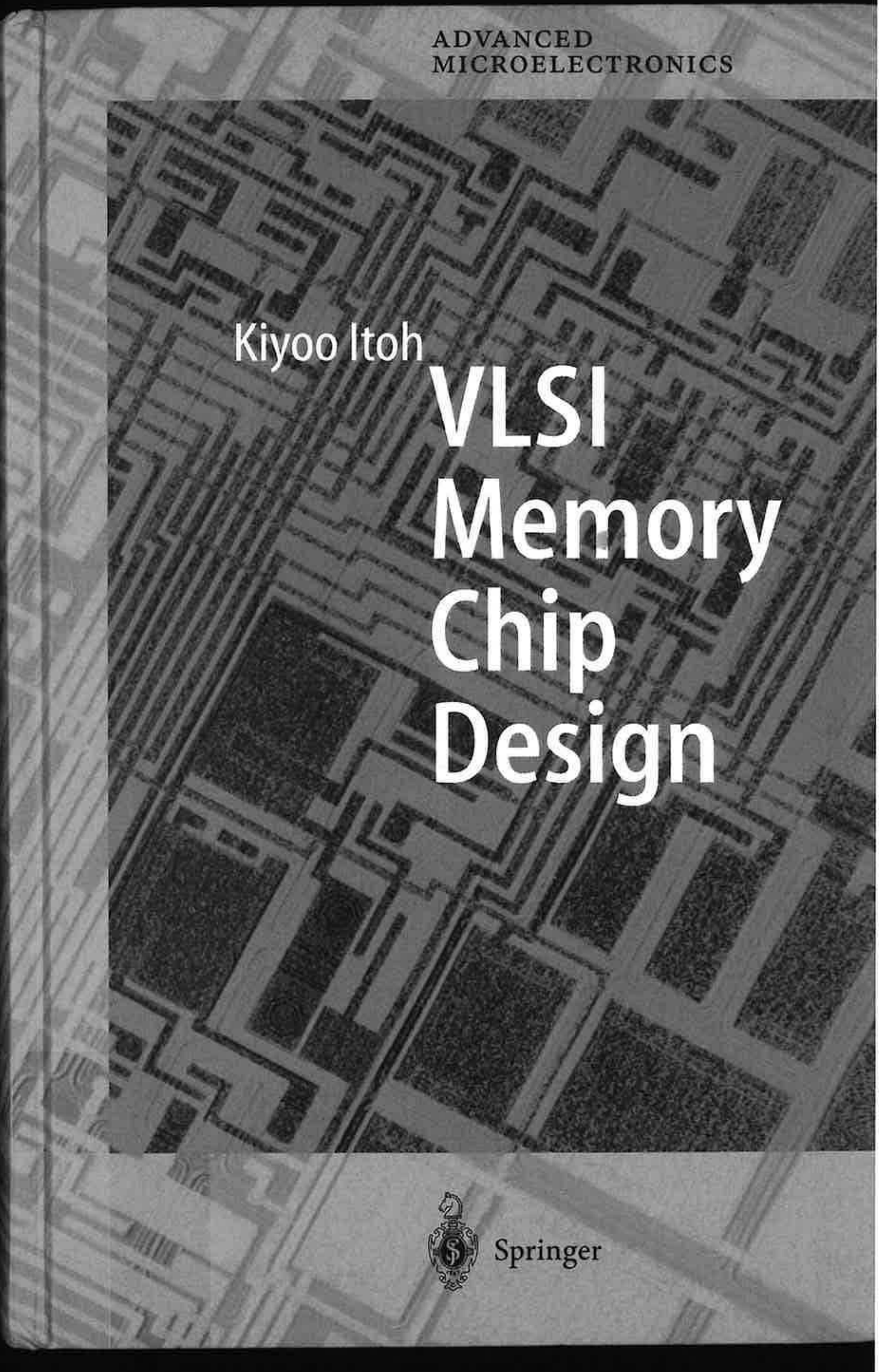


ADVANCED
MICROELECTRONICS

Kiyoo Itoh

VLSI Memory Chip Design

Springer

RONICS

rovides systematic information on
manufacturing of microelectronic
rchers or engineers in their fields,
:h as wafer processing, materials,
/LSI implementation, and subsys-
ı physics and engineering and the
as research scientists.

ss Communication

ooks/ssam/

Kiyoo Itoh

VLSI Memory Chip Design

With 416 Figures and 26 Tables

Springer

Dr. Kiyoo Itoh
Hitachi Ltd., Central Research Laboratory
1-280, Higashi-Koigakubo
Kokubunji-shi
Tokyo185-8601
Japan
e-mail: k-itoh@crl.hitachi.co.jp

Series Editors:

Dr. Kiyoo Itoh
Hitachi Ltd., Central Research Laboratory
1-280 Higashi-Koigakubo
Kokubunji-shi
Tokyo 185-8601
Japan

Professor Takayasu Sakurai
Center for Collaborative Research
University of Tokyo
7-22-1 Roppongi, Minato-ku,
Tokyo 106-8558
Japan

Library of Congress Cataloging-in-Publication Data

Itoh, Kiyoo, 1941-
VLSI memory chip design / Kiyoo Itoh.
p. cm. -- (Springer series in advanced microelectronics ; 5)
Includes bibliographical references and index.
ISBN 3540678204 (alk. paper)
1. Semiconductor storage devices--Design and construction. 2. Integrated circuits--Very large scale integration--Design and construction. I. Title. II. Series.

TK7895.M4 I876 2001
621.39'732--dc21
00-068735

ISSN 1437-0387
ISBN 3-540-67820-4 Springer-Verlag Berlin Heidelberg New York

Springer-Verlag Berlin Heidelberg New York
a member of Springer Science+Business Media

http://www.springer.de

Printed in Germany

Typesetting by the author.
Data conversion and figure processing by LE-TEX Jelonek, Schmidt & Vöckler GbR, 04229 Leipzig.

Cover concept by eStudio Calmar Steinen using a background picture from Photo Studio "SONO". Courtesy of Mr. Yukio Sono, 3-18-4 Uchi-Kanda, Chiyoda-ku, Tokyo.
Cover design: *design & production* GmbH, Heidelberg

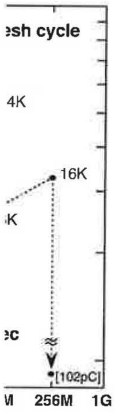
Printed on acid-free paper SPIN: 12187942 57/3180 - 5 4 3

ance, the basic technologies
peripheral circuits are des-
cy are explained. Chapter 4
M which strongly influences
, in the chip. The relation-
ving/sensing is explained in
on-chip voltage generators
tors are essential for power-
apter 6 discusses subsystem-
mportant in providing wide
S. Chapters 7 and 8 describe
sizing the importance of the
f lowering power-supply vol-
old-current reduction which

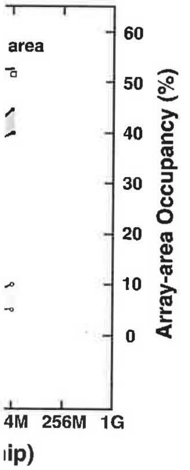
leagues and the office admi-
Ohta, at Hitachi Ltd. They
to finalize my work. Special
inuing support and patience

Kiyoo Itoh

Contents



f data lines [3.3, 3.4]. A C_D of 200 fF



se amplifiers in a DRAM chip [3.4]

y of a word line made of a resistive
r of divisions has been determined
ltant area penalty. In the 64 Kb
short and the speed requirement
e, as shown in Fig. 3.40a [3.9], was
, poly-Si was replaced by a lower-

resistance polycide, in order to double or quadruple the number of memory cells to be driven. To further reduce the *RC* delay, a poly-Si or polycide word line strapped with a low-resistance aluminum line in each string of 64–128 cells [3.17], as shown in Fig. 3.40b, has been widely accepted in commercial chips since the 1 Mb generation. In the 64 Mb generation, even a hybrid division of the two types shown in Fig. 3.40 has been proposed. However, the aluminum-strapped word-line structure suffers from some drawbacks. It becomes difficult to achieve fine patterning of aluminum at a tight pitch of word lines on a hilly surface, while still connecting to the poly-Si line at the bottom, as the memory cell is miniaturized. In addition, even the word-line structure starts to create quite a large *RC* delay as many memory cells are connected to a word or subword line. A multidivided word-line scheme using a hierarchical word-line structure [3.18, 3.19], as shown in Fig. 3.25, solves these problems, and is discussed later in detail.

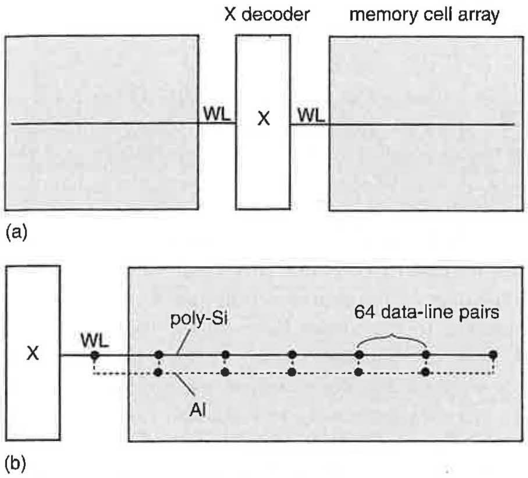


Fig. 3.40. Reductions in word-line delay [3.4, 3.9, 3.17]. (**a**) 64 Kb–256 Kb. 64 Kb: poly-Si (30 $\Omega/\square$), 64 data-line pairs per WL. 256 Kb: polycide (1–5 $\Omega/\square$), 256 data-line pairs per WL. (**b**) 1 Mb and beyond: poly-Si (50 $\Omega/\square$), Al (0.1 $\Omega/\square$)

3.6 Read and Relevant Circuits

3.6.1 The Address Buffer

In the NMOS era of the 16–256 Kb generations, a differential address buffer using an on-chip reference voltage [3.14] was widely used. Since the 1 Mb

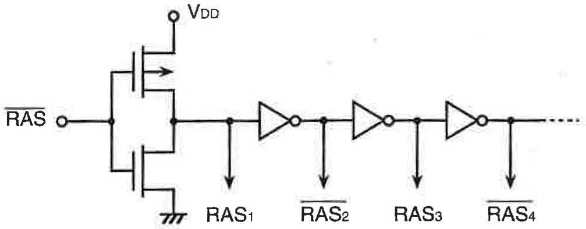
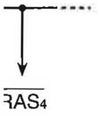


Fig. 3.41. The $\overline{\mathrm{RAS}}$ clock buffer [3.21]

generation, various simple CMOS address buffers have been proposed. In principle, the same circuit configuration is applicable to both the row and column address buffers. Recently, however, a high-speed column address buffer has been especially important to enhance the data throughput by shortening the column address access time t_{AA}, which is dominated by the address buffer (about 40% of t_{AA} [3.20]).

Figure 3.41 shows a typical $\overline{\mathrm{RAS}}$-clock buffer [3.21] to control the row address buffers. To discriminate between the TTL logic levels of over 2.4 V or below 0.8 V for an input $\overline{\mathrm{RAS}}$ signal, the logical threshold voltage is adjusted to be about 1.6 V by tuning the channel-width ratios of the NMOS and the PMOS at the first input stage. A multistage CMOS circuit controls the row internal circuits, using pulses with differing polarities and delays.

Figure 3.42 shows an address buffer that features a cross-coupled differential amplifier [3.21]. It enables an almost constant speed due to a differential circuit configuration, independently of the power-supply noise. Address buffers consisting of inverters, similar to the above $\overline{\mathrm{RAS}}$ buffer, have also been widely accepted. In general, however, the operation of inverter-type buffers is susceptible to power-supply noise [3.20]. For example, as soon as the input logic signals to many buffers are simultaneously switched to the other logic state so that the ground (V_{SS}) line voltage is instantaneously raised and maximized, the speed of each buffer is degraded, with a reduced NMOS gate–source voltage. For a 64 Mb design [3.20], a simultaneous switching of 13 address buffers caused a V_{SS} noise of 0.4 V and a speed difference of 2.3 ns between the inverter and cross-coupled types, although the difference depended on the quality of the V_{SS} layout. In the figure, the input voltage of address Ai is compared with a reference V_{REF} (1.6 V) to discriminate between a high logic level (H) and a low logic level. The resultant differential signal developed between N_1 and N_2 is quickly amplified to a full V_{DD} by the cross-coupled amplifier, as a result of the application of $\overline{\mathrm{RAS}}_2$. After that, the complementary addresses $\mathrm{a_i}$ and $\overline{\mathrm{a_i}}$ are generated by the application of $\overline{\mathrm{RAS}}_4$ to address latches (ALCs). Note that during standby periods (i.e. $\overline{\mathrm{RAS}}$: H) both $\mathrm{a_i}$ and $\overline{\mathrm{a_i}}$ are kept low and there is no current path in the buffer.



ffers have been proposed. In
icable to both the row and co-
-speed column address buffer
ata throughput by shortening
minated by the address buffer

ffer [3.21] to control the row
TL logic levels of over 2.4 V or
l threshold voltage is adjusted
ratios of the NMOS and the
MOS circuit controls the row
larities and delays.
tures a cross-coupled differen-
ant speed due to a differential
er-supply noise. Address buf-
e $\overline{RAS}$ buffer, have also been
cation of inverter-type buffers
example, as soon as the input
ly switched to the other logic
is instantaneously raised and
ded, with a reduced NMOS
0], a simultaneous switching
4 V and a speed difference of
types, although the difference
he figure, the input voltage of
1.6 V) to discriminate between
e resultant differential signal
ied to a full V_{DD} by the cross-
ion of $\overline{RAS}_2$. After that, the
ed by the application of $\overline{RAS}_4$
tandby periods (i.e. $\overline{RAS}$: H)
rent path in the buffer.

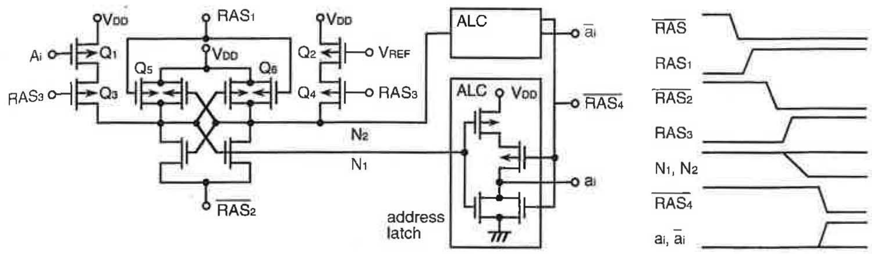


Fig. 3.42. The cross-coupled address buffer [3.21]

Figure 3.43 shows a typical address transition detector (ATD) [3.22], although the variations are shown in Fig. 7.18. Exclusive OR of a_0' and the address delayed by τ generates a short pulse every a_0' transition. All the short pulses generated from all the address input transitions are summed up to one ATD pulse, $\overline{EQ}$. Thus, an ATD pulse is generated at any address transition so as to control internal circuits instead of external clocks. If any address transition is quickly detected by ATD and the resultant ATD signal precharges the I/O line in advance, a data line will be selected using addresses just after the transition, so that a data on the data line is outputted on the I/O line without waiting for the I/O precharging. Thus, a long I/O precharging time can be concealed. The ATD signal reduces the power dissipation of main amplifier by cutting the dc current during periods when it is not needed.

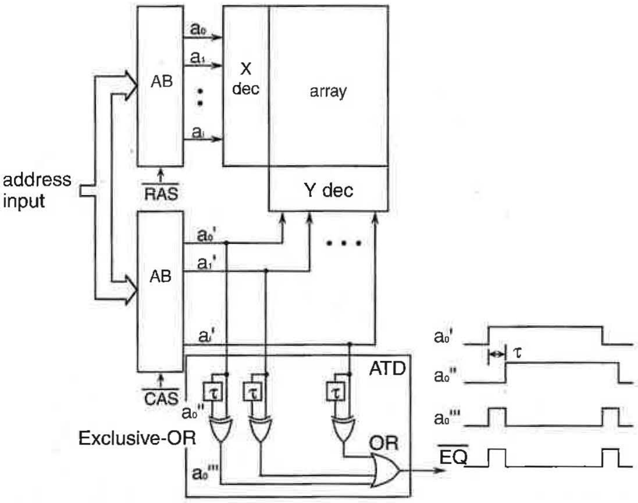


Fig. 3.43. The address transition detector (ATD) scheme [3.22]. AB, address buffers

3.6.2 The Address Decoder

Major concerns for the address-decoder block are power dissipation, speed, and area, because the block includes a huge number of circuits and occupies quite a large segment of the chip.

There are two kinds of decoder; row decoders and column decoders. In DRAM design, unlike SRAM designs, the circuit configurations of the two are totally different. Each row decoder must be a dynamic circuit, while each column decoder can be a static circuit, as explained previously. Note that to precharge all the row decoders without any dc current path, all of the complementary addresses are fixed at a low level during a precharge period, as shown in Fig. 3.42.

Figure 3.44 shows dynamic and static decoders, exemplified by two-bit address signals. There are two kinds of dynamic decoder in DRAM applications; NOR and NAND decoders. First, all of the output nodes (X_0–X_3) are precharged to V_{DD} by transistors Q_P, while keeping all addresses low. Then, according to the succeeding valid address signals, the output nodes are discharged or kept high. Obviously, in NOR decoders all output nodes except for a selected one are discharged, while in NAND decoders all output nodes except for a selected one are kept high. Thus, NOR decoders suffer from a drawback of the large charging and discharging power. The power increases with memory capacity, because an increased number of the nodes – for example, a few thousands, for multimegabit DRAMs – is involved. On the other hand, in NAND decoders only one node is discharged or charged up, independently of memory capacity. NAND decoders, however, suffer the drawback of a slower speed, because the node is discharged by serially connected (i.e. stacked) transistors. The number of stacked transistors is also limited by the body effect of the transistor. Static NAND decoders for the column are simple, as shown in Fig. 3.44c.

Figure 3.45 shows applications of dynamic decoders to the row and static decoders to the column. In dynamic decoders, a word line WL is activated by an RX pulse that is applied after the decoder output Xi has been settled. In the selected decoder, the Q_W gate–drain (i.e. RX terminal) capacitance C_{GD} is large, because the Q_W gate stays at the high level of $V_{DD} - V_T$. Thus, an RX pulse positively going from 0 V to V_{DD} can boost the Q_W gate voltage. The boost ratio is large, because a diode Q_D isolates the Q_W gate from the node Xi capacitance. Due to the resulting boosted gate voltage, to higher than $V_{DD} + V_T$, the word line is quickly driven to V_{DD}. In the non-selected decoders, an RX pulse application never raises the Q_W gate voltages, since the gate voltages are 0 V and thus their C_{GD} values are almost zero. For NOR decoders, even the heavily capacitive Q_W gate is quickly discharged by at least one transistor of the decoder. For NAND decoders, however, to accomplish rapid decoding, Q_W is driven with the help of a small CMOS inverter, whose input capacitance is small enough to be quickly driven even by stacked transistors. Despite the area penalty, the inverter added to each

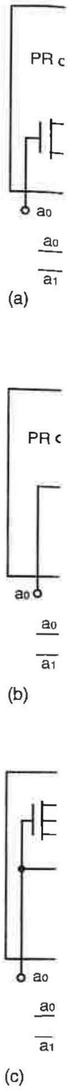


Fig.
(a) D

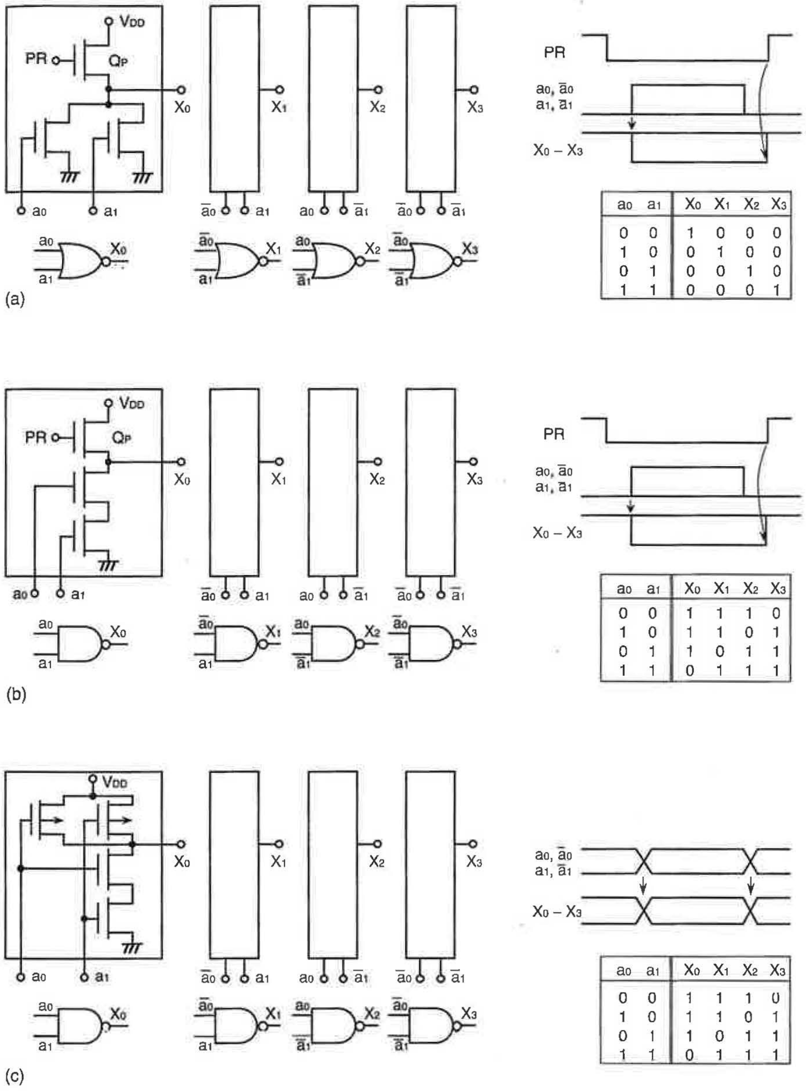
deco
reas
ders

ower dissipation, speed,
of circuits and occupies

nd column decoders. In
onfigurations of the two
namic circuit, while each
d previously. Note that
current path, all of the
ring a precharge period,

, exemplified by two-bit
ecoder in DRAM appli-
e output nodes (X_0–X_3)
eping all addresses low.
gnals, the output nodes
coders all output nodes
ND decoders all output
s, NOR decoders suffer
rging power. The power
sed number of the nodes
RAMs – is involved. On
is discharged or charged
ders, however, suffer the
scharged by serially con-
acked transistors is also
NAND decoders for the

lers to the row and static
d line WL is activated by
t Xi has been settled. In
rminal) capacitance C_{GD}
l of $V_{DD} - V_T$. Thus, an
st the Q_W gate voltage.
es the Q_W gate from the
l gate voltage, to higher
V_{DD}. In the non-selected
Q_W gate voltages, since
ues are almost zero. For
ate is quickly discharged
D decoders, however, to
help of a small CMOS
o be quickly driven even
e inverter added to each



a0	a1	X0	X1	X2	X3
0	0	1	0	0	0
1	0	0	1	0	0
0	1	0	0	1	0
1	1	0	0	0	1

a0	a1	X0	X1	X2	X3
0	0	1	1	1	0
1	0	1	1	0	1
0	1	1	0	1	1
1	1	0	1	1	1

a0	a1	X0	X1	X2	X3
0	0	1	1	1	0
1	0	1	1	0	1
0	1	1	0	1	1
1	1	0	1	1	1

Fig. 3.44. Decoders and operations, exemplified by two address bits [3.4]. (**a**) Dynamic NOR; (**b**) dynamic NAND; (**c**) static NAND

decoder never increases the decoder power because it is a CMOS inverter. The reason why NMOS NOR decoders have been replaced by CMOS NAND decoders since the 1 Mb generation is just to reduce the ever-increasing decoder

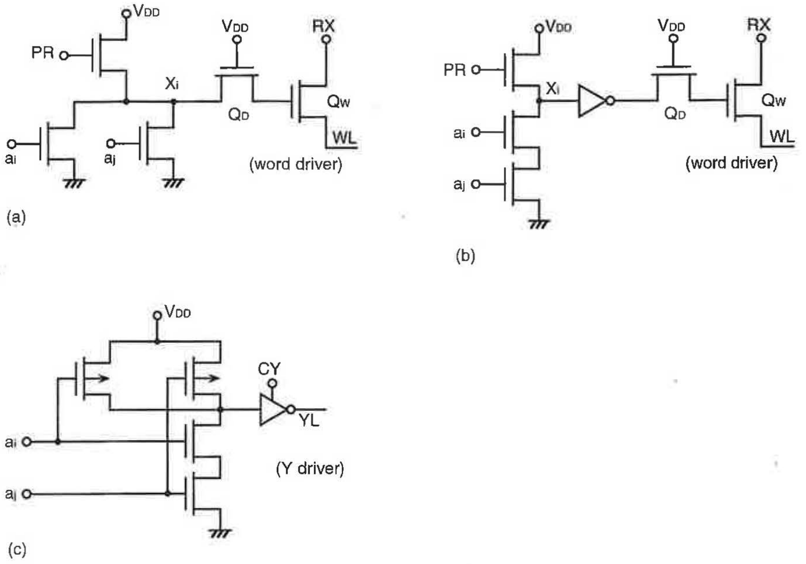
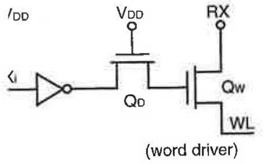


Fig. 3.45. Applications of decoders to word and column drivers [3.4]. (**a**) Dynamic NOR; (**b**) dynamic NAND; (**c**) static NAND

power. Even for a small memory capacity chip of 1 Mb, CMOS NAND decoders have reduced the decoder power down to 4% of that needed for NMOS NOR decoders [3.7, 3.23] as shown in Fig. 3.17. However, to improve the performance of CMOS NAND decoders further, it is essential to reduce the number of stacked transistors. This is realized by predecoding schemes [3.21], as follows.

A predecoding scheme achieves a faster decoding and area reduction of a decoder while reducing the number of stacked transistor in a CMOS NAND decoder. In addition, it reduces the input capacitance and the necessary address input lines of the decoder. Figure 3.46 compares predecoding schemes [3.4]. Direct decoding, 2 bit predecoding, and 3 bit predecoding are shown in Figs. 3.46a–c, respectively. A circle in the figures denotes a transistor connection. For example, when a high level is applied to a_0 and a_1 in Fig. 3.46a, decoders X_0, X_4, and X_8 are selected as a result of NAND decoding. Each 2 bit predecoder can select one of four address input lines coming to the decoders by using two sets of complementary addresses from two address buffers, while each 3 bit predecoder can select one of eight address input lines by using three sets of complementary addresses. Here, let us cite an example of a total external address bits of 6 bits (A_0–A_5). The numbers of address lines to the decoders for direct decoding, 2 bit decoding, and 3 bit decoding are 12, 12, and 16, respectively. The numbers of transistors



ı drivers [3.4]. (a) Dynamic

Mb, CMOS NAND deco-
ıf that needed for NMOS
Iowever, to improve the
is essential to reduce the
edecoding schemes [3.21],

ding and area reduction
d transistor in a CMOS
ıpacitance and the neces-
46 compares predecoding
ınd 3 bit predecoding are
ıe figures denotes a tran-
is applied to a_0 and a_1
ed as a result of NAND
four address input lines
lementary addresses from
select one of eight address
ddresses. Here, let us cite
ts (A_0–A_5). The numbers
ing, 2 bit decoding, and
ıe numbers of transistors

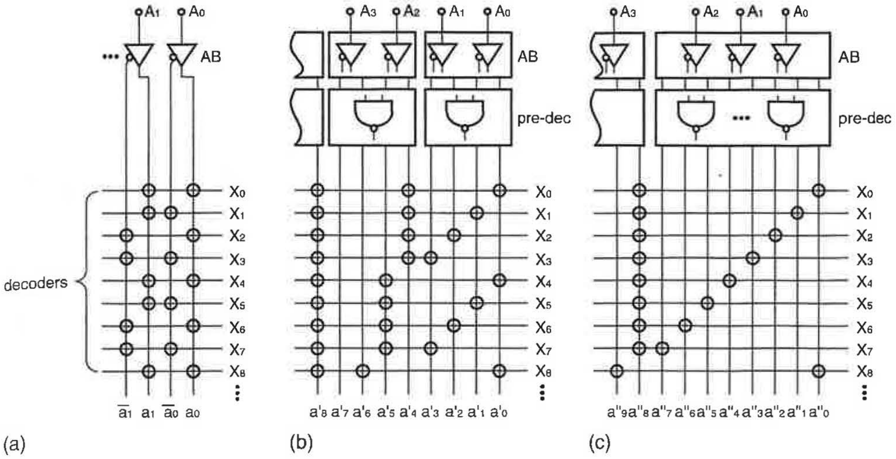


Fig. 3.46. Predecoding [3.4]. (**a**) no predecoding; (**b**) 2-bit predecoding; (**c**) 3-bit predecoding

connected to each address line are 32, 16, and 8, and the numbers of transistors consisting of each decoder are six, three, and two, in the same order. Hence, the predecoding schemes achieve a higher speed with reduction of the address-line capacitance and the number of stacked transistors needed for each NAND decoder, given an acceptable number of address lines. They also reduce the decoder area. The resulting improvement in speed offsets an additional delay caused by the predecoders. Here, predecoding schemes with more than 4 bits are not practical because of a rapid increase in the number of address lines.

Figure 3.47 shows a reduction in the delay of an address line [3.24], which is an example of the buffer insertions shown in Fig. 3.20. Quite a long delay time is developed, despite the aluminum address line, because the line becomes resistive and capacitive due to fine patterning, and is loaded by the distributed gate capacitances of the decoder transistors, as shown in Fig. 3.47a. However, the delay is reduced by the multidivided decoder shown in Fig. 3.47b. The resulting block decoder is driven by a buffer. Each block is constructed so as to correspond to a subarray and only one block is selectively activated by the subarray activation pulse Φ_i.

3.6.3 The Word Driver

A word driver needs to be designed carefully – more so than a column driver – because its load has the following unique features:

1. *A Boosted Word Voltage.* The need for a boosted word voltage, for full write and read operations, means that row decoders and word drivers

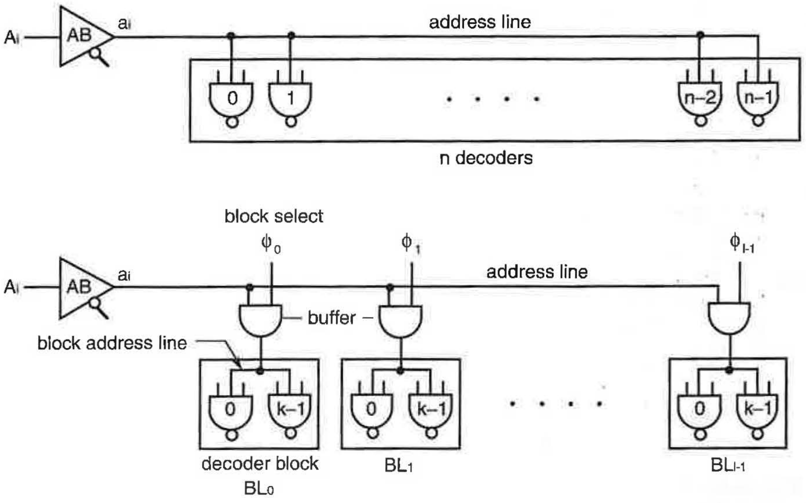
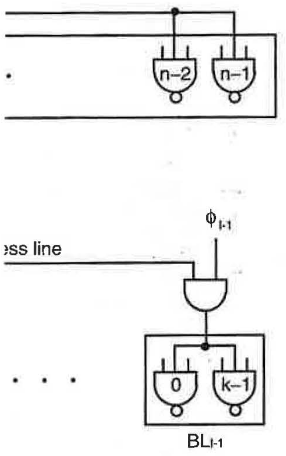


Fig. 3.47. The delay reduction of an address line running on decoders (*upper*) by insertion of buffers (*lower*) [3.4, 3.24]

have complicated designs. On the other hand, column-relevant circuits, such as column decoders and drivers, do not need any boosting. Even without boosting, an amplified signal voltage from a data line can be transmitted to the I/O line, and a data-input voltage of V_{DD} can be fully transmitted from the I/O line to the data line with the help of the CMOS sense amplifier.

2. *A Large Word-Line Capacitance and Resistance.* The electrical characteristics of a word line differ from those of column line YL in the shared Y decoder scheme. Word-line capacitance is quite heavy, because of connections with many memory cells. On the other hand, column-line capacitance is light, because a small number of transistors, equal to the number of data-line divisions (i.e. q in Fig. 3.14), are connected to a YL line. Thus, a larger word driver is needed. As for line resistance, there is also a large difference between the two. A word line made of poly-Si or polycide is resistive for a folded data-line cell, while a column line is made of aluminum.
3. *Loss of Stored Information.* If the stored charges of a non-selected cell are allowed to escape to a data line through the transfer transistor, the refresh time or S/N ratio of the cell is degraded, as discussed in Chap. 4. Thus, noise suppression is essential on non-selected word lines. Moreover, to avoid loss of stored information, data-line precharging must be started after completely turning off the word pulse. Such considerations are not



ning on decoders (*upper*) by

, column-relevant circuits,
need any boosting. Even
from a data line can be
oltage of V_{DD} can be fully
vith the help of the CMOS

ce. The electrical charac-
lumn line YL in the sha-
is quite heavy, because of
e other hand, column-line
f transistors, equal to the
.4), are connected to a YL
for line resistance, there is
rd line made of poly-Si or
vhile a column line is made

rges of a non-selected cell
he transfer transistor, the
d, as discussed in Chap. 4.
cted word lines. Moreover,
recharging must be started
uch considerations are not

needed for the column. Here, an explanation of the column driver is omitted in what follows, because the driver is almost the same as in Fig. 3.45.

A Basic Word Driver. Figure 3.48 shows the basic unit of conventional word drivers [3.13, 3.21, 3.23]. Each word line is divided into two, to reduce word-line delay (see Fig. 3.40), and the resulting divided word line has its own word driver, Q_W. Since a CMOS NAND decoder cannot be placed at a tight word-line pitch, it is shared with two sets (left and right) of four word drivers, although only the right section is shown in the figure. Address signals X_j and X_k are inputted from 2-bit predecoders to the decoder. Each of the four word drivers selected by the decoder is selectively driven by decoded row select lines RX (RX_0–RX_3), enabling the corresponding word line to be driven. The RX drivers in Fig. 3.48b provide a boost word voltage to one of RXs as a result of two-address bit (a_0, a_1) decoding, as follows. At first, node WDL is precharged almost to V_{DD} during the precharge period (i.e. $\overline{RP}$: H) and all address signals and thus all RX lines are fixed at 0 V. When the addresses have been valid after starting activation with $\overline{RP}$: L, a clock Φ_B generated by the $\overline{RAS}$ buffer is applied, so that only one selected RX line is driven at a high enough voltage, boosted by C_B.

The latch transistor Q_L in the word driver suppresses noise on each non-selected word line while fixing the word line at 0 V as follows. During precharge period all of the non-selected word lines are fixed at 0 V because the Q_Ls are turned on. When one set of four word lines is selected by a decoder, the gate voltages of the corresponding four Q_Ls are changed from V_{DD} to 0 V and the Q_Ls are thus cut off. At this moment, the gate voltages of the corresponding four word-driver transistors Q_W are increased from 0 V to $V_{DD} - V_T$. During this process, the word lines are not at any floating voltage that easily couples noise to the lines, because the four word lines WL_0–WL_3 are fixed at the voltages (i.e. 0 V) of RX_0–RX_3 through the respective Q_Ws. After that, for example, RX_0 is selected and a V_{DH} pulse is sent to WL_0, fixing the remaining non-selected word lines at 0 V. Note that all of the word lines belonging to the non-selected decoders continue to be fixed at 0 V, because the Q_Ls are turned on. Thus, the noise coupled to each non-selected word line, even during signal amplifications performed at a large voltage swing of V_{DD} or a half-V_{DD} on data lines, can be sufficiently suppressed. To further reduce noise on the word lines, another scheme of an additional transistor, which is controlled by address signals, on each word line has been proposed [3.25].

High-speed driving of the RX line is also important, because a long delay is developed by the heavy loading of the large Q_Ws and the long RX line running along a memory array. An RX driver placed at the end of the subarrays in Fig. 3.49a increases the line delay. However, a RX driver for each subarray in Fig. 3.49b [3.26], which is similar to the scheme in Fig. 3.47, shortens the delay. In this scheme, only one subarray is selected by the address signals and a subarray selection signal Φ_i. The node WDL of the RX line in Fig. 3.48 corresponds to a line WDL that is common to a number of subarrays in

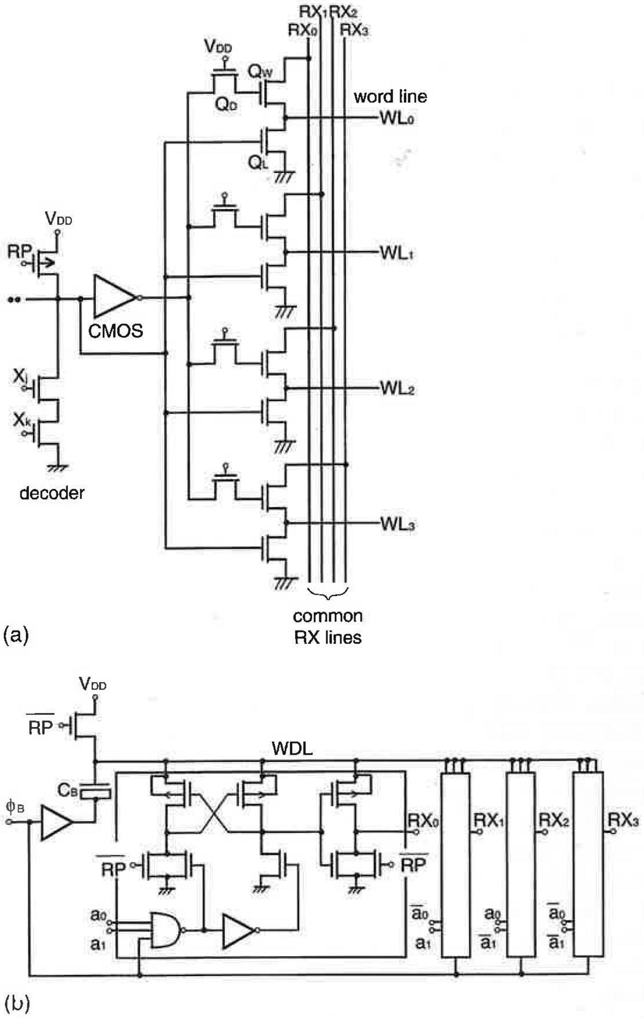
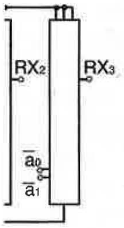


Fig. 3.48. The configuration of word drivers and relevant circuits [3.4, 3.13, 3.21, 3.23]. (**a**) Word drivers; (**b**) RX drivers

Fig. 3.49b. As soon as $\overline{\text{RAS}}$ activation starts, the heavily capacitive WDL line is boosted. After that, the addresses are valid and a subarray is thus selected so that a V_{DH} pulse is applied to a common RX line belonging to the subarray. Consequently, the boosting time of the WDL line can be concealed and the RX line that the RX driver must drive is shortened to one subarray. Thus, the driving speed is improved.



rs and relevant circuits

e heavily capacitive WDL
id and a subarray is thus
n RX line belonging to the
VDL line can be concealed
shortened to one subarray.

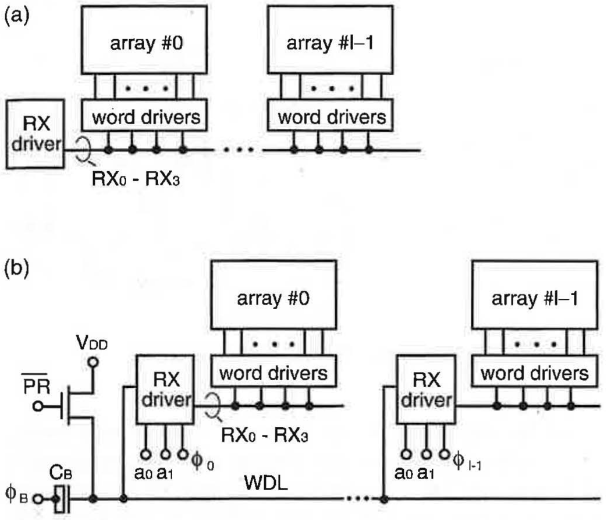
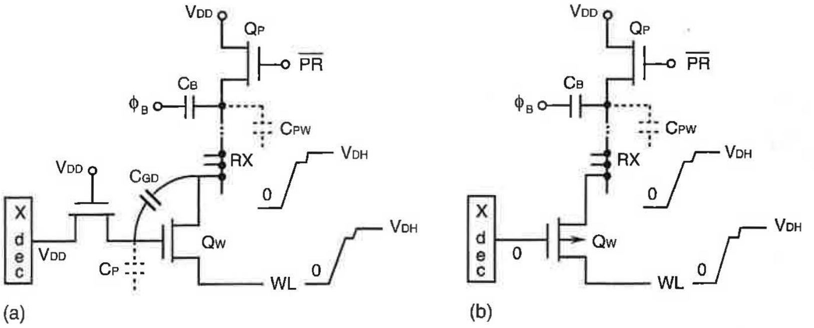


Fig. 3.49. Driving schemes of RX lines [3.4, 3.26]. (**a**) Direct driving of RX lines; (**b**) selective driving of multidivided RX lines

The Voltage-Stress Relaxed Word Driver. This is the word driver that must operate at the highest voltage in a DRAM chip. Therefore, many circuits have been proposed to relax the voltage stress applied at normal and/or burn-in tst operations. They are categorized as the use of PMOS transistor in the word driver instead of NMOS transistor, the changing of the boost ratio according to V_{DD}, and the use of a well-regulated boosted dc voltage.

The PMOS Output Transistor [3.4]. The stress voltages applied to the word-driver output transistor are high, because the word-line voltage must be higher than $V_{DD} + V_{TM}$ (where V_{TM} is the threshold voltage of the memory-cell transistor). However, the PMOS transistor can relax the voltage stress, ensuring high reliability, as follows. For the NMOS transistor in Fig. 3.50a, in order that an increased voltage V_{DH} ($> V_{DD} + V_{TM}$) boosted by C_B at the drain (i.e. RX) is available at the source (i.e. word line), the gate voltage must be boosted to a level that is somewhat higher than $V_{DH} + V_{TD}$ by utilizing C_{GD}. Here, V_{TD} is threshold voltage of the output transistor, which is usually smaller than V_{TM} due to the narrow channel effect, and so on. This dual boosting raises the gate voltage to a considerably high voltage. The PMOS output transistor in Fig. 3.50b does not need dual boosting, since the activation is performed only by lowering the gate voltage from V_{DH} to 0 V. Hence, the PMOS transistor reduces the gate voltage by at least V_{TD}. In actual practice, the difference in gate voltage between the PMOS and NMOS transistors becomes larger, due to variations in the boost ratio at the

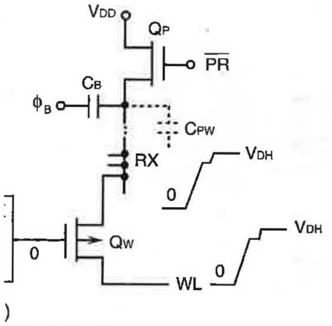


(a) (b)

Fig. 3.50. Word boosting for NMOS (**a**) and PMOS (**b**) drivers [3.4].

NMOS transistor caused by variations in the V_{DD} and fabrication-process. The PMOS transistor may allow the gate–source voltage to be large enough to achieve a faster speed despite the lower conductance of the PMOS, compared to the NMOS transistor, in which the gate–source voltage is usually insufficient. Moreover, for lower-V_{DD} operation, the boost ratio must be larger, since there is a minimal threshold voltage for V_{TD} to prevent a degradation of the V_{DH} level caused by a subthreshold current; this is discussed in Chap 8. Thus, the PMOS transistor has been widely used instead of the NMOS transistor.

The Varied Boost-Ratio Driver. There are some operating modes in which MOS devices must not break down even when the operating voltage varies widely. These are the burn-in test mode, which ensures device reliability by applying an increased voltage, and battery operations, which require a wide voltage margin. Figure 3.51a shows a word driver for boosting a well-regulated low-voltage V_{DL}, which is lowered from V_{DD} ($= 3.3\,\mathrm{V}$) using an on-chip voltage down-converter [3.27]. The node WDL, which corresponds to WDL in Figs. 3.48 and 3.49, is boosted by activating $\overline{\Phi}_{\mathrm{B}}$ to a low level after it has been precharged to V_{DL}. An excessively high stress voltage is applied to devices in the burn-in test mode, since the boost ratio is almost constant. A voltage down-converter dedicated to the test mode relaxes the stress voltage in a high-voltage region as follows. The reference voltage V_{RN} increases when V_{DD} is increased. Eventually, however, it is clamped at a voltage of two MOS-diode drops and thus V_{DL} becomes equal to $r_1 V_{\mathrm{RN}}$. Here, r_1 is a resistance division ratio. Although V_{DL} increases slightly with V_{DD} because of a slight increase in the diode drops caused by the increased diode current, V_{DL} is determined in turn by the other voltage down-converter for the burn-in mode. The converter generates a boosted voltage, monitoring the threshold voltage V_{TM} of the memory-cell transistor as follows. The input voltage V_{RB} of the comparator is given by



IOS (b) drivers [3.4].

V_{DD} and fabrication-process.
ce voltage to be large enough to
tance of the PMOS, compared
urce voltage is usually insuffi-
oost ratio must be larger, since
o prevent a degradation of the
is is discussed in Chap 8. Thus,
tead of the NMOS transistor.

me operating modes in which
n the operating voltage varies
h ensures device reliability by
erations, which require a wide
er for boosting a well-regulated
$(= 3.3\,\mathrm{V})$ using an on-chip
L, which corresponds to WDL
ing $\overline{\Phi}_{\mathrm{B}}$ to a low level after it
igh stress voltage is applied to
oost ratio is almost constant.
mode relaxes the stress voltage
ce voltage V_{RN} increases when
nped at a voltage of two MOS-
$_1V_{\mathrm{RN}}$. Here, r_1 is a resistance
with V_{DD} because of a slight
creased diode current, V_{DL} is
onverter for the burn-in mode.
nitoring the threshold voltage
The input voltage V_{RB} of the

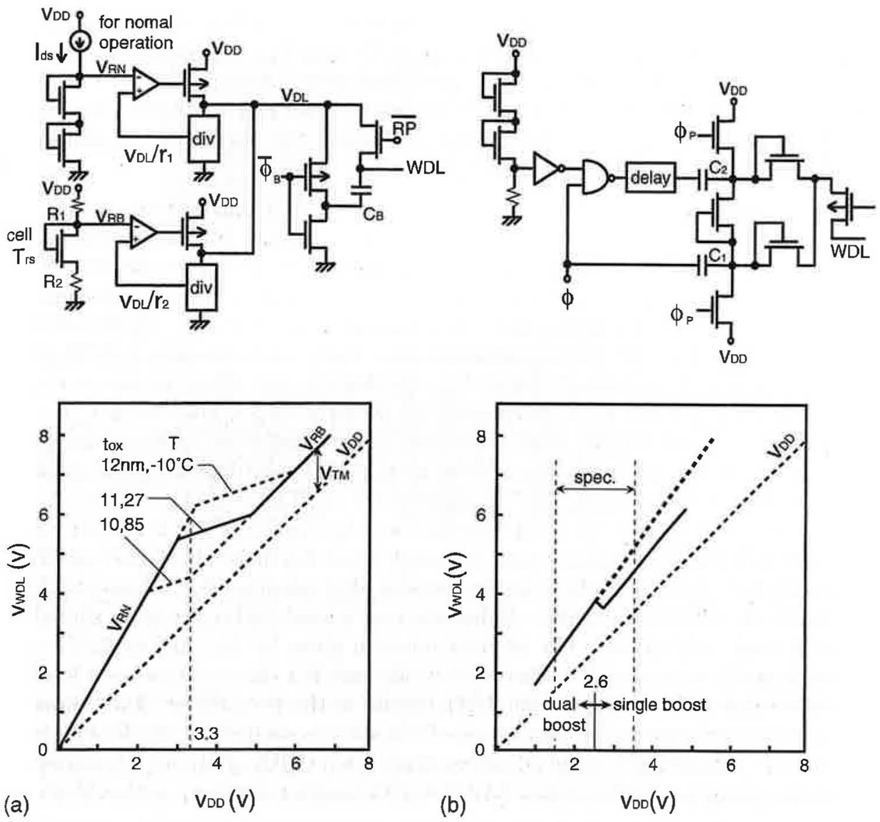


Fig. 3.51. Varied boost ration word drivers [3.4, 3.27, 3.28]. Switching of raised voltages (a) and boost number (b)

$$V_{\mathrm{RB}} = \frac{R_2}{R_1+R_2}\left\{V_{\mathrm{DD}} + \frac{R_1+R_2}{R_2}V_{\mathrm{TM}}\right\},$$

$$\therefore\ V_{\mathrm{WDL}} = BV_{\mathrm{DL}} = B(r_2V_{\mathrm{RB}}),$$

where $V_{\mathrm{DD}} \gg V_{\mathrm{TM}}$, r_2 is a resistance division ratio, and B is a boost ratio. Hence,

$$V_{\mathrm{WDL}} = V_{\mathrm{DD}} + r_2B\,V_{\mathrm{TM}},$$

for $Br_2R_2/(R_1+R_2) = 1$. An additional voltage r_2BV_{TM} is thus V_{TM} for $r_2B = 1$, which minimizes the necessary boosted voltage. The boosted word voltage V_{WDL} can track variations of V_{TM} because a memory-cell transistor is used.

Figure 3.51b shows a voltage up-converter for a battery operation [3.28]. It features dual boosting in the low-V_{DD} region, but single boosting in the high-V_{DD} region. Thus, a raised S/N ratio even in the low-V_{DD} region, and the suppression of an excessive stress voltage in the high-V_{DD} region are obtained. This is essential to ensure the wide operational range of $V_{DD} = 2.6 \pm 1$ V that is inherent in battery operations. A voltage up-converter [3.29] described in Chap. 5 is a variation of this concept, in which the boost ratio is almost continuously changed over a wide range of V_{DD} values.

The Raised dc Supply-Voltage Driver [3.29–3.32]. A quasi-static supply-voltage (V_{DH}) generator, discussed in Chap. 5, would not only minimize the stress voltage to devices, but also enable fast operation, as follows. Figure 3.52a shows a word driver using the dc power supply V_{DH}. One decoder is shared with four word lines through transistors controlled by the decoded signals Φ_{X0}–Φ_{X3}. At the beginning of activation, the non-selected three of the four signals, which were all at V_{DD}, go down to 0 V while the remaining selected one goes to V_{DD}. After that, the decoder output goes down to 0 V and the selected PMOS word transistor is thus turned on, with a change in the gate voltage from V_{DH} to 0 V, so that the word line is activated at V_{DH}. During this process, the remaining three word lines are latched at 0 V. This driver eliminates the need to drive a heavily capacitive RX line and thus eventually realizes a higher speed, although it needs a little additional time to ensure the sequence of the enabling decoder after enabling Φ_{Xi}. Figure 3.52b shows the other word driver. A decoder and a level shifter are both shared with four word drivers. The decoder function given by Φ_{Xi} in Fig. 3.52a is replaced by a decoded RX scheme, in which each RX driver comprises a level shifter and an inverter (see Fig. 3.54), similar to the word driver. The timing requirement between the application of RX and the enabling of the decoder is relaxed, permitting even an advanced application of RX. Without a boosting effect, which is available at the NMOS word transistor combined with a MOS-

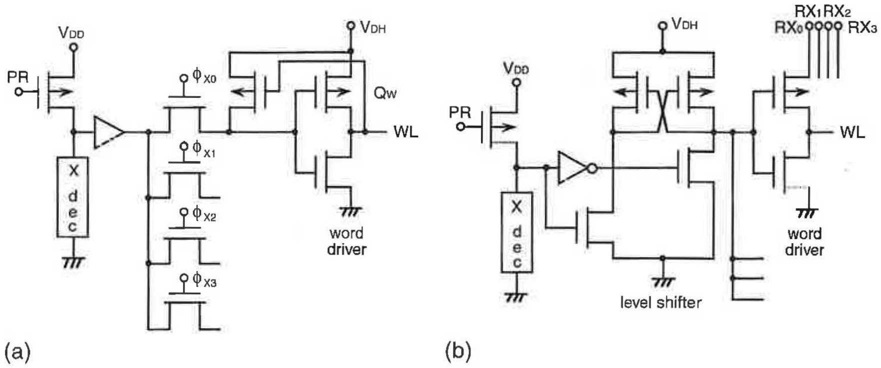
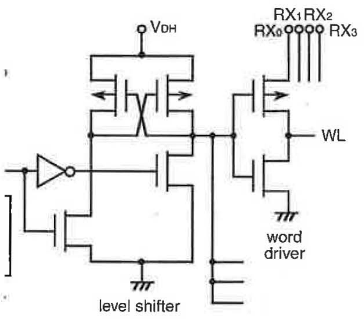


Fig. 3.52. A word driver using a raised dc supply voltage [3.4, 3.29–3.32]. (**a**) Static driver; (**b**) dynamic driver

in the low-V_{DD} region, and the
high-V_{DD} region are obtained.
range of $V_{DD} = 2.6 \pm 1$ V that
-converter [3.29] described in
ich the boost ratio is almost
values.

3.32]. A quasi-static supply-
5, would not only minimize
fast operation, as follows. Fi-
wer supply V_{DH}. One decoder
tors controlled by the decoded
ion, the non-selected three of
vn to 0 V while the remaining
der output goes down to 0 V
us turned on, with a change
the word line is activated at
word lines are latched at 0 V.
ly capacitive RX line and thus
eeds a little additional time to
er enabling Φ_{Xi}. Figure 3.52b
level shifter are both shared
given by Φ_{Xi} in Fig. 3.52a is
h RX driver comprises a level
o the word driver. The timing
the enabling of the decoder is
on of RX. Without a boosting
nsistor combined with a MOS-



oltage [3.4, 3.29–3.32]. (a) Static

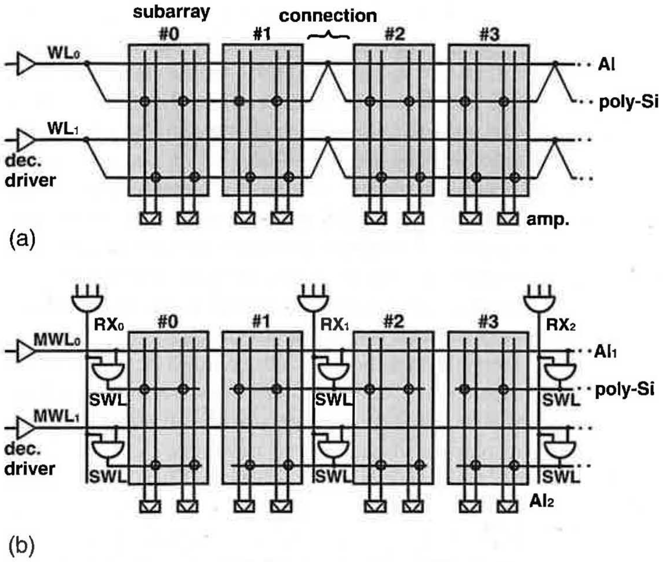


Fig. 3.53. The concept behing partial activation of a multidivided word line [3.18, 3.19, 3.33]. (**a**) No division; (**b**) multidivision

diode shown in Fig. 3.48, an increased word voltage is quickly outputted. This is due to the use of a V_{DH} level shifter.

The Reduction of Word-Line Delay. Even if word-driver speed is improved by using a PMOS transistor, the need to reduce a large word-line delay remains. Although the aluminum-strapped poly-Si or polycide word line, as shown in Fig. 3.53a, has been popular since the 1 Mb generation, a large *RC* word-line delay has been prominent in the 64 Mb generation and beyond. This is because of the ever-finer and longer aluminum line and the ever-increasing number of memory cells connected to a word line. Unfortunately, an increase in the number of word-line divisions causes an area penalty, with a resulting increased number of word drivers.

One solution is the hierarchical word-line structure [3.18, 3.19, 3.33], as conceptually shown in Fig. 3.53b. This is an example of the delay reduction scheme in Fig. 3.19d. One word line is divided into several by the small subword-line (SWL) drivers. All of the subword lines in a row are commonly controlled by a main word line (MWL). Therefore, they are simultaneously activated by selecting a main word line and all of the row select lines (RX_0, RX_1, etc.). The choice of aluminum lines for MWL and RX would improve the speed of the simple aluminum strapping in Fig. 3.53a, despite poly-Si or polycide subword lines. In simple strapping, a word line is heavily loaded with a huge number of memory cells in a row. The resulting heavy capacitance develops a long delay even on a low-resistance aluminum line. In this case the total delay is approximately the sum of a large MWL delay and the subword-line delay. On the other hand, in the hierarchical structure a main word line is loaded by only quite a small number of AND logic gates, which is the same as the number of word-line divisions. Considering that the number of memory cells connected to one subword line ranges from 256 to 512, the loading for the main word line is quite light, which enables an extremely small delay compared with the subword-line delay. An RX line also enables a small delay because of the aluminum material. The delay could be further shortened if it is driven by a RX driver located at each data-line division. Eventually, the total delay of the hierarchical structure is almost confined to the subword-line delay. Thus, the delay is less than that for simple strapping.

Figure 3.54 shows an actual hierarchical structure applied to a 256 Mb chip [3.18, 3.19, 3.33]. It relaxes the MWL pitch to one-quarter so that aluminum wiring is enabled even on the top surface of the substrate, and it allows SWL drivers to be placed alternately to meet the tight word-line layout pitch. Eight-row word lines, each of which is divided into a number of subword lines (SWL_{00}, $\overline{\mathrm{SWL}_{10}}$, ...), are controlled by a set of complimentary main word lines (MWL, $\overline{\mathrm{MWL}}$), which are driven by a row decoder and a set of word drivers. One set of four SWL drivers is located at each end of each subarray. One of four SWL drivers is activated by one of the decoded RX lines from a RX driver. The full use of NMOS transistors realizes a small

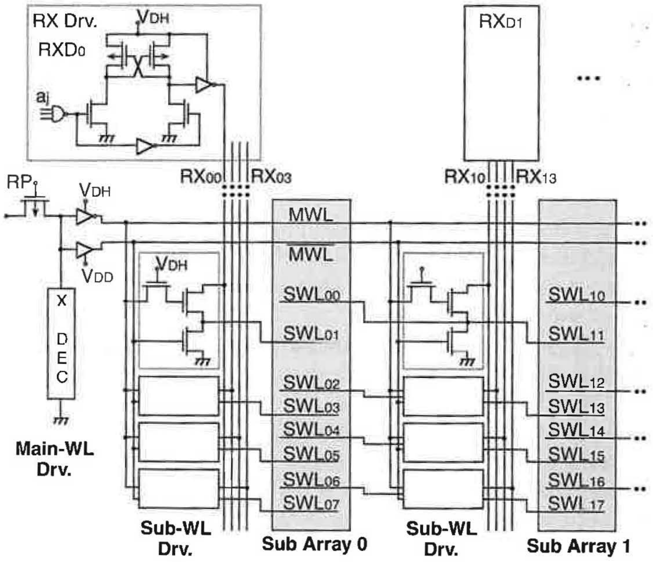


Fig. 3.54. Hierarchical word-line architecture [3.18, 3.19]