Red Neuronal de Hopfield con técnicas de procesamiento estocástico paralelo–secuencial

F.Colodro, A.Torralba y L.G.Franquelo
Dpto. de Ingeniería Electrónica, de Sistemas y Automática
Escuela Superior de Ingenieros
Avda. Reina Mercedes, s/n, SEVILLA–41012
Tlf: (95) 4556851, Fax: (95) 4556849
e–mail: pcolr@gte.esi.us.es

Abstract— **En este artículo se presenta la realización de una Red Neuronal Estocástica de Hopfield (SHNN) con un gran número de unidades. Originalmente , la SHNN propuesta por van de Bout en [1], requiere tiempos de convergencia grandes al acumularse en el estado neuronal los pulsos estocásticos que codifican las sinapsis secuencialmente. En otras realizaciones ([2]–[3]) los pulsos sinápticos son acumulados en paralelo pero encuentran limitado el número máximo de neuronas de la red en arquitecturas multichip a 100 unidades aproximadamente. Este hecho se debe a las limitaciones que hoy por hoy la_tecnología impone en el nñero de pines de I/O en los circuitos integrados. A continuación se propone una realizaclón multichip que permite establecer un compromiso entre los problemas planteados anteriormente. La arquitectura, basada en la utilización de una estrategia mixta paralelo–secuencial, reduce el número de de líneas de interconexión al valor de k y aumenta la velocidad de convergencia respecto a la SHNN secuencial por un factor de k. Para evaluar el comportamiento de la red se ha simulado y resuelto un problema de partición.**

I. Introducción

Las Redes Neuronales de Hopfield (HNN), caracterizadas por estar constituidas por una única capa completamente conectada, han demostrado ser adecuadas para la resolución de problemas de optimización. Diferentes tecnologias analógicas y digitales han sido utilizadas en su realización física. Una técnica utilizada recientemente es la lógica estocástica, que presenta algunas ventajas respecto de realizaciones anteriores, siendo la más significativa, la realización de la operación producto mediante una simple puesta AND.

En [1] se presenta la Red Neuronal Estocástica de Hopfield (SHNN) cuya arquitectura es reproducida en la figura 1. Dicho circuito resuelve la ecuación discretizada de carga de la HNN:

$$u_i(t+\delta t) = u_i(t) + (\sum_{j=0}^{n-1} G_{ij}v_j(t) + I_i) \times \delta t \quad (1)$$

$$u_i(t) = f(u_i(t)) \quad i = 0, 1, ..., n-1 \quad (2)$$

La señal binaria s_{ij}, que pulsa con probabilidad proporcional al producto $G_{ij}v_j$, ha sido obtenida multiplicando mediante una puerta AND las señales estocásticas que codifican los valores G_{ij} y v_j. Por la simplicidad de la realización del producto se podrían calcular todos los productos $G_{ij}v_j$ de la red simultáneamente y reducir drásticamente el tiempo de convergencia. No obstante, las neuronas del circuito de la figura 1 acumulan una synapsis $G_{ij}v_j$ por ciclo de reloj debido a la dificultad de realizar la suma en paralelo de las señales estocásticas. Por lo tanto, n ciclos de reloj son requeridos en la evaluación de $\sum_{j=0}^{n-1} G_{ij}v_j(t) \times \delta t$.

Para reducir el tiempo de convergencia de la red, el circuito digital (circuito F) propuesto en [2]–[3] es usado. El circuito permite realizar la suma de los pulsos synápticos s_{ij} en paralelo y acumularlos en el estado neuronal i–ésimo en un sólo ciclo de reloj. Esta realización reduce el tiempo de convergencia de la SHNN paralela por un factor de n. Notamos que para la conexión de las neuronas de la red sólo es necesario realimentar las señales estocásticas que pulsan con probabilidad proporcional al estado neuronal. Por lo tanto, en una realización multichip de n neuronas con acumulación completamente paralela de las ssinapsis, necesitaría n pines I/O para la comunicación entre chips. Con las presentes limitaciones tecnológicas, redes soportadas en arquitecturas multichip no podrían ser realizadas con más de 100 neuronas. Este tipo de redes son pequeñas

comparadas con los problemas de optimización encontrados en aplicaciones reales.

Una SHNN con técnicas mixtas de procesamiento paralelo y secuencial es propuesta en este artículo para la realización de redes con un gran número de neuronas. La estrategia presentada es un compromiso entre el tiempo de convergencia, el coste en área de silicio y la factibilidad de la realización.

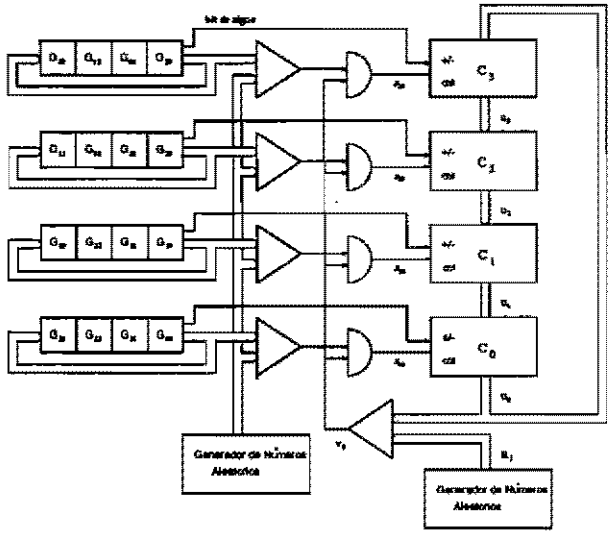

Figure 1: Red Neuronal Estocástica de Hopfield (SHNN).

II. El circuito F

El problema principal para calcular $\sum_{j=0}^{n-1} G_{ij} v_j$ usando lógica estacástica es la suma. A continuación citamos algunas de las soluciones que se pueden encontrar en la literatura:

- En [1] el autor realiza la suma por multiplexación temporal de los pulsos estacásticos.
- En [6], un circuito analógico es propuesto.
- En [7] la secuencia temporal de los pulsos estocásticos es transformada por medio de una función exponencial. Por tanto, la operación suma es reemplazada por la operación producto que se realiza mediante puertas lógicas ANDs.

En [2]–[3] los autores proponen el uso de un circuito digital (circuito F) que permite la realización de la suma paralela de los pulsos sinápticos. El circuito F es una red combinacional que recibe n bits de igual peso como entrada y genera como salida una palabra de $(d = log_2 n)$ bits que codifica en binario la suma de los bits de entrada. Este tipo de circuito fue llamado contador (n,d) por Dadda [8]. La implementación directa del circuito F usando lógica de dos niveles, tal como una PLA, supondría una complejidad combinacionaa al número n de pulsos sinápticos, haciendo esta opción impracticable incluso para valores no muy grandes de n. En [3] dos realizaciones del circuito F con complejidad $O(n)$ son propuestos. La realización del circuito llamado paralelo usa sumadores totales (figuras 2.*a* y 2.*b*) requiere un único ciclo de reloj para calcular las sumas:

$$F^{+} = \sum_{G_{ij}v_j>0} G_{ij} v_j \tag{3}$$

$$F^{-} = \sum_{G_{ij}v_j<0} G_{ij} v_j \tag{4}$$

El tiempo de retardo máximo y el número de sumadores totales que el circuito F requiere para distintas longitudes de la palabra de entrada son mostrados en la tabla 1.

Tabla 1

n	n^o de sumadores	retardo máximo (puertas de 2 niveles)
3	1	1
7	4	3
15	11	5
31	26	7
63	57	9
127	120	10

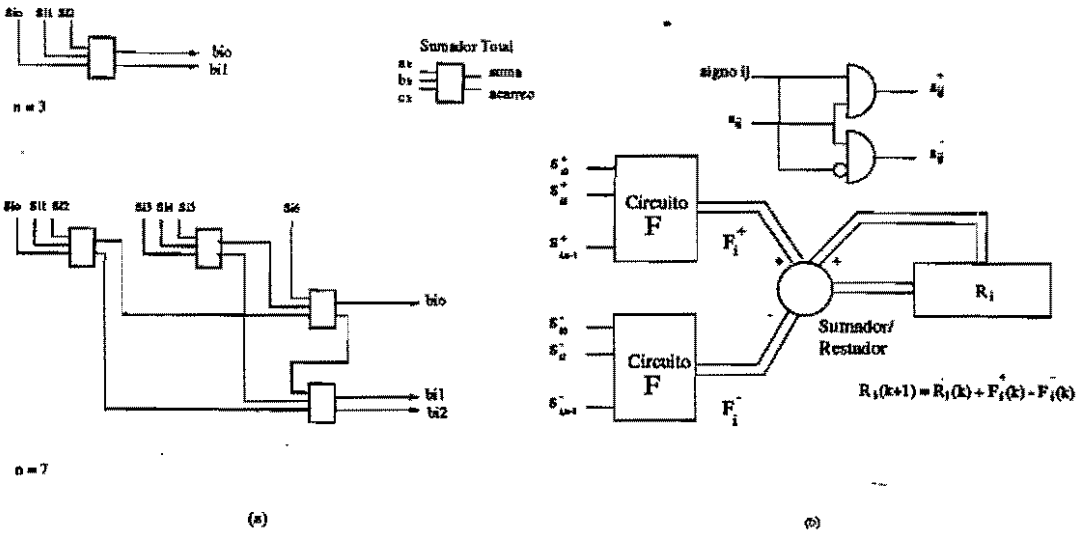

Figure 2: a)Circuito F usando sumadores totales. b) Neurona de una SHNN con acumulación en paralelo de los pulsos sinápticos.

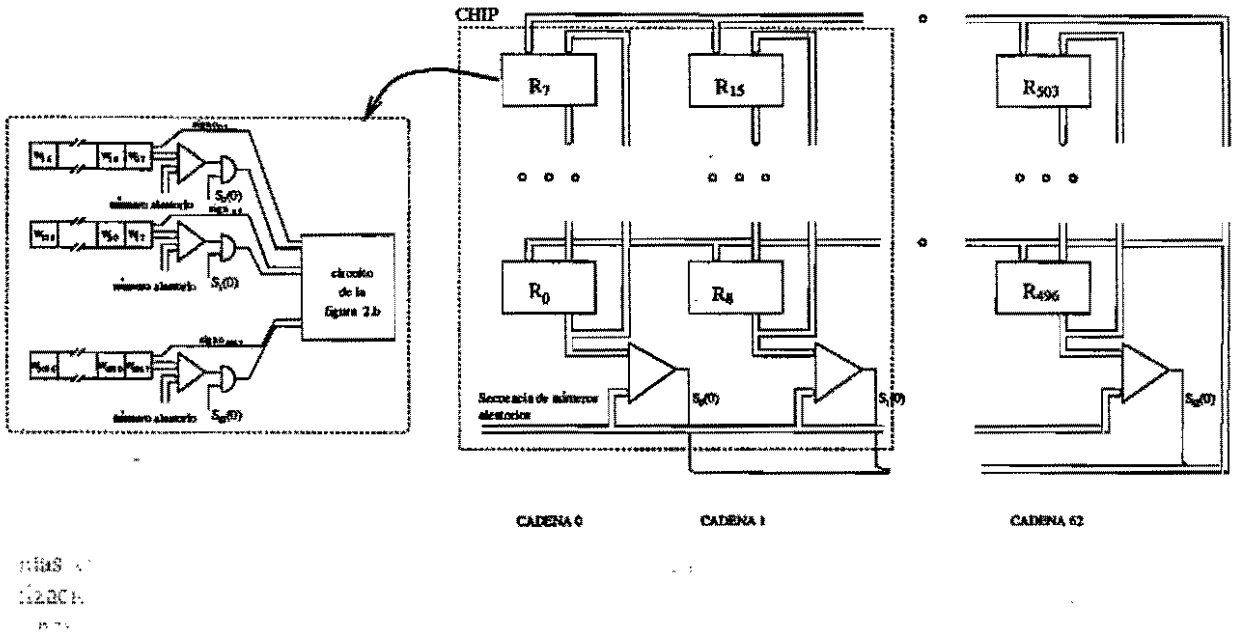

Figure 3: Realización de una SHNN mixta paralelo–secuencial con un gran número de unidades.

III. SHNN CON PROCESAMIENTO MIXTO PARALELO–SECUENCIAL

Supongamos que queremos realizar una red Hopfield para resolver un problema de partición de 504 nodos. La realización secuencial de la red propuesta por van der Bout requeriría 504 ciclos de reloj para acumular la contribución $\sum_j G_{ij} v_j \times \delta t$ al estado neuronal.

Por otra parte, si se usará los circuitos propuestos en las figuras 2.*a* y 2.*b*, un único ciclo de reloj se requeriría (o dos si los pulsos positivos o negativos fueran acumulados en diferentes ciclos de reloj). El tiempo de convergencia de la red sería mejorado por un factor de 504. No obstante, la realización multichip de tal red neuronal requeriríria más de 504 pins de entrada y salida por chip.

Una nueva red mixta paralelo–secuencial es propuesta en este articulo para resolver los problemas que plantean las arquitecturas completamente secuenciales o paralelas. Como ejemplo tomaremos la red de 504 neuronas propuesta anteriormente. La nueva arquitectura está constituida por 63 cadenas de 8 neuronas cada una (figura 3). Los registros (R_0,R_1,...,R_{503}) contienen los estados neuronales (u_0,u_1,...,u_{503}). Por claridad en la descripción del circuito, en la figura 3 no se muestran las unidades de almacenamiento de las sinapsis. La entrada de offset I_i puede ser modelada como el peso $G_{n+1,i}$ de la neurona $(n+1)$–ésima, cuya salida es forzada a permanecer constantemente a 1. Una posible realización multichip localizaría 2 cadenas en cada chip, como se muestra en la figura 3. Se hace notar que solamente 64 pines externos de I/O son necesarios en cada chip para la interconexión de neuronas.

Para evaluar el comportamiento de la SHNN con procesamiento mixto paralelo–secuencial en casos reales, un prototipo está actualmente siendo diseñado con 504 neuronas en la tecnología CMOS 1.0μ. La red será implementada como una arquitectura multichip. El chip tiene 16 neuronas y la estructura de una neurona está mostrada en la figura 2. En el estado actual de diseño, la unidad que conforma a la neurona está concluida, habiendose aprovechado el material existente de un trabajo previo, la realización de una red estocástica de Hopfield completamente paralela [3]. En el momento actual, estamos trabajando en la estrategia de secuenciamiento para la obtención de la red mixta paralelo–secuencial, en el nuevo dimensionamiento del chip y en la sustitución de las unidades de almacenamiento de las sinapsis (en la red completamente paralela se usaron biestables *flip-flop*) por memorias RAM, las cuales son macroceldas muy optimizadas que el fabricante suministra, con vistas a reducir el área de silicio. En la siguiente sección se comentan los tiempos de respuesta de la nueva red, obtenidos por simulación, y una estima preliminar del área ocupada por la neurona.

Una vez inicializados, los contenidos de los registros de las cadenas son continuamente desplazados cíclicamente, tal que el registro del fondo de la cadena contiene el estado neuronal $k_{mod(N)} + m \times N$ en el ciclo k, donde N es el número de estados por cadena (8 en este ejemplo). Este valor es comparado con un número aleatorio para generar la señal estocástica $S_m(k)$. $S_m(k)$ es realimentado al resto de neuronas y multiplicado por el peso w_{ij} para actualizar el contenido de R_r, donde i y j son

$$i = k_{mod(N)} + m \times N \tag{5}$$

$$j = (k+r)_{mod(N)} + \lceil \frac{r}{N} \rceil \times N \tag{6}$$

Los pesos sinápticos en el chip están organizados en cadenas que se desplazan cíclicamente. Para el ejemplo anterior hay 63 cadenas por estado neuronal de 8 pesos cada una.

Debido al carácter estocástico de las señales , el circuito propuesto sigue la dinámica de la ecuación de carga (1), dejando que transcurra un número considerable de ciclos de reloj. La función de transferencia neuronal f puede ser modificada ajustando la función de distribución de la secuencia aleatoria de números.

Para el caso de 504 neuronas, la red propuesta puede actualizar cada neurona en $O(8)$ ciclos de reloj. La arquitectura de la SHNN con procesamiento mixto paralelo–secuencial es una solución prometedora para construir grandes redes con un reducido tiempo de convergencia a un costo razonable.

IV. RESULTADOS

La dinámica de las dos neuronas más lentas es mostrada en la figura 4 cuando la red es usada para partición de grafos de 504 nodos. Los resultados han sido obtenidos por simulación. Todas las neuronas fueron inicializadas con números aleatorios en el rango de [-20..20]. Notamos

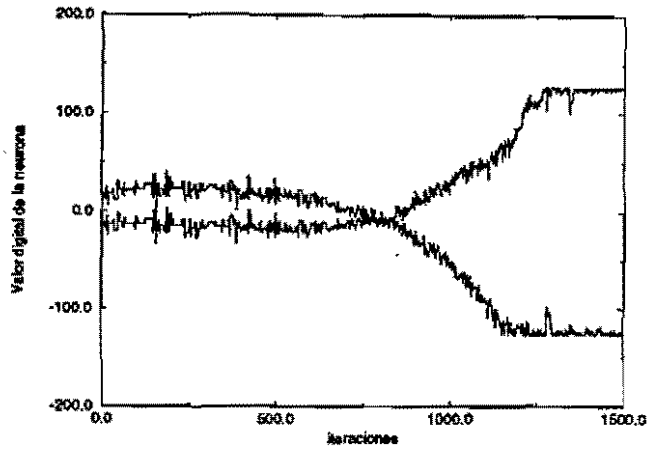

Figure 4: Resultados de simulación: Evolución temporal de las dos neuronas más lentas del ejemplo propuesto.

que sólo 1200 ciclos de reloj son necesarios para converger. La arquitectura de [1], basada en procesamiento secuencial, requeriría 64×1200 ciclos en el mismo caso.

El área estimada para una neurona es $1.2mm^2$ usando la tecnología CMOS de $1.0\mu m$. El área reservada para el almacenaje de los pesos sinápticos es de $12.5mm^2$. Por tanto, un chip de 16 neuronas (figura 3) ocupará aproximadamente $82mm^2$.

V. Conclusiones

Una estrategia para la realización hardware de SHNNs con un gran número de neuronas ha sido presentada. Lógica estocástica fue usada por eficiencia de área. Arquitecturas precedentes estuvieron limitadas a grandes tiempos de convergencia y/o limitaciones en el número de pines I/O. La SHNN con procesamiento mixto paralelo–secuencial ha sido propuesta para encontrar un compromiso entre eficiencia de área, número de pines I/O y tiempo de convergencia. La limitación más importante de esta arquitectura es el área consumida para el almacenaje de los pesos sinápticos, dificultad compartida por todas las arquitecturas de redes neuronales. Una red de n neuronas puede ser diseñada en una estructura multichip con el circuito propuesto, si el número de líneas de interconexión entre chips es N, $\frac{n}{N}$ ciclos de reloj son consumidos en la suma de la euación (1).

References

[1] D.E. van den Bout and T.K.Miller III, "A digital architecture employing stochasticism for the simulation of Hopfield neural nets". *IEEE Trans. Circuits and Systems*, vol. 36, pp. 732–738, May 1989.

[2] A.Torralba, F.Colodro. "Towards a fully parallel Stochastic Hopfield Neural Network". *Proc. of the ISCAS'93*, pp. 2741–2743, May 1993.

[3] A.Torralba, F.Colodro. "Two digital circuits for a Fully Parallel Stochastic Neural Network". IEEE. *Trans. of Neural Network*, Sep 1995.

[4] D.E. van den Bout and T.K.Miller III, "TInMANN: The Integer Markovian Artificial Neural Network".

[5] M.S.Melton, T.Phan, D.S.Reeves, D.E. van den Bout, "The TInMANN VLSI chip". *IEEE Trans. Neural Networks*, vol.3, no. 3, May 1992.

[6] Y.Kondo and Y.Sawada, "Functional abilities of a stochastic logic neural network". *IEEE Trans. Neural Networks*, vol. 3, no. 3, May 1992.

[7] C.Janer and J.M.Quero, "Fully parallel summation in a new Stochastic Neural Network architecture". *IEEE Trans. Int. Conf. in Neural Network*, San Francisco, 1993.

[8] L.Dadda. "Some schemes for parallel multipliers". *Alta Freq.*, vol. 19, pp. 349–356, May 1965.