Design, fabrication and testing of the package-on-package microcircuit

K R Marcinkevich¹, P V Lunkov¹, D A Kirienko¹, V V Putrolaynen¹, M A Belyaev¹ and A V Yartsev²

¹Institute of Physics and Technology, Petrozavodsk State University, 33 Lenin str., 185910, Petrozavodsk, Russia
²GS Nanotech, 11 Industrialnaya str., Gusev, 238050, Russia

E-mail: marcink-kristina@yandex.ru

Abstract. The paper presents the results of the development of a package-on-package multichip microcircuit with three-dimensional integration of processor and memory dies. The stages of fabrication the upper part of the microcircuit with several memory dies, which are mounted by the wire bonding method, and the lower part of the microcircuit, on which the processor die is attached by the flip-chip method, are described. After the fabrication of the microcircuit by combining the upper and lower substrates, a multi-stage functional testing of the processor and memory dies is performed using a test board and software loaded into the microcircuit memory.

1. Introduction

In recent decades, the trend towards miniaturization of electronic devices pushes manufacturers to search for solutions to reduce the size of printed circuit boards and component base [1]. At the same time, there is a demand to reduce the power consumption of devices along with increasing their performance [2]. One of the options for solving these problems is to reduce the topological norms in the processes of manufacturing silicon dies. However, by now the microelectronic industry has come close to the fundamental limitations of miniaturization of field-effect transistors [3] - the main elements of digital logic and memory. Therefore, planar technologies for manufacturing silicon dies are being replaced by three-dimensional ones, in which a plurality of planar layers connected by vertical channels are made on a silicon wafer [4]. Such technologies, called monolithic three-dimensional integration, are currently successfully used in the manufacture of DRAM [5] and NAND [6] memory dies, but are not used in the manufacture of three-dimensional logic integrated circuits due to the high heat emission and the large number of input/output (I/O) lines.

In the context of the development of electronic devices, three-dimensional integration of logic and memory integrated circuits is of great interest, since most modern devices include these components. However, there are currently no successful examples of monolithic 3D integration of such devices. Instead, three-dimensional integration at the package-level is used [4], in which several dies are connected at the stage of packaging. Dies can be connected using vertical channels formed by deep silicon etching [6], or using intermediate inserts or substrates [7, 8]. The second option has more flexibility both in layout options and in the selection of dies manufacturers, allowing to change the number and model of processor, DRAM, NAND and NOR memory dies in the microcircuit. The three-
dimensional integration method called "package-on-package" [7,9] is very popular among such packaging methods.

A “package-on-package” microcircuit consists of two substrates that are connected to each other using solder balls. The upper substrate contains the memory dies, and the lower one contains the processor die. There are solder balls on the underside of the bottom substrate for attaching to the motherboard. The location of the processor die at the bottom is caused by the requirements for a large number of I/O lines for communication with peripheral devices located on the motherboard. The solder balls connecting the substrates together are used to establish communication between the processor and memory. One of the main advantages of such vertical integration is the reduction in the size of the motherboard (and the final device) because of the fact that memory microcircuits do not occupy separate places on the motherboard [10]. The length of the conductors between the processor and the memory is also reduced, which makes it possible to increase the speed of data exchange between them [10], while reducing energy losses from resistance heating. In addition, the use of “package-on-package” microcircuits allows to reduce the number of layers on the motherboard, thanks to fewer intersecting lines compared to direct mounting of processor and memory microcircuits on the motherboard.

This paper describes the development, fabrication and testing of “package-on-package” microcircuits using processor, NOR, NAND and DRAM memory dies. The paper describes the main stages of developing a microcircuit using a computer-aided design system and the stages of its fabrication with a description of the sequence of applied technological processes. In addition, the paper presents a procedure for testing a microcircuit after its fabrication, including testing the basic functions of the processor core, NOR, NAND and DRAM memory, as well as interfaces for interaction with peripheral devices.

2. Microcircuit development

The development of the microcircuit was carried out taking into account its purpose as a control computing module as part of a modular system for collecting and analyzing information. This module must interact with external peripheral devices for collecting and aggregating data using network protocols of the Ethernet standard. For processing and analyzing data, including using artificial neural networks, hardware acceleration is used with the help of a tensor computing module connected via the USART interface. The storage of primary and processed data is carried out by data storage modules, interaction with which is carried out via the SATA interface. Also, a 9-axis inertial sensor is connected directly to the control computational module using the SPI interface.

To implement the presented tasks, a sufficiently powerful processor with hardware implementation of the specified interfaces is required. In addition, the processor should have reduced power consumption, since due to the packaging method, the heat generated from the processor can be efficiently dissipated only through the motherboard on which the microcircuit is mounted. Heat sink through the upper radiator is difficult due to the presence of the upper substrate with memory microcircuits and, as a consequence, significant thermal resistance between the processor die and the upper side of the microcircuit. Taking into account the above, when developing the microcircuit, a 64-bit single-core processor with MIPS architecture was used, which has a maximum power consumption of 2 W when the core is running at 400 MHz. frequency. The Linux operating system was chosen to control the modular system for collecting and analyzing information.

Based on the results of the preliminary assessment of memory requirements, it was found that to effectively solve the assigned tasks it is necessary to have: 1 GB of random-access memory, 512 MB of non-volatile memory for storing operating system files and program modules, and 32 MB of non-volatile memory for storing boot software and service information required for the processor work. To do this, it was proposed to place on the upper substrate 2 dies of DDR3L random-access memory with a volume of 512 MB each, one die of NAND memory with a volume of 512 MB and two dies of NOR memory with a volume of 16 MB each.

When developing a package-on-package microcircuit, one of the main difficulties is the design of substrates for attaching dies. During this design it is necessary to solve a number of problems. The first
problem is related to the density of the memory dies on the upper substrate. Memory dies have a significant area (DDR - 10x8 mm, NAND - 9x7 mm, NOR - 3x3 mm) and are mounted using the wire bonding method, in which part of the substrate area is occupied by contact pads, while the substrate size is only 21x21 mm. The second problem is associated with a large number of I/O lines of the processor, which is mounted on a substrate using the flip-chip method, and the need to organize heat sink through the substrate to the motherboard. The third challenge relates to the need to use the bottom substrate to make electrical connections to both the top substrate and the motherboard. An effective solution to these problems was achieved through the use of the Cadence SIP Layout software package, which allows to develop substrates for multichip microcircuits, taking into account various dies sizes, the location of contact pads, the mounting method (wire bonding, flip-chip), and analyze the result on a three-dimensional model of the microcircuit. Figure 1a shows an image of the upper and lower substrate of the developed microcircuit. During the design, it was found that to solve the set tasks, the upper substrate should have 6 copper layers, and the lower one – 8 layers. Using the Cadence Sigrity power circuit analysis package, “hot spots” were identified and eliminated; the current in them exceeded the recommended one and caused a drop voltage more than by 1%. The traces and vias have been optimized. Figure 1b shows the simulation results in the upper copper layer of the lower substrate.

![Figure 1](image.png)

**Figure 1.** Image of the upper and lower substrates (a), developed in Cadence SIP Layout, with DDR (1), NOR (2), NAND (3) memory and processor (4) dies attached, and the results of voltage simulation in the upper copper layer with using Cadence Sigrity (b).

### 3. Microcircuit fabrication

Package-on-package microcircuits fabrication is a multi-stage process: thinning and dicing silicon wafers into separate dies, attaching of dies on a substrate, creating electrical connections between dies and a substrate, sealing with a mold compound, mounting solder balls, dicing strips into separate substrates, connection of the upper and lower substrate into a single microcircuit.

Silicon wafers were thinned using a grinding and polishing machine DGP 8761. This operation was performed to reduce the thickness of the dies from 900 µm to 180-250 µm, which made it possible to reduce the final thickness of the microcircuit. Before the thinning operation, the face side of the wafer (figure 2a) was laminated with a protective film to avoid damage to the functional layer. The thinning process was monitored using built-in thickness gauges and included the stages of coarse and fine grinding and dry polishing. After thinning, the plates were mounted on a frame with a film for further dicing into individual dies (figure 2b). The film on the frame, on which the DDR and NAND memory
dies were mounted, had an additional adhesive layer, which was used later when mounting the dies on the substrate. This type of film is called DAF (die attach film). After mounting of the wafer on the frame, the protective film, which was glued before thinning, was removed from the front side of the wafer. To do this, it was exposed to ultraviolet radiation, which led to the destruction of the photosensitive adhesive layer.

![Figure 2](image)

**Figure 2.** Lamination of the front side of a silicon memory wafer before thinning (a) and a thinned wafer mounted on a frame with the film before dicing (b).

All silicon wafers were diced using diamond circular saws by two-stage method (step cut). First, most of the wafer thickness was diced with a wider saw, and then the remainder was diced with a narrower saw. This method makes it possible to reduce the size of the chips of the silicon wafer. If the film on the frame had an additional adhesive layer (DAF film), then it was also cut at the second stage. The dicing of silicon memory wafers was carried out on the DFD 6361 HC machine, and processor wafers were diced on the DFD 6560 machine. Figure 3 shows the cutting lines of the silicon memory and processor wafers. It should be noted that before dicing the processor wafers, the grooving of the surface polyamide layers in the area of the planned cutting line was used with the help of a laser beam on the DFL 7160 machine. If you do not carry out the grooving procedure and immediately start dicing with diamond saws, then the size of the resulting chips could exceed acceptable values, up to the formation of cracks in the wafer. This could also lead to the appearance of hidden defects in the form of delamination of the wafer layers. After dicing, the film on the frame was exposed to ultraviolet radiation to facilitate the separation of the die from the film by reducing adhesion.

At the next stage, the dies were attached on the substrates and electrical connections were created between the contact pads of the dies and the substrate. The size of the substrates was only 21x21 mm, therefore, to optimize production processes, several substrates were combined into one, the so-called strip, which in our case consisted of 18 single substrates (2 rows of 9 pieces). It should be noted that not all dies on the wafer were functional and some of them had internal defects that were identified during testing at the manufacturing plant. Dies that were working correctly were marked as known good die on the map used on the die mounting machine. Datacon 2200 EVO was used to mount dies and passive components (resistors and capacitors) on substrates. It had an optical positioning system, a set of accessories for separating dies from a film and placing them on a substrate, and a set of dispensers for applying liquid glue, flux, or solder paste. The memory and processor dies were attached using different mounting methods, so the further description of technological processes for them varies.
The memory dies were attached on the upper strip in three stages: at the first stage, passive components (resistors and capacitors) were attached using solder paste by the surface mount method. These components were then soldered by reflowing of the solder paste in an oven. At the second stage, two NOR memory dies were attached. Since the film on the frame with NOR memory dies did not have an additional adhesive layer, liquid glue was used to mount the dies, which was applied to the strip using a dispenser. To separate the dies from the film, a needle ejector was used - a special tool that presses on the die from under the film during the rise of the die using a vacuum gripper. It is used to stretch the film and facilitate the separation of the die from the film, which could otherwise lead to damage (cracking, chipping, etc.). After separation of the die, it was transferred to the attachment site and glued to the strip. After attaching all the NOR memory dies, the strip was placed in an oven to polymerize the glue. At the third stage, NAND and DDR memory dies were attached. At this stage, no liquid glue was applied to the strip, since there was a film adhesive layer on the bottom of the dies thanks to the use of a DAF film. After separation of the dies from the film (the process is similar to the separation of NOR memory dies), they were attached on the strip with a force sufficient for the adhesive film to glue evenly to the surface of the strip without the formation of air bubbles. Thereafter, the strip was placed in an oven to polymerize the adhesive film layer. Figure 4 shows the images of a single substrate and a strip after attaching all the memory dies.

**Figure 3.** Image of cutting lines of a memory (a) and a processor (b) wafers.

**Figure 4.** Image of a single substrate (a) and a strip (b) with attached NOR (1), NAND (2) and DDR (3) memory dies.

After attaching the memory dies between the contact pads of the dies and the strip, wire connections were created using the technology of thermosonic wire bonding. Immediately before the wire bonding, the strips with dies were subjected to plasma treatment to remove contamination from the surface of the
contact pads and to activate the surface to improve the wire bonding process. The strips were placed in the ASM Eagle Express GoGu machine where they were preheated to reduce the ultrasonic power during bonding. At the first stage of wire bonding, a ball was formed at the free end of the wire because of melting under the action of an electric discharge. Then, the formed ball was pressed against the contact of the die by a bonding capillary with the supply of ultrasound, forming a welded joint. Then the capillary rose and formed a loop of wire, which ended at the second contact on the contact pad of the strip. Here the capillary pressed the wire again and, under the influence of ultrasound energy, formed a second welded joint. After the second connection was made, the wire was broken so that a new wire bonding cycle could be started. To create connections between the required contact pads, a “die-strip” connection map was loaded into the software of the wire bonding machine. The search for the required contact pads was carried out using the optical system of the wire bonding machine. Figure 5 shows an image of NOR, NAND and DDR memory dies after the thermosonic wire bonding procedure.

At the next stage, the strips were sealed using an epoxy mold compound to create mechanical and chemical protection of dies and welded joints from external influences. To improve the adhesion of the liquid mold compound to the strip surface, it was cleaned in an oxygen-hydrogen plasma. The Fico AMS-i306 machine was used to seal the strips. Immediately before placing the strips into the mold, they were heated to minimize the temperature difference with the molten mold compound and to avoid sudden temperature changes that can lead to large warping of the strips. After heating, the strips were placed into the mold in two pieces and were poured under pressure with molten mold compound. Then, heat treatment was carried out to complete the process of polymerization of the mold compound, which contributed to the partial relaxation of mechanical stresses in the strip caused by the difference in the coefficients of thermal expansion of the materials used (silicon, strip material, mold compound). Figure 6 shows the image of the strips before and after sealing with a mold compound.

![Figure 5](image1.png)

**Figure 5.** Image of NOR (a), NAND (b), DDR (c) dies together with wire electrical connections after the wire bonding process.

![Figure 6](image2.png)

**Figure 6.** Strips before (a) and after (b) sealing with mold compound.
The processor was mounted on the lower strip by the flip-chip method, in which the die was soldered to the strips using solder balls mounted by the manufacturer at the stage of fabricating silicon wafers. Before attaching the dies, the strips were treated with an oxygen-argon plasma to clean the contact pads and improve the quality of soldering. The attachment of the processor die and passive components (resistors and capacitors) was carried out using the Datacom 2200 EVO machine. First, using dispensers, solder paste was applied to the strip for mounting passive components and flux for soldering the die. Then, using vacuum grippers, the passive components and the die were attached on the substrate. After that, the strips were placed into an oven to reflow the solder paste and solder balls. Then the space between the die and the strip was filled with a special glue (underfill), which reduces the likelihood of mechanical destruction of the solder balls due to differences in the coefficient of thermal expansion between silicon and the substrate material during the heating/cooling cycles of the microcircuit during its operation. To polymerize the underfill, the strip was again placed into the oven. Since on the upper side of the lower strip there are contact pads for connecting the lower and upper parts of the "package-on-package" microcircuit using solder balls, the lower strip is not sealed with a mold compound. Figure 7 shows the image of the strip after mounting the processor dies.

![Figure 7. Image of a strip with attached processor dies.](image1)

The following technological processes are the same for the bottom and top strips. Firstly, the solder balls were mounted on the underside of the strips. The balls on the upper strip will be used to connect the substrates to each other, and on the lower one - to connect the fabricated microcircuit to the motherboard. Before placing the balls, the strips were treated in an oxygen-argon plasma to clean the contact pads. Then, using a Shibuya SBM 351 machine, flux was applied on the contact pads of the strips and solder balls were mounted. Then the strips were placed into a Rehm V8 Nitro oven to melt the balls and solder them to the contact pads. After that, using the dicing machine the strips were diced into separate substrates. Figure 8 shows the image of the resulting lower and upper substrates.

![Figure 8. Upper (a) and lower (b) substrates of a “package-on-package” microcircuit.](image2)
At the last stage of fabrication the microcircuit, the lower and upper substrates were soldered to each other. Before this, the lower substrate was treated in an oxygen-argon plasma to clean the contact pads before soldering. Using a Datacom 2200 EVO machine, flux was applied to the contact pads of the lower substrate, and the upper substrate was mounted on the lower one. The assembly was then placed into an oven to reflow the solder balls and form a soldered joint. Figure 9 shows the image of the microcircuit after soldering, side view. Then the space between the microcircuits was filled with glue to mechanically protect the bottom die and improve thermal contact between the two substrates. Polymerization of the glue was carried out in the oven.

![Image of the "package-on-package" microcircuit after soldering, side view.](image)

**Figure 9.** Image of the "package-on-package" microcircuit after soldering, side view.

4. Microcircuit testing

To test the fabricated microcircuit, a test board was used, which had ports for connecting peripheral devices with USB, SATA, PCI Express, Ethernet, I2C interfaces. The testing process was controlled using a workstation that was connected to the test board using USB and Ethernet interfaces. The boot software was installed into the NOR processor memory, with the help of which the test software was loaded into the microcircuit memory and the commands were transmitted from the workstation.

The first group of tests checked the operation of the cache memory, the translation lookaside buffer, the control and arithmetic coprocessor. The procedures of writing and reading a predefined sequence of data were used to check the data cache of the 1st and 2nd levels. The 1st level instruction cache was checked by filling in branch instructions with writing and reading from memory. The operation of the translation lookaside buffer was checked both through the operations of writing and reading data into the buffer, and through checking the mechanism of translating virtual addresses. To check the control coprocessor, the command to read the processor identifier was executed. The arithmetic coprocessor was tested using the matrix addition operation. After that, the operation of the processor timers were tested using the following procedures: writing of the counting interval into the boot registers, starting and stopping counting, checking with enabled and disabled interrupts, checking in single and cyclic counting modes, checking the generation of a non-maskable interrupt, checking the timer outputs on the block port lines of discrete signals, interrupt priority check.

The next step was functional testing of DDR, NOR and NAND. Firstly, we checked the processor's operation with DDR RAM. After measuring the amount of available memory, the correct operation of the data bus and the address bus was checked using sequential data writing and reading according to the "walking ones" algorithm. Then the complex testing of DDR memory was carried out by means of
writing and reading data according to the algorithms of “walking ones and zeros”, “moving inversions” and “modulo-X”. At the last stage, the mode of direct memory access was tested with the transfer of data blocks of different sizes and with a change in the offset of the address of receiving and sending data.

Then the NOR memory and the operation of the SPI interface were tested using the following procedures: reading the memory device identifier, reading and checking the status register, multiple data readings with the calculation of the checksum in standard and fast reading modes, checking the checksum written into the memory of the boot program, writing and erasing data in the free memory area, writing into the status register, checking the operation of interrupts at the end of the cycle on the SPI interface.

When checking NAND memory, the identifier of the memory device was read and checked, the number of bad blocks of memory was requested, which was compared with the threshold value allowed for the correct operation of the microcircuit. Then, multiple tests were performed with reading data and calculating a checksum, as well as sequential writings into the memory of predefined constants, followed by verification using the read data.

After testing the basic functionality of the processor and memory, we checked the operability of the interfaces for interacting with peripheral devices: USB, SATA, Ethernet, PCI Express, I2C. To test the USB interface, a USB flash drive connected to the test board and the values of the block registers were read, the device identifier was checked, and a connection to the USB flash drive was established.

To check the SATA interface, SATA disks were connected to the test station, after which test procedures were performed: checking the connection of devices, performing a self-diagnosis procedure, reading device parameters (model, serial number, amount of memory, logical sector size, etc.), writing data into various sectors of the device with subsequent verification using the sector reading procedure, reading sectors in multi-sector mode, verification of reading and writing operations in direct memory access mode.

When testing the Ethernet interface, we used three transmission speed modes (10, 100 and 1000 Mbit s) with forced and automatic speed settings. During testing, the following checks were performed: checking the device identifier, transmitting and receiving the tuning data packet, changing the size of the transmitted data, changing the number of descriptors for receiving and transmitting, checking the data buffer padding, checking interrupts for receiving data packets, checking the JUMBO FRAME mode with transmission data up to 8192 bytes.

To test the PCI Express interface, the expansion board connected to the test board was used. During testing the following procedures were carried out: initializing the device on the PCI Express bus, checking the connection with the device, checking the set operating mode and baud rate, searching for connected devices and reading their identifiers.

To test the I2C protocol, we used the EEPROM memory installed on the test board. In the course of testing, the connections with the device were checked, indicating the address and type of the device, the amount of available memory was read, and the procedure for writing the memory with test data was performed, followed by verification by reading the data.

After successfully passing all the tests, the barebox boot loader was installed in the NOR memory of the microcircuit, with the help of which the Linux operating system was installed in the NAND memory of the microcircuit.

Acknowledgement
This research is implemented in Petrozavodsk State University with financial support by the Ministry of Science and Higher Education of Russia within Agreement no. 075-11-2019-088 of 20.12.2019 on the topic “Creating the high-tech production of mobile microprocessor computing modules based on SiP and PoP technology for smart data collection, mining, and interaction with surrounding sources”.

References
[1] Taylor R, Baron D and Schmidt D 2015 The world in 2025 - Predictions for the next ten years
2015 10th International Microsystems, Packaging, Assembly and Circuits Technology Conference, IMPACT 2015 - Proceedings pp 192-5

[2] Mittal S 2014 A survey of techniques for improving energy efficiency in embedded computing systems Int. J. Comput. Aided Eng. Technol. 6 440-59

[3] Mamaluy D and Gao X 2015 The fundamental downscaling limit of field effect transistors Appl. Phys. Lett. 106 193503

[4] John H and Lau P D 2016 3DIC Integration and Packaging (New York: McGraw-Hill Education)

[5] Belmonte A, Oh H, Rassoul N, Donadio G L, Mitard J et al. 2021 Capacitor-less, Long-Retention (>400s) DRAM Cell Paving the Way towards Low-Power and High-Density Monolithic 3D DRAM pp 28.2.1-28.2.4

[6] Kim C, Kim D H, Jeong W, Kim H J, et al. 2018 A 512-Gb 3-b/Cell 64-Stacked WL 3-D-NAND Flash Memory IEEE J. Solid-State Circuits 53 124-33

[7] Chen M F, Chen F C, Chiou W C and Yu D C H 2019 System on integrated chips (SoIC(TM) for 3D heterogeneous integration Proceedings - Electronic Components and Technology Conference pp 594-9

[8] Hsu F C, Lin J, Chen S M, Lin P Y, Fang J, Wang J H and Jeng S P 2018 3D Heterogeneous Integration with Multiple Stacking Fan-Out Package Proceedings - Electronic Components and Technology Conference pp 337-42

[9] Yoshida A, Taniguchi J, Murata K, Kada M, Yamamoto Y, Takagi Y, Notomi T and Fujita A 2006 A study on package stacking process for Package-on-Package (PoP) Proceedings - Electronic Components and Technology Conference pp 825-30

[10] Menon A R, Karajgikar S and Agonafer D 2009 Thermal design optimization of a package on package Annual IEEE Semiconductor Thermal Measurement and Management Symposium pp 329-35