DISCRETE COSINE TRANSFORM GENERATOR FOR VLSI SYNTHESIS

Jill Hunter, John V McCanny

DSiP Laboratories, School of Electrical Engineering and Computer Science(j.hunter@ee.qub.ac.uk),
The Queen's University of Belfast, BELFAST, BT9 5AH, Northern Ireland.

ABSTRACT

A generator for the automated design of Discrete Cosine Transform (DCT) cores is presented. This can be used to rapidly create silicon circuits from a high level specification. These compare very favourably with existing designs. The DCT cores produced are scaleable in terms of point size as well as input/output and coefficient wordlengths. This provides a high degree of flexibility. An example, 8-point 1D DCT design, produced occupies less than 0.92 mm^2 when implemented in a 0.35μ double level metal CMOS technology. This can be clocked at a rate of 100MHz.

1. INTRODUCTION

The rapid design of application specific DSP cores is an increasing requirement for the creation of complex, next generation integrated circuits. Single chip solutions incorporating numerous DSP blocks are typically needed in an assortment of multi-media products. These comprise a combination of image compression, speech recognition and communication capabilities and increasingly must be designed, tested and manufactured in a much reduced time.

The DCT is a key computational block required in many signal and image processing applications. In image processing the 2D 8-point DCT forms the basis of many of the primary compression standards, including JPEG, MPEG1, MPEG2, H.261 and H.263. Other image processing applications employ a range of DCT point sizes for 1D, 2D and more recently 3D transforms. Radiological archiving of diagnostic images and progressive image transmission have both been examined with a range of DCT transform sizes up to 16x16. The DCT can also be embedded with vector quantizers to further enhance the performance of image compression systems. Such systems have undergone trials in low bit-rate video-phone and video-conference facilities. The Shape Adaptive DCT (SADCT) has also been proposed for coding of low bit-rate applications. Here the DCT architecture needs to be both regular and completely scaleable. Speech processing applications also employ DCT cores. These are used directly in some speech coding systems, as well as in filters within spectrum analysers. A summary of typical DCT transform sizes used in popular DSP/image processing applications is presented in Table 1. Typical operating speeds range from 64kbps for auditory applications to broadcast video at 250Mbps. In the case of HDTV, a maximum operating speed of around 1.3Gbps is required.

The purpose of this paper is to present methods for the rapid silicon design of DCT circuits. The generator described provides a non-specialist with the means to create, on-demand layouts for application specific DCT's tailored to exact specifications. This has been created through the hierarchical use of parameterised libraries of hardware description language models which leads to a completely scaleable generator. The DCT core module can additionally be connected in a variety of physical configurations allowing considerable scope for area/speed trade-offs. This is important in order to cover the full spectrum of performance and bandwidth requirements.

APPLICATION	**POINT SIZE** *1D DCT*	*2D DCT*
Image Compression	- - -	4x4 8x8 16x16
Digital Filtering	- - -	8x8 16x16
LMS Filtering	8 12 16 32 64 128	- - -
Transmultiplexing	3 4 7 14 64 72 128	- - -
Vector Quantization	4 9 16 25	2x2 5x5 6x6 4x4 8x8 16x16
Speech Coding	32 64 128 256 512	- - -

Table 1 Typical DCT applications

2. DCT GENERATOR ARCHITECTURE

For an input sequence of data values, $x(n)$,the 1-D DCT, $y(n)$, is defined as

$$y(n) = \alpha(k) \sum_{n=0}^{N-1} x(n) \cos \frac{(2n+1)\pi k}{2N} \qquad (1)$$

with $\alpha(0) = \frac{1}{\sqrt{N}}$ and $\alpha(k) = \sqrt{\frac{2}{N}}$ for $\{0 \le k < N\}$.

Real-time implementation of the DCT operation is highly computationally intensive. Accordingly, much effort has been directed to the development of suitable cost effective VLSI architectures to perform this [6,7]. Traditionally the focus has been on reducing the number of multiplications required. Additional design criteria has included minimising the complexity of control logic, memory requirements, power

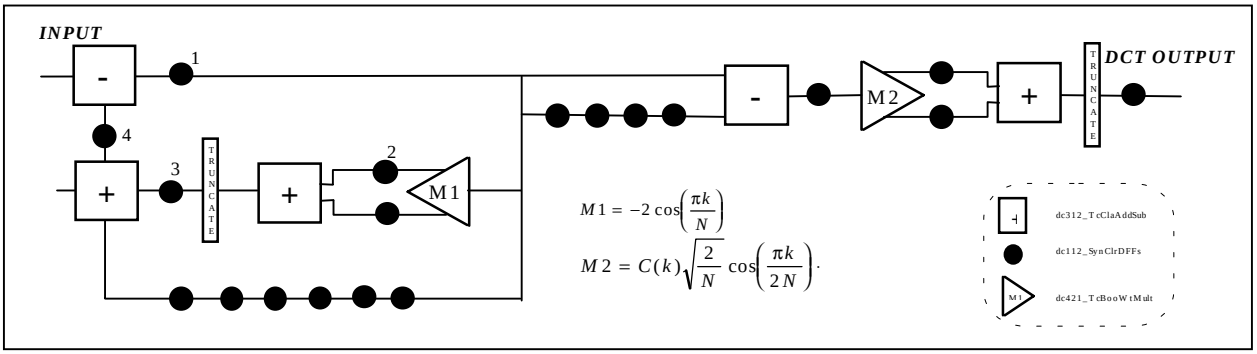


Figure 1 DCT_quad module

consumption and complexity of interconnect. These have all been taken into consideration in the design of the DCT generator.

Of direct interest for the DCT generator are scaleable architectures. These tend to be derived from recursive algorithms which result in either butterfly, systolic array or lattice type structures. The algorithm presented by Cvetkovic [1] allows for the generation of higher order DCT's from identical lower order modules. However the butterfly structure is only suitable for even point-size transforms. Additionally global communications are employed and these result in a complex layout. Wang and Chen [2] have developed a series of linear systolic arrays for computing 1-D N-point and 2-D NxN-point DCT's. The architecture developed is completely scaleable and attractive in terms of regularity and complexity. Such designs are however expensive in terms of circuit area. A novel lattice implementation is presented by Chiu and Liu [3]. This computes the DCT from a frame recursive point of view and the resulting structures are claimed to be the fastest know thus far. It is this lattice organisation, which is used as the basis of the work presented in this paper. Area optimality and communication locality are the key advantages of this structure. Additionally the scaleable nature of the time-recursive algorithm can be exploited and used to radically simplify the architectural mapping. A full custom VLSI design using this lattice organisation has been presented by Srinivasan and Liu [8].

The DCT architecture developed is modular and based on a generic processing element (DCT_quad). Such elements can be arrayed to produce a required *N*-point DCT circuit when combined using RAM sequencing blocks. The complexity of implementation is low; only 2 multipliers are required and an *N*-point 2D DCT can be computed using only *N*/4 modules.

2.1 DCT_quad Module

The structure for the basic DCT_quad module is shown in Figure 1. This block accepts data in a bit parallel, word serial manner typical of that required in real-time image and speech processing systems. Equation (1) can be used to derive the equivalent transfer function i.e.

$$H_c(z) = (-1)^k \alpha(k) \sqrt{\frac{2}{N}} \cos\left(\frac{\pi k}{2N}\right) \cdot \frac{(1 - z^{-1})}{1 - 2\cos\left(\frac{\pi k}{N}\right) z^{-1} + z^{-2}} \qquad (2)$$

The DCT_quad module directly implements this transfer function. It can be used to compute the *k* individual DCT channel outputs depending on the particular multiplier coefficient encoded .

In order to increase the system throughput rate the design has been pipelined. However, this can create problems in a bit parallel implementation. The introduction of the 4 pipelined stages in the recursive loop produces a four cycle delay in the feedback loop. On the face to it, this appears to negates any potential benefits. However, this can be exploited to advantage. Whilst a single DCT output value can only be produced every four cycles, there are 3 additional channels which can be used to process 3 other separate data streams. The DCT_quad block can therefore process 4 data channels without the addition of any new circuitry. Moreover, data can be processed by each circuit element on each clock cycle. The total time required to process an N-point 1-D DCT using these methods is 4(N+1) cycles. Here new data is input to the circuit every 4 cycles under the control of the input data sequencer. This comprises a dual-port RAM and a 0-3 counter (Figure 2). It was decided to include this block as a separate entity from the DCT_quad module; some applications require input data to be accessible from a central memory. In such cases, circuitry is included which can access the data in the desired scheduling sequence. Additionally, for larger DCT circuits, it is known that more compact circuit layouts are achieved using pre-designed RAM cells.

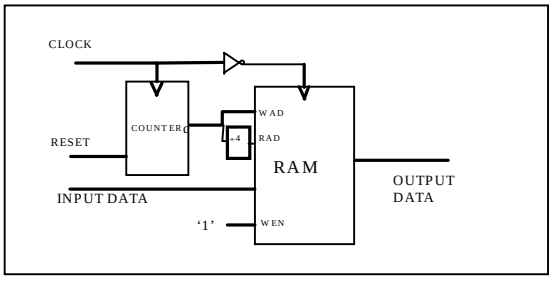


Figure 2 RAM Sequencer Block

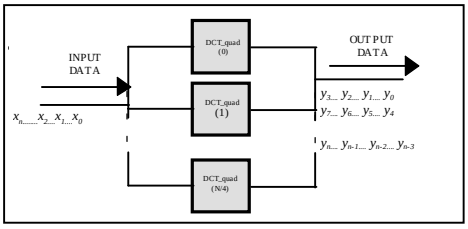


Figure 3 N/4 DCT_quad 1D DCT

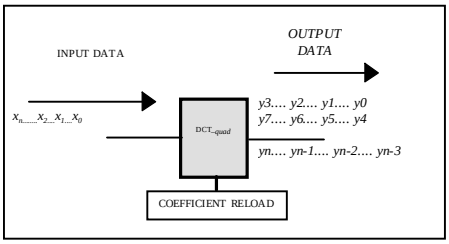


Figure 4 1 DCT_quad 1D DCT

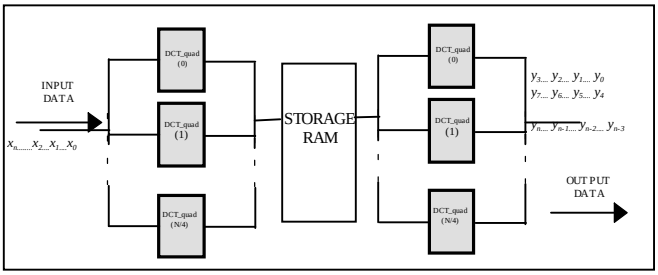


Figure 5 N/2 DCT_quad 2D DCT

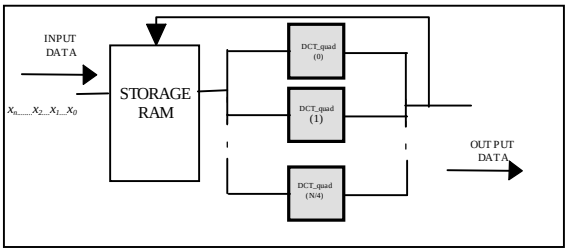


Figure 6 N/2 DCT_quad 2D DCT

The RAM block used is parameterisable and can be defined in terms of stored data wordlength and number of input data words i.e. N, the DCT transform size. A typical RAM block which stores 8 x 9-bit input words and occupies 0.047mm^2 in 0.35µ CMOS DLM technology.

2.2 DCT_quad Sub-Modules

The most critical sub-module within the DCT_quad architecture is the multiplier and accordingly, attention has been focused on the types of multiplier which could be used to implement this. Of the three principal styles reported in literature (combinational multiplier, bit-serial and distributed arithmetic), none appears to consistently offer outstanding comparative results. Theoretical studies [4] suggest that bit-serial is the preferred approach for DCT implementation. Such designs exhibit a high degree of concurrency, have a large dynamic range yet have narrow signal processing bandwidths. However, practical DCT circuits reported do not confirm these findings. Accordingly a combination of architectural design issues needs to be addressed. The hierarchical design methodology used [5] tends to favour the use of combinational multipliers and this approach has been adopted. These have been implemented using architectural blocks from *ISS Ltd.'s* hierarchical library [5]. For the adder and subtractor blocks, a two's complement carry-lookahead add/subtract module has been employed. The subtraction of the binary inputs is accomplished by placing a set of invertors on the data input lines. The multipliers used are Booth-encoded Wallace trees circuts, which are attractive in terms of performance and silicon area. Delay circuitry is also used to facilitate dynamic operation

3. DCT CORE LAYOUT

The DCT_quad block can be connected in a variety of physical configurations to implement both 1-D and 2-D DCT circuits. For simplicity, the input scheduling ROM for these circuits has been omitted. Figure 3 indicates the most basic mode of connection for a 1-D DCT. In this N/4 quad modules are connected in parallel to complete the structure. Each module generates 4 DCT coefficients and the total time required to calculate the complete 1D block DCT is 4(N+1) cycles. The first valid recursive data values emerge after 4N cycles. A single quad module can alternatively be used to implement the DCT as shown in Figure 4. This reduces the circuit area to around a quarter of it's value in the orientation in Figure 3. In this case the total time required to compute a full 1-D block is N(N+1) cycles. When N/4 does not return an integer result, the value is rounded up to the next whole number. Additional multiplier coefficient reload circuitry is therefore required. The most straightforward 2-D DCT circuit can be generated if two identical 1-D DCT blocks are cascaded, the data from the second block being processed immediately after it emerges from the first (Figure 5). An N-point 2-D DCT implemented in this manner requires N/2 DCT_quad blocks and an additional storage RAM if the structure is to be used for block transforms. Using this design a full block 2-D DCT output is available after $8N^2+10$ cycles with subsequent outputs available $4N^2+5$ cycles later. These times are inclusive of all in-circuit resets. It is also possible to reuse the original circuit to calculate the 2-D DCT (Figure 6). This is a standard technique used to reduce hardware overhead but increases circuit latency.

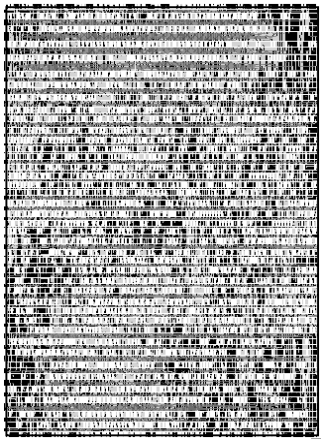
4. DCT PERFORMANCE

The DCT generator has been used to create many different DCT silicon designs . These include a JPEG and MPEG compatible 1D 8-point DCT circuit. In this case, the DCT_quad block has been connected in it's most basic configuration as shown in Figure 3. The circuit accepts 9-bit input data and produces a 12-bit output with 12- and 10-bit internal multiplier coefficients. Simulation and synthesis has been carried out using the *Synopsys* suite and a layout for the circuit has been developed using the *Compass Design Automation* tools along with their Passport libraries. The resulting design and corresponding circuit statistics are shown in Figure 7. It has been calculated that a 2D 8-point DCT with internal RAM sequencing will occupy 2.5mm^2. This can be clocked up to 100MHz and consequently the demanding performance requirements of HDTV can be met. This also compares favourably with recent designs reported in literature. Madisetti and Willson [6] describe a 100MHz 2-D 8x8 DCT targeted at HDTV applications which occupies 10mm^2 in an 0.8µ process. A 200MHz 13mm^2 (0.8µm CMOS) 2-D DCT macrocell has also been presented by Matsui et. al. [7].

Power consumption data has been derived using the *EPIC Powermill* tool. The DCT_quad block described has a power consumption of 282mW when clocked at 100MHz and targeted at a two-layer, high density, 0.35µ process. Supporting research on smaller arithmetic modules has indicated that with correlated input data (e.g. real-time video), the reported power consumption is much less, up to 50% in some cases. Fixed-point circuit accuracy was also assessed. This was measured using a combination of the *Synopsys* simulation tool supported by *Matlab* functions. The signal-noise ratio (SNR) was measured after DCT processing of the standard test image Lena. The value of 43dB obtained exceeds the 40dB required for most many video compression standards.

5. SUMMARY

A new DCT generator has been presented which facilitates the rapid design of application specific DCT cores. The transforms produced are portable across many silicon foundries. Worked examples confirm that the cores are comparable in area, performance and power consumption to those based on more conventional methods; design times (typically 1 day in the case of an 8-point core), are reduced by several orders of magnitude. Most significantly, no constraints are placed on the DCT point size and area/speed trade-offs can be rapidly computed and analysed. At present, a one DCT_quad design, containing only two multipliers, is being used to implement an 8-point 2D DCT. This will be presented at a later date.



1D 8-point DCT	
Technology	0.35µ CMOS
Chip Area	0.92mm^2
Nos. of Transistors	50k
Max. Clock Speed	100MHz
Data Rate	1.8GBps
PSNR	43dB

Figure 7 1D 8point DCT circuit

6. REFERENCES

[1] Z. Cvetkovic and M V. Popovic "New Fast Recursive Algorithms for the Computation of Discrete Cosine and Sine Transforms" IEEE Trans. Signal Processing, vol.40,no. 8, 1992

[2] C. Wang and C. Chen "High-Throughput VLSI Architectures for the 1-D and 2-D Discrete Cosine Transforms" IEEE Trans. Circuits Systems for Video Tech.,vol.5, no. 1,1995

[3] C.T. Chiu and K.J. Ray Liu "Real-Time Parallel and Fully Pipelined Two-Dimensional DCT Lattice Structures with Application to HDTV Systems" IEEE Trans. Circuits and Systems for Video Technology, vol. 2, no 1, March 1992

[4] D.Crook and J. Fulcher "A Comparison of Bit Serial and Bit Parallel DCT Designs" VLSI Design, vol. 3, no 1

[5] J.McCanny, D.Ridge, Y.Hu, J.Hunter "Hierarchical VHDL Libraries for DSP ASIC Design" IEEE Proceedings, ICASSP-97, Munich, vol.1, pp.675-679

[6] A. Madisetti and A. Willson, "A 100MHz 2-D 8x8 DCT/IDCT processor for HDTV applications" IEEE Trans. Circuits, Systems for Video Tech. vol. 5,no. 2**,** April 1995

[7] M. Matsui, M. Hara, Y.Uetani, L.Kim, T.Nagamatsu, Y.Wantanabe, K.Masuda, T.Sakurai "A 200 MHz 13mm^2 2-D DCT macrocell using sense-amplifying pipeline flip-flop scheme" IEEE Jour. Solid State Circuits, vol.29, no.12 Dec 1994

[8] V. Srinivasan and K.J. Ray Liu "VLSI Design of High-Speed Time-Recursive 2-D DCT/IDCT Processor for Video Applications" IEEE Trans. on Circuits and Systems for Video Technology, vol. 6., no. 1.,February 1996