Energy-Efficient Memory Hierarchy for Motion and Disparity Estimation in Multiview Video Coding

Felipe Sampaio[#1], Bruno Zatt[#2], Luciano Agostini[*3], Sergio Bampi[#4]
[#]*PPGC/PGMICRO, Instituto de Informática, Universidade Federal do Rio Grande do Sul*
Porto Alegre, RS - Brasil
[1]fmsampaio@inf.ufrgs.br
[2]bzatt@inf.ufrgs.br
[4]bampi@inf.ufrgs.br
[*]*GACI, Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas*
Pelotas, RS - Brasil
[3]agostini@inf.ufpel.edu.br

Abstract— **This work presents an energy-efficient memory hierarchy for Motion and Disparity Estimation on Multiview Video Coding employing a Reference Frames-Centered Data Reuse (RCDR) scheme. In RCDR the reference search window becomes the center of the motion/disparity estimation processing flow and calls for processing all blocks requesting its data. By doing so, RCDR avoids multiple search window retransmissions leading to reduced number of external memory accesses, thus memory energy reduction. To deal with out-of-order processing and further reduce external memory traffic, a statistics-based partial results compressor is developed. A content-adaptive reference frame compression scheme is proposed to additionally reduce the external memory communication, while minimizing the error propagation along the MVC structure. The on-chip video memory energy is reduced by employing a statistical power gating scheme and candidate blocks reordering. Experimental results show that our reference-centered memory hierarchy outperforms the state-of-the-art by providing reduction of up to 71% for external memory energy, 88% on-chip memory static energy, and 65% on-chip memory dynamic energy. The reference frame compression improves the external memory savings by 69.5%, on the average.**

Keywords— **Multiview Video Coding, 3D-Video, Low-Power Design, On-Chip Video Memory, Memory Hierarchy, Energy Efficiency, Motion Estimation, Disparity Estimation.**

I. INTRODUCTION AND RELATED WORKS

The state-of-the-art Multiview Video Coding (MVC) [1] standard provides 20%-50% increased coding efficiency in comparison to the H.264/AVC [2]. Besides new syntax elements to support multiview video representation, the key coding tool in MVC is the inter-view prediction, which uses the Disparity Estimation (DE) search to capture the objects displacement due to different camera positions. Along with the Motion Estimation (ME), the DE represents the most energy consuming module in the MVC encoder (more than 90% of the overall energy) [3]. ME/DE is used to search an image region (candidate block) that presents the best matching in the reference frame (previously decoded frames). The search is performed within a search window (SW) using a search algorithm like TZ Search [4]. This search window is typically fetched from external memory and stored in an on-chip video memory. Even for fast search algorithms, frequent memory accesses and large on-chip memory requirements lead to high energy consumption. Moreover, since the memory energy contributes to approximately 90% of the total ME/DE energy consumption [3], on/off-chip memory energy reduction is mandatory to meet the constraints and design requirements posed by the mobile battery-powered devices.

Several already published solutions aimed to deal with memory restrictions regarding the memory issues in the ME/DE on MVC. Some of these works proposed strategies to cache the ME/DE reference samples in two ways: (a) traditional macroblock (MB) centered data reuse (MBDR) [3] and (b) reference-centered data reuse (RCDR) strategies [6][7]. However, the MBDR-based approaches suffer with the increased number of reference frames and search window size required by the MVC encoding. Meanwhile, the RCDR-based solutions do not properly consider the impact of partial results (RCDR penalty) memory traffic and on-chip video memory size. This way, *one goal of our work is to design a RCDR-based memory hierarchy, which is promised to be energy efficient, which takes care of the partial results storage*.

Another set of works aim to reduce the external memory accesses by applying techniques to compress the reference samples before storing them in the external memory. In this case, the ME/DE must recover the original samples (by decompression) to use them as reference in the block matching process. Lossless compression techniques were proposed in [8][9]. Moreover, lossy solutions based on quantization were proposed in [10][11]. Video content characteristics are firstly exploited in [12]. However, these schemes have not been designed for the MVC memory constraints, which require much more ME/DE memory accesses. Besides, as the MVC prediction structure has relatively more dependencies than single-view encoders, errors are propagated along the GOP (Group of Pictures) not only along the temporal neighboring frames, but also for the neighboring views. Therefore, *another goal of our work is to compress the reference data while minimizing the error propagation along the MVC prediction structure*.

Previous works also aimed to reduce the memory related energy consumption of ME/DE on MVC encoders. In [3], we presented a low power ME/DE architecture that uses the concept of search map and dynamic search window formation. The work [13] extends the previous work by employing a multi-sleep state model on-chip memory to reduce the leakage energy. Latest, in [8] we proposed an on-chip memory power management based on such techniques, like the search direction elimination, to reduce even more the on-chip memory energy. These dynamic search window based memory hierarchies lead to irregular memory access and on-chip memory misses. With this in mind, *this work aims to allow regular external memory access pattern, which is more energy efficient, by fetching and storing the entire search window*. Besides, low-power techniques are proposed to reduce the on-chip memory energy consumption.

This work proposes a reference-centered memory hierarchy for ME/DE on MVC targeting low-energy consumption at both on-chip storage and off-chip memory access. The memory hierarchy is composed of an on-chip video memory, an on-chip memory power gating control, a partial search results compressor and a content-adaptive reference frame compression. Additionally, a customized search control unit is

proposed to exploit the search behavior to achieve further energy reduction.

This paper is organized as follows: Section 2 presents the reference-centered based memory hierarchy and all proposed techniques in details; Section 3 shows and discusses the results and the comparisons with related works; finally, Section 4 concludes the paper.

II. Reference-Centered Memory Hierarchy

Fig. 1 presents the architecture of our video memory hierarchy employing reference-centered data reuse scheme.

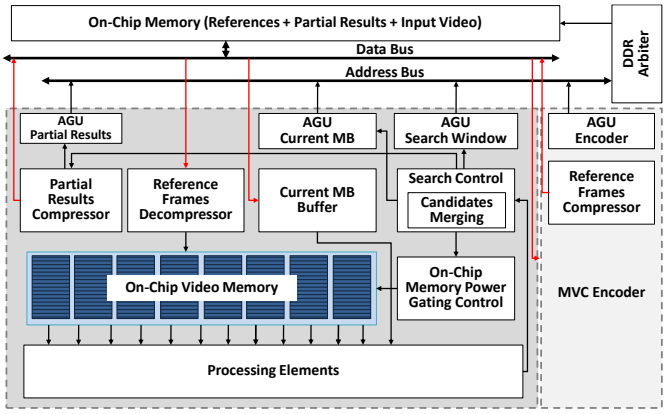

Fig. 1 Memory hierarchy block diagram

To control the ME/DE search and, consequently, the memory access pattern a Search Control unit is defined. Firstly, the Search Control sends search window requests to the memory. These requests are represented in video positions format. Therefore Address Generation Units (AGUs) are used to translate the requests to multiple external memory positions. Before the data is stored on-chip, the search window samples must be decompressed to be recovered (lossy or lossless). Then, a burst of candidate block positions is generated by the Search Control according to the TZ Search algorithm. These candidate positions are rearranged by the energy-aware candidates merging unit in order to reduce the number of on-chip memory line switching. An on-chip memory power gating control monitors the search statistics and power-gates the on-chip memory lines accordingly. The candidates are processed by an array of processing elements (not described in this work). The best matching candidate and its SAD cost (Sum of Absolute Differences) are forwarded to the search control. Due to the out-of-order processing inherent to RCDR, the temporary motion/disparity vectors and SAD (Sum of Absolute Differences) values must be stored for mode decision. These partial results are compressed using statistic-based non-uniform quantization and Huffman coding. As the partial results compressor employs variable-length coding, the partial results data is only sent to external memory once the local buffer is full. A specific AGU is implemented for partial results data.

A. Reference-Centered Data-Reuse

The RCDR uses inverted dependence logic between reference frames and current MB. In this approach, the reference search window is fetched from external memory and those MBs requiring that specific data are processed. In other words, the reference data "calls" the MBs to be processed. For this reason, we define the term *dependent frames* for those frames "called" by a given reference frame.

Fig. 2 depicts the distinctions between search window-based MBDR and RCDR. Observe that in MBDR each current MB requests up to four search windows demanding 4x increased on-chip memory. Additionally, as those search windows are required more times in the future (to encode other frames), external memory data retransmission is needed. In contrast, on-chip storage of a single search window is required for RCDR resulting in reduced on-chip memory. Moreover, in RCDR the search window is requested and read from external memory a single time. Indeed, current MBs belonging to dependent frames are accessed multiple times. However, this represents a small impact for both on/off-chip memories, as demonstrated in results section. The GDV (Global Disparity Vector) is taken into consideration to locate the current MBs positions, as shown in Fig. 2.

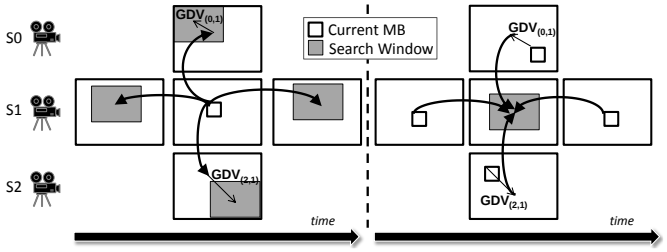

Fig. 2 (a) MBDR vs. (b) RCDR schemes

B. Partial Results Compressor

The partial results are composed of two distinct data types, (i) motion/disparity vectors and (ii) SAD values, that present distinct numerical range and statistical behavior. For this reason, we discuss them separately.

To concentrate these vectors in a reduced range, a median spatial predictor (above and left MBs) was proposed. The differential vectors distribution is concentrated in a small range around DVx=0. For 54 values that represent $\mu \pm 2\sigma = 95.8\%$ of differential vectors occurrences, a Huffman table was generated according to the techniques presented in [15]. The differential vectors values out of this range are represented by a special Huffman value followed by the vector value in 8-bit binary representation.

The SAD values are spread along a wide numerical range. For such distribution, quantization is required. To reduce the impact of quantization errors, we employ a non-uniform quantization designed for a Gaussian distribution according to Lloyd algorithm and a Lloyd-Max refinement [16]. The quantizer employs 512 levels optimized for minimum mean square error (MMSE). After quantization, the quantized SADs are encoded using a 189-entries Huffman table.

C. Content-Adaptive Reference Frame Compression

The proposed scheme compresses the samples after they are completely encoded and reconstructed (i.e., after the Deblocking Filter). At this point, the reconstructed samples are stored in the external memory that are later used as reference in the ME/DE of the subsequent frame(s).

Our scheme is applied to every 4x4 block of a reconstructed Macroblock. Initially, a simplified intra-prediction using only 4x4 blocks is performed to eliminate the spatial redundancies intrinsic to the reconstructed reference samples. In order to avoid additional computation, the proposed scheme inherits the best 4x4 intra mode calculated by the MVC mode decision. Note that our compressor uses the best 4x4 mode regardless of the mode selected for encoding the MB (that may be intra 16x16 or inter-frame/view). The simplified intra-prediction in our scheme is compliant to the H.264/AVC definition for 4x4 blocks: 9 possible modes using thirteen neighboring samples, when available. Then, the residue (difference between the reconstructed and the predicted samples) is calculated. The residue values distribution is much more concentrated when compared to the reconstructed

samples. Exploiting this concentrated distribution, the Huffman-based entropy encoder is applied. Since the intra-prediction exploits the spatial correlation of the image, the heterogeneous (textured) blocks tend to generate spreader distributions of values, which are not desirable for Huffman coding. To better deal with such blocks, the proposed scheme implements a non-linear quantization to further minimize the range of representation.

After the quantization, the initial samples cannot be recovered identically due to the range discretization. It leads to MVC encoder drops on rate-distortion efficiency. Regarding the high number of dependencies on the prediction structures (temporal and disparity domains), these errors may propagate along all ME/DE operations inside the GOP. To handle with this issue, we propose a content-adaptive strategy to adapt the Huffman table and the quantization step according to the image characteristic.

Our content adaptation scheme classifies the 4x4 blocks in four homogeneity groups *HG=[G0,G1,G2,G3]* according to homogeneity degree, which is measured using the statistical variance over the original blocks. Note, the variance calculation is performed using the original blocks, thus avoiding the data dependencies within the encoder loop. Four different non-linear quantizations were defined: *nLev(G0)=8*, *nLev(G1)=16*, *nLev(G2)=32* and *nLev(G3)=256* (lossless), where *nLev* is the number of quantization intervals (levels). They were adapted to achieve the best possible efficiency (joint error and residue minimization) for the specific homogeneity property of each group. Allied to the quantization design, four different static Huffman tables were designed to have the best possible fit with the quantized coefficients of each group. The Huffman encoder is composed of: 8-entry table for *G0*, 16-entry table for *G1*, 32-entry table for *G2*, and 256-entry table for *G3*.

D. On-chip Video Memory Organization

Our on-chip video memory is logically defined as a circular buffer organized in a 2D-array fashion to provide direct matching to video data. It is composed by B logical memory banks that rotate after each search step to avoid retransmission of overlapping SW. This organization, however, is not suitable for physical implementation once ME/DE requires MB parallel read. The physical organization is composed of 16 parallel 128-bits wide SRAM banks to store 16 reference pixels per bank line. Each memory line stores and feed one complete MB in parallel. Each bank is further divided in sectors of n lines representing one search window column. The total number of lines is defined by the number of MBs in the search window. Note that differently from the logical organization, MBs columns are not shifted for every search step. For that the memory sectors are renamed accordingly.

We propose a statistical power gating scheme that employs multiple SRAM sleep modes in order to reduce the static energy consumption due to the leakage current. This scheme does not require image properties extraction or MB-level memory access prediction in order to provide a light-weight (but still efficient) solution. Four power states are implemented [17]: *S0=OFF (Vdd=0)*, *S1= Data Retentive (Vdd=Vdd*0.3)*, *S2= Data Retentive (Vdd=Vdd*0.5)* and *S3= ON (Vdd= Vdd)*. Where each state has an associated wakeup energy cost ($WE_{S0} > WE_{S1} > WE_{S2} > WE_{S3}=0$). For this reason, regions that are frequently accessed are mapped to *S3*, unused regions to *S0*, and other regions are mapped to *S1-S2* according to run-time statistics.

Although static energy is becoming dominant in submicron on-chip memories, dynamic energy reduction significantly contributes to overall energy. To avoid frequent on-chip memory line switching, we also define an energy-aware candidate blocks merging strategy. As far as multiple dependent MBs are searching simultaneously in the same search window, multiple search points are requested multiple times. Our candidate blocks merger receives all search points generated by the search control and rearranges them in order to process together repeated candidates and avoid unnecessary SRAM line switching (address line switching, bitline pre-charge, sense amplifiers switching, output buffer switching). Moreover, the new processing order follows the left-right and up-down fashion so the processing can start even before the rightmost column is updated for each search step

III. Results and Discussions

The experimental results were generated using real video sequences and coding settings recommended by JVT [19] running on the MVC reference software [4]. A customized energy simulator was used to measure the energy consumption of our approach and related solutions. The SRAM leakage and wake-up energies values were calculated based on [18]. The off-chip memory energy savings were evaluated by using the MT46H64M16LF LPDDR 1 Gigabit memory [19]. Note, the experimental results include the wakeup energy overhead.

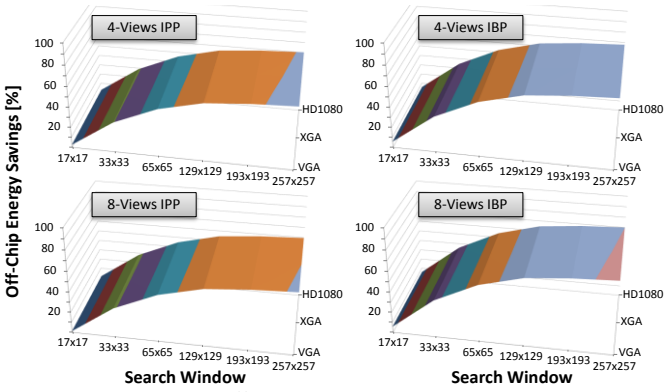

Fig. 3 Off-Chip Memory Energy Savings

Fig. 3 presents the off-chip energy savings for changing search window size and video resolution, under four distinct scenarios, compared to Level-C [3]. Observe that the energy savings scale well with the increase in number of views and search window. Additionally, our solution does not suffer with frame resolution increase. Higher savings happen in case of "IBP" due to more intense search window reuse, i.e., each search window is used by an increased number of current MBs. These results include energy reduction due to the partial results compression. Our compressor leads to 53.2% (average) external energy reduction for partial results communication (compared to the non-compression scenario).

When compared to the non-regular access pattern solution proposed in [3], the proposed RCDR strategy is capable to reduce the DDR energy consumption in 30%. The effective read energy of the RCDR is 2.4x higher than [3], since only the required reference samples are fetched from the off-chip memory. However, the irregular fashion of the memory access of [3] leads with high energy consumption due to the required page activations and, this way, the DDR burst read is not well exploited. In terms of savings, the page activation energy is 98% reduced by the RCDR approach. The page activation energy is dominant in the overall energy consumed by [3]. This way, even the RCDR requiring more read operations, the overall DDR energy consumption is reduced when compared to the irregular search based related work.

Compared to the Level-C, our on-chip video memory size is significantly reduced (see Fig. 4) because there is no need to

simultaneously store multiple search windows on chip. Note that our on-chip memory grows smoothly with the search window increase. Moreover, compared to the search window storage, the cost (considered in Fig. 4) for storing current MBs (that may reach 9 MBs for 8-views "IBP") is negligible. This cost is amortized as the search window increases. The reduced on-chip memory size directly leads to less static energy consumption, as shown in Fig. 5. Compared to [3], 77% energy reduction is reached without employing our power-gating technique. If the power-gating is used, further 88% of reduction is reached outperforming [3] in 61 %. These results refer to 4-views "IBP" scenario.

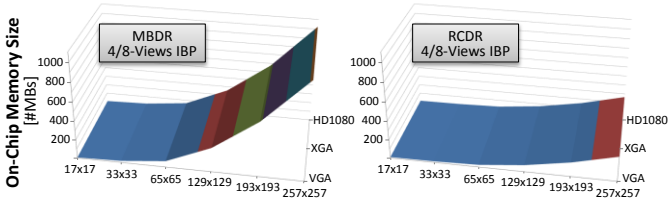

Fig. 4 On-chip memory size reduction

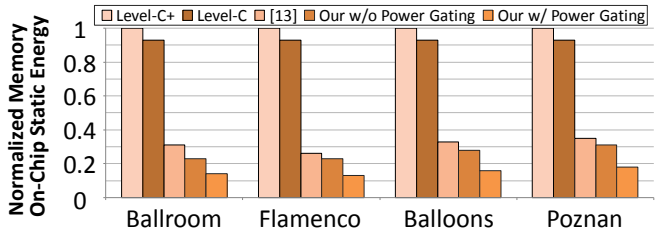

Fig. 5 On-Chip Static Energy Savings (Leakage)

In terms of on-chip dynamic energy, our candidate merging strategy reduces the energy consumption in 65%. At the best of authors' knowledge, this is the first application specific technique to address the dynamic on-chip memory energy consumption for the ME/DE.

Tab. 1 presents the comparison of our content-adaptive scheme with state-of-the-art [12][10][11][8]. It can be noted that none of the works have performed evaluations considering the MVC encoding. In this sense, the error propagation path along the ME/DE references considered in the related works is shorter than that considered in this work, due to the multiple view coding and complex prediction structures.

TABLE I
REFERENCE FRAME COMPRESSOR RESULTS AND COMPARISONS

Parameter	Lossy				Lossless	
	Our *(content-adaptive)*	**[12]**	**[10]**	**[11]**	**Our *(G3-only)***	**[8]**
Target	**MVC**	non-MVC	non-MVC	non-MVC	**MVC**	non-MVC
Content-Adaptive?	**Yes**	Yes	No	No	**No**	No
External Memory Savings	**69.5%**	25-50%	21-31%	17-24%	**51.3%**	24%
BD-PSNR	**-0.01 dB**	-0.04 dB	*N.I.*	-0.01 dB	**0 dB**	0 dB
BD-BR	**0.18%**	1.36-3.92%	0.38-21%	0.7%	**0%**	0%

Tab. 1 shows that our content-adaptive reference compressor is able to reduce the error propagation to achieve as negligible rate-distortion drops as the related single-view reference frame compressor algorithms [12][10][11]. Our content adaptation surpasses the adaptive scheme of [12] in all aspects: i.e., external memory reduction (39% of savings) and rate-distortion efficiency (0.03dB increased BD-PSNR). An additional column in Tab. 1 was inserted to compare [8] with a lossless non-adaptive version of our scheme using only the *G3* configuration. In this comparison, even without content adaptation, our scheme achieves the best results, surpassing the related work external memory savings of [8] by 27.3%.

IV. CONCLUSIONS

A memory hierarchy for Motion and Disparity Estimation on Multiview Video Coding was presented. It exploits a reference-centered data reuse scheme along with partial results compression and memory access scheduling in order to reduce external memory energy. An on-chip video memory organization with line-level power gating and candidates merging scheme is presented targeting on-chip energy reduction. Our memory architecture provides up to 71% off-chip memory energy reduction. On-chip memory-related energy is reduced on 88% and 65% for static and dynamic energies, respectively. The content-adaptive reference frame compression scheme provides 69.5% reduction in the external memory accesses, the best among the related works.

REFERENCES

[1] Joint Draft 8.0 on Multiview Video Coding, JVT-AB204, 2008.

[2] P. Merkle, et al. "Efficient Prediction Structures for Multiview Video Coding." In: IEEE TCSVT, v. 17, n. 11, pp. 1461-1473, nov. 2007.

[3] B. Zatt, M. Shafique, F. Sampaio, L. Agostini, S. Bampi, J. Henkel, "Run-time adaptive energy-aware motion and disparity estimation in multiview video coding", IEEE DAC, pp. 1026-1031, 2011.

[4] JMVC Reference Software , Sep. 2009.

[5] C.-Y. Chen, et al. "Level C+ Data Reuse Scheme for Motion Estimation With Corresponding Coding Orders." In: TCSVT, v. 16, n. 4, p. 553-558, april. 2006.

[6] P.-K. Tsung, et al. "System Bandwidth Analysis of Multiview Video Coding with Precedence Constraint". IEEE ISCAS p. 1001-1004, 2007.

[7] T.-C. Chen, et al, "Single Reference Frame Multiple Current Macroblocks Scheme for Multi-Frame Motion Estimation in H.264/AVC", In IEEE ISCAS, 2005, pp. 1790 – 1793.

[8] D. Silveira, et al, "Memory bandwidth reduction in video coding systems through context adaptive lossless reference frame compression," In: SPL, pp.1-6, 2012.

[9] Z. Wang; et al, "Memory efficient lossless compression of image sequences with JPEG-LS and temporal prediction," In: PCS, pp.305-308, 2012.

[10] Z. Ma and A. Segall. "Frame buffer compression for low-power video coding", In: IEEE ICIP, pp.757-760, 2011.

[11] A. Gupte, et al. "Memory Bandwidth and Power Reduction Using Lossy Reference Frame Compression in Video Encoding", In: IEEE TCSVT, v. 21, n.2, pp.225-230, Feb. 2011.

[12] L. Song, et al. "An adaptive bandwidth reduction scheme for video coding", In: IEEE ISCAS, pp.401-404, 2010.

[13] B. Zatt, M. Shafique, S. Bampi, J. Henkel, "A Low-Power Memory Architecture with Application-Aware Power Management for Motion & Disparity Estimation in Multiview Video Coding", IEEE ICCAD, pp. 40-47, 2011.

[14] M. Shafique, B. Zatt, F. L. Walter, S. Bampi, J. Henkel, "Adaptive Power Management of On-Chip Video Mamory for Multiview Video Coding", IEEE DAC, pp. 866-875, 2012.

[15] Huffman, D.A., "A Method for the Construction of Minimum-Redundancy Codes," IRE, vol.40, no.9, pp.1098-1101, Sept. 1952.

[16] Max, J.; , "Quantizing for minimum distortion," Information Theory, IRE Transactions on , vol.6, no.1, pp.7-12, March 1960.

[17] H. Singh et al., "Enhanced leakage reduction techniques using intermediate strength power gating", IEEE Transaction on Very Large Scale Integration, vol. 15, no. 11, pp. 1215-1224, 2007.

[18] S.l Rodriguez, B. Jacob,"Energy/power breakdown of pipelined nanometer caches (90nm/65nm/45nm/32nm", ISLPED, pp. 25-30, 2006.

[19] JVT. "Com. Test Cond. for Multiview Video Coding". JVT-T207, 2007.

[20] Micron. "1Gb: x16, x32 Mobile LPDDR SDRAM". Available at: <www.micron.com>. Last Accessed: December, 2012.