Lossless Compression Algorithm and Architecture for Reduced Memory Bandwidth Requirement with Improved Prediction Based on the Multiple DPCM Golomb-Rice Algorithm

Imjae Hwang[1], Juwon Yun[1], Woonam Chung[1], Jaeshin Lee[1],
Cheong-Ghil Kim[2], Youngsik Kim[3] and Woo-Chan Park[1,*]

[1] *Sejong University, Seoul, Korea*
[2] *Namseoul University, Cheonan, South Korea*
[3] *Korea Polytechnic University, Gyeonggi Province, South Korea*
E-mail: ijhwang@rayman.sejong.ac.kr; jwyun@rayman.sejong.ac.kr; woonam.chung@gmail.com; jaeshin74@naver.com; cgkim@nsu.ac.kr; kys@kpu.ac.kr; pwchan@sejong.ac.kr
[*] *Corresponding Author*

Received 30 November 2020; Accepted 22 June 2021;
Publication 21 September 2021

Abstract

In a computing environment, higher resolutions generally require more memory bandwidth, which inevitably leads to the consumption more power. This may become critical for the overall performance of mobile devices and graphic processor units with increased amounts of memory access and memory bandwidth. This paper proposes a lossless compression algorithm with a multiple differential pulse-code modulation variable sign code Golomb-Rice to reduce the memory bandwidth requirement. The efficiency of the proposed multiple differential pulse-code modulation is enhanced by selecting the

Journal of Web Engineering, Vol. 20_6, 1813–1828.
doi: 10.13052/jwe1540-9589.2065

optimal differential pulse code modulation mode. The experimental results show compression ratio of 1.99 for high-efficiency video coding image sequences and that the proposed lossless compression hardware can reduce the bus bandwidth requirement.

Keywords: Lossless image compression, hardware architecture, memory bandwidth reduction.

1 Introduction

As the resolution of the latest mobile devices and graphic processor units (GPUs) increases rapidly, the memory bandwidth needed to access the images stored in the frame buffer also increases. This increased amount of memory access influences overall performance and power consumption [1, 2, 9–11], as does memory bandwidth utilization [3, 12]. To this end, lossless frame buffer compression and memory bandwidth reduction methods could be effective ways of equipping a high-performance bus as an IP.

Many studies have been conducted for this purpose [4–6]; the lossless compression algorithm with differential differential pulse code modulation Golomb-Rice encoding (DDPCM-GR), the high-throughput lossless image-compression algorithm with differential pulse code modulation variable sign code GR-encoding (DPCM-VSC GR), and the recompression algorithm with multiple DPCM mode average semi-fixed length coding (MDA-SFL) were proposed in [4, 5], and [6], respectively.

This paper proposes a lossless compression algorithm with a multiple DPCM (MDPCM) variable sign code GR to reduce the memory bandwidth requirement. The algorithm uses MDPCM for prediction and variable sign code (VSC) GR for entropy coding [5].

The structure of the paper is as follows. Section 2 provides an overview of previous studies related to the subject of this work. Section 3 introduces the proposed MDPCM-GR algorithm. Section 4 presents the proposed lossless compression hardware architecture. Section 5 presents the proposed algorithm and hardware performance verification results. The conclusion is given in Section 6.

2 Background

In this section, we review works related to our proposed algorithm and hardware architecture. All the included studies suggest that they have improved the performance based on the DPCM algorithm.

2.1 DPCM Algorithm

DPCM [7] is a widely adopted algorithm for reversible data compression. It consists of the prediction stage and entropy coding of prediction error results. The size of the input pixels' data can be reduced using the spatial locality characteristics of the image.

DDPCM [8] is the algorithm proposed by ATI to reduce the size of Z data in 3D graphics by improving the DPCM. The DDPCM algorithm calculates the differential values for the x-axis and y-axis directions based on the results of the DPCM algorithm.

2.2 DDPCM-GR

DDPCM-GR [4] is the lossless compression algorithm that performs DDPCM as a prediction function and GR encoding for entropy coding for fixed-k, where k is 2. For the processing block in the original image, DDPCM is applied. For the error data from DDPCM, GR encoding is performed. GR encoding is an entropy coding method that applies unary encoding into quotient q, where input value N is divided by parameter M, and the remainder r is stored as binary data.

2.3 DPCM-VSC GR

DPCM-VSC GR [5] is the high-throughput compression algorithm that performs DPCM as a prediction function and VSC GR coding for entropy coding. For the processing block in the original image, 2D DPCM is applied. The quotient value is calculated by dividing the resulting value of DPCM by the $ 2 ∧ k $ value, where k = 0, 1, 2, 3. VSC values and unary code values are obtained by performing GR encoding and VSC.

The lengths of the compressed data and the original block are compared after all stages have been completed. If the length of the compression data is smaller than the length of the original block, compressed data are generated by packing all related data. The related data consist of the K value, the DPCM mode value, the first factor value, the remaining value, the unary data value, and the variable data for sign. If the original block is longer than the compressed data, compressed data is not used and the original block is maintained.

3 Proposed Lossless Compression Algorithm

This section introduces the execution flow of the proposed multiple DPCM GR (MDPCM-GR) algorithm for compression and decompression.

3.1 Overall Process of the Proposed Compression and Decompression Algorithm

The proposed MDPCM-GR algorithm consists of a compressor that compresses the original data and a decompressor that restores the compressed data. The compression process proceeds as follows. First, MDPCM is executed based on the original block data, after which the DPCM mode with the lowest cost is selected from among the DPCM costs calculated. Finally, VSC GR [5] encoding is performed on the error values from MDPCM. Decompression follows the inverse process of compression. Figure 1 shows diagrams of these processes.

3.2 Multiple DPCM Algorithm

The proposed MDPCM algorithm can select a DPCM mode with minimum cost by performing four different directional DPCMs; such as horizontal and vertical DPCM and DDPCM.

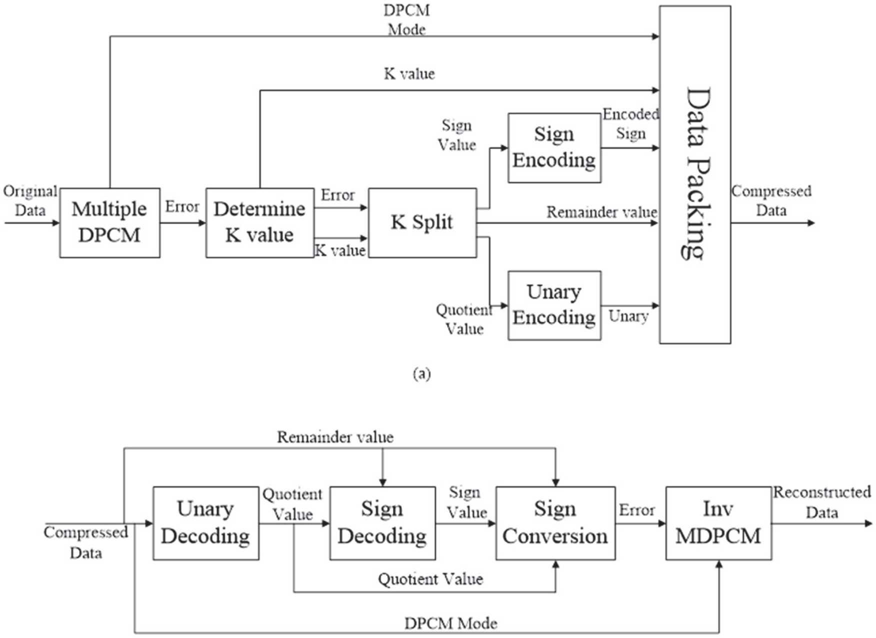

Figure 1 Diagram of the proposed compression algorithm (a) Compression algorithm (b) Decompression algorithm.

DPCM is applied to the 2D image in the horizontal and vertical directions. The equations of the horizontal DPCM (Equation (1)) and vertical DPCM (Equation (2)) are as follows:

$$Err_h(i,j) = \begin{cases} P(i,j) - P(i-1,j), & if\, i \neq 0 \\ P(i,j) - P(i,j-1), & if\, i = 0 \end{cases} \tag{1}$$

$$Err_v(i,j) = \begin{cases} P(i,j) - P(i,j-1), & if\, j \neq 0 \\ P(i,j) - P(i-1,j), & if\, j = 0 \end{cases} \tag{2}$$

The proposed MDPCM uses vertical and horizontal DDPCM, which are performed based on vertical and horizontal DPCM, respectively. When performing DDPCM, the optimal prediction function is selected as the prediction function result, and the entropy encoding cost for the prediction function result is calculated based on the sum of the absolute values of the DPCM results.

The benefits of the proposed MDPCM are as follows. First, because the DPCM and DDPCM are applied both horizontally and vertically, it is predicted for images that have both horizontal and vertical spatial localities. Second, because the DDPCM, which performs an additional DPCM, is applied to the proposed algorithm, it is predicted for the image with rapid pixel difference. Finally, the overhead for applying MDPCM is only 2 bits, which is very low.

4 Proposed Lossless Compression Hardware Architecture

This section introduces the proposed lossless compression hardware architecture and describes each module of the configuration and its operational flow in detail.

4.1 Overall Hardware Architecture of the Proposed Lossless Compression

As shown in Figure 2, the proposed lossless compression hardware architecture consists of the following units: an MDPCM unit for MDPCM execution; a sign converter unit for converting a negative value from the DPCM results into a positive value; data packing unit for packing compressed data; parallel VSC encoding unit; a K splitter unit; a cost calculation unit for the result of dividing K values; and a parallel GR encoder unit for GR parallel encoding.

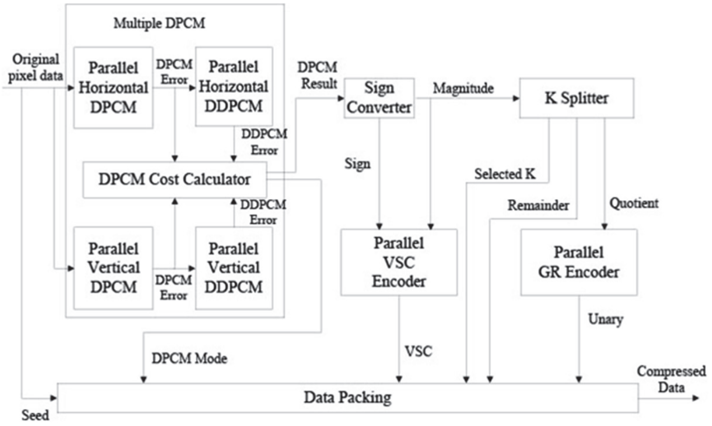

Figure 2 Overall proposed lossless compression hardware architecture.

The process is as follows: to compress the original image, the first pixel data of the original pixel data is transmitted as a seed to the data packing unit. In addition, the original pixel data are transmitted to the MDPCM unit for MDPCM execution. For the transmitted pixel data, both parallel horizontal DPCM and parallel vertical DPCM are performed in the MDPCM unit. Each DPCM error calculated by MDPCM is transmitted to the DPCM's cost calculation unit and the associated DDPCM unit. The parallel horizontal DDPCM unit and parallel vertical DDPCM unit perform DDPCM in parallel based on each DPCM error received. The results of DDPCM performed in parallel are transmitted to the DPCM cost calculator unit.

Based on the received DDPCM result, the DPCM cost calculator determines the cost for each DPCM error. It also selects the DPCM mode with the lowest cost among the calculated costs. After the lowest-cost DPCM mode is selected, it is transmitted to the data packing unit, and the error value of the DPCM is transmitted to the sign converter unit to be converted by the sign converter. The parallel GR encoder generates unary data from the received data as well as VSC data. The generated data are finally transmitted to the data packing unit.

4.2 Overall Hardware Architecture of the Proposed Lossless Decompression

Figure 3 shows the overall proposed lossless decompression hardware architecture, which consists of the following units: a parallel InvMDPCM unit for performing parallel inverse MDPCM; a sign converter unit for sign

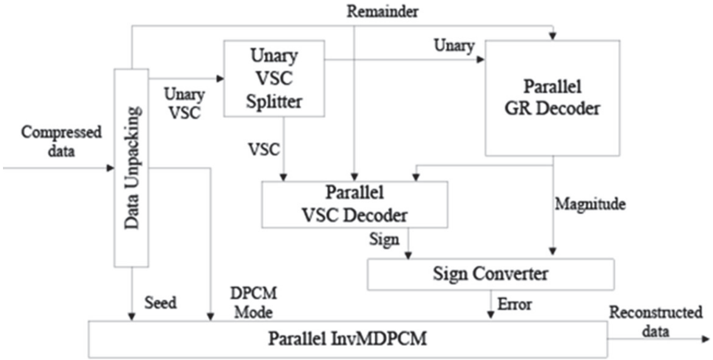

Figure 3 Overall hardware architecture of the proposed lossless decompression.

conversion through sign data and magnitude data; a parallel GR decoder unit for GR decoding in parallel; a zero detector unit to check if the DPCM error value is 0; a parallel VSC decoder unit for performing VSC decoding in parallel; a data unpacking unit that decompresses the compressed data; and a variable unary/VSC splitter unit that splits VSC and unary data.

The hardware of the proposed lossless decompression is performed as follows. First, to decompress the compressed data, the data unpacking unit unpacks the input data. The unsplit unary VSC data received from the unpacking unit are split into unary data and VSC data in the unary VSC splitter unit. The split unary data is transmitted to the parallel GR decoder unit and the zero detector unit, while split VSC data are transmitted to the parallel VSC decoder unit. The zero detector unit checks whether the DPCM error value is zero based on the unary data and the remainder data. The zero-detection result is delivered to the parallel VSC decoder device, which reconstructs the signed data through the received zero-detection result and decodes the VSC.

The unary data are decoded in parallel by the parallel GR decoder unit to restore the quotient data. The magnitude data are restored based on the restored quotient data and the remainder data. When the sign and magnitude data are restored, the sign converter unit reconstructs the DPCM error value. Finally, the parallel InvMDPCM unit works in parallel to DPCM mode based on seed and DPCM error values to reconstruct the pixel data.

5 Experimental Results

In this section, the simulation results of the proposed algorithm and hardware architecture are summarized. The performance will be measured by the compression rate of the algorithm and the bus bandwidth in the hardware structure.

5.1 Compression Ratio

In this paper, we measure the compression ratio (CR) of compressed images to verify the efficiency of the lossless compression algorithm. CR refers to compressed image size per original image size and the higher the CR the more efficient the compression algorithm is. The equation for the CR is as follows.

$$CR = \frac{Original\ image\ size}{Compressed\ image\ size} \tag{3}$$

Seven image sequences in Class A and B of benchmarks of high-efficiency video coding (HEVC) were used to obtain CR measurements in the proposed lossless compression algorithm and hardware architecture. The proposed lossless compression algorithm was compared with those developed in [4, 5], and [6].

Table 1 shows the average CR of the proposed algorithm and of [4, 5], and [6] for all frames of the HEVC image sequences. The experimental results for the HEVC image sequences show that the average CR values are 1.99, 1.63, 1.92, and 1.91 for the proposed lossless compression algorithm, [4, 5], and [6] respectively. The proposed lossless compression algorithm achieves higher CR in all HEVC image sequences than the other algorithms.

The proposed MDPCM algorithm has 2-bit overhead requirement to store DPCM mode information. As shown in the experimental results of Table 1,

Table 1 CR for HEVC image sequences

Image	Proposed	[4]	[5]	[6]
Traffic	2.13	1.70	2.04	2.03
People on street	2.15	1.70	2.01	2.08
Kimono	2.17	1.76	2.11	2.05
Park scene	1.90	1.62	1.86	1.82
Cactus	1.81	1.56	1.77	1.73
Basketball drive	2.00	1.62	1.94	1.90
BQ terrace	1.80	1.46	1.71	1.73
Avg.	1.99	1.63	1.92	1.91

it is clear that the increased compression rate of the prediction function results of the proposed MDPCM algorithm is sufficient to fulfil the overhead requirement.

5.2 Performance Analysis for Bus Bandwidth

The CR of the proposed compression algorithm may be limited by bus bandwidth because the data transmitted over the bus should be aligned to its width. Therefore, the data transmitted over the bus may lose the CR by as much as the bandwidth of the aligned data.

To analyze this, we measured the actual average CR of the traffic scene with five different bus bandwidths of 8, 16, 32, 64, and 128 bits. Figure 4 shows the experimental results of the actual CRs on them. Here, the actual CR is limited to a 128-bit bus width, the most common bandwidth. However, the actual CR on the 8-bit bus width is similar to the ideal case of the 1-bit one, which means that the performance of the proposed lossless compression hardware is good on the 8-bit bus bandwidth. Consequently, the proposed lossless compression hardware can reduce bus bandwidth requirement.

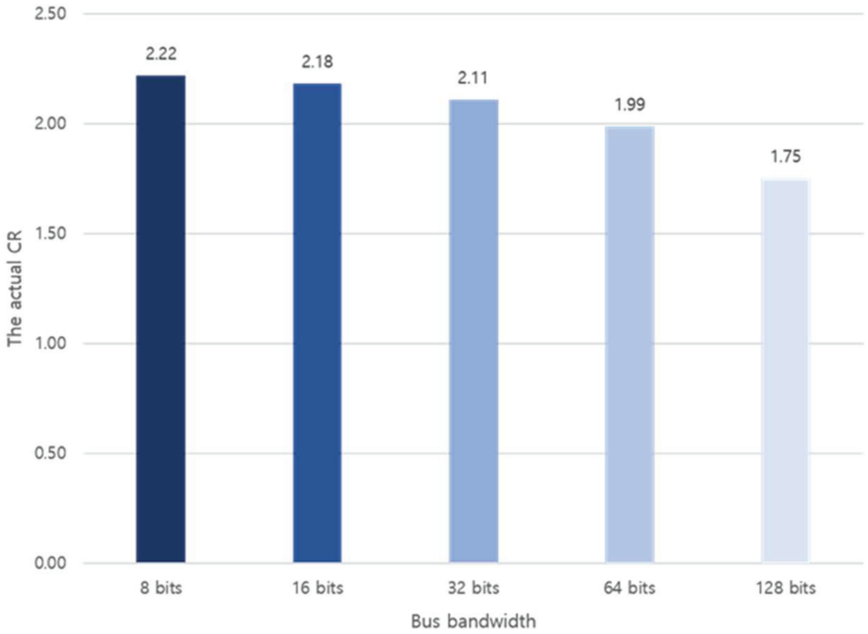

Figure 4 The experimental results of actual CR on five different bus bandwidth.

6 Conclusion

In this paper, a lossless compression algorithm and hardware architecture are proposed to reduce memory bandwidth requirements. The proposed hardware architecture has a high processing speed because the compression and decompression steps can be performed in parallel. In addition, high CR was achieved by improving the existing algorithm's prediction function. The prediction function also has improved performance, reflecting the advantages of both DPCM and DDPCM. We used the HEVC image sequence as a benchmark to verify the performance of the proposed algorithm and hardware architecture.

As a result of the experiment, we achieved an average CR of 1.99 in Classes A and B of the HEVC image sequence. Through this, the proposed lossless compression architecture was verified to achieve a higher CR than the comparison algorithms. In addition, as this study's algorithm achieves a high CR, the proposed lossless compression hardware can reduce bus bandwidth requirements. Thus, the proposed lossless compression hardware is suitable for applications requiring high memory bandwidth and memory access, such as GPUs and high-resolution video processors.

Acknowledgement

This research was supported in part by the MOTIE (Ministry of Trade, Industry & Energy) (10080568, Development of the Embedded GP-GPU for Deep Learning based Circumstances Recognition) and KSRC (Korea Semiconductor Research Consortium) support program for the development of the future semiconductor device. The EDA tool was supported by the IC Design Education Center (IDEC), Korea.

References

[1] D. Burger, J. R. Goodman, A. Kagi, 'Limited bandwidth to affect processor design', IEEE Micro, vol. 17, no. 6, pp. 55–62, Nov. 1997.

[2] H. David, E. Gorbatov, U. R. Hanebutte, R. Khanna, C. Le, 'RAPL: memory power estimation and capping', In 2010 ACM/IEEE International Symposium on Low-Power Electronics and Design (ISLPED), pp. 189–194, Aug. 2010.

[3] ARM Mali GPU OpenGL ES Application Optimization Guide, Available on. https://developer.arm.com/docs/dui0555/b/optimization-checklist/the-checklist/reduce-memory-bandwidth-usage

[4] H.-S. Kim, J.-H. Lee, H.-J. Kim, S.-H. Kang, W.-C. Park, 'A Lossless Color Image Compression Architecture Using a Parallel Golomb-Rice Hardware Codec', IEEE Transactions on Circuits and Systems for Video Technology, vol. 21, no. 11, pp. 1581–1587, Nov. 2011.

[5] J. Lee, J. Yun, J. Lee, I. Hwang, D. Hong, Y. Kim, C. G. Kim, W.-C. Park, 'An Effective Algorithm and Architecture for the High-Throughput Lossless Compression of High-Resolution Images', IEEE Access, Vol. 7, Issue 1, pp. 138803–138815. Sep 2019.

[6] L. Guo, D. Zhou, S. Goto, 'A new reference frame recompression algorithm and its VLSI architecture for UHD TV video codec', IEEE Transactions on Multimedia, vol. 16, pp. 2323–2332, Dec. 2014.

[7] A. D. Mitra, P. K. Srimani, 'Differential pulse-code modulation', Int. J. Electron., vol. 46, pp. 633–637, Jun. 1972.

[8] S. Morein, 'ATI radeon hyperz technology', In Proceedings of the Graphics Hardware, 2000.

[9] D. Silveira, G. Povala, L. Amaral, B. Zatt, L. Agostini, M. Proto, 'Efficient reference frame compression scheme for video coding system: algorithm and VLSI design', Journal of Real-Time Image Processing 16, pp. 391–411, 2019.

[10] Yu-Hsuan Lee, Tzu-Chieh Chen, Hsuan-Chi Liang, Jian-Xiang Liao, 'Algorithm and Architecture Design of FAST-C Image Corner Detection Engine', Very Large Scale Integration (VLSI) System IEEE Transaction on, vol. 29, no. 4, pp. 788–799, 2021.

[11] Sungchul Yoon, Sungho Jun, Yongkwon Cho, Kilwhan Lee, Hyukjae Jang, Tae Hee Han, 'Optimized Lossless Embedded Compression for Mobile Multimedia Applications', Electronics, vol. 9, p. 868, 2020.

[12] Yu-Hsuan Lee, Cheng-Hung Kuei, Yue-Zhan Kao, Shih-Song Fan Jiang, 'Algorithm and VLSI Architecture Designs of A Lossless Embedded Compression Encoder for HD Video Coding Systems', Journal of Circuits, Systems and Computers, 2020.

Biographies

Imjae Hwang received the B.S. and Ph.D. degree in Internet engineering from Sejong University, Seoul, Korea in 2012. He is currently doctoral student in Computer engineering, Sejong University, Seoul, Korea. His current research interests include 3-D rendering processor, high performance computing, real-time ray tracing and lossless compression.

Juwon Yun was born in South Korea, in 1986. He received the B.S. degree from the Department of Game and Multimedia Engineering, Korea Polytechnic University, Siheung, South Korea, in 2013, and the M.S. and Ph.D. degrees in computer engineering from Sejong University, Seoul, South Korea, in 2020. His current research interests include sound tracing, game engine, mobile GPU, and computer graphics.

Woonam Chung is Senior Engineer, Sejong University. He received his BS, MS and PhD in Computer Science from Yonsei University, Korea. His current research interests include, Global illumination, real-time ray tracing GPUs, 3D graphics algorithms and applications.

Jaeshin Lee received the B.S. degree in electronic engineering from Kyunghee University, Yongin, Korea, in 1997 and the ph.D in computer engineering from Sejong University, Seoul, Korea in 2017. She joined Samsung Electronics Co., Ltd in 1997 and worked as a principal engineer for 6 years until 2020. Her current research interests include image compression algorithm and parallel processing architecture, hardware IP and SOC Architecture targeting power reduction, performance enhancement, and area reduction.

Cheong-Ghil Kim received the B.S. in Computer Science from University of Redlands, CA, U.S.A. in 1987. He received the M.S. and Ph.D. degree in Computer Science from Yonsei University, Korea, in 2003 and 2006, respectively. Currently, he is a professor at the Department of Computer Science, Namseoul University, Korea. His research areas include Multimedia Embedded Systems, Mobile AR, and 3D Contents. He is a member of IEEE.

Youngsik Kim received the B.S., M.S., and Ph.D. degree in Dept. Computer Science from the Yonsei University, Korea, in 1993, 1995, and 1999 respectively. He had worked for System LSI, Samsung Electronics Co. Ltd from Aug. 1999 to Feb. 2005 as a senior engineer. Since March 2005 he has been working for Dept. of Game & Multimedia Engineering in Korea Polytechnic University. His research interests are in 3D Graphics and Multimedia Architectures, Game Programming, and SOC designs.

Woo-Chan Park received the B.S., M.S., and Ph.D. degrees in Computer science from Yonsei University, Seoul, Korea, in 1993, 1995, and 2000, respectively. From 2001 to 2003, he was a Research Professor with Yonsei University. He is currently a Professor of Computer engineering, Sejong University, Seoul. His current research interests include ray tracing processor architecture, 3-D rendering processor architecture, real-time rendering, advanced computer architecture, computer arithmetic, lossless image compression hardware, and application-specific integrated circuit design.