Simulação de um neurónio 100% analógico com processamento de corrente

Pedro Kulzer, António Branco, Dinis Santos

Resumo - **Projectou-se e simulou-se um neurónio 100% analógico, implementável em tecnologia CMOS de 1.2μm. O neurónio consiste em circuitos simples de processamento de corrente, desde as entradas das sinapses até à saída do axónio. Apenas a entrada do peso é em tensão, para atingir o máximo de rapidez de resposta. O neurónio tem um limiar de disparo adaptativo, bem como um circuito de aprendizagem não-supervisionada, autónoma e local.**

Abstract - **A 100% analog neuron, implementable in CMOS 1.2μm technology, was designed and simulated. The neuron consists of simple current processing circuits, right from the synapses' inputs to the axon's output. Only the weight input is in voltage form, to achieve the fastest possible response. The neuron has an adaptive firing threshold, as well as an autonomous, local, non-supervised learning circuit.**

I. INTRODUÇÃO*

Neste trabalho projectou-se um neurónio analógico em tecnologia CMOS de 1.2μm, com as suas diversas componentes: sinapses, limiar adaptativo, gerador de sigmóide e circuito de aprendizagem autónomo. Para construir uma rede completa, basta aglomerar o conjunto desejado destes neurónios analógicos e realizar as ligações entre eles.

II. MOTIVAÇÃO PARA ESTE TRABALHO

A motivação para este trabalho surgiu pela necessidade de se implementar uma rede neuronal treinável e suficientemente rápida para o processamento de sinais até 80 Mhz ou mesmo mais. Isto implica tempos de resposta da rede da ordem dos 10ns, o que apenas poderá ser conseguido através de neurónios completamente analógicos. A vantagem destes é que não há relógios nem acessos a memórias ou outros circuitos digitais "lentos". Apenas existe um tempo de resposta mínimo determinado pelo tempo de estabelecimento das saídas da rede analógica, e portanto, dos neurónios analógicos. O comportamento é semelhante ao de um circuito com amplificadores operacionais, dando origem a um tempo de estabelecimento global do sistema analógico.

Alguns argumentos que favorecem a opção por esta implementação 100% analógica são os que se seguem.

A. Tempo de resposta independente do nº de neurónios

As realizações digitais em computador exigem a execução das tarefas de cada neurónio, um de cada vez. A menos que se use multi-processamento, o tempo de resposta de cada camada é claramente proporcional ao número de neurónios em cada uma delas. Nas realizações analógicas em VLSI essa limitação não existe, já que se processa uma camada duma só vez, obtendo-se um puro multi-processamento.

B. Implementação fácil em VLSI

O facto de cada neurónio 100% analógico ser totalmente independente de sinais de controlo exteriores torna-o num módulo ideal para ser implementado numa estrutura VLSI com *layout* repetitivo. As implementações digitais ou mistas exigem a utilização de pistas que transportam sinais de controlo (*clocks*, *strobes*, *refresh*, etc.) a cada neurónio de forma individual, o que pode aumentar bastante a complexidade do projecto e exigir estudos de topologias de ligações que facilitem o trabalho.

O neurónio 100% analógico apenas necessita que lhe forneçam as linhas dos sinais de entrada, que também poderão ser apenas locais, facilitando assim o projecto do *layout*. Na Fig.1 mostra-se um exemplo de *layout*.

C. Tolerância a falhas

Justamente devido à natureza distribuída do *layout* em VLSI duma rede neuronal, essa rede será bastante resistente a falhas do próprio *chip* (poeiras e danos posteriores). Assim, perspectiva-se a possibilidade de aproveitamento quase total dos *chips* fabricados em cada *wafer* de silício. Só não se poderão aproveitar aqueles que ficaram demasiado danificados (partes periféricas da *wafer*, pistas

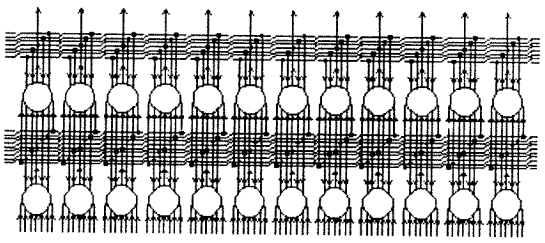


Fig. 1 - Exemplo de implementação de *layout* em que todas as ligações são locais (em forma de campo receptivo), estendendo-se as pistas de distribuição dos sinais pelo comprimento de apenas alguns neurónios. Desta forma, a área ocupada pelas pistas é mínima e o desenho facilitado.

* Trabalho realizado no âmbito da disciplina de VLSI analógico.

indispensáveis danificadas).

As redes existentes nos *chips* serão também resistentes a danos sofridos posteriormente durante a aplicação prática. Mesmo danificadas, estas redes continuarão a funcionar relativamente bem, dependendo da extensão dos danos (5 a 10%), podendo mesmo reorganizar-se e reaprender de forma a desempenhar melhor as suas funções com a funcionalidade que lhe resta. Isto é extremamente importante para aplicações espaciais (satélites, sondas, veículos lunares). Pelo contrário, num sistema computacional convencional, a ocorrência de uma falha numa única pista ou *bit* de memória compromete inevitavelmente o funcionamento.

D. Aprendizagem não-supervisionada incorporada

Uma rede que utilize uma aprendizagem não-supervisionada possui a potencialidade de distribuir essa aprendizagem por regras locais aos neurónios, o que também elimina a necessidade de pistas de sinais de supervisão da aprendizagem.

E. Adaptação às imprecisões dos neurónios

A presença da aprendizagem numa rede neuronal elimina à partida a necessidade de precisões impensáveis em VLSI. Qualquer não-linearidade, *offset* ou imprecisão de outra natureza, desde que mantida abaixo dum máximo crítico, será compensada pela aprendizagem efectuada em cada neurónio, à semelhança do que acontece num amplificador operacional realimentado: a saída é mais ou menos independente do valor exacto e da linearidade do ganho.

III. OBJECTIVOS

Propõe-se a implementação de um neurónio completamente analógico e autónomo, sem a necessidade de quaisquer circuitos externos. Este neurónio vai ter de obedecer aos seguintes requisitos:

A. Funcionamento dos transístores MOS na região de inversão forte

Para manter um baixo consumo optámos por uma implementação em tecnologia CMOS. Pretendemos um consumo por neurónio tão baixo quanto possível, sem que isso implique excessiva falta de precisão.

No entanto, a velocidade de processamento desejada não é tão baixa que permita um funcionamento na região de inversão fraca, pelo que se optou pelo funcionamento na região de inversão forte (acima da tensão de limiar, V_t).

B. Processamento de corrente

Nas implementações referidas na literatura, os sinais de entrada e de saída no neurónio são sinais de tensão, o que impõe limites estritos ao valor das capacidades de carga. As velocidades de processamentos são baixas, da ordem das centenas de ns (aproximadamente de 200ns no *chip* de Choi [2]). Como se pretendem tempos da ordem dos 10 ns, se possível com a tecnologia actual, é quase inevitável a opção pelo funcionamento dos circuitos no chamado modo de corrente Na Fig.2 apresenta-se um modelo de neurónio em modo de corrente.

Uma simplificação adicional consiste na eliminação de praticamente todos os circuitos de conversão corrente-tensão e tensão-corrente, excepto nas entradas.

C. Simplicidade de implementação

Como queremos obter um neurónio o mais simples possível mas funcional, de forma a conseguir níveis de integração em VLSI o mais elevados possível, teremos de manter o nível de complexidade por neurónio também o mais baixo possível. Isso será especialmente crítico nas sinapses, já que cada neurónio pode ter centenas ou até milhares de sinapses, ocupando estas a maior parte da área do *chip*.

Sempre que possível, utilizaram-se dimensões mínimas para os transístores. Isso terá vantagens no espaço ocupado, bem como na minimização das capacidades parasitas o que aumenta a velocidade, sem ter desvantagens resultantes da transcondutância menor destes transístores, já que isso vai ser pouco importante.

D. Velocidade elevada

Como estamos interessados em obter a máxima velocidade de resposta possível para estes neurónios, tivemos de ter algum cuidado no projecto e teste dos vários circuitos. A alínea anterior já implica, só por si, uma maior velocidade de resposta do que em circuitos mais complexos, devido ao menor número de transístores envolvidos. No entanto, é possível que tenhamos de sacrificar alguma simplicidade para obter maior rapidez, e vice-versa. As dimensões dos transístores poderão não ser mínimas para se conseguir esse objectivo.

IV. PROCESSAMENTO DE CORRENTE

Teremos um modelo exterior em que as entradas e as saídas são sinais de corrente. É óbvio que no seu interior também se vai manter a operação em modo de corrente, pelo menos no que respeita às partes do circuito em que se exige a máxima velocidade de resposta.

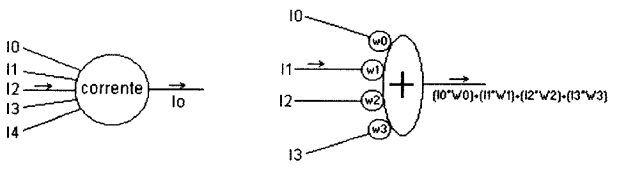


Fig. 2 - Modelo do neurónio a operar em modo de corrente. Os sinais de entrada são multiplicados pelos pesos e somados no corpo do neurónio. Se este tivesse uma função de saída linear, teríamos à saída uma soma pesada das entradas.

No multiplicador de *Gilbert* utilizado por Choi [2], são empregues as técnicas de *cascode* e fontes de corrente anteriores para melhorar a sua resposta, o que aqui já está implícito na própria arquitectura.

Os aspectos mais importantes a testar foram os tempos de resposta, a linearidade e as distorções.

V. SINAPSE

A sinapse é o lugar biológico onde se transmite um potencial excitatório ou inibitório ao corpo do neurónio, cuja intensidade depende directamente do factor eficiência da sinapse. Este processo pode ser visto como uma multiplicação aproximada, do sinal de entrada por um valor a que chamamos *peso*. Isto significa que o neurónio vai receber um potencial que será uma soma pesada dos sinais de entrada.

Cada sinapse vai ser constituída por um circuito multiplicador que vai receber um sinal de entrada e um sinal de peso, gerando uma saída que é o produto dos dois. Como não se conseguem multiplicar duas correntes, temos de utilizar um dos circuitos multiplicadores habituais. Podemos escolher entre o multiplicador de *Gilbert* e multiplicadores mais simples e com menos quadrantes de operação. O multiplicador tem que obedecer o melhor possível às seguintes especificações:

A. Simplicidade de implementação

Não queremos um circuito complicado e altamente preciso, bastando-nos um circuito que cumpra a sua função com o mínimo de precisão e transístores. Assim, no caso no modelo desejado em que apenas se querem saídas positivas, podemos utilizar um simples amplificador de transcondutância.

B. Entrada em corrente

O amplificador de transcondutância proposto possui uma entrada em tensão que terá de ser convertida numa entrada de corrente. Para isso, basta ligar essa entrada em corrente directamente ao par diferencial.

A função de transferência deste multiplicador (ignorando o efeito de corpo), será da forma:

$$i_{out} = g_{m_{Global}} \cdot \Delta v = \frac{i_{in}}{V_{GS} - V_t} \cdot \Delta v \qquad (1)$$

VGS é a tensão média nos transístores do par diferencial, e v é a diferença de potencial entre os dois condensadores.

C. Corrente mínima igual a zero

No que respeita à corrente mínima de entrada, esta será de 0A (ausência de actividade), o que pode causar problemas no funcionamento do par diferencial para correntes muito pequenas. De qualquer forma, esta é uma exigência que terá de ser cumprida pois as entradas podem variar entre zero e um valor máximo. Se houver problemas demasiados na zona de baixas correntes de entrada, teremos eventualmente de recorrer a multiplicadores cujo circuito esteja permanentemente polarizado a uma dada corrente mínima maior que zero.

D. Distribuição das velocidades pelas entradas

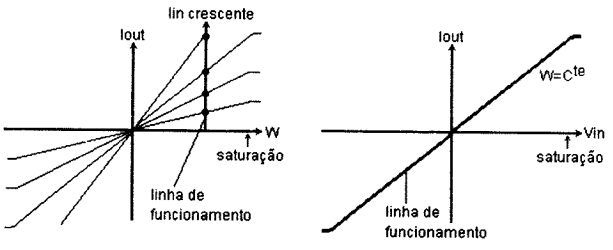
Já que este multiplicador tem de ser mesmo implementado pelo amplificador de transcondutância proposto, existe o problema da entrada diferencial em tensão. Esta será a única entrada em modo de tensão, o que reduzirá a velocidade de resposta de todo o circuito a essa entrada. Para minimizar este problema aparentemente sem solução, podemos ligar o sinal mais crítico à entrada mais rápida. Em princípio, vai interessar ter velocidade elevada para as entradas em corrente das sinapses.

As entradas em tensão referentes aos pesos não vão ser críticas pois estes variarão mais lentamente, além de atingirem estados aproximadamente estacionários durante os processos de aprendizagem da respectiva rede onde estarão integrados.

Uma vantagem resultante desta configuração das entradas, é que com um dado valor de peso abaixo dum limite máximo que provocaria saturação, o circuito virtualmente não satura enquanto se aumentar a entrada em corrente. Além de algumas não-linearidades, o circuito responde com uma corrente de saída aproximadamente proporcional à de entrada, aparentemente sem limitações de saturação. A única limitação é devida ao ponto em que as polarizações deixam de funcionar correctamente, devido a tensões dreno-fonte elevadas demais para a alimentação conseguir produzir. Isto é ilustrado em forma de gráficos no Grf.1.

E. Peso numa só capacidade

A nossa proposta é de armazenar cada peso numa só capacidade, sob a forma duma carga positiva. Isto simplifica grandemente o circuito de actualização dos pesos, apesar de degradar ligeiramente a resposta do par diferencial, devido ao ponto central já não estar à massa para sinais. O circuito resultante e as respectivas equações são mostrados na Fig.3.



Grf. 1 - À esquerda mostra-se a linha de funcionamento do multiplicador por nós proposto. Aí vê-se claramente que, para um dado peso abaixo do valor que provocaria saturação, obtém-se uma resposta linear para qualquer valor da corrente de entrada também abaixo de um valor máximo. Este valor máximo será apenas fixado pela tensão de alimentação. À direita temos a linha de funcionamento dum multiplicador em que o sinal de entrada é aplicado às entradas do par diferencial. Qualquer que seja o valor do peso, há um valor limite de tensão de entrada.

Na sinapse implementada por Barranco [1] também se utiliza um único condensador para o armazenamento do respectivo peso.

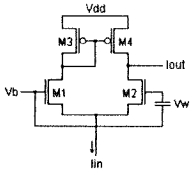
F. Ausência de refrescamento dos pesos

Como o objectivo da implementação destes neurónios analógicos era o de colocá-los a trabalhar a velocidades de aprendizagem da ordem dos 100Mhz, certamente que as fugas dos condensadores não terão qualquer importância significativa. As actualizações deles serão tão rápidas, que as fugas são constantemente compensadas. Além disso, não tem qualquer interesse prático conseguir manter os valores dos pesos durante longos períodos de *stand-by*, já que estes neurónios se destinam a serem integrados num sistema altamente adaptativo que vê o ambiente a mudar constantemente, pelo que terá de ser capaz de actualizar os pesos muito rapidamente. Esta última consideração elimina a necessidade de refrescamento. Mesmo que fosse necessário um refrescamento para processos mais lentos, poder-se-ia utilizar uma DAC neuronal como em Barranco [1].

G. Mínima dissipação de potência

A dissipação de potência é proporcional à corrente e tensão de alimentação da sinapse. Durante as experiências posteriores, vamos procurar valores razoáveis para estes dois parâmetros. Podemos desde já adiantar que a tensão *standard* de 3.3V utilizada nos processadores de baixo consumo, é perfeitamente adequada para a alimentação desta sinapse. Desta forma, garante-se uma parcela de 1.1V para a polarização de todos os transístores na zona de saturação (desde que a tensão de limiar, V_t, destes seja inferior àquele valor).

Com esta tensão de alimentação e uma corrente de entrada máxima de 5μA, o consumo seria de 16.5μW por sinapse. Se o sistema final contivesse 100 neurónios com 30 sinapses cada, o consumo total máximo (com todos os neurónios activos) seria de 50mW. Este consumo refere-se apenas às sinapses e teria de ser acrescido do consumo do circuito de geração de sigmóide, que também terá um consumo da mesma ordem de grandeza. Assim, o consumo de potência do sistema pode atingir os 100mW.



$$I_{out} = K \cdot I_{in} \cdot V_w$$

,em que

$$K = \frac{1}{V_{GS} - V_t}$$

Fig. 3 - No multiplicador proposto, o valor do peso é armazenado sob a forma de um condensador.

VI. DIMENSIONAMENTO E SIMULAÇÃO NO SPICE

Em todo o trabalho, o simulador de circuitos eléctricos PSPICE foi utilizado para analisar o comportamento, bem como para obter valores experimentalmente razoáveis para as diversas grandezas.

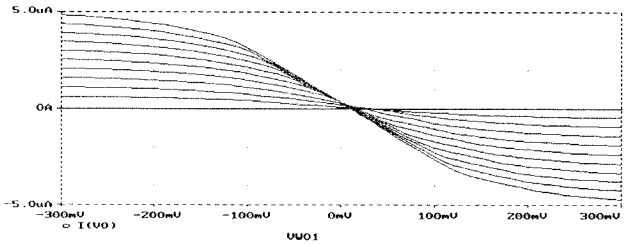
Quanto à corrente máxima de entrada, experimentámos inicialmente 5μA, o que poderá eventualmente ser baixo demais para atingir a velocidade de resposta desejada. Com esta tensão de alimentação, a tensão que se deverá colocar numa das entradas do par diferencial, será de 2V.

O valor máximo do peso corresponde à máxima diferença de potencial entre as entradas do par diferencial que ainda não provoca demasiada saturação. Este valor será certamente da ordem das centenas de millivolt, já que os transístores estarão polarizados no limite da sua zona de saturação (1.1V), pelo que qualquer perturbação poderá levá-los para fora dessa zona. Os resultados da simulação no Grf.2 mostram que esses limites são da ordem da centena de millivolt.

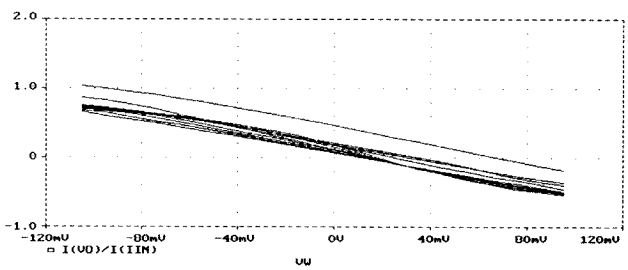
Podemos ainda observar a relação linear entre o peso e a saída do multiplicador, bem como o valor da constante K. Enquanto que a primeira curva devia ser uma linha recta com valores proporcionais apenas ao peso, já no segundo caso esperamos uma linha horizontal de valor igual àquela constante. Os resultados são mostrados nos gráficos Grf.3 e Grf.4.

Este valor experimental de K aproxima-se do valor teórico que se situará algures perto de:

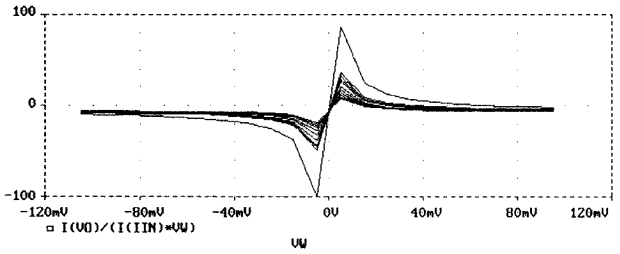
$$K_{teórico} = \frac{1}{1.07 - 0.77} = 3.3 \qquad (2)$$



Grf. 2 - Varrimento do valor do peso, obtido no PSPICE. Pode observar-se que a saída do multiplicador satura para pesos superiores a 100mV em valor absoluto. Assim, o valor máximo para os pesos será de 100mV.



Grf. 3 - Estas curvas representam o quociente da corrente de saída e de entrada, que se agrupam numa zona bem delimitada. A curva mais afastada das restantes é a referente a uma corrente de entrada de 0.01A. Para as correntes maiores já se verificou um menor erro.



Grf. 4 - Estas curvas representam o quociente da corrente de saída pela de entrada e pelo valor do peso. Desta vez já se obteve parte duma linha horizontal que tende para um valor de ordenada que só pode ser visto fazendo um *zoom*. A curva mais afastada das restantes é mais uma vez referente a uma corrente de entrada de 0.01A. Para as correntes maiores já se verificou um menor erro.

Podemos agora definir um *factor de rejeição da entrada*, para comparações futuras. Para peso nulo, designar-se-á aqui por *Zero Weight Input Rejection Ratio (ZWIRR)*:

$$ZWIRR = \frac{Iin}{Iout}\Big|_{Vw=0} \quad (3)$$

$$= \frac{2\mu A}{360n} = 13.9 (22.9dB)$$

O respectivo gráfico da variação da saída nestas condições, é mostrado no Grf.5.

Da mesma forma se podia introduzir o conceito de *Zero Input Weight Rejection Ratio*. No entanto, no caso presente esse valor seria infinito, uma vez que a corrente de saída é zero quando a corrente de entrada é zero.

Nas secções seguintes apresentam-se os resultados da simulação da resposta deste multiplicador isolado, para impulsos rápidos na entrada. Igualmente interessante é a observação da variação da velocidade d e resposta com os pesos, embora, como referido acima, este parâmetro seja menos crítico.

VII. CIRCUITO DE EXTRACÇÃO DA SAÍDA DO MULTIPLICADOR

Este circuito servirá ao mesmo tempo para permitir a junção de várias sinapses num ponto comum. Na Fig.4 podem-se ver os transístores M5 e M6 acrescentados para o efeito. Tal como já fizemos para o multiplicador isolado, vamos investigar os valores dos factores de amplificação obtidos por esta configuração com espelho de corrente na saída.

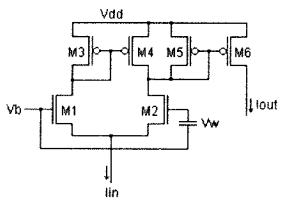
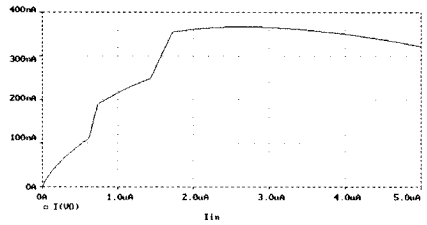


Fig. 4 - Os transístores M5 e M6 formam o circuito de extracção da corrente proveniente do multiplicador.



Grf. 5 - Corrente de saída com peso nulo. Esta corrente é baixa mas diferente de zero, o que pode ser útil, na medida em que um peso acidentalmente colocado a zero ainda pode aprender algo, pelo facto de ainda ter alguma acção.

Desta vez estes factores de amplificação já contém mais uma componente referente ao espelho de corrente de saída:

$$K = \frac{1}{V_{GS} - V_t} \cdot G_I \text{ , c/ } G_I = \frac{W_{M6}}{L_{M6}} \cdot \frac{L_{M5}}{W_{M5}} = \frac{2.4}{6} \cdot \frac{24}{2.4} = 4 \quad (4)$$

em que GI é o ganho em corrente fornecido pelo espelho de corrente da saída.

Um valor teórico para K seria:

$$K = \frac{1}{1.07 - 0.77} \cdot 4 = 13.3$$

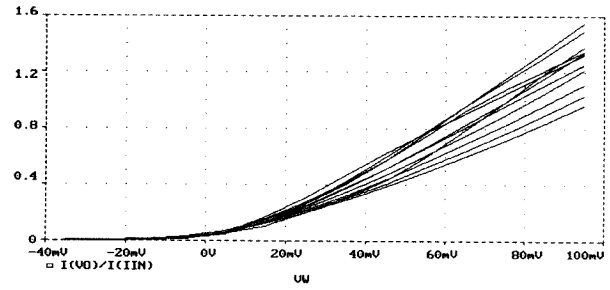
também considerando um peso nulo. Este valor é da ordem de grandeza dos valores experimentais. Neste caso, temos um *Zero Weight Input Rejection Ratio* de:

$$ZWIRR = \frac{5\mu}{180n} = 27.8 (28.9dB)$$

Até este valor melhorou em relação ao anterior do multiplicador isolado (22.9dB). Isto será devido à distorção inicial da curva da corrente de saída do extractor, onde as correntes iniciais são mais baixas do que no multiplicador isolado original. No Grf.6 ainda se mostram as curvas correspondentes ao factor de amplificação da corrente de entrada para a de saída, para efeitos de apreciação da dispersão da mesma, relativamente ao multiplicador isolado.

VIII. JUNÇÃO DE SINAPSES

A junção de sinapses efectua-se muito simplesmente pela soma das suas correntes no ponto de entrada deste circuito extractor. Dessa soma resultará uma corrente que pode ser positiva ou negativa. Corrente positiva significa corrente no sentido de dentro para fora do extractor. Como é evidente, este tipo de extractor não funciona para correntes negativas (isto é, de fora para dentro), pelo que para essas correntes a saída é simplesmente zero. Na Fig.5 mostra-se a forma de interligação da várias sinapses. Cada sinapse introduz uma capacidade parasita resultante dos seus dois transístores de saída, o que reduz a velocidade de resposta global. Nos gráficos Grf.7 a Grf.11 mostram-se diversas situações de funcionamento que mais importa considerar.



Grf. 6 - Curvas referentes ao factor de amplificação que ainda contém o peso.

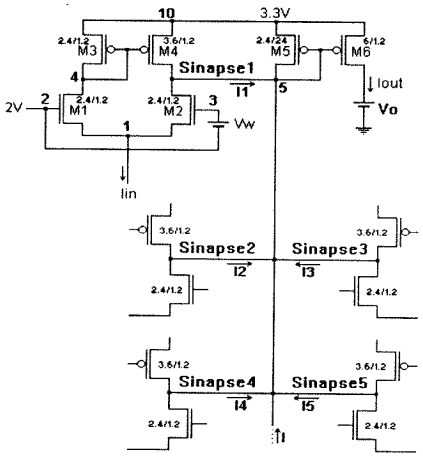
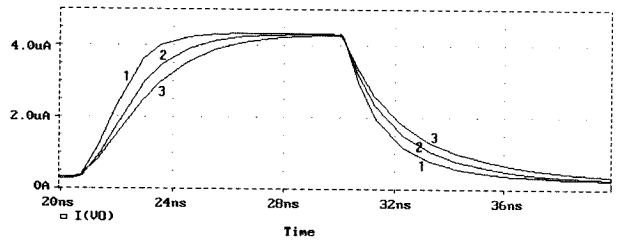
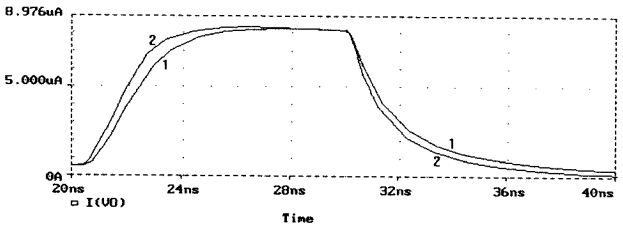


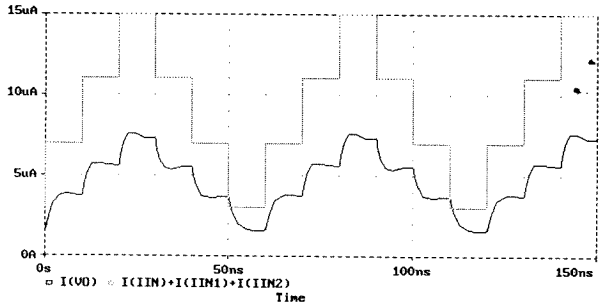
Fig. 5 - Aspecto final do circuito extractor de corrente com todas as sinapses associadas, penduradas no nó 5.



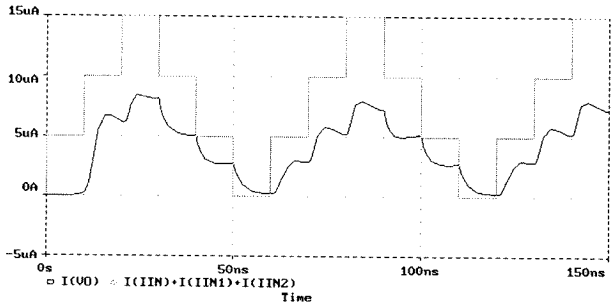
Grf. 7 - Respostas para uma só sinapse (1) e de duas e três sinapses - (2) e (3), sendo apenas uma sinapse excitada e as restantes não-excitadas. Pode-se concluir que, quantas mais sinapses se juntarem mais lenta fica a resposta à saída do extractor. Isto deve-se essencialmente às capacidades que cada sinapse acarreta para o nó 5 (Fig.5).



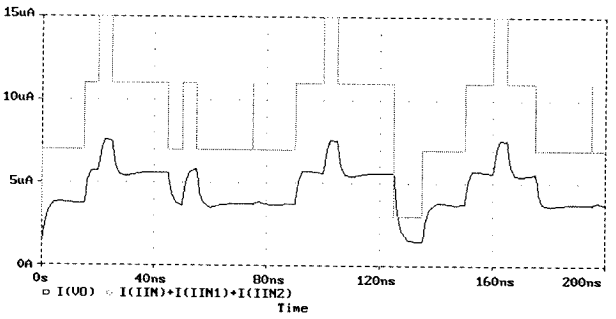
Grf. 8 - Respostas para apenas uma sinapse excitada (1) e para ambas excitadas (2) numa junção de duas sinapses. Quando ambas são excitadas, a velocidade de resposta é maior do que apenas uma sinapse excitada. Isso será devido à maior corrente disponível para carregar as capacidades parasitas às tensões para as quais o nó 5 (Fig.5) tem de variar.



Grf. 9 - Resposta para a excitação individual sobreposta de três sinapses unidas no nó 5 (Fig.5). Acima da curva da corrente de saída, está representada a curva da soma das excitações pesadas de entrada, para efeitos de comparação. Todos os pesos estavam a 0.1V, pelo que a soma directa das entradas (Iin+Iin1+Iin2) é semelhante à soma das correntes das sinapses (corrente de saída).



Grf. 10 - Resposta semelhante à anterior, mas com correntes de entrada a partirem do zero. Observe-se que a variação mais lenta da corrente de saída.



Grf. 11 - Resposta do circuito com três sinapses juntas. Todos os pesos são iguais a 0.1V e as correntes de entrada partem de 1μA. Mais uma vez, a corrente de saída segue as transições da curva da soma directa das entradas (Iin+Iin1+Iin2).

IX. GERADOR DE SIGMÓIDE

Na Fig.6 está representado o circuito do gerador de sigmóide. Este circuito segue o extractor de corrente das sinapses.

Para eliminar ou minimizar o *offset* na saída, podemos alterar as dimensões de M4, M3 ou M5. A modificação que resultou numa menor distorção na linearidade deste circuito, foi a feita no transístor M4 e resumiu-se a aumentar a relação W/L de forma conveniente: L=2.4μm e W=9.6μm. Para se conseguir o factor de ganho unitário desejado, teremos de fazer com que M2 desvie a corrente de M4 mais depressa, de forma a que M3 corte mais depressa. A modificação necessária para atingir este

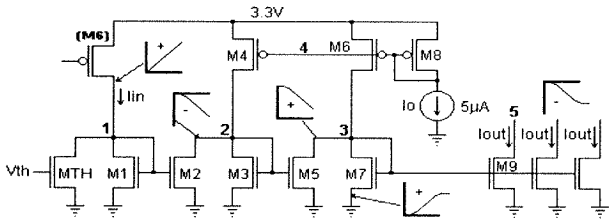


Fig. 6 - Esquema do circuito gerador de sigmóide. O transístor MTH impõe o limiar, já que só a partir duma corrente de entrada maior que a corrente em MTH, é que M1 iniciará a condução. M2 e M4 realizam a limitação superior de corrente, que terá de ser invertida por M5, M6 e M7 para se obter uma corrente a entrar em M9. No esquema estão representados os gráficos da corrente em cada linha, para a respectiva entrada de corrente a subir. Como se pode ver, as ramificações do axónio deste neurónio são feitas adicionando tantos transístores a seguir a M9 quantos se queira.

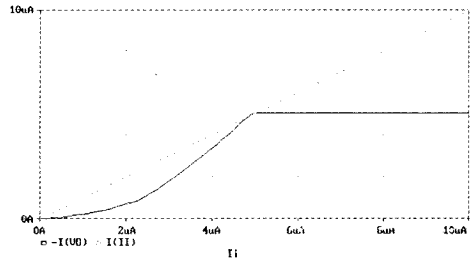
aquele objectivo foi um aumento da relação W/L de M2: L=2.4µm e W=13.2µm.

Por último, desejamos ainda corrigir a máxima corrente de saída para 5µA, tal como desejado inicialmente. Para isso, basta alterar as dimensões de M7 ou M9. Decidiu-se modificar apenas M9, que passa a ter as dimensões L=2.4µm W=6.0µm. Após estas alterações, pode-se observar o resultado final no Grf.12.

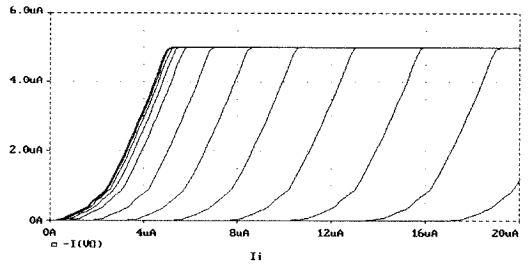
Para completar o circuito de activação do neurónio, resta apenas verificar o funcionamento do limiar de disparo realizado através do transístor MTH e a respectiva tensão de porta VTH. Esta tensão será gerada por um condensador cuja carga será actualizada por um circuito especial de controlo adaptativo deste limiar. Tudo o que MTH faz, é desviar parte da corrente de entrada, de forma a que o circuito de sigmóide seja activado com desvio da corrente de entrada, ou seja, só a partir duma certa corrente de entrada. Aplicando uma tensão variável na porta de MTH, podemos observar as translações sucessivas que a função do circuito sofre, no Grf.13.

O facto do *MOSFET* só reagir a partir de V_t não constitui qualquer problema, já o circuito que vai regular o valor da tensão VTH adaptativamente, corrigirá esse problema de forma automática. A mesma coisa acontecerá com a não-linearidade entre essa tensão e o limiar de disparo resultante.

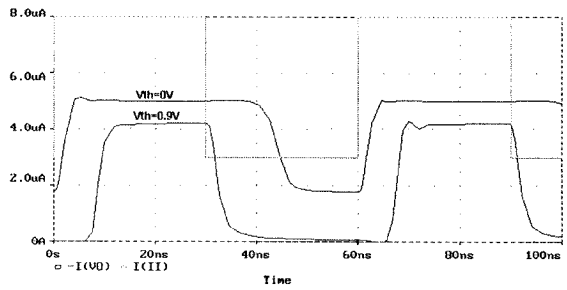
Nos gráficos Grf.14 e Grf.15 mostram-se os resultados das simulações que pretendem testar a rapidez de resposta do circuito completo.



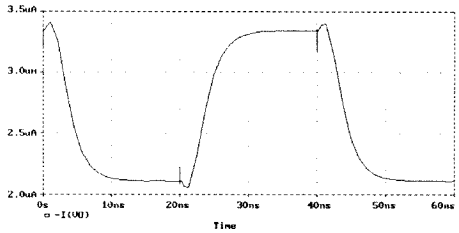
Grf. 12 - Função de transferência com todas as correcções introduzidas excepto a não-linearidade pronunciada na zona inicial da curva.



Grf. 13 - A função sigmóide sofre translações com uma relação quadrática em relação à tensão de limiar, como seria de esperar devido à característica quadrática da corrente do *MOSFET* em relação à sua tensão de porta.



Grf. 14 - Observa-se que o atraso diminui com a presença duma tensão de *threshold*. Significa que o pior caso é para limiares nulos. Os tempos medidos foram os seguintes: $T_{atraso\ subida}$=1.3ns, T_{subida}=3ns, $T_{atraso\ descida}$=7ns, $T_{descida}$=4.5ns. Isto significa que este circuito leva cerca de 4.3ns a responder a uma subida da entrada, e cerca de 11.5ns para uma descida, no pior caso. O circuito limita bem a corrente de saída a 5µA quando a de entrada ultrapassa esse valor.



Grf. 15 - Esta é a resposta à tensão do limiar de disparo. Com uma corrente de entrada de 4µA, os tempos obtidos nesta simulação são os seguintes: $T_{atraso\ subida}$=1.5ns, T_{subida}=7ns, $T_{atraso\ descida}$=1.5ns, $T_{descida}$=7ns. Estes tempos são da mesma ordem de grandeza dos observados para as respostas à entrada em corrente. De qualquer forma, não necessitariam de ser tão curtos, já que um neurónio adapta o seu limiar de forma mais lenta do que os restantes processos de adaptação

X. CONTROLO AUTOMÁTICO DO LIMIAR DE DISPARO

O controlo automático do limiar de disparo consiste em modificar a tensão de *threshold* aplicada ao transístor que desvia a corrente que entra no gerador de sigmóide. Esta modificação deverá ser tal, que origine o fenómeno da *habituação* e *recuperação* dum neurónio.

Para conseguir simular um comportamento semelhante a este, basta adicionar malhas capacitivas que acumularão a carga equivalente ao limiar de disparo, sob a forma duma tensão VTH. Na Fig.7 mostra-se o aspecto do circuito de adaptação do limiar de disparo, já com todas as dimensões.

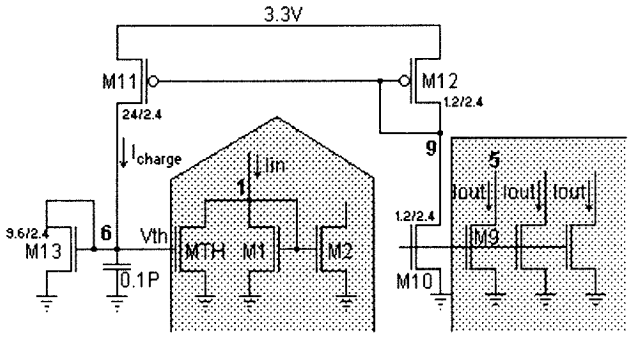


Fig. 7 - O circuito que controla automaticamente a tensão de *threshold* VTH é constituído pelos transístores M10, M11, M12 e M13, e pelo condensador de 0.1pF. M10, M11 e M12 realimentam corrente para o condensador que se carrega com maior ou menor velocidade consoante o valor da saída do neurónio detectado por M12. Saídas elevadas carregam o condensador mais rapidamente, pelo que o limiar sobe mais depressa. Se a saída for reduzida, o condensador carregará mais lentamente. O condensador também estará sempre a descarregar (muito menos) através de M13, pelo que se não houver saída este fará o limiar descer.

Por razões de economia de área, utilizou-se um condensador de baixo valor. As dimensões dos transístores M11 e M13 foram determinadas empiricamente. Para se colocar uma resistência em vez de M13, seria necessária um valor de 15MΩ, o que é inviável em VLSI, além de apresentar a desvantagem de poder descarregar totalmente o condensador.

Os valores relevantes a determinar são as variações de tensão e corrente em pontos importantes.

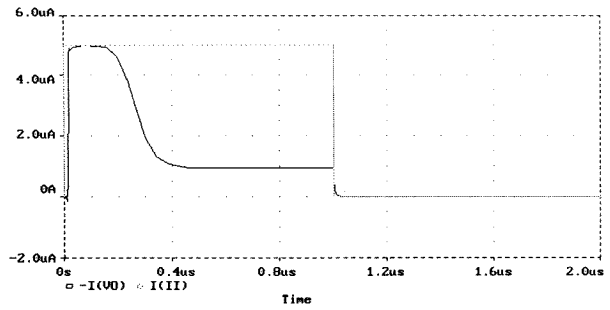
Este é apenas um exemplo. Para cada problema particular é necessário escolher convenientemente os tempos.

Duma forma geral, para alterar estes tempos de resposta do limiar, tem-se de actuar apenas nas relações (L/W) dos transístores M11 e M13.

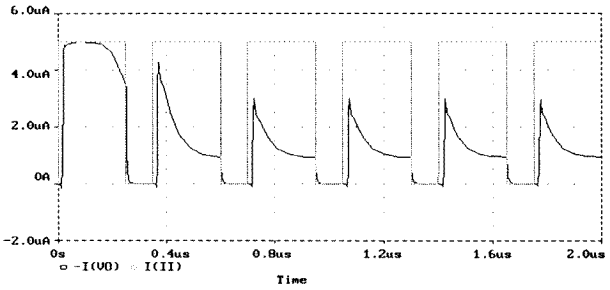
Nos gráficos Grf.16 e Grf.17 mostram-se duas situações em que o circuito de adaptação do limiar entra em acção.

XI. CIRCUITO DE APRENDIZAGEM

Este circuito implementará a *regra de Hebb modificada* que deverá obedecer ao requisitos que se seguem.



Grf. 16 - Curva da corrente de saída foi obtida com as dimensões dos transístores indicadas no texto: M11 – L=24μm, W=2.4μm; M13 – L=9.6μm, W=2.4μm. Como se pode observar pela curva da corrente de saída, esta desceu para aproximadamente 1μA após a habituação completa. Aí se manteve, até que a entrada foi a zero.



Grf. 17 - Resposta do neurónio a impulsos de corrente negativos com uma amplitude de 5μA e à frequência de cerca de 2.9MHz. Ao fim do segundo impulso, já o neurónio estava habituado, pelo que a partir daí ele mantinha a saída mais reduzida. Sempre que o impulso desaparecia, o limiar recuperava algo, até que o próximo impulso o voltava a descer.

- Coincidência de actividade na sinapse e na saída do neurónio implica uma fortificação (aumento de peso) dessa sinapse. Isto é válido para todas as sinapses.
- Soma de todos os pesos dum neurónio constante, ou seja, os pesos serão normalizados. Isto significa que a subida dum peso devida à condição anterior implica uma descida controlada de todos os outros.

Esta *regra de Hebb* é uma regra de aprendizagem não-supervisionada, e tal como já explicado no capítulo das redes neuronais biológicas, consiste simplesmente no incremento da força das sinapses através das actividades simultâneas da saída e entrada. Em termos matemáticos, isto reduz-se à equação (5), ou seja, se a entrada e saída estão activas, então o peso será aumentado em módulo e esse aumento será proporcional ao produto das duas actividades.

$$|\Delta W| = \varepsilon(X \cdot Y) \tag{5}$$

Se a saída ou entrada forem nulas, então pela lei do anulamento do produto não há modificação do peso. Assim, se o peso é positivo, continuará a crescer, e vice-versa. Por outras palavras, a *regra de Hebb* apenas reforça as sinapses, sem lhes mudar o sinal. O objectivo é permitir a existência de sinapses negativas (inibitórias). A constante é uma *constante de aprendizagem* que determina a percentagem de modificação do peso, em relação ao resultado directo da multiplicação. Esta constante é geralmente menor que a unidade (atenuadora).

Para evitar um crescimento indefinido das sinapses, poder-se-ia proceder de duas maneiras distintas:

- O valor do peso satura num máximo.
- A soma dos módulos dos pesos mantém-se constante.

Esta última forma de controlar os pesos é biologicamente plausível, além de ter certas vantagens para o reconhecimento de padrões de estímulos: sempre que um conjunto de pesos sobre, os outros descem. No final ficam apenas alguns pesos activos correspondentes ao padrão de estímulos treinado, enquanto que os restantes ficam a zero (sinapse inexistente).

Esta regra é conseguida através dum circuito multiplicador que multiplica cada entrada pela saída. O problema é a exigência de multiplicação de correntes, além deste circuito ter de ser repetido para cada sinapse. Este circuito está representado na Fig.8.

Um projecto extremamente cuidadoso nesta importantíssima parte da aprendizagem, é indispensável para se obter um comportamento correcto de uma rede neuronal construída com base nestes neurónios. Além da dependência dos restantes factores do projecto, o comportamento global desta rede depende essencialmente da correcta e eficaz concepção dos mecanismos de adaptação (aprendizagem).

XII. CONCLUSÕES E COMENTÁRIOS FINAIS

Este trabalho destinou-se a verificar a implementabilidade de um neurónio 100% analógico em modo de corrente. Esse modo de corrente foi apenas conseguido para as entradas e saídas, bem como para todas as operações internas deste neurónios, exceptuando as entradas dos pesos. Este neurónio destinar-se-ia a áreas de aplicação que exijam o máximo em velocidade de resposta duma rede neuronal. Implementações mais lentas, com menor consumo e mais *standard* poderão ser aplicadas em áreas onde a velocidade não fosse o factor determinante da *performance* dos respectivos sistemas. Aqui poder-se-iam utilizar os transístores na zona de inversão fraca, reduzindo o consumo em várias ordens de grandeza.

Com tecnologias mais modernas em que os menores tamanhos conduzem a capacidades menores, será possível alcançarem-se velocidades ainda maiores, mesmo utilizando os transístores na zona de inversão fraca. Isto pode ter um grande interesse para a implementação de redes de grandes dimensões e baixo consumo. Só assim é que conseguirá estudar o comportamento de redes grandes e utilizá-las em aplicações reais.

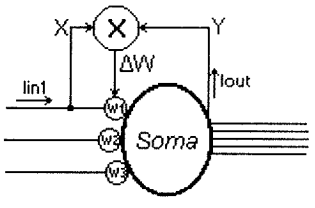


Fig. 8 - Circuito auxiliar de aprendizagem para a primeira sinapse cujo peso é W1. Este circuito multiplica a corrente de saída pela de entrada da respectiva sinapse, fornecendo um sinal de modificação do peso desta sinapse. Este sinal de modificação é proporcional àquela multiplicação, mas pode vir afectado duma constante de atenuação (constante de aprendizagem). Para cada sinapse seria necessário um circuito idêntico. Note-se que a extracção da corrente de saída exige mais um transistor por axónio, devido ao modo de funcionamento em corrente.

REFERÊNCIAS

[1] L.B.Barranco, E.S.Sinencio, A.R.Vásquez, "A CMOS analog adaptive BAM with on-chip learning and weight refreshing", IEEE Transactions on Neural Networks, Vol. 4, N.3, May 1993.

[2] J.Choi, S.H.Bang, B.Sheu, "A programmable analog VLSI neural network processor for communication receivers", IEEE Transactions on Neural Networks, Vol. 4, N.3, May 1993.

[3] J.Alspector, B.Gupta, R.B.Allen, "Performance of a stochastic learning microchip", NIPS - Neural Information Processing Systems vol. 1, Morgan Kaufmann Publ, 1989.

[4] M.Brownlow, L.Tarassenko, A.F.Murray,.A.Hamilton, H.M.Reekie, "Pulse-firing neural chips for hundreds of neurons", NIPS - Neural Information Processing Systems vol. 2, Morgan Kaufmann Publ., 1990.

[5] J.P.A.Carreira, "Circuitos com processamento de corrente para conversão analógico-digital e digital-analógico de sinais de alta frequência"; Tese de Mestrado, Instituto Superior Técnico da Universidade Técnica de Lisboa, 1993.

[6] T.D.Chiueh, R.M.Goodman, "VLSI implementation of a high-capacity neural network associative memory", NIPS - Neural Information Processing Systems vol. 2, Morgan Kaufmann Publ., 1990.

[7] S.P.Deweerth, C.A.Mead, "An analog VLSI model of adaptation in the vestibulo-ocular reflex", NIPS - Neural Information Processing Systems vol. 2, Morgan Kaufmann Publ., 1990.

[8] H.P.Graf, L.D.Jackel, "Analog electronic neural network circuits", IEEE Circuits and Devices, pp. 44-55, July 1989.

[9] P.R.Gray, R.G.Meyer, "Analysis and design of Analog Integrated Circuits", John Wiley & Sons, Inc., third edition, 1993.

[10] S.Grossberg, "Studies of mind and brain - neural principles of learning, perception, development, cognition and motor control", BSPS - Boston Studies in the Philosophy of Science vol. 70, D.Reidel Publishing Company, 1982.

[11] A.Hartstein, R.H.Koch, "A self-learning neural network", NIPS - Neural Information Processing Systems vol.1, Morgan Kaufmann Publ, 1989.

[12] J.R.Mann, S.Gilbert, "An analog self-organizing neural network chip", NIPS - Neural Information Processing Systems vol.1, Morgan Kaufmann Publ, 1989.

[13] C.Mead, "Analog VLSI and Neural Systems", Addison Wesley.

[14] J.L.Meador, C.S.Cole, "A low-power CMOS circuit which emulates temporal electrical properties of neurons", NIPS - Neural Information Processing Systems vol.1, Morgan Kaufmann Publ, 1989.

[15] A.Moopenn, T.Duong, A.P.Thakoor, "Digital-Analog hybrid synapse chips for electronic neural networks", NIPS - Neural Information Processing Systems vol. 2, Morgan Kaufmann Publ, 1990.

[16] P.Mueller, J.V.Spiegel, D.Blackman, T.Chiu, T.Clare, J.Dao, C.Donham, T.Hsieh, M.Loinaz, "A programmable analog neural computer and simulator", NIPS - Neural Information Processing Systems vol.1, Morgan Kaufmann Publ, 1989.

[17] A.F.Murray, A.Hamilton, L.Tarassenko, "Programmable analog pulse-firing", NIPS - Neural Information Processing Systems vol.1, Morgan Kaufmann Publ, 1989.

[18] B.Nabet, R.B.Darling, R.B.Pinter, "Analog implementation of shunting neural networks", NIPS - Neural Information Processing Systems vol.1, Morgan Kaufmann Publ, 1989.

[19] S.Ryckebusch, J.M.Bower, C.A.Mead, "Modeling small oscillating biological networks in analog VLSI", NIPS - Neural Information Processing Systems vol.1, Morgan Kaufmann Publ, 1989.

[20] J.L.Ryckebusch, M.A.Mahowald, C.A.Mead, "Winner-take-all networks of O(N) complexity", NIPS - Neural Information Processing Systems vol.1, Morgan Kaufmann Publ, 1989.

[21] S.Satyanarayana, Y.Tsividis, "A reconfigurable analog VLSI neural network chip", NIPS - Neural Information Processing Systems vol. 2, Morgan Kaufmann Publ, 1990.