An 8-Gb 12-Gb/s/pin GDDR5X DRAM for Cost-Effective High-Performance Applications

Martin Brox, Mani Balakrishnan, Martin Broschwitz, Cristian Chetreanu, Stefan Dietrich, Fabien Funfrock, Marcos Alvarez Gonzalez, Thomas Hein, Eugen Huber, Daniel Lauber, Milena Ivanov, Maksim Kuzmenka, Christian N. Mohr, Juan Ocon Garrido, Swetha Padaraju, Sven Piatkowski, Jan Pottgiesser, Peter Pfefferl, Manfred Plan, Jens Polney, Stephan Rau, Michael Richter, Ronny Schneider, Ralf Oliver Seitter, Wolfgang Spirkl, Marc Walter, Jörg Weller, *Member, IEEE*, and Filippo Vitale

***Abstract*— The graphic DRAM interface standard GDDR5X is developed as an evolutionary extension to the widely available GDDR5. The implementation presented here achieves a data rate of 12 Gb/s/pin on a single-ended signaling interface with 32 IOs for a total memory bandwidth of 48 GB/s. The GDDR5X DRAM relies on the quad data rate operation enabled by a phase-locked loop (PLL), a receiver with a pre-amplifier in a dual-regulation loop and a one-tap digital feedback equalizer (DFE). To support lower performance modes, an additional GDDR5-like operation is provided, which bypasses the PLL. The interface is realized on a conventional high-volume DRAM process to provide a cost-efficient, discrete package 8-Gb DRAM for high-performance graphic cards and compute applications.**

***Index Terms*— GDDR5, GDDR5X, graphic DRAM, high-speed memory, wireline transceiver.**

I. Introduction

GDDR5 has emerged as the leading DRAM interface for applications, requiring high system bandwidth such as graphic cards, game consoles, or high-performance computing applications. However, recently, it has become obvious that per-pin data rate of GDDR5 DRAMs is no longer increasing. Therefore, the industry has initiated the development of post-GDDR5 memories. In this paper, we will introduce a new DRAM interface which intends to be an evolutionary extension of GDDR5. It offers the possibility to increase per-pin bandwidth by a factor of 2 while reusing the important parts of the established GDDR5 ecosystem. The initial target for the device presented in this paper is a data rate of 12 Gb/s/pin. This paper is organized as follows. Section II motivates why we started this development. Section III introduces the basic ideas behind GDDR5X. GDDR5X is striving for high data rates on a DRAM. Therefore, in Sections IV–VI, we describe details of the implementation of the clocking system, receiver, and transmitter. In Section VII, we present a set of measurement data, and Section VIII provides a summary and an outlook on future developments.

Manuscript received April 27, 2017; revised July 2, 2017; accepted July 28, 2017. Date of publication August 30, 2017; date of current version December 26, 2017. This paper was approved by Guest Editor Takashi Kono. *(Corresponding author: Martin Brox.)*

The authors are with Micron Semiconductor (Deutschland) GmbH, 81739 Munich, Germany (e-mail: mbrox@micron.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSSC.2017.2737945

II. Motivation and Silicon Implementation

GDDR5X is targeting the market for graphic DRAMs. In this field, the leading standard is GDDR5. GDDR5 has been around for nearly ten years and has been very successful. First, devices were introduced at a data rate of 6 Gb/s/pin [1] while cards readily available on the market have reached 8 Gb/s/pin. However, even the fastest GDDR5 [2] is only running at 9 Gb/s/pin which barely exceeds the speed of the parts already commercially available. Any further substantial increase is limited by the command clock frequency. Reaching 12 Gb/s/pin, which is the target of this GDDR5X device, requires a clock period of 333 ps for the command decoder. This is difficult in a DRAM process where MOS transistors are lagging in performance compared to a logic process by some years. We conclude that GDDR5 has limited headroom for further scaling.

One possible path to extend the memory bandwidth beyond this limit is available under the term of the high bandwidth memory (HBM) [3]–[5]. In an HBM device, the number of pins per device is vastly increased which as a consequence allows to reduce the per-pin data rate significantly. The most recent implementation of this standard promises to reach a data rate of 2 Gb/s/pin while providing in total 1024 IOs. Using this information, we can compare two graphics sub-systems each presenting 8 GB of memory.

A system built up from GDDR5 requires eight pieces of 8-Gb DRAM each running at the data rate of 8 Gb/s/pin. In total, this system will provide 256 GB/s of bandwidth. Built up from HBM2, a typical system would be constructed from two cubes of four high-stacked HBM2. At the data rate of 2 Gb/s/pin, this system will provide 512 GB/s. Hence, HBM2 is able to offer the increased performance as requested by the industry. However, this performance comes at a high cost which becomes apparent from the sketch of a typical HBM system (Fig. 1). An HBM system is built on a Si-interposer carrying the logic die (SoC) and the logic buffer die of the DRAM stack (I/O). The DRAMs themselves are stacked using Through Silicon Via technology (TSV) on top of the logic die. This complex system raises concern of testability, stability, handling, and not the least of cost.

We regard this as the chance to introduce another graphics memory beyond GDDR5 which being an extension of GDDR5 is named GDDR5X. The goals are stated

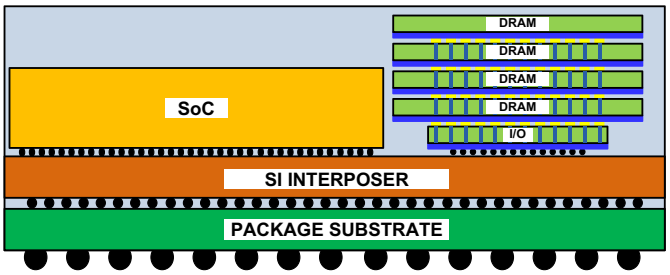

Fig. 1. Illustration of a typical HBM2-system. The SoC and the HBM buffer die (I/O) are placed on a Si-interposer with the DRAM-dies stacked on the top of the buffer die and connected by TSV.

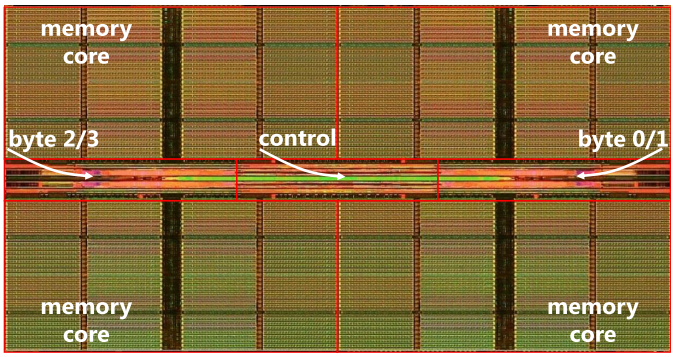

Fig. 2. Die photograph of the 8-Gb-GDDR5X DRAM implementation. The organization of the die follows an ODIC style. An 8-Gb-GDDR5 in the same process node looks virtually identical.

quite simply: provide headroom to double GDDR5's per-pin data rate while evolutionary extending this successful approach. GDDR5X should maintain the general architecture, IO impedance and termination, and the general command protocol. Most importantly, it needs to be a discrete component to enable established packaging, handling, and testing. Fig. 2 shows the resulting die photograph of the 8-Gb GDDR5X SGRAM implementation. The process is an established, unmodified 20-nm DRAM process with single copper and dual-aluminum metallization. In the same process node, the die looks very similar to a more conventional GDDR5. The peripheral circuitry is placed inside the central stripe where all bond pads are located. The command/address control interface is located in the center of this region, while two bytes each of the DQ interfaces are placed to the left and right in an outer-DQ/inner-control (ODIC) style arrangement. Memory cores are placed above and below this central axis.

III. Basics of GDDR5X

This section describes the basic ideas behind GDDR5X. Fig. 3(a) sketches the behavior of the relevant signals at the peak data rate [quad data rate (QDR) mode]. Command (CMD) and addresses (ADD) are provided in GDDR5 style. To overcome the core- and command-clock limitation following the established scheme from the double data rate (DDR) evolution, the prefetch is doubled relative to GDDR5. This allows to maintain the established core access timing (tCCD) and clock frequency of GDDR5 while doubling the data rate. As a result, the memory core of the GDDR5X device can be virtually identical to a GDDR5 core using

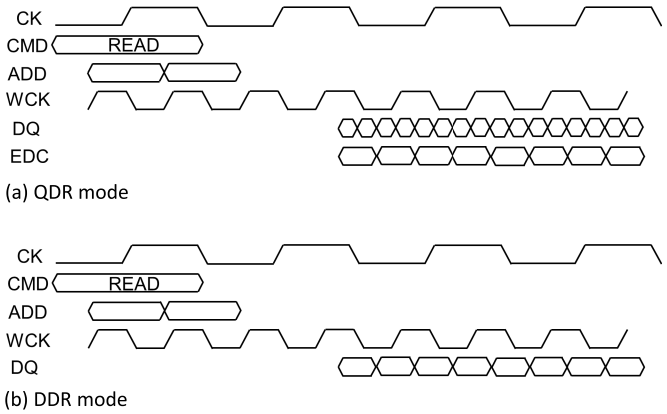

Fig. 3. Timing diagram for the relevant signals of GDDR5X in (a) QDR and (b) DDR mode. Command (CMD), addresses (ADD), data-clock (WCK), and error-detect (EDC) are operated in the same way as GDDR5. In the full-speed GDDR5X-mode, data (DQ) are driven QDR, while at lower speed DQ are driven DDR.

the same bank structure and row/column design (see Fig. 2). Just the number of IO-lines and IO amplifiers needs to be doubled to support the doubling of the prefetch. Repeating the previous calculation, a tCK of 666 ps is sufficient reach the 12-Gb/s/pin target. To further stay close to the established GDDR5 ecosystem, the data-clock (WCK) frequency of GDDR5X is maintained at twice the command clock (CK) frequency. The only change is the data rate of the data (DQ) bus which doubles compared to GDDR5 and thus requires QDR operation. On the other hand, the data rate of the error detect (EDC) pin does not change compared to GDDR5 such that EDC is operated in DDR mode. EDC is the return channel for the error detect information of the GDDR5 and GDDR5X protocol. In this way, a virtually error-free return channel for the error detect information is provided.

QDR operation relies on frequency doubling inside the DRAM since the WCK itself does only provide half the number of the required edges. In this device, this functionality is implemented by a phase-locked loop (PLL). However, a graphics device needs to support low-frequency operation in addition to the nominal, high-frequency mode. In a graphic sub-system, this low-power/low-frequency mode is required for any state which does not request full performance from the 3-D engine of a graphic card to enhance battery lifetime. For this purpose, GDDR5X implements a second mode which bypasses the PLL and is operating in DDR mode for which the PLL is not needed [Fig. 3(b)].

The signaling scheme (Fig. 4) closely follows GDDR5 as well, albeit with a reduced supply voltage VDDQ of 1.35 V compared to 1.5 V for GDDR5. GDDR5X implements a 60-Ω pull-up and a 40-Ω pull-down as a CMOS driver against a 60-Ω on-die high level termination on the receiving side. In this way, the data eye is centered at a VREFD of 0.7 × VDDQ with a swing of $+/-0.3 \times$ VDDQ.

IV. Clock System Architecture

Reaching highest data rates requires a careful design of the clocking system for the data path. Fig. 5 shows the

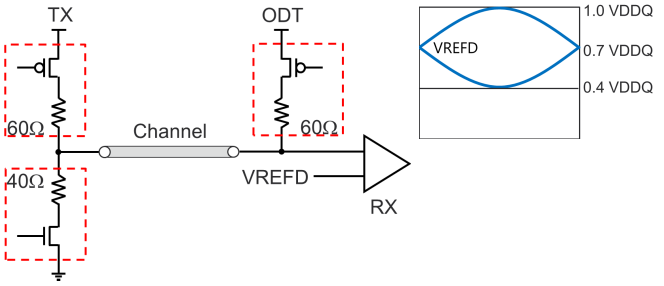

Fig. 4. Illustration of the signaling approach and levels of GDDR5X. The GDDR5-style of a calibrated 60-Ω pull-up/40-Ω pull-down driver and 60-Ω termination is reused. VDDQ is set to 1.35 V.

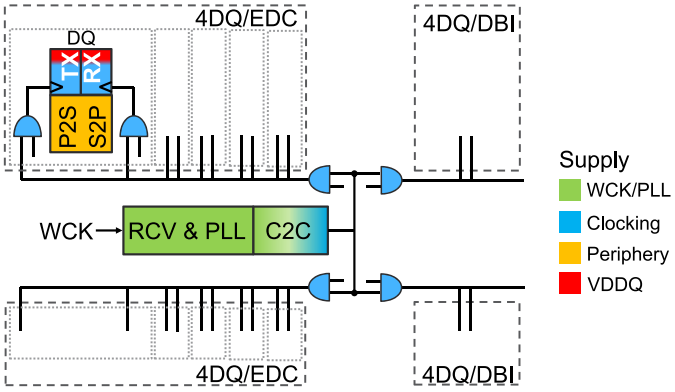

Fig. 5. Illustration of the clocking scheme and supply hierarchy for a shared double byte (e.g., byte 0/1).

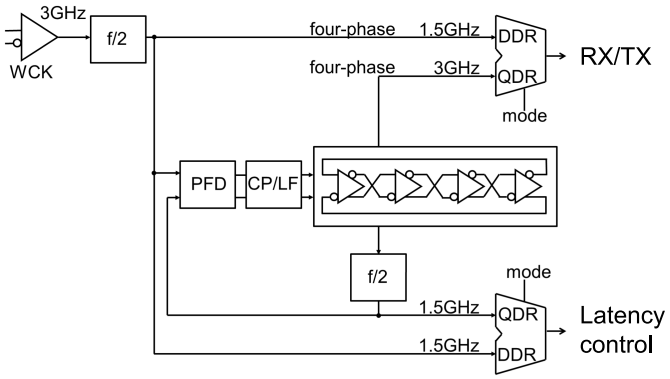

Fig. 6. Block diagram of central data-clock generation. QDR mode is supported by a PLL. The PLL is bypassed in DDR mode to avoid any relevant low-frequency limitation.

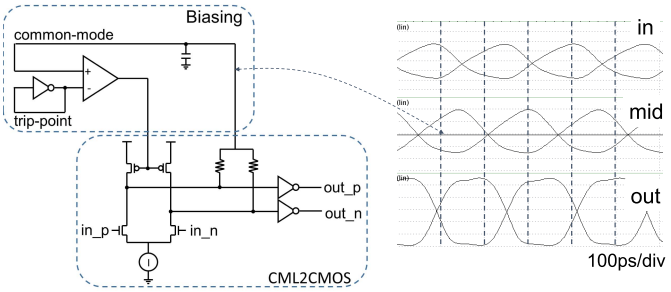

Fig. 7. Schematics of the central CML–CMOS conversion circuit with the associated biasing block. (Right) Simulation example.

implementation of the WCK clock system for two bytes. The high-speed clocks are distributed in two disjoint regions to the left and to the right of the center (Fig. 2). The clock is driven across two bytes with eight DQs, one EDC, and one DBI per byte. Clock receiver and PLL are designed in current mode logic (CML). The rest of the clock tree is designed in CMOS for best scaling of clocking power with frequency. The CMOS clock is distributed to the four corners each serving four DQ and one support pin—either EDC or DBI. The CMOS supply voltage is down-converted from the external VDD supply and final up-conversion to VDDQ is performed directly at the pad. To separate noise sources as much as possible, the chip implements three isolated on-die CMOS domains: a first one for the PLL, a second one for the CMOS clock distribution, and a third one for all remaining peripheral circuits. In this way, all externally visible timing is best isolated from voltage noise generated by core or data path activities.

The architecture of the central clock generation is sketched in Fig. 6. Following the clock receiver, the clock frequency is divided by 2 in a conventional CML divider to generate a four-phase clock at 1.5 GHz. These clocks can directly be used for receiving and transmitting data and for the read/write latency control in DDR mode [6]. In QDR mode, they serve as the reference clock for the phase-frequency detection (PFD) of the PLL. The clock from the voltage-controlled oscillator (VCO) is used in QDR mode for receiving and transmitting. To reach 12 Gb/s/pin requires a frequency of 3 GHz of the four-phase clock. Thus, the VCO clock is run through another divider into the PFD. Latency control in QDR mode requires as well a clock at the frequency of the command clock of 1.5 GHz [6]. Therefore, the divided clock is well suited for this purpose.

The initial frequency divider is needed in QDR mode to support the special training requirements of GDDR5X. Like in GDDR5, during the first mandatory training sequence, the CK command clock and the WCK data path clock need to be correctly aligned. This training step requires a data path clock running at the frequency of the command clock in phase with the final QDR-clock. It needs to work without the PLL to avoid waiting for the PLL to continuously relock while sweeping the WCK phase. Thus, the divided clock can be used since it meets the frequency requirement and is phase-locked by the PLL to the VCO clocks. Receiver, divider, and PLL are designed in CML and converted to CMOS at the central bus driver.

Fig. 7 shows the schematic of the CML–CMOS converter implemented as a pMOS loaded differential amplifier driving directly into a CMOS inverter. The gate of the pMOS is regulated such that the common mode of the output of the amplifier is held at the trip point of an inverter replicating the output inverter. This scheme can operate over the relevant frequency range from high-speed down to 50 MHz for low-speed operation. Hence, special provisions to support low-frequency modes [2] are not required. Also, the scheme eliminates the need for floating capacitors which are inefficient to design

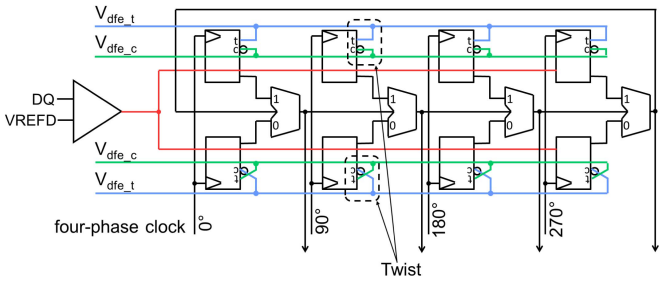

Fig. 8. Block diagram of the two-stage receiver with a differential amplifier as the first stage and a loop-unrolled DFE as the second stage.

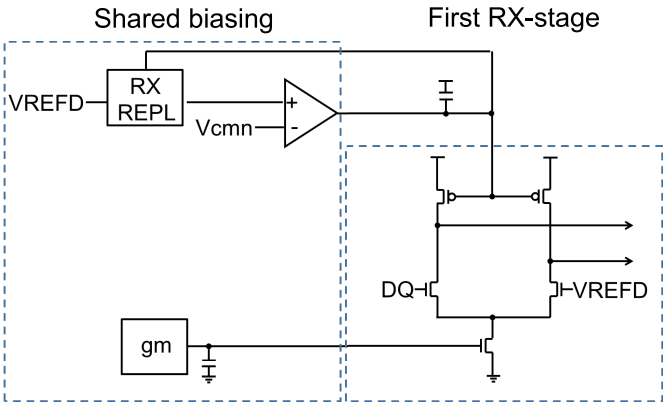

Fig. 9. Schematics of the first stage of the receiver with its dual-loop biasing circuitry. Biasing is shared between five receivers.

in the absence of high-quality MIM-capacitors as is typical for a cost-sensitive DRAM process. The simulation example shows that from input to output CMOS levels are reached. The middle node—the input to the inverter—is moving around the common mode level held at the inverter trip point by the regulation loop.

V. Receiver

The receiver implements a two-stage design (Fig. 8). The first stage is a regulated pre-amplifier driving a loop-unrolled one-tap DFE sampler as the second stage.

The first stage (Fig. 9) operates solely from VREFD and is not split for DFE. It implements two independent regulation loops. The amplifier itself is an nMOS differential amplifier with a pMOS load. The current for the amplifier is derived from a constant gm-stage to control amplification over process, voltage and temperature variations (PVT). The gate voltage of the pMOS is derived via a replica of the receiver to hold the common mode of the output of the first stage (Vcmn) at a value best suited for sampling in the next stage. For current savings, the biasing circuitry is shared between multiple DQs. The design omits the implementation of DFE at the pre-amplifier. This scheme avoids to split the first stage into two amplifiers operating at VREFD plus and VREFD minus the DFE voltage. An nMOS receiver may get critical at VREFD minus the DFE voltage since the input voltage can drop quite low. In addition, the receiver operates without reshaping techniques and in particular does not implement a continuous time linear equalizer (CTLE) which is widely used in high-speed links. However, a CMOS-CTLE implementation requires floating

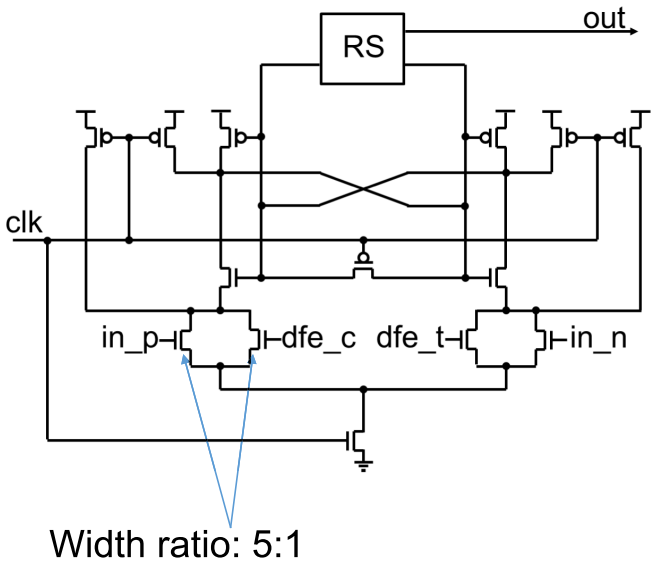

Fig. 10. Schematics of the second stage sampler with embedded DFE.

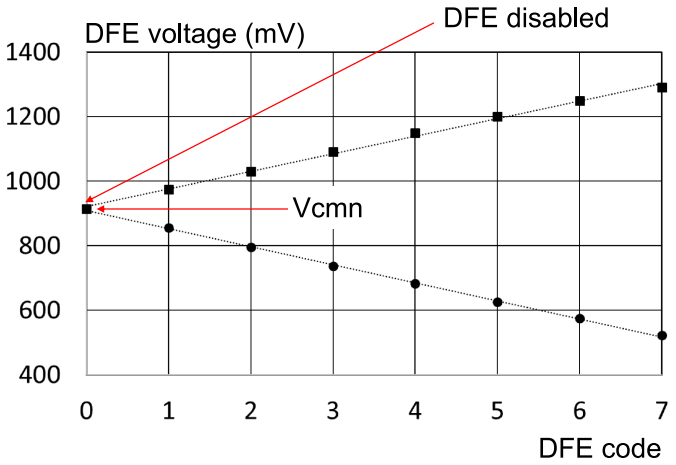

Fig. 11. Absolute values of the two voltages dfe_t and dfe_c to implement the DFE-function. At a code of 0, voltages dfe_t and dfe_c are identical and are replica biased to the common mode of the output of the first stage (Vcmn). DFE is disabled for the code.

capacitors with low parasitic capacitance. Such high-quality floating capacitors are not readily available in a DRAM process. Thus, our analysis indicated that for our process and bandwidth target the benefit of CTLE is overcompensated by the overall bandwidth degradation through adding significant parasitic capacitance to the amplifier.

Signal recovery is off-loaded to the second stage which implements a one-tap DFE in a conventional loop-unrolled architecture. Key modification is to design the DFE into the sampler itself. Besides the standard differential input pair, we add parallel nMOS transistors which are statically biased by control voltages dfe_t and dfe_c (Fig. 10). The size of the parallel devices needs to be only one fifth of the input pair. In the two branches for the two DFE states, connections between sampler and control voltages are twisted to favor receiving either the one or the zero as required by the DFE.

The control voltage for the DFE strength is digitally programmable. Fig. 11 shows the control voltage dependent on their digital control code. A code of 0 generates

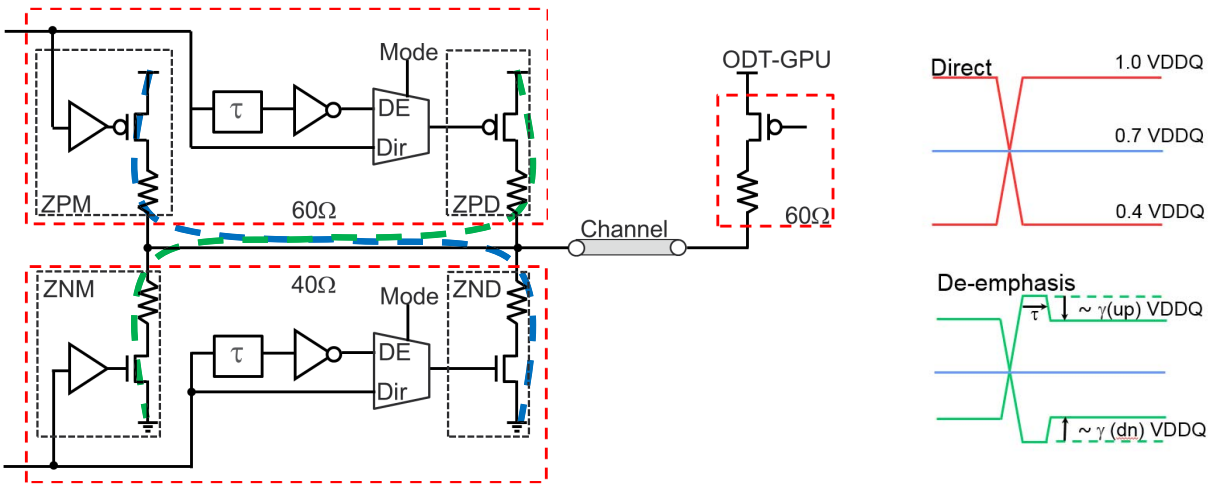

Fig. 12. Block diagram of the de-emphasis implementation of the transmitter. Both pull-up and pull-down are split in main and de-emphasis branch. Green and blue lines indicate active paths for de-emphasized pull-down and pull-up, respectively. The diagram on the left illustrates the target for the transmitter operation on the direct and de-emphasized mode. γ denotes the strength of the de-emphasis.

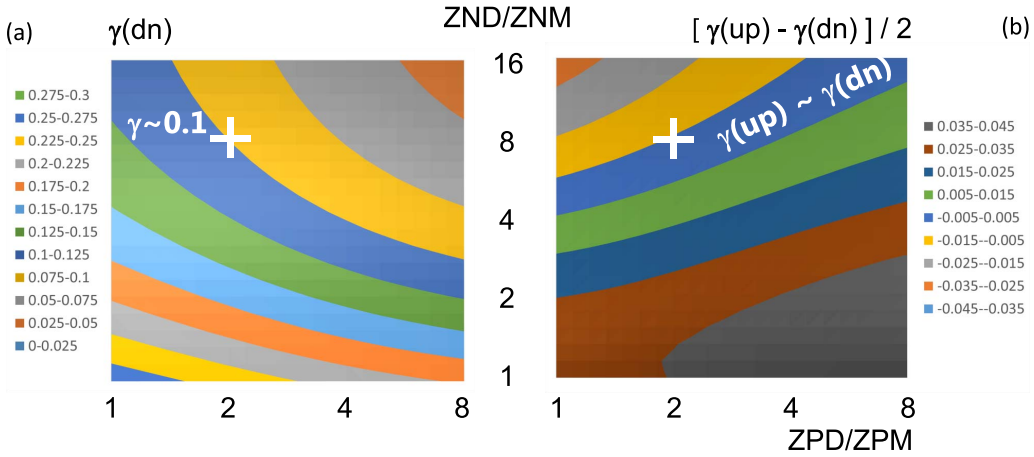

Fig. 13. Calculation results for the set of equations governing the de-emphasis. Axis are the ratio of de-emphasis to main branch for the pull-up (X) and pull-down (Y). Color codes give (a) γ value for pull-down and (b) difference between γ for pull-up and pull-down. The white cross indicates the chosen design point.

both voltages identically such that the sampler is not offset. Using this state, the system is effectively operated without DFE. With increasing digital code, a regulation circuit generates the two complimentary DFE-voltages with increasing difference. The common mode of the output is replica biased to be held at the common mode of the first stage of the receiver. At the highest calibration code, the system acts like a VREFD offset of ±80 mV at the input. On system level, the best value for the code needs to be determined and programmed by the controller during boot-up of the system.

VI. Transmitter

The transmitter implements a one-tap de-emphasis. The goal was to find an implementation which can be added in a simple way without increasing design overhead for the transmitter. In addition, it needs to be optimized for the asymmetric GDDR5X interface. The basic idea of the design is sketched in Fig. 12. Pull-up and pull-down are sliced into two sections: a main part M and a de-emphasis part D. In the direct mode without de-emphasis, both branches are driven in parallel. In de-emphasis operation, for pull-up operation the main driver of the pull-up is activated against the de-emphasis driver of the pull-down. This is indicated by the path sketched in blue. For pull-down operation, the driver activates the path sketched in green, namely, the drivers of the main pull-down and the de-emphasis pull-up paths.

To calculate the required sizing for these four impedances, pull-up and pull-down for main and de-emphasis each, we follow two rules:

1) in the direct mode, implement the nominal 60-Ω/40-Ω impedance of GDDR5X;
2) in the de-emphasis mode, realize a symmetric de-emphasis both for the pull-down and the pull-up.

De-emphasis ratios for the pull-down and the pull-up are denoted by γ (dn) and γ (up), respectively.

The rules given above lead to a set of equations by straightforward consideration of the respective voltage dividers and

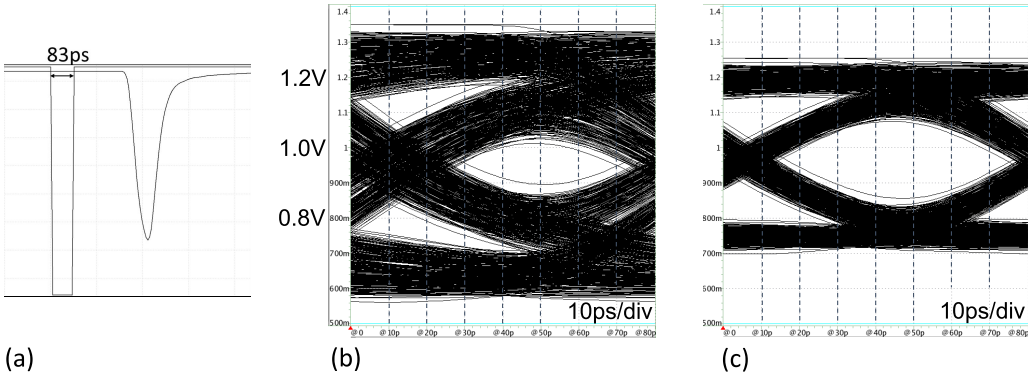

Fig. 14. (a) Pulse response of the evaluated, simple channel. Simulation results for the transmitter in (b) direct and (c) de-emphasis mode.

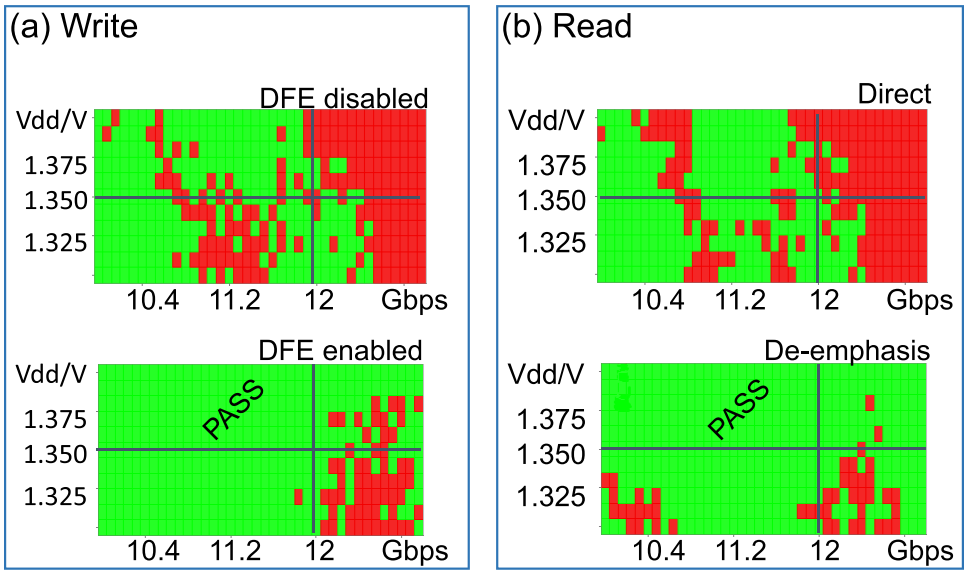

Fig. 15. Experimental results for full-array IO-stress patterns for (a) write and (b) read in an application like environment with GDDR5X compliant transmitter and termination. Data for write compare enabled/disabled DFE, while data for read compare enabled/disabled de-emphasis.

including the 60-Ω pull-up termination

$$40\Omega = \frac{ZNM \times ZND}{ZNM + ZND} \tag{1}$$

$$60\Omega = \frac{ZPM \times ZPD}{ZPM + ZPD} \tag{2}$$

$$0.4 + \gamma\,(dn) = \frac{ZNM}{ZNM + \frac{60\Omega \times ZPD}{60\Omega + ZPD}} \tag{3}$$

$$1.0 - \gamma\,(up) = \frac{ZND}{ZND + \frac{60\Omega \times ZPM}{60\Omega + ZPM}}. \tag{4}$$

The solution to this set of equations is best visualized graphically (Fig. 13). The axis for the plots is the ratio of de-emphasis to main impedance for the pull-up on the x-axis and for the pull-down on the y-axis. The colors are coding calculation results for the following:

1) the de-emphasis of the pull-down on;
2) the difference between pull-up and pull-down de-emphasis.

The goal was to de-emphasize pull-up and pull-down to a similar value. This goal is fulfilled in the blue region in the upper half of the right hand chart. A ratio of 2 for the pull-up and 8 for the pull-down falls into this regime. The same ratio achieves a pull-down de-emphasis of around 0.1 or, in terms of voltage, 0.1 × VDDQ. From the preliminary investigations, 0.1 turned out to be a good value for the de-emphasis for simple channels. The advantage of the set of ratios of 2 and 8 for driver design is the binary ratio both for pull-up and pull-down. A driver, whose strength is programmed by a digital code during calibration [7], can easily realize these binary ratios by a binary shift of the digital code.

The last part to mention for the transmitter is the implementation of the delay needed for the de-emphasis operation. Here, the conventional solution (see [8]) would suggest a digital delay of one symbol for the delayed and inverted driver activation. However, for ease of implementation, the de-emphasis branch of the transmitter implements a pure inverter delay. This avoids digital overhead in the serial-to-parallel conversion of the data path which might slow down the overall speed of the data path. The delay needs to be adjusted to the highest target data rate. As such the driver is optimized for top speed but is still fully functional at lower speed.

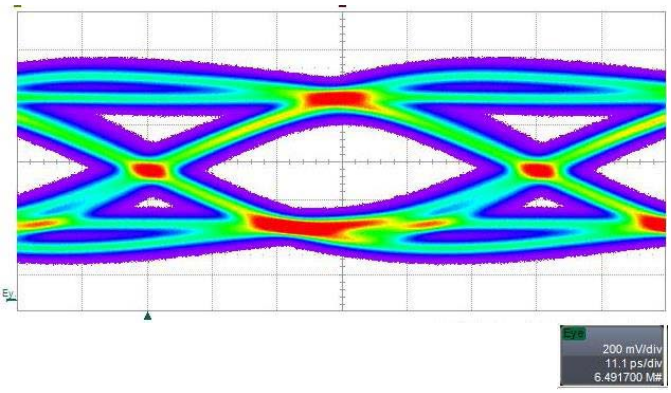

Fig. 16. Read data eye on ATE at 15 Gb/s/pin.

As an example, Fig. 14 shows simulated data eyes across an arbitrary, band-limited graphics channel without cross-coupling. The DRAM itself is fully modeled including package and power delivery in the package and the chip. Fig. 14(a) shows the pulse response of this simple channel. The data eye in direct mode is shown in Fig. 14(b), while the eye in de-emphasis mode is presented in Fig. 14(c). A clear improvement in eye-opening is apparent in simulation for this simple channel.

VII. Measurements

This section presents a subset of measurements of the GDDR5X device. Here, we initially restrict to an environment with a 40Ω channel as is typical for a graphics application, and GDDR5X compliant driver and termination impedances. By using a realistic channel and specification-compliant impedances, the benefits of DFE on the receiver or de-emphasis on the transmitter become visible. Fig. 15(a) compares the write operation to the memory between the cases of DFE-enabled and DFE-disabled. In the voltage and frequency range of interest error-free operation can only be achieved with DFE being enabled. Fig. 15(b) presents a similar comparison for the transmitter. Here, as well, error-free operation is only achieved after activating the de-emphasis mode of the driver.

Fig. 16 finally gives an outlook on the capabilities of this DRAM device in a memory tester (ATE) environment with perfect signaling and clocking conditions. This measurement shows the transmitter eye in more detail, yet operating at 15 Gb/s/pin. Thus, this GDDR5X implementation on a DRAM process can operate far beyond the initial target of 12 Gb/s/pin.

VIII. Summary

In this paper, we have presented an evolutionary extension of the GDDR5 DRAM interface to enable higher bandwidth for discrete DRAM components. We have implemented GDDR5X on a DRAM reaching 12 Gb/s/pin in an application like environment using a mixed CML and CMOS clocking system with a PLL, a one-tap DFE receiver with the DFE realized on the sampler, and a de-emphasized transmitter optimized for the asymmetric VDDQ termination.

This device demonstrates that data rates far beyond 10 Gb/s/pin can be implemented on a commercial-grade DRAM process. Hence, GDDR5 is not the end of the development for high data rate DRAM. Beyond GDDR5X, the development of the next graphics standard GDDR6 has already started. Since the targets of GDDR6 are similar to our GDDR5X design, we believe that the experience gained from GDDR5X will help to accelerate the introduction of GDDR6 into the market.

References

[1] S. J. Bae *et al.*, "A 60 nm 6 Gb/s/pin GDDR5 graphics DRAM with multifaceted clocking and ISI/SSN-reduction techniques," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, San Francisco, CA, USA, Feb. 2008, pp. 278–613.

[2] H.-Y. Joo *et al.*, "A 20 nm 9 Gb/s/pin 8 Gb GDDR5 DRAM with an NBTI monitor, jitter reduction techniques and improved power distribution," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, San Francisco, CA, USA, Feb. 2016, pp. 314–315.

[3] D. U. Lee *et al.*, "A 1.2V 8 Gb 8-channel 128 GB/s high-bandwidth memory (HBM) stacked DRAM with effective microbump I/O test methods using 29 nm process and TSV," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, San Francisco, CA, USA, Feb. 2014, pp. 432–433.

[4] J. C. Lee *et al.*, "A 1.2V 64 Gb 8-channel 256 GB/s HBM DRAM with peripheral-base-die architecture and small-swing technique on heavy load interface," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, San Francisco, CA, USA, Feb. 2016, pp. 318–319.

[5] K. Sohn *et al.*, "A 1.2 V 20 nm 307 GB/s HBM DRAM with at-speed wafer-level IO test scheme and adaptive refresh considering temperature distribution," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, San Francisco, CA, USA, Apr. 2016, pp. 316–317.

[6] R. Kho *et al.*, "A 75 nm 7 Gb/s/pin 1 Gb GDDR5 graphics memory device with bandwidth improvement techniques," *IEEE J. Solid-State Circuits*, vol. 45, no. 1, pp. 120–133, Jan. 2010.

[7] D. U. Lee *et al.*, "Multi-slew-rate output driver and optimized impedance-calibration circuit for 66 nm 3.0 Gb/s/pin DRAM interface," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, San Francisco, CA, USA, Feb. 2008, pp. 280–613.

[8] C. Kim *et al.*, *High-Bandwidth Memory Interface*. Berlin, Germany: Springer, 2013, pp. 51–59.

Martin Brox received the Diploma and Ph.D. degrees from the University of Münster, Münster, Germany, in 1988 and 1992, respectively.

In 1988, he joined Siemens Corporate Research, Munich, Germany, and in 1992, he moved to the IBM/Siemens/Toshiba DRAM development project. In 1997, he joined Siemens, Munich, Germany, which later became Infineon and Qimonda, where he was responsible for RDRAM, GDDR3, and GDDR5. In 2009, he joined Elpida (now part of Micron), Munich, where he focuses on GDDR5, GDDR5X, and GDDR6.

Dr. Brox was a member of the program committees of ISSCC and ESSCIRC.

Mani Balakrishnan received the B.S. degree from the Coimbatore Institute of Technology, Coimbatore, India, in 2000, and the M.S. degree from the University of Southern California, Los Angeles, CA, USA, in 2003.

He was with Biomorphic, San Jose, CA, USA, from 2004 to 2006, where he was involved in research on the CMOS image sensor. He was with Intel India, Bengaluru, India, from 2006 to 2012, where he was involved in the design and development of high-speed IO blocks for the Intel Architecture Group. In 2013, he joined Elpida (now Micron Technology, Inc.), Munich, Germany. He is involved in I/O high-speed design.

Martin Broschwitz was born in Hoya, Germany, in 1972. He received the "Diplom-Physiker" (M.S. equivalent) degree and the Ph.D. degree in experimental physics from Technical University Braunschweig, Braunschweig, Germany, in 1998 and 2003, respectively.

He joined Infineon Technologies AG, Munich, Germany, in 2004 as a Product Engineer with focus on production test program development. He joined the GDDR5 Product Development Team in 2007 and was involved in GDDR3 speed testing. Since 2009, he has been with Elpida, Munich, Germany, where he is involved in test pattern and coverage development for GDDR5 memories in the high-speed region.

Cristian Chetreanu received the "Diplom-Ingenieur" (M.S. equivalent) degree in electrical engineering from ETH Lausanne, Lausanne, Switzerland, in 1996.

He joined Infineon Technologies AG, Munich, Germany, in 1999 as a Wafer Test Process Engineer for embedded DRAMs. Since 2002, he has been a Test Process Engineer for component testing of graphics DRAMs. Since 2009, he has been with Elpida, then Micron, Munich, for DRAM characterization.

Stefan Dietrich was born in Munich, Germany, in 1965. He received the Diploma degree in physics from the Technical University of Munich, Munich, in 1993, and the Ph.D. degree from the University of Augsburg, Augsburg, Germany, in 1996.

In 1996, he joined the Memory Products Division, Infineon Technologies (formerly Siemens Semiconductors), Munich, which became Qimonda AG, where he was involved in the development of high-speed graphics dynamic memories and emerging memory platforms such as CBRAM and PCRAM. In 2009, he joined Elpida Memory, Munich, where he was responsible for data path design of GDDR5 and the consulting of GDDR3 graphics memory. In 2014, he joined Micron Semiconductor, Munich (Elpida Memory was transferred to Micron Semiconductor), as the Design Lead of an 8-Gb, 5.5-Gbps GDDR5 design, where he is responsible for high-speed graphics data path design for GDDR5, GDDR5x, and GDDR6 designs, respectively. He holds more than 50 U.S. and foreign patents.

Dr. Dietrich is a member of the German Physical Society.

Fabien Funfrock received the Diploma degrees in physics from the ENSPG (now known as Phelma/INP), Grenoble, France, and from the University of Karlsruhe, Karlsruhe, Germany, in 1998.

From 2000 to 2009, he was with Infineon (later Qimonda), Munich, Germany, as a Test Engineer, where he was involved in the analysis and optimization of commodity and graphic DRAM, and then became a Staff Engineer. In 2009, he joined Elpida, Munich Design Centre, Munich (now Micron), where he was involved in GDDR5, and since 2013 he has been with the Design Department.

Marcos Alvarez Gonzalez received the M.S. degree in computer science and engineering from the Universidad de A Corunna, Corunna, Spain, in 2008, and the M.S. degree in computer science and artificial intelligence from Arizona State University, Tempe, AZ, USA, in 2010.

He joined Micron Technology Inc., Boise, ID, USA, in 2010 as a Quality and Reliability Assurance Engineer. Since 2015, he has been with Micron Semiconductor Deutschland, Munich, Germany, as a Senior Product Engineer.

Thomas Hein received the Diploma degree in information technology from the Technical University of Dresden, Dresden, Germany, in 1995.

In 1995, he joined Siemens Semiconductors, Munich, Germany, which became Infineon Technologies and later Qimonda AG, Munich, where he led the design of multiple multi-bank MDRAM, SGRAM, GDDR1/3/4/5 designs. From 2009 to 2014, he was with Elpida Memory (Europe) GmbH, Munich, Germany. He was the Design Lead of the 8G GDDR5X. In 2014, he joined Micron Semiconductor (Deutschland) GmbH, Munich, where he is currently involved in the definition and design of various high-speed GDDR5, GDDR5X, and GDDR6 DRAMs. His research interests include DRAM design, new DRAM architectures, chip packaging, and high-speed interfaces.

Eugen Huber received the "Diplom-Ingenieur" (M.S. equivalent) degrees in electrical engineering from the University for Applied Sciences Munich, Munich, Germany, and the Technical University of Munich, Munich, in 1990 and 1996.

He joined Siemens Semiconductor Division (later on Infineon, Qimonda), Munich, in 1998, as a Test Process Engineer for wafer and component test for logic chips used in automotive applications. In 2007, he joined the GDDR5 team as a Test Engineer with focus on bench test system enabling. Since 2009, he has been with Elpida (now Micron), Munich Design Center, Munich.

Daniel Lauber, photograph and biography not available at the time of publication.

Milena Ivanov received the master's degree in telecommunications from the Technical University in Sofia, Sofia, Bulgaria, in 1996, and the Diploma degree in electrical engineering from the Technical University of Munich, Munich, Germany in 2005.

She joined the Memory Products Division, Infineon Technologies (which became Qimonda AG), Munich, in 2005, where she was involved in the development of emerging memories, particularly, conductive bridging RAM. In 2009, she joined Elpida (now Micron), Munich, where she was involved in the development of GDDR5, GDDR5X, and GDDR6. She has authored or co-authored several U.S. patents.

Maksim Kuzmenka was born in Minsk, Belarus, in 1972. He received the M.S. degree in electrical engineering from the Belarusian State University of Informatics and Radioelectronics, Minsk, in 1993.

In 1993, he joined BELOMO, Minsk, where he was involved in switched mode power supply development. In 2001, he joined Infineon Technologies and later Qimonda, Munich, Germany, as a signal integrity and mixed-signal Design Engineer. In 2009, he joined Elpida Memory, Munich, with focus on DRAM I/O and power management circuit design. Since 2014, he has been with Micron Memory, Munich, where he is involved in mixed signal circuit design.

Christian N. Mohr received the B.S. degree in electrical engineering from the University of Arkansas, Fayetteville, AR, USA, in 1998.

He joined Micron Technology, Inc., Boise, ID, USA, in 1998 as a DRAM Design Engineer. He has been involved in the design process for multiple memory architectures including SDR, DDR, DDR2, DDR3, DDR4, and RLDRAM. He is currently with Micron, where he focuses on the design and testing methodologies for these and various other devices.

Juan Ocon Garrido received the M.S. degree in electrical engineering and the M.S. degree in physics from the University of Granada, Granada, Spain, completed in 2002 and 2004, respectively.

He joined the Memory Products division of Infineon Technologies AG in 2003, in the Product Engineering group, being responsible for burn-in stress and test coverage definition for DDR, DDR2, DDR3 and the first GDDR5. With the Qimonda AG Graphic Memory group he moved to Elpida's Munich Design Center, now part of Micron Technology Inc., where he has been involved in the development of GDDR5, GDDR5X, and GDDR6.

Swetha Padaraju was born in Hyderabad, India, in 1987. She received the M.Sc. degree in information and communication engineering, majoring in microelectronics, from the Technical University of Darmstadt, Darmstadt, Germany, in 2012.

She joined the Graphics Memory Division, Elpida Memory GmbH, Munich, Germany, in 2012, which was later acquired by Micron Technology Inc., Boise, ID, USA. She is an Analog Design Engineer and is mainly responsible for power systems.

Sven Piatkowski received the Physical Designer degree from the Siemens Technical Academy, Berlin, Germany, in 2002.

He joined Infineon Technologies AG, Munich, Germany, in 2002, where he was involved in the P&R block level implementation for DRAM and FLASH products. In 2004, he joined P&R Design Group, as the Team Lead, where he was responsible for the development of various semi-custom blocks and standard cell libraries for the worldwide DRAM and FLASH products. In 2006, he joined Qimonda AG as a Senior Engineer, where he started establishing a semi-custom flow with special focus on place and route, static timing analysis, and layout verification. In 2008, he joined the GDDR5 team and played a major role in supporting the first implementation of a graphic DRAM in this semi-custom flow. He joined Elpida (now part of Micron), Munich, in 2009, where he is responsible for standard library development, floor planning, place and route, and layout verification.

Jan Pottgiesser received the Diploma degree from Ruhr-University Bochum, Germany, in 1994.

He joined Siemens-Halbleiter, Munich, Germany, which became Infineon Technologies AG, in 2001 as the Team Lead of the Test Chip Design Group, where he was responsible for the development of various full-custom and semi-custom test chips. He joined Qimonda AG in 2006 as a Staff Engineer, where he started establishing a semi-custom flow with special focus on place and route, static timing analysis, and layout verification. In 2008, he joined the GDDR5 team and played a major role in supporting the first implementation of a graphic DRAM in this semi-custom flow. He joined Elpida (now Micron Technology), Munich, in 2009, where he was involved in standard library development, floor planning, place and route, clock-tree synthesis, layout verification, and STA.

Peter Pfefferl received the B.S. degree in electrical engineering from Gesamthochschule Kassel, Kassel, Germany, in 1983.

He joined Siemens, Munich, Germany, in 1983, where he was involved in mainframe development, especially in CPU and main memory system development. In 1994, he joined the DRAM Development Alliance, a joint project among Siemens, Toshiba, and IBM, where he was the part of 256M product design. In 1997, he joined Siemens (later Infineon, Qimonda), Munich, where he was involved in several design areas of commodity DRAMs. Since 2009, he has been with Micron (formerly Elpida), Munich Design Center, Munich, where he is involved in the design of GDDR5, GDDR5x, and GDDR6 graphics DRAMs.

Manfred Plan was born in Augsburg, Germany, in 1960. He received the diploma degree in electrical engineering from the Technical University of Munich, Munich, Germany, in 1987.

He was a Research Assistant and Project Ing. of integrated biosensors with the TUM/Fraunhofer Institute, Munich, and he joined the R&D Department, Siemens AG, Munich, in 1988, where he was with different departments for product development, and also with the University of California, Berkeley, CA, USA, and in the production line in Villach, Austria. From 1990 to 1998, he was with the Consumer Electronics Division (Megatext, picture in picture, TV scanrate converter), where he was involved in the full-custom and semi-custom design of chips, memory cores, and interfaces, and from 1998 to 2000, he was with Siemens/Infineon Technologies AG, Munich, where he was involved in the semi-custom design of a crypto microprocessor IC. In 2000, he joined the Memory Product Development Division, Infineon/Qimonda AG, where he was involved in the design and development of embedded DRAMs, reduced latency and high-speed graphics DRAM (RLDRAM, GDDR3, GDDR5) with focus on core-, array-, DLL, chip verification and full-custom design. From 2009 to 2014, he was with Elpida Memory (Europe) GmbH (now Micron Technology, Inc.), Munich, where he was involved in core circuit development for DDR4 and GDDR5 up to 7 Gbps, and since 2014, he has been with Micron Technology Inc., where he was involved in GDDR5, GDDR5X and GDDR6.

Mr. Plan is a member of FEANI.

Jens Polney received the Diploma degree in electrical engineering from the University of Magdeburg, Magdeburg, Germany, in 1997.

In 1998, he joined Siemens Semiconductors, Munich, which became Infineon Technologies and Qimonda AG, Munich, Germany, where he was involved in SDR, DDR1, DDR2, DDR3 and Rambus DRAM designs. From 2009 to 2014, he was with Elpida Memory (Europe) GmbH, Munich, Germany. In 2014, he joined Micron Semiconductor (Deutschland) GmbH, Munich, where he is currently a Design and Verification Engineer at various high-speed DDR4, GDDR5, GDDR5X, and GDDR6 DRAMs.

Stephan Rau received the B.Eng. degree in electrical engineering and information technology from the University of Applied Sciences, Munich, Germany, in 2007.

He joined Qimonda AG in 2007, where he was with the Library Development CAD Group. He joined Micron Technology, Inc. (at this time Elpida Memory Europe GmbH, Munich) in 2011, where he is responsible for PNA, RTL coding, library characterization, synthesis, and STA. Since then, he has been involved in the development of different Graphic DRAMs (GDDR5, GDDR5X, and GDDR6).

Michael Richter received the Diploma degree in electrical engineering from the Technical University of Munich, Munich, Germany, in 1982.

He joined Siemens Semiconductors, Munich, in 1984, where he was involved in the application-specific integrated circuit (ASIC) design and design support for ASIC customers. He was with Infineon Technologies AG, Munich, where he was responsible for the product definition of smart card ICs and served as a Program Manager of a high-speed crypto IC project. He was with Qimonda AG, where he was responsible for the product definition and standardization of GDDR5, which he continued with Elpida Memory, Munich, for GDDR5X and HBM, and since 2014, he has been with Micron Technology Inc., for GDDR6.

Ronny Schneider was born in Soemmerda, Germany, in 1976. He received the M.S. degree in physics from Friedrich-Schiller-University, Jena, Germany, in 2001.

He joined Infineon Technologies Corp., Burlington, VT, USA, in 2001, where he was involved in circuit design and failure analysis for DDR and DDR2 SDRAMs. From 2004 to 2006, he was the Team Leader of a DDR2 SDRAM design, Xi'an, China. Since 2006, he was with Qimonda AG, Munich, Germany, where he was responsible for the full-chip verification of the GDDR5 SDRAM products. In 2009, he joined Elpida Memory (Europe) GmbH, Munich, Germany. With the acquisition of Elpida by Micron, he was with Micron Semiconductor (Deutschland) GmbH, Munich, where his responsibilities included full-chip verification of all GDDR5/5X/6 SDRAM products and circuit design, and was the Design Team Lead of several GDDR5 SDRAM products.

Ralf Oliver Seitter, photograph and biography not available at the time of publication.

Wolfgang Spirkl received the Diploma and Ph.D. degrees in physics from Ludwig-Maximilians-University, Munich, Germany, in 1986 and 1990, respectively.

From 1998 to 2009, he was with Siemens (later Infineon, Qimonda), Munich, as a Principal Engineer, where he was involved in the verification of embedded, network, commodity, and graphic DRAM. In 2009, he joined the Munich Design Centre, Elpida, Munich, where he worked on GDDR5, and he joined Micron with the merger in 2014.

Marc Walter received the Dipl.Ing. degree in electrical engineering from RWTH Aachen University, Aachen, Germany, focusing on computer science, in 2000.

He joined Infineon Technologies, Munich, Germany, in 2000, where he was a Physical Design Engineer of customer specific circuits with the Embedded Memory Group. In 2002, he moved to the Center of Competence Design for Testability (DfT) in product development at Infineon/Qimonda, where he was responsible for the conception, elaboration, and implementation of DRAM test methodologies and DfT circuits for DDR2, DDR3, and GDDR5 products. In 2009, he joined Elpida (now Micron Technology, Inc.), Munich, and continued his work on GDDR5, DDR4, and recently GDDR5X.

Jörg Weller (M'07) was born in Munich, Germany, in 1968. He received the Diploma degree in electrical engineering (M.S.E.E.) from the Technical University of Munich, Munich, in 1994.

He joined Siemens AG, Munich, in 1994, where he was responsible for design and layout of a 4 M-bit DRAM, and with production wafer testing for RDRAMs in 1998. In 2001, he joined Infineon AG, Munich, where he started design analysis for graphics DRAMs. He joined the spin-off of Qimonda AG in 2006, where he was involved in design analysis of GDDR3 and GDDR5 interfaces. In 2009, he joined Elpida Memory Europe, Munich, where he was involved in graphic DRAM development. With the transition to Micron in 2012, he has been involved in GDDR5, GDDR5x, and GDDR6 graphic DRAM.

Filippo Vitale received the master's degree in electronic engineering from the University of Naples "Federico II," Naples, Italy.

In 2001, he joined the Memory Product Group, STMicroelectronics, Naples, which became Numonyx BV, then Micron Semiconductor, as a CAD Engineer, involved in the development and deployment of design automation methodologies mainly in the field of analog and mixed signal verification. Since 2014, he has been with the Munich Design Center of Micron, where he is involved in the development of GDDR5, GDDR5X, and GDDR6.