chip, and a 256MB DRAM has a rate four times higher. As memory sizes continue to increase, it's likely that soft error rates will also increase.

Unfortunately, the PC industry has largely failed to recognize this cause of memory errors. Electrostatic discharge, power surges, or unstable software can much more easily explain away the random and intermittent nature of a soft error, especially right after a new release of an operating system or major application.

Although cosmic rays and other radiation events are perhaps the biggest cause of soft errors, soft errors can also be caused by the following:

- **Power glitches or noise on the line**—This can be caused by a defective power supply in the system or by defective power at the outlet.
- **Incorrect type or speed rating**—The memory must be the correct type for the chipset and match the system access speed.
- **RF (radio frequency) interference**—Caused by radio transmitters in close proximity to the system, which can generate electrical signals in system wiring and circuits. Keep in mind that the increased use of wireless networks, keyboards, and mouse devices can lead to a greater risk of RF interference.

- **Static discharges**—These discharges cause momentary power spikes, which alter data.
- **Timing glitches**—Data doesn't arrive at the proper place at the proper time, causing errors. Often caused by improper settings in the BIOS Setup, by memory that is rated slower than the system requires, or by overclocked processors and other system components.
- **Heat buildup**—High-speed memory modules run hotter than older modules. RDRAM RIMM modules were the first memory to include integrated heat spreaders, and many high-performance DDR and DDR2 memory modules now include heat spreaders to help fight heat buildup.

Most of these problems don't cause chips to permanently fail (although bad power or static can damage chips permanently), but they can cause momentary problems with data.

How can you deal with these errors? The best way to deal with this problem is to increase the system's fault tolerance. This means implementing ways of detecting and possibly correcting errors in PC systems. Three basic levels and techniques are used for fault tolerance in modern PCs:

- Nonparity
- Parity
- ECC

Nonparity systems have no fault tolerance at all. The only reason they are used is because they have the lowest inherent cost. No additional memory is necessary, as is the case with parity or ECC techniques. Because a parity-type data byte has 9 bits versus 8 for nonparity, memory cost is approximately 12.5% higher. Also, the nonparity memory controller is simplified because it does not need the logic gates to calculate parity or ECC check bits. Portable systems that place a premium on minimizing power might benefit from the reduction in memory power resulting from fewer DRAM chips. Finally, the memory system data bus is narrower, which reduces the amount of data buffers. The statistical probability of memory failures in a modern office desktop computer is now estimated at about one error every few months. Errors will be more or less frequent depending on how much memory you have.

This error rate might be tolerable for low-end systems that are not used for mission-critical applications. In this case, the extreme market sensitivity to price probably can't justify the extra cost of parity or ECC memory, and such errors then must be tolerated.

Parity Checking

One standard IBM set for the industry is that the memory chips in a bank of nine each handle 1 bit of data: 8 bits per character plus 1 extra bit called the parity bit. The parity bit enables memory-control circuitry to keep tabs on the other 8 bits—a built-in cross-check for the integrity of each byte in the system.

Originally, all PC systems used parity-checked memory to ensure accuracy. Starting in 1994, most vendors began shipping systems without parity checking or any other means of detecting or correcting errors on the fly. These systems used cheaper nonparity memory modules, which saved about 10%–15% on memory costs for a system.

Parity memory results in increased initial system cost, primarily because of the additional memory bits involved. Parity can't correct system errors, but because parity can detect errors, it can make the user aware of memory errors when they happen.

Since then, Intel and other chipset manufacturers have put support for ECC memory in many chipsets (especially so in their higher-end models). The low-end chipsets, however, typically lack support for either parity or ECC. If more reliability is important to you, make sure the systems you purchase have this ECC support.

How Parity Checking Works

IBM originally established the odd parity standard for error checking. The following explanation might help you understand what is meant by odd parity. As the 8 individual bits in a byte are stored in memory, a parity generator/checker, which is either part of the CPU or located in a special chip on the motherboard, evaluates the data bits by adding up the number of 1s in the byte. If an even number of 1s is found, the parity generator/checker creates a 1 and stores it as the ninth bit (parity bit) in the parity memory chip. That makes the sum for all 9 bits (including the parity bit) an odd number. If the original sum of the 8 data bits is an odd number, the parity bit created would be a 0, keeping the sum for all 9 bits an odd number. The basic rule is that the value of the parity bit is always chosen so that the sum of all 9 bits (8 data bits plus 1 parity bit) is stored as an odd number. If the system used even parity, the example would be the same except the parity bit would be created to ensure an even sum. It doesn't matter whether even or odd parity is used; the system uses one or the other, and it is completely transparent to the memory chips involved. Remember that the 8 data bits in a byte are numbered 0 1 2 3 4 5 6 7. The following examples might make it easier to understand:

```
Data bit number:     0 1 2 3 4 5 6 7     Parity bit
Data bit value:      1 0 1 1 0 0 1 1     0
```

In this example, because the total number of data bits with a value of 1 is an odd number (5), the parity bit must have a value of 0 to ensure an odd sum for all 9 bits.

Here is another example:

```
Data bit number:     0 1 2 3 4 5 6 7     Parity bit
Data bit value:      1 1 1 1 0 0 1 1     1
```

In this example, because the total number of data bits with a value of 1 is an even number (6), the parity bit must have a value of 1 to create an odd sum for all 9 bits.

When the system reads memory back from storage, it checks the parity information. If a (9-bit) byte has an even number of bits, that byte must have an error. The system can't tell which bit has changed or whether only a single bit has changed. If 3 bits changed, for example, the byte still flags a parity-check error; if 2 bits changed, however, the bad byte could pass unnoticed. Because multiple bit errors (in a single byte) are rare, this scheme gives you a reasonable and inexpensive ongoing indication that memory is good or bad.

The following examples show parity-check messages for three types of older systems:

For the IBM PC:	`PARITY CHECK x`
For the IBM XT:	`PARITY CHECK x     yyyyy (z)`
For the IBM AT and late model XT:	`PARITY CHECK x     yyyyy`

where `x` is 1 or 2:

1 = Error occurred on the motherboard.

2 = Error occurred in an expansion slot.

In this example, `yyyyy` represents a number from 00000 through FFFFF that indicates, in hexadecimal notation, the byte in which the error has occurred.

Also, (z) is (S) or (E):

(S) = Parity error occurred in the system unit.

(E) = Parity error occurred in an optional expansion chassis.

Note

An expansion chassis was an option IBM sold for the original PC and XT systems to add more expansion slots.

When a parity-check error is detected, the motherboard parity-checking circuits generate a nonmaskable interrupt (NMI), which halts processing and diverts the system's attention to the error. The NMI causes a routine in the ROM to be executed. On some older IBM systems, the ROM parity-check routine halts the CPU. In such a case, the system locks up, and you must perform a hardware reset or a power-off/power-on cycle to restart the system. Unfortunately, all unsaved work is lost in the process.

Most systems do not halt the CPU when a parity error is detected; instead, they offer you the choice of rebooting the system or continuing as though nothing happened. Additionally, these systems might display the parity error message in a different format from IBM, although the information presented is basically the same. For example, most systems with a Phoenix BIOS display one of these messages:

```
Memory parity interrupt at xxxx:xxxx
Type (S)hut off NMI, Type (R)eboot, other keys to continue
```

or

```
I/O card parity interrupt at xxxx:xxxx
Type (S)hut off NMI, Type (R)eboot, other keys to continue
```

The first of these two messages indicates a motherboard parity error (Parity Check 1), and the second indicates an expansion-slot parity error (Parity Check 2). Notice that the address given in the form xxxx:xxxx for the memory error is in a segment:offset form rather than a straight linear address, such as with IBM's error messages. The segment:offset address form still gives you the location of the error to a resolution of a single byte.

You have three ways to proceed after viewing this error message:

- You can press S, which shuts off parity checking and resumes system operation at the point where the parity check first occurred.
- You can press R to force the system to reboot, losing any unsaved work.
- You can press any other key to cause the system to resume operation with parity checking still enabled.

If the problem occurs, it is likely to cause another parity-check interruption. It's usually prudent to press S, which disables the parity checking so you can then save your work. In this case, it's best to save your work to a floppy disk to prevent the possible corruption of the hard disk. You should also avoid overwriting any previous (still good) versions of whatever file you are saving because you could be saving a bad file caused by the memory corruption. Because parity checking is now disabled, your save operations will not be interrupted. Then, you should power the system off, restart it, and run whatever memory diagnostics software you have to try to track down the error. In some cases, the POST finds the error on the next restart, but you usually need to run a more sophisticated diagnostics program—perhaps in a continuous mode—to locate the error.

Systems with an AMI BIOS display the parity error messages in one of the following forms:

```
ON BOARD PARITY ERROR ADDR (HEX) = (xxxxx)
```

or

```
OFF BOARD PARITY ERROR ADDR (HEX) = (xxxxx)
```

These messages indicate that an error in memory has occurred during the POST, and the failure is located at the address indicated. The first one indicates that the error occurred on the motherboard, and the second message indicates an error in an expansion slot adapter card. The AMI BIOS can also display memory errors in one of the following manners:

```
Memory Parity Error at xxxxx
```

or

```
I/O Card Parity Error at xxxxx
```

These messages indicate that an error in memory has occurred at the indicated address during normal operation. The first one indicates a motherboard memory error, and the second indicates an expansion slot adapter memory error.

Although many systems enable you to continue processing after a parity error and even allow disabling further parity checking, continuing to use your system after a parity error is detected can be dangerous. The idea behind letting you continue using either method is to give you time to save any unsaved work before you diagnose and service the computer, but be careful how you do this.

Note that these messages can vary depending not only on the ROM BIOS but also on your operating system. Protected mode operating systems, such as most versions of Windows, trap these errors and run their own handler program that displays a message different from what the ROM would have displayed. The message might be associated with a blue screen or might be a trap error, but it usually indicates that it is memory or parity related.

Caution

When you are notified of a memory parity error, remember the parity check is telling you that memory has been corrupted. Do you want to save potentially corrupted data over the good file from the last time you saved? Definitely not! Be sure you save your work with a different filename. In addition, after a parity error, save only to a floppy disk if possible and avoid writing to the hard disk; there is a slight chance that the hard drive could become corrupt if you save the contents of corrupted memory.

After saving your work, determine the cause of the parity error and repair the system. You might be tempted to use an option to shut off further parity checking and simply continue using the system as though nothing were wrong. Doing so is like unscrewing the oil pressure warning indicator bulb on a car with an oil leak so the oil pressure light won't bother you anymore!

Error-Correcting Code (ECC)

ECC goes a big step beyond simple parity-error detection. Instead of just detecting an error, ECC allows a single bit error to be corrected, which means the system can continue without interruption and without corrupting data. ECC, as implemented in most PCs, can only detect, not correct, double-bit errors. Because studies have indicated that approximately 98% of memory errors are the single-bit variety, the most commonly used type of ECC is one in which the attendant memory controller detects and corrects single-bit errors in an accessed data word (double-bit errors can be detected but not corrected). This type of ECC is known as *single-bit error-correction double-bit error detection (SEC-DED)* and requires an additional 7 check bits over 32 bits in a 4-byte system and an additional 8 check bits over 64 bits in an 8-byte system. If the system uses SIMMs, two 36-bit (parity) SIMMs are added for each bank (for a total of 72 bits), and ECC is done at the bank level. If the system uses DIMMs, a single parity/ECC 72-bit DIMM is used as a bank and provides the additional bits. RIMMs are installed in singles or pairs, depending on the chipset and motherboard. They must be 18-bit versions if parity/ECC is desired.

ECC entails the memory controller calculating the check bits on a memory-write operation, performing a compare between the read and calculated check bits on a read operation, and, if necessary, correcting bad bits. The additional ECC logic in the memory controller is not very significant in this age of inexpensive, high-performance VLSI logic, but ECC actually affects memory performance on writes. This is because the operation must be timed to wait for the calculation of check bits and, when the system waits for corrected data, reads. On a partial-word write, the entire word must first be read, the affected byte(s) rewritten, and then new check bits calculated. This turns partial-word write operations into slower read-modify writes. Fortunately, this performance hit is very small, on the order of a few percent at maximum, so the tradeoff for increased reliability is a good one.

Most memory errors are of a single-bit nature, which ECC can correct. Incorporating this fault-tolerant technique provides high system reliability and attendant availability. An ECC-based system is a good choice for servers, workstations, or mission-critical applications in which the cost of a potential memory error outweighs the additional memory and system cost to correct it, along with ensuring that it does not detract from system reliability. If you value your data and use your system for important (to you) tasks, you'll want ECC memory.

RAM Upgrades

Adding memory to a system is one of the most useful upgrades you can perform and also one of the least expensive—especially when you consider the increased performance of Windows and Linux when you give them access to more memory. In some cases, doubling the memory can practically double the speed of a computer.

The following sections discuss adding memory, including selecting memory chips, installing memory chips, and testing the installation.

Upgrade Options and Strategies

Adding memory can be an inexpensive solution; the cost of mainstream memory is extremely low, and adding more memory can give your computer's performance a big boost.

How do you add memory to your PC? You have two options, listed in order of convenience and cost:

- Adding memory in vacant slots on your motherboard
- Replacing your current motherboard's memory with higher-capacity memory

If you decide to upgrade to a more powerful computer system or motherboard, you usually can't salvage the memory from your previous system. Most of the time it is best to plan on equipping a new board with the optimum type of memory that it supports.

Be sure to carefully weigh your future needs for computing speed and a multitasking operating system against the amount of money you spend to upgrade current equipment.

To determine at what point you should add memory, you can use the Performance Monitor (Perfmon.msc) built into Windows. You can launch it from the Start; Run dialog box or from a command prompt. To check memory usage, select Memory as the Performance object and enable the following counters:

- **Pages/Sec**—This counter measures the number of times per second that the system uses virtual (swapfile) memory rather than physical memory. A value above 20 indicates a potential problem. Check the virtual memory settings; if the counter remains above 20, and it is not during periods of heavy disk or file access, then you should consider installing more memory.

- **Committed Bytes and Available Bytes**—Committed Bytes tracks virtual memory in use; Available Bytes tracks physical memory available. Add more memory if you run short of available bytes.
- **Cache Bytes**—Measures the amount of RAM used for file system cache. Add more RAM if this amount drops below 4MB.

Tip

It is normal to see very high Pages/sec counts during periods of heavy disk or file access, such as when running a malware scan, indexing operation, defragmentation, etc. If the Available Bytes value doesn't decrease as the Pages/sec value increases, then the high Pages/sec numbers are probably due to application-generated disk access, and is not indicative of any sort of insufficient memory problem or even a bottleneck in memory.

Before you add RAM to a system (or replace defective RAM chips), you must determine the memory modules required for your system. Your system documentation has this information.

If you need to replace a defective memory module or add more memory to your system, there are several ways to determine the correct module for your system:

- **Inspect the modules installed in your system**. Each module has markings that indicate its capacity and speed. RAM capacity and speed were discussed in detail earlier in this chapter. You can write down the markings on the memory module and use them to determine the type of memory you need. Check with a local store or an online memory vendor for help.
- **Look up your system using the online memory-configuration utility provided by your preferred memory vendor**. Originally, these configuration utilities were primarily for users of name-brand systems. However, most vendors have now added major motherboard brands and models to their databases. Therefore, if you know your system or motherboard brand and model, you can find the memory that is recommended.
- **Download and run analysis software provided by the memory module maker or from a third party**. SiSoftware Sandra and similar programs use the SPD chip on each module to determine this information.
- **Consult your system documentation**. I list this option last for a reason. If you have installed BIOS upgrades, you might be able to use larger and faster memory than your documentation lists as supported by your system. You should check the latest tech notes and documentation available online for your system and check the BIOS version installed in your system to determine which memory-related features it has. A BIOS upgrade might enable your system to use faster memory.

Adding the wrong modules to a system can make it as unreliable as leaving a defective module installed and trying to use the system in that condition.

Note

Before upgrading an older Pentium (P5 class) system beyond 64MB of RAM, be sure your chipset supports caching more than 64MB. Adding RAM beyond the amount supported by your L2 cache controller slows performance rather than increases it. Pentium II and later processors, including the AMD Athlon, Duron, and Sempron families, have the L2 cache controller integrated in the processor (not the chipset), which supports caching up to 4GB and beyond on most newer models.

Purchasing Memory

When purchasing memory, there are some issues you need to consider. Some are related to the manufacturing and distribution of memory, whereas others depend on the type of memory you are purchasing. This section covers some of the issues you should consider when purchasing memory.

Suppliers

Many companies sell memory, but only a few companies actually make memory. Additionally, only a few companies make memory chips, but many more companies make memory modules such as SIMMs, DIMMs, and RIMMs. Most of the companies that make the actual RAM chips also make modules containing their own chips. Other companies, however, strictly make modules; these companies purchase memory chips from several chip makers and then produce modules with these chips. Finally, some companies don't make either the chips or modules. Instead, they purchase modules made by other companies and relabel them.

I refer to memory modules made by the chip manufacturers as *first-party modules*, whereas those made by module (but not chip) manufacturers I call *second-party modules*. Finally, those that are simply relabeled first- or second-party modules under a different name are called *third-party modules*. I always prefer to purchase first- or second-party modules if I can because they are better documented. In essence, they have a better pedigree and their quality is generally more assured. Not to mention that purchasing from the first or second party eliminates one or more middlemen in the distribution process as well.

First-party manufacturers (where the same company makes the chips and the modules) include Micron (www.crucial.com), Infineon (formerly Siemens), Samsung, Mitsubishi, Toshiba, NEC, and others. Second-party companies that make the modules (but not the chips) include Kingston, Viking, PNY, Simple Tech, Smart, Mushkin, and OCZ Technologies. At the third-party level you are not purchasing from a manufacturer but from a reseller or remarketer instead.

Most of the large manufacturers don't sell small quantities of memory to individuals, but some have set up factory outlet stores where individuals can purchase as little as a single module. One of the largest memory manufacturers in the world, Micron, sells direct to the consumer at www.crucial.com. Because you are buying direct, the pricing at these outlets is often competitive with second- and third-party suppliers.

Considerations in Purchasing DIMMs

When you are purchasing DIMMs, here are the main things to consider:

- Do you need SDR, DDR, DDR2, or DDR3 versions?
- Do you need ECC or non-ECC?
- Do you need standard (unbuffered) or registered versions?
- What speed grade do you need?
- Do you need a specific column address strobe (CAS) latency?

Currently, DIMMs come in SDR (SDRAM), DDR, DDR2 and DDR3 versions. They are not interchangeable because they use completely different signaling and have different notches to prevent a mismatch. High-reliability systems such as servers can use ECC versions, although most desktop systems use the less-expensive non-ECC types. Most systems use standard unbuffered DIMMs, but file server or workstation motherboards designed to support very large amounts of memory might require registered DIMMs (which also include ECC support). Registered DIMMs contain their own memory registers, enabling the module to hold more memory than a standard DIMM. DIMMs come in a variety of speeds, with the rule that you can always substitute a faster one for a slower one, but not vice versa.

Another speed-related issue is the column address strobe (CAS) latency. Sometimes this specification is abbreviated CAS or CL and is expressed in a number of cycles, with lower numbers indicating higher speeds (fewer cycles). The lower CAS latency shaves a cycle off a burst mode read, which marginally improves memory performance. Single data rate DIMMs are available in CL3 or CL2 versions. DDR DIMMs are available in CL2.5 or CL2 versions. DDR2 DIMMs are available in CL 3, 4 or 5. DDR3 DIMMs are available in CL 7, 8, and 9. With all memory types, the lowest CL number is the fastest (and usually the most expensive) memory type. You can mix DIMMs with different CAS latency ratings, but the system usually defaults to cycling at the slower speeds of the lowest common denominator.

Considerations in Purchasing Obsolete Memory

Many people are surprised to find that obsolete memory types cost much more than that used by current systems. This is because of simple supply and demand, what is least popular generally costs the most. This can make adding memory to older systems cost prohibitive.

Most Pentium systems after 1995 used EDO SIMMs that were non-ECC and rated for 60ns access time. If your system is older than that, you might need FPM memory instead of EDO. The FPM and EDO types are interchangeable in many systems, but some older systems do not accept the EDO type. Some Pentium 4 systems use RIMMs, which are available in 184-pin and 232-pin versions. Although they appear to be the same size, they are not interchangeable. If the system supports ECC, you might need (or want) ECC versions. You can mix ECC and non-ECC modules, but in that case the system defaults to non-ECC mode.

Tip

Instead of buying "new" obsolete memory for older systems, check with computer repair shops, Craigslist, or other users who might have a collection of old parts.

High-reliability systems might want or need ECC versions, which have extra ECC bits. As with other memory types, you can mix ECC and non-ECC types, but systems can't use the ECC capability.

Replacing Modules with Higher-Capacity Versions

If all the memory module slots on your motherboard are occupied, your best option is to remove an existing bank of memory and replace it with higher-capacity modules.

However, just because higher-capacity modules are available to plug into your motherboard, don't automatically assume the higher-capacity memory will work. Your system's chipset, BIOS, and OS set limits on the capacity of the memory you can use. Check your system or motherboard documentation to see which size modules work with it before purchasing the new RAM. You should make sure you have the latest BIOS for your motherboard when installing new memory.

If your system supports dual- or triple-channel memory, you must use modules in matched pairs or triples (depending on which type your system supports) and install them in the correct location on the motherboard. You should consult your motherboard manual for details.

Installing Memory Modules

When you install or remove memory, you are most likely to encounter the following problems:

- Electrostatic discharge
- Improperly seated modules
- Incorrect memory configuration settings in the BIOS Setup

To prevent electrostatic discharge (ESD) when you install sensitive memory chips or boards, you shouldn't wear synthetic-fiber clothing or leather-soled shoes because these promote the generation of static charges. Remove any static charge you are carrying by touching the system chassis before you begin, or better yet, wear a good commercial grounding strap on your wrist. You can order one from any electronics parts store. A grounding strap consists of a conductive wristband grounded at the other end through a 1-meg ohm resistor by a wire clipped to the system chassis. Be sure the system you are working on is unplugged.

Caution

Be sure to use a properly designed commercial grounding strap; do not make one yourself. Commercial units have a 1-meg ohm resistor that serves as protection if you accidentally touch live power. The resistor ensures that you do not become the path of least resistance to the ground and therefore become electrocuted. An improperly designed strap can cause the power to conduct through you to the ground, possibly killing you.

Follow this procedure to install memory on a typical desktop PC:

1. Shut down the system and unplug it. As an alternative to unplugging it, you can turn off the power supply using the on/off switch on the rear of some power supplies. Wait about 10 seconds for any remaining current to drain from the motherboard.
2. Open the system. See the system or case instructions for details.
3. Connect a static guard wrist strap to your wrist and then to a metal portion of the system chassis, such as the frame. Make sure the metal plate on the inside of the wrist strap is tight against the skin of your wrist.
4. Some motherboards feature an LED that glows as long as the motherboard is receiving power. Wait until the LED dims before removing or installing memory.
5. Move obstructions inside the case, such as cables or wires, out of the way of the memory modules and empty sockets. If you must remove a cable or wire, note its location and orientation so you can replace it later.
6. If you need to remove an existing module, flip down the ejector tab at each end of the module and lift the module straight up out of the socket. Note the keying on the module.
7. Note the specific locations needed if you are inserting modules to operate in dual-channel mode. The sockets used for dual-channel memory might use a different-colored plastic to distinguish them from other sockets, but ultimately you should consult the documentation for your motherboard or system to determine the proper orientation.
8. To insert a module into a socket, ensure that the ejector tabs are flipped down on the socket you plan to use. DIMMs are keyed by notches along the bottom connector edges that are offset from the center so they can be inserted in only one direction, as shown in Figure 6.16.

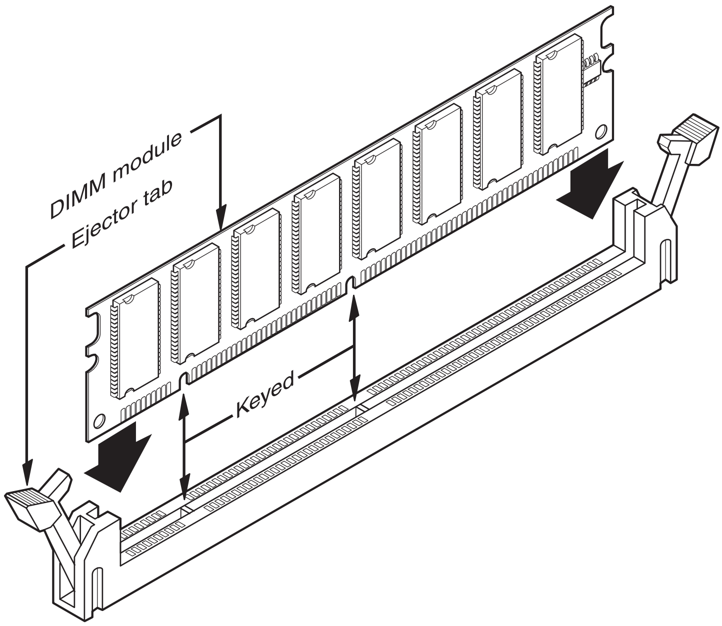


Figure 6.16 DIMM keys match the protrusions in the DIMM sockets. SDR/DDR/DDR2/DDR3 DIMM keys are similar but not exactly the same.

9. Push down on the module until the ejector tabs lock into place in the notch on the side of the module. It's important that you not force the module into the socket. If the module does not slip easily into the slot and then snap into place, it is probably not oriented or aligned correctly. Forcing the module could break it or the socket. If installing RIMMs, you need to fill any empty RIMM sockets with continuity modules. Refer to Figure 6.14 for details.
10. Replace any cables or wires you disconnected.
11. Close the system, reconnect the power cable, and turn on the PC.

The SIMMs used in older systems are oriented by a notch on one side of the module that is not present on the other side, as shown in Figure 6.17. The socket has a protrusion that must fit into this notched area on one side of the module. This protrusion makes installing a SIMM backward impossible unless you break the connector or the module. Figure 6.18 details the notch and locking clip.

After installing the memory and putting the system back together, you might have to run the BIOS Setup and resave with the new amount of memory being reported. Most newer systems automatically detect the new amount of memory and reconfigure the BIOS Setup settings for you. Most newer systems also don't require setting any jumpers or switches on the motherboard to configure them for your new memory.

After configuring your system to work properly with the additional memory, you might want to run a memory-diagnostics program to ensure that the new memory works properly.

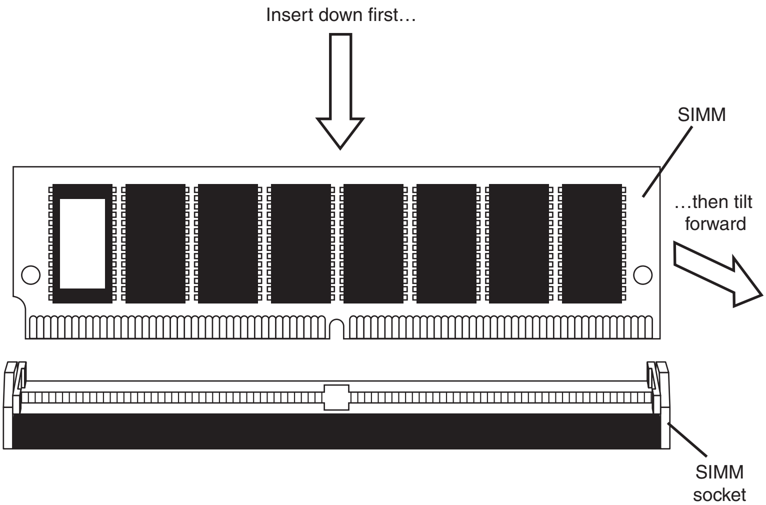


Figure 6.17 The notch on this SIMM is shown on the left side. Insert the SIMM at a 45° angle and then tilt it forward until the locking clips snap into place.

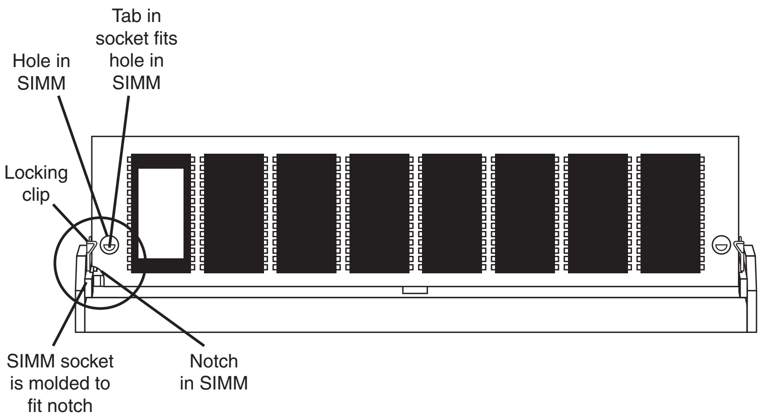


Figure 6.18 This figure shows the SIMM inserted in the socket with the notch aligned, the locking clip locked, and the hole in the SIMM aligned with the tab in the socket.

Troubleshooting Memory

Memory problems can be difficult to troubleshoot. For one thing, computer memory is still mysterious to people because it is a kind of "virtual" thing that can be hard to grasp. The other difficulty is that memory problems can be intermittent and often look like problems with other areas of the system, even software. This section shows simple troubleshooting steps you can perform if you suspect you are having a memory problem.

To troubleshoot memory, you first need some memory-diagnostics testing programs. You already have several and might not know it. Every motherboard BIOS has a memory diagnostic in the POST that runs when you first turn on the system. In most cases, you also receive a memory diagnostic on a utility disk that came with your system. Many commercial diagnostics programs are on the market, and almost all of them include memory tests.

When the POST runs, it not only tests memory, but also counts it. The count is compared to the amount counted the last time BIOS Setup was run; if it is different, an error message is issued. As the POST runs, it writes a pattern of data to all the memory locations in the system and reads that pattern back to verify that the memory works. If any failure is detected, you see or hear a message. Audio messages (beeping) are used for critical or "fatal" errors that occur in areas important for the system's operation. If the system can access enough memory to at least allow video to function, you see error messages instead of hearing beep codes.

See the disc accompanying this book for detailed listings of the BIOS beep and other error codes, which are specific to the type of BIOS you have. These BIOS codes are found in the Technical Reference section of the disc in printable PDF format for your convenience. For example, most Intel motherboards use the Phoenix BIOS. Several beep codes are used in that BIOS to indicate fatal memory errors.

If your system makes it through the POST with no memory error indications, there might not be a hardware memory problem, or the POST might not be able to detect the problem. Intermittent memory errors are often not detected during the POST, and other subtle hardware defects can be hard for the POST to catch. The POST is designed to run quickly, so the testing is not nearly as thorough as it could be. That is why you often have to boot from a standalone diagnostic disk and run a true hardware diagnostic to do more extensive memory testing. These types of tests can be run continuously and be left running for days if necessary to hunt down an elusive intermittent defect.

Fortunately several excellent memory test programs are available for free download. Here are some I recommend:

- **Microsoft Windows Memory Diagnostic**—http://oca.microsoft.com/en/windiag.asp
- **Memtest86**—www.memtest86.com

Not only are these free, but they are available in a bootable CD format, which means you don't have to install any software on the system you are testing. The bootable format is actually required in a way since Windows and other OSs prevent the direct access to memory and other hardware required for testing. These programs use algorithms that write different types of patterns to all of the memory in the system, testing every bit to ensure it reads and writes properly. They also turn off the processor cache in order to ensure direct testing of the modules and not the cache. Some, such as Windows Memory Diagnostic, will even indicate the module that is failing should an error be encountered. Note that a version of the Windows Memory Diagnostic is also included with Windows 7/Vista. It can be found as part of the Administrative tools, as well as on the bootable install DVDs under the Repair option.

One problem with software based memory diagnostics is that they do only pass/fail type testing; that is, all they can do is write patterns to memory and read them back. They can't determine how close the memory is to failing—only whether it worked. For the highest level of testing, the best thing to have is a dedicated memory test machine, usually called a *module tester*. These devices enable you to insert a module and test it thoroughly at a variety of speeds, voltages, and timings to let you know for certain whether the memory is good or bad. Versions of these testers are available to handle all types of memory modules. I have defective modules, for example, that work in some systems (slower ones) but not others. What I mean is that the same memory test program fails the module in one machine but passes it in another. In the module tester, it is always identified as bad right down to the individual bit, and it even tells me the actual speed of the device, not just its rating. Companies that offer memory module testers include Tanisys (www.tanisys.com), CST (www.simmtester.com), and Innoventions (www.memorytest.com). They can be expensive, but for a high volume system builder or repair shop, using one of these module testers can save time and money in the long run.

After your operating system is running, memory errors can still occur, typically identified by error messages you might receive. Here are the most common:

- **Parity errors**—The parity-checking circuitry on the motherboard has detected a change in memory since the data was originally stored. (See the "How Parity Checking Works" section earlier in this chapter.)
- **General or global protection faults**—A general-purpose error indicating that a program has been corrupted in memory, usually resulting in immediate termination of the application. This can also be caused by buggy or faulty programs.
- **Fatal exception errors**—Error codes returned by a program when an illegal instruction has been encountered, invalid data or code has been accessed, or the privilege level of an operation is invalid.
- **Divide error**—A general-purpose error indicating that a division by 0 was attempted or the result of an operation does not fit in the destination register.

If you are encountering these errors, they could be caused by defective or improperly configured memory, but they can also be caused by software bugs (especially drivers), bad power supplies, static discharges, close proximity radio transmitters, timing problems, and more.

If you suspect the problems are caused by memory, there are ways to test the memory to determine whether that is the problem. Most of this testing involves running one or more memory test programs.

Another problem with software based diagnostics is running memory tests with the system caches enabled. This effectively invalidates memory testing because most systems have what is called a *write-back cache*. This means that data written to main memory is first written to the cache. Because a memory test program first writes data and then immediately reads it back, the data is read back from the cache, not the main memory. It makes the memory test program run very quickly, but all you tested was the cache. The bottom line is that if you test memory with the cache enabled, you aren't really writing to the SIMM/DIMMs, but only to the cache. Before you run any memory test programs, be sure your processor/memory caches are disabled. Many older systems have options in the BIOS Setup to turn off the caches. Current software based memory test software such as the Windows Memory Diagnostic and Memtest86 automatically turn off the caches on newer systems.

The following steps enable you to effectively test and troubleshoot your system RAM. Figure 6.19 provides a boiled-down procedure to help you step through the process quickly.

First, let's cover the memory-testing and troubleshooting procedures.

1. Power up the system and observe the POST. If the POST completes with no errors, basic memory functionality has been tested. If errors are encountered, go to the defect isolation procedures.
2. Restart the system and then enter your BIOS (or CMOS) Setup. In most systems, this is done by pressing the Del or F2 key during the POST but before the boot process begins (see your system or motherboard documentation for details). Once in BIOS Setup, verify that the memory count

is equal to the amount that has been installed. If the count does not match what has been installed, go to the defect isolation procedures.

3. Find the BIOS Setup options for cache and then set all cache options to disabled. Figure 6.20 shows a typical Advanced BIOS Features menu with the cache options highlighted. Save the settings and reboot to a bootable floppy or optical disc containing the memory diagnostics program.

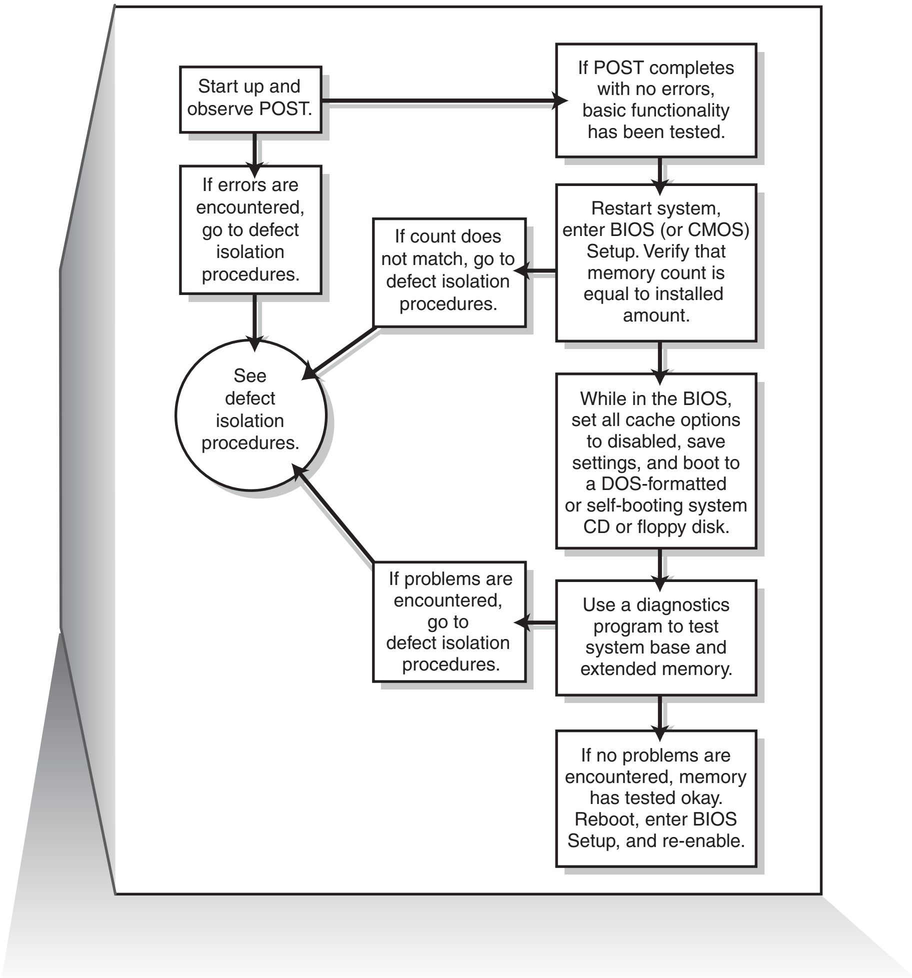


Figure 6.19 Testing and troubleshooting memory.

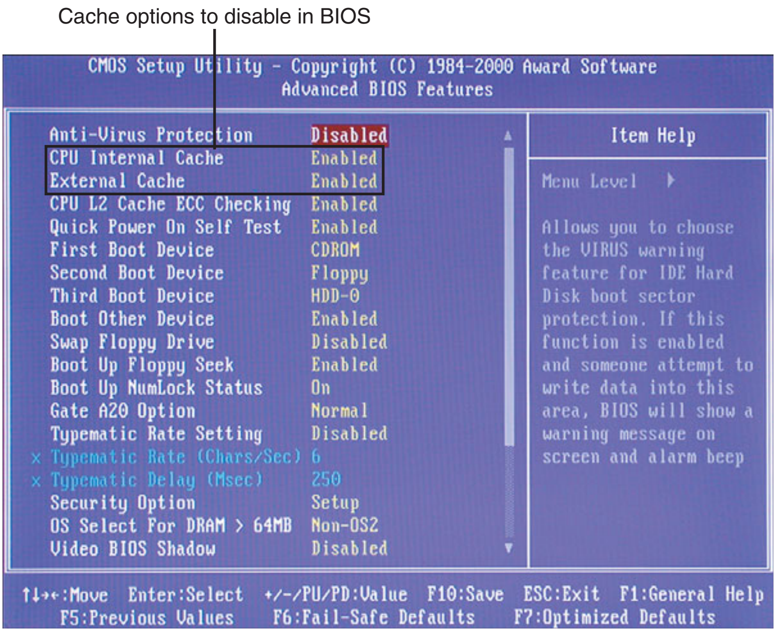


Figure 6.20 The CPU Internal (L1) and External (L2) caches must be disabled in the system BIOS Setup before you test system memory; otherwise, your test results will be inaccurate.

4. Follow the instructions that came with your diagnostic program to have it test the system base and extended memory. Most programs have a mode that enables them to loop the test—that is, to run it continuously, which is great for finding intermittent problems. If the program encounters a memory error, proceed to the defect isolation procedures.
5. If no errors are encountered in the POST or in the more comprehensive memory diagnostic, your memory has tested okay in hardware. Be sure at this point to reboot the system, enter the BIOS Setup, and reenable the cache. The system will run very slowly until the cache is turned back on.
6. If you are having memory problems yet the memory still tests okay, you might have a problem undetectable by simple pass/fail testing, or your problems could be caused by software or one of many other defects or problems in your system. You might want to bring the memory to a module tester for a more accurate analysis. Some larger PC repair shops have such a tester. I would also check the software (especially drivers, which might need updating), power supply, and system environment for problems such as static, radio transmitters, and so forth.

Memory Defect Isolation Procedures

To use these steps, I am assuming you have identified an actual memory problem that is being reported by the POST or disk-based memory diagnostics. If this is the case, see the following steps and Figure 6.21 for the steps to identify or isolate which module is causing the problem.

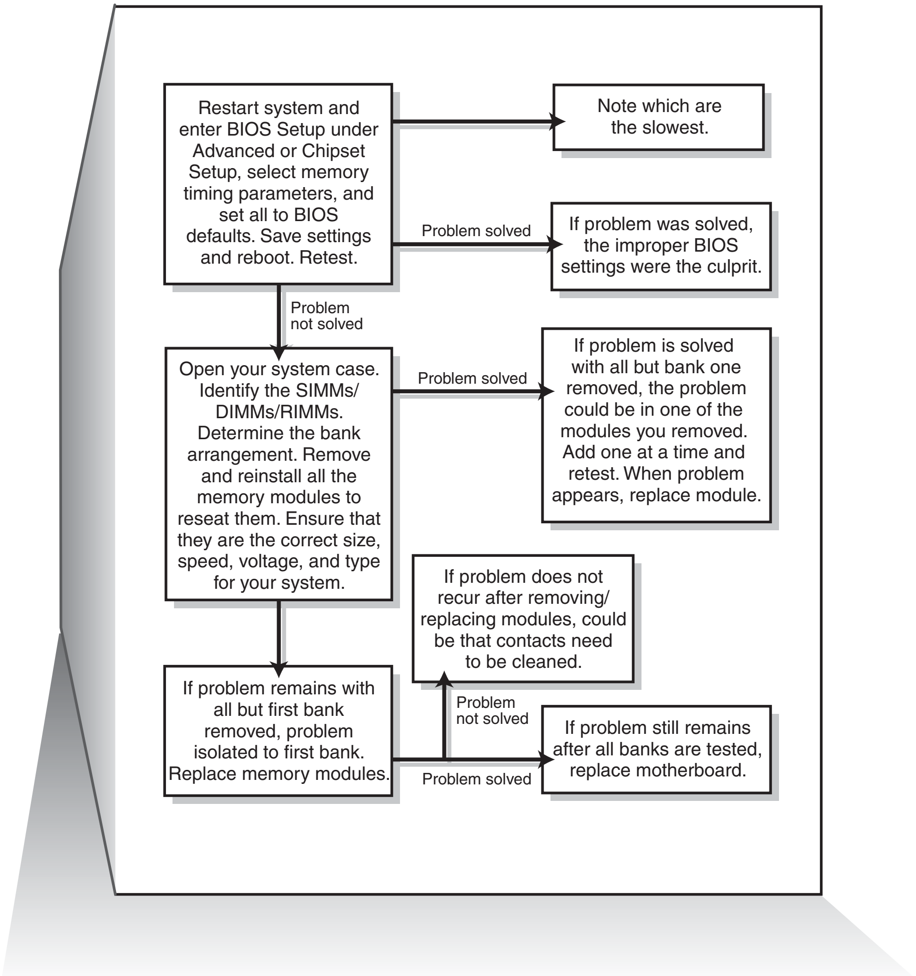
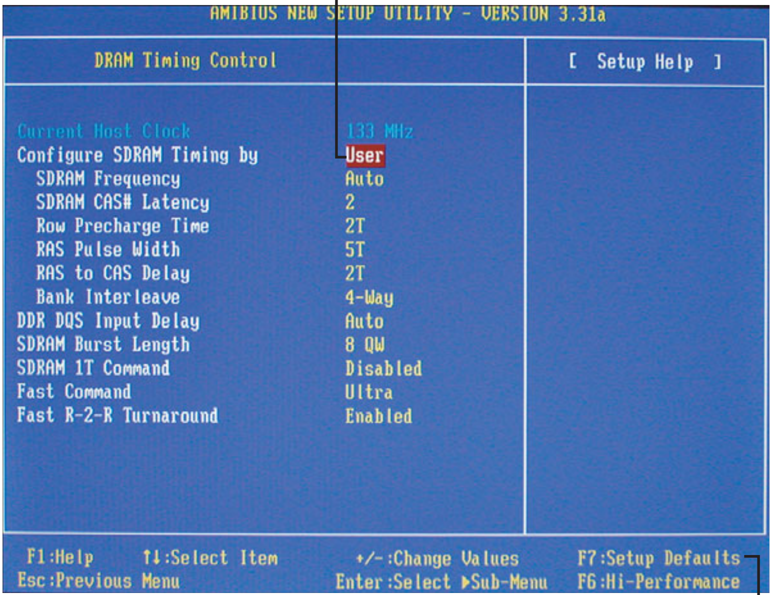


Figure 6.21 Follow these steps if you are still encountering memory errors after completing the steps in Figure 6.19.

1. Restart the system and enter the BIOS Setup. Under a menu usually called Advanced or Chipset Setup might be memory timing parameters. Select BIOS or Setup defaults, which are usually the slowest settings. If the memory timings have been manually set, as shown in Figure 6.20, reset the memory configuration to By SPD.
2. Save the settings, reboot, and retest using the testing and troubleshooting procedures listed earlier. If the problem has been solved, improper BIOS settings were the problem. If the problem remains, you likely do have defective memory, so continue to the next step.
3. Open the system for physical access to the modules on the motherboard. Identify the bank arrangement in the system. Using the manual or the legend silk-screened on the motherboard, identify which modules correspond to which banks. Remember that if you are testing a multichannel system, you must be sure you remove all of the modules in the same channel.

4. Remove all the memory except the first bank and then retest using the troubleshooting and testing procedures listed earlier (see Figure 6.22). If the problem remains with all but the first bank removed, the problem has been isolated to the first bank, which must be replaced.

Change this setting to SPD to revert to the module's default memory timings.



Press F7 to use Setup Defaults for memory and other system timings.

Figure 6.22 This system is using user-defined memory timings, which could cause the memory to be unstable.

5. Replace the memory in the first bank (preferably with known good spare modules, but you can also swap in others that you have removed) and then retest. If the problem still remains after testing all the memory banks (and finding them all to be working properly), it is likely the motherboard itself is bad (probably one of the memory sockets). Replace the motherboard and retest.
6. At this point, the first (or previous) bank has tested good, so the problem must be in the remaining modules that have been temporarily removed. Install the next bank of memory and retest. If the problem resurfaces now, the memory in that bank is defective. Continue testing each bank until you find the defective module.
7. Repeat the preceding step until all remaining banks of memory are installed and have been tested. If the problem has not resurfaced after you have removed and reinstalled all the memory, the problem was likely intermittent or caused by poor conduction on the memory contacts. Often simply removing and replacing memory can resolve problems because of the self-cleaning action between the module and the socket during removal and reinstallation.

The System Logical Memory Layout

The original PC had a total of 1MB of addressable memory, and the top 384KB of that was reserved for use by the system. Placing this reserved space at the top (between 640KB and 1,024KB, instead of at the bottom, between 0KB and 640KB) led to what is often called the *conventional memory barrier*. The constant pressures on system and peripheral manufacturers to maintain compatibility by never breaking from the original memory scheme of the first PC has resulted in a system memory structure that is (to put it kindly) a mess. Almost two decades after the first PC was introduced, even the newest systems are limited in many important ways by the memory map of the first PCs.

The original PC used an Intel 8088 processor that could run only 16-bit instructions or code, which ran in what was called the *real mode* of the processor. These early processors had only enough address lines to access up to 1MB of memory, and the last 384KB of that was reserved for use by the video card as video RAM, other adapters (for on-card ROM BIOS or RAM buffers), and finally the motherboard ROM BIOS.

The 286 processor brought more address lines, enough to allow up to 16MB of RAM to be used, and a new mode called protected mode that you had to be in to use it. One area of confusion was that RAM was now noncontiguous; that is, the operating system could use the first 640KB and the last 15MB, but not the 384KB of system reserved area that sat in between.

When Intel released the first 32-bit processor in 1985 (the 386DX), the memory architecture of the system changed dramatically. There were now enough address lines for the processor to use 4GB of memory, but this was accessible only in 32-bit protected mode, in which only 32-bit instructions or code could run. Unfortunately, it took 10 years for the industry to transition from 16-bit to 32-bit operating systems and applications. From a software instruction perspective, all the 32-bit processors since the 386 are really just faster versions of the same.

When AMD released the first x86-64 processor in 2003 (Intel followed suit in 2004), the 64-bit era was born. In addition to 16-bit and 32-bit modes, these chips have a 64-bit mode as well. 64-bit processors have three distinctly different modes, with different memory architectures in each. For backward compatibility, 64-bit processors can run in 64-bit, 32-bit, or 16-bit modes, and 32-bit processors can run in 32-bit or 16-bit modes, each with different memory limitations. For example, a 64-bit processor running in 32-bit mode can only address 4GB of RAM, and a 64-bit or 32-bit processor running in 16-bit mode can only address 1MB of RAM. All Intel-compatible PC processors begin operation in 16-bit real mode when they are powered on. When a 32-bit or 64-bit operating system loads, it is that operating system code that instructs the processor to switch into 32-bit or 64-bit protected mode.

When an operating system such as Windows is loaded, the processor is switched into 32-bit protected mode early in the loading sequence. Then, 32-bit drivers for all the hardware can be loaded, and then the rest of the operating system can load. In 32-bit protected mode, the operating systems and applications can access all the memory in the system up to 4GB. Similarly, on a 64-bit operating system, the system switches into 64-bit protected mode early in the boot process and loads 64-bit drivers, followed by the remainder of the operating system.

The 32-bit editions of Windows support 4GB of physical memory (RAM). What many don't realize is that the PC system hardware uses some or all of the fourth gigabyte for the BIOS, motherboard resources, memory mapped I/O, PCI configuration space, device memory (graphics aperture), VGA memory, and so on. This means that if you install 4GB (or more) RAM, none of it past 4GB will be seen at all, and most or all of the fourth gigabyte (that is, the RAM between 3GB and 4GB) will be disabled because it is already occupied by other system hardware. This is called the *3GB limit*, which is analogous to the 640K memory limit we had on 16-bit systems in the 1980s. The 16-bit addressing supported 1MB, but the upper 384K was already in use by the system hardware (BIOS, video, adapter ROM, and so on).

Figure 6.23 shows the memory map for a modern system using an Intel G45 chipset, which supports a maximum of 16GB of RAM. For a 32-bit OS, the line labeled "Top of usable DRAM (32-bit OS)" is at 4,096MB. Note that the PCI memory range, FLASH, APIC (Advanced Programmable Interrupt Controller), and Reserved areas take up a total of 770MB of the memory below 4GB. You can also see the 384K (0.375MB) of memory below 1MB that is used by the system as well. This means that if you are running a 32-bit OS, even if you have 4GB of RAM installed, the amount usable by the OS would be 4,096MB – 770MB – 0.375MB, which is 3,325.625MB (or about 3.24GB, rounded down).

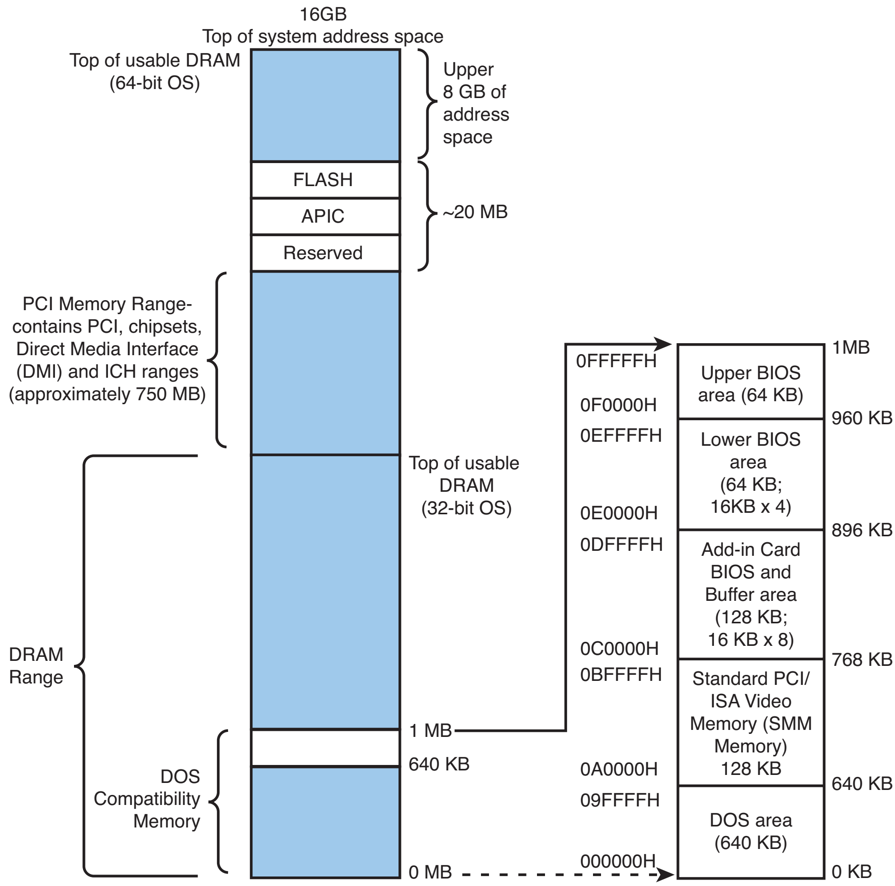


Figure 6.23 Memory map for a system using an Intel G45 chipset.

Can any of that unused memory between 3GB and 4GB be reclaimed? For those running a 32-bit OS, the answer is no. However, if you are running a 64-bit OS on a system that supports memory remapping (primarily a function of the motherboard chipset and BIOS), then the answer is yes. Most newer motherboard chipsets have a feature that can remap the otherwise disabled RAM in the fourth GB to the fifth (or higher) GB, where it will be both visible to and usable by a 64-bit OS. Note, however, that if the motherboard doesn't support remapping, then even when a 64-bit OS is being run, the memory will be lost.

Note that the 3GB limit is not as strictly defined as it was with the 640K limit. This means that if you do install 4GB, you might get to use as much as 3.5GB of it, or possibly as little as 2.5GB or less. It depends largely on the types of buses in the system as well as the type and number of video cards installed. With a single low-end video card, you may have access to 3.5GiB. However, on a newer system with two or more PCIe x16 slots, and especially with two or more high-end PCI Express video cards installed, you may drop the usable limit to something close to 2GiB.

For running 32-bit editions of Windows, I used to recommend installing a maximum of 3GB RAM, because most if not all of the fourth GB is unusable. However, on systems that support dual-channel memory, it is often just cheaper to install two 2GB modules to get 4GB than it is to install two 1GB modules and two 512MB in order to get 3GB. On desktop systems that support dual-channel memory, you would not want to install three 1GB modules, because in that case not all the memory would run in dual-channel mode.

Index

Numbers

A

B

C

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W

X

Y

Z