

**SYSTEM AND METHOD FOR HAVING CORRECT-BY-CONSTRUCTION
TIMING CLOSURE FOR FACE-TO-FACE BONDING**

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] The present application claims the benefit of and priority to U.S. Provisional Patent Application No. 63/518,988, filed on August 11, 2023, entitled "INTEGRATED CIRCUIT HAVING MEMORIES AND A SHARED WRITE PORT", identified by Docket Number P23-133-US-PSP, the entire contents of which is incorporated herein by reference in its entirety.

[0002] The present application claims the benefit of and priority to U.S. Provisional Patent Application No. 63/602,733, filed on November 27, 2023, entitled "METHOD AND SYSTEM FOR KNOWN-GOOD-DIE TESTABILITY OF FACE-TO-FACE BONDED CHIPLETS", the entire contents of which is incorporated herein by reference in its entirety.

[0003] The present application claims the benefit of and priority to U.S. Provisional Patent Application No. 63/602,737, filed on November 27, 2023, entitled "SYSTEM AND METHOD FOR HAVING CORRECT-BY-CONSTRUCTION TIMING CLOSURE", the entire contents of which is incorporated herein by reference in its entirety.

[0004] The present application claims the benefit of and priority to U.S. Provisional Patent Application No. 63/567,649, filed on March 20, 2024, entitled "ASSEMBLY HAVING A FACE-TO-FACE BONDED CHIPLET", identified by Docket Number P24-052-US-PSP, the entire contents of which is incorporated herein by reference in its entirety.

[0005] The present application claims the benefit and priority to U.S. Provisional Patent Application No. 63/637,742, filed on April 23, 2024, entitled "INTEGRATED CIRCUIT HAVING MICROVAULT MEMORIES", identified by Docket Number P24-081-US-PSP, the entire contents of which is incorporated herein by reference in its entirety.

[0006] The present application claims the benefit of and priority to U.S. Provisional Patent Application No. 63/637,764, filed on April 23, 2024, entitled "FEFET STRUCTURES ON INTEGRATED CIRCUITS", identified by Docket Number P24-082-US-PSP, the entire contents of which is incorporated herein by reference in its entirety.

[0007] The present application claims the benefit of and priority to U.S. Provisional Patent Application No. 63/674,471, filed on July 23, 2024, entitled "SYSTEM, METHOD, AND APPARATUS FOR WAFER-SCALE MEMORY", identified by Docket Number P24-135-US-PSP, the entire contents of which is incorporated herein by reference in its entirety.

BACKGROUND*Relevant Field*

[0008] The present disclosure relates to integrated circuits. More particularly, the present disclosure relates to systems and methods for having correct-by-construction timing closure.

Description of Related Art

[0009] Chiplets refer to miniature chips that are designed to work as a single entity while using advanced packaging technology. These miniaturized chips are created by dividing the larger chip into several smaller chips, each with its own function or capability. The concept originated from the semiconductor industry's need to overcome the physical restrictions of traditional monolithic chip designs and achieve higher levels of integration. The idea behind chiplets is to create a modular system of interconnected and interchangeable chips that can be combined in different configurations to create advanced computing systems with improved performance, power efficiency, and functionality.

[0010] Chiplets can be based on different architectures, such as CPU, GPU, memory, or IO, and can be assembled and stacked in a variety of ways, depending on the specific application requirements. One of the advantages of the chiplet approach is the ability to mix and match different chiplets from different manufacturers to create custom solutions that meet specific computing needs. This approach also allows for faster time to market, reduced development costs, and increased flexibility, as chiplets can be upgraded or replaced without the need for a complete system redesign.

[0011] The use of chiplets may be used in various industries, including consumer electronics, cloud computing, and data centers, where the demand for high-performance computing and energy efficiency is high. Chiplets are expected to play a significant role in the future of computing and are likely to unlock new possibilities for creating more powerful and/or sophisticated electronic devices.

[0012] SUMMARY

[0013] Described herein is a system for achieving correct-by-construction timing closure, which consists of a first semiconductor device with a thin interface module, an address register that registers an address for address interconnects based on an address clock from an address-clock interconnect, a data register that registers data for data interconnects based on a data clock from a data-clock interconnect, and multiple bonds on the surface of the first

semiconductor device, which are connected to the address and data interconnects and designed to interface with a bonded second semiconductor device.

[0014] This system is designed to achieve correct-by-construction timing closure between the first and second semiconductor devices, as an optional feature. Optionally, the first and second semiconductor devices in this system can be face-to-face bonded together. The first and second semiconductor devices in this system can be stacked together as an optional feature. Another optional feature of this system is that the address-clock interconnect and the data-clock interconnect may be coupled to merge the address clock and the data clock.

[0015] In one embodiment, the data register in the system can be a read data register, which is designed to receive read data through data interconnects. Optionally, this system may also include a write data register that is configured to register write data and provide it to multiple write data interconnects, in which the plurality of bonds also encompasses these write data interconnects. In another embodiment, the data register can be a write data register, designed to provide write data to the data interconnects. This system may additionally comprise at least two read data registers that are designed to receive read data from multiple read data interconnects, where the plurality of bonds also includes these read data interconnects.

[0016] In a further embodiment, the first semiconductor device in the system can have a read-interface register, where the address register and data register together form this read-interface register. This read-interface register is designed to register a read address from read interconnects and read data from these read interconnects based on a read clock from a read-clock interconnect. The read interconnects encompass the address interconnects and data interconnects, while the address clock and data clock form the read clock. Also, the address-clock interconnect and data-clock interconnect form the read-clock interconnect. Additionally, the first semiconductor device may include a write-interface register designed to register a write address and write data from write interconnects according to a write clock from a write-clock interconnect. A plurality of read bonds can be placed on the surface of the first semiconductor device, which are operatively connected to the read interconnects of the read-interface register, and the plurality of bonds also consists of these read bonds. Furthermore, a plurality of write bonds can be placed on the surface of the first semiconductor device, which are operatively connected to the write interconnects of the write-interface register.

[0017] In one embodiment, the system includes a second semiconductor device with a read peripheral, which has read-peripheral interconnects coupled to its surface and is configured to connect to a plurality of read bonds, thereby connecting the read peripheral of the second semiconductor device to the read-interface register of the first semiconductor

device. In another embodiment, the system includes a second semiconductor device with a write peripheral, which has write-peripheral interconnects coupled to its surface and is configured to connect to a plurality of write bonds, thereby connecting the write peripheral of the second semiconductor device to the write-interface register of the first semiconductor device. In yet another embodiment, the read interconnects consist of read-address interconnects and read-data interconnects. Furthermore, the system may have a plurality of read bonds, including read-address bonds disposed on the surface of the first semiconductor device and operatively coupled to the read-address interconnects of the read-interface register, as well as read-data bonds disposed on the surface of the first semiconductor device and operatively coupled to the read-data interconnects of the read-interface register. Additionally, the plurality of bonds can include a read-clock bond, with the read-clock interconnect coupling the read-clock bond to the read-interface register as an optional feature.

[0018] In one embodiment, the system features write interconnects, which consist of write-address interconnects and write-data interconnects. Additionally, the system may include a plurality of write bonds comprising a plurality of write-address bonds and write-data bonds, which are located on the surface of the first semiconductor device and connected to the respective write-address and write-data interconnects of the write-interface register. Optionally, the system's bonds may also include a write-clock bond, with the write-clock interconnect connecting this bond to the write-interface register.

[0019] A method for forming a semiconductor device with correct-by-construction timing closure involves forming a thin interface module on a base die, creating an address register and a data register within the thin interface module, and forming a series of bonds on the surface of the first semiconductor device that connect to the address and data interconnects. These bonds are then bonded to a second semiconductor device. In a potential enhancement to this method, the thin interface module can be situated in close proximity to the bonds, specifically, less than a predetermined distance. This enables correct-by-construction timing closure between the first and second semiconductor devices. In one embodiment, the method involves bonding a plurality of bonds with the second semiconductor device by face-to-face bonding the first semiconductor device to the second semiconductor device. In another embodiment, the bonding process involves stacking the first semiconductor device and the second semiconductor device together. Additionally, the thin interface module can be formed on the base die by creating it under the plurality of bonds. Alternatively, the plurality of bonds can be formed over the thin interface module. Furthermore, the forming acts in the method can be implemented in accordance with a netlist.

[0020] In the method, a plurality of bonds and an address register are formed in a spaced relation to each other such that communication between the bonds and the address register takes less than a predetermined amount of time. Additionally, the address interconnects and the data interconnects are vertically wired to the plurality of bonds. Furthermore, both the first and second semiconductor devices are bonded together. This bonding could optionally be achieved using a face-to-face bonding technique. Lastly, the first and second semiconductor devices can be stacked together in the method.

[0021] In one embodiment, the method includes merging the address-clock interconnect and the data-clock interconnect together, which consequently merges the address clock and the data clock. In another embodiment, the method involves receiving read data via the data interconnects, and the data register used is a read data register. Additionally, this embodiment may optionally include registering write data into a write data register and providing the write data to a plurality of write data interconnects, where the plurality of bonds also includes the write data interconnects. In yet another embodiment, the method comprises providing write data to the data interconnects, with the data register functioning as a write data register. Furthermore, this embodiment may optionally involve receiving first read data in a first read data register from a first memory module on the second semiconductor device and receiving second read data in a second read data register from a second memory module on the second semiconductor device.

[0022] In one embodiment, the method includes a first semiconductor device with a read-interface register, formed by an address register and a data register, configured to register a read address from read-address interconnects and read data from read interconnects in accordance with a read clock from a read-clock interconnect, as well as a write-interface register configured to register a write address and write data from write interconnects in accordance with a write clock from a write-clock interconnect. Additionally, the first semiconductor device has a plurality of read bonds and write bonds disposed on its surface, which are operatively coupled to the read and write interconnects of the read-interface and write-interface registers, respectively.

[0023] In another embodiment, the method further includes a second semiconductor device with a read peripheral having read-peripheral interconnects coupled to its surface and configured to couple to the plurality of read bonds on the first semiconductor device, thereby connecting the read peripheral of the second semiconductor device to the read-interface register of the first semiconductor device. In yet another embodiment, the method also comprises a second semiconductor device containing a write peripheral with write-peripheral interconnects

coupled to its surface and configured to couple to the plurality of write bonds on the first semiconductor device, thereby connecting the write peripheral of the second semiconductor device to the write-interface register of the first semiconductor device. Furthermore, in an alternate embodiment, the read interconnects of the method consist of read-address interconnects and read-data interconnects. Lastly, in another embodiment, the plurality of read bonds consists of read-address bonds and read-data bonds disposed on the surface of the first semiconductor device, which are operatively coupled to the read-address interconnects and read-data interconnects of the read-interface register, respectively. In one embodiment of the method, the plurality of bonds can include a read-clock bond, which allows the read-clock interconnect to couple the read-clock bond to the read-interface register. In another embodiment, the write interconnects in the method can consist of write-address interconnects and write-data interconnects.

[0024] A further embodiment of the method can comprise a plurality of write-address bonds disposed on the surface of the first semiconductor device and coupled to the write-address interconnects of the write-interface register, as well as a plurality of write-data bonds disposed on the surface of the first semiconductor device and coupled to the write-data interconnects of the write-interface register. Additionally, in another embodiment, the plurality of bonds can include a write-clock bond, which enables the write-clock interconnect to couple the write-clock bond to the write-interface register. A method of using the system according to any other embodiments includes providing the address to the address register and clocking the address clock, thereby loading the address register with the address. In one embodiment, a method of using a system can involve providing data to a data register and then clocking the data clock to load the data register with the data.

[0025] A system for achieving correct-by-construction timing closure can comprise a first semiconductor device that provides a thin interface module on a base die, which includes an address registering mechanism for address interconnects in accordance with an address clock from an address-clock interconnect, a data registering mechanism for data interconnects in accordance with a data clock from a data-clock interconnect, and multiple bonding mechanisms for communicating with a second semiconductor device via address and data interconnects.

[0026] BRIEF DESCRIPTION OF THE DRAWINGS

[0027] These and other aspects will become more apparent from the following detailed description of the various embodiments of the present disclosure with reference to the drawings wherein:

[0028] Fig. 1 is a block diagram of an integrated circuit that may be part of a semiconductor device such as a chiplet in accordance with an embodiment of the present disclosure;

[0029] Fig. 2 shows a perspective view of an assembly having the integrated circuit of Fig. 1 implemented on a semiconductor device that is electrically connected to another device to form the assembly in accordance with an embodiment of the present disclosure;

[0030] Fig. 3 shows a block diagram illustrating the memory address space of the integrated circuit of Fig. 1 in accordance with an embodiment of the present disclosure;

[0031] Fig. 4 shows a block diagram illustrating the memory address space with the signal interfaces of the integrated circuit of Fig. 1 in accordance with an embodiment of the present disclosure;

[0032] Fig. 5 shows an illustration of an integrated circuit that may be part of a semiconductor device such as a chiplet in accordance with an embodiment of the present disclosure;

[0033] Fig. 6 shows a perspective of an assembly having the integrated circuit of Fig. 1 implemented on a semiconductor device that is electrically connected to a system-on-a-chip in accordance with an embodiment of the present disclosure;

[0034] Figs. 7A and 7B show a block diagram of a system employing correct-by-construction timing closure for an application chiplet having read and write registers and a memory chiplet in accordance with an embodiment of the present disclosure;

[0035] Fig. 8 shows a block diagram of a system employing correct-by-construction timing closure for an application chiplet having a common read/write address register and a memory chiplet in accordance with an embodiment of the present disclosure; and

[0036] Fig. 9 shows a flow chart diagram of a method for ensuring that correct-by-construction timing closure is achieved in accordance with an embodiment of the present disclosure.

[0037] **DETAILED DESCRIPTION**

[0038] Fig. 1 shows a block diagram of an integrated circuit 100 that may be packaged as a bondable chiplet (e.g., face-to-face chiplet bondable) in accordance with an embodiment of the present disclosure. The integrated circuit (IC) 100 includes a modules group 106 consisting of modules 108, 110, 112, and 114. The IC 100 also features a shared write port 102, configured to write to the modules group 106 using a write peripheral 104. Additionally, it includes read peripherals 116, 118, 120, and 122 and read ports 124, 126, 128, and 130, configured to read from the modules 108, 110, 112, 114.

[0039] The write port 102 may be configured to provide a single write address space for all of the modules group 106 where each of the modules 108, 110, 112, 114 has a dedicated read port 124, 126, 128, 130 respectively. The integrated circuit 100 may be packaged as part of a chiplet configured to be electrically connected to another integrated circuit device (e.g., another chiplet, or IC package, with or without electrical contacts, electrical bumps, etc.). The chiplet may be electrically connected to another device including, for example, by bonding, soldering, wafer-to-wafer bonding, face-to-face chiplet bonding, chiplet-to-wafer bonding, chiplet-to-interposer bonding, and/or may be connected together with an interposer or other interfacing technology. None, one, or more of interposers may be used or other interfacing technologies that are common to heterogeneous 3D system-in-package solutions may be utilized in electrically connecting a chiplet to another device.

[0040] Each read port (124, 126, 128, 130) in the chiplet may feature electrical contacts on a side of the chiplet or on multiple sides of the chiplet. The read ports 124, 126, 128, 130 may use multi-cycle pipelined circuitry. Upon bonding to another device (e.g., wafer, chiplet, chip, SOC, package, FPGA, etc.), the electrical contacts may line up in a manner that provides dedicated access to specific modules of the modules 108, 110, 112, 114. For instance, a processing/computing element may have exclusive access to module 108 via the read port 124, which may contain the neural network weights in a register file. The computing/processing elements of the bonded device may be an SoC or FPGA. Or, for example, may be as shown in Fig 2. as an example. Similarly, a different processing/computing element may have exclusive read access to module 110 via the read port 126, which includes a different register file. In this specific embodiment, this arrangement of the electrical contacts ensures that each computing/processing element has the dedicated access it needs to carry out its specific computation efficiently thereby providing a compact, modular, and scalable system that allows different processing elements to maintain dedicated access to specific modules 108, 110, 112, 114. Without dedicated access, different processing elements might have to queue up to use the same resource which would slow down overall processing speed. By providing dedicated access, the proposed chiplet ensures that each processing element can operate at its maximum capability without interference from other computing elements in this specific embodiment.

[0041] The write peripheral 104 is a peripheral circuitry responsible for processing and writing data into the memory cells found within the modules 108, 110, 112, 114. The write peripheral 104 may include dedicated contacts so that a chip electrically connected (e.g., bonded) to a chiplet of the integrated circuit, such that the write port 102 is accessible via a shared write logic system that involves utilizing a shift register-based, different voltage design,

preferably high voltage design, that has a shared write address and data components. This shared write logic system is designed to be accessed via a bonded chip, another bonded chiplet, and/or via other circuitry in the same package as the integrated circuit 100. A shift register could allow the system to move data through a series of stages, with each subsequent stage receiving the data from the previous stage. By utilizing a shift register, the system can increase the data throughput while maintaining a low rate of data transfers. The shared write address space refers to the location where data is written in the chiplet.

[0042] In another embodiment, an interlock 132 may disable the read ports 124, 126, 128, 130 while data is being written to the modules group 106 via the write port 102. Likewise, the interlock 132 may disable the write port 102 when read operations are being carried out on the read ports 124, 126, 128, 130. The written data can later be accessed concurrently by all processing elements that need to read the data via a respective one of the read ports 124, 126, 128, 130. This ensures that all processing elements have the most commonly used data available to them without regard to other reads being concurrently carried out by other processing elements.

[0043] The write peripheral 104 circuit includes a write driver. This unit receives the data to be written and converts it into suitable signals that can change the state of the memory cells. Depending on the type of memory technology used, these signals could involve voltage levels, current pulses, or other types of energy. The shared write logic system may be high voltage due to the specific voltage requirements of the chiplet. The write driver must provide enough power to reliably change the state of the memory cells, but it must also operate within suitable parameters to avoid causing damage or unnecessary wear.

[0044] The write peripheral 104 circuit may also feature a data buffer or write buffer. This component temporarily stores the data to be written, allowing the write operation to be performed at an optimal pace. By balancing the speed of incoming data with the speed at which the memory cells can be written, the write buffer helps prevent data loss and optimizes system performance.

[0045] The write peripheral 104 may also include, in some embodiment, a write control unit that orchestrates the sequence of operations in the write process. It generates control signals to activate the write driver at the appropriate times, controls the flow of data from the write buffer, and coordinates the timing of the write operations. By synchronizing these various activities, the write control unit ensures efficient and reliable write operations.

[0046] The write peripheral 104 may also include data encoding mechanisms to improve reliability and data integrity. For example, before the data is written to the memory

cells, these mechanisms encode it in a way that allows potential errors to be detected, and in some cases, corrected when the data is later read. This can be helpful in systems where data integrity has a higher priority, such as in servers or scientific research devices.

[0047] The write peripheral 104 may also include a timing unit that serves as the system's heartbeat, supplying clock signals that synchronize the operation of the system's various components. In some systems, it may include components like oscillators, clock generators, or phase-locked loops. The timing unit may ensure that all operations occur at the suitable time relative to each other.

[0048] The IC 100 may be implemented as a face-to-face bonded chiplet, with modules 108, 110, 112, and 114 formed from a non-volatile memory. In some specific embodiments, the IC 100 may also feature a dynamic allocation circuitry to allocate memory blocks to the modules group 106 based on the usage of the modules group 106 (e.g., each module 108 may include dynamic allocation circuitry for dynamically allocating a range of read locations for a respective processing element).

[0049] The IC 100 may feature a plurality of clocks in some specific embodiments, with each clock of the plurality of clocks feeding a respective module of the plurality of modules, providing each respective module with decoupled timing relative to the other modules of the plurality of modules. The modules group 106 may be arranged in any topology known to one of ordinary skill in the relevant art. Bit-cell density can be up to 10 times more dense than embedded SRAM cells in the modules group 106.

[0050] The IC 100 may be formed on a chiplet that includes a first side and a second side, with the second side configured for bonding to a second semiconductor device. The IC 100 may include a high voltage write logic adjacent to the first side of the chiplet. A decoder circuitry, a driver circuitry, and a register circuitry may be formed on the silicon substrate portion of the chiplet, while the modules group 106 may be formed on a second layer portion of the chiplet. Stacked structures, CMOS over and under arrays, or other structures known to one of ordinary skill in the relevant art may be utilized. The second semiconductor device may comprise a plurality of processing elements. Each processing element includes a respective interface to communicate with a respective module of the plurality of modules on the modules group 106 when the second semiconductor device is bonded to the chiplet.

[0051] The silicon substrate traditionally serves as the initial stage of IC fabrication, focusing on the creation of active components, particularly transistors. Techniques like diffusion, ion implantation, oxidation, and material deposition are employed to fashion the intricate structures of transistors. These processes operate at small scales. The application of

photolithography, etching, and implantation techniques enables the definition of transistor structures with precision. The silicon substrate's significance lies in its ability to establish the fundamental building blocks necessary for signal processing, amplification, and control within the IC. This layer is sometimes called Front-End-Of-The-Line ("FEOL"). However, in some embodiments, the modules group 106 isn't implemented in this first layer portion of a chiplet.

[0052] Next in the manufacturing process, a second layer may be added that traditionally takes on the role of interconnect fabrication, facilitating the electrical connections between various IC components. This phase traditionally focused on the creation of passive components, including interconnects, vias, and metal-insulator-metal (MIM) capacitors. The second layer processes typically differ from the processes used on the silicon substrate in terms of precision and scale. The interconnects are formed by depositing and patterning metal layers, typically aluminum or copper, to construct the wiring network. Dielectric layers, such as silicon dioxide or low-k dielectrics, are introduced to insulate the interconnects and prevent signal interference between different wiring layers. The second layer's traditional function is to establish the necessary interconnections that enable the routing and distribution of electrical signals throughout the IC. However, as described herein, circuitry may be utilized within this second layer (sometimes referred to as Back-End-Of-The-Line ("BEOL")). The second portion of the chiplet may be used for implementing modules group 106, in some specific embodiments.

[0053] Alternate embodiments of the IC 100 may be implemented as a stacked die, a monolithic design, TSVs, or silicon through vias. In a stacked die design, several dies may be stacked on top of each other, with each die performing different functions, such as memory and processing. The stacked die may communicate through wire bonds, microbumps, or bump-less bonds. In a monolithic design, the various functions and modules of the IC 100 may be integrated onto a single die, forming a more compact and power-efficient design. Alternate embodiments of the IC 100 may be implemented as a stacked die or a monolithic design. The stacked dies may fused together through bonding of microbumps or bump-less bonds. Stacks of more than two dies may use silicon through vias (TSVs), in specific embodiments.

[0054] Additionally, the IC 100 may include one or more interlocks 132 to prevent conflicts in reading and writing data. The modules group 106 may be formed from a variety of non-volatile or semi-volatile (e.g., very long refresh periods) memory technologies, such as Static Random-Access Memory (SRAM), Ferroelectric Field Effect Transistor (FeFET), Ferroelectric Random Access Memory (FeRAM), Resistive Random Access Memory

(ReRAM), Spin-Orbit Torque (SOT) Memory, Spin Transfer Torque (STT) Memory, charge trap, floating gate memories, and/or Schottky diodes.

[0055] The modules group 106 may utilize a Static Random-Access Memory (SRAM) topology. The SRAM topology may employ a cross-coupled flip-flop structure (e.g., latching flip-flops), ensuring the stored data remains intact as long as power is supplied. Thus, in some specific embodiments, the modules group 106 may utilize heterogeneous types of memory including volatile and non-volatile memory types.

[0056] The modules group 106 may utilize a Flash Memory topology. The Flash memory is a non-volatile memory technology used in applications where data persistence is needed, such as solid-state drives (SSDs) and USB flash drives. The flash memory topology disclosed herein features a matrix of memory cells, each consisting of a floating-gate transistor or charge trap device. The modules group 106 may also use wear-leveling techniques to prolong the lifespan of the memory cells. Additionally or alternatively, conventional DRAM memory may be used.

[0057] The modules group 106 may utilize a Ferroelectric Random-Access Memory (FeRAM) topology. The FeRAM topology utilizes a ferroelectric material capable of retaining polarization states. One such memory topology may, in specific embodiments, utilize a FeFET to retain state information and program the ferroelectric material. These ferroelectric materials may be used to retain state information and act as a memory bit cell.

[0058] The modules group 106 may utilize a Phase Change Memory (PCM) topology, which is a non-volatile memory technology that utilizes reversible phase changes in materials to store data. The PCM topology may include any phase change material, for example a chalcogenide alloy or a chalcogenide glass housed within a memory cell.

[0059] The modules group 106 may utilize a Resistive Random-Access Memory (ReRAM) topology, which is a non-volatile memory technology based on resistive switching phenomena. The ReRAM topology may utilize a thin-film material that exhibits reversible changes in resistance upon the application of electrical stimuli.

[0060] The modules group 106 may utilize a Spin-Orbit Torque (SOT) Magnetic Random-Access Memory topology. SOT-MRAM is a type of non-volatile memory that utilizes spin-orbit torque to switch the magnetic state of a storage element. The SOT-MRAM topology may incorporate a magnetic tunnel junction (MTJ) structure and leverages the spin-orbit coupling effect to write and read data. The magnetic tunnel junction may have a dielectric layer between a magnetic fixed layer and a magnetic free layer. Writing may be done by switching magnetization of the free magnetic layer by injecting an in-plane current in an

adjacent SOT layer. Reading may be done by putting current into the magnetic tunnel junction. The SOT-MRAM can optimize the spin-orbit materials by using current-driven switching schemes while minimizing write energy consumption, in some specific embodiments.

[0061] The modules group 106 may utilize a Spin Transfer Torque (STT) Magnetic Random-Access Memory topology. The STT-MRAM is another type of non-volatile memory that relies on spin transfer torque to manipulate the magnetic state of a storage element. The STT-MRAM topology can use a magnetic tunnel junction (MTJ) structure, where the magnetization orientation determines the stored data. Additionally, the orientation of a magnetic layer in a magnetic tunnel junction or spin valve can be changed using a spin-polarized current, for example.

[0062] The IC 100 may include a single write peripheral 104 with a dedicated clock, or each module 108, 110, 112, 114 may have its own dedicated write peripheral utilizing a shared clock (not shown in Fig. 1). Additionally, the modules group 106 may be organized into separate partitions, each with a dedicated read peripheral 116, 118, 120, 122 having an independent clock.

[0063] Another possible embodiment of the IC 100 includes an interface (e.g., the same, different, higher or lower voltage) to enable data transfer external to the packaging of the IC 100. The IC 100 may also include an integrated microcontroller unit (MCU) or a digital signal processor (DSP) for processing data within the IC in yet additional specific embodiments.

[0064] Fig. 2 shows a perspective view of an assembly 200 of the integrated circuit 212 of Fig. 1 implemented on a chiplet 230 that is bonded to a second device 226 in accordance with an embodiment of the present disclosure. The integrated circuit 212 is the circuitry within the chiplet 230. The second device 226 may be a chiplet, semiconductor wafer, semiconductor package, encased circuitry, etc. For example, the second device 226 may be an AI accelerator such that each processing unit has read access to one module (or a predetermined set) of the modules group 236. In yet another embodiment, the second device 226 may be a network controller where there is an offload circuit to read the data from each of the modules to process incoming/outgoing packets, etc. The assembly 200 includes a modules group 236 having a plurality of modules, including a first module 232 and a second module 234. Fig. 2 shows several modules, however, for clarity, only modules 232, 234 have reference numbers. The integrated circuit 212 further comprises a shared write port 222. The shared write port 222 interfaces into the write peripheral 202.

[0065] Although the second device 226 may use the shared write port 222 via an address & data bus with a clock and an enable signal to write data to any modules within the modules group 236, other ways of writing data may be considered. For example, serial connections, parallel connections, various buses, or ports, may be used, such as a DDR (Double Data Rate) Interface, an SRAM (Static Random-Access Memory) Interface, a NAND Flash Memory Interface, a NOR Flash Memory Interface, an HBM (High Bandwidth Memory) Interface, a GDDR (Graphics Double Data Rate) Interface, an NVMe (Non-Volatile Memory Express) Interface, SPI, I2C, etc. Each of the modules has a read port with a read address 218 (to send an address to a module 234) and read 220 (which is the data read from the module 232).

[0066] The modules group 236 is formed on a chiplet 230 having two sides including a surface 228 that can be bonded to and complement a second device 226. The chiplet 230 may be formed by forming circuitry on a silicon substrate 204 and then by adding a second layer 206. In other embodiments, these layers may be reversed and/or other layers may be added, removed, etc. The read address 218 and read data 220 are used for reading the module 232.

[0067] Although the second device 226 may use an address & data bus with a clock and a enable signal to read data from the module 232, other ways of reading data may be considered. For example, serial connections, parallel connections, various buses, or ports, may be used, such as a DDR (Double Data Rate) Interface, a SRAM (Static Random-Access Memory) Interface, a NAND Flash Memory Interface, a NOR Flash Memory Interface, a HBM (High Bandwidth Memory) Interface, a GDDR (Graphics Double Data Rate) Interface, a NVMe (Non-Volatile Memory Express) Interface, SPI, IC2, etc.

[0068] All of the read ports (e.g., 218 and 222) are configured to be inactive when a write operation is applied to the shared write port 222. The read ports may also be configured to process reads concurrently with each other. The shared write port 222 is configured to write to an address space, where the shared write port 222 is configured to write to the first module 232 via a first portion of the address space and write to the second module 234 via a second portion of the address space. Each module of the plurality of modules 236 includes an independent read port for concurrent reading via a respective independent read port of any of the plurality of modules.

[0069] Each read port for a respective module may include contacts for circuitry found within the second device 226 to interface via metallic contacts. Thus, there may be metallic contacts on the top layer 208 that are configured to interface with metallic contacts on the

surface 228 of the chiplet 230 such that the metallic contacts allow for a read space that is coextensive with a read space of a module of the modules 236. The read spaces of the modules group 236 may all be coextensive with each other (as is described with reference to Figs. 3 and 4).

[0070] In one embodiment, the read peripheral for the first module 232 is implemented on a silicon substrate 204 (sometimes referred to as a Front-end-of-the-line). The second layer 206 (sometimes call the Back-end-of-the-line) may be built next in the manufacturing process on top of the silicon substrate 204 (and any circuitry) and may contain the respective memory bit cells. In an alternative embodiment, the read peripheral for the first module 232 is implemented in the second layer 206 and is disposed between the modules group 236 and the surface 228 of the chiplet 230.

[0071] The modules group 236 may be configured to process write commands only during reset. The write commands may be “slow write” commands. That is, the modules group 236 may have very low write speeds relative to its read speed. The write logic may be frozen (or disabled) when the modules group 236 are used for reading data. In some specific embodiments, the integrated circuit 212 provides functionality to allocate memory blocks to the modules group 236 based on the usage of the modules group 236. In other embodiments, the memory addresses are fixed along with the allocation. The integrated circuit 212 may be implemented as a face-to-face bonded chiplet 230. The face-to face bonding may be bump-less wafer bonding.

[0072] The modules group 236 can have a single write peripheral 202. In other embodiments, each module of the modules group 236 may have a dedicated write peripheral that utilizes a shared clock. In yet other embodiments, the modules group 236 may also be organized into separate partitions each with partition having a dedicated read peripheral, where each dedicated read peripheral has an independent clock. The partitions may be one, two, or more modules of the modules group 236.

[0073] The write peripheral 202 circuitry's overall architecture may include a series of different components, including write driver, address decoders, sense amplifiers, data input latches, data bus, etc. and/or some combination thereof. Write drivers or write buffers, may be tasked with transferring data onto the memory cell. They may enhance the input signal to achieve a level appropriate for the memory cell. Address decoders may be used to interpret the memory address that is fed as an input where the data needs to be written. By activating the specific row and column of the memory array linked to that address, they may be used to select the target memory cell. Sense amplifiers may be used to identify and boost the signal from the

memory cells during read operations, also participate in refreshing the memory cell post data write in write operations. The write operation is instigated by a write enable signal. When a write command is initiated, this signal propels the write drivers and decoders into the writing process. Data input latches may be used as temporary storage units, retaining the data set to be written into the memory until the write operation is implemented. A data bus with a transmission route, can be used to facilitate the movement of data from the data input latches to the memory cells.

[0074] A write operation to the modules group may be performed through a priority arbitration circuit that facilitates the modules to be accessed in a predetermined order, and the shared write port 222 may be configured to write to a virtual address space that is mapped onto a physical memory space. The integrated circuit 212 may include a high voltage write logic used within the write peripheral 202, and the second semiconductor device 226 may comprise a plurality of processing elements, whereby each processing element includes a respective interface to communicate with a respective module of the modules group 236. Furthermore, the chiplet 230 may include an interface to the shared write port 222 on the second side to thereby interface with a complementary interface on the second semiconductor device 226.

[0075] The integrated circuit 212 may also include a power gating circuitry that selectively powers down a module of the modules 236 when not in use. Additionally, the integrated circuit 212 may have a write peripheral 202 of the modules group 236 connected to a dedicated I/O pad to enable data transfer external to the package of the integrated circuit.

[0076] The integrated circuit 212 may utilize multiple modules of the modules group 236 grouped together. These modules may be synchronized with one another in specific embodiments. In some cases, all the modules are synchronized, while in other instances, only specific modules are to be synchronized. For instance, the circuit on a second device 226 may be synchronized with a specific module when reading data from one of the modules in the module group 236.

[0077] To synchronize the modules, the integrated circuit 212 may use various timing technologies. In some cases, a plurality of clocks may feed each respective module of the modules group 236, thereby allowing each module to have decoupled timing relative to the other modules in the group. This decoupling ensures that any delay in one module will not affect the functioning of other modules. It is worth noting that the clocks used may or may not need to be synchronized. In some cases, a common clock can be used to synchronize the modules. In yet other embodiments, the clock signal or signals may be provided by the second device 226.

[0078] In alternative embodiments, other synchronization techniques can be used, such as phase comparison of the clock signals or a phase-locked loop (PLL) synchronization method. Another embodiment for synchronizing the modules in the IC could use delay-locked loop (DLL) synchronization. In this method, a delay element is added to the clock signal path, and the output is compared to the input clock signal. The feedback loop adjusts the delay element until the output of the DLL matches the input, resulting in synchronization of the clock signals.

[0079] In another embodiment, the integrated circuit 212 could use a combination of different synchronization techniques to achieve synchronization between the modules. For example, some modules may use PLL synchronization while others use clock delay lines or DLL synchronization, depending on their specific design parameters. Additionally, the integrated circuit 212 can also use redundant synchronization techniques to ensure reliability and redundancy in case one method fails. For example, the integrated circuit 212 could use both PLL synchronization and DLL synchronization simultaneously, so that if one method fails, the other can still maintain synchronization.

[0080] Fig. 3 shows a block diagram 300 illustrating the memory address space of the integrated circuit of Fig. 1 in accordance with an embodiment of the present disclosure. The memory address space includes a write address space 316 and read data address spaces 310, 312, 314.

[0081] The write address space 316 consists of various units where data, e.g., weights, and/or instructions can be stored. These units are referred to as memory addresses. The module group 302 includes multiple memory modules 304, 306, 308. The write address space 316 may be distributed among the memory modules 304, 306, 308 such that the write address space 316 spans from 0 to $N*M-1$. As shown in Fig. 3, the modules group 302 has N memory modules 304, 306, 308, where N is a positive integer, and each module has a memory size of M . The total number of unique write memory addresses in the write address space will be $N*M$, which can be referenced by an integer from 0 to $N*M-1$.

[0082] Starting at 0, memory addresses of the write address space 316 are ordered sequentially up to $N*M-1$. In other words, the first address is 0 and the final address is $N*M-1$, encompassing a total of $N*M$ addresses. This ordering can be linear (each address increases by one) or some other specified pattern depending.

[0083] The write memory addressing can be implemented in a variety of ways based on the system architecture. One method used in a specific embodiment is to use the base and limit registers. The base register holds the smallest legal physical write memory address, and

the limit register specifies the size of the range. Therefore, to generate a logical address, you would add the base to the relative address. In other embodiments, a memory addressing scheme may be used where the base used is set to be 0. Yet additional write addressing techniques will be appreciated by one of ordinary skill in the relevant art.

[0084] For any device that writes to the modules group 302, each memory module can possess a unique set of write memory addresses such all memory addresses within the modules group 302 is unique with respect to writing data. e.g., the first module starting at 0 and the last one ending at $N*M-1$. This allocation, in some embodiments, may be dependent on the memory management system of the device writing data to the modules 304, 306, 308, which could range from simple fixed partitioning schemes to more complex dynamic partitioning models.

[0085] For instance, in a straightforward linear model where each module (304, 306, or 308) has an equal size of M addresses, the first module 304 would possess write addresses 0 to $M-1$, the second module would have write addresses M to $2M-1$, the third module would have write addresses $2M$ to $3*M-1$, and so forth. The N th module 308, therefore, would possess write addresses from $(N-1)*M$ to $N*M-1$.

[0086] It is contemplated that one of ordinary skill in the relevant art may use other implementations of write memory addresses from 0 to $N*M-1$ that depends on various factors such as the hardware architecture, operating system, memory management schemes, and the nature of the programs being run on the system, etc.

[0087] The modules group 302 has different read data address spaces 310, 312, 314. These read address spaces 310, 312, 314 may have overlapping addresses spaces, may have contiguous address spaces, or may have coextensive address spaces. The read address spaces 310, 312, 314 may be independent relative to each other. The system includes three independent read address spaces, labeled as read address spaces 310, 312, and 314. Each of these read address spaces is distinct from the others, meaning that reads can be performed in each space without affecting the others.

[0088] The read address spaces 310, 312, 314 may be defined as contiguous blocks of memory addresses, each with its own starting address and ending address. In modules group 302, each read address space 310, 312, 314 may have a range of addresses that corresponds to values from 0 to $M-1$, where M is a maximum value determined by the size of the modules 304, 306, 308 being used. Additionally, write address space 316 may be defined with a range of addresses from 0 to $N*M-1$, where N is the total number of modules in the modules group 302.

[0089] In one embodiment, allowing one processing unit to interface with each read address space 310, 312, 314, the concurrent reads may be implemented as described herein. The independence of the read address spaces 310, 312, 314 ensures that each processing unit can access its desired data without causing any interference or conflict with other processing units.

[0090] Fig. 4 shows a block diagram illustrating the memory address space with the signal interfaces of the integrated circuit of Fig. 1 in accordance with an embodiment of the present disclosure. The signals used in Fig. 4 may be used with any embodiment described herein. However, one of ordinary skill in the relevant art will appreciate that different signaling schemes may be used.

[0091] The modules group 402 includes modules 404, 406, 408 that share a common write peripheral 411. The write peripheral 411 includes a write address bus that includes the address of the data being written, a write data bus that includes the data, a write clock cause the writes to occur (e.g., either on a leading or trailing edge of the clock signal, etc.). The writes only occur if the write enable signal indicates a write should occur. Any logic may be used, e.g., high voltage may correspond to 1 and a low voltage may correspond to 0, or vice versa. In some embodiments, the write peripheral 411 may be on the chiplet 230 and in other embodiments, the write peripheral 411 is on the second device 226.

[0092] The modules group 402 has modules 404, 406, 408 where each has a respective read peripheral 410, 412, 414. Each of the read peripheral 410, 412, 414 has a read address bus to send an address for reading, a read data bus to receive the data, a read clock which is the clock used to control the timing of the output of the digital data, and an output enable that is a precondition to outputting data. Any logic may be used, e.g., high voltage may correspond to 1 and a low voltage may correspond to 0, or vice versa. In yet additional embodiments, multi-bit or analog data storage may be used. In some embodiments, one or more of the read peripherals 410, 412, 414 may be on the chiplet 230 and in other embodiments, one or more of the read peripherals 410, 412, 414 are on the second device 226.

[0093] Fig. 5 shows an illustration of an integrated circuit 500 that may be part of a semiconductor device such as a chiplet in accordance with an embodiment of the present disclosure. The integrated circuit 500 may be disposed on a semiconductor device, such as a chiplet, that has a silicon substrate 506 and a second layer portion 508. Within the integrated circuit 500, there may be an array section forming the module 502 where a three-dimensional column array of memory bit cells 522 has the necessary components to store memory in non-volatile, semi-volatile memory, or a memory format as described herein.

[0094] The integrated circuit 500 may include a modules group having a plurality of modules including a first module and a second module, etc. even though only a single module 502 is shown. The memory bit cells 522 are written to by the shared write port 512, 516, which includes both a write address bus line 512 and a write data bus 516. These buses run through the second layer 508 and can be connected to a second semiconductor device via an interposer. The second device has electrical contacts that complement those on the surface 518, allowing it to be electrically coupled to the write address and data buses. The memory bit cells 522 can be read from via the read port 524, 526, which includes a read address bus line 524 and a read data bus 526. Both of these buses can also run through the second layer 508 to the second semiconductor device coupled to the surface 518, which also has complementary electrical contacts to allow it to be electrically coupled to the read address and data buses.

[0095] Various kinds of memory technologies may be used for the memory bit cells 522, such as a vertical connectivity fabric structure formed from non-volatile memory unit cells arranged in a three-dimensional column array 522. The memory bit cells 522 may utilize one or more of a cross-point, 3D NANDs, 3D NORs, 3D ANDs, and/or a stacked planar layer.

[0096] In some embodiments, the integrated circuit 500 is electrically connected to a second semiconductor device (not shown in Fig. 5) comprising another integrated circuit, which may be a system-on-chip or a Field-Programmable-Gate-Array. In some embodiments, the memory bit cells 522 may be formed from various non-volatile memory types, such as FeFET, FeRAM, ReRAM, SOT, or STT. Additionally, alternatively, or optionally, the memory bit cells may be formed from non-volatile memory unit cells having 2-terminal devices, 3-terminal devices, or 4-terminal devices. The memory bit cells 522 can also be formed from volatile, non-volatile, or semi-volatile memory, such as SRAM.

[0097] For example, the memory unit bit cells 522 may be formed from ferroelectric materials, such as a ferroelectric tunnel junction, a diode, a capacitor, a single-gate transistor, or a dual-gate transistor. Alternatively, the memory unit bit cells 522 may be formed from memristive materials, such as at least one ReRAM, or magnetic materials, such as at least one spin-orbit-torque device or at least one spin-transfer-torque device. Moreover, the non-volatile memory unit cells 522 may also be formed from phase-change materials or anti-ferroelectric materials.

[0098] In some alternative embodiments, the non-volatile memory unit cells 522 can be formed from other types of materials, such as phase change materials, anti-ferroelectric materials, or multi-bit PCM materials. The non-volatile unit cells can be formed utilizing different structures, such as resistive random-access memory (RRAM) technology, magnetic

random-access memory (MRAM) technology, or ferroelectric random-access memory (FRAM) technology.

[0099] Moreover, in some implementations, 3D NAND technology may be utilized to form the memory unit bit cells 522. For example, the memory unit bit cells 522 may be formed from stacked memory layers where each layer includes a plurality of memory cells that can be accessed using shared bit lines. In such a case, the read port 524, 526 may be coupled to the bit lines, and the write port 512, 516 may be coupled to the word lines that control the access to each layer.

[00100] In another embodiment, the 3D connectivity fabric structure can be built with stacked layers of either NAND gates, NOR gates, or AND gates, and in some cases, different types of logic gates may be combined to optimize the structure's functionality. In addition, the 3D connectivity fabric structure may be formed utilizing through-silicon-via (TSV) technology, which allows the vertical interconnection of the different layers of the structure. This may be utilized to stack two or more chiplets together including those disclosed herein.

[00101] Additionally, the non-volatile memory unit cells may include 2-terminal devices, such as a capacitive or a memristive device with or without an additional selector device such as a diode in series, 3-terminal devices, such as a floating-gate transistor, a transistor with an access gate, or 4-terminal devices, such as a transistor with two access gates. The type and configuration of the non-volatile memory unit cells 522 may depend on the specific application requirements, including the speed, power consumption, and reliability of the circuit. The memory unit cell may include or be a single ferroelectric transistor, 6T SRAM cell, or any other SRAM cell variant known to one of ordinary skill in the relevant art. The memory unit cell may be a combination of many different devices, including, but not limited to, one or more of a transistor, a memristor, a capacitor, etc.

[00102] In some embodiments of the present disclosure, a ferroelectric material can be utilized to form the non-volatile memory unit cells 522. The ferroelectric material may be implemented as any kind of device, including, but not limited to, a thin-film device, such as a ferroelectric tunnel junction, a capacitor, a single-gate transistor, or dual-gate transistors, etc.

[00103] In another embodiment, the non-volatile memory unit cells 522 may be formed from a memristive material, such as a Metal Oxide Memristor (MOM), Conductive-Bridging RAM (CBRAM), or valence change memory (VCM), each of which provides different benefits regarding power consumption, speed, endurance, etc.

[00104] Moreover, in some embodiments, the non-volatile memory unit cells 522 may be formed from a magnetic material, such as spin-orbit-torque (SOT) devices, spin-transfer-torque (STT) devices, or perpendicular magnetic tunnel junctions (p-MTJ).

[00105] In one embodiment, the modules group may include many modules where each of which can be accessed through dedicated read ports 524, 526 with a dedicated read peripheral 520 while sharing the same write port 512, 516 and shared write peripheral 510. The shared write port 512, 516 can be configured to selectively write to one or more of the plurality of modules within the modules group including the memory bit cells 522. Each of the modules may have the same or different sizes, and different module sizes may be configured to optimize the utilization of the memory array with different operating scenarios, etc.

[00106] Furthermore, the integrated circuit 500 may be formed utilizing different manufacturing processes and techniques, which include but not limited to, a CMOS or Bipolar-CMOS-DMOS (BCD) process, a silicon-on-insulator (SOI) process, a FinFET process, a silicon germanium (SiGe) process, a gallium arsenide (GaAs) process, etc.

[00107] Fig. 6 shows a perspective of an assembly 600 having the integrated circuit of Fig. 1 implemented on a semiconductor device, such as the chiplet 230, that is electrically connected to a system-on-a-chip ("SOC") 610 in accordance with an embodiment of the present disclosure. The semiconductor device, in this embodiment, is the chiplet 230 that is electrically connected to a system-on-a-chip ("SOC") 610.

[00108] Referring to Fig. 6, the SOC 610 includes a silicon substrate 602 on which a plurality of processing elements is formed, including a processing element 606. The processing elements can communicate with each other through a Network-on-Chip ("NOC") 604, which is a communication fabric that directs data transfer between the processing elements. The communication fabric can take various forms, including buses, switches, NOCs, etc. The NOC 604 in the SOC 610 directs data traffic between the various nodes (e.g., the processing element 606) and links, which provide the communication paths between the nodes.

[00109] The plurality of processing elements including the processing element 606 processing elements can be any suitable type of processors capable of executing instructions, including microprocessors, graphics processing units (GPUs), digital signal processors (DSPs), or application-specific integrated circuits (ASICs).

[00110] Additionally, the SOC 610 may comprise various modules, such as module 232, which are grouped together to provide memory functionality to the assembly 600 as described here. The modules in modules group 236 can be coupled to a respective processing element provide it readable memory. In some embodiments, the coupling between the module (e.g.,

module 232) and the processing element (e.g., 606) can be achieved through interconnects on the silicon substrate 602.

[00111] After the circuitry is formed on the silicon substrate 602, a second layer 608 can be disposed on top of the substrate. The second layer 608 can be any suitable material, such as an insulating material, a metal, a dielectric, or interconnect layer, and it may be bonded to the chiplet 230, e.g., to establish an electrical connection between the silicon substrate 602 and the chiplet 230. The bonding can be done using any suitable technique, including but not limited to, adhesives, soldering, or welding, etc.

[00112] In general, the assembly 600 provides a means of integrating the chiplet 230, which can include the integrated circuit of Fig. 1, with the SOC 610. Integrating the chiplet 230 provides various advantages, such as enhanced functionality, higher performance, and lower power consumption. Moreover, the integration of the chiplet 230 with the SOC 610 can be accomplished in various ways, depending on the particular application and design objectives of the system.

[00113] The assembly 600 can incorporate various variations and modifications, depending on the specific requirements of the system. For example, the processing elements 606 formed on the silicon substrate 602 can vary in their number, type, and arrangement. Similarly, the modules in modules group 236 can vary in their number, type, and function.

[00114] Furthermore, the second layer 608 can be modified to include additional functionality. For instance, the second layer 608 can include passive components, such as resistors, capacitors, and inductors, or active components, such as transistors or diodes. Incorporating these components in the second layer 608 can further enhance the functionality and performance of the system.

[00115] In another variation, the assembly 600 can incorporate a heterogeneous integration approach, where the chiplet 230 is fabricated using a different technology than that used for the SOC 610. This approach allows for the optimal use of different fabrication technologies for different parts of the system, resulting in improved performance and reduced power consumption.

[00116] Figs. 7A and 7B show a block diagram of a system 700 employing correct-by-construction timing closure for an application chiplet 702 with an interface circuit 701 having read registers 726a, 726b and a write register 728, and a memory chiplet 704 in accordance with an embodiment of the present disclosure. Thus, the system 700 may include two separate semiconductor devices, e.g., the application chiplet 702 and the memory chiplet 704 that may have been formed on two separate dies. The interface circuit 701 (which can be referred to as

interface module) can include registers 726a, 726b, 728 to allow the application chiplet 702 to have a correct-by-construction timing closure with the group of memories 706 on the memory chiplet 704.

[00117] The memory chiplet 704 may include a group of modules 706 having modules 708a, 708b and 708c that are independently accessible via read peripherals 722a, 722b, and 722c, respectively. The memory chiplet 704 may also include a shared write peripheral 710 where data may be written to a memory location within the group of modules 706. The interconnects found within the memory chiplet 704 may be coupled to the surface via bonds to facilitate communication between the semiconductor devices 702, 704, for example when they are bonded together in a stack configuration in one specific embodiment. A designer of the application chiplet 702 may place the interface module 701 using a netlist on a location within a predetermined distance from a surface of the application chiplet 702. That is, the signal time and characteristic may be predefined to work the memory of the memory chiplet 704. For example, the travel time of a signal from a read register 726 may be less than a predetermined time.

[00118] The interface module 701 may include one or more read registers 726a, 726b and a write register 728 that may communicate with the memory chiplet 704 via several bonds as shown in Fig. 7B.

[00119] The read register 726a can interface and communicate with other circuitry within the application chiplet 702 via a read application programming interface (“API”) 714a. That is, the read API 714a may be a bus where data may be requested by external circuitry (e.g., via a CPU) to provide data from the memory chiplet 704 to the other circuitry (e.g., the exemplary CPU).

[00120] The read register 726a may be controlled via a read clock that is received via a read-clock interconnect 712a. The read-clock interconnect 712a may also be coupled to a read-clock bond 738a which can be connected to the read peripheral 722a on the memory chiplet 704 to provide clocking for the memory contained therein. The interconnects may be connected together by the read-clock bond 738a on the surface of the application chiplet 702 and a respective bond (not explicitly shown in Fig. 7B) on the memory chiplet 704.

[00121] The read register 726a may be coupled to a read-address interconnect 716a (which may include multiple parallel connections) that communicates a read address that has been loaded into the read register 726a. The read-address interconnect 716a is coupled to a plurality of read-address bonds 740a so that the read address can be received by the memory

chiplet 704. The read address is a value that the read peripheral 722a can translate to query a location within the module 708a.

[00122] The read register 726a also includes a read-data interconnect 718a that can receive the data from the module 708a via a plurality of read-data bonds 742a. The data may be held by the register 726a for communication to other circuitry via the read API 714a.

[00123] The read register 726a may also include a read-data enable interconnect 720a to enable the output of data from the module 708a. The read-data enable interconnect 720a is coupled to read-data enable bond 744a so that when stacked, the chiplets 702, 704 are in electrically communication with each other.

[00124] A read register 726b may be similar or identical to the read register 726a on the application chiplet 702 that can communicate with other circuitry through a read application programming interface (API) 714b. The read API 714b acts as a bus where external circuitry, such as a CPU, can request data from the memory chiplet 704.

[00125] The read register 726b is controlled by a read clock, which is received through a read-clock interconnect 712b. The read-clock interconnect 712b is also connected to a read-clock bond 738b, which provides clocking for the memory chiplet 704. The read-clock bond 738b connects the interconnects on the application chiplet 702 and the memory chiplet 704.

[00126] The read register 726b is connected to a read-address interconnect 716b, which transmits a loaded read address. The read-address interconnect 716b is connected to multiple read-address bonds 740b, allowing the memory chiplet 704 to receive the read address. The read address is used by the read peripheral 722b to query a location within the module 708b.

[00127] The read register 726b also has a read-data interconnect 718b, which receives data from the module 708. The data is stored in the register 726b and can be communicated to other circuitry through the read API 714b.

[00128] Additionally, the read register 726b includes a read-data enable interconnect 720b, which enables the output of data from the module 708b. The read-data enable interconnect 720b is connected to a read-data enable bond 744b, ensuring electrical communication between the stacked chiplets 702 and 704.

[00129] The interface module 701 also includes a write-interface register 728 that receives write data and a write address via a write application programming interface 724 to write the write data to the group of modules 706. The write-interface register 728 receives a write clock via a write write-clock interconnect 730 that is also coupled to the surface of the semiconductor device via write-clock bond 746. The write-interface register 728 also has coupled to it write-address interconnects 732 to provide a write address from the write-interface

register 728 to a plurality of write-address bonds 748 on the surface. The write-interface register 728 is also coupled to write-data interconnect 734 to send data via a plurality of write-data bonds 750. The write-interface register 728 is also coupled to a write-data enable interconnect 736 which send a write enable signal via a write-enable bond 752.

[00130] Fig. 8 shows a block diagram of a system 800 employing correct-by-construction timing closure for an application chiplet having a common read/write address register 812 and a memory chiplet 802 in accordance with an embodiment of the present disclosure. The system 800 includes comprises various interconnected components and may be part of a stack of chiplets that are bonded together.

[00131] The interface module 806 comprises circuitry responsible for interfacing between different system components. This logic incorporates the address register 810, the read data register 816, and the write data register 820 to facilitate memory transfer between the chiplets. The memory chiplet 802 is responsible for data storage within the system 800. The read/write select interconnect 826 serves as a mechanism or signal used to select between read and write operations, providing flexibility and control over the system's functionalities. The read/write select interconnect 826 is coupled to read/write bond 836. The clock serves as a timing signal, facilitating synchronization across the system, including the address register 810, read data register 816, and write data register 820 and the memory 802 via the clock interconnect 808.

[00132] The address interconnects 804 facilitate the transmission of the address signals via a plurality of bonds 828 to the memory chiplet 802. The address may be for reading or writing depending upon the value of the read/write select signal on the read/write select interconnect 826. The address register 810 serves as a storage mechanism for the address received via the API 812 by loading the address to the address interconnects 804 and communicates the address to the memory 802 via address interconnect bonds 830. The address register API 812 acts as an application programming interface related to the address register which may be accessed by other circuitry. To synchronize operations, the clock uses the clock interconnect 808 to transmit clock signals, enabling the loading of an address into the address register 810.

[00133] In the embodiment shown in Fig. 8, there are two data registers utilized: the read data register 816 and the write data register 820. The read data register 816 receives data in the read mode when selected by the read/write select 826, while the write data register 820 provides the memory chiplet 802 with data to be written.

[00134] The read data register 816 serves as a dedicated register for storing read data, ensuring data integrity and accessibility. It is connected to the memory 802 through the read-data bonds 832 via the read data interconnect 814. Additionally, there is an application programming interface, the read-data register API 818, associated with the read data register 816. This API allows seamless integration and interaction with other circuitry.

[00135] Similarly, there is a write data register 820 responsible for storing write data and a write data register API 822 associated with it. The write data register 820 acts as a dedicated register for storing write data to be written to the memory chiplet 802. The write data register API 822 serves as an application programming interface for receiving write data from other circuitry. The write data signals are transmitted through the write-data interconnects 824 via write-data bonds 834, enabling smooth data transfer during write operations to the memory chiplet 802.

[00136] In the embodiment shown in Fig. 8, the read data register 816 and the write data register 820 may be connected to the memory chiplet 802 through bonds that are conductive connections, such as conductive pads. The memory chiplet 802 includes complementary conductive connections as well (e.g., conductive pads). These conductive pads may be metal pads, which are commonly used in integrated circuit technology. Metal pads can act as the contact points for interfacing the chiplets together.

[00137] A netlist may be used to represent a comprehensive representation of the interconnections between various components present within the system, providing a visual depiction of circuitry arrangements. The system incorporates a plurality of bonds, denoting multiple physical connections that facilitate seamless communication between components or devices.

[00138] Fig. 9 shows a flow chart diagram of a method 900 for ensuring that correct-by-construction timing closure is achieved in accordance with an embodiment of the present disclosure.

[00139] The method 900 may include forming a thin interface module on a base die (block 902). For example, device may form a thin interface module on a base die, as described above. As also shown in Fig. 9, the method 900 may include forming an address register in the thin interface module, the address register configured to register an address for address interconnects in accordance with an address clock from an address-clock interconnect (block 904). For example, device may form an address register in the thin interface module, the address register configured to register an address for address interconnects in accordance with an address clock from an address-clock interconnect, as described above. As further shown in

Fig. 9, the method 900 may include forming a data register in the thin interface module, the data register configured to register data for data interconnects in accordance with a data clock from a data-clock interconnect (block 906). For example, device may form a data register in the thin interface module, the data register configured to register data for data interconnects in accordance with a data clock from a data-clock interconnect, as described above.

[00140] The method 900 may include forming a plurality of bonds disposed on a surface of the first semiconductor device, the plurality of bonds operatively coupled to the address interconnects and the data interconnects (block 908). For example, the device may form a plurality of bonds disposed on a surface of the first semiconductor device, the plurality of bonds operatively coupled to the address interconnects and the data interconnects, as described above. The bonds may be vertically wired to the surface using EDA software (e.g., one that uses netlists) where there is a predetermined maximum distance that the thin interface module can be from the plurality of bonds. There may also be constraints that the wires must be vertically routed directly to the surface of the chiplet.

[00141] The method 900 may stack the first and second semiconductor devices together (block 910), bond the plurality of bonds with the second semiconductor device (block 912) such as via complementary bonds, and bond the first and second semiconductor devices together (block 914)

[00142] Various alternatives and modifications can be devised by those skilled in the art without departing from the disclosure. Accordingly, the present disclosure is intended to embrace all such alternatives, modifications, and variances. Additionally, while several embodiments of the present disclosure have been shown in the drawings and/or discussed herein, it is not intended that the disclosure be limited thereto, as it is intended that the disclosure be as broad in scope as the art will allow and that the specification be read likewise. Therefore, the above description should not be construed as limiting, but merely as exemplifications of particular embodiments. And those skilled in the art will envision other modifications within the scope and spirit of the claims appended hereto. Other elements, steps, methods, and techniques that are insubstantially different from those described above and/or in the appended claims are also intended to be within the scope of the disclosure.

[00143] The embodiments shown in the drawings are presented only to demonstrate certain examples of the disclosure. And the drawings described are only illustrative and are non-limiting. In the drawings, for illustrative purposes, the size of some of the elements may be exaggerated and not drawn to a particular scale. Additionally, elements shown within the

drawings that have the same numbers may be identical elements or may be similar elements, depending on the context.

[00144] Where the term "comprising" is used in the present description and claims, it does not exclude other elements or steps. Where an indefinite or definite article is used when referring to a singular noun, e.g., "a," "an," or "the," this includes a plural of that noun unless something otherwise is specifically stated. Hence, the term "comprising" should not be interpreted as being restricted to the items listed thereafter; it does not exclude other elements or steps, and so the scope of the expression "a device comprising items A and B" should not be limited to devices consisting only of components A and B. This expression signifies that, with respect to the present disclosure, the only relevant components of the device are A and B.

[00145] Furthermore, the terms "first," "second," "third," and the like, whether used in the description or in the claims, are provided for distinguishing between similar elements and not necessarily for describing a sequential or chronological order. It is to be understood that the terms so used are interchangeable under appropriate circumstances (unless clearly disclosed otherwise) and that the embodiments of the disclosure described herein are capable of operation in other sequences and/or arrangements than are described or illustrated herein.

What is Claimed is:

1. A system for achieving correct-by-construction timing closure, the system comprising:

a first semiconductor device having a thin interface module, the thin interface module comprising:

an address register configured to register an address for address interconnects in accordance with an address clock from an address-clock interconnect;

a data register configured to register data for data interconnects in accordance with a data clock from a data-clock interconnect; and

a plurality of bonds disposed on a surface of the first semiconductor device, the plurality of bonds operatively coupled to the address interconnects and the data interconnects, the plurality of bonds configured to interface with a bonded second semiconductor device.

2. The system according to claim 1, wherein the system is configured for achieving correct-by-construction timing closure between the first and second semiconductor devices.

3. The system according to claim 1, wherein the first and second semiconductor devices are face-to-face bonded together.

4. The system according to claim 1, wherein the first and second semiconductor devices are stacked together.

5. The system according to claim 1, wherein the address-clock interconnect and the data-clock interconnect are coupled together to thereby merge the address clock and the data clock.

6. The system according to claim 1, wherein the data register is a read data register configured to receive read data via the data interconnects.

7. The system according to claim 6, further comprising:

a write data register configured to register write data to provide the write data to a plurality of write data interconnects, wherein the plurality of bonds includes the write data interconnects.

8. The system according to claim 1, wherein the data register is a write data register configured to provide write data to the data interconnects.

9. The system according to claim 8, further comprising:

at least two read data registers configured to receive the read data from a plurality of read data interconnects, wherein the plurality of bonds includes the read data interconnects.

10. The system according to claim 1, the first semiconductor device further comprising:
a read-interface register, wherein the address register and the data register form the read-interface register, the read-interface register configured to register a read address from read interconnects and read data from the read interconnects in accordance with a read clock from a read-clock interconnect, wherein the read interconnects include the address interconnects and the data interconnects, the address clock and the data clock form the read clock, the address-clock interconnect and the data-clock interconnect form the read-clock interconnect;
a write-interface register configured to register a write address and write data from write interconnects in accordance with a write clock from a write-clock interconnect;
a plurality of read bonds disposed on a surface of the first semiconductor device, the plurality of read bonds operatively coupled to the read interconnects of the read-interface register, wherein the plurality of bonds includes the plurality of read bonds; and
a plurality of write bonds disposed on the surface of the first semiconductor device, the plurality of write bonds operatively coupled to the write interconnects of the write-interface register.
11. The system according to claim 10, the system further comprising:
the second semiconductor device comprising:
a read peripheral having a read-peripheral interconnects coupled to the surface of the second semiconductor device and configured to couple to the plurality of read bonds to thereby couple the read peripheral of the second semiconductor device to the read-interface register of the first semiconductor device.
12. The system according to claim 10, the system further comprising:
the second semiconductor device comprising:
a write peripheral having a write-peripheral interconnects coupled to the surface of the second semiconductor device and configured to couple to the plurality of write bonds to thereby couple the write peripheral of the second semiconductor device to the write-interface register of the first semiconductor device.
13. The system according to claim 10, wherein the read interconnects comprise:
read-address interconnects; and
read-data interconnects.

14. The system according to claim 13, wherein the plurality of read bonds comprises:
a plurality of read-address bonds disposed on a surface of the first semiconductor device, the plurality of read-address bonds operatively coupled to the read-address interconnects of the read-interface register; and
a plurality of read-data bonds disposed on the surface of the first semiconductor device, the plurality of read-data bonds operatively coupled to the read-data interconnects of the read-interface register.
15. The system according to claim 14, wherein the plurality of bonds includes a read-clock bond, wherein the read-clock interconnect couples the read-clock bond to the read-interface register.
16. The system according to claim 10, wherein the write interconnects comprise:
write-address interconnects; and
write-data interconnects.
17. The system according to claim 16, wherein the plurality of write bonds comprises:
a plurality of write-address bonds disposed on the surface of the first semiconductor device, the plurality of write-address bonds coupled to the write-address interconnects of the write-interface register; and
a plurality of write-data bonds disposed on the surface of the first semiconductor device, the plurality of write-data bonds coupled to the write-data interconnects of the write-interface register.
18. The system according to claim 17, wherein the plurality of bonds includes a write-clock bond, wherein the write-clock interconnect couples the write-clock bond to the write-interface register.
19. A method for forming a first semiconductor device having correct-by-construction timing closure, the method comprising:
forming a thin interface module on a base die;
forming an address register in the thin interface module, the address register configured to register an address for address interconnects in accordance with an address clock from an address-clock interconnect;
forming a data register in the thin interface module, the data register configured to register data for data interconnects in accordance with a data clock from a data-clock interconnect;

forming a plurality of bonds disposed on a surface of the first semiconductor device, the plurality of bonds operatively coupled to the address interconnects and the data interconnects; and

bonding the plurality of bonds with a second semiconductor device.

20. The method according to claim 19, further comprising forming the thin interface module in spaced relation that is less than a predetermined distance to the plurality of bonds thereby achieving correct-by-construction timing closure between the first and second semiconductor devices.

21. The method according to claim 19, wherein the act of bonding the plurality of bonds with the second semiconductor device comprises face-to-face bonding the first semiconductor device to the second semiconductor device.

22. The method according to claim 19, wherein the act of bonding the plurality of bonds with the second semiconductor device comprises stacking the first semiconductor device and the second semiconductor device together.

23. The method according to claim 19, wherein the act of forming the thin interface module on the base die comprises forming the thin interface module under the plurality of bonds.

24. The method according to claim 19, wherein the act of forming the plurality of bonds comprises forming the plurality of bonds over the thin interface module.

25. The method according to claim 19, wherein the forming acts are implemented in accordance with a netlist.

26. The method according to claim 19, the method further comprising forming the plurality of bonds and the address register in spaced relation to each other such that communication between the plurality of bonds and the address register is less than a predetermined amount of time.

27. The method according to claim 19, further comprising vertically wiring the address interconnects and the data interconnects to the plurality of bonds.

28. The method according to claim 19, further comprising bonding the first and second semiconductor devices together.

29. The method according to claim 28, wherein the bonding is face-to-face bonding.

30. The method according to claim 19, further comprising stacking the first and second semiconductor devices together.

31. The method according to claim 19, further comprising merging the address-clock interconnect and the data-clock interconnect together to thereby merge the address clock and the data clock.
32. The method according to claim 19, further comprising receiving read data via the data interconnects, wherein the data register is a read data register.
33. The method according to claim 32, further comprising:
registering write data into a write data register; and
providing the write data to a plurality of write data interconnects, wherein the plurality of bonds includes the write data interconnects.
34. The method according to claim 19, further comprising providing write data to the data interconnects, wherein the data register is a write data register.
35. The method according to claim 34, further comprising:
receiving first read data in a first read data register from a first memory module on the second semiconductor device; and
receiving second read data in a second read data register from a second memory module on the second semiconductor device.
36. The method according to claim 19, wherein the first semiconductor device further comprising:
a read-interface register, wherein the address register and the data register form the read-interface register, the read-interface register configured to register a read address from read-address interconnects and read data from the read interconnects in accordance with a read clock from a read-clock interconnect, wherein the read interconnects include the address interconnects and the data interconnects, the address clock and the data clock form the read clock, the address-clock interconnect and the data-clock interconnect form the read-clock interconnect;
a write-interface register configured to register a write address and write data from write interconnects in accordance with a write clock from a write-clock interconnect;
a plurality of read bonds disposed on a surface of the first semiconductor device, the plurality of read bonds operatively coupled to the read interconnects of the read-interface register, wherein the plurality of bonds includes the plurality of read bonds; and
a plurality of write bonds disposed on the surface of the first semiconductor device, the plurality of write bonds operatively coupled to the write interconnects of the write-interface register.

37. The method according to claim 36, wherein the second semiconductor device comprises a read peripheral having read-peripheral interconnects coupled to the surface of the second semiconductor device and configured to couple to the plurality of read bonds to thereby couple the read peripheral of the second semiconductor device to the read-interface register of the first semiconductor device.
38. The method according to claim 36, wherein the second semiconductor device comprises a write peripheral having write-peripheral interconnects coupled to the surface of the second semiconductor device and configured to couple to the plurality of write bonds to thereby couple the write peripheral of the second semiconductor device to the write-interface register of the first semiconductor device.
39. The method according to claim 36, wherein the read interconnects comprise:
read-address interconnects; and
read-data interconnects.
40. The method according to claim 36, wherein the plurality of read bonds comprises:
a plurality of read-address bonds disposed on a surface of the first semiconductor device, the plurality of read-address bonds operatively coupled to the read-address interconnects of the read-interface register; and
a plurality of read-data bonds disposed on the surface of the first semiconductor device, the plurality of read-data bonds operatively coupled to the read-data interconnects of the read-interface register.
41. The method according to claim 40, wherein the plurality of bonds includes a read-clock bond, wherein the read-clock interconnect couples the read-clock bond to the read-interface register.
42. The method according to claim 36, wherein the write interconnects comprise:
write-address interconnects; and
write-data interconnects.
43. The method according to claim 42, wherein the plurality of write bonds comprises:
a plurality of write-address bonds disposed on the surface of the first semiconductor device, the plurality of write-address bonds coupled to the write-address interconnects of the write-interface register; and

a plurality of write-data bonds disposed on the surface of the first semiconductor device, the plurality of write-data bonds coupled to the write-data interconnects of the write-interface register.

44. The method according to claim 43, wherein the plurality of bonds includes a write-clock bond, wherein the write-clock interconnect couples the write-clock bond to the write-interface register.

45. A method of using the system according to one of claim 1-18, comprising:
providing the address to the address register; and
clocking the address clock to thereby load the address register with the address.

46. A method of using the system according to one of claim 1-18, comprising:
providing the data to the data register; and
clocking the data clock to thereby load the data register with the data.

47. A system for achieving correct-by-construction timing closure, the system comprising:

a first semiconductor means for providing a thin interface module on a base die, the thin interface module comprising:

an address means for registering an address for address interconnects in accordance with an address clock from an address-clock interconnect;

a data means for registering data for data interconnects in accordance with a data clock from a data-clock interconnect; and

a plurality of bonding means for communicating with a second semiconductor device via address interconnect means and the data interconnect means.