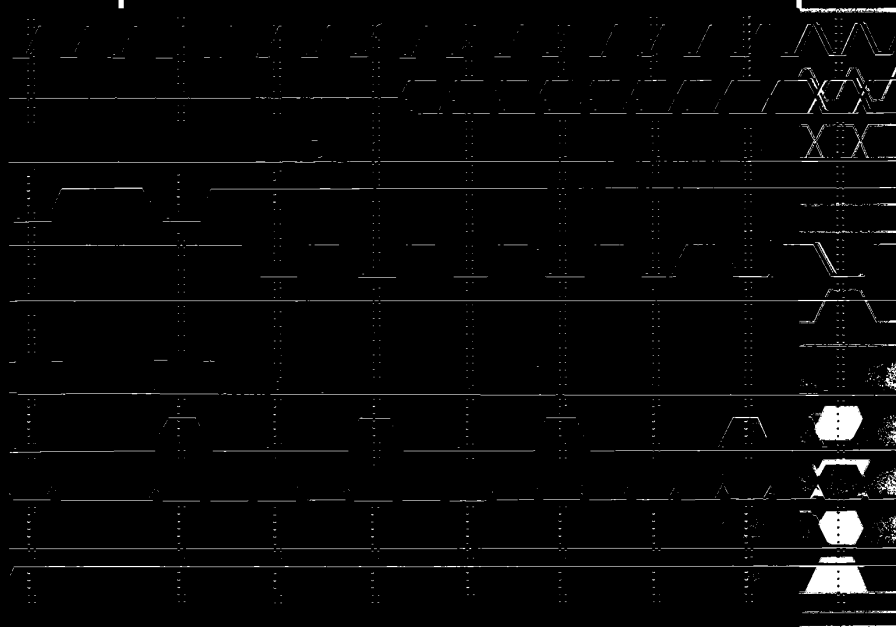




# HIGH PERFORMANCE MEMORIES

New architecture DRAMs and SRAMs  
evolution and function



**Betty Prince**



Copyright © 1996, 1999 by John Wiley & Sons Ltd.  
Baffins Lane, Chichester,  
West Sussex PO19 1UD, England

National 01243 779777  
International (+44) 1243 779777

e-mail (for orders and customer service enquiries): cs-books@wiley.co.uk

Visit our Home Page on <http://www.wiley.co.uk> or <http://www.wiley.com>

All Rights Reserved. No part of this book may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except under the terms of the Copyright, Designs and Patents Act 1988 or under the terms of a licence issued by the Copyright Licensing Agency, 90 Tottenham Court Road, London W1P 9HE, UK, without the permission in writing of the Publisher, with the exception of any material supplied specifically for the purpose of being entered and executed on a computer system, for the exclusive use by the purchaser of the publication.

Neither the author(s) nor John Wiley & Sons Ltd accept any responsibility or liability for loss or damage occasioned to any person or property through using the material, instructions, methods or ideas contained herein, or acting or refraining from acting as a result of such use. The author(s) and Publisher expressly disclaim all implied warranties, including merchantability or fitness for any particular purpose. There will be no duty on the author(s) or Publisher, with the exception of any material supplied specifically for the purpose of being entered and executed on a computer system, for the exclusive use by the purchaser of the publication.

Designations used by companies to distinguish their products are often claimed as trademarks. In all instances where John Wiley & Sons is aware of a claim, the product names appear in initial capital or all capital letters. Readers, however, should contact the appropriate companies for more complete information regarding trademarks and registration.

#### *Other Wiley Editorial Offices*

John Wiley & Sons, Inc., 605 Third Avenue,  
New York, NY 10158-0012, USA

Wiley-VCH Verlag GmbH  
Pappelallee 3, D-69469 Weinheim, Germany

Jacaranda Wiley Ltd, 33 Park Road, Milton,  
Queensland 4064, Australia

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01,  
Jin Xing Distripark, Singapore 0512

John Wiley & Sons (Canada) Ltd, 22 Worcester Road,  
Rexdale, Ontario M9W 1L1, Canada

#### *Library of Congress Cataloging-in-Publication Data*

#### *British Library Cataloguing in Publication Data*

A catalogue record for this book is available from the British Library

ISBN 0 471 986100

Typeset in 10/12pt Palatino by Keyword Typesetting Services Ltd., Wallington, Surrey

Printed and bound in Great Britain by Bookcraft (Bath) Ltd.

This book is printed on acid-free paper responsibly manufactured from sustainable forestation, for which at least two trees are planted for each one used for paper production.

## Cont

### Acknowledgements About the author Introduction

### 1 Overview

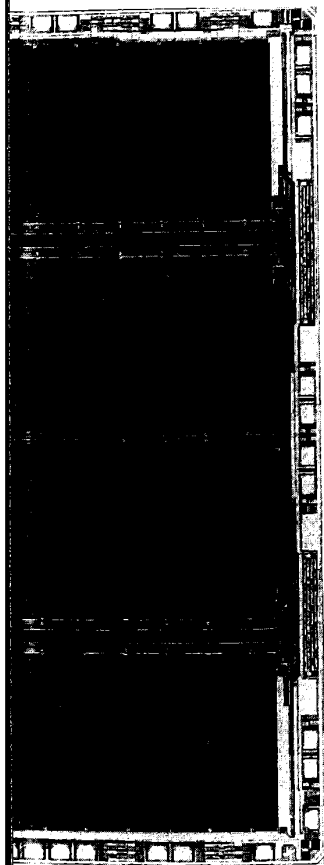
- 1.1 Overview
- 1.2 New
- 1.3 Men  
1.3.1  
1.3.2  
1.3.3
- 1.4 Effect
- 1.5 Effect  
Bibliography

### 2 High Performance

- 2.1 The
- 2.2 System
- 2.3 System  
2.3.1  
2.3.2  
2.3.3  
2.3.4  
2.3.5
- 2.4 Overview
- 2.5 Main  
2.5.1  
2.5.2  
2.5.3  
2.5.4
- 2.6 DRAM  
2.6.1  
2.6.2  
2.6.3  
2.6.4
- 2.7 Peripherals  
2.7.1
- 2.8 Conclusion  
2.8.1
- 2.9 Comments

es to MOS such as bipolar and  
 apath for the specific memory  
 d being less so that the device  
 cture requires additional wire  
 layers for interconnects on the  
 DRAMs and SRAMs increase,  
 non. It is not unusual for a  
 divisions, as shown in the die  
 onous SRAM from Cypress

bank. Some new architectures,  
 Each small bank is faster than a



ing multiply divided architecture  
 ctor)

single large bank would have been. Synchronous DRAMs, DRAMs from Rambus, Inc. and Mosys Inc., are a few of the parts that use small independent banks on a chip to gain speed. A block diagram of a two-bank synchronous DRAM is shown in Figure 1.4.

Wider buses are used to improve the bandwidth which is the rate at which data can be moved into or out of the memory. A RAM with a 32-bit wide bus clearly has 32 times the data throughput rate of a RAM with a 1-bit wide bus.

The limitations to bus width are package size, board space for the external traces, and ringing due to high peak currents when the wide outputs swing. All of these limitations are currently being addressed. Miniature packages and small memory modules address the package size and board space issue. New low voltage operating ranges and low swing interfaces are addressing the peak current issue.

Another method of speeding up the memory is to use alternative technologies such as bipolar or GaAs. It is not uncommon in CMOS circuits for elements of the architecture requiring high drive current such as outputs and line drivers to be made of bipolar circuit elements, a technology called BiCMOS.

The disadvantage of BiCMOS is higher standby power consumption, larger silicon area and a more expensive process resulting in higher cost. In systems where speed in the GHz range is required, and where the expense of appropriate cooling techniques is not prohibitive, high speed alternative technologies such as GaAs can be used.

New interface architectures are also being used to improve the bandwidth of the RAMs. The most common is the synchronous interface. Making a memory synchronous puts it under the control of the system clock. "Wait states", during which the processor must wait for the output data from the RAM, can be reduced or eliminated. With a synchronous RAM the input addresses can be latched into the memory, freeing the processor to perform other tasks until the data is available after a known number of cycles. While the intrinsic speed of the RAM does not increase by the addition of a synchronous interface, the effective speed in the system increases since the processor no longer waits for the RAM to operate.

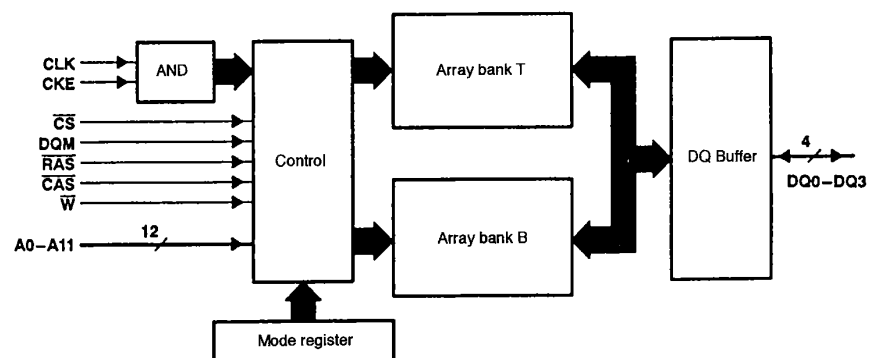


Figure 1.4 Block diagram of a two-bank synchronous DRAM (source: Texas Instruments)

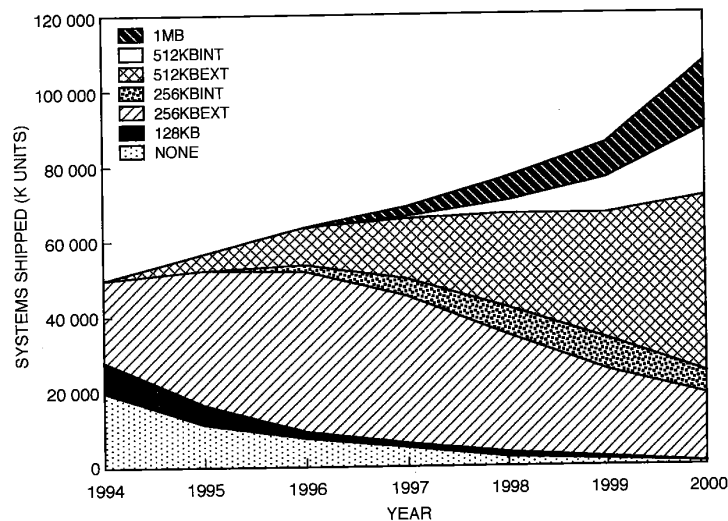


**Figure 2.3** Shipments of systems with different size caches (thousand units) (source: Memory Strategies International)

1996 timeframe and moving to the 512KB and 1MB cache as newer applications require the larger memory.

Fast DRAMs are not expected to replace cache in any but the low end systems in the timeframe of this estimate for several reasons. The SRAMs will continue to be faster than the DRAMs while the speed capability of the processors already exceeds that of the fast SRAMs and continues to increase maintaining a performance advantage for the cache hierarchy architecture.

Since the cache SRAM is a small memory bank that sits close to the processor the speed loss in transmission is minimized. The trend toward smaller packages and multichip modules will also tend to get the cache SRAM even closer to the processor. A small amount of SRAM positioned close to the processor will continue to be faster than a larger amount of DRAM in the main memory bank which can not all be close to the processor.

Not only are delays in the wiring between the DRAM and the processor difficult to avoid, the wiring will also tend to be an unacceptable source of EMI radiation at very high speeds.

### 2.3.3 Synchronous and Asynchronous SRAMs in Fast Caches

As faster SRAMs have been required, the synchronous SRAM has been developed. The addition of input and output registers puts the cache under the control of the system clock for fast burst accesses. The penalty paid is the initial multiple cycle delay, or latency, at the beginning of each burst.

An analysis adapted from Micron Technology on the system benefits of the asynchronous compared to the synchronous burst SRAMs in the same technology is shown in Table 2.2.

The number sequence in the table represents the number of cycles for first through fourth access of the memory. For example a 20 ns asynchronous SRAM has 2-1-1-1 performance on a 33 MHz bus. This means that it takes two cycles to set up the first access and then one cycle for each succeeding access. The wait state in the "burst" part of the cycle is zero. For most system designers the latency in the first cycle is less important than the number of cycles to access the rest of the burst.

We assume a 15 ns asynchronous SRAM in 0.5 micron technology as a baseline. At 40 MHz the 15 ns asynchronous SRAM and the burst SSRAM both provide a two-cycle initial latency and no latency during the burst. At this frequency the asynchronous SRAM would be used.

At 50 MHz the asynchronous SRAM is 3-2-2-2, while at 50 MHz the burst SSRAM in the same technology is 2-1-1-1 with no wait state in the burst. The ability to avoid a wait state means that at 50 MHz most systems would switch to the flowthrough burst SSRAM if the price is competitive. They would not normally switch to the pipelined burst SRAM at 50 MHz since the initial latency is longer than that of the flowthrough part and both have 0 wait states during the burst at this speed.

At 75 MHz the flowthrough burst SSRAM cycle becomes 3-2-2-2, and the pipelined synchronous burst SRAM is still 3-1-1-1 with no wait state in the burst. At 75 MHz therefore many systems would find the pipelined burst SSRAM more attractive.

While some systems might have a problem with the initial three-cycle latency of the pipelined part, the alternative is a double-buffered system or a more expensive cache configuration which is prohibitively costly except in the highest end systems.

The above rationale assumes, of course, that the cost differential between the parts at the trade-off speed is essentially zero. The cost differential in volume, however, of all three parts should be similar since they are assumed to be in the same technology. Package cost might cause some price differential.

The conclusion is that a 15 ns asynchronous SRAM could give 0 wait states to about 40 MHz, while a 9 ns burst SSRAM in the same 0.5 micron technology could give 0 wait states to 66 MHz and a 7 ns pipelined burst SSRAM could give 0 wait states over about 75 MHz.

Similar reasoning leads to the conclusion that fast DRAMS are unlikely to replace SRAM cache in any but the lowest end systems due to the inherent slow speed and resulting higher latency of the DRAM [6]. For example, with a 50 MHz bus, a pipelined burst SSRAM has a burst read speed of 50 MHz with a 3 cycle latency (3-1-1-1) and a Hyperpage Mode (EDO) DRAM has a burst read speed of 25 MHz

with a 6 cycle of late

### 2.3.4 Cache

An analysis the cache si

The mark station, foll by price rai

Processor shown in T

In the 48 speed. For chip and th

The 586 cache SRA support a 6 up time. Th the faster i internal clo

**Table 2.2** Analysis of SRAM types by bus frequency

| System bus (MHz)  | 33      | 50      | 60      | 66      | 75      | 83      | 100     |
|-------------------|---------|---------|---------|---------|---------|---------|---------|
| Async SRAM        | 2-1-1-1 | 3-2-2-2 | 3-2-2-2 | 3-2-2-2 | 3-2-2-2 | 3-2-2-2 | 3-2-2-2 |
| Sync burst SRAM   | 2-1-1-1 | 2-1-1-1 | 2-1-1-1 | 2-1-1-1 | 3-2-2-2 | 3-2-2-2 | 3-2-2-2 |
| Pipelined SB SRAM | 3-1-1-1 | 3-1-1-1 | 3-1-1-1 | 3-1-1-1 | 3-1-1-1 | 3-1-1-1 | 3-1-1-1 |

Source: adapted from Micron Technology [2].

Processor

486DX2

486DX4

586

586

586

586

P6

<sup>a</sup>Some Pentium

Table 2.6 SRAM speed capability

| Processor  | Bus    | TAC (ns) | Type              | Mode        |
|------------|--------|----------|-------------------|-------------|
| 486DX2/DX4 | 33MHz  | 20       | Asynchronous      | Flowthrough |
| 586        | 66MHz  | 9        | Burst synchronous | Flowthrough |
| 586        | 75MHz  | 7        | Burst synchronous | Pipeline    |
| 686        | 100MHz | 4        | Burst synchronous | Pipeline    |

The 512KB cache workstation, that was originally forecast to be the first user of the 586 class processors, was expected to run at 66 MHz on the bus and required a 7ns SRAM in the cache. Parity was also used in this system so the SRAM was required to have  $\times 9$  or  $\times 18$  organization.

A 256Kb SRAM organized  $32K \times 9$  was developed with a design target of 7 ns speed. These SRAMs proved difficult to build with the required speed and were too small for the application. Early  $64K \times 18$  synchronous SRAMs were then built for the high end 512KB workstation cache.

It then appeared that the first users of the Pentium would be PCs not workstations. To improve SRAM supply and reduce the price, a  $32K \times 32$  Burst SSRAM was developed by many suppliers with flowthrough access time of 9–10 ns but that could run in pipelined mode at 6–7 ns in the system. This was followed by 8 ns flowthrough parts that ran at 4–5ns in pipelined mode and a density increase to  $64K \times 32$ . Parity was also dropped since it made the SRAM more expensive and was not needed in a typical PC. Parity was retained for parts used in workstations.

High speed interfaces, such as HSTL, GTL and CTT, have begun to appear on SRAMs. Intel combined the L2 cache and the P6 processor in a single package with a GTL type of interface for the high end systems. Compact packages such as the BGA (ball grid array) were developed to reduce package inductance.

As the industry moves to faster processors a faster speed can be provided by a combination of technology and high speed interfaces. The pipelined speed of 4 ns is required at 100 MHz and can be provided by the 0.35 micron technology SRAMs.

### 2.3.5 SRAM Use Based on Processor Speed

For processors of 50 MHz and above the most cost-effective solution for 586 class systems is the synchronous SRAM, since an asynchronous SRAM must be in a more expensive technology, such as BiCMOS, to have an equivalent speed with an otherwise identical synchronous SRAM. Above 66 MHz a synchronous burst pipelined part is preferred to an equivalent technology synchronous flowthrough part because of the reduced access time of the pipelined burst part at essentially the same cost.

A 0.5 micron technology  $32K \times 32$  synchronous burst pipelined SRAM can provide a 5 ns access time during the burst. The initial latency, however, is 3 cycles rather than the 2 cycles of the flowthrough part. The flowthrough access time of the same part would be about 7 ns.

Table 2.8 Applications segments for fast DRAMs

## Computer:

Main memory (workstation, PC, notebook)  
 Graphics subsystems (frame buffers, video)  
 Add-on modules (SIMMs, DIMMs)  
 Peripherals (printers, fax)

## Consumer:

Mobile systems (notepad, communicator)  
 Video (TV, VCR)  
 Games (multimedia)

## Communications:

Switching systems  
 Set top (ATM, MPEG2)

## Industrial:

Medical (X-ray video, scanners)  
 Embedded processors

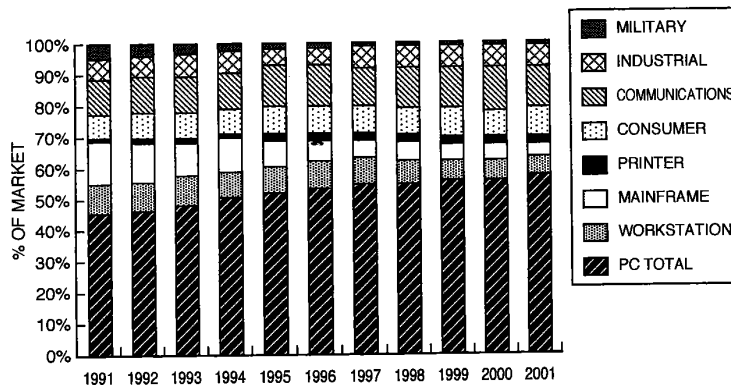


Figure 2.4 DRAM market by major applications segment (%\$) (source: Memory Strategies International)

### 2.5.1 Mainframe and Supercomputer Applications

Mainframe and supercomputer applications fall into two types, those that have single, fast, large main memories and those that are multiple smaller workstations operating in a multiprocessor arrangement. These latter are much more common and will be considered as part of the discussion on workstations.

The older type has a very large fast main memory in the gigabyte density range. These main memories use very deep bit wide organizations of DRAMs.

For example a 160 MHz, 512MB system with 128-bit bus could use two interleaved boards in main memory, each board having 128 16M $\times$ 1 Synchronous DRAMs (SDRAMs) with 80 MHz burst access time. The interleaving provides the speed