DESIGN TRADE-OFF IN MERGED DRAM LOGIC FOR VIDEO SIGNAL PROCESSING

Sunho Chang and Lee-Sup Kim

Korea Advanced Institute of Science and Technology (KAIST)
373-1 CHiPS, Gusong-dong Yusong-gu, Taejon Korea

ABSTRACT

The trade-off in designing *merged DRAM logic (MDL)* is explored for video signal processing. Computing requirements and memory bandwidths are quantitatively analyzed in the programmable MDL architecture. The *number of processing elements (NPE)* and the *number of bus width (NBW)* are obtained as a function of macro block rate, clock frequency, data rate, and number of clock/memory access cycles. Optimal MDL design parameters are determined from the minimum cost and design metrics: *DRAM access rate (DAR)* and *area ratio of DRAM (ARD)*.

1. INTRODUCTION

Merging DRAM with Logic (MDL) has been a solution in designing system-on-a chip, especially with below 0.18 μm CMOS technologies. Many computation-intensive video signal processing have been implemented with MDL that have high data bandwidth, low DRAM latency, and low power dissipation. To enhance the merging effects, trade-off of performance, area, and power must be considered in designing MDL.

Conventionally, two types of MDL have been implemented. One is merging a dedicated signal processor with DRAM [1]; the other is merging a general-purpose processor with DRAM [2]. In both cases, design trade-off in the MDL has not been analytically considered. Additionally, there has been not much study to enhance merging effects of DRAM and logic.

In this paper, we have explored trade-off in designing MDL for video signal processing. At first, real-time computing requirements and memory bandwidths are quantitatively analyzed in the programmable MDL architecture. From the analysis, two relations, the *number of processing elements (NPE)* and the *number of bus width (NBW),* are obtained as a function of design parameters. Secondly, design constraints (time, area, and power) are described by the MDL design parameters. Finally, design trade-off of MDL is determined in terms of design cost function and design metrics: *DRAM access rate (DAR)* and *area ratio of DRAM (ARD).*

This paper is organized as follows. In Section 2, MDL design considerations are explored. In Section 3, design constraints are considered and MDL design metrics is described. Experimental results are written in Section 4. Finally, Section 5 concludes this paper.

Acknowledgements: This work has been supported in part by the Korea Science and Engineering Foundation (KOSEF) through MICROS research center in KAIST.

2. EXPLORING MDL DESIGN

2.1 Definition

Several parameters in video signal processing are defined to determine real-time requirements on the processing throughput and memory bandwidth in the MDL:

- Horizontal and vertical pel size: H x V
- Frame rate (frames/sec): F
- Macro-block (MB) size in pel: M x M.

From the parameters, pel rate (samples/sec) is

$$Pelrate(samples/sec) = H \times V \times F. \quad (1)$$

The macro block rate (MBs/sec) can be defined as

$$MBrate\ (MBs/sec) = [H/M] \times [V/M] \times F \quad (2)$$

where [ ] means the nearest upper integer. Then, the macro block time is given by

$$MBtime = 1 / (MBrate) = 1 / ([H/M] \times [V/M] \times F). \quad (3)$$

The *Pelrate*, *MBrate*, and *MBtime* for each level of MPEG2 are summarized in Table 1.

Table 1. Macro block rate for MPEG2 four levels

picture level	Formula	MP@LL	MP@ML	MP@H1440	MP@HL
pels/line	H	352	720	1440	1920
lines/frame	V	288	480	1080	1080
frames/s	F	30	30	60	60
pel rate(samples/s	PR=HxVxF	3041280	10368000	93312000	1.24E+08
MB rate(MBs/s)	MR=PR/MB	11880	40500	364500	486000
MB time(us)	1/MR	84.2	24.7	2.7	2.1

The macro block cycle time is defined as

$$MBcycletime = MBtime / Ncycles \quad (4)$$

where *Ncycles* refers to the number of cycles for the processing of one macro-block (MB) in basic processing elements. For real-time operation the following condition must be met in the MDL system:

$$MDLcycletime \le MBcycletime. \quad (5)$$

If there exists *n-parallel processing (NPP)*, (5) is

$$MDLcycletime \le MBtime / (Ncycles / NPP). \quad (6)$$

As clock frequency (Fclk) is the inverse of *MDLcycletime*, NPP can be obtained as

$$NPP \ge MBrate \times Ncycles / Fclk. \quad (7)$$

From (7) the minimally required parallel processing is

$$NPPmin = MBrate \text{ x } Ncycles / Fclk. \quad (8)$$

2.2 Processing and Clock Frequency in MDL

Based on the definition mentioned in 2.1, computational requirement for video signal processing is analyzed to decide the proper number of processing element and clock frequency. If there is N processing element (NPE), the *speed-up*, using analytical results from [3], is given by

$$S(NPE) = \frac{NPE}{1 + \frac{NPE - 1}{Rp}} \quad (9)$$

where *Rp* is a constant that depends on the capability of computing in the MDL. As the *Rp* increases, S(NPE) approaches NPE. By applying (9) into (8), the minimally required NPE is represented as

$$NPE\ \min = \frac{(Rp-1) \cdot MBrate \cdot Ncycles,compute}{Rp \cdot Fclk - MBrate \cdot Ncycles,compute} \quad (10)$$

where *Ncycles,compute* is the number of clock cycles for computing. After *Ncycles,compute* is obtained from the basic processing elements, *NPEmin* and *Fclk* is determined from (10).

2.3 Memory Access and Data Rate in MDL

For real-time operation, the cycle time of merged *memory element (ME)* must meet the following condition:

$$MEcycletime \le MBcycletime. \quad (11)$$

If the condition (11) is not met, the memory bandwidth still becomes bottleneck even though external memories are merged with logic. Since the video signal data to be handled is enormous, wide bandwidth is required. If there are *n-times basic bus width (NBW)*, (11) is

$$MEcycletime \le MBtime / (Ncycles,memory / S(NBW)) \quad (12)$$

where *Ncycles,memory* is the number of memory cycles required to access merged DRAM. The *S(NBW), speed-up due to NBW*, is defined as

$$S(NBW) = \frac{NBW}{1 + \frac{NBW - 1}{Rb}} \quad (13)$$

where *Rb* is a constant that depends on the capability of communication in the MDL. As the *Rb* increases, S(NBW) approaches NBW. From (12), S(NBW) can be

$$S(NBW) \ge MBrate \text{ x } Ncycles,memory / Drate \quad (14)$$

where *Drate,* the inverse of *Mecycletime,* is the data rate of the merged DRAM. From (14), the minimally required bus width is

$$NBW\ \min = \frac{(Rb-1) \cdot MBrate \cdot Ncycles,memory}{Rb \cdot Drate - MBrate \cdot Ncycles,memory} \quad (15)$$

After *Ncycles,memory* is obtained from the MDL, *NBWmin* and *Drate* is determined from (15).

3. TRADE-OFF IN MDL DESIGN

3.1 Design Constraints

We define MDL design constraints; time, area, and dynamic power as a function of NPE, NBW, and *Fclk* as follows.

i) $$Time(NPE, NBW) = 2((\lceil SBD/NBW \rceil + 1 + Ld) + CA/NPE + Lp)/Fclk \quad (16)$$

where SBD is the size of basic data block, CA is computation amounts, *Ld* is the latency of DRAM, and *Lp* is the latency of processing elements.

ii) $$Area(NPE, NBW) = \sqrt[\rho]{\frac{NBW}{NBW\ \max}} \cdot Adram(NBW\ \max) + NPE \cdot Ape + NPE \cdot NBW \cdot Abus + Aothers \quad (17)$$

where *Adram* is the area of DRAM; *Ape* is the area of processing element; *Abus* is the area of bus; and *Aothers* is the area of SRAM and controller. The constant ρ is a factor of reduced DRAM area due to decreasing NBW [4].

iii) $$Power\_dynamic(NPE, NBW, Fclk) = (NPE \cdot Cpe + NPE \cdot NBW \cdot Cbus + NBW \cdot Cmemory) \cdot V^2 \cdot Fclk = \alpha \cdot C \cdot V^2 \cdot Fclk \cdot Area(NPE,NBW) \quad (18)$$

where α is the node transition activity factor, C is the node capacitance. From three design constraints, we define *MDL cost* as

$$MDL\ Cost(NPE, NBW, Fclk) = Time(NPE, NBW, Fclk)^a \cdot Area(NPE, NBW)^b \cdot Power(NPE, NBW, Fclk)^c \quad (19)$$

where a, b, and c are integers to emphasize one of the design constraints for a specific target application.

3.2 DRAM Access Rate (DAR) and Area Ratio of DRAM (ARD)

To figure out merging effects and implementing cost, we define design metrics: *DAR* and *ARD*. At first, *DAR (DRAM Access Rate)* is defined as the ratio of clock cycles for transferring data to clock cycles for computing data.

$$DAR(NPE, NBW) = 2 (\lceil SBD/NBW \rceil + 1 + Ld) / (CA/NPE + Lp). \quad (20)$$

Even though NBW is widened enough, there exits *the minimum DAR, DAR_0,* due to the latency of DRAM, which is

$$DAR_0 = 2 (1 + Ld) / (CA/NPE + Lp). \quad (21)$$

Therefore, the change of DAR (ΔDAR) can be written as

$$\Delta DAR = DAR - DAR0. \quad (22)$$

The ΔDAR represents the merging effect of DRAM with logic according to the variation of NBW. Secondly, *ARD (Area Ratio of DRAM to logic)* is defined as

$$ARD(NPE, NBW) = \frac{\sqrt[\rho]{\frac{NBW}{NBW\ \max}}\, Adram(NBW\ \max)}{NPE \cdot Ape + NPE \cdot NBW \cdot Abus + Aothers}. \quad (23)$$

The DAR and ARD in the MDL are shown in Fig.1 for several design parameters.

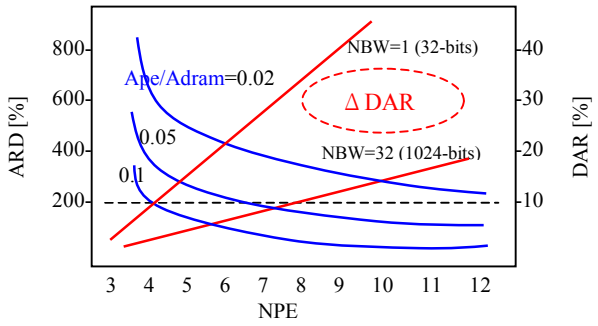

Fig.1 DAR and ARD in the MDL

The *Aothers* is assumed as 0.1Adram and *Ape* is assumed as three kinds of *0.02Adram*, *0.05Adram*, and *0.1Adram*. As the NPE increases, the ΔDAR becomes larger when NBW is widened, and the ARD is decreased depending on the ratio of *Ape/Adram*.

4. EXPERIMENTAL RESULTS

Exploring trade-off of MDL has been experimented on the programmable MDL for the fundamental video signal processing: *inverse discrete cosine transform (IDCT)* and *motion compensation (MC)*.

4.1 Programmable MDL Architecture

As shown in Fig. 2, basic digital signal processing elements consist of 32-bit *arithmetic logic unit (ALU)*, *multiplier and accumulator (MAC)*, and *barrel shifter (BS)* with 8-bit wise split function. The synchronous DRAM and multi-port SRAM [5] are adopted into the memory system.

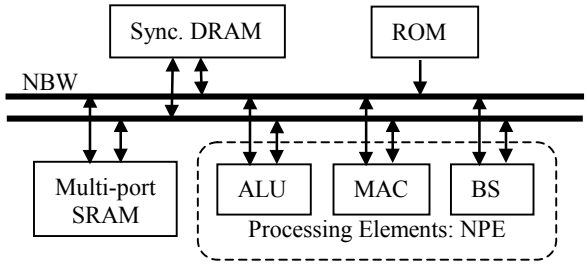

Fig.2 Block diagram of the programmable MDL

4.2 Product of NPE and Clock Frequency

4.2.1 Number of Cycles to Compute IDCT

The 1-D IDCT computation for a 8x8 basic data block can be written as

$$Y(k) = \frac{c(k)}{2}\sum_{i=0}^{3}(X(i) \pm X(7-i)\cdot\cos((2i+1)\cdot k\pi/16) \quad (24)$$

where the ± sign is used for even and odd values of k, respectively. From (24), the 8-point 1-line IDCT can be written as two linear matrixes. Then, the computations are

4Additions(or Subtractions)+ 4(Division by 2)
+16Matrix calculation
= 4 (ADD + DIV + 4cMACs) (25)

where 4cMACs indicates four continuous-MAC operations. Based on the 8-bit wise split function of the processing elements, the number of computations to compute IDCT for one MB is

Ncycles,compute IDCT
= (ADD + DIV + 4cMAC) x *2 (- and +)*
x *8 (lines/8x8 block)* x *2 (horizontal and vertical)*
x *4 (basic blocks/MB)*
= 128 (ADD + DIV + 4cMACs). (26)

4.2.2 Number of Cycles to Compute MC

For each macro block, MC is performed after IDCT. The *motion vector (MV)* decoding is a process to decode an encoded MV that addresses each reference macro block of a reference picture. The MV is encoded by using the *difference (delta)* with respect to the *last encoded MV (PMV)*. The *delta* can be obtained from the summation of 4-additions, 1-multiplication, and 1-shift operation [6]. The MV can be obtained from the summation of 3-additions, 1-multiplication, and 1-shift operation. Therefore, the number of computations for one MB is

Ncycles,MV
= (4ADDs+MUL+SFT) x *4MV*
+(3ADDs+MUL+SFT) x *4MV.* (27)

In the case of four-MV, the predictions for the input streams are

P(x,y,t)
= [p(x-u1, y-v1, t-1) + p(x-u2, y-v2, t-1)] /2
+ [p(x-u3, y-v3, t-1) + p(x-u4, y-v4, t-1)] /2 (28)

where (u#,v#) means MV#. After predictions, the frame addition is performed on the IDCT results e(x,y,t) and prediction results p(x,y,t) like

Frame Addition (x,y,t) = e(x,y,t) + p(x,y,t). (29)

From (28) and (29), the number of computations for one macro-block is written as

Ncycles,pred/add =
(4Additions + 2 (Division by 2)) x *MBsize/4.* (30)

The total number of clock cycles to compute one MB is obtained from the sum of (26) and (30), and given by

Ncycles,compute
= Ncycles,computeIDCT + Ncycles,computeMC
= Ncycles,computeIDCT + Ncycles,MV
+ Ncycles,pred/add
=156ADDs + 136DIVs + 128-4cMACs+ 8SFTs
+ (4ADDs + 2DIVs) x *MBsize/4.* (31)

The clock frequency obtained with varying NPE for four-levels of MPEG is shown in Table 2.

Table 2. Clock frequency [MHz] with varying NPE

Level	NPE=1	NPE=2	NPE=3	NPE=4
HL	575	285	195	150
H1440	432	216	144	108
ML	48	24	16	12
LL	15	8	5	4

The clock frequency for several *Rp* is shown in Fig. 3.

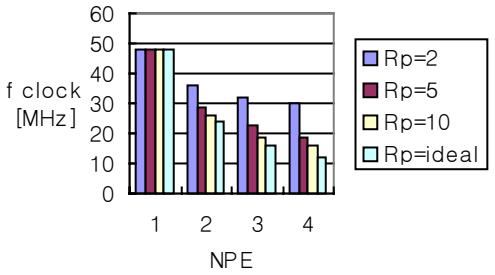

Fig.3 Clock frequency for several Rp with varying NPE @ML

4.3 Product of NBW and Data Rate

At first, the number of clock cycles to access embedded memories [5] for one macro-block IDCT is

$$Ncycles,memoryIDCT = Ncycles,DRAM\_IDCT + Ncycles,SRAM\_IDCT = (MBsize/4 + Ld) + 2(MBsize/4 + Ls) \times NSA \quad (32)$$

where *Ls* is latency of SRAM; NSA (number of SRAM access) is the number of intermediate *load* and *store* cycles that are consumed for IDCT. Secondly, in the case of B-prediction (worst case), *read previous picture DRAM* by MV1, *read future picture DRAM* by MV2, and finally *read IDCT output* from 8-port SRAM should be performed. Hence, the *Ncycles,memory MC* can be written as

$$Ncycles,memoryMC = (MB(P)size/4 + Ld) + (MB(F)size/4 + Ld) + (MB(S)size/4 + Ls). \quad (33)$$

Now the total number of clock cycles to access one MB is obtained from the sum of (32) and (33) like

$$Ncycles,memory = 3(MBsize/4 + Ld) + (MBsize/4 + Ls)(2NSA+1). \quad (34)$$

The data rate of DRAM obtained with varying NBW for four-levels of MPEG is shown in Table 3.

Table 3. Inverse data rate [ns] with varying NBW

	NBW=1	NBW=2	NBW=4	NBW=8
HL	1.1	2.3	4.5	9.0
H1440	1.5	3.1	6.1	12.2
ML	13.5	27.0	54.1	108.1
LL	46.1	92.2	184.3	368.7

The inverse data rate for several *Rb* is shown in Fig. 4

4.4 Optimal MDL Design Parameters

Optimal MDL design parameter is obtained from the minimum cost and design metrics: DAR and ARD for various candidates of design parameters. The results are shown in Table 4.

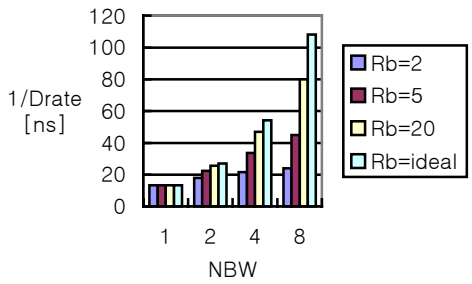

Fig. 4 Inverse data rate for several Rb with varying NBW @ML

Table 4. Optimal MDL design parameters for IDCT/MC

Parameter	MP@LL	MP@ML	MP@H1440	MP@HL
NPE	1	1	4	4
NBW [x32]	1	1	8	8
Fclk [MHz]	15	48	108	150
1/Drate [ns]	46	14	12	9

5. CONCLUSION

The trade-off of MDL design for video signal processing has been explored. From the quantitative analysis of computing and memory access requirements, the relations of NPE (number of processing element), NBW (number of bus width), *Fclk* (clock frequency), and *Drate* (data rate) are obtained. Optimal MDL design parameter is determined from the minimum cost and design metrics; DAR (DRAM Access Rate) and ARD (Area Ratio of DRAM) for various candidates of design parameters. The methodology of this work can be effectively applied to enhance the merging effects of DRAM and logic for video signal processing.

6. REFERENCES

[1] W. K. Luk, et al.,``Development of a High Bandwidth Merged Logic/DRAM Multimedia Chip," *ICCD*, Oct. 1997.

[2] D. Patterson, T. Anderson et al., "Intelligent RAM(IRAM): Chips that Remember and Compute," *ISSCC Digest of Technical Papers*, pp. 224-225, Feb. 1997.

[3] H. Stone, "Multiprocessor Performance," *High Performance Computer Architecture*, Addison Wesley, 1987.

[4] K. Kawabata, "eRAM Technology," *Embedded Systems Conference West*, 1997.

[5] S.Chang, J.S.Kim, and L.S.Kim, "25.6Gbits/s Horizontally and Vertically Accessible Embedded Multi-port SRAM," *Electronics Letters*, 1999, Vol.35 No.21, pp. 1823-1825

[6] T.Masaki, et al,"VLSI Implementation of IDCT and Motion Compensator for MPEG2 HDTV Video Decoding," *IEEE Trans. Circuit and Systems for Video Technology*, 1995, Vol.5, No.5, pp. 387-395.