

MEMÓRIAS SEMICONDUTORAS

Prof. Wang Jiang Chau (PSI/EPUSP)
v. 2017

0. INTRODUÇÃO

Os sistemas digitais modernos necessitam de uma grande capacidade de armazenagem além de acesso a grandes quantidades de informação em alta velocidade. Discos e fitas são meios pelos quais a informação pode ser armazenada mas são considerados lentos devido à existência de componentes mecânicos associados a eles. As memórias semicondutoras, sejam como subsistemas de um circuito integrado (CI) ou como dispositivos comerciais isolados, são aquelas utilizadas, atualmente, como memórias primárias de qualquer sistema de porte, chegando a capacidade de até vários *Mbytes*. É necessário, então, se manter o custo por *bit* baixo, buscando-se alta densidade, baixo consumo de potência e operação rápida.

Em geral, a capacidade de armazenamento é dada em termos de *bits*, uma vez que um *flip-flop* ou outro tipo de registrador é utilizado para armazenar cada *bit* separadamente. Um *byte* (8 ou 9 *bits*) ou um conjunto de *bytes*, conhecido como palavra, é armazenado em uma determinada posição identificada por um endereço. Uma característica chave de sistemas de memória é que uma única palavra é acessada em um único endereço durante cada ciclo de operação de memória.

As memórias semicondutoras comerciais ou utilizados em circuitos dedicados subdividem-se em dois grandes grupos: as memórias de acesso aleatório, RAMs, para as quais, os dados são lidos ou escritos em qualquer ordem pelo mesmo tempo, e as memórias ROM que são de apenas leitura. Na realidade, as memórias ROM são também de acesso aleatório mas as RAMs foram assim chamadas (e mantem a nomenclatura) para serem diferenciadas das memórias de acesso dependente do tempo (fitas e discos). Para cada um destes grupos existem diversas estratégias de implementação que serão tópicos das próximas seções. Na nossa discussão e nos exemplos que serão apresentados, nos ateremos à tecnologia CMOS que é a mais utilizada atualmente, porém nos referiremos também a outras tecnologias existentes quando necessário.

A organização preferida para a maioria das memórias de grande porte é mostrada na **figura a** em que as unidades básicas de armazenagem, chamadas de células, guardam as informações em forma de *bits*. Podem ser utilizados também em conjunto, como um banco, para armazenar uma palavra ou *byte*. No caso da figura, estamos tratando de uma memória RAM, mas a arquitetura seria a mesma para ROMs exceto pela realização das células. Uma célula é utilizada para leitura ou escrita através da seleção de uma linha pela decodificação

binária da informação de endereço. Por exemplo, o decodificador da figura apresenta 2^N linhas de saída que são acionadas de acordo com o código binário das N entradas.

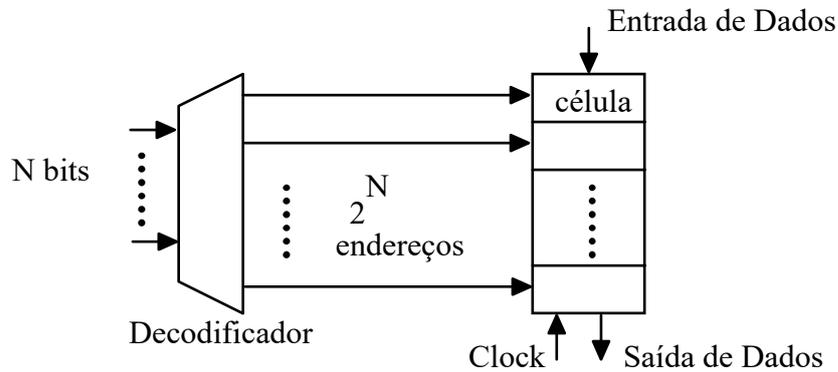


Figura A

1. DECODIFICADORES

Seja para RAMs ou ROMs, os decodificadores são elementos que podem vir a ser determinantes para o tempo de acesso e consumo de potência total da memória. Há diversos esquemas para se implementar tal decodificador sendo uma solução trivial a configuração da **figura b** com um banco de portas lógicas AND formando o decodificador de 8 linhas.

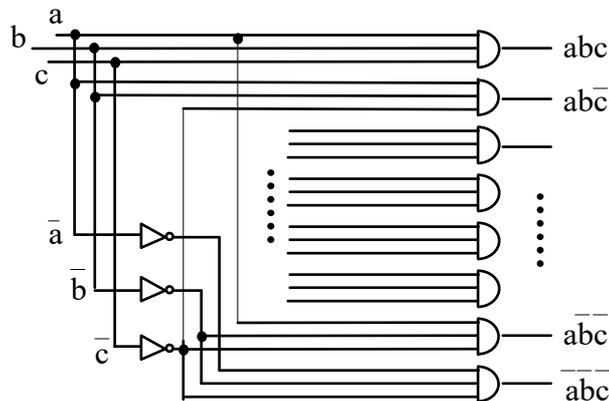


Figura B

A lógica AND pode ser implementada por diferentes estilos de projeto, como pseudo-NMOS, por exemplo (os alunos terão a oportunidade de conhecer diversas técnicas de projeto na disciplina de PSI3452), que permite que o leiaute tenha uma estrutura previsível e regular. Entretanto, o alto *fanin* das portas é um problema grave e pode comprometer a velocidade do decodificador. Outras alternativas são: implementação em forma de portas lógicas regulares CMOS

(de biblioteca de células-padrão), que permitiria um *fanin* menor às custas, porém, de vários níveis lógicos como mostra a figura c(a) ou, então, em uma implementação com a pré-decodificação de um *bit* como mostrada na figura c(b). O segundo AND serve como um segundo estágio e como o comando da linha de palavra. Nestes casos, para efeito de leiaute, algum cuidado deve ser tomado para que exista um casamento entre a altura do circuito correspondente a cada linha do decodificador com cada célula de uma coluna do núcleo da memória.

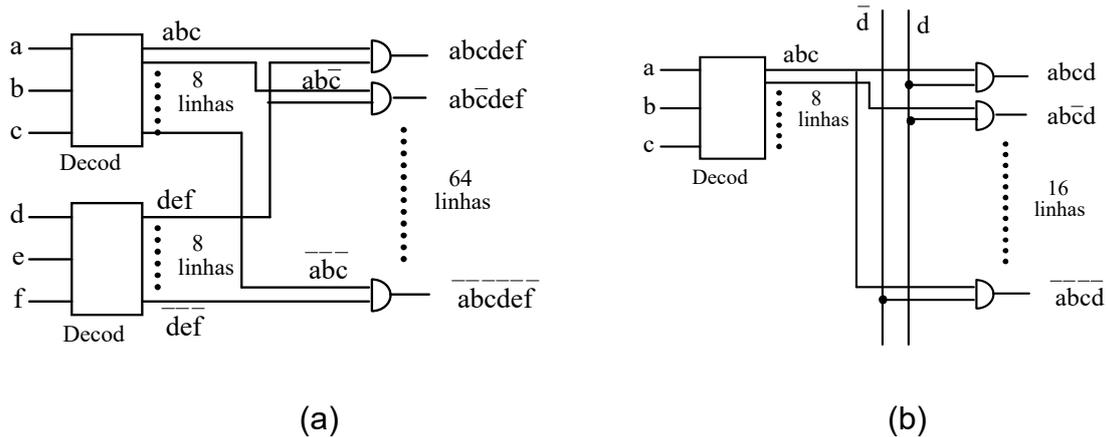


Figura C

Na prática, para se ter um leiaute mais equilibrado, a maioria das memórias de grande porte apresentam uma organização com seleção de célula tanto por linhas como por colunas, baseada em decodificadores para as duas dimensões, como mostra a figura d. Esta configuração é certamente mais equilibrada do que o caso em que a decodificação dos n *bits* de endereço de fileiras (2^n linhas de seleção) e m *bits* de endereço de colunas (2^m linhas) fosse feita com um total de $2^{(n+m)}$ linhas em apenas um dos eixos.

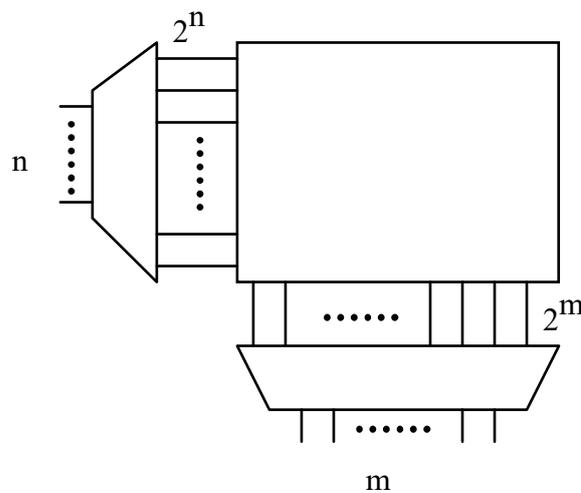


Figura D

2. MEMÓRIAS RAM

As memórias RAM são assim conhecidas (memória de acesso aleatório) porque os endereços da memória podem ser acessados em qualquer ordem e no mesmo tempo para leitura e escrita. A rede de armazenamento, ou o *core*, de uma RAM é formada por células arranjadas em colunas e fileiras para que as conexões sejam compartilhadas na forma mostrada na **figura e**. As linhas horizontais, conhecidas como linhas de palavra (*word line*), são controladas de fora do *core* e as linhas verticais, conhecidas como linhas de *bit* (*bit line*), são os caminhos por onde passam os sinais armazenados. Uma célula é utilizada para leitura ou escrita através da seleção de uma coluna e uma linha.

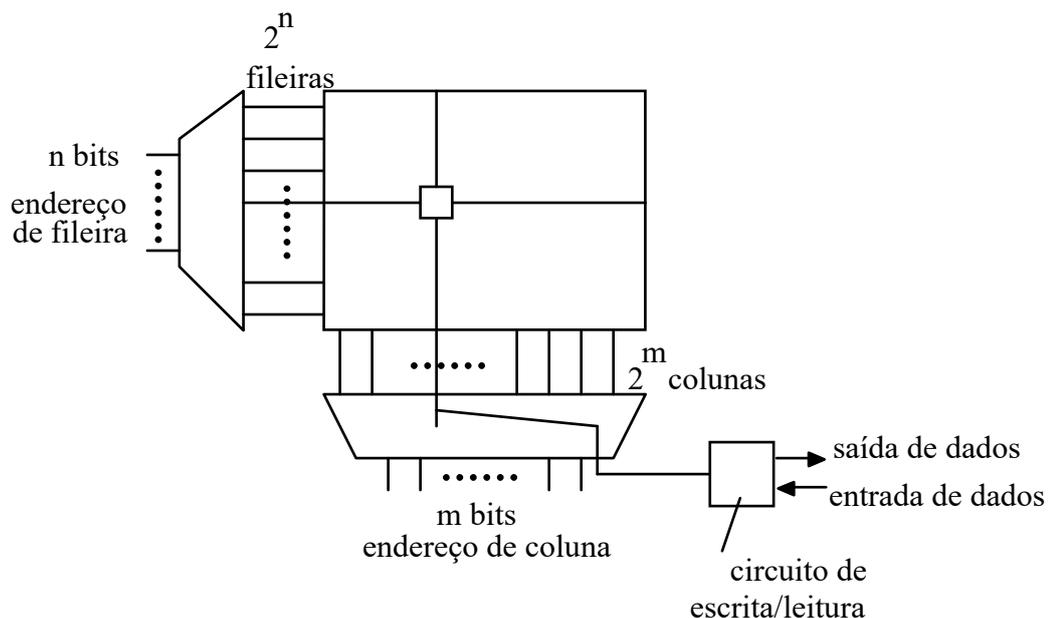


Figura E

Cada célula, em uma versão mais complexa (porém, de construção trivial), poderia ser um registrador tipo D, sendo que memórias com grande capacidade de armazenamento poderiam ser construídas utilizando-se muitos destes registradores. Na prática, porém, esta forma não é econômica: cada *flip-flop* requer um certo número de transistores, conexões para o acesso de dados, além da lógica para a seleção do registrador ao se armazenar ou retirar dados (*write* e *read*), o que torna este tipo de memória excessiva em termos de área ocupada - exigiria um total de 10 a 30 transistores por *bit* para cada célula.

A estratégia básica para a redução de custos na construção de memórias de alta capacidade consiste na redução ao máximo da complexidade das células de armazenagem. Em todas as memórias semicondutoras modernas, as células mantêm a sua função lógica, porém, outras propriedades, como a quantização de amplitudes, a capacidade de regeneração de níveis lógicos, isolamento entrada-saída e *fanout*, são sacrificadas, sendo deslocadas para circuitos periféricos da memória que são compartilhados por todas as outras

células. Este processo de simplificação permite a obtenção de células com até 6 transistores.

Existem dois tipos de memórias RAM, as estáticas (SRAM) e as dinâmicas (DRAMs). O que diferencia um tipo do outro é a forma como a informação é armazenada nas células. Nas memórias estáticas os dados estão guardados por um circuito biestável o que garante a integridade da informação frente a ruídos e oscilações, enquanto que nas RAMs dinâmicas a célula é um circuito passivo e a informação é armazenada em capacitores, tornando-as mais instáveis. Devido às características diferentes de suas células, os circuitos periféricos utilizados pelas DRAMs e SRAMs são também diferentes, o que será mostrado nas próximas seções. Por outro lado, o decodificador é o mesmo para os dois tipos de memória.

As SRAMs devido à sua estabilidade são mais imunes a ruídos e mais rápidas, além de ter uma complexidade menor no que se refere ao controle dos sinais. Porém devido à célula básica conter um biestável, ela ocupa uma área consideravelmente maior. As DRAMs são mais complexas devido à necessidade de esquemas de refrescamento para evitar a perda de dados, decorrente de corrente de fugas, porém ocupa menos área.

RAMs estáticas são usados em aplicações de alta velocidade tais como memórias *cache* e *buffers* de computadores ou na memória principal de supercomputadores. Na prática existem muitas variedades de SRAMs que satisfazem diferentes aplicações e nichos específicos sendo difícil enumerar todas as possíveis realizações. A alta velocidade, porém, é sempre uma importante característica- atualmente, SRAMs de diversas tecnologias com velocidades de poucos ns estão disponíveis comercialmente com uma dissipação de potência que varia de $< 0,1$ W em processos CMOS para 2 a 5 W para tecnologia ECL.

As RAMs dinâmicas, de seu lado, são utilizadas para mercados de alto-volume e baixo custo. As suas grandes vantagens são a alta densidade, o baixo consumo de potência e o baixo custo por *bit* de armazenamento. Elas são consideravelmente maiores que as SRAMs e a dissipação de potência típica de um DRAM é de 100 a 500 mW no seu estado ativo e 0,25 a 10mW no estado de espera.

A cada geração, a capacidade das memórias quadruplicam, seguindo a Lei de Moore. O parâmetro tempo de acesso de memória apresentou uma melhoria exponencial também, porém em um fator de 1,3, ou seja uma taxa mais reduzida do que a evolução da lógica. Enquanto que no começo dos anos 80, os tempos típicos de lógica e de acesso da memória eram próximos, a distância entre os dois aumentou bastante com o correr dos anos, tornando-se o tempo de acesso o gargalo para o aumento do desempenho dos sistemas. Técnicas arquiteturais foram necessárias e a incorporação de memórias cache em sistemas (micro-)processados foi uma decorrência deste desenvolvimento.

Devido à complexidade das RAMs, geralmente, elas não são projetadas dedicadamente como subsistemas específicos de um *chip* (de um sistema).

Normalmente, adota-se RAMs comerciais (ou blocos pré-projetados) pois os seus fabricantes têm uma experiência específica acumulada muito grande o que não se encontra no projetista médio de sistemas. As RAMs só são incluídas em *chips* como subsistemas quando são pequenas ou se deseja, realmente, aumentar o desempenho do sistema como é o exemplo de microprocessadores com memórias *cache* associadas (baseadas em RAMs estáticas).

2.1. RAMS ESTÁTICAS (SRAMS)

Apresentamos nesta seção alguns esquemas com diferentes disposições de transistores para a construção das células do *core* de SRAMs; além disso, mostraremos como o problema de atraso em nós de alta capacitância é equacionado. A **figura f** apresenta o diagrama estrutural simplificado de uma SRAM típica onde pode-se perceber que as linhas de *bit* (de dados) são pré-carregados a um certo potencial e equalizados. O objetivo da pré-carga é promover sempre a descarga ao invés da carga de um nó, pois trata-se de um processo mais rápido, envolvendo transistores tipo n.

O *chip* aguarda uma transição em um de seus endereços, no pino de habilitação de escrita (*WE*, *write enable*) e na seleção de pastilha (*CS*, *chip select*) e quando isto ocorre, um sinal de relógio interno é gerado para controlar a temporização dentro da pastilha. Primeiro, o sinal de pré-carga é desligado e o endereço é fornecido aos *buffers* de endereçamento que comandam os decodificadores de linhas e colunas, conhecidos também como decodificadores X e Y. Os decodificadores irão ativar uma linha de palavra conectando a fileira correspondente a transistores de controle de sinais de dados das células correspondentes; na leitura, as células desta fileira comandam as linhas de *bit*, $\overline{\text{BIT}}$ e BIT , e uma coluna é selecionada. Na escrita, a coluna da célula a ser programada é selecionada primeiro para, então, o sinal ser transferido.

Após a seleção de fileira de células ser feita, ocorre uma variação lenta do potencial nas linhas de *bit* devido à capacitância total da linha- a capacitância da interconexão adicionada à capacitância de difusão dos drenos dos transistores de passagem ligados à linha de *bit* é grande, como mostraremos mais adiante. Aumentar a capacidade de carga e descarga das células não é solução, pois teríamos que aumentar o tamanho dos transistores para isto; portanto, há aumento na ocupação de área, além do aumento de capacitâncias envolvidas, e então do tempo de atraso (eliminando ou reduzindo a pretensa vantagem do aumento da capacidade de carga).

Normalmente, o tamanho das células é determinado pelas vias de contato e pelas linhas de interconexão (alimentação, terra, linha de palavra e de *bit*) e elas são projetadas de um forma otimizada com os dispositivos do menor tamanho possível. Se os transistores na célula de memória forem projetados mais fortes, com o intuito de melhorar o tempo de acesso, haveria também um aumento nas capacitâncias de difusão e da linha de *bit* na mesma medida, com o efeito final de um aumento das dimensões sem a esperada melhoria no desempenho.

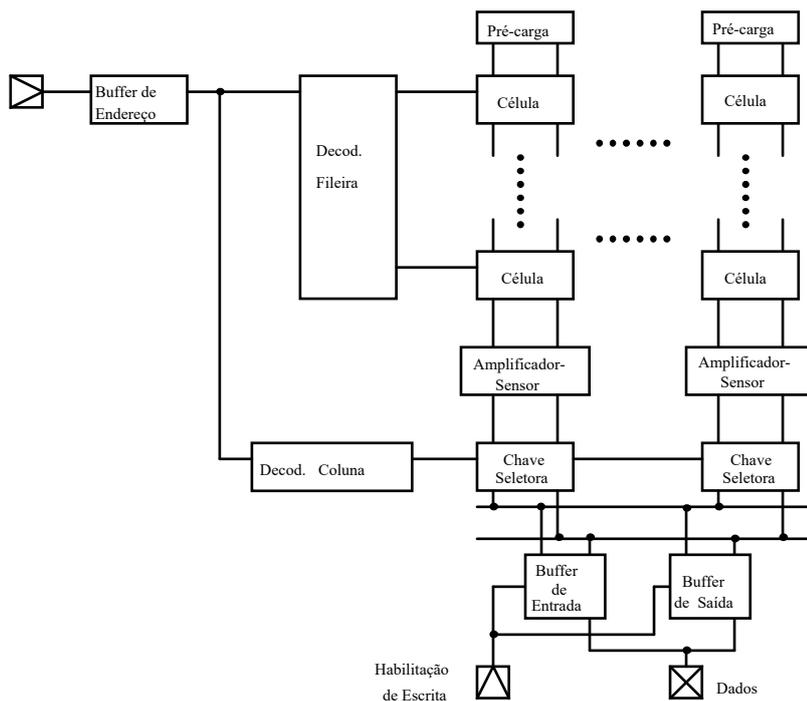


Figura F

Para minimizar os atrasos nas linhas de *bit* de alta capacitância, emprega-se amplificadores-sensores que são capazes de detectar uma variação de potencial à sua entrada a apenas uma fração da escala total em que a fonte de alimentação trabalha. Então, após a célula de memória estabelecer uma pequena diferença de potencial entre as linha de *bit*, o amplificador-sensor é ativado, e a diferença é rapidamente amplificada para uma excursão normal de tensão e fornecida ao *buffer* de saída.

O número de amplificadores depende dos multiplexadores de coluna: se os seletores de coluna forem colocados à saída dos amplificadores-sensores, mais amplificadores são necessários (um amplificador por coluna) aumentando dissipação de potência e área. Se os seletores forem colocados diretamente nas linhas de *bit* e à entrada dos amplificadores-sensores, as entradas dos amplificadores serão mais lentas e o tempo de acesso se degradará por causa do grande diferencial de sinal que é necessário devido aos transistores de passagem. Como compromisso de projeto, alguma multiplexação de coluna deve ser feita tanto antes como depois da amplificação e, como regra geral, a fim de se casar a dimensão do amplificador-sensor com a das linhas de *bit*, o amplificador deve ser o mais simples possível (veja a **figura j**).

2.1.1. CÉLULA DE SRAM

A **figura g(a)** mostra uma célula genérica de uma RAM estática. O circuito consiste de um dois inversores acoplados de forma cruzada conectados às linhas de BIT e $\overline{\text{BIT}}$ através de transistores de passagem. O dispositivo de

carga pode ser tanto um resistor, um transistor n- tipo depleção ou enriquecimento ou, então, um transistor tipo p. O propósito do dispositivo de carga é compensar o efeito de fuga de cargas nos drenos dos transistores de passagem e de "pull-down".

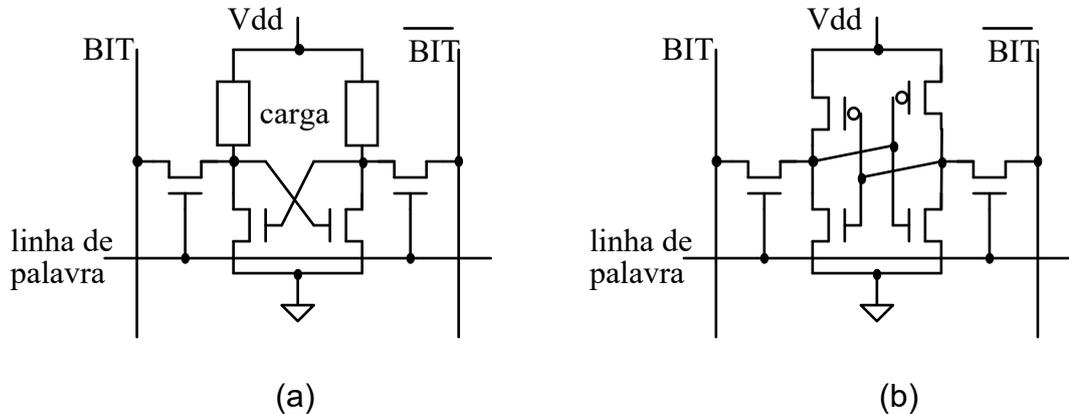


Figura G

Para a escrita de uma célula, o dado é colocado na linha de BIT enquanto o dado invertido é colocado na linha $\overline{\text{BIT}}$, após o qual a linha de palavra é selecionada. A operação de leitura, por sua vez, é iniciada com a pré-carga das linhas de BIT e $\overline{\text{BIT}}$, para daí, com a seleção da fileira pela linha de palavra, ou o sinal de BIT ou de $\overline{\text{BIT}}$ se descarregará por um dos transistores "pull-down".

O esquema típico de célula CMOS para SRAMs é mostrado na **figura g(b)**. A fim de se combinar confiabilidade e dimensões pequenas, algumas condições de contorno de projeto devem ser observadas. Seguiremos o projeto de células com linhas de BIT totalmente pré-carregadas apesar de que existem outras fórmulas com pré-carga parcial para otimização de desempenho ou ausência de pré-carga. O projeto a ser apresentado é seguro, sendo bastante tolerante a variações de processo porém não é o mais rápido que poderia se conseguir. A explicação a seguir é razoavelmente breve mas serve para ilustrar os critérios de projeto.

2.1.1.a. Operação de Leitura

No nosso exemplo, ambas as linhas de BIT e $\overline{\text{BIT}}$ são pré-carregadas em V_{dd} (nível lógico '1'); a adoção de valor de pré-carga em $V_{dd} / 2$ pode trazer vantagens de um maior equilíbrio entre as leituras de '0' e '1', porém é mais complexa. A pré-carga é feita no momento correto (por exemplo, na fase ϕ_1 de um sistema com relógio de dupla fase), enquanto todos os transistores de passagem ligados a todas linhas BIT e $\overline{\text{BIT}}$ estão abertas (não conduzindo). Na fase de leitura, a linha da célula a ser lida é selecionada, ativando os transistores de passagem. Agora, para o exemplo da **figura h(a)**, se um valor '1' estiver armazenado (em X), isto significa que T_1 e T_4 estão conduzindo, portanto

a linha $\overline{\text{BIT}}$ se descarregará para a terra também por T_4 e, assim, o dado armazenado reaparecerá nas linhas de BIT e $\overline{\text{BIT}}$.

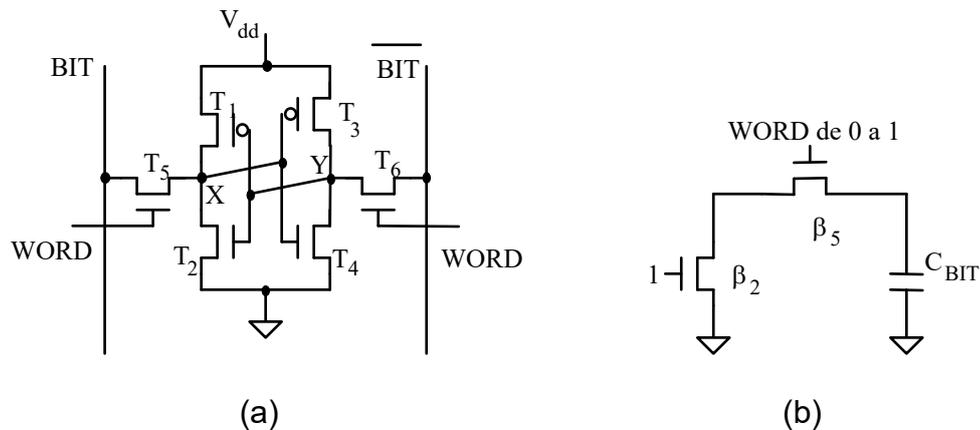


Figura H

O transistor p pode ser de dimensão mínima uma vez que ele precisa apenas sobrepor os efeitos de fuga. Deve-se notar que a relação entre as dimensões dos transistores de passagem T_5 e T_6 e aquelas dos transistores "pull-down" T_2 e T_4 , respectivamente, vai determinar a velocidade de descarga da linha de *bit*. Para analisar o efeito acima, vamos assumir o modelo da **figura h(b)**, onde o caminho entre os transistores T_5 e T_2 está representado. Um transistor de dimensão "unitária" mínima é escolhido como transistor de passagem para a redução da carga da linha de palavra e minimizar o tamanho da célula. Se o transistor de descarga T_2 tiver também a dimensão "unitária" então,

$$\beta_5 = \beta_2 = \beta_n, \text{ onde } \beta_n \text{ corresponde ao transistor de dimensão "unitária".}$$

O tempo de descarga será então, proporcional a τ_F , onde

$$\tau_F = \frac{4 \cdot C_{\text{BIT}}}{\beta_n \cdot V_{\text{dd}}} \cdot (\text{veja que depende da capacitância total na linha BIT})$$

Além das considerações de velocidade (para o caso de escrita, também, como veremos depois) deve-se verificar a relação entre a dimensão do transistor de passagem e o transistor de descarga associado para que, durante a operação de leitura, as linhas pré-carregadas não causem erro de escrita. Tal erro ocorreria se o transistor de passagem fosse, por exemplo, infinitamente grande, fazendo com que os dois nós de armazenamento - X e Y da **figura h(a)** - segurasse o potencial " $V_{\text{dd}} - V_{\text{tn}}$ ", onde V_{tn} é a tensão de limiar do transistor de passagem tipo n. O nó que deveria armazenar o nível '0' seria puxado para o nível lógico '1', uma lógica totalmente errada, apagando-se a informação da célula. Deve-se projetar as dimensões dos dois transistores n em série de tal forma que o nó X ou Y para o nível '0' apresente uma tensão abaixo de V_{tn} por segurança. Por final, outras considerações como o compartilhamento de cargas entre os nós internos e as linhas de *bit* devem ser levadas em conta também.

2.1.1.b. Operação de Escrita

As linhas de *bit* são também pré-carregadas em V_{dd} e quando a seleção de coluna é ativada ou a linha BIT ou $\overline{\text{BIT}}$ é descarregado para o nível lógico '0' dependendo do valor nas linhas de dados (I/O) que agem como um sorvedouro de corrente neste estado. Após a validação da linha de palavra, o valor das linhas de *bit* são escritas na célula.

Considerando a operação de escrita, a relação entre as dimensões dos transistores deve ser calculada de tal forma que o chaveamento possa ocorrer. Para ilustrar o fato, vamos assumir que o nó X da **figura h(a)** tenha um '0' armazenado enquanto o nó Y tenha '1', o que significa que T_2 e T_3 estão em estado de condução. Desejamos chavear X para '1' e Y para '0', ou seja, a linha de BIT apresenta '1' e $\overline{\text{BIT}}$ '0'. Uma vez que Y deve ser puxado para '0', a relação entre T_3 e T_6 deve ser de tal forma que T_1 possa ser ativada (aproximadamente $0,7.V_{dd}$ à sua porta, maior que a tensão de transição do transistor tipo p). Por outro lado o valor de X será determinado pela relação entre os transistores T_2 e T_5 , e deve ser pequeno o suficiente para ativar T_4 . Esta realimentação positiva garante o chaveamento correto da célula.

2.1.1.c. Variações de células SRAMs

O modelo acima de célula de memória RAM não é a única existente, porém é a mais comum dentro da tecnologia CMOS. Esta, comparada a outras tecnologias apresenta um consumo de potência estática bastante baixo devido apenas a correntes de fuga de junção.

Existem outras configurações como, por exemplo, células de memória com duas portas de acesso que são utilizadas em estruturas de microprocessadores como a esquematizada na figura i(a). Pode-se haver leituras independentes nesta configuração enquanto a escrita é feita para uma mesma célula. Este é um uso particular em uma arquitetura com operações com três endereços ($A \text{ op } B \rightarrow C$). Neste esquema, a célula aciona duas vias, colocando uma carga extra nos transistores.

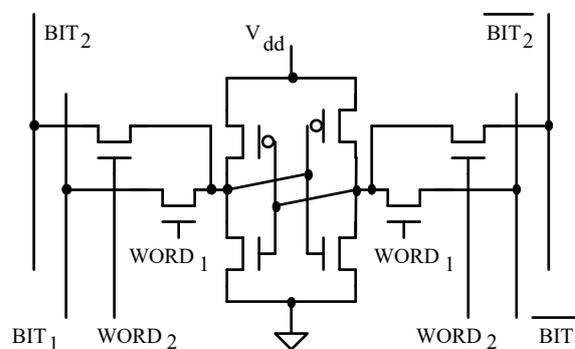


Figura I

2.1.2. AMPLIFICADORES-SENSORES DE SRAM

O tempo de acesso de memórias tende a ser longo devido à alta capacitância nas linhas de *bit* resultante do grande número de dispositivos da rede de memória ligados a elas. Amplificadores-sensores são utilizados para acelerar este processo sendo projetados de forma a sentir pequenas excursões de voltagem nos seus terminais. Normalmente, várias linhas de coluna são ligadas a um mesmo amplificador-sensor como forma de otimizar a área, exceto em casos em que a alta velocidade é prioritária tendo-se, então um amplificador por coluna. As ROMs e DRAMs, por sua vez, utilizam, também, um amplificador para cada coluna, mas isto é devido à configuração de suas redes que exige a ativação de todas as linhas de saída ao mesmo tempo (para o refrescamento, no caso das DRAMs). A saída dos amplificadores-sensores são conectados a *buffers* de saída através de transistores de passagem controlados pela seleção de coluna.

No circuito mostrado na **figura j**, um par de inversores acoplados pela fonte, com uma carga ativa, é utilizado como amplificador-sensor convertendo o sinal de duas para uma via. Os transistores p , T_3 e T_4 , funcionam em uma configuração conhecida espelho de corrente, onde se utiliza o princípio de que a corrente de dois transistores idênticos é a mesma se as respectivas tensões entre a porta e a fonte forem iguais e estiverem operando na região de saturação. O espelho de corrente é um bloco muito comum na tecnologia bipolar uma vez que há a necessidade de casamento entre dois transistores e, nesta tecnologia, tal casamento é oferecido pelo próprio potencial de junção. A tecnologia MOS também é utilizada a despeito de possíveis problemas devido a efeitos de variações de V_t , sendo, então, os transistores T_1 e T_2 construídos de tal forma que as dimensões e as tensões de limiar sejam casadas.

Do ponto de vista lógico, a estrutura acima funciona da seguinte forma: com a pré-carga, ambas as linhas de BIT e $\overline{\text{BIT}}$ carregam o valor lógico '1'. Estando o sinal de habilitação do amplificador-sensor (SAE) ativado, o transistor T_5 funciona como uma fonte de corrente forçando T_1 e T_2 a conduzirem, estabelecendo-se em X um nível intermediário entre o nível '0' e '1'. Se a linha BIT caminhar para o '0' devido à lógica da célula, então, o circuito sai do equilíbrio e X caminhará também para '0' rapidamente. Por outro lado, se BIT tender a '1' o nível lógico de X também tenderá a '1'.

Alguns cuidados de projeto do amplificador-sensor devem ser tomados: normalmente, procura-se ter uma corrente de polarização, gerado pela fonte de corrente o mais alto possível para se garantir que os transistores trabalhem em saturação, região onde o ganho é maior e uma grande corrente de saída possa ser produzida. Tem-se, assim, um chaveamento rápido além de uma maior excursão de tensão à saída. Deve-se atentar, porém, para o fato de que quanto maior as correntes no circuito maior serão a área e a dissipação de potência.

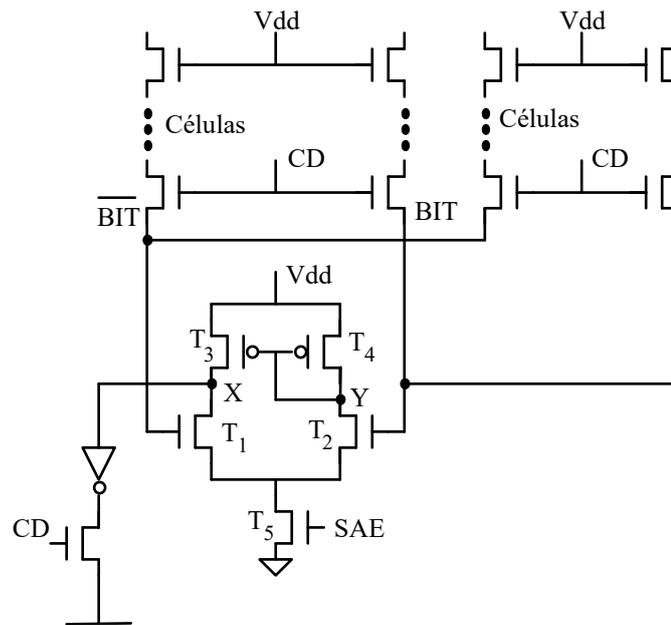
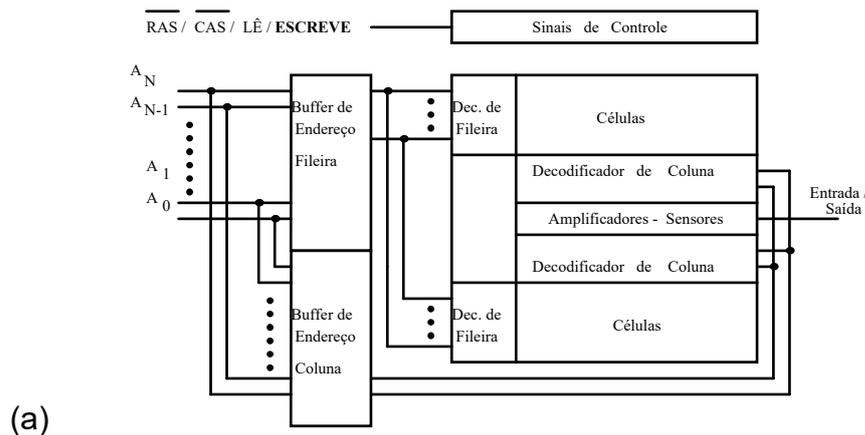


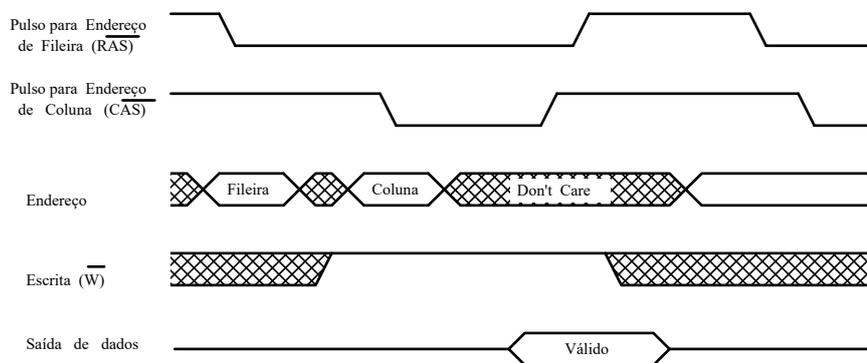
Figura J

Algumas das grandes vantagens do amplificador-sensor da **figura j** são o seu alto ganho em modo diferencial e o baixo ganho em modo comum. Isto significa que, se as entradas de T_1 e T_2 se moverem em sentido oposto, a saída variará enormemente, mas se as entradas de T_1 e T_2 se moverem para valores mais altos ou baixos conjuntamente, a saída não varia. Isto é significativo para células de memória que possuem duas linhas de *bit* correndo uma próxima à outra: qualquer ruído afetará as duas linhas simultaneamente da mesma forma tendo pouca influência na saída. Portanto, o amplificador-sensor baseado em espelho de corrente permite uma alta velocidade na transmissão de dados e alta imunidade a ruídos. De outro lado, ele tem uma pequena amplificação de voltagem e uma excursão de saída limitada uma vez que o ganho não é mantido após os transistores saírem da região de saturação. A excursão mínima de potencial do amplificador é limitada pelo descasamento de tensão de limiar entre T_1 e T_2 uma vez que o par diferencial amplifica a diferença de potencial de limiar entre os dois transistores como se fosse uma diferença entre os seus terminais.

Existem diversas variações na configuração de amplificadores-sensores por espelho de corrente. É possível que se queira um sinal à saída com duas vias para alimentar novos circuitos com entradas diferenciais como, por exemplo, outro estágio amplificador, um *latch* ou um *buffer* de saída (com entrada diferencial). Neste caso, como os nós X e Y não conseguem apresentar uma boa excursão, principalmente devido às limitações do transistor T_4 que tem o dreno e a porta em curto, então, uma possibilidade é a utilização de um par de amplificadores simétricos de tal forma que no segundo amplificador, ao contrário do primeiro, a linha de BIT está ligado à entrada do transistor T_1 e $\overline{\text{BIT}}$ está ligado à entrada do transistor T_2 . Isto permite que os nós X de cada amplificador seja usado para um nó da saída diferencial. Outras variações incluem a colocação da saída X, através da pré-carga, em um estado intermediário de alta



(a)



(b)

Figura L

As memórias dinâmicas são extremamente complexas devido à necessidade de técnicas especiais para se garantir a integridade dos dados. Por exemplo, algumas memórias comerciais contêm mais de vinte fases de sinais de relógio gerados internamente para efetuar as diversas funções de escrita, leitura e refrescamento. Outra característica das RAMs é que, geralmente, os pinos de endereçamento de fileiras e colunas não são separados, porém multiplexados no tempo, como mostra a **figura L(a)**. Objetiva-se, assim, compatibilizar o número de pinos com a densidade da memória e minimizar a contagem destes pinos, e, também, portanto, o custo de encapsulamento e da placa. O esquema da figura divide a memória em duas metades com o objetivo de melhor organizar os acessos às células, embora, em dispositivos comerciais, encontre-se divisões em número bem maior de blocos. Inicialmente a seleção de fileira é realizada utilizando-se os N bits de endereço que são passados pelo *buffer* de endereço de fileira. Os dados armazenados nas células dinâmicas, cuja linha de palavra foi ativada, são colocados nas linhas de coluna (de *bit*) correspondente sendo selecionado, então, pelo decodificador de coluna. Observe que os bits de endereçamento de fileira são os mesmos da de coluna. Durante o ciclo de leitura, após o endereço de fileira ser registrado pelo *buffer*, ele é desconectado dos pinos e qualquer variação nos bits de endereço não afeta mais os resultados até o próximo ciclo.

O diagrama de tempos da **figura I(b)** mostra a ocorrência de uma operação de leitura. A função de multiplexação no tempo dos endereços de fileira e coluna para um mesmo conjunto de pinos não pode ser realizado por meios estáticos assim como outros sinais como o de pré-carga das linhas de coluna. Estas funções dinâmicas são realizadas através de dois sinais externos de controle de endereço de fileira e coluna, \overline{RAS} e \overline{CAS} . Em um ciclo de leitura o sinal de escrita, \overline{W} , é mantido alto durante uma operação $\overline{RAS}/\overline{CAS}$. A saída será mantida em alta impedância até, que passado o tempo de acesso, o dado válido apareça nela.

No ciclo de escrita, o sinal \overline{W} está baixo durante a operação $\overline{RAS}/\overline{CAS}$. O *bit* de entrada de dados precisa estar correto na borda de descida de \overline{W} ou \overline{CAS} , aquela que ocorrer por último. Independente de qualquer outra atividade, toda célula deve ser refrescada a certo intervalo de tempo. Isto é feito de tal forma que cada fileira é acessada seqüencialmente com os amplificadores-sensores realizando a regeneração dos sinais. A seleção das fileiras é feita através de um circuito contador externo e nestas condições usa-se o sinal "RAS-only".

2.2.1. CÉLULAS DE DRAMS

A primeira célula de memória a ser usada amplamente é a mostrada na **figura m(a)**. Esta célula, ao contrário das estáticas, não necessita de projeto de relação geométrica (de condutância) entre os transistores internos sendo que todos os transistores podem ser realizados na dimensão mínima para minimizar a área. A capacitância onde o valor lógico da célula é armazenado é mostrada explicitamente na figura como C_g , porém, na realidade, ela é a própria capacitância parasitária do nó de entrada do transistor T_2 . Observe que neste esquema as linha de leitura W_{BIT} e escrita R_{BIT} são separados o que evita a destruição da carga acumulada dinamicamente, independentemente do controle de tempo. É comum, no entanto, se ter estas duas linhas unidas para economizar área, desde que os sinais de leitura e escrita, RD e WR , sejam ativados em tempos diferentes. Ainda assim, não haverá leitura destrutiva uma vez que o transistor T_2 isola a carga acumulada da linha de dados. No entanto, devido às correntes de fuga, o refrescamento ainda é necessário de tempos em tempos para reforçar o nível lógico em C_g . As linhas de *bit* são pré-carregadas tanto na leitura como na escrita, sendo que o sinal lido será sempre invertido em relação ao valor armazenado. Pode-se notar pelas linhas pontilhadas que um esquema com portas múltiplas para leitura pode ser facilmente implementado a partir do modelo básico.

Como a leitura não é destrutiva, os amplificadores-sensores podem ser dimensionados apenas visando o problema de refrescamento ao contrário do esquema da **figura m(b)** onde o problema de leitura destrutiva aparece. Este esquema apresenta, no entanto, grandes vantagens em matéria de ocupação de área uma vez que apenas um transistor T e um capacitor C_g são necessários na célula (a área do capacitor corresponde apenas à área da porta de um transistor (sem dreno e fonte). O segundo capacitor é apenas demonstrativo da

capacitância parasitária da própria linha de dados. Este esquema é o mais utilizado hoje para as RAMs de grandes dimensões. Existem, no entanto, muitas variações quanto à realização mais detalhada desta célula dependendo do método para a implementação do capacitor, do tipo de condutores para as linhas e colunas, etc.

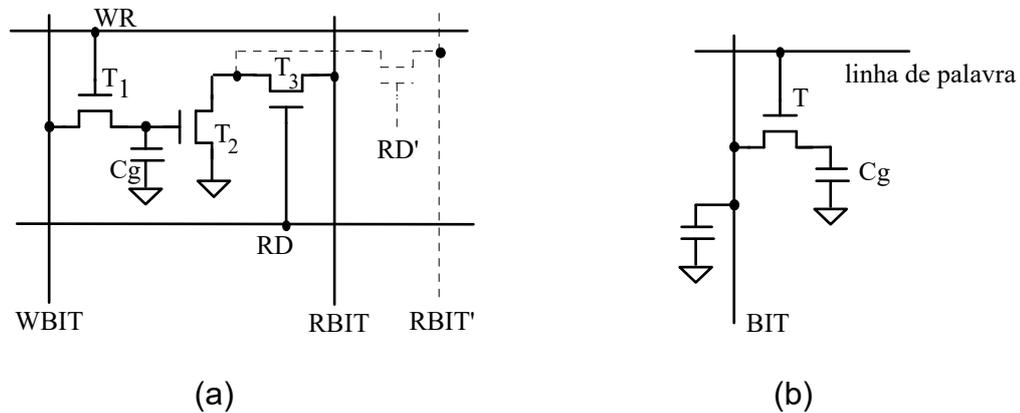


Figura M

Devido ao critério de minimização de área, C_g é normalmente muito pequeno, na ordem de 30 a 100 fF. Na leitura, a carga armazenada em C_g é compartilhada com a capacitância da linha de coluna de 10 a 20 vezes maior o que poderia reduzir um valor lógico '1' na mesma proporção. Portanto, os dados armazenados devem ser regenerados sempre que uma leitura é feita além do refrescamento constante que continua sendo necessário. A simplicidade aparente da célula é enganosa, pois a sua utilização acarreta uma série de dificuldades que um projetista médio não estaria apto a resolver. O projetista deveria optar ou pela solução mais simples apresentada anteriormente ou, então, usar células estáticas e só adotar a célula de um transistor apenas nos casos em que o desempenho geral do sistema esteja afetado.

2.2.2. FUNCIONAMENTO E AMPLIFICADORES-SENSORES

Como foi mencionado acima, o projeto e o funcionamento de uma DRAM é extremamente complexo. Tão logo o sinal de controle de fileira, RAS, for ativado, o mesmo ocorre para uma série de sinais internos de relógio. A **figura n** mostra o esqueleto de uma memória DRAM típica, exemplo no qual apenas alguns sinais de relógio aparecem. Inicialmente, antes do sinal de acesso de fileira ser levantado, o sinal de pré-carga é ativado de tal forma que todas as linhas de *bit* são carregados em um valor que, normalmente, é $V_{dd}/2$. Em seguida, uma das muitas linhas de palavra, WL, é selecionada conectando cada capacitor à sua linhas de *bit* correspondente, ao mesmo tempo que uma das linhas compensadoras de palavra (*dummy word line*), DWL0 ou DWL1, ligada à linha de dados oposta, é também ativada. Existem diversas formas de se carregar estas células de compensação como por exemplo através de uma tensão de referência ligada à essas células. O modelo da figura é, também,

bastante comum, onde a pré-carga é feita através do própria equalização entre os capacitores de dados e de compensação. Assumindo que as linhas de dados \overline{BL} e \overline{BL} são simétricas e que, por força do amplificador sensor, elas vão possuir os níveis '0' ou '1' (V_{dd}) de forma alternada, temos, após a leitura, uma redistribuição de cargas igualitária entre as duas linhas de dados e entre as células ligadas a DWL0 e DWL1 no valor $V_{dd}/2$ - isto após todas as linhas de palavra, WL, serem desativadas e as linhas compensadoras, DWL0 e DWL1, e o sinal de equalização, EQ, serem ativados,. Este esquema é extremamente positivo no que refere ao consumo de potência uma vez que não se necessita de nenhuma fonte independente de carga.

Pode-se perceber no esquema da figura que apenas uma linha de dado é ativado dentro de um par sendo a outra usada para a compensação. A rigor, as células compensadoras não seriam necessárias uma vez que a própria capacitância associada à linha de dados já seria suficiente para manter o valor entre '0' e V_{dd} . Porém, as células compensadoras são importantes para manter a simetria (os capacitores são do mesmo tamanho daqueles que armazenam dados), diminuir a suscetibilidade dos amplificadores sensores a variações de nível da pré-carga da linha de dados e aumentar a velocidade e confiabilidade do sistema.

Após a pré-carga, quando a leitura é feita, o valor de $V_{dd}/2$ da célula compensadora é comparada com o valor $V_{dd}/2+\Delta$ ou $V_{dd}/2-\Delta$ da outra linha de dados através do amplificador-sensor que amplifica a diferença de potencial em \overline{BL} e \overline{BL} . O diagrama de tempos da **figura n** mostra um exemplo onde o valor '1' armazenado na célula ligada à linha de palavra WL6 é lido quando o sinal SAE é ativado; o amplificador-sensor entra em ação, com a linha \overline{BL} indo rapidamente a '1', puxado pelo transistor tipo p do lado esquerdo, enquanto a linha \overline{BL} vai rapidamente a '0', puxado pelo transistor tipo n do lado direito.

O amplificador-sensor deve ser projetado com certo cuidado, em particular, a sua fonte de corrente (definido simplificada pelo sinal SAE) que drena a corrente para os pares cruzados de transistores. Quando o sinal RAS é ativado, a corrente drenada em direção à terra deve ser inicialmente pequena de tal forma que o transistor com a porta ativada no potencial mais alto, $V_{dd}/2$ (em relação a $V_{dd}/2-\Delta$, ou $V_{dd}/2-\Delta$ quando comparado a $V_{dd}/2$), seja o responsável pela condução tendo, assim, o potencial junto ao seu dreno sendo puxado para '0'. Após uma diferença razoável entre os sinais \overline{BL} e \overline{BL} ter se estabelecido, o chaveamento deve ser rápido. Se o chaveamento for muito abrupto logo de início, os dois transistores tipo n drenarão a corrente e ambos irão colocar os seus nós junto aos drenos no nível lógico '0'. Do ponto de vista do funcionamento, uma vez que a leitura é feita, o sinal SAE pode ser desativado e a linha de *bit* pode ser pré-carregada novamente.

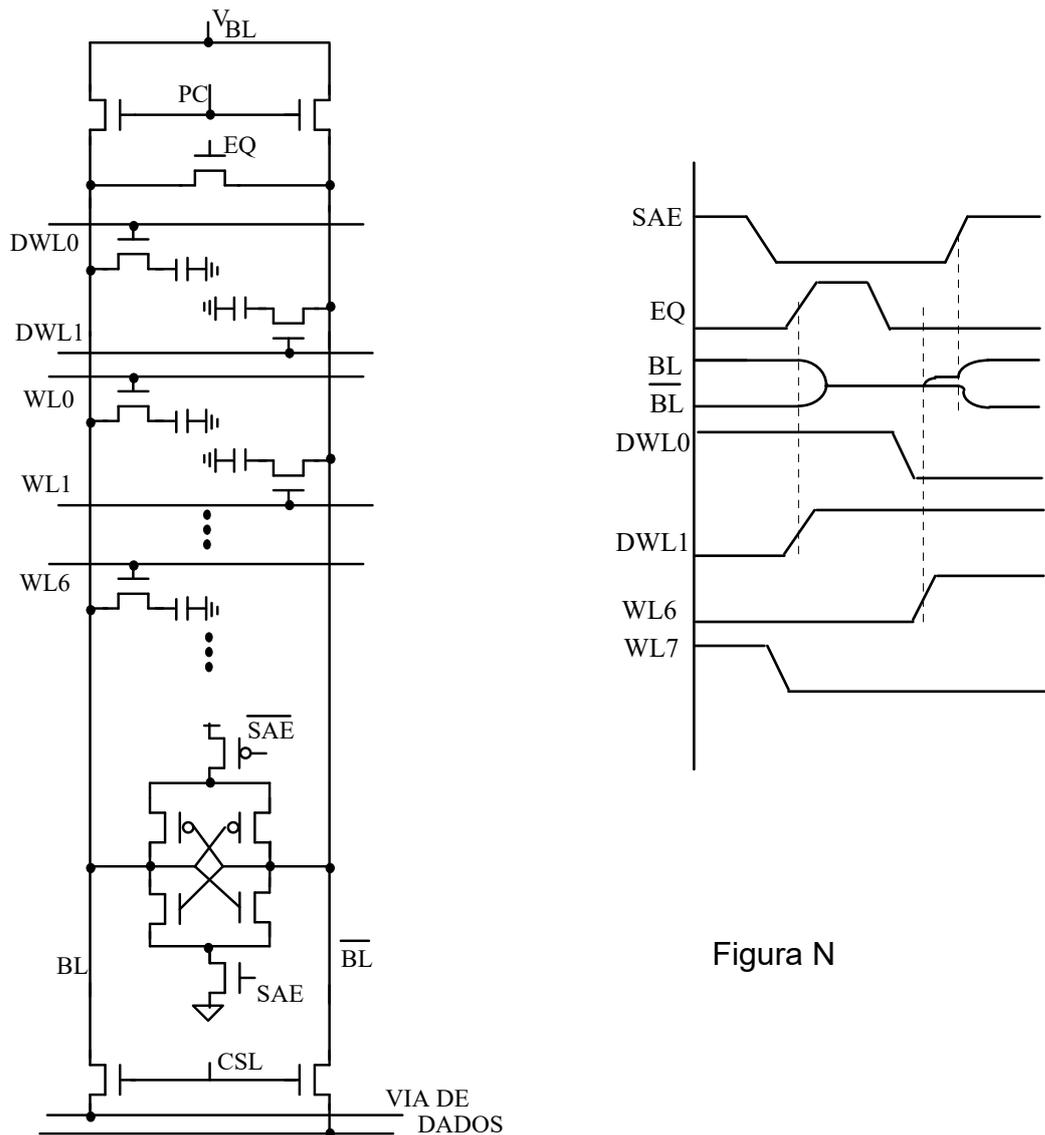


Figura N

Pré-carregar a linha de *bit* para o valor $V_{dd}/2$ traz algumas conseqüências negativas apesar de estarmos trabalhando com a região de maior ganho do transistor. Muitos optam em utilizar uma pré-carga para um valor maior (até mesmo V_{dd}) uma vez que é mais fácil o transistor n se descarregar do que o transistor p atingir o valor '1'. Se desejarmos que o sinal da célula de memória seja V_{dd} quando se tratar do valor lógico '1', o sinal na linha de *bit* terá de ser puxada até o valor $V_{dd}+V_{TN}$, uma vez que o efeito de perda do valor pelo perda de nível do transistor de passagem n da célula deve ser compensado.

A estrutura apresentada na figura para o amplificador-sensor é essencial em DRAMs uma vez que a regeneração do sinal é necessária tanto no refrescamento como na leitura (quando destrutiva). Os amplificadores-sensores para SRAMs apresentados nas seções anteriores, ao contrário, não são capazes de realizar tal função uma vez que os amplificadores apenas controlam o sinal de saída não se importando com os sinais de *bit* - o acoplamento é

capacitivo, o que impede o sinal de saída amplificado ser reescrito nas células de memória, porém, a resposta é mais rápida. Um amplificador alternativo para as DRAMs, para tornar o tempo de acesso mais rápido, é aquele com a inserção de um transistor de passagem entre cada par de transistor n e p (do biestável), isolando a linha de *bit* e a de saída resistivamente. Neste esquema, a saída é drenada mais rapidamente devido à menor capacitância apesar de o estabelecimento de sinais em BL e \overline{BL} se dar mais lentamente, mas isto pode ocorrer sem problemas, em paralelo, enquanto o sinal de coluna seleciona e trabalha o dado de saída.

Notar que outro aspecto importante para os amplificadores-sensores é quanto ao leiaute final - cada um deste circuito deve estar disponível para cada duas linhas ou uma coluna uma vez que o refrescamento se dá para uma fileira toda simultaneamente. Portanto, as células devem ser projetadas com cuidado visando o casamento do tamanho delas com os amplificadores e os transistores de seleção (comandados pela linha de seleção CSL, decorrente do CAS).

3. ROMS E VARIAÇÕES

As memórias de apenas leitura (ROMs) são utilizadas principalmente para o armazenamento de constantes multiplicadoras, informação de controle e instruções de programa em sistemas digitais. Do ponto de vista comportamental, são blocos ou dispositivos que fornecem uma saída fixa para cada combinação existente de entradas binárias. Sendo um bloco universal, pode ser utilizada, também, para implementar funções Booleanas quaisquer, porém outras soluções são, normalmente, preferidas quando realizadas como bloco interno de um CI, pois as ROMs tendem a ter um tempo de acesso lento além de desperdiçar espaço. As ROMs carregam a informação de forma “definitiva” e a sua rede de dados/células pode ser implementada, normalmente, como um plano OR de uma PLA através da inserção de um transistor junto ao nó correspondente ao cruzamento das linhas de palavra e de dados. Tal configuração pode ser vista na **figura o(a)** em um exemplo onde as palavras 1000, 0100, 0101 e 0010 estão armazenadas nos quatro endereços da ROM implementada como dois conjuntos de portas NOR (na técnica pseudo-NMOS).

A estratégia de implementação acima é típica de um projeto de CI onde a ROM é uma de suas partes (para microprogramação, por exemplo) e o próprio projetista já define a o conteúdo da memória antes da fabricação. Quando se trata de sistemas de uso mais geral em que o usuário define uma programação particular, a implementação pode ser feita através uma rede repleta de transistores sendo o conteúdo da ROM definido através das máscaras de contato ou de metal possibilitando, assim, ao fabricante ter os componentes pré-fabricados e em estoque. A **figura o(b)** apresenta um leiaute possível em diagrama *stick* para a o exemplo da **figura o(a)** onde os contatos foram definidos apenas para os transistores onde o valor ‘1’ está armazenado.

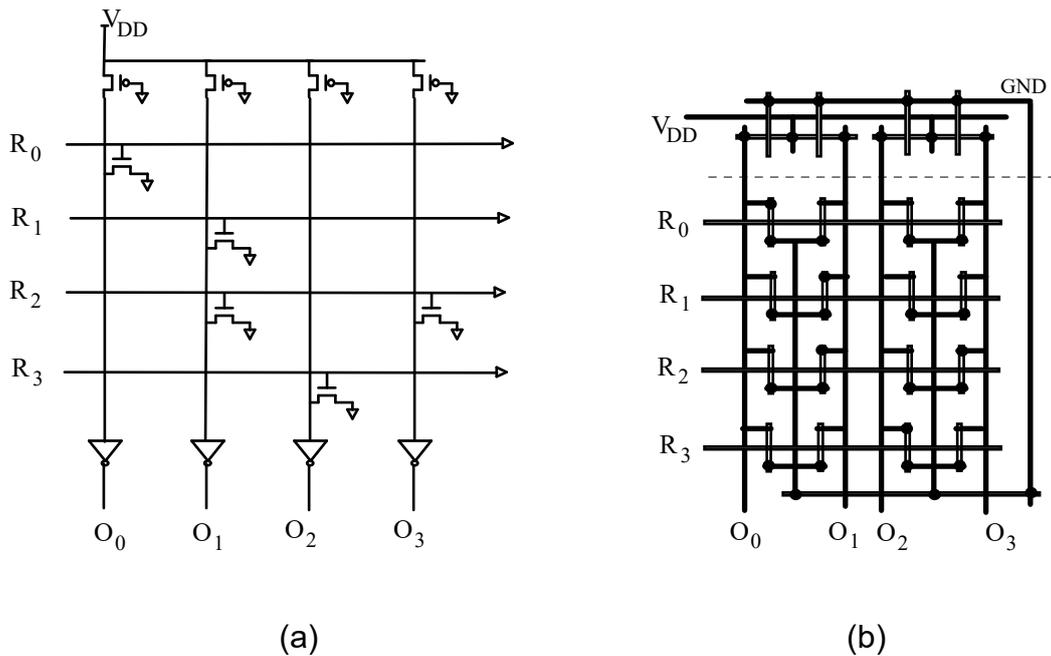


Figura 0

3.1. DECODIFICAÇÃO E AMPLIFICADORES-SENSORES

As memórias ROM assim como as RAMs são universais compreendendo todas as 2^n fileiras de endereçamento portanto um decodificador equivalente aos apresentados nas seções anteriores se aplica. Comparado a uma PLA, o decodificador corresponde ao plano AND desta estrutura porém com todas as possíveis linhas de palavra. Na verdade, para manter a geometria do bloco ou da pastilha favorável e equilibrada, faz-se a decodificação por fileiras e por colunas, como no caso das RAMs, e o leiaute tende a ser, na realidade, bem mais complicado do que o apresentada na **figura 0(b)**.

O amplificador-sensor de uma ROM se diferencia bastante daquelas vistas nas seções anteriores, pois as ROMs têm linhas de dados simples ao invés de pares de linhas com sinais invertidos como encontrada nas RAMs. Portanto, os amplificadores são ou de portas simples (com uma única entrada e única saída) ou com porta dupla de entrada (com alguma tensão de referência em uma das entradas). No segundo caso, amplificadores diferenciais como as apresentadas nas seções anteriores podem ser utilizados tendo uma tensão de referência ($V_{dd}/2$, por exemplo) à uma entrada enquanto a outra está conectada à saídas de células (um amplificador por saída).

Um exemplo de um amplificador-sensor com entrada única é mostrado na **figura p**, assumindo-se, neste caso, que a ROM é parte de um CI (ou mesmo uma EPROM), cujo acionamento de dá pelo pino e sinal *chip enable*, CE. Pode-se notar pela figura que, quando a ROM não esta' seleccionada (CE é '0'), o transistor T_1 está fora de operação e o transistor T_5 está ligado, aterrando o nó X. O amplificador está, neste caso, em um estado de espera, de economia de

energia. Quando CE é '1', o chip é selecionado, o transistor p, T_1 , passa a conduzir enquanto o transistor n, T_5 , sai de cena. Enquanto nenhuma coluna for selecionada (e portanto, os transistores da rede de programação das ROMs estão isolados dos amplificadores), pode-se observar que não existe nenhum caminho de X para terra e como, inicialmente, a saída do inversor I_2 estava em '0', a tendência é de transistor T_3 puxar o valor X para '1'. Isto não ocorre completamente devido à realimentação existente que provoca o desligamento do transistor T_3 e V_x fica em um valor maior, mas em torno, de $V_{dd}/2$. O valor exato de V_x é definido pelas geometrias de T_2 e T_4 e à saída do amplificador temos o valor '0'. Com a seleção de uma coluna, se o transistor da célula da ROM não estiver conduzindo, a saída do amplificador sensor permanecerá em '0' pois V_x continua sem caminho para terra. Se ao contrário, algum dado estiver carregado na célula, o transistor da linha selecionada puxará V_x para o valor '0' e a seqüência de inversores colocará '1' à saída. O fato de V_x corresponder, inicialmente, à região de máximo ganho do inversor I_1 não é acidental, pois, assim, V_x irá rapidamente em direção a '0'. O transistor T_2 , no entanto, evita que tal valor seja realmente atingido, o que acaba sendo vantajoso em novas leituras de palavras da ROM, com um tempo de recuperação mais rápido do que com o valor intermediário de V_x .

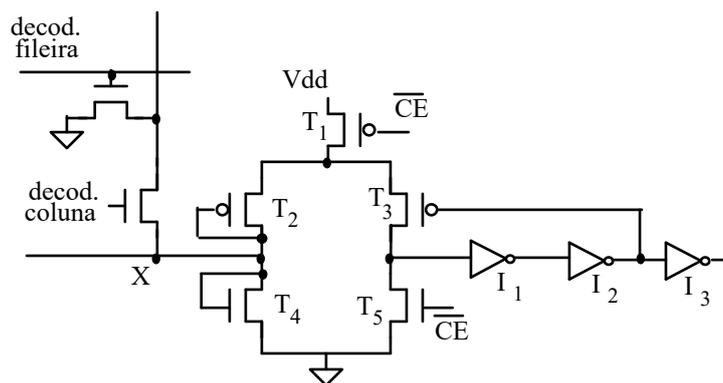


Figura P

3.2. PROMS, EPROMS, EEPROMS e FLASH

Ao contrário das RAMs, as memórias ROMs se caracterizam pela não volatilidade de seus dados, isto é, a informação permanece mesmo com a falta de alimentação elétrica. As ROMs descritas anteriormente são aquelas programadas por máscara durante a fabricação e que não podem ser modificadas após a programação. Existem outras categorias de ROMs que são programáveis em campo, normalmente, dispositivos comerciais de prateleira. As PROMs, por exemplo, são dispositivos que podem ser programados através de pulsos elétricos aplicados seletivamente, com a função de destruírem conexões entre linhas e colunas dentro do dispositivo. A programação é permanente após a destruição destes contatos.

Outra classe de ROMs é aquela que além de oferecer a programação em campo como as PROMs, permite, também, a reprogramação de dados. Devido à condição de não volatilidade, técnicas de processamento tecnológico mais complexas são, normalmente, necessárias. Existem dois tipos destes dispositivos: as com apagamento por luz ultravioleta, EPROMs, e as apagáveis eletricamente, EEPROMs (ou E²PROMs). As EPROMs são construídas com transistores canal n em uma estrutura com duas portas de silício policristalino sobrepostas (uma acima da outra, com uma separação de óxido entre elas). A porta G1, mais próxima do canal a ser formado, é flutuante, não estando conectada a nada, enquanto a porta mais externa, G2, serve de linha de palavra assim como para efeito de programação. A programação é realizada através da acumulação de cargas na porta G1 através da aplicação de alto potencial (~25V) no dreno e na porta (com fonte e substrato aterrados) provocando uma corrente muito alta próxima à superfície do transistor. Os elétrons com alta energia conseguem passar a barreira do óxido e se acumulam em G1. Estas cargas negativas não permitem, em uma operação normal com 5V junto à porta G2, que o canal seja formado o que significa que o transistor está programado em estado não condutivo com a saída da linha de dados sempre em '1'. Sem a programação, o funcionamento do transistor é normal, exceto por uma tensão de limiar, V_T , mais alta que o normal. A luz ultravioleta é utilizada no apagamento da informação uma vez que ela tem uma capacidade de tornar o óxido de silício condutivo e as cargas armazenadas podem, assim, ser retiradas. Para a utilização desta técnica é necessária a construção de uma janela que permita a passagem da luz ultravioleta de desprogramação.

As E²PROMs, por sua vez, são programadas em um transistor de duas portas semelhante ao do EPROM, porém, com uma camada de óxido de porta G1 bastante fina (~10 nm). O efeito verificado, neste caso, é o de tunelamento (*Fowler-Nordheim tunneling*) que ocorre para uma tensão de porta aproximadamente do dobro de uma operação normal com as cargas se acumulando também na porta G1. A aplicação reversa do mesmo potencial é suficiente para desprogramar o transistor. A tecnologia flash é semelhante, porém, para maior rapidez, as escritas e leituras são feitas em blocos.

BIBLIOGRAFIA

- Lance A. Glassner e Daniel W. Dobberpuhl, "The Design and Analysis of VLSI Circuits", Addison-Wesley Publishing Company, 1985.
- Phillip E. Allen e Douglas R. Holberg, "CMOS Analog Circuit Design", Holt, Rinehart and Winston, Inc. , 1987.
- David A. Hodges e Horace G. Jackson, "Analysis and Design of Digital Integrated Circuits", McGraw-Hill, 1988, segunda edição.
- Neil Weste e Kamran Eshraghian, "Principles of CMOS VLSI Design - A Systems Perspective", Addison-Wesley Publishing Company, 1988.
- H. B. Bakoglu, "Circuits, Interconnections and Packaging for VLSI", Addison-Wesley Publishing Company, 1990.

- Douglas A. Pucknell e Kamran Eshraghian, "Basic VLSI Design - Systems and Circuits", Prentice Hall, 1994, terceira edição.
- Edson Fregni e Antonio M. Saraiva, "Engenharia de Projeto Lógico Digital: Conceito e Prática", Editora Edgard Blücher, 1995.