

NeuroMatrix® NM6403 DSP with Vector/Matrix Engine

Dmitri Fomine^a, Vladimir Tchernikov^a, Pavel Vixne^a and Pavel Chevtchenko^a

^aResearch Center MODULE, 3 Eight March 4th Street, Box 166, Moscow, 125190, Russia,
tel. +7-095-152-9335, fax. +7-095-152-3168, e-mail: dfomine@module.ru

ABSTRACT

The paper describes the architecture of the NeuroMatrix® NM6403 DSP designed for image processing, signal processing and neural networks emulation [1,2]. The paper includes a brief description of the processor structure and its instruction set. The NM6403 is the first DSP based on NeuroMatrix® Core (NMC) comprises an original 32-bit VLIW RISC processor and a 64-bit SIMD Vector co-processor (VCP). In contrast to other modern general purpose DSPs and microprocessors with SIMD units such as: Texas Instruments c64xx, Intel Pentium MMX, Motorola AltiVec PowerPC G4 and Analog Devices TigerSHARC, the new DSP performs variable bit-length vector/matrix arithmetic, logic and saturation operations. The main NMC operation is matrix by vector multiplication. The NM6403 supports shared memory mode for two 64-bit external data buses. Two byte-width communication ports simplify the multiprocessor systems design. The NM6403 has been designed by RC "Module" (www.module.ru) and produced by Samsung 0.5 μ m CMOS technology. The peak performance - up to 14.400 MMACs (million multiplication and accumulations per second) has been achieved at a 50MHz clock rate, 3.3V operating voltage and PBGA256 package.

1 INTRODUCTION

The NM6403 is a high performance DSP with elements of VLIW and SIMD architectures [3]. NeuroMatrix® architecture has a native support for 1-, 2-, 3- up to 64-bit data processing. Each of these data types is critical to standard DSP algorithms, image processing, voice compression and the next generation of wireless protocols. The flexible operands width and ability to scale performance let designers trade off precision and performance to suit their applications.

2 THE PROCESSOR STRUCTURE

The NM6403 is intended for processing of 32-bit scalar data and variable bit length vector data packed into 64-bit data words. The block diagram is depicted in Fig. 1.

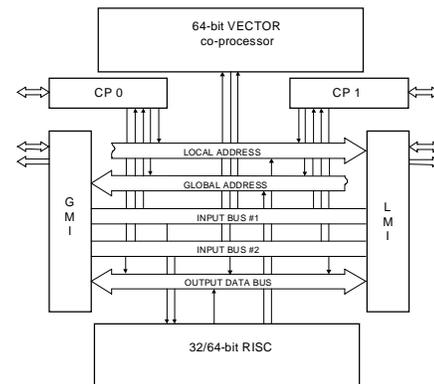


Fig. 1. Processor Block Diagram

The processor is comprised of the following functional units:

RISC processor A 5-stage pipelined 32-bit RISC performs scalar arithmetical/logical and shift operations with 32-bit data and general control functions.

Vector Coprocessor (VCP) - performs arithmetical and logical operations with 64-bit vectors - packed words of variable bit length vector data.

LMI and GMI - two identical 64-bit programmable local and global interfaces to external SRAM/DRAM. Each interface supports up to two memory banks and can function in a "shared-memory" mode. The NM6403 supports 32-bit internal addresses. The total external memory range is 4Giga 32-bit words.

CP0 and CP1 - two communication ports are hardware compatible with TI TMS320C4x DSP and provide bi-directional data transfer. Each CP

includes a DMA unit that performs 64-bit data transfer between the port and external memory connected to the global and (or) local buses.

The NM6403 DSP has five buses for rapid data exchange between the functional units. Instruction fetch is performed by 64-bit words. Each word is either one 64-bit or two 32-bit instruction(s).

2.1 RISC Processor

The processor is a 5-stage pipelined 32-bit RISC with the original instruction set. It operates with 32- and 64-bit wide instructions (usually two operations are executed by each instruction). All internal units of the RISC are 32-bit wide.

2.2 Vector Coprocessor

The NeuroMatrix® architecture [4] provides a unique flexibility of choice; the desired level of performance and precision for 2-D MAC procedure:

$$Y_m = U_m + \sum_{n=1}^N X_n \times W_{n,m}$$

According to application requirements, designers can select the necessary length (precision) of operands and of products. The number of MACs (multiplication and accumulations) depends on the length and number of operands. The highest performance - 14.400 MMACs is achieved with one-bit length operands at a 50MHz clock rate. It is possible to increase the precision of calculations using any operand length up to 32-bits. In this case, the performance is 50 MMACs with a 64-bit result. The VCP includes an active matrix which looks like an array multiplier (Fig. 2).

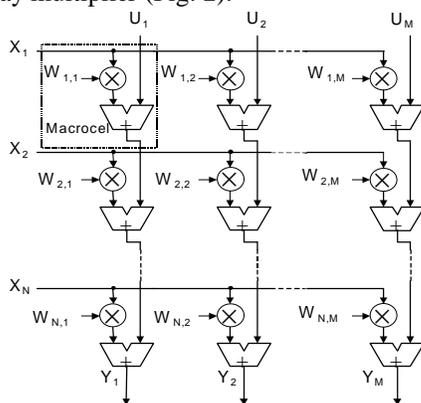


Fig. 2. Active Matrix

The structure comprises cells that include a 1-bit memory (flip-flop) surrounded by several logical elements. The software designer can combine the cells into several macrocells using two 64-bit programmable registers: DB - data boundary and MB - MAC boundary. These registers define the borders between rows and columns with macrocells. Each macrocell performs the multiplication of variable input words by preloaded coefficients (Wi) and accumulates the result from the macrocells in the column above it. The columns simultaneously calculate the results in one processor cycle.

The example of VCP configuration for 8-bit data (Xi) and coefficients (Wij) processing is shown in Fig. 3. This results in a peak performance of 1.200 MMACs which is achieved by parallel execution of 24 MACs with 21-bit results in one 20-nsec processor cycle.

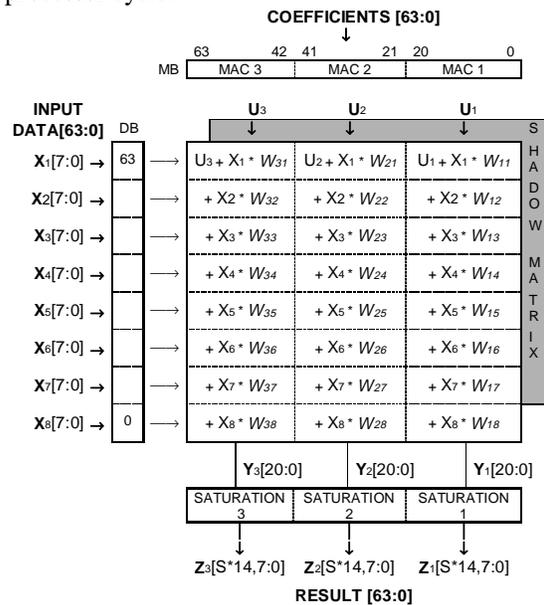


Fig. 3. NeuroMatrix® Engine

The number of MACs depends on the length and number of words packaged into a 64-bit block. The engine configuration can change dynamically during the calculations. You can start the application with maximum precision and nominal performance and then dynamically increase the performance by reducing the data-word lengths. The diagram at Fig. 4 shows the effects of speed vs. word length on the [5] peak performance of VCP.

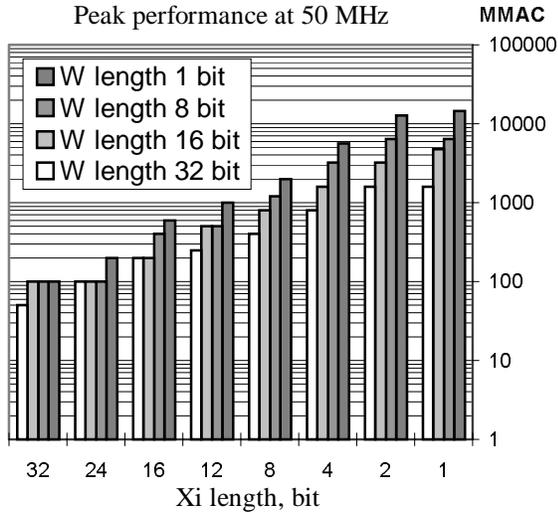


Fig. 4. Speed/Precision Trade-Off

The active matrix offers outstanding performance with "Boolean" arithmetic. So, 1-bit by 1-bit "Boolean" multiplication delivers 50.000+ MOPS at a 50MHz clock rate. There is another interesting feature of using the "binary" coefficients. If the coefficients have binary values - "all 1" or "all 0", the matrix becomes a powerful switch (router). The remapping of bit positions into a 64-bit input data word is performed in one processor cycle.

To load new coefficients to the active matrix, 32 clock cycles are needed. To avoid the delays due to refreshing the coefficients, the shadow matrix is used. The new coefficients are loaded to the shadow matrix in a background mode and then copied to the active matrix in one clock cycle.

To avoid arithmetic overflow, the saturation function (Fig. 5) with user-programmable saturation boundaries is used:

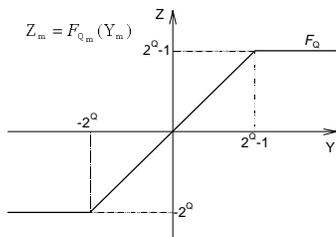


Fig. 5. Saturation Function

The saturation function reduces the number of significant bits of MAC products.

3 INSTRUCTION SET

The NM6403 instruction set is divided into two major types: Scalar Instructions (Fig. 6)

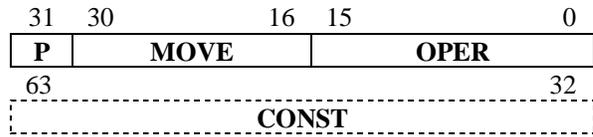


Fig. 6. Scalar Instruction Code

and Vector Instructions (Fig. 7).

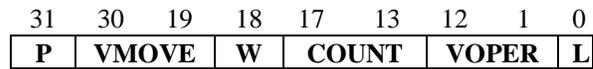


Fig. 7. Vector Instruction Code

The scalar instructions are RISC-like instructions, which also support conditional branch, call, and return instructions. The processor supports immediate instructions with 32-bit addressing, base, indexed, and relative addressing. The vector instructions have a special field to define the number (from 1 up to 32) of repeats of their executions. This solution supports short hardware loops and essentially increases the code density.

The assembly code of an inner loop of the 3x3 Convolution Filter can be written as:

```

nb1 = 80008000h;
sb = 02020202h;
<L>
  rep 24 wfifo = [ar6++], ftw;
  WTW_REG(gr2);
  rep 32 data = [ar0++],ftw with vsum , data, 0;
  WTW_REG(gr2);
  rep 32 data = [ar1++],ftw with vsum , data, afifo;
  WTW_REG(gr2);
  rep 32 data = [ar2++] with vsum , data, afifo;
  if > delayed goto L with gr7--;
  ar6 = gr6;
  rep 32 [ar4++gr4] = afifo;

```

In the first two lines the active matrix of 4 columns 16-bits each and 8 rows 8-bits each is initialized. In the third line the weights are loaded into the shadow matrix. The next line copies the weights from the shadow matrix to the active one. Then calculation of sums of 3 elements in the first line of 3x3 mask is performed. The sixth line copies new weights from the shadow matrix to the active one. Then we calculate the sums of 3 elements in the second line of

3x3 mask and add them to the previous ones. The eighth line, copies new weights from the shadow to the active matrix. Then calculation of sums of 3 elements in the third line of 3x3 mask and their addition to the previous ones is performed. Finally, the tenth line contains a delayed conditional jump. The term "delayed" means that two more lines are executed before the jump itself occurs. The first of the delayed commands restores the pointer to the weights array in external memory, the second one stores the result of the calculations to the memory.

Some benchmark comparisons between NM6403 and TI's C80 is shown in Table 1. The competitor's data are taken from [6].

Table 1. Convolution Filters

Convolution Filter	NM6403 Cycles/pixel	TMS320C80 Cycles/pixel
3x3	1.8	2.1
5x5	2.6	7.3
7x7	4.3	n/a
9x9	5.7	n/a

It is important to note that the number of cycles for NM6403 increases as a linear function.

The technique of using NM6403 DSP for neural networks acceleration can be found in [5].

4 APPLICATIONS AND BENCHMARKS

Due to its flexibility and high performance, NM6403 has a broad range of applications:

- digital signal processing (FFT, DFT, WHT);
- image processing (Convolution Filters);
- neural net and vector/matrix acceleration;
- telecommunications;
- embedded systems;
- large super parallel computer systems.

The results of running the standard DSP functions are shown in Table 2. The benchmarks were written in NM6403 assembler and compiled using the NM6403 Software Development Kit. The results have been obtained using the NeuroMatrix® NM1 PCI Board .

The parameters of the functions are: Sobel Transform - frame size: 384x288 bytes, FFT 256-

points - 32-bit data, Walsh Hadamard Transform (WHT) - 21 step, initial data - 5-bit.

Table 2. NM6403 Benchmarks

	Pentium II, 300 MHz	Pentium MMX, 200 MHz	TMS320C40, 50 MHz	NM6403, 40 MHz
Sobel (frs/sec)	n/a	21	6.8	68
FFT (usec)	200	n/a	464	102
WHT (sec)	2.58	2.80	n/a	0.45

5 Conclusion

The NM6403 is a new class of fixed point DSPs. Its key element, the NeuroMatrix® Core provides programmable operand width and offers scaleable performance from 50 MMAC (32-bit data, 64-bit product) up to 14.400 MMAC (1-bit data, 8-bit products) at 50 MHz clock rates. The power consumption at 50MHz and 3.3V is less than 1 W, which is especially important in embedded systems.

References

- [1]. Peter Clarke, "Neural-emulator IC promises scalability", *EETimes* 1998, April 27, Issue 1004, pp. 37-38.
- [2]. Peter Clarke, "Pact eyes multimedia, telecom", *EETimes* 1999, October 25, http://www.eet.com/story/core_competency/OEG19991025S0005
- [3]. Markus Levy, "1999 DSP-architecture Directory", *EDN Access* 1999, April 15, pp. 67-68, 102.
- [4]. *Patent 2131145 Russian Federation*, "Processor, device for saturation functions calculation, computing device and adder", RC Module, June 16, 1998.
- [5]. P.A. Chevtchenko, D.V. Fomine, V.M. Tchernikov, P.E. Vixne, "Using of microprocessor NM6403 for neural net emulation", *SPIE Proceedings Vol. 3728*, 1999, pp.242-252.
- [6]. *Texas Instruments Europe*, "Implementation of an image processing library for the TMS320C8X (MVP)", BPRA059, July 1997.