A case for multi-channel memories in video recording

Eero Aho, Jari Nikara, Petri A. Tuominen, and Kimmo Kuusilinna
Nokia Research Center
Tampere, Finland
eero.aho@nokia.com

Abstract – **In video recording, ever increasing demands on image resolution, frame rate, and quality necessitate a lot of memory bandwidth and energy. This paper presents and evaluates such a potential memory load in future handheld multimedia devices. Based on the achieved simulation results, the multi-channel memories provide the capability for high bandwidth without excessive overhead in terms of energy consumption. A full HDTV (1080p) quality video recording with H.264/AVC encoding at 30 frames per second (fps) is found here to require 4.3 GB/s memory bandwidth. According to the simulations, this memory requirement can be fulfilled with four 32-bit memory channels operating at 400 MHz and consuming 345 mW of power. As another example, 400 MHz 8-channel memory configuration is able to provide the required bandwidth for video recording with up to 3840x2160@30 fps. Die stacking is the technology thought to be able to provide the required bandwidth, sufficiently low power consumption, and the multi-channel memory organization.**

I. Introduction

Video recording with H.264/MPEG-4 AVC video encoding [1] requires a huge execution memory bandwidth. For example, a software H.264/AVC encoder implementation processing HDTV quality stream (720p@30fps) has been estimated to require memory access at the rate of 5570 GB/s [2]. However, most of the bandwidth can be supplied by the cache memory. As shown later in this paper, the bandwidth requirement for the whole video recording chain (720p@30fps) can be diminished down to 1.9 GB/s with appropriate caching. Nevertheless, future multimedia devices still have increased bandwidth requirements arising from bigger resolutions and larger frame rates that make the execution memory design a challenging task.

In low-power hand-held devices, multi-channel memories provide a way to attain the required bandwidth. The enabling technology for multi-channel memory organization is die stacking or 3D integration [3], [4], [5].

This paper evaluates the multi-channel memory benefits when the main memory load is predominantly from a single source (cache misses), produced e.g. with a symmetric multi-processor (SMP). The use case is video recording including H.264/MPEG-4 AVC encoding and the simulated multi-channel memory configuration has up to eight channels. Each of the channel contains a theoretical next generation mobile DDR (Double Data Rate) SDRAM.

Details of the video recording use case are introduced in Section II. Section III describes the simulated memory architecture model and analysis results are given in Section IV. Section V concludes the paper.

II. Memory Load: Video Recording

There are a multitude of alternative implementations of the video recording use case which trade the memory bandwidth requirements to something else, typically local embedded memories and custom hardware. It should be emphasized that this paper intentionally does not attempt to optimize the de/encoding algorithms. For instance, the use of reference frames is strictly limited in contemporary practical implementations which is not the case here.

The analyzed video recording use case is depicted in Fig. 1. The use case was prepared according to the literature review [2], [6], [7], [8], [9]. The subsequent analysis is based on the assumption that the use case is implemented in a symmetric multi-core environment (SMP) with multi-level cache architecture and external main execution memory.

The use case contains two main processing parts: image processing and video coding (Fig. 1). At the beginning, the video stream originates from the image sensor and it is buffered in execution memory. After various processing steps, including H.264 encoding, the video stream is multiplexed with the corresponding audio stream and stored in removable media. While this process is ongoing, the stream must also be presented on the device display. The memory traffic to execution memory is indicated either as the function of the currently processed image size [pixels] or the maximum stream size [Mbits/s]. It is further assumed that the device display is capable of presenting WVGA images. The video encoding exhibits an implementation dependent constant factor that is estimated to be six for the purposes of this analysis.

978-3-9810801-5-5/DATE09 © 2009 EDAA

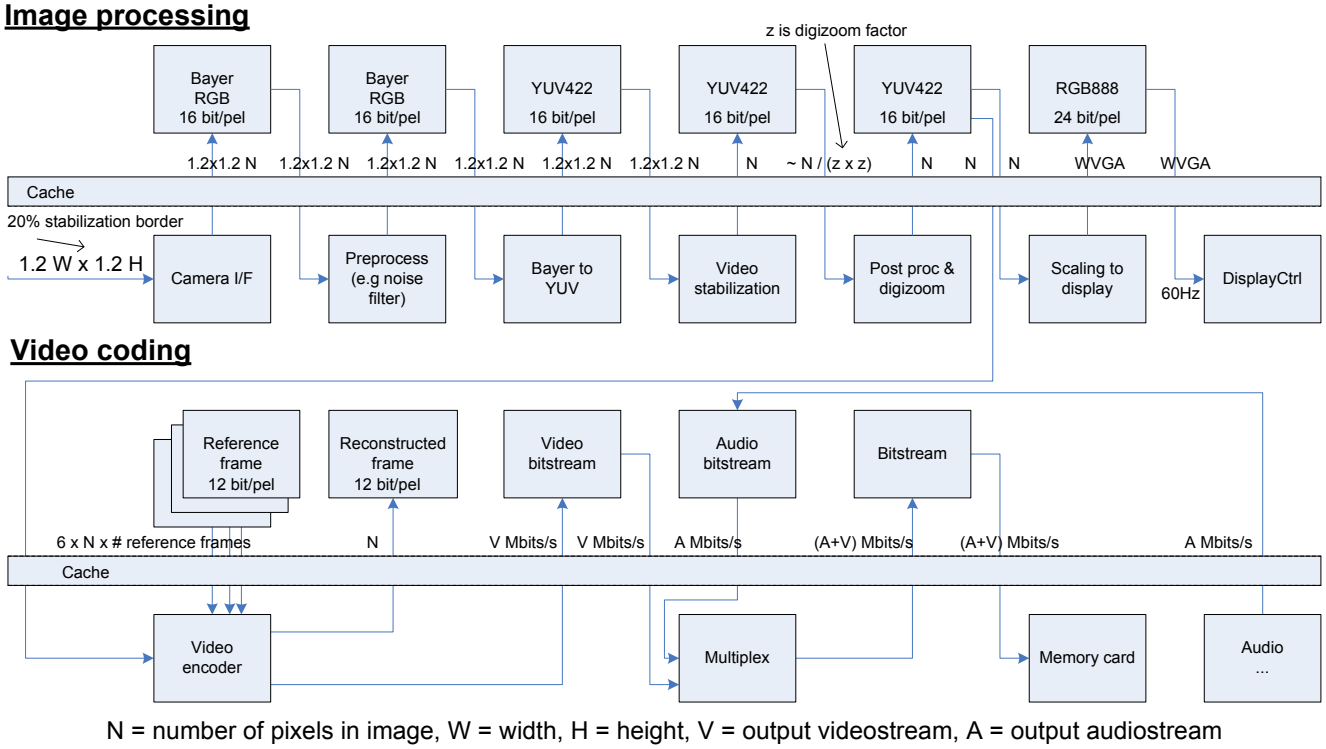

Fig. 1. Video recording use case.

The analyses in this paper assume that the cache is large enough to provide hits for any other memory access than the ones depicted in Fig. 1. The memory load from instructions is relatively insignificant and not taken into account here.

In the video recording use case, the single most memory intensive part is the video encoding. This is mainly due to the fact that video encoding utilizes not only spatial redundancy in information but time redundancy as well. That is, previously encoded (and then reconstructed) pictures are needed to encode the current frame. Since several pictures may be needed and their sizes are considerable, it is not reasonable to assume that the relevant picture data would always be present in the cache. The situation is made even worse by the fact that the reference frames can be from the past and/or the future.

TABLE I tabulates the memory bandwidth requirements for the shown video recording use case and for the five HD compatible encoding levels defined by H.264/AVC [1]. The image size is 1280x720, 1920x1088, or 3840x2160 pixels (pels). In addition, the maximum frame rate that needs to be supported ('Limits') is either 30 fps or 60 fps. The maximum bitrate defines the largest possible output stream bandwidth [Mb/s]. Bayer RGB and YUV422 encodings use 16 bits to store one pixel and, correspondingly, H.264 encoded frames require 12 bits (YUV420) and the displayed RGB888 format needs 24 bits per pixel.

Reads and writes are assumed identical operations with respect to examining the memory bandwidth. Therefore, the bandwidth numbers for each processing step combine the traffic caused by both consumption and production of data. The DisplayCtrl processing is assumed to have constant memory requirements regardless of original image size. This is natural since the display size stays the same in this use case. The total data memory load for one frame is the sum of the image processing and video encoding parts. As an example, for 1080 HD at 60 fps, the total execution memory bandwidth requirement is estimated to be 8.6 GB/s. Test material for big resolution videos (720p, 1080p, 2160p) can be found, for example, in [10].

III. Memory Architecture and Simulation Model

For the purposes of building a simulation model, the multi-channel memory architecture is illustrated in Fig. 2. The architecture was modeled and simulated with a commercially available SystemC electronic system level (ESL) design environment. The models are untimed transaction level models (TLMs) associated with separate timing and power information.

The parts of the overall system loading the memory, e.g., symmetric multi-processing cores (SMP), hardware accelerators (HWA), caches ($), etc., have been abstracted

TABLE I. MEMORY BANDWIDTH REQUIREMENT FOR THE STAGES OF THE VIDEO RECORDING USE CASE ($M = 1024^2$).

	H.264/AVC Level	3.1	3.2	4	4.2	5.1
	Format	720p HD	720p HD	1080p HD	1080p HD	UHD
IMAGE TYPES	Width [pel]	1280	1280	1920	1920	3840
	Height [pel]	720	720	1088	1088	2160
	Limits [fps]	30	60	30	60	30
	Max bitrate [Mb/s]	13.35	19.07	19.07	47.68	228.88
	Bits per pel	16	16	16	16	16
	Camera I/F [Mb]	20.25	20.25	45.90	45.90	182.25
IMAGE PROCESSING	Preprocess [Mb]	40.50	40.50	91.80	91.80	364.50
(numbers in bits	Bayer to YUV [Mb]	40.50	40.50	91.80	91.80	364.50
per frame)	Video stabilization [Mb]	34.31	34.31	77.78	77.78	308.81
	Post proc & digizoom [Mb]	28.13	28.13	63.75	63.75	253.13
	Scaling to display [Mb]	25.05	25.05	42.86	42.86	137.55
	DisplayCtrl [Mb]	10.99	10.99	10.99	10.99	10.99
	Image proc. total (1 frame) [Mb]	199.72	199.72	424.87	424.87	1621.72
	Bits per pel	12	12	12	12	12
VIDEO	Nb of reference frames	5	5	5	5	5
CODING	Video encoder [Mb]	330.94	330.80	749.73	749.90	2982.22
(numbers in bits	Multiplex [Mb]	0.95	0.67	1.35	1.67	16.02
per frame)	Memory card [Mb]	0.47	0.34	0.67	0.84	8.01
	Video coding total (1 frame) [Mb]	332.36	331.81	751.75	752.41	3006.24
	Data Mem. load (1 frame) [Mb]	532.08	531.54	1176.63	1177.28	4627.97
TOTAL	Data Mem. load (1s) [Mb]	15962.46	31892.20	35298.77	70636.82	138838.96
	Data Mem. load [MB/s]	1995	3987	4412	8830	17355

into a *load model*, which generates just read and write access requests to the memory subsystem. The memory subsystem contains M parallel channels of memory controller (MC), DRAM interconnect and bank cluster (BC). A bank cluster contains one or more memory banks.

A load model encapsulates everything else but the memory controllers, DRAM interconnects, and bank clusters. The use case is running on the SMP that acts as a memory master causing *master transactions*, i.e., memory access request that are mapped onto channels. The load model is based on the video recording use case (Fig. 1) that represents very regular and foreseeable memory access behavior, i.e., it needs relatively large data amounts resulting in several memory accesses to sequential memory locations. Within the load model, the processing chain of the video recording is described as a state machine. Each state results in memory access requests. The use case is assumed to be mapped completely onto the SMP and therefore, the memory access requests are always through a cache.

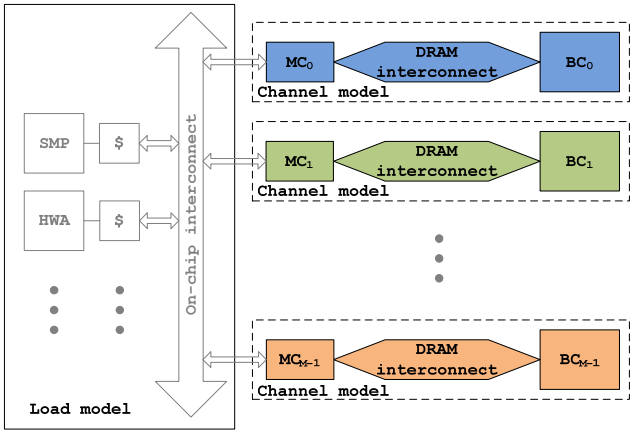

Fig. 2. A block diagram of the simulated multi-channel memory architecture.

TABLE II. MEMORY MAPPING OVER CHANNELS.

Address	0	16	32	…	16x(*M*-1)	16*M*	…
BC	BC_0	BC_1	BC_2	…	BC_{M-1}	BC_0	…

A memory controller, DRAM interconnect, and bank cluster form an entity called *channel model*. The delay and power consumption figures in the simulations are attained from the channel model. The memory controller takes care of memory mappings onto banks, rows and columns of the bank cluster. Another task of the controller is to manage all the DRAM operations: precharges, activations, reads, writes, refreshes, and power downs. For example, the read access to a closed page (row) includes the precharge operation, the activation of the correct row, and the read operation of the correct column. When data is read from an open page, only the read operation is needed. The memory controller takes also care of the data refresh, done periodically for all DRAM banks, and the power down schemes used to save energy when normal operations are not performed. For maximum energy savings, it is assumed that bank clusters go to power down states after the first idle clock cycle.

The memory controller is connected to the bank cluster with the DRAM interconnect. The bank clusters are based on our best estimations on the next generation mobile DDR SDRAM. Estimation is necessary since no 3D integration compatible standard memory components exist at this time. The capacity of a bank cluster is 512 Mb. The address space of the memory is divided to banks, rows, and columns. The bank cluster contains four banks. The word width of a data access is 32 bits. The interface clock frequency of the evaluated next generation mobile DDR SDRAM is restricted from 200 to 533 MHz according to DDR2 specification [11]. However, the memory is double data rate (DDR), i.e., the data is accessed on both of the clock edges.

The timing and power parameters for the used next generation mobile DDR SDRAM are estimated according to the contemporary Mobile DDR SDRAM devices that use clock frequencies from 133 to 200 MHz [12], [13], [14]. The parameters with clear connection to clock frequency are extrapolated accordingly. The other parameters are used exactly as they are denoted in the utilized Mobile DDR SDRAM datasheet for 200 MHz [12]. As an exception, the used operating voltage is projected to be 1.35 V according to the current development in newer processes [15].

The data can be mapped to the channels in several ways. In this study, the maximum bandwidth for a single use case is desired. Therefore, the data for the channels is interleaved in such a way that all the channels can be used in a single master transaction. TABLE II depicts the memory mapping over the channels. Byte addressable memory is used, minimum DRAM burst size is four, and word length is 32 bits (4 bytes). This makes minimum practical interleaving granularity 16 (= 4x4). For example, addresses from 0 to 15 are located in bank cluster zero and addresses from 16 to 31 in bank cluster one.

In the modeling, the interface power consumption is not taken into account. Instead, the analysis assumes the estimate for the interface power per channel as [16].

$$interface\ power = nr\ of\ \ pins \times C \times V^2 \times f_{clk} \times activity. \quad (1)$$

The number of pins toggling during a burst, *nr of pins*, is assumed to be 36 (data bus and data strobe signals) [12]. For the capacitance value, denoted *C*, the expected value for 3D chip-to-chip connection is 0.4 pF that is an average capacitance value of the bonding techniques (wire bonding, flip chip, and tape automated bonding) given in [17]. The voltage *V* is the I/O voltage, estimated for next generation devices as 1.2 V. Clock frequency is denoted by f_{clk} and *activity* is fixed to be 50%. As an example, with 400 MHz clock frequency, these assumptions result in the approximate interface power of 5 mW per channel.

IV. ANALYSIS

This section summarizes the result involving the most critical measures: access time and power consumption. The purpose is to keep focus on trends rather than in certain specific results in order to have better understanding of different configuration options.

In all the evaluations, DRAM open page policy is used. The address multiplexing type defines how the DRAM input address is mapped to bank address, row address, and column address. The shown results utilize Row-Bank-Column (RBC) address multiplexing type since somewhat better performance were achieved compared to the Bank-Row-Column (BRC) multiplexing type.

The first analysis is made to determine the clock frequency that just fulfills the requirements of the high definition video recording use case. The least demanding HDTV quality H.264/AVC level is 3.1 implying the resolution of 1280x720 with the frame rate of 30 fps. The obtained total access times with a varying clock frequency are shown in Fig. 3. The main trend is that close to 2x speedup can be achieved by using double clock frequency or double the number of exploited channels.

The frame rate of 30 fps means that the real time requirement for a single frame is 33 ms, which is highlighted with a red line in Fig. 3. When considering the 1-channel configuration, the first two frequencies 200 and 266 MHz cannot meet the performance requirements, since already the memory access time is clearly over the real-time requirement. Also the first clock frequency with the 1-channel configuration meeting the requirement from the access time perspective (333 MHz, marked “marginal” in

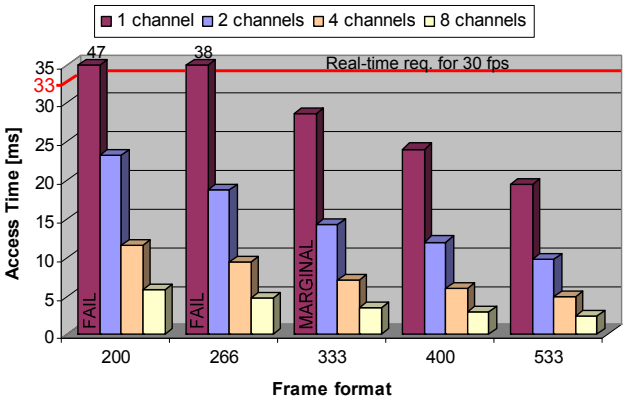

Fig. 3. Effect of memory clock frequency on memory access time. One frame encoded (1280x720@30fps).

Fig. 3), is on the edge because the memory access time cannot in reality be driven too close to real-time requirements. The system rarely runs only a single use case and some margin is needed also for data processing.

Therefore, one conclusion from this access time analysis is that at least two channels are required to satisfy the real-time requirements of the 720p HDTV with all the examined DDR2 clock frequencies. Alternatively, one may conclude that in order to keep the channel count open, at least 400 MHz memory is required for the high definition video recording.

Let us next assume the memory clock frequency as 400 MHz. Fig. 4 illustrates how well such execution memory manages to serve different video recording loads. As was already seen in Fig. 3, H.264/AVC level 3.1 is achievable with all interleaving schemes. Level 3.2 (1280x720@60 fps) requires at least two channels (Fig. 4). Encoding the full HDTV (1920x1080 pixels) with frame rate of 30 fps requires approximately 2.2 times more memory bandwidth compared to 720p@30fps. In order to be on the safe side regarding the real time requirements, 1080p@30fps employs at minimum four channels. The frame format 1920x1080@60fps and 3840x2160@30fps need all eight channels. In any case, 2160p@30fps format starts to be already on doubtful whether any of the memory configurations is capable of meeting the real time requirements.

Fig. 5 presents the power consumptions of the memory subsystem with respect to the different frame formats. The memory clock frequency is still 400MHz. The Figure contains the computed (with (1)) interface power with the expected chip-to-chip pin capacitance 0.4 pF and 1.2 V I/O voltage. The interface power is marked with dark on top of the bars. Bars with zero values mean that the memory subsystem configuration cannot meet the real time requirements with a 15 % margin for the data processing.

As mentioned, strict power saving modes are assumed; the bank clusters go to power down states after the first idle clock cycle. Therefore, the increase in power consumption is moderate when comparing multi-channel to single-channel configuration. With a single channel, average power consumption for 720p@30fps is 150 mW whereas 8-channel configuration demands 205 mW. Video recording for a frame format 1080p@30fps with four channels consumes 345 mW and 3840x2160@30fps with 8-channel configuration requires a huge amount of power, up to 1280 mW.

For comparison, the Cell Broadband Engine (Cell BE) contains a dual XDR DRAM memory interface. The XDR memory interface operating with 1.6 GHz clock frequency

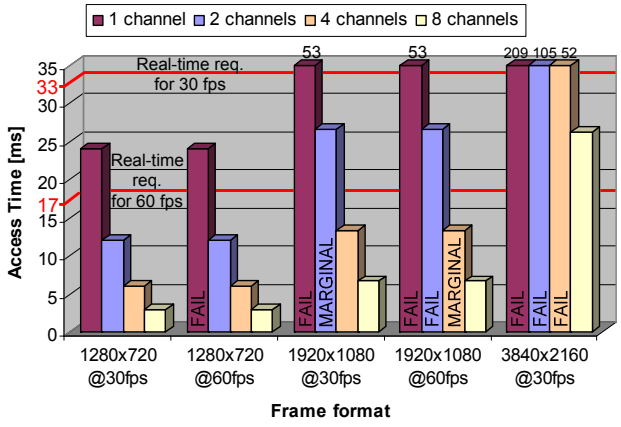

Fig. 4. Effect of encoding format on memory access time (Clock frequency is 400MHz).

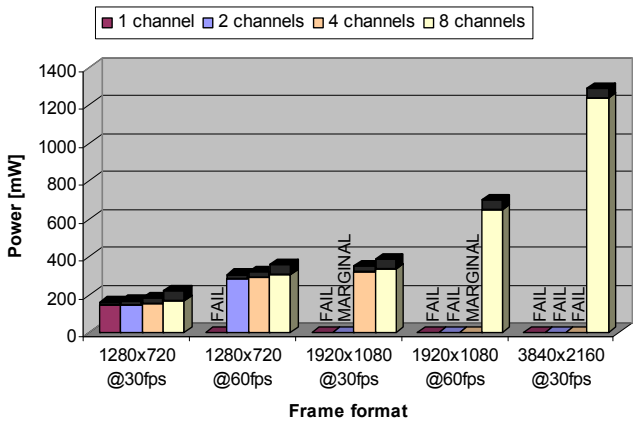

Fig. 5. Effect of encoding format on memory power consumption (Clock frequency is 400MHz).

acquires 25.6 GB/s bandwidth and consumes typically power of 5 W [18]. According to this study, the proposed theoretical next generation mobile DDR SDRAM with eight channels and 400 MHz clock frequency has similar bandwidth (25.0 GB/s) but power consumption from 4% to 25% of the XDR value, depending on the used encoding format (Fig. 5).

V. CONCLUSIONS

Multi-channel memories are an attractive option for future mobile multimedia devices since they provide the capability for high bandwidth but, when used properly, do not introduce excessive overhead compared to single-channel memories in terms of energy consumption. When the memory bandwidth requirements become high enough, multiple channels are the only way to go.

To summarize the simulation results over the different H.264 levels, we conclude that the multi-channel memory subsystem configuration scales well for future needs. With 400 MHz clock frequency, the H.264 level 3.2 (1280x720p@60 fps) is already the high definition video recording use case that clearly needs several channels from the execution memory. Respectively, the H.264 level 4 (1920x1080p@30 fps) requires the 4-channel configuration. The 8-channel configuration is needed for the H.264 level 4.2 (1920x1080p@60 fps) and it is capable of performing the memory accesses up to the H.264 level 5.2 (3840x2160p@30 fps). For reference, XDR memory can give comparable bandwidth but consumes up to 25 times more power than the proposed one.

In future systems, where the memory loads exceed the HDTV requirement, novel policies, advanced control mechanisms, and reorganization of traditional memory management are needed to keep the power consumption manageable. For instance, it may be necessary to divide very large multi-channel memories into independent channel clusters, each consisting of reasonable number of channels. Naturally, aggressive use of power-down modes is necessary for energy efficient operation with handheld devices.

REFERENCES

[1] ITU-T Recommendation H.264, "Advanced video coding for generic audiovisual services," *International Telecommunication Union*, March 2005.

[2] T-C Chen, S-Y Chien, Y-W Huang, C-H Tsai, C-Y Chen, T-W Chen, and L-G Chen, "Analysis and architecture design of an HDTV720p30 frames/s H.264/AVC encoder," *IEEE Trans Circuits Syst. Video Technol.*, vol. 16, no. 6, June 2006, pp. 673-688.

[3] P. Jacob, O. Erdogan, A. Zia, P. M. Belemjian, R. P. Kraft, and J. F. McDonald, "Predicting the performance of a 3D processor-memory chip stack," *IEEE Design Test Computers*, vol. 22, no. 6, Nov. 2005, pp. 540-547.

[4] G. L. Loi, B. Agrawal, N. Srivastava, S.-C. Lin, T. Sherwood, and K. Banerjee, "A thermally-aware performance analysis of vertically integrated (3-D) processor-memory hierarchy," *Proc. ACM/IEEE Design Automation Conf.*, July 2006, pp. 991-996.

[5] C. C. Liu, I. Ganusov, M. Burtscher, and S. Tiwari, "Bridging the processor-memory performance gap with 3D IC technology," *IEEE Design Test Computers*, vol. 22, no. 6, Nov. 2005 pp. 556-564.

[6] H. H. Chen, C.-K. Liang, Y.-C. Peng, and H.-A. Chang, "Integration of digital stabilizer with video codec for digital video cameras," *IEEE Trans Circuits Syst. Video Technol.*, vol. 17, no. 7, July 2007, pp. 801-813.

[7] T. Kurashige, J. Shiokawa, H. Chiba, N. Yamamoto, T. Kitade, H. Tarumizu, H. Kami, and T. Imai, "Development of MPEG camera," *Proc. IEEE Int. Symp. Consumer Electronics*, Dec. 1997, pp. 218-221.

[8] K. Illgner, H.-G. Gruber, P. Gelabert, J. Liang,Y. Yoo, W. Rabadi, and R. Talluri, "Programmable DSP platform for digital still cameras," *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Mar. 1999, pp. 2235-2238.

[9] A. Gentile, S. Vitabile, L. Verdoscia, and F. Sorbello, "Image processing chain for digital still cameras based on the SIMPil architecture," *Proc. Int. Conf. Parallel Processing Workshops*, June 2005, pp. 215-222.

[10] DCS Videobench, Online: http://www.videobench.rd. tut.fi/

[11] Wikipedia, DDR2 SDRAM, Online: http://wikipedia.org/wiki/DDR2_SDRAM.

[12] Micron, 512Mb: x16, x32 Mobile DDR SDRAM, Data Sheet, Mar. 2008.

[13] Micron, Calculating DDR Memory System Power, Technical Note, May 2005.

[14] Micron, Low-Power Versus Standard DDR SDRAM, Technical Note, Jan. 2007.

[15] International Technology Roadmap for Semiconductors, 2007 Edition, System Drivers, Report, 2007.

[16] T. Wolf and M. A. Franklin, "Performance models for network processor design," *IEEE Trans. Parallel Distrib. Syst.*, vol. 17, no. 6, June 2006, pp. 548-561.

[17] R. Kisiel and Z. Szczepanski, "Trends in assembling of advanced IC packages," *Journal Telecomm. and Inform. Technol.*, no. 1, 2005, pp. 63-69.

[18] W.-Y. Yip, S. Best, W. Beyene, and R. Schmitt, "System Co-Design and Co-Analysis Approach to Implementing the XDR Memory System of the Cell Broadband Engine Processor: Realizing 3.2 Gbps Data Rate per Memory Lane in Low Cost, High Volume Production," *Proc. Asia South Pacific Design Automation Conf.*, Jan. 2007, pp. 858-865.