Low-Power Motion Estimation Processor with 3D Stacked Memory

Shuping ZHANG‡†, Jinjia ZHOU†, Nonmembers, Dajiang ZHOU†, Member, Shinji KIMURA†, Senior Member, and Satoshi GOTO†, Fellow

SUMMARY  Motion estimation (ME) is a key encoding component of almost all modern video coding standards. ME contributes significantly to video coding efficiency, but it also consumes the most power of any component in a video encoder. In this paper, an ME processor with 3D stacked memory architecture is proposed to reduce memory and core power consumption. First, a memory die is designed and stacked with ME die. By adding face-to-face (F2F) pads and through-silicon-via (TSV) definitions, 2D electronic design automation (EDA) tools can be extended to support the proposed 3D stacking architecture. Moreover, a special memory controller is applied to control data transmission and timing between the memory die and the ME processor die. Finally, a 3D physical design is completed for the entire system. This design includes TSV/F2F placement, floor plan optimization, and power network generation. Compared to 2D technology, the number of input/output (IO) pins is reduced by 77%. After optimizing the floor plan of the processor die and memory die, the routing wire lengths are reduced by 13.4% and 50%, respectively. The stacking static random access memory contributes the most power reduction in this work. The simulation results show that the design can support real-time 720p @ 60 fps encoding at 8 MHz using less than 65 mW in power, which is much better compared to the state-of-the-art ME processor.

key words: 3DIC design, motion estimation processor, low power design, memory stacking

1. Introduction

With the development of semiconductor technology, portable devices became more powerful. Meanwhile, the performance of cameras integrated in portable devices improved as well. The most obvious feature is that capture resolution has increased from 0.3 megapixel (MP) to 1.3, 3, 8, and even 10 MP. With such a powerful camera, 720p and 1080p video recording and playback have been common capabilities of newer portable devices. Furthermore, smartphones equipped with 20-MP cameras [1], [2], 41-MP cameras [3], and 4K video recording capability [2] have been released. Benefiting from the portability of smartphones and the vibrant picture quality of video, multiple users are recording their lives on video. At the same time, power consumption has been a bottleneck: many portable devices need to be charged once or even twice per day, which would be inconvenience for users. Users prefer long battery life in their portable devices. The popularity of video capture and playback by portable devices is increasing, and therefore, a low-power video codec is required.

Much of the applicable research has focused on the reduction of power consumption in the video codec itself; such research has shown good results. Sze et al. [4] implemented a full real-time 720p H.264 decoder in 65 nm by using a variety of techniques such as multiple voltage, frequency domains, frame level dynamic voltage, and frequency scaling. In this study, the core power consumption of the video decoder was reduced to 1.8 mW. Lin et al. [5] implemented a 1080p @ 30 fps H.264/AVC encoder in 130 nm by applying several techniques, including complexity reduction and cross-stage hardware sharing to optimize the encoder core power at 242 mW.

However, although the video encoder/decoder core power can be significantly reduced, the total power consumption is still high when the total system uses an external dynamic random access memory (DRAM). Many works have been focused on minimizing DRAM power by decreasing the DRAM bandwidth requirement [6], [7]. Unfortunately, conventional 2D integrated circuit process technology has encountered a bottleneck in reducing power consumption. Now, many researchers are attempting to solve the DRAM problem by using 3D large-scale integration (LSI). One example of modifying the regularity of the DRAM architecture involved the development of an industrial high-performance 8-Gb 3D double data rate type 3 (DDR3) memory [8]. In addition, Samsung has applied 3D-TSV technology to its 30-nm-class DRAM products to keep pace with Moore’s Law and industry projections. Many researchers are studying wide IO memory [9] and the hybrid memory cube (HMC) [10] to improve performance. This research is not limited to the memory area; 3D LSI design is also being studied. A 64-core processor [11] with stacked memory was designed; its maximum throughput is 63.8 GB/s. Zhao et al. [12] introduced a five-tier stacked H.264 application using on-chip DRAM stacking. Although memory power is not given in [12], we can figure it out by [13] using the characteristics given in [12]. Thus, the memory power is 492.5 mW, which is still considerably high.

In this paper, a motion estimation (ME) processor with 3D stacked memory architecture is proposed to reduce the memory power. ME is a key encoding component of almost all modern video coding standards. In a video encoder, most of the power is consumed by the ME because it requires a large memory bandwidth. As profiled in [17], ME uses more than 50% of the total computation time in an H.264/AVC encoder when the encoder is configured to use single-direction...
full search with a search range (SR) of 32. The ME used in this design has a SR of 211 and two reference frames, which will consume more than 50% of the total computation time in [17]. In addition, the ME dominates 76.62% logic area of the H.264 encoder in [18], which implies that the ME occupies more than 70% of power consumption in H.264 encoder.

The rest of this paper is organized as follows: Sect. 2 presents the architecture of this design. Section 3 gives an overview of the entire design flow of this work. Section 4 describes the details of the physical design. The simulation results are illustrated in Sect. 5. Finally, Sect. 6 presents our conclusions.

2. Architecture Design

Memory bandwidth and memory access power are now becoming problems in [19]. Therefore, this work tries to find a solution of the bandwidth problem. F2F stacked 3D architecture for video encoder component (ME) is first proposed in this work. Many 3D architectures have been published, but different applications result into different technical difficulties. To demonstrate the feasibility of 3DLSI architecture for video encoder, SRAM stacked 720p ME architecture is presented in this section.

2.1 2-Die Stacked 3D Architecture

Figure 1 shows the side view of the stacked dies. Two dies (ME processor and memory) of the same size are stacked face-to-face. The ME processor die is positioned on the top because of the following considerations: First all IO cells are in ME processor die, and are connected to the landing pads on the backside of the ME processor die (the upper surface of the 3D chip) by through-silicon-via (TSV) technology. The landing pads are connected to the lead by wire bonding. Second, the ME processor consumes more power than the memory, so that the ME processor die generates more heat than the memory die. Thus, better cooling can be provided for the ME processor when it is positioned on the top die.

Based on this design, IO pins are not needed for the memory die because all data transmission and power delivery to the memory die are accomplished through the face-to-face (F2F) bonding pads. Therefore, without the limitation of IO pins, a memory with 128/256-bit-width IO can be applied to this design.

Another advantage of this 2-die stacked 3D architecture is that we can fully utilize 2D electronic design automation (EDA) tools. Once we decide the interface between the two dies (i.e., the locations of the F2F pads), the usual 2D design flow can be applied to the ME processor die and memory die separately.

2.2 ME Processor Architecture

The top-level block diagram of the ME processor is shown in Fig. 2. Based on our previous work [19], the ME processor contains a hierarchical integer motion estimation (IME) component and a fractional motion estimation (FME) component. The IME component is separated into an IME coarse search engine (IMEC) and an IME refinement search engine (IMER). IMEC, IMER, and FME work in parallel in a macro block (MB)-level pipeline, and a memory scheduler component issues data requests to static random access memory (SRAM) interfaces and schedules MB-level tasks.

Two independent memory interfaces are employed to provide connectivity to the memory die. SRAM (A) is a 256-bit interface for buffering the reference frame, while SRAM (B) is a 128-bit interface for buffering the source frame and motion vectors (MVs). Data is stored to SRAM (A) after reference frame compression [7].

Two caches implemented with on-chip SRAMs are employed for serving reference frame data to the IMEC and IMER. The 24-KB IMEC cache consists of 16 data memory banks with independent read addresses. Each bank is implemented with a single-port read, single-port write (1R1W) SRAM of 256 × 48 bits. The 512-KB IMER cache is also composed of 16 data memory banks. Each bank is implemented with a 1R1W SRAM with 2048 × 128 bits.

The ME processor can provide a max throughput of 1.59 Gpixels/s @ 210 MHz. Therefore, real-time 720p @ 60 fps video encoding can be supported at 8 MHz. Based on our previous ME engine [20], we removed the DRAM interface, modified the cache system, and added a special memory controller.

The main features of the ME processor including the cache size and search range, are same with [19], so that we can stack more SRAM memories or DRAM memories for...
2.3 Memory Architecture

The ME processor requires two memories for data access. Compared to 2D ME [19], external DRAMs are completely replaced by the SRAM memories (memory (A) and memory (B)) in this design. There is not any DRAM memory applied in the whole system. The bit widths of its two SRAM interfaces are 256 and 128, respectively, and therefore, two wide IO memories are customized for this ME processor. The memory is composed of many normal SRAM blocks. With the limitation in chip size and the large size of SRAM blocks, the capacity is limited. Memory (A) is designed to be 14.25 Mb because one 720p frame requires 7.032-Mb memories, and two reference frames are stored in memory (A). Memory (B) is designed to be 8 Mb because one source frame and some MVs are stored in Memory (B).

The synthesis results show that the memory die can run at 300 MHz. Therefore, the bandwidth can be as high as 14 GB/s. However, because of the limitation of the memory size, the whole system can support no more than 720p. 512-Kb sub-banks that make up a 128-bit-width memory bank. The total capacity of memory (A) is 8 Mb. These four banks share another data bus, and all sub-banks in memory (B) share another address bus. The data bus and address bus are 128-bit and 16-bit, respectively. Table 1 summarizes the specifications of these two memories.

The SRAM bank has three modes: read, write and stand-by. In both read mode and write mode, the internal power and leakage power are consumed, but in stand-by mode, only the leakage power is consumed.

Since the memory banks share the data and address bus, when data operations occur, all the memory banks will consume internal power and leakage power, even though there is only one bank with data access one time. To reduce the power consumption of the memory die, all memory banks can switch to stand-by mode automatically until there is data access on the corresponding bank. This approach reduces power consumption by 13.89%.

2.4 Memory Controller Architecture

The DDR controller is integrated in the original ME processor [19]. Since external DRAMs are completely replaced by the SRAM memories. In this work, a special SRAM memory controller is designed to replace the DDR controller. There are two main functions of the memory controller. First, it can control data transmission between the ME processor and the two memories. Second, it is capable of controlling the timing of the ME processor and the two memories.

The special memory controller includes two independent controllers (controllers (A) and (B)), as shown in Fig. 4. The controller responds to data access requests from the memory scheduler in the ME processor. Controller (A) undertakes the data transmission between memory (A) and the ME processor, while memory B and the ME processor are
connected by controller B. Both memories are compatible with the burst mode, whose burst length is 8.

There are some other benefits to this design. First, all interfaces between the ME processor die and memory dies are included in this module. To change the interface, we only need to modify this module. Second, the ME processor can be easily integrated into this design.

3. 3D Design Flow

The entire design flow is presented in this section. To completely utilize the 2D EDA tools, we propose a die-level 3D LSI architecture. The basic flow is similar to the conventional 2D design flow, but several special steps are added to complete the 3D design. Here, we classify our design flow into two parts: front-end design flow and back-end design flow.

3.1 Front-end Design Flow

The front-end design focuses on the functional and logical parts of the overall design. The goal of the front-end design is to create a gate-level design with the correct logic function.

The DDR memory controller is replaced by a novel SRAM controller in this work. After completing this modification, the timing should be verified so that it is the same as that of the original ME processor.

Figure 5 shows the partition and synthesis flow. Two SRAM memories with negligible logic are grouped into the memory module for the memory die, while the novel memory controller and ME processor are partitioned in the processor module for the processor die. A top module is defined to contain the memory module and processor module.

After partitioning, 65-nm process technology is employed to synthesize the register transfer level codes. To check whether this design meets the timing constraints, the codes are synthesized twice. The processor module and memory module are synthesized first. The top module (containing the memory module and processor module) is synthesized next. Both of the synthesis results show that the timing constraint is satisfied.

The max operating frequency of this design is 300 MHz. Since the operating frequency is set at 8 MHz, there are plenty of margins of timing to the circuits. The Synopsys design compiler is applied to generate the gate-level net list in this work.

3.2 Back-end Design Flow

Figure 6 shows the back-end design flow. The net list file and timing constraint file, which are generated in the front-end design, are imported to the EDA tool. The Cadence Encounter is used for the back-end design. The ME processor die and the memory die run the back-end design. Thus, we need to separate the generated net list into two net list files before we begin the back-end design. In addition, we need to prepare many files, including a 65-nm technology file, timing library, and physical library before beginning the back-end design.

In the back-end design flow, we optimize the floor plan so that the EDA tools can obtain good results in subsequent steps such as standard cell placement, clock tree synthesis (CTS), and auto wire routing. After we complete the floor plan, we need to design the power network, which is presented in Sect. 4.5. Third, we place TSV and F2F pads manually. TSV is used in the IO area for connecting the IO cell to the landing pad. Consequently, TSV placement is accomplished only in the ME processor die. The rest of the steps can be performed automatically by the EDA tools. We have written an autorun script to improve design efficiency. For autorouting, we have created some special settings to reserve the top metal layer for F2F bonding.

After the script is finished, we obtain the routing result. The routing result will be analyzed to see whether there were any violations. We perform several iterations with different configurations, and finally obtain a result without violations.

4. 3D Physical Design

The physical design is also introduced in [12], but only 3D bonding and power delivery are mentioned. In this section, we will show how to implement 3D physical design in details using 2D EDA tools. We first define F2F pads and
TSVs by modifying the library exchange format (LEF) file and the timing library. Then, the floor plan of the memory die and the ME processor die, the placement of the F2F pads and TSVs, and the power network are presented. Benefiting from the 3D architecture, the footprint of this design is 5 mm × 5 mm.

4.1 F2F Pad and TSV Definitions

The F2F pad and TSV shown in Fig. 7 are the necessary components in 3D physical design. However, 2D EDA tools do not support them. Therefore, we define them as cells for processing by EDA tools easily.

The side view of the defined TSV is shown in Fig. 7(a). According to Tezzaron’s 3D packaging technology, the diameter of the TSV is 2 µm, while the diameter of the landing pad on the first metal layer is 5 µm. A large landing pad allows for a larger misalignment of TSV, so that it can improve the yield.

Figure 7(b) illustrates the shape of the customized F2F pad. The size of the F2F pad is 3.4 µm × 3.4 µm. Since the electrical characteristics of the F2F bond pads are very similar to a typical via, the capacitance and the resistance of the F2F bond pad is negligibly small. The F2F pad has two pins. The pin on the top metal layer connects to the other die, while the other pin on the low metal layer connects to the signal, power, or ground (P/G).

4.2 Floor Plan of Memory Die

Both memory A and memory B are integrated in the memory die. Memory B includes 16 memory blocks, while memory A includes 30 memory blocks, i.e., there are 46 available memory blocks in total. In addition, a backup block is set in the memory die. These 47 blocks are placed in a 4384-µm × 4640-µm core area, as shown in Fig. 8. To minimize the chip size, the blocks are placed as close to each other as possible.

The floor plan is a key step in the back-end design flow. It will significantly and directly affect auto placement and routing results. Several ideas are proposed to optimize the floor plan of the memory die.

First, to minimize the wire length, the blocks, which share the same address bus and data bus, are lumped together. As described in Sect. 2.3, since all banks from the same SRAM (SRAM (A) or (B)) share the address bus, placing these banks together can considerably reduce the wire length. As shown in Fig. 7, the blocks covered by the white box are placed together, since all blocks belong to SRAM (A) and share the address bus.

Second, the triangle in the corner of the block indicates the orientation of the block and the locations of the pins of the block. The location and the order of the pins in a block are fixed. However, the orientation of the block can be set by the user. By positioning the pins of each block as close as possible and in the same order, routing congestion can be reduced.

Finally, within the SRAM (SRAM (A) or (B)), the blocks that share the same part of the data bus are placed together to reduce the length of the routing wire. As described in Sect. 2.3, the four blocks covered by the red box in Fig. 8 are the sub-bank 0 of each bank in SRAM (B), and they share the same low 32-bit data bus, so they neighbor each other.

Consequently, after optimizing the floor plan, the total routing wire length is reduced by more than 50%.

4.3 Placement of F2F Pads and TSVs

In this design, TSV is used in the IO area of the ME processor die. Figure 9 shows the TSVs inserted in the location of the IO pads. To improve the yield, redundant TSVs (32 per IO pad) are placed to connect one IO pad. The pitch of the redundant TSV is 10 µm. There are 10 input cells, 14 output cells, and 32 P/G cells in the processor die, i.e., 56 IO pads. Therefore, the total number of TSVs is 1792.
The locations of the F2F pads are determined by the floor plan of the memory die as shown in Fig. 10(a). To reduce the routing wire length, F2F signal pads are placed in the white boxes, where the pins of the memory blocks are gathered. To reduce routing congestion, the order of the F2F signal pads is the same as the pins in the blocks. Figure 10(b) shows the F2F signal pads. The pitch is 5 \( \mu \)m. There are 803 F2F signal pads in total.

4.4 Floor Plan of Processor Die

Figure 11 shows the floor plan of the ME processor die. Since the locations of the F2F pads are decided by the memory die, the floor plan of the ME processor is optimized based on the fixed F2F pads.

As described in Sect. 2.2, the IMEC cache is composed of 16 memory banks, and the IMER cache also consists of 16 memory banks. Each bank is composed of two RAM blocks. In order to reduce the routing wire length, these two blocks are put together. In addition, 2.2 M logic gates and 72 cache blocks are placed in a 3340-\( \mu \)m \( \times \) 3400-\( \mu \)m core area. The blocks of the IMER cache are placed at the bottom, while the blocks of the IMEC cache are placed at the top. The blocks are placed from the periphery to the interior with proper orientation. Compared to the floor plan automatically generated by Encounter, the routing wire length is reduced by 13.4%.

The IO cells are placed primarily at the left and right sides of the IO area because the top and bottom sides are occupied by F2F P/G pads. There are only 24 control signal cells and 32 P/G cells in the processor die because most of the pins connected to the memory die are replaced by F2F signal pads. Compared to [19], the number of IO pins is reduced by 77%.

4.5 Power Network

A strong power delivery network ensures reliable operation of circuits on a chip, especially in a 3D IC. Thus, we build a strong power network to ensure power delivery. Figure 12 shows the power network of the processor die. The core area is surrounded by a wide P/G core ring that is bridged inside and outside. Inside the core area, power rails that connect to the power ring horizontally supply power to the standard cells. Furthermore, the power stripes, which vertically connect to the power ring and power rails, are set in small interval to reduce the IR-drop. To enhance the power supply to the RAM blocks, block rings are also added. These are not shown in the figure.

The F2F P/G pads deliver power from the ME processor die to the memory die. The pads are directly connected to the power ring. The power network of the memory die is similar to that of the processor die, except that there is no P/G cell, and the F2F P/G pads are located on the power ring. Therefore, the power network of the memory die is not introduced again.

5. Simulation Results and Comparison

The proposed architecture is synthesized using the Synopsys design compiler using a 65-nm standard cell library. Then, Cadence Encounter is used for the back-end design including the floor plan, power network generation, TSV and F2F placement, CTS, and auto placement and routing. Figure 13 shows the layout of the ME processor die and the memory die. The physical characteristics of the entire design are summarized in Table 2. The simulations are based on these layouts.
5.1 3D Power Analysis

3D power analysis is performed by the Cadence Encounter Power System (EPS). The power analysis flow is similar to the 2D flow. This is because the method for analyzing power involves calculating the power of each cell at a given toggle rate. The activity ratio of the memory is 60% on average while the read/write ratio is 99/1.

The physical library, timing library, technology file, and also the net list file, design exchange format (DEF) file, and standard delay format file are imported into EPS. Then, the analysis method is set to static, and the corner mode is set to normal (1.2 V at 25°C). The frequency is read from the SDC timing constraint file. We obtain the toggle rate from ModelSim and set it as the parameter of input activity. Finally, EPS reports that the total power is 64.85 mW. The power of the processor die and memory die are 37.67 mW and 27.18 mW, respectively.

5.2 3D IR-Drop Simulation

To ensure the reliable operation of circuits in this work, the voltage inside the chip should be larger than 1 V. To check whether the voltage is satisfied with this requirement all over the chip, we do the IR-drop simulation by EPS. The voltage of 1.05 V is selected as simulation condition, since this is the worst case in this library.

A 2D IR-drop simulation is performed first. We created a cell list to create a power grid library. Then, we set the analysis method to static mode, the temperature to 25°C, and set the locations of source power. Here, the source power of the processor die is the power cell, and the source power of the memory die is the F2F power pad. Then, we import the result files of the power analysis. After finishing these steps, EPS analyzes the IR-drop automatically.

The average and worst IR-drop of the processor die are 2.007 µV and 5.478 mV, respectively. The average and maximum IR-drop of the memory die are 0.642 µV and 1.555 mV, respectively. Note that the voltage of the power net is 1.05 V, and therefore, there is a sufficient margin for the 3D stacked case.

A true 3D IR-drop simulation is also carried out to demonstrate that the power supply network can accommodate the requirements of the entire system. The simulation method is shown in Fig. 14.

First, we duplicate the technology file by renaming the metal layers and vias. For example, MET1 is renamed as MET1M, and VIA12 is renamed as VIA12M. Then the duplicated technology file and the original technology file are combined, and the metal layers are reorganized to generate a new technology file with double metal layers. At the same time, the timing libraries and cell libraries (LEF files) are duplicated by renaming the cells and the libraries in the same way.

After that, we export the DEF file of the processor die and the memory die separately. For the memory die, we modified the DEF file by renaming all of the cells and layers inside the file, as mentioned above. The duplicated technology file, timing libraries, cell libraries, and the modified DEF file of the memory die are imported into EPS so that it can perform the power analysis again. The new power result is the same as the previous one.

Third, the DEF file of the processor die and the modified DEF file of the memory die are combined to generate a new 3D DEF file containing all details of the entire design. Then the 3D DEF file, all net list files, the new technology file, and the original and duplicated timing and cell libraries are imported to EPS. Finally, we set the power cell as the source power, and import the power result of the processor die and the new power result of the memory die. After finishing these steps, EPS analyzes the IR-drop automatically.

The simulation result shows that under the condition of 1.05 V and 25°C, the average IR-drop is 1.289 µV, while the worst IR-drop of the entire design is 5.742 mV. It proves that the IR-drop is under our threshold, which is 50 mV.

5.3 3D Thermal Analysis

3D thermal analysis is done by HotSpot [22]. HotSpot provides a common thermal model for analyzing, and is capable of modeling stacked 3-D chips. It also provides reasonable heat spreader and heat sink model for a typical, high-performance microprocessor. This heat spreader and heat sink model is used in thermal analysis. The thermal interface material (TIM) is assumed to thermal grease, whose thermal conductivity is set to four.

First, the core area of each die was divided into many small areas basing on the floor plan. Second, we counted...
the power of each standard cell, block, and assigned it to different area by the location of the standard cell and block. Then we got the power distribution of each die. The floor plan and the power distribution information is imported to Hotspot as initial data. Then analysis mode is set to grid, the ambient temperature is set to 25°C, and the frequency is set to 8 MHz.

The analysis results are shown in Fig. 15. The thermal problem can be negligible, since this is a low power design. We can see the maximum temperature of this design in 25.05°C. It is hardly to distinguish from the ambient temperature. It proves that this design does not suffer the thermal problem.

5.4 Power Comparison

This work focuses on the 3DLSI solution for ME processor. Thus we compare the proposed architecture with 2D ME and other 3D ME. We do not compare this design with [4], [5], [12] or [21] for their different functions. Reference [4] is a video decoder that does not have an ME processor. Reference [5] is a 2D full encoder that includes not only ME processor but also other parts. Reference [12] introduced an H.264 full encoder with 3D DRAM stacking while [21] introduced a 14-core, 25 power domains application processor with DRAM stacking for multimedia mobile applications. An H.264 codec is also integrated in this chip. Neither of two 3D papers gives the logic gate count of the ME portion, or indicates the area of the ME portion in the chip layout, Thus we cannot figure out the power consumption of the ME portion. Though ME occupies the most power consumption in H.264 encoder, but not all. It is not fair to compare with ME only processor, especially the low power 3D ME.

The power will be normalized by Eq. (1), since other works used different process technologies. Here, $V$ is the operating voltage, and $C$ is the capacitance. The capacitance ratio can be calculated for the processor technology in a simple manner:

$$\text{Power}_A = \text{Power}_B \times \left( \frac{C_A}{C_B} \right) \times \left( \frac{V_A}{V_B} \right)^2$$

(1)

Table 3 shows the specifications of this design and a power comparison to other works.

An ME processor with system-in-silicon architecture is introduced in [22]. This ME needs 760 clock cycles to process one MB. Its core power and memory power are 2383 mW and 190 mW, respectively. The total power is 2573 mW. After normalizing by Eq. (1), the normalized core power and normalized memory power are 382.5 mW and 50 mW, respectively. The total energy efficiency is 6.952 nJ/pixel. Benefiting from the optimization of the algorithm for the ME processor, the core power of this design is much lower than [22].

The 2D implementation of the ME processor is introduced in [19]. A DDR3 memory controller is integrated in [19], which means that the DDR3 memory is used for
Table 3 Design specification and power comparison.

<table>
<thead>
<tr>
<th>Memory type</th>
<th>Technology</th>
<th>Frequency</th>
<th>Throughput</th>
<th>Core power</th>
<th>Energy efficiency</th>
<th>Memory power</th>
<th>Energy efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>System-in-Silicon DRAM / on chip</td>
<td>180 nm/110 nm</td>
<td>200 MHz/25 MHz @ 1.8 V</td>
<td>1080p @ 30 fps</td>
<td>2.383 mW</td>
<td>6.149 n/pixel</td>
<td>190 mW</td>
<td>1.895 n/pixel</td>
</tr>
<tr>
<td>External DDR3</td>
<td>40 nm</td>
<td>210 MHz @ 1.1 V</td>
<td>420p @ 48 fps</td>
<td>622 mW</td>
<td>0.742 n/pixel</td>
<td>1836 mW</td>
<td>27.18 mW</td>
</tr>
<tr>
<td>3D stacked memory / on chip</td>
<td>65 nm</td>
<td>8 MHz @ 1.2 V</td>
<td>720p @ 60 fps</td>
<td>382.5 mW</td>
<td>0.681 n/pixel</td>
<td>2183 mW</td>
<td>27.18 mW</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Memory type</th>
<th>Technology</th>
<th>Frequency</th>
<th>Throughput</th>
<th>Core power</th>
<th>Energy efficiency</th>
<th>Memory power</th>
<th>Energy efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>System-in-Silicon DRAM / on chip</td>
<td>180 nm/110 nm</td>
<td>200 MHz/25 MHz @ 1.8 V</td>
<td>1080p @ 30 fps</td>
<td>2.383 mW</td>
<td>6.149 n/pixel</td>
<td>190 mW</td>
<td>1.895 n/pixel</td>
</tr>
<tr>
<td>External DDR3</td>
<td>40 nm</td>
<td>210 MHz @ 1.1 V</td>
<td>420p @ 48 fps</td>
<td>622 mW</td>
<td>0.742 n/pixel</td>
<td>1836 mW</td>
<td>27.18 mW</td>
</tr>
<tr>
<td>3D stacked memory / on chip</td>
<td>65 nm</td>
<td>8 MHz @ 1.2 V</td>
<td>720p @ 60 fps</td>
<td>382.5 mW</td>
<td>0.681 n/pixel</td>
<td>2183 mW</td>
<td>27.18 mW</td>
</tr>
</tbody>
</table>

1: Power normalized to 65 nm (P65 = P180 / 6.25 = P110 / 3.8 = P40 x 1.9)
2: power is modeled by [13]
3: Energy efficiency = Norm. power / throughput

6. Conclusion

In this paper, an ME processor with 3D stacked memory architecture is proposed to reduce memory power and provide higher bandwidth. ME is a key encoding component of almost all modern video coding standards. ME is also the component that consumes the most power in a video encoder. By adding F2F pads and TSV definitions, 2D EDA tools are extended to support the proposed 3D stacking architecture. Furthermore, a special memory controller is designed to control the data transmission and the timing between the memory die and the ME processor die. Finally, a 3D physical design is completed for the entire system; this design includes a floor plan optimization of the two dies, TSV/F2F placement, and power network generation. Comparing the 3D design to the 2D-technology-based ME, the number of IO pins is reduced by 77%. After optimizing the floor plan of the processor die and memory die, the routing wire lengths are reduced by 13.4% and 50%, respectively. The stacking SRAM memory contributes more than 90% of power reduction comparing to 2D design. The simulation results show that the power consumption of the entire design is 64.85 mW, which is much better than that of a state-of-the-art ME processor. This work demonstrates the feasibility of the F2F stacked 3D architecture for video encoder.

Problems about high-memory-bandwidth requirement and high-power consumption can be addressed by the proposed scheme for realizing 8K × 4K low-power encoder. Our future work is to implement 8K × 4K HEVC encoder with 3DLSI technology.

Acknowledgments

This research was supported by the regional innovation strategy support program of MEXT and the Waseda University Graduate Program for Embodiment Informatics (FY2013–FY2019).

References


data storage. The throughput of the 2D ME is 4320p @ 48 fps, while the operating frequency is 210 MHz @ 1.1 V. The core power is 622 mW; after normalizing by Eq. (1), the normalized core power is 1181.8 mW. The memory power is 1836 mW, which is modeled by [13]. Compared to the 2D design, the capacity of memory in this design is reduced by a significant amount. Benefiting from the 3D LSI technology, the total energy efficiency is increased by 60%.

After optimizing the algorithm, which is introduced in [20] including rhombus window full search, reference pixel fetching mechanism, directional 5T12S (5 transform points among 12 search points) search scheme, the ME processor can process one MB in 28 clock cycles, so that it can encode 720p @ 60 fps video sequences with two reference frames at 8 MHz. Under this operating frequency, the power consumption of the ME processor die and the memory die are as shown above. As mentioned in Sect. 4.1, the capacitance and the resistance of the F2F bond pad is negligibly small. Thus, large power-consuming drivers are not required in this design. The power of F2F pad is negligible and it is included in the total power. We can see the energy efficiency of ME processor between 2D and 3D is almost the same. In 2D design, off-chip DRAM memory is applied because of the limit of chip size. In 3D design, however, customized SRAM memory is stacked with the ME processor. The energy efficiency of memory in this design is improved by more than 50%. The memory die contributes the most power reduction in this work. Because of the small memory size, frequency reduction, and 3D integration, the energy efficiency of this design is 1.173 nJ/pixel, which is only 16.87% of [22] and 61.9% of [19].
Shinji Kimura received the B.E., M.E. and Dr. of Eng. degrees in information science from Kyoto University, Kyoto, Japan in 1982, 1984, and 1989, respectively. He was an Assistant Professor at Kobe University from 1985, was an Associate Professor at Nara Institute of Science and Technology from 1993. He has been a Professor of Waseda University from 2002. He is interested in the formal hardware verification, low power synthesis and reconfigurable hardware.

Satoshi Goto received the B.E. and the M.E. Degrees in Electronics and Communication Engineering from Waseda University in 1968 and 1970 respectively. He also received the Dr. of Engineering from the same University in 1978. He joined NEC Laboratories in 1970 where he worked for LSI design, Multimedia system and Software as GM and Vice President. Since 2002, he has been Professor, at Graduate school of Information, Production and Systems of Waseda University at Kitakyushu. He served as GC of ICCAD, ASPDAC, VLSI-SOC, ASICON and ISOCCC and was a board member of IEEE CAS society. He is IEEE Life Fellow and IEICE Fellow. He is Visiting Professor at Shanghai Jiao Tang University and Tsinghua University of China and Member of Science Council of Japan.