Department of Electronic Engineering, King's College London.

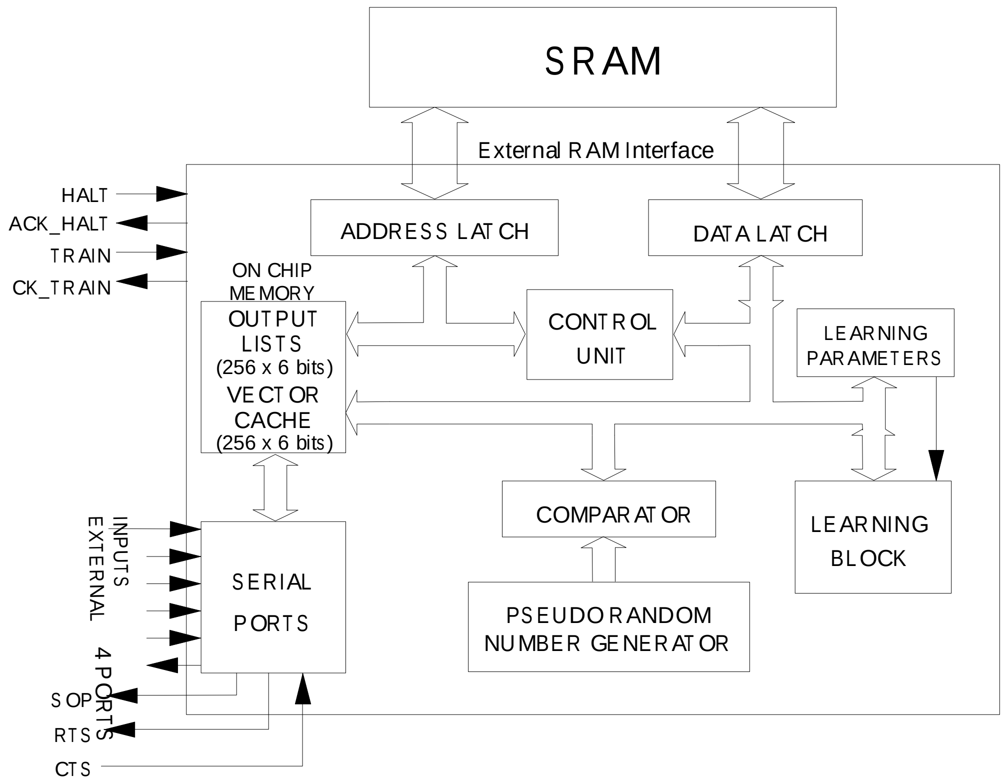
pRAM-256 VLSI Neural Network Processor
pRAM artificial neuron with learning in hardware

General Description

The *p*RAM-256 is a versatile neural network processor with an on-chip learning unit. It offers the flexibility of a software solution with the speed of hardware. Connections between the *p*RAM neurons are reconfigurable which allows a network's architecture to be modified at any time. The *p*RAM-256 can complete one pass of the training process, training all 256 *p*RAMs, in less than 0.25 ms when operating at the maximum clock speed of 33 MHz. Because of the high number of *p*RAMs supported by the *p*RAM-256, a typical neural network can be built using a single *p*RAM Module. Several *p*RAM Modules can operate in parallel so that larger networks can be built. The *p*RAM-256 is fabricated using an advanced sub-micron gate array semi-custom technology from GEC Plessey Semiconductors. The use of a 68 pin PGA package allows a compact neural network to be built into existing and future systems. Interfaces to EISA and VME bus systems have been defined.

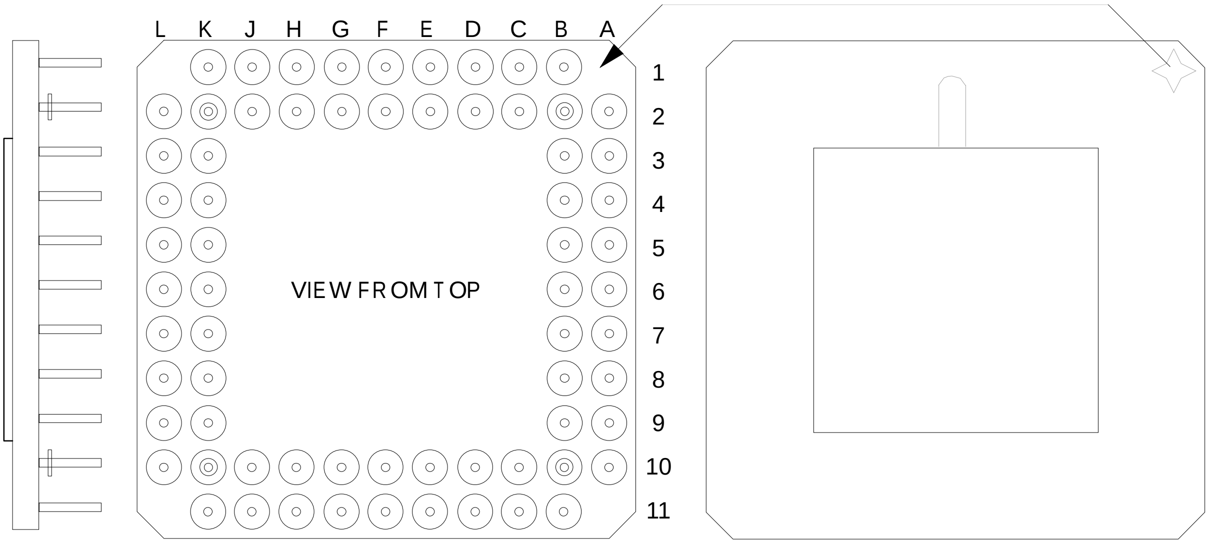
Features

- ☑ 256 *p*RAMs, each with 6 inputs
- ☑ Configurable connections between *p*RAMs
- ☑ On-chip Reinforcement Learning Unit
- ☑ Learning can be Global, Local or Competitive within the same unit
- ☑ A non-learning cycle for all 256 *p*RAMs takes 5120 cycles: 0.154ms at 33 MHz
- ☑ A learning cycle for all 256 *p*RAMs takes 8192 cycles: 0.246ms at 33 MHz
- ☑ External static RAM used for efficient weight storage



Architecture of the pRAM-256

Pin Connections



68 pin PGA package

A2	WE	B9	PENALTY	F10	NC	K4	A4
A3	RRW	B10	TRAIN	F11	NC	K5	A5
A4	RTS	B11	ACK_TRAIN	G1	D10	K6	A7
A5	CTS	C1	D3	G2	D9	K7	A9
A6	NORTH	C2	D2	G10	NC	K8	GND
A7	SOUTH	C10	NC	G11	NC	K9	A12
A8	V_{dd}	C11	HALT	H1	D11	K10	A15
A9	REWARD	D1	V_{dd}	H2	GND	K11	ACK_HALT
A10	CLK	D2	D4	H10	NC	L2	A1
B1	D1	D10	GND	H11	V_{dd}	L3	A3
B2	D0	D11	RESET	J1	D13	L4	V_{dd}
B3	SOP1	E1	D6	J2	D12	L5	A6
B4	GND	E2	D5	J10	NC	L6	A8
B5	PRAM_OUT	E10	NC	J11	NC	L7	A10
B6	EAST	E11	NC	K1	D14	L8	A11
B7	WEST	F1	D8	K2	D15	L9	A13
B8	EXT	F2	D7	K3	A2	L10	A14

Table of pin assignment

The CMOS process used to fabricate the *p*RAM-256 is fully static. Therefore the device may be halted and operated in a standby mode with no loss of data. The *p*RAM-256 may also be operated at supply voltages below 5V (e.g. 3V) with a much lower power consumption but with a reduced operating speed.

Pin Descriptions

Name	Type	Descriptions
D0 to D15	BI-DIR	Bi-directional data bus connected to external SRAM and controlled by WE. During normal operation, connection pointers and memory contents (α) are transferred from the SRAM to the *p*RAM-256 using this bus. When HALT is active and a valid address is set, data may be transferred to the internal registers at the rising edge of RRW.
A1 to A15	TRI (A1,A2 - BI-DIR)	Address bus to control the transfer of data to or from the *p*RAM-256 from the external SRAM. A3 to A15 are outputs which are tri-stated when HALT is active. A1 and A2 are normally outputs but when HALT is active, these become inputs which are used to address the internal registers .
WE	TRI	This is the write enable signal from the *p*RAM-256 to external SRAM It is tri-stated when HALT is active. Active LOW.
SOP1	O/P	**Start Of *p*RAM 1**: This signal is set HIGH by the *p*RAM-256 to indicate the start of the first *p*RAM process. It is set LOW at the end of the first *p*RAM process.
RTS	OPEN DRAIN	**Ready To Start**: This signal synchronises the inter-module communications and indicates that the processing of one of the 256 *p*RAMs has been completed. It should be connected to the CTS and the RTS pins of the other *p*RAM-256 modules (if present). An open drain gate is used to simplify the connection. If a single *p*RAM-256 is used, this output may be ignored. RTS goes LOW when each of the 256 *p*RAM processes starts and goes HIGH when the process is completed.
CTS	I/P	**Clear to Start**: This signal tells the *p*RAM-256 that all other modules are ready to begin processing the next *p*RAM. It is active HIGH. If a single *p*RAM-256 is used, this pin should be tied HIGH or to RTS.
pRAM_OUT	O/P	This is the output of the *p*RAM currently being processed. The data is valid as soon as RTS goes HIGH. It should be connected to a serial input (NORTH, SOUTH, EAST or WEST) of another module (if present), or this signal may be read by external circuitry. 256 bits of data are output between active transitions of SOP1.
NORTH	I/P	Serial input which accepts the *p*RAM outputs from another module. Data on the serial input is latched into on-chip memory at the trailing edge of CTS.
EAST SOUTH WEST	I/P	As NORTH. Unused inputs may be used for external inputs as EXT.
EXT	I/P	Serial input for external inputs. It can accept up to 256 bits of data in one SOP1 frame. Data is latched at the trailing edge of CTS.
HALT	I/P	Used to halt the module at the end of the current pass. It is active LOW.
ACK_HALT	O/P	This signal acknowledges the halt request. If training is disabled, ACK_HALT is set at the end of PASS 1. However, if training is enabled, ACK_HALT is set at the end of PASS 2. It is active HIGH. When halted, all control and data lines except A1 and A2 are tri-stated to allow an external device to access the SRAM or to write to the *p*RAM-256 internal registers.
TRAIN	I/P	Training is enabled at the end of the current pass, when TRAIN is set to HIGH.
ACK_TRAIN	O/P	This signal acknowledges the enable training request. This is set at the end of PASS 1. It is active HIGH.
RESET	I/P	Master reset. Active LOW.
CLK	I/P	Clock signal, maximum 33MHz, CMOS levels.
REWARD PENALTY	I/P	External environment inputs used during reinforcement training. These must be held in a constant state by external circuitry during PASS 2. Active HIGH.
RRW	I/P	This is the write enable control of the three internal registers, ρ, ρλ and FBPL. ρ and ρλ are the learning rate and decay rate used by the on-chip learning unit. FBPL controls the selection of the feedback polynomial of the pseudo random number generator. The module must be halted before data can be transferred into these registers. ρ or ρλ are written when RRW is high, FBPL is written when RRW is low; the appropriate register address must be present on the address bus (A1-A2).

Internal register addresses

Three internal registers are implemented inside the chip. Two of these registers, ρ and ρλ, are for the learning and decay rates of the on-chip learning unit. The third register, FBPL, selects the feedback polynomial of the pseudo random number generator. These registers are 16-bit write-only registers. They can be accessed by first halting the chip and, when HALT_ACK is TRUE, asserting the corresponding address onto the address bus. Data on the data bus will be transferred to the selected register on the rising edge of the RRW signal (ρ and ρλ) or when RRW is low, for the FBPL register. When A1:A2 = 11 RRW can be either state.

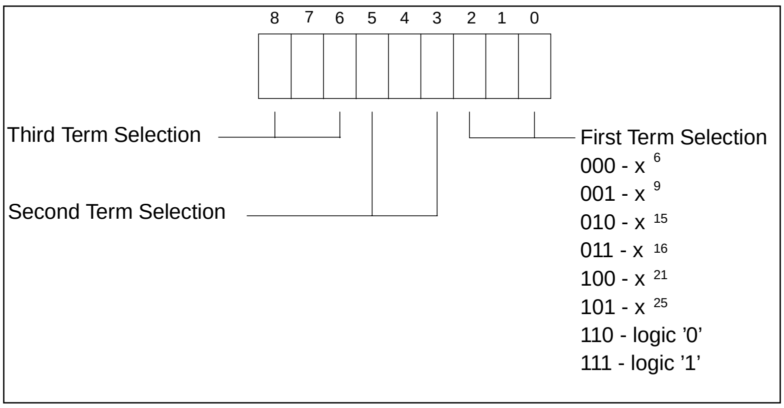
Register	A1	A2	Function
ρ	0	0	Learning rate
ρλ	1	0	Decay rate
FBPL	0	1	Feedback polynomial selection

Addresses of the internal registers

The feedback polynomial comprises five terms which are selected from the following tap points: $\{x^6, x^9, x^{15}, x^{16}, x^{21}, x^{25}\}$. This set of tap points has been selected to provide the highest number of irreducible polynomials. The possible polynomials are listed in the following table. Since the x^0 and x^{31} terms must be included in all irreducible polynomials, only three terms need be selected. The selection is achieved by writing to the FBPL register. Three bits are required for the selection of each term, therefore a 9-bit number is required to specify the feedback polynomial selected. A '1' will be injected into the shift register after a RESET for the purpose of auto-starting the random number generator.

$x^0 + x^6 + x^9 + x^{15} + x^{31}$	0x088	$x^0 + x^9 + x^{15} + x^{21} + x^{31}$	0x111	$x^0 + x^6 + x^{16} + x^{25} + x^{31}$	0x158
$x^0 + x^6 + x^9 + x^{16} + x^{31}$	0x0C8	$x^0 + x^6 + x^{16} + x^{21} + x^{31}$	0x118	$x^0 + x^9 + x^{16} + x^{25} + x^{31}$	0x159
$x^0 + x^6 + x^9 + x^{21} + x^{31}$	0x108	$x^0 + x^9 + x^{16} + x^{21} + x^{31}$	0x119	$x^0 + x^{15} + x^{16} + x^{21} + x^{31}$	0x11A
$x^0 + x^6 + x^{15} + x^{16} + x^{31}$	0x0B0	$x^0 + x^6 + x^{15} + x^{25} + x^{31}$	0x150	$x^0 + x^{15} + x^{16} + x^{25} + x^{31}$	0x15A
$x^0 + x^6 + x^{15} + x^{21} + x^{31}$	0x110	$x^0 + x^9 + x^{15} + x^{25} + x^{31}$	0x151		

The available irreducible feedback polynomials

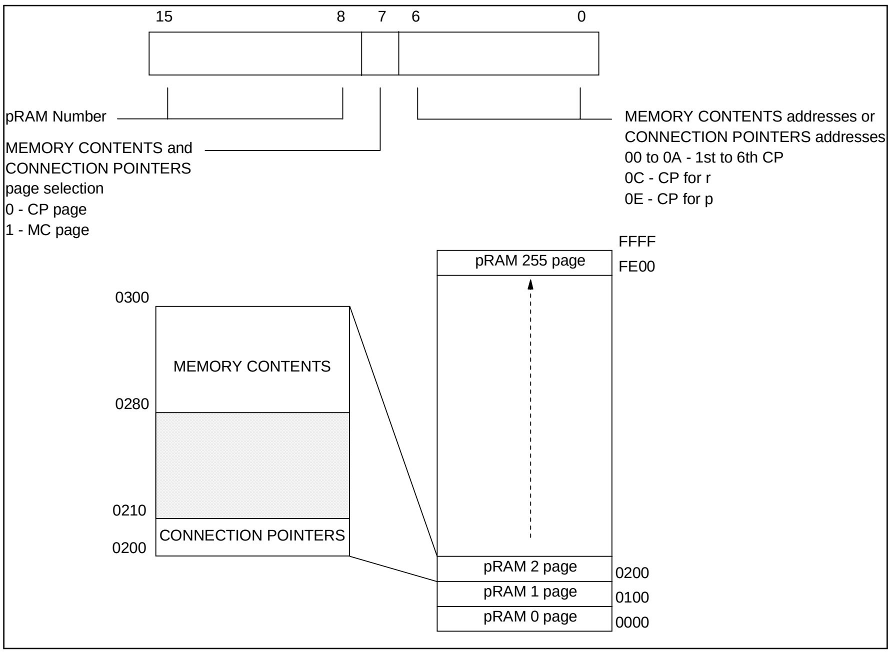


Definition of the FBPL register

Memory mapping

The memory address space requires minimal address decoding. It is divided into 512 pages with a page size of 128 words (256 bytes) and the address bus is 16 bits wide. The *p*RAM number within the *p*RAM-256 defines the high byte of the address. The low byte is used to address the weights ($\alpha_{\underline{u}}$) or the connection pointers of that *p*RAM. Bit 7 of the low byte is used to select the page for $\alpha_{\underline{u}}$ (when '1') or the page for connection pointers (when '0'). The memory map of the external memory and the definition of the RAM address are shown in the following figure. The first Connection Pointer (CP) to be fetched is at the lowest address and defines the most-significant bit of the weight address ($\alpha_{\underline{u}}$).

High speed static RAM is desirable for the external memory. The access time of the SRAM should be less than 25ns when the chip is clocked at 33MHz.



Address definition and External memory management

Defining the Network Architecture

One of the most important features of the Serial Update Architecture is the use of reconfigurable interconnections between the *p*RAMs. Every input of the *p*RAM is associated with a Connection Pointer. A Connection Pointer defines the routing by which a *p*RAM's input is connected to the output of another *p*RAM or to an external input. Thus, the network architecture is described by a Connection Table which could be in ROM for a stand-alone network, or in RAM to allow a host computer to reconfigure the network at any time. The Connection Table can be written in such a way that a variety of network topologies can be built using the same hardware. The concept of a Connection Pointer is

extended to the reward and penalty inputs to enable both Global and Local learning strategies in the on-chip learning scheme.
A Connection Pointer is a 12 bit binary number which defines the device and the internal address of the data source. The definition of a Connection Pointer is shown in the following table. The Device number (bits 8-11) defines the source of the data for each pRAM input and is one of the following: within the *p*RAM-256 Module itself, in an adjacent pRAM-256 module, the External Input, Vcc or GND. Vcc and GND allow a constant '1' or '0' to be presented to any pRAM input.
The Internal Address specifies the *p*RAM number within the module if the Device is set to be a *p*RAM-256 Module or the external input number if the Device is set to be an External Input. Where Vcc or GND are specified, bits 7-0 are not used.

BIT	11 - 8	7 - 0
Function	Device location	*p*RAM number or External input number

Definition of a Connection Pointer

bit				Data source
11	10	9	8	
0	0	0	0	Local Chip
0	0	0	1	North Chip
0	0	1	0	East Chip
0	0	1	1	South Chip
0	1	0	0	West Chip
0	1	0	1	GND
0	1	1	0	VCC
0	1	1	1	Global REWARD
1	0	0	0	Global PENALTY
1	0	0	1	Negated local chip data
1	0	1	0	Negated north chip data
1	0	1	1	Negated east chip data
1	1	0	0	Negated south chip data
1	1	0	1	Negated west chip data
1	1	1	0	External input
1	1	1	1	Negated external input

Definition of data source in connection pointer

Functional Description

The *p*RAM-256 processes a network of 256 *p*RAM neurons and performs training of the *p*RAM weights in a single package. It supports multi-module operations, therefore applications which require large-scale neural networks can be implemented by the use of a number of *p*RAM-256 devices.

Each *p*RAM-256 has 256 *p*RAMs which are processed serially at high speed.

The operation of a *p*RAM Module is performed in two passes, PASS1 and PASS2.

PASS1: is a process which calculates the new outputs for the 256 *p*RAMs. This requires 256 non-learning *p*RAM cycles.

PASS2: is an optional process which updates the weights of the 256 *p*RAMs. These weights are stored in external SRAM. This pass uses 256 learning *p*RAM cycles. PASS2 may be omitted if training is not required, in which case TRAIN is LOW. Faster *p*RAM processing is possible when training is disabled.

Non-learning *p*RAM cycle: This process calculates the new *p*RAM outputs. In a non-learning cycle, the *p*RAM-256 fetches the six Connection Pointers from the SRAM and decodes them to generate an Input Vector. The Input Vector then forms part of the address used by the *p*RAM-256 to fetch the selected weight (α) from the SRAM. Finally, the *p*RAM-256 executes the *p*RAM algorithm by comparing α to a random number and storing the result (1 or 0) in the internal Output List. The result is broadcast at the same time to the other modules from *p*RAM_OUT. The Input Vector formed in this cycle is saved in an on-chip cache memory for PASS2 (if enabled).

Learning *p*RAM cycle: This process updates a *p*RAM weight. In a learning cycle, the *p*RAM-256 fetches the Connection Pointers for r and p from the SRAM and decodes them to generate the reward and penalty environment signals for the on-chip learning unit. By using the Input Vector, which was generated in the corresponding non-learning *p*RAM cycle above, α is fetched from the SRAM and updated according to the learning rule. The updated α is then saved in the external SRAM.

Connection Pointer: A 12 bit binary number which specifies the source of data for *p*RAM inputs and the *p*RAM reward and penalty inputs. The Connection Pointer table is held in the external SRAM; this table must be defined before *p*RAM processing starts and may be redefined at any time by asserting the HALT input and waiting for HALT_ACK.

Design Engineer: Dr Chi-Kwong Ng
Technical enquiries: Professor T G Clarkson, Department of Electronic Engineering, King's College London, Strand, London WC2R 2LS, Tel: 020 7848 2367, Fax: 020 7848 2559, Email: tgc@kcl.ac.uk