SYSTEM FOR CONTROLLING TEMPERATURES OF MEMORY AND METHOD OF OPERATING SAME

Abstract

A system (for controlling temperatures in a memory) includes: a high bandwidth memory (HBM) including core dies, the HBM being arranged into portions, each of the portions including memory cells, the HBM further including: a first sensing unit configured to generate a first environmental signal corresponding to a first transistor in a corresponding first one of the memory cells in a first one of the portions; and a second sensing unit configured to generate a second environmental signal corresponding to a second transistor in a corresponding second one of the memory cells in a second one of the portions; and a controller configured to perform non-monolithic temperature control of the HBM based on the one or more first environmental signals and the one or more second environmental signals such that temperature of one or more of the portions is differentially controlled.

Inventors:	**FAN; Philex Ming-Yan (Hsinchu, TW), HUANG; Chia-En (Hsinchu, TW), WANG; Yih (Hsinchu, TW), CHANG; Jonathan Tsung-Yung (Hsinchu, TW)**
Applicant:	**TAIWAN SEMICONDUCTOR MANUFACTURING COMPANY, LTD.** (Hsinchu, TW)
Family ID:	**1000008586955**
Appl. No.:	**19/185905**
Filed:	**April 22, 2025**

Related U.S. Application Data

parent US continuation 18439982 20240213 parent-grant-document US 12283336 child US 19185905

parent US continuation 17680044 20220224 parent-grant-document US 11901035 child US 18439982

us-provisional-application US 63220033 20210709

Publication Classification

Background/Summary

PRIORITY CLAIM [0001] The present application is a continuation of U.S. patent application Ser. No. 18/439,982, filed Feb. 13, 2024, now U.S. Pat. No. 12,283,336, issued Apr. 22, 2025, which is a continuation of U.S. patent application Ser. No. 17,680,044, filed Feb. 24, 2022, now U.S. Pat. No. 11,901,035, issued filed Feb. 13, 2024, that claims the priority of U.S. Provisional Application No. 63/220,033, filed Jul. 9, 2021, each of which is incorporated herein by reference in its entirety.

BACKGROUND
[0002] An integrated circuit ("IC") includes one or more semiconductor devices. One way in which to represent a semiconductor device is with a plan view diagram referred to as a layout diagram.

Description

BRIEF DESCRIPTION OF THE DRAWINGS
[0003] Aspects of the present disclosure are best understood from the following detailed description when read with the accompanying figures. It is noted that, in accordance with the standard practice in the industry, various features are not drawn to scale. In fact, the dimensions of the various features may be arbitrarily increased or reduced for clarity of discussion.
[0004] FIG. **1** is a block diagram of a semiconductor device, in accordance with some embodiments.
[0005] FIG. **2**A is a cross-sectional view of a semiconductor device, in accordance with some embodiments.
[0006] FIG. **2**B is a three-quarter perspective view of an HBM, in accordance with some embodiments.
[0007] FIGS. **2**C-**2**E are corresponding block diagrams of corresponding sensing units, in accordance with some embodiments.
[0008] FIGS. **3**A-**3**D are block diagrams of corresponding HBMs, in accordance with some embodiments.
[0009] FIGS. **4**A-**4**G are block diagrams of corresponding DDV FS devices, in accordance with some embodiments.
[0010] FIGS. **5**A-**5**B are circuit diagrams of corresponding temperature sensors, in accordance with some embodiments.
[0011] FIGS. **6**A-**6**C are circuit diagrams of corresponding threshold voltage detectors, in accordance with some embodiments.
[0012] FIGS. **7**A-**7**E are corresponding flowcharts, in accordance with some embodiments.
[0013] FIGS. **7**F-**7**I are corresponding flowcharts showing blocks of FIG. **7**A in more detail, in accordance with some embodiments.
[0014] FIG. **8** is a flowchart of a method of manufacturing a semiconductor device, in accordance with some embodiments.

[0015] FIG. **9** is a block diagram of an electronic design automation (EDA) system in accordance with some embodiments.

[0016] FIG. **10** is a block diagram of an integrated circuit (IC) manufacturing system, and an IC manufacturing flow associated therewith, in accordance with some embodiments.

DETAILED DESCRIPTION

[0017] The following disclosure provides many different embodiments, or examples, for implementing different features of the provided subject matter. Specific examples of components, materials, values, steps, operations, materials, arrangements, or the like, are described below to simplify the present disclosure. These are, of course, merely examples and are not intended to be limiting. Other components, values, operations, materials, arrangements, or the like, are contemplated. For example, the formation of a first feature over or on a second feature in the description that follows may include embodiments in which the first and second features are formed in direct contact, and may also include embodiments in which additional features may be formed between the first and second features, such that the first and second features may not be in direct contact. In addition, the present disclosure may repeat reference numerals and/or letters in the various examples. This repetition is for the purpose of simplicity and clarity and does not in itself dictate a relationship between the various embodiments and/or configurations discussed.

[0018] Further, spatially relative terms, such as "beneath," "below," "lower," "above," "upper" and the like, may be used herein for ease of description to describe one element or feature's relationship to another element (s) or feature (s) as illustrated in the figures. The spatially relative terms are intended to encompass different orientations of the device in use or operation in addition to the orientation depicted in the figures. The apparatus may be otherwise oriented (rotated 90 degrees or at other orientations) and the spatially relative descriptors used herein may likewise be interpreted accordingly.

[0019] In some embodiments, a high-bandwidth memory (HBM) includes a stack of core dies/chips/ICs, e.g., dynamic random access memory (DRAM) core dies and a differentiated DVFS device, where DV FS is an acronym for dynamic voltage and frequency scaling (DVFS). Banks of the core dies include sensing units (each including a temperature sensor and/or a threshold voltage detector). In general, frequent read or write activity at a memory cell in a core die raises the temperature of the memory cell. An unchecked/uncontrolled increase in temperature can lead to thermal runaway. When the temperature of a memory cell in a core die exceeds a permissible maximum temperature, the value of the data bit stored in the memory cell is susceptible to becoming corrupted, e.g., due to increased leakage, or the like. According to another approach, a technique referred to as monolithic DVFS is used to prevent overheating in an HBM. The monolithic DVFS technique according to the other approach monitors temperatures in each of the core dies of the HBM. When a portion of any core die exceeds a permissible temperature, the monolithic DVFS technique of the other approach reduces the clock of all the bit cells of all the core dies, wherein the clock is reduced by decreasing the frequency of the clock that is supplied to all of the core dies. The monolithic DVFS technique according to the other approach is described as a monolithic type of thermal throttling because the same clock provided to all of the bit cells of all the core dies thereby treating the HM B as if the HMB is a monolith, i.e., not treating the HBM of the other approach as including different portions which behave differently.

[0020] Often, however, the portion of a given core die which exceeds the permissible temperature is a minority portion of the given core die. As a result, the monolithic thermal throttling according to the other approach not only cools the minority portion of the given core die, but also unnecessarily cools the majority portion of the given core die and the entireties of the other core dies, resulting in unnecessarily diminished performance of the majority portion of the given core die and the entireties of the other core dies, which is wasteful and inefficient. In contrast to the monolithic thermal throttling of the other approach, in terms of portions of the HBM whose temperatures are to be controlled, the differentiated thermal throttling of some embodiments, e.g.,

bank-wide, group-wide, channel-wide, core-wide, or the like, is more granular, and thus is less wasteful and more efficient.

[0021] FIG. **1** is a block diagram of a semiconductor device **100** in accordance with at least one embodiment of the present disclosure.

[0022] In FIG. **1**, semiconductor device **100** includes, among other things, a processor **102** and a high-bandwidth memory (HBM) **104** operationally coupled together by a bus **106**. In some embodiments, semiconductor device **100** is a system-in-package (SIP) type of semiconductor device. In some embodiments, semiconductor device **100** is a compute-in-memory (CIM) system. An example of semiconductor device **102** is shown in FIG. **2**A, or the like.

[0023] FIG. **2**A is a cross-sectional view of a semiconductor device **200**, in accordance with some embodiments.

[0024] More particularly, semiconductor device **200** is a SIP type of semiconductor device. SIP **200** includes a processor **202** and an HBM **204** operationally coupled together by a bus **206**. SIP **200** is an example of semiconductor device **100** of FIG. **1**. In some embodiments, SIP **200** is a compute-in-memory (CIM) system. Processor **202** is an example of processor **102** of FIG. **1**. HBM **204** is an example of HBM **104** of FIG. **1**. Bus **206** is an example of bus **106** of FIG. **1**. In some embodiments, processor **202** is a central processing unit (CPU). In some embodiments, processor **202** is a graphics processing unit (GPU). In some embodiments, processor **202** is a visual processing unit (VPU). In some embodiments, processor **202** is a general purpose graphics processing unit (GPGPU).

[0025] In FIG. **2**A, SIP **200** further includes an interposer structure **208** and a printed circuit board (PCB) **210**. Interposer structure **208** is formed over PCB **210**. Interposer **208** is formed of a semiconductor, e.g., silicon. Interposer **208** includes portions of signal lines which comprise bus **206**, and portions of signal lines **216**. In some embodiments, signals on bus **206** are communicated internally with respect to SIP **200**. Accordingly, in such embodiments, the signal lines of SIP **200** are described as intra-SIP signal lines. In some embodiments, the signals on at least some of signal lines **216** are communicated externally with respect to SIP **200**. Accordingly, in such embodiments, signal lines of SIP **200** are generally described as extra-SIP signal lines though not every signal line **216** necessarily carries a signal which is communicated externally with respect to SIP **200**.

[0026] In FIG. **2**A, HBM **204** includes a stack **214** of four core dies **218**(**1**), **218**(**2**), **218**(**3**) and **218**(**4**). In some embodiments, HBM **204** includes a buffer die (FIG. **2**B) under stack **214**. In some embodiments, stack **214** includes two or three or five or more core dies. HBM **204** further includes one or more differentiated dynamic voltage and frequency scaling (DDVFS) devices (see FIGS. **3**A-**3**D).

[0027] FIG. **2**B is a three-quarter perspective view of HBM **204**, in accordance with some embodiments.

[0028] In FIG. **2**B, each of core dies **218**(**1**)-**218**(**4**) includes memory cells MC which are arranged/organized into two channels of eight banks B**0**-B**7**. Each channel is arranged into two groups of banks. Accordingly, HBM **204** includes eight channels **222**(**0**)-**222**(**7**). More particularly, core die **218**(**0**) is arranged into channels **222**(**0**) and **222**(**1**); core die **218**(**1**) is arranged into channels **222**(**2**) and **222**(**3**); core die **218**(**2**) is arranged into channels **222**(**4**) and **222**(**5**); and core die **218**(**3**) is arranged into channels **222**(**6**) and **222**(**7**).

[0029] Also, accordingly, for example, core die **218**(**3**) includes four groups of banks. More particularly, channel **222**(**6**) of core die **218**(**3**) includes: group **224**(**0**) comprised of banks B**0**-B**3**; and group **222**(**4**) comprised of banks B**4**-B**7**. Channel **222**(**7**) of core die **218**(**3**) includes: group **224**(**1**) comprised of banks B**0**-B**3**; and group **222**(**3**) comprised of banks B**4**-B**7**. Relative to the X-axis shown in FIG. **2**B, group **224**(**2**) is on the left side of channel **222**(**6**) and group **224**(**0**) is on the right side of channel **222**(**6**). Similarly, group **224**(**3**) is on the left side of channel **222**(**7**) and group **224**(**1**) is on the right side of channel **222**(**7**). In some embodiments, groups **224**(**2**) and **224**(**3**) are referred to as left-side groups and groups **224**(**0**) and **224**(**1**) are referred to as right-side

groups.

[0030] In some embodiments, an aggregation of one or more but fewer than all of the banks in an HBM is referred to as a lot. In some embodiments, an aggregation of one or more but fewer than all of the groups in an HBM is referred to as a batch. In some embodiments, an aggregation of one or more but fewer than all of the channels in an HBM is referred to as a bundle. In some embodiments, an aggregation of one or more but fewer than all of the core-dies in an HBM is referred to as a collection.

[0031] In FIG. **2**B, HBM **204** further includes arrays **226(1)** and **226(2)** of through-silicon vias (TSVs) **228**, or the like, and arrays **230(1)** and **230(2)** of TSVs **232**, or the like. Relative to the X-axis, arrays **226(1)** and **230(1)** are between the left and right groups of each of channel **222(6)** of core die **218(3)**, channel **222(4)** of core die **218(2)**, channel **222(2)** of core die **218(1)**, and channel **222(0)** of core die **218(0)**. Similarly, arrays **226(2)** and **230(2)** are between the left and right groups of each of channel **222(7)** of core die **218(3)**, channel **222(6)** of core die **218(2)**, channel **222(3)** of core die **218(1)**, and channel **222(1)** of core die **218(0)**.

[0032] Relative to the Y-axis shown in FIG. **2**B, array **230(1)** is between array **226(1)** and array **230(2)**. Similarly, array **230(2)** is between array **226(2)** and array **230(1)**.

[0033] In some embodiments, TSVs **228** of array **226(1)** provide power system (PS) voltages to even-numbered channels **222(0)**, **222(2)**, **222(4)** and **222(6)**, and TSVs **228** of array **226(2)** carry PS voltages to odd-numbered channels **222(1)**, **222(3)**, **222(5)** and **222(7)**.

[0034] In some embodiments, TSVs **232** of array **230(1)** carry input/output (I/O) signals and/or control signals to even-numbered channels **222(0)**, **222(2)**, **222(4)** and **222(6)**, and TSVs **230** of array **232(2)** carry I/O signals and/or control signals to odd-numbered channels **222(1)**, **222(3)**, **222(5)** and **222(7)**.

[0035] In FIG. **2**B, memory cells MC in each of banks B**0**-B**7** are arranged into arrays. Each memory cell MC is a dynamic random access memory (DRAM) cell. In some embodiments, memory cells MC are a type of memory cell other than a DRAM cell.

[0036] As shown in an exploded view **223** included in FIG. **2**B, each memory cell MC includes a capacitor C**1** and a first field-effect transistor (FET) FET**1**. Capacitor C**1** is used to store data voltages, e.g., bit values. That is, capacitor C**1** stores a voltage representing a logical zero, i.e., a 0 bit, or a voltage representing a logical one, i.e., a 1 bit. Transistor FET**1** is a switch used to transmit data voltages to capacitor C**1**. In exploded view **223**, DRAM cell MC is assumed to be in the i-th row and the j-th column of the array. Accordingly, switch FET**1** electrically coupled between i-th bitline B(i) and a first terminal of capacitor C**1**, and a second terminal of capacitor C**1** is electrically coupled to a low system voltage, e.g., ground, as contrasted with a high system voltage, VDD. In some embodiments, ground is VSS. The gate electrode of switch FET**1** is electrically coupled to a j-th row-enable line RE(j). Capacitor C**1** is electrically coupled to bitline BL(i) when switch FET**1** is turned on by the signal on row-enable line RE(j).

[0037] In FIG. **2**B, each of banks B**0**-B**7** includes corresponding sensing units S**0**-S**7** (see FIGS. **2**C-**2**D). For example, bank B**0** of group **224(0)** of channel **222(6)** of core die **218(3)** includes sensing unit S**0**. Physically/spatially, sensing units S**0**-S**7** are located proximally to corresponding banks B**0**-B**7**. Each of sensing units S**0**-S**7** generates environmental information, e.g., temperature and/or threshold voltage, or the like. Because of the proximity of sensing units S**0**-S**7** to corresponding banks B**0**-B**7**, the environmental information generated by sensing units S**0**-S**7** is regarded as being representative of all of memory cells MC in corresponding banks B**0**-B**7**. For example, because of the proximity of sensing unit S**2** of bank B**2** of group **224(0)** of channel **226(6)** of core die **218(3)** to bank B**2** of group **224(0)** of channel **226(6)** of core die **218(3)**, the environmental information generated by sensing unit S**2** of bank B**2** of group **224(0)** of channel **226(6)** of core die **218(3)** is regarded as being representative of all of memory cells MC in bank B**2** of group **224(0)** of channel **226(6)** of core die **218(3)**.

[0038] As shown in FIG. **2**B: relative to the Y-axis, sensing units S**0**-S**7** are centrally located in

corresponding banks B**0**-B**7**; and, relative X-axis, sensing units S**0**-S**7** are located at the right side of corresponding banks B**0**-B**7**. In some embodiments, relative to the X-axis and the Y-axis, sensing units S**0**-S**7** are centrally located in corresponding banks B**0**-B**7**. In some embodiments, relative to the X-axis and the Y-axis, sensing units S**0**-S**7** are located positions other than as noted above.

[0039] For simplicity of illustration, not all of the components of HBM **204** are shown in FIG. **2**B. Also, for simplicity of illustration, not all components shown in HMB **204** are labelled in FIG. **2**B. For example, in FIG. **2**B, HBM **204** further includes one or more DDVFS devices (see FIGS. **3**A-**3**D). In some embodiments, each of core dies **218**(**0**)-**218**(**3**) includes one or more DDVFS devices. In some embodiments, one or more DDVFS devices are located in a buffer die **212**.

[0040] In some embodiments, HBM **204** is compliant with a standard JESD-**235** as defined by the standards-setting organization known as the Joint Electron Device Engineering Council (JEDEC). In some embodiments, HBM **204** is compliant with a standard JESD-**235**A as defined by the JEDEC. In some embodiments, HBM **204** is compliant with a standard for HMBs defined by another standards-setting organization. Details of typical HBMs, albeit HMBs which do not include, e.g., sensors S**0**-S**7** and DDVFS devices such as disclosed herein, are found, e.g., in U.S. Pat. No. 10,180,906, granted Jan. 15, 2019, and U.S. Pre-Grant Publication No. 202100247978, published Aug. 12, 2021, the entireties of each of which are hereby incorporated by reference.

[0041] FIGS. **2**C-**2**E a corresponding block diagrams of corresponding sensing units, in accordance with some embodiments.

[0042] Sensing unit **235**C of FIG. **2**C is an example of each of sensing units S**0**-S**7** of FIG. **2**B. Accordingly, sensing unit **235**C is also labeled "Sx" in FIG. **2**C.

[0043] In FIG. **2**C, sensing unit **235**C includes a temperature sensor **236** and a threshold voltage sensor **238**. Temperature sensor **236** (see FIGS. **5**A-**5**B) is configured to sense a temperature and output a signal T which is representative of the sensed temperature. Because of the proximity of temperature sensor **236** to a first given instance of memory cell MC of a given bank, signal T is regarded as being representative of the temperature of all instances of memory cell MC in the given bank. The temperature signal T is an example of an environmental signal. Threshold voltage detector **238** (see FIGS. **6**A-**6**C) is configured to detect a threshold voltage and output a signal Vt which is representative of the sensed threshold voltage. Because of the proximity of threshold voltage sensor **238** to a given instance of transistor FET**1** in a second given instance of memory cell MC of a given bank, signal Vt is regarded as being representative of the threshold voltage of all instances of transistor FET**1** in all instances of memory cell MC in the given bank. The threshold voltage Vt is an example of an environmental signal. In some embodiments, the first given instance of memory cell MC in the given bank is the same as the second given instance of memory cell MC in the given bank.

[0044] In some embodiments, sensing unit **235**C of FIG. **2**C includes temperature sensor **236** but does not include threshold voltage sensor **238**. Such embodiments are shown in FIG. **2**D as sensing unit **235**D.

[0045] In some embodiments, sensing unit **235**C of FIG. **2**C includes threshold voltage sensor **238** but does not include temperature sensor **236**. Such embodiments are shown in FIG. **2**E as sensing unit **235**E.

[0046] FIGS. **3**A-**3**D are block diagrams of corresponding HBMs, in accordance with some embodiments.

[0047] Each of HBMs **304**A, **304**B, **304**C and **304**D of corresponding FIGS. **3**A-**3**D is an example of HBM **204** of FIG. **2**B.

[0048] In some respects, FIGS. **3**A-**3**D are simplified representations of corresponding HBMs **304**A-**304**D as compared to HBM **204** of FIG. **2**B. But in some respects, FIGS. **3**A-**3**D are more detailed representations of corresponding HBMs **304**A-**304**D as compared to HBM **204** of FIG. **2**B.

[0049] In FIGS. **3**A and **3**C-**3**D, each of corresponding HBMs **304**A, **304**C and **304**D includes: a

core die **218**(*x*); a channel **222**(*x*); a bank Bx; and a sensing unit **235**C. Core die **218**(*x*) is an example of each of core dies **218**(**0**)-**218**(**3**) of FIG. **2**B. Channel **228**(*x*) is an example of each of channels **222**(**0**)-**222**(**7**) of FIG. **2**B. Bank Bx is an example of each of banks B**0**-B**7** of FIG. **2**B. In FIG. **3**B, HBM **304**B includes banks B_α, B_β, B_γ and B_δ. Where banks B_α, B_β, B_γ and B_δ are included in left-side groups **224**(*x*), banks B_α, B_β, B_γ and B_δ correspond to banks B**4**-B**7**. Where banks B_α, B_β, B_γ and B_δ are included in right-side groups **224**(*x*), banks B_α, B_β, B_γ and B_δ correspond to banks B**0**-B**3**.

[0050] In FIGS. **3**A-**3**D, each of corresponding HBMs **304**A, **304**B, **304**C and **304**D further includes a differentiated dynamic voltage and frequency scaling (DDVFS) device **340** (see FIGS. **4**A-**4**G). In each of FIGS. **3**A-**3**D, DDVFS device **340** is included in core die **218**(*x*). In some embodiments, core die **218**(*x*) includes multiple instances of DDVFS device **340**.

[0051] In some embodiments, one or more instances of DDVFS device **340** are in a buffer die (FIG. **2**B) rather than in corresponding instances of core die **218**(*x*). In some embodiments, one or more instances of DDVFS device **340** are in a buffer die (FIG. **2**B) in addition to one or more instances of DDVFS device **340** in corresponding instances of core die **218**(*x*). In some embodiments, one or more instances of DDVFS device **340** are in processor **202** (FIG. **2**A) rather than in corresponding instances of core die **218**(*x*). In some embodiments, one or more instances of DDVFS device **340** are in processor **202** (FIG. **2**A) in addition to one or more instances of DDVFS device **340** in corresponding instances of core die **218**(*x*). In some embodiments, one or more instances of DDVFS device **340** are locations other than the locations noted above.

[0052] In FIGS. **3**A-**3**D, each of corresponding HBMs **304**A-**304**D is arranged so that DDVFS device **340** receives: temperature T from temperature sensor **236** (see FIGS. **5**A-**5**B) of sensing unit **235**C; and threshold voltage Vt from threshold voltage detector **238** of sensing unit **235**C. Because of the proximity of temperature sensor **236** to a first given instance of memory cell MC of bank Bx, signal T is regarded as being representative of the temperature of all instances of memory cell MC in bank Bx. Because of the proximity of temperature sensor **236** to a given instance of transistor FET**1** in a second given instance of memory cell MC of bank Bx, signal Vt is regarded as being representative of the threshold voltage of all instances of transistor FET**1** in all instances of memory cell MC in bank Bx. In some embodiments, the first given instance of memory cell MC in bank Bx is the same as the second given instance of memory cell MC in bank Bx.

[0053] In each of FIGS. **3**A-**3**D, DDVFS device **340** is configured to adjust: a local clock signal (see FIGS. **4**A-**4**G) based on temperature T and/or a threshold voltage Vt; and/or adjust a local power supply (PS) voltage (see FIGS. **4**A-**4**G) based on temperature T and/or threshold voltage Vt. Each of the local clock signal and the local PS voltage is an example of a transistor-temperature-affecting (TTA) parameter. By adjusting the local clock signal and/or the local PS voltage, DDVFS device adaptively adjusts/controls temperatures of corresponding instances of memory cell MC (see FIG. **2**B).

[0054] In general, to reduce temperature, DDVFS device **340** decreases a frequency of the local clock signal and/or decreases a value of the local PS voltage. In general, a reduction in the frequency of the local clock signal and/or in value of the local PS voltage tends also to reduce the performance of instances of memory cell MC which receive the local clock signal and/or the local PS voltage. Conversely, where current temperature is below a maximum permissible temperature such that circumstances would permit a safe rise in temperature of instances of memory cell MC which receive the local clock signal and/or the local PS voltage, in general, performance of the of instances of memory cell MC which receive the local clock signal and/or the local PS voltage is improved by DDVFS **340** increasing a frequency of the local clock signal and/or increasing a value of the local PS voltage. The DVFS technique according to the other approach, which uses monolithic thermal throttling to provide the same clock signal to all the bit cells of all the core dies, treats the HMB according to the other approach as a monolith, i.e., does not treat the HBM of the other approach as including different portions which behave differently, increases a frequency the

clock signal only when the temperature of all portions of all core dies in the HBM can be raised safely. Often, however, the portion of a given core die which exceeds the permissible temperature is a minority portion of the given core die. As a result, while the monolithic thermal throttling according to the other approach cools the minority portion of the given core die, the monolithic thermal throttling according to the other approach also unnecessarily prevents a majority portion of the given core die and the entireties of the other core dies from undergoing a safe increase in temperature associated with an increase in the frequency the clock signal and the value of the PS voltage signal, resulting in unnecessarily diminished performance of the majority portion of the given core die and the entireties of the other core dies, which is wasteful and inefficient. In contrast to the monolithic thermal throttling of the other approach, in terms of portions of the HBM which are to be cooled, the differentiated thermal throttling of each of HBM **304**A (bank-wide, discussed below), HBM **304**B (group-wide, discussed below), HBM **304**C (channel-wide, discussed below) and HBM **304**D (core-wide, discussed below) is more granular, and thus is less wasteful and more efficient.

[0055] As discussed in more detail below, FIGS. **3**A-**3**D differ from each other, for example, in terms of which components of corresponding HBMs **304**A-**304**D are configured to receive the local clock signal and/or the local voltage from DDVFS device **340**, and in terms of how many instances of sensing unit **235**C transmit temperature T and/or threshold voltage Vt to DDVFS device **340**.

[0056] Turning specifically to FIG. **3**A, HBM **304**A is configured so that each instance of bank Bx receives the local clock signal and the local supply voltage from a corresponding instance of DDVFS device **340**. Accordingly, HBM **304**A is configured so that each instance of sensing unit **235**C, i.e., each bank Bx, transmits temperature T and threshold voltage Vt to DDVFS device **340**. Because each instance of bank Bx receives the local clock signal and the local supply voltage from DDVFS device **340** in FIG. **3**A, the local clock signal is labeled CLK_bnk and the local PS voltage is labeled PSV_bnk in FIG. **3**A. Each of the signals CLK_bnk and PSV_bnk is an example of a TTA parameter.

[0057] The context for the adjective "local" in FIG. **3**A is bank-specific, otherwise described as bank-by-bank. Accordingly, the arrangement of HBM **304**A results in bank-wide adjustment of corresponding temperatures amongst memory cells (not shown) in each instance of bank Bx, i.e., the arrangement of HBM **304**A results in a bank-wide type of differentiated thermal throttling. Such bank-wide differentiated thermal throttling contrasts with the monolithic thermal throttling of the other approach. The DVFS technique according to the other approach uses monolithic thermal throttling, i.e., provides the same clock signal to all of the bit cells of all the core dies thereby treating the HMB according to the other approach as a monolith, i.e., not treating the HBM according to the other approach as including different portions which behave differently. Often, however, the portion of a given core die which exceeds the permissible temperature is a minority portion of the given core die. As a result, the monolithic thermal throttling according to the other approach not only cools the minority portion of the given core die, but also unnecessarily cools the majority portion of the given core die and the entireties of the other core dies, resulting in unnecessarily diminished performance of the majority portion of the given core die and the entireties of the other core dies, which is wasteful and inefficient. In contrast to the monolithic thermal throttling of the other approach, in terms of portions of the HBM whose temperatures are to be controlled, e.g. cooled, the bank-wide differentiated thermal throttling of HBM **304**A is more granular, and thus is less wasteful and more efficient.

[0058] Turning specifically to FIG. **3**B, HBM **304**B is configured so that each instance of group **224**(x) receives the local clock signal and the local supply voltage from a corresponding instance of DDVFS device **340**. As such, each of banks B_α, B_β, B_γ and B_δ in group **224**(x) receives the same local clock signal and the same local supply voltage from DDVFS device **340**. Accordingly, HBM **304**B is configured so that sensing unit **235**C of one of banks B_α, B_β, B_γ and B_δ

transmits temperature T and threshold voltage Vt to DDVFS device **340**. In FIG. **3**B, bank B_γ is shown as transmitting temperature T and threshold voltage Vt to DDVFS device **340**. In some embodiments (not shown), one of banks B_α, B_β and B_δ transmits temperature T and threshold voltage Vt to DDVFS device **340** rather than bank B_γ. Because each instance of group **224**(*x*) receives the local clock signal and the local supply voltage from DDVFS device **340** in FIG. **3**B, i.e., because each of banks B_α, B_β, B_γ and B_δ in group **224**(*x*) receives the same local clock signal and the same local supply voltage from DDVFS device **340** in FIG. **3**B, the local clock signal is labeled CLK_grp and the local PS voltage is labeled PSV_grp in FIG. **3**B. Each of the signals CLK_grp and PSV_grp is an example of a TTA parameter.

[0059] The context for the adjective "local" in FIG. **3**B is group-specific, otherwise described as group-by-group. Accordingly, the arrangement of HBM **304**B results in group-wide adjustment of corresponding temperatures amongst memory cells (not shown) in each instance of group **224**(*x*), i.e., the arrangement of HBM **304**B results in a group-wide type of differentiated thermal throttling. Such group-wide differentiated thermal throttling contrasts with the monolithic thermal throttling of the other approach. In contrast to the monolithic thermal throttling of the other approach, in terms of portions of the HBM which are to be cooled, the group-wide differentiated thermal throttling of HBM **304**B is more granular, and thus is less wasteful and more efficient.

[0060] Regarding FIG. **3**B, because HBM **304**B is configured so that sensing unit **235**C of one of banks B_α, B_β, B_γ and B_δ transmits temperature T and threshold voltage Vt to DDVFS device **340**, in some embodiments (not shown), at least one but not all of banks B_α, B_β, B_γ and B_δ in a given instance of group **224**(*x*) includes a corresponding sensing unit **245**C.

[0061] Turning specifically to FIG. **3**C, HBM **304**C is configured so that each instance of channel **222**(*x*) receives the local clock signal and the local supply voltage from a corresponding instance of DDVFS device **340**. As such, each instance of group **224**(*x*) in channel **222**(*x*), and thus each corresponding instance of bank Bx in channel **222**(*x*), receives the same local clock signal and the same local supply voltage from DDVFS device **340**. Accordingly, HBM **304**C is configured so that sensing unit **235**C of one of banks Bx in one of groups **224**(*x*) transmits temperature T and threshold voltage Vt to DDVFS device **340**. Because each instance of channel **222**(*x*) receives the local clock signal and the local supply voltage from DDVFS device **340** in FIG. **3**C, the local clock signal is labeled CLK_chan and the local PS voltage is labeled PSV_chan in FIG. **3**C. Each of the signals CLK_chan and PSV_chan is an example of a TTA parameter.

[0062] The context for the adjective "local" in FIG. **3**C is channel-specific, otherwise described as channel-by-channel. Accordingly, the arrangement of HBM **304**C results in channel-wide adjustment of corresponding temperatures amongst memory cells (not shown) in each instance of channel **222**(*x*), i.e., the arrangement of HBM **304**C results in a channel-wide type of differentiated thermal throttling. Such channel-wide differentiated thermal throttling contrasts with the monolithic thermal throttling of the other approach. In contrast to the monolithic thermal throttling of the other approach, in terms of portions of the HBM which are to be cooled, the channel-wide differentiated thermal throttling of HBM **304**C is more granular, and thus is less wasteful and more efficient.

[0063] Regarding FIG. **3**C, because HBM **304**C is configured so that sensing unit **235**C of one of banks Bx in one of groups **224**(*x*) transmits temperature T and threshold voltage Vt to DDVFS device **340**, in some embodiments (not shown), not every instance of group **224**(*x*) in a given instance of channel **222**(*x*), and thus not every instance of bank Bx in the given instance of channel **222**(*x*), includes a corresponding sensing unit **245**C.

[0064] Turning specifically to FIG. **3**D, HBM **304**D is configured so that each instance of core-die **218**(*x*) receives the local clock signal and the local supply voltage from a corresponding instance of DDVFS device **340**. As such, each instance of channel **222**(*x*) in core die **218**(*x*), and thus each corresponding instance of group **224**(*x*) in core die **218**(*x*), and thus each corresponding instance of bank Bx in core die **218**(*x*), receives the same local clock signal and the same local supply voltage

from DDVFS device **340**. Accordingly, HBM **304**D is configured so that sensing unit **235**C of one of banks Bx in one of groups **224**(*x*) in one channels **222**(*x*) in one of core dies **218**(*x*) transmits temperature T and threshold voltage Vt to DDVFS device **340**. Because each instance of core die **218**(*x*) receives the local clock signal and the local supply voltage from DDVFS device **340** in FIG. **3**D, the local clock signal is labeled CLK_core and the local PS voltage is labeled PSV_core in FIG. **3**D. Each of the signals CLK_core and PSV_core is an example of a TTA parameter.

[0065] The context for the adjective "local" in FIG. **3**D is core-specific, otherwise described as core-by-core. Accordingly, the arrangement of HBM **304**D results in core-wide adjustment of corresponding temperatures amongst memory cells (not shown) in each instance of core die **218**(*x*), i.e., the arrangement of HBM **304**D results in a core-wide type of differentiated thermal throttling. Such core-wide differentiated thermal throttling contrasts with the monolithic thermal throttling of the other approach. In contrast to the monolithic thermal throttling of the other approach, in terms of portions of the HBM which are to be cooled, the core-wide differentiated thermal throttling of HBM **304**D is more granular, and thus is less wasteful and more efficient.

[0066] Regarding FIG. **3**D, because HBM **304**D is configured so that sensing unit **235**C of one of banks Bx in one of groups **224**(*x*) in one channels **222**(*x*) in one of core dies **218**(*x*) transmits temperature T and threshold voltage Vt to DDVFS device **340**, in some embodiments (not shown), not every instance of channel **222**(*x*) in a given instance of core die **218**(*x*), and thus not every instance of group **224**(*x*) in the given instance of channel **222**(*x*), and thus not every instance of bank Bx in the given instance of channel **222**(*x*), includes a corresponding sensing unit **245**C.

[0067] Regarding FIGS, **3**A-**3**B, in some embodiments (not shown), each of corresponding HBMs **304**A-**304**D includes sensing unit **235**D of FIG. **2**D rather than sensing unit **235**C of FIG. **2**C. It is to be recalled that sensing unit **235**D includes temperature sensor **236** but does not include threshold voltage sensor **238**. Accordingly, in such embodiments, each of corresponding HBMs **304**A-**304**D (see FIGS. **4**B, **4**C and **4**F) is arranged so that DDVFS device **340** receives only temperature T from temperature sensor **236**.

[0068] Regarding FIGS, **3**A-**3**B, in some embodiments (not shown), each of corresponding HBMs **304**A-**304**D includes sensing unit **235**E of FIG. **2**E rather than sensing unit **235**C of FIG. **2**C. It is to be recalled that sensing unit **235**E includes threshold voltage sensor **238** but does not include temperature sensor **236**.. Accordingly, in such embodiments, each of corresponding HBMs **304**A-**304**D (see FIGS. **4**D, **4**E and **4**G) is arranged so that DDVFS device **340** receives only threshold voltage Vt from temperature sensor **236**.

[0069] FIGS. **4**A-**4**G are block diagrams of corresponding DDVFS devices **442**A-**442**G, in accordance with some embodiments.

[0070] In FIGS. **4**A-**4**G, each of corresponding DDVFS devices **442**A-**442**G is an example of DDVFS device **340** of FIGS. **3**A-**3**D.

[0071] In FIG. **4**A, DDVFS device **442**A includes: a temperature comparator **444**; a threshold voltage comparator **446**; a controller **448**A; a clock generator **454**; and a PS voltage (PSV) generator **456**. Controller **448**A includes a clock adjuster **450** and a PS voltage adjuster **452**.

[0072] Temperature comparator **444** is configured to receive temperature T from temperature sensor **236** of sensing unit **235**C of FIGS. **3**A-**3**D, and a signal MAX_temp representing a maximum permissible temperature for any instance of memory cell MC of HBMs **304**A-**304**D of corresponding FIGS. **3**A-**3**D. Temperature comparator **444** is further configured: to make a first comparison by comparing temperature T against temperature-reference MAX_temp; and to generate a first output signal representing the first temperature-comparison results, where the first output signal indicates which of temperature T and temperature-reference MAX_temp is larger.

[0073] Temperature comparator **444** is further configured to receive a signal MOD_temp representing a moderate temperature for any instance of memory cell MC of HBMs **304**A-**304**D of corresponding FIGS. **3**A-**3**D. In a circumstance in which temperature T is less than MAX_temp, such that a rise in temperature T could be safely accommodated by instances of memory cell MC in

FIGS. **3**A-**3**D which receive the corresponding local clock signal and/or the local PS voltage, in general, performance of the of instances of memory cell MC which receive the corresponding local clock signal and/or the local PS voltage is improved by DDVFS **340** increasing a frequency of the corresponding local clock signal and/or increasing a value of the corresponding local PS voltage. Accordingly, temperature comparator **444** is further configured: to make a second comparison by comparing temperature T against temperature-reference MOD_temp; and to generate a second output signal representing the second temperature-comparison results, where the second output signal indicates which of temperature T and temperature-reference MOD_temp is larger.

[0074] Threshold voltage comparator **446** is configured to receive threshold voltage Vt from threshold voltage detector **238** of sensing unit **235**C of FIGS. **3**A-**3**D, and a signal MIN_Vt representing a minimum permissible threshold voltage for any instance of transistor FET**1**, and thus any instance of memory cell MC, of HBMs **304**A-**304**D of corresponding FIGS. **3**A-**3**D.

[0075] Threshold voltage comparator **446** is further configured: to make a comparison by comparing threshold voltage Vt against voltage-reference MIN_Vt; and to generate an output signal representing the voltage-comparison results, where the output signal indicates which of threshold voltage Vt and voltage-reference MIN_Vt is larger.

[0076] In FIG. **4**A, controller **448**A is configured to receive the first and second temperature-comparison results from temperature comparator **444** and the voltage-comparison results from threshold voltage comparator **446**. More particularly, each of clock adjuster **450** and PS voltage adjuster **452** of controller **448**A is configured to receive the first and second temperature-comparison results from temperature comparator **444** and the voltage-comparison results from threshold voltage comparator **446**.

[0077] Clock adjuster **450**, and thus controller **448**A, is configured to cause the frequency of a local clock signal CLK_local to be increased or decreased by CLK generator **454** based on the first and second temperature-comparison results from temperature comparator **444** and/or the voltage-comparison results from threshold voltage comparator **446**.

[0078] Local clock signal CLK_local is an example of CLK_bnk of FIG. **3**A, CLK_grp of FIG. **3**B, CLK_chan of FIG. **3**C, or CLK_core of FIG. **3**D, or the like. M ore particularly, Clock adjuster **450** is configured to generate control signals CLK_up and CLK_down and thereby control CLK generator **454** correspondingly to increase or decrease the frequency of local clock signal CLK_local.

[0079] Signals CLK_up and CKL_down have mutually exclusive states. In some embodiments, when control signal CLK_up is in a high-state (which indicates that the frequency of local clock signal CLK_local is to be increased), then control signal CLK_down is in a low-state (which indicates that the frequency of local clock signal CLK_local is not to be decreased). Conversely, when control signal CLK_up is in the low-state (which indicates that the frequency of local clock signal CLK_local is not to be increased), then control signal CLK_down is in the high-state (which indicates that the frequency of local clock signal CLK_local is to be decreased). In some embodiments (not shown), clock adjuster **450** outputs a single control signal to CLK generator **454** rather than control signals CLK_up and CLK_down. In such embodiments, the single control signal has states which correspondingly indicate whether the frequency of the local clock signal CLK_local is to be increased or decreased.

[0080] In some embodiments, when the first temperature-comparison results indicate that temperature T is above MAX_temp, then clock adjuster **450** outputs control signal CLK_up in the low-state and control signal CLK_down in the high-state. Such a combination of states of control signals CLK_up and CLK_down, causes CLK generator **454** to alter, i.e., to decrease, the frequency of the local clock signal CLK_local.

[0081] In some embodiments, when the first voltage-comparison results indicate that threshold voltage is below MIN_Vt, then clock adjuster **450** outputs control signal CLK_up in the low-state and control signal CLK_down in the high-state. Such a combination of states of control signals

CLK_up and CLK_down, causes CLK generator **454** to alter, i.e., to decrease, the value of the local PS voltage PSV_local.

[0082] In some embodiments, when the first temperature-comparison results indicate that temperature T is below MAX_temp and the second temperature-comparison results indicate that temperature T is below MOD_temp, then clock adjuster **450** outputs control signal CLK_up in the high-state and control signal CLK_down in the low-state. Such a combination of states of control signals CLK_up and CLK_down, causes CLK generator **454** to alter, i.e., to increase, the frequency of the local clock signal CLK_local up to a permissible maximum value.

[0083] PS voltage adjuster **452**, and thus controller **448**A, is configured to cause the value of a local PS voltage PSV_local to be increased or decreased by PSV generator **456** based on the first and second temperature-comparison results from temperature comparator **444** and/or the voltage-comparison results from threshold voltage comparator **446**. Local PS voltage PSV_local is an example of PSV_bnk of FIG. **3**A, PSV_grp of FIG. **3**B, PSV_chan of FIG. **3**C, or PSV_core of FIG. **3**D, or the like. More particularly, PS voltage adjuster **452** is configured to generate control signals PSV_up and PSV_down and thereby control PSV generator **456** correspondingly to increase or decrease the value of local PS voltage PSV _local.

[0084] Signals PSV_up and PSV_down have mutually exclusive states. In some embodiments, when signal PSV_up is in a high-state (which indicates that the value of local PS voltage PSV_local is to be increased), then signal PSV_down is in a low-state (which indicates that the value of local PS voltage PSV_local is not to be decreased). Conversely, when signal PSV_down is in the high-state (which indicates that the value of local PS voltage PSV_local is to be decreased), then signal PSV_up is in the low-state (which indicates that the value of local PS voltage PSV_local is not to be increased). In some embodiments (not shown) PS voltage adjuster **452** outputs a single control signal to PSV generator **456** rather than control signals PSV_up and PSV_down. In such embodiments, the single control signal has states which correspondingly indicate whether the value of the local PS voltage PSV_local is to be increased or decreased.

[0085] In some embodiments, when the first voltage-comparison results indicate that threshold voltage Vt is below MIN_Vt, then PSV adjuster **452** outputs control signal PSV_up in the low-state and control signal PSV_down in the high-state. Such a combination of states of control signals PSV_up and PSV_down, causes PSV generator **456** to alter, i.e., to decrease, the value of the local PS voltage PSV_local.

[0086] In some embodiments, when the first temperature-comparison results indicate that temperature T is above MAX_temp, then PSV adjuster **452** outputs control signal PSV_up in the low-state and control signal PSV_down in the high-state. Such a combination of states of control signals PSV_up and PSV_down, causes PSV generator **456** to alter, i.e., to decrease, the value of the local PS voltage PSV_local.

[0087] Optionally, in some embodiments, DDVFS **442**A is configured with a feedback loop (shown as a phantom line, i.e., a dashed line) which feeds the frequency, ω, of local clock signal CLK_local back to clock adjuster **450**. In such embodiments, clock adjuster **450**, and thus controller **448**A, is configured to cause the frequency of a local clock signal CLK_local to be increased or decreased by CLK generator **454** based in part on (A) the feedback value of local clock signal CLK_local and (B) the first and second temperature-comparison results from temperature comparator **444** and/or the voltage-comparison result from threshold voltage comparator **446**.

[0088] Optionally, in some embodiments, DDVFS **442**A is configured with a feedback loop (shown as a phantom line, i.e., a dashed line) which feeds the value of PSV_local back to PS voltage adjuster **452**. In such embodiments, PS voltage adjuster **452**, and thus controller **448**A, is configured to cause the value of a local PS voltage PSV_local to be increased or decreased by PSV generator **456** based in part on (A) the feedback value of local PSV voltage PSV_local and (B) the first and second temperature-comparison results from temperature comparator **444** and/or the

voltage-comparison result from threshold voltage comparator **446**.

[0089] In some embodiments, DDVFS device **442**A of FIG. **4**A includes temperature comparator **444** but does not include threshold voltage comparator **446**. Such embodiments are shown in FIGS. **4**B, **4**C and **4**F as corresponding DDVFS devices **442**B, **442**C and **442**F.

[0090] In some embodiments, DDVFS device **442**B of FIG. **4**B, and more particularly, controller **445**B, includes clock adjuster **450** but does not include PS voltage adjuster **452**. Such embodiments are shown as DDVFS device **442**C of FIG. **4**C. In some embodiments, DDVFS device **442**B of FIG. **4**B, and more particularly, controller **445**B, includes PS voltage adjuster **452** but does not include clock adjuster **450**. Such embodiments are shown as DDVFS device **442**F of FIG. **4**F.

[0091] In some embodiments, DDVFS device **442**A of FIG. **4**A includes threshold voltage comparator **446** but does not include temperature comparator **444**. Such embodiments are shown in FIGS. **4**D, **4**E and **4**G as corresponding DDVFS devices **442**D, **442**E and **442**G.

[0092] In some embodiments, DDVFS device **442**D of FIG. **4**D, and more particularly, controller **445**D, includes PS voltage adjuster **452** but does not include clock adjuster **450**. Such embodiments are shown as DDVFS device **442**E of FIG. **4**E. In some embodiments, DDVFS device **442**D of FIG. **4**D, and more particularly, controller **445**D, includes clock adjuster **450** but does not include PS voltage adjuster **452**. Such embodiments are shown as DDVFS device **442**G of FIG. **4**G.

[0093] FIGS. **5**A-**5**B are circuit diagrams of corresponding temperature sensors **522**A-**522**B, in accordance with some embodiments.

[0094] Each of temperature sensors **522**A-**522**B includes a bias current generator **558**, a bipolar junction transistor (BJT) and a readout circuit **560**. Bias current generator **558** is electrically coupled to a node nd**1**. Details of typical bias current generators are found, e.g., in U.S. Pat. No. 8,183,910, granted May 22, 2012, or U.S. Pat. No. 9,166,067, granted Oct. 20, 2015, the entireties of each of which are hereby incorporated by reference. Readout circuit **560** is electrically coupled to node nd**1** and is configured to generate temperature T. In some embodiments, readout circuit **560** includes a buffer circuit.

[0095] In FIG. **5**A, transistor BJT**1** is an NPN type of BJT which is electrically coupled between node nd**1** and a low system voltage, e.g., ground, as contrasted with a high system voltage, VDD. In some embodiments, ground is VSS. Transistor BJT**1** has a diode configuration in which the base terminal of BJT**1** is electrically coupled to the collector terminal, i.e., to node nd**1**.

[0096] In FIG. **5**B, transistor BJT**2** is a PNP type of BJT which is electrically coupled between node nd**1** and ground. Transistor BJ T**2** has a diode configuration such that the base terminal of BJT**2** is electrically coupled to the emitter terminal, i.e., to ground.

[0097] FIGS. **6**A-**6**C are circuit diagrams of corresponding threshold voltage detectors **622**A-**622**C, in accordance with some embodiments.

[0098] Each of threshold voltage detectors **622**A-**622**C includes a proportional to absolute temperature (PTAT) current generator **564** and one or more field-effect transistors (FETs). PTAT current generator **564** is electrically coupled to a node nd**2**. In some embodiments (not shown), each of voltage detectors **622**A-**622**C includes a complementary to absolute temperature (CTAT) circuit. Details of typical PTAT current generators are found, e.g., in U.S. Pat. No. 8,183,910, granted May 22, 2012, or U.S. Pat. No. 9,166,067, granted Oct. 20, 2015, the entireties of each of which are hereby incorporated by reference.

[0099] In FIG. **6**A, threshold voltage detector **638**A includes a transistor FET**2**, which is an N-type FET transistor, and which is electrically coupled between node nd**2** and a low system voltage, e.g., ground, as contrasted with a high system voltage, VDD. In some embodiments, ground is VSS. Transistor FET**2** has a diode configuration in which the gate terminal of FET**2** is electrically coupled to the drain terminal, i.e., to node nd**2**. The signal on node nd**2** represents threshold voltage Vt.

[0100] In FIG. **6**B, threshold voltage detector **638**B includes a transistor FET **3**, which is a P-type FET transistor, and which is electrically coupled between node nd**2** and ground. Transistor FET**3**

has a diode configuration in which the gate terminal of FET**3** is electrically coupled to the drain terminal, i.e., to ground. The signal on node nd**2** represents threshold voltage Vt.

[0101] Regarding FIG. **6**C, threshold voltage detector **638**C is a combination of threshold voltage detectors **638***a* and **638**B. More particularly, threshold voltage detector **638**C includes both FET**2** of FIG. **6**A and FET**3** of FIG. **6**B. In FIG. **6**C, diode-configured FET**2** is electrically coupled between node nd**2** and a node nd**3**, and diode-configured FET**3** is electrically coupled between node nd**3** and ground. The signal on node nd**2** represents threshold voltage Vt.

[0102] FIGS. **7**A-**7**E are corresponding flowcharts **700**A-**700**E, in accordance with some embodiments.

[0103] More particularly, flowcharts **700**A-**700**E are corresponding methods of differentiated thermal throttling of a high bandwidth memory (HBM). Examples of an HBM include HBM **204** of FIG. **2**B, HBMs **204**A-**304**D of corresponding FIGS. **3**A-**3**D, or the like. Flowchart **700**A includes block **702**A, **704**A and **706**A.

[0104] At block **702**A, from each of first and second sensing units which are arranged correspondingly in core dies, first and second instances of a temperature signal and/or first and second instances of a threshold voltage signal are received. The first and second instance of the temperature signal and the first and second instances of the threshold voltage signal are associated with first and second transistors in corresponding first and second memory cells of corresponding first and second banks of corresponding core dies.

[0105] Regarding block **702**A, examples of the first and second sensing units are corresponding instances of sensing unit **235**C of FIG. **2**C, or the like, where sensing unit **235**C is an example of each instance of each of sensing units S**0**-S**7** in FIG. **2**B. Examples of the first and second temperature signals are corresponding instances of temperature T generated by corresponding instances of temperature sensor **236** of FIG. **2**C which are in corresponding instances of sensing unit **235**C, or the like. Examples of the first and second threshold voltage signals are corresponding instances of threshold voltage Vt generated by corresponding instances of threshold voltage detector **238** of FIG. **2**C in corresponding instances of sensing unit **235**C, or the like. Examples of the first and second transistors in corresponding first and second memory cells of corresponding first and second banks of corresponding core dies are first and second instances of transistor FET**1** in corresponding first and second instances of memory cell MC of corresponding first and second ones of banks B**0**-B**7** of corresponding ones of core dies **218**(**0**)-**218**(**3**) of FIG. **2**B. From block **702**A, flow proceeds to block **704**A.

[0106] At block **704**A, a first clock and/or a first PS voltage of one or more memory cells in a first set is/are adjusted based correspondingly on the first temperature (see block **702**A) and/or the first threshold voltage (see block **702**A). The first set is comprised of corresponding one or more banks of the HBM. Such adjustment thereby changes corresponding temperatures in the one or more memory cells of the first set. In some embodiments, both of the first clock and the first PS voltage of the one or more memory cells in the first set are adjusted. In some embodiments, the first clock and/or the first PS voltage is/are adjusted based on the first temperature and the first threshold voltage. In some embodiments, the first clock and/or the first PS voltage is/are adjusted based on only the first temperature. In some embodiments, the first clock and/or the first PS voltage is/are adjusted based on only the first threshold voltage.

[0107] Regarding block **704**A, examples of adjusting the first clock include decreasing or increasing a frequency of the first clock, or the like. An example of adjusting the first PS voltage includes decreasing a value/magnitude of the first PS voltage, or the like. Examples of changing a temperature in a memory cell include decreasing or increasing a temperature of the memory cell. Examples of the first clock include a first instance of local clock signal CLK_local in each of FIGS. **4**A-**4**D, local clock signal CLK_bnk in FIG. **3**A, local clock signal CLK_grp in FIG. **3**B, local clock signal CLK_chan in FIG. **3**C, local clock signal CLK_core in FIG. **3**D, or the like. Examples of the first PS voltage include a first instance of local PS voltage PSV_local in each of FIGS. **4**A-

4D, local PS voltage PSV_bnk in FIG. **3**A, local PS voltage PSV_grp in FIG. **3**B, local PS voltage PSV_chan in FIG. **3**C, local PS voltage PSV_core in FIG. **3**D, or the like. From block **704**A, flow proceeds to block **706**A.

[0108] At block **706**A, a second clock or a second PS voltage of one or more memory cells in a second set is/are adjusted based correspondingly on the second temperature (see block **702**A) and/or the second threshold voltage (see block **702**A). The second set is comprised of corresponding one or more banks of corresponding of the HBM. Such adjustment thereby changes corresponding temperatures in the one or more memory cells of the second set. In some embodiments, in terms of membership, the first and second sets are non-overlapping of each other. In some embodiments, the second clock and/or the second PS voltage is/are adjusted based on the second temperature and the second threshold voltage. In some embodiments, the second clock and/or the second PS voltage is/are adjusted based on only the second temperature. In some embodiments, the second clock and/or the second PS voltage is/are adjusted based on only the second threshold voltage.

[0109] Regarding block **706**A, examples of adjusting the second clock include decreasing or increasing a frequency of the second clock, or the like. An example of adjusting the second PS voltage includes decreasing a value/magnitude of the second PS voltage, or the like. Examples of changing a temperature in a memory cell include decreasing or increasing a temperature of the memory cell.

[0110] Examples of the second clock include a second instance of local clock signal CLK_local in each of FIGS. **4**A-**4**D, local clock signal CLK_bnk in FIG. **3**A, local clock signal CLK_grp in FIG. **3**B, local clock signal CLK_chan in FIG. **3**C, local clock signal CLK_core in FIG. **3**D, or the like. Examples of the second PS voltage include a second instance of local PS voltage PSV_local in each of FIGS. **4**A-**4**D, local PS voltage PSV_bnk in FIG. **3**A, local PS voltage PSV_grp in FIG. **3**B, local PS voltage PSV_chan in FIG. **3**C, local PS voltage PSV_core in FIG. **3**D, or the like.

[0111] Turning now to FIG. **7**B, flowchart **700**B is a bank-wide type of differentiated thermal throttling of an HBM. An example of the HBM is HBM **304**A of FIG. **3**A.

[0112] In FIG. **7**B, flowchart **700**B includes blocks **702**B, **704**B and **706**B. Flowchart **700**B assumes a scenario in which a first set of one or more of the memory cells in the HBM is contained within at least a first lot of one or more of the banks of the HBM, and a second set of one or more of the memory cells is contained within at least a second lot of one of the banks of the HBM. In terms of membership, in some embodiments, the first and second lots are non-overlapping of each other.

[0113] Block **702**B is similar to block **702**A, the latter having been discussed above in the context of FIG. **7**A. At block **702**B, from each of first and second sensing units, first and second instances of a temperature signal and/or first and second instances of a threshold voltage signal are received. The first and second instance of the temperature signal and the first and second instances of the threshold voltage signal are associated with first and second transistors in corresponding first and second memory cells of the corresponding first and second lots of corresponding one or more banks. From block **702**B, flow proceeds to block **704**B.

[0114] Block **704**B is similar to block **704**A, the latter having been discussed above in the context of FIG. **7**A. At block **704**B, a first clock and/or a first PS voltage of one or more memory cells in the first bank is/are adjusted based correspondingly on the first temperature (see block **702**B) and/or the first threshold voltage (see block **702**B). Such adjustment thereby achieves bank-wide adjustment of corresponding temperatures in the one or more memory cells of the first lot. In some embodiments, both of the first clock and the first PS voltage of the first bank are adjusted. An example of bank-wide adjustment is the arrangement of DDVFS device **340** in HBM **304**A of FIG. **3**A. In some embodiments, the first clock and/or the first PS voltage is/are adjusted based on the first temperature and the first threshold voltage. In some embodiments, the first clock and/or the first PS voltage is/are adjusted based on only the first temperature. In some embodiments, the first

clock and/or the first PS voltage is/are adjusted based on only the first threshold voltage. From block **704**B, flow proceeds to block **706**B.

[0115] Block **706**B is similar to block **706**A, the latter having been discussed above in the context of FIG. **7**A. At block **706**B, a second clock and/or a second PS voltage of one or more memory cells in the second bank is/are adjusted based correspondingly on the second temperature (see block **702**B) and/or the second threshold voltage (see block **702**B). Such adjustment thereby achieves bank-wide adjustment of corresponding temperatures in the one or more memory cells of the second lot. In some embodiments, both of the second clock and the second PS voltage of the one or more memory cells in the second bank are adjusted. An example of bank-wide adjustment is the arrangement of DDVFS device **340** in HBM **304**A of FIG. **3**A. In some embodiments, the second clock and/or the second PS voltage is/are adjusted based on the second temperature and the second threshold voltage. In some embodiments, the second clock and/or the second PS voltage is/are adjusted based on only the second temperature. In some embodiments, the second clock and/or the second PS voltage is/are adjusted based on only the second threshold voltage.

[0116] Turning now to FIG. **7**C, flowchart **700**C is a group-wide type of differentiated thermal throttling of an HBM. An example of the HBM is HBM **304**B of FIG. **3**B.

[0117] In FIG. **7**C, flowchart **700**C includes blocks **702**C, **704**C and **706**C. Flowchart **700**C assumes a scenario in which a first set of one or more of the memory cells in the HBM is contained within at least a first batch of one or more groups of the HBM, and a second set of one or more of the memory cells is contained within at least a second batch of one or more groups of the HBM. In terms of membership, in some embodiments, the first and second batches are non-overlapping of each other.

[0118] Block **702**C is similar to block **702**B, the latter having been discussed above in the context of FIG. **7**B. At block **702**C, from each of first and second sensing units, first and second instances of a temperature signal and/or first and second instances of a threshold voltage signal are received. The first and second instance of the temperature signal and the first and second instances of the threshold voltage signal are associated with first and second transistors in corresponding first and second memory cells of the corresponding first and second batches. From block **702**C, flow proceeds to block **704**C.

[0119] Block **704**C is similar to block **704**B, the latter having been discussed above in the context of FIG. **7**B. At block **704**C, a first clock and/or a first PS voltage of one or more memory cells in the first group is/are adjusted based correspondingly on the first temperature (see block **702**C) and/or the first threshold voltage (see block **702**C). Such adjustment thereby achieves group-wide adjustment of corresponding temperatures in the one or more memory cells of the first batch. In some embodiments, both of the first clock and the first PS voltage of the first group are adjusted. An example of group-wide adjustment is the arrangement of DDVFS device **340** in HBM **304**B of FIG. **3**B. In some embodiments, the first clock and/or the first PS voltage is/are adjusted based on the first temperature and the first threshold voltage. In some embodiments, the first clock and/or the first PS voltage is/are adjusted based on only the first temperature. In some embodiments, the first clock and/or the first PS voltage is/are adjusted based on only the first threshold voltage. From block **704**C, flow proceeds to block **706**C.

[0120] Block **706**C is similar to block **706**B, the latter having been discussed above in the context of FIG. **7**B. At block **706**C, a second clock and/or a second PS voltage of one or more memory cells in the second group is/are adjusted based correspondingly on the second temperature (see block **702**C) and/or the second threshold voltage (see block **702**C). Such adjustment thereby achieves group-wide adjustment of corresponding temperatures in the one or more memory cells of the second batch. In some embodiments, both of the second clock and the second PS voltage of the one or more memory cells in the second group are adjusted. An example of group-wide adjustment is the arrangement of DDVFS device **340** in HBM **304**B of FIG. **3**B. In some embodiments, the second clock and/or the second PS voltage is/are adjusted based on the second temperature and the

second threshold voltage. In some embodiments, the second clock and/or the second PS voltage is/are adjusted based on only the second temperature. In some embodiments, the second clock and/or the second PS voltage is/are adjusted based on only the second threshold voltage.

[0121] Turning now to FIG. **7**D, flowchart **700**D is a channel-wide type of differentiated thermal throttling of an HBM. An example of the HBM is HBM **304**C of FIG. **3**C.

[0122] In FIG. **7**D, flowchart **700**D includes blocks **702**D, **704**D and **706**D. Flowchart **700**D assumes a scenario in which a first set of one or more of the memory cells in the HBM is contained within at least a first bundle of one or more channels of the HBM, and a second set of one or more of the memory cells is contained within at least a second bundles of one or more channels of the HBM. In terms of membership, in some embodiments, the first and second bundles are non-overlapping of each other.

[0123] Block **702**D is similar to block **702**B, the latter having been discussed above in the context of FIG. **7**C. At block **702**D, from each of first and second sensing units, first and second instances of a temperature signal and/or first and second instances of a threshold voltage signal are received. The first and second instance of the temperature signal and the first and second instances of the threshold voltage signal are associated with first and second transistors in corresponding first and second memory cells of the corresponding first and second bundles of corresponding one or more channels. From block **702**D, flow proceeds to block **704**D.

[0124] Block **704**D is similar to block **704**B, the latter having been discussed above in the context of FIG. **7**C. At block **704**D, a first clock and/or a first PS voltage of one or more memory cells in the first channel is/are adjusted based correspondingly on the first temperature (see block **702**D) and/or the first threshold voltage (see block **702**D). Such adjustment thereby achieves channel-wide adjustment of corresponding temperatures in the one or more memory cells of the first bundle. In some embodiments, both of the first clock and the first PS voltage of the first channel are adjusted. An example of channel-wide adjustment is the arrangement of DDVFS device **340** in HBM **304**C of FIG. **3**C. In some embodiments, the first clock and/or the first PS voltage is/are adjusted based on the first temperature and the first threshold voltage. In some embodiments, the first clock and/or the first PS voltage is/are adjusted based on only the first temperature. In some embodiments, the first clock and/or the first PS voltage is/are adjusted based on only the first threshold voltage. From block **704**D, flow proceeds to block **706**D.

[0125] Block **706**D is similar to block **706**B, the latter having been discussed above in the context of FIG. **7**C. At block **706**D, a second clock and/or a second PS voltage of one or more memory cells in the second channel is/are adjusted based correspondingly on the second temperature (see block **702**D) and/or the second threshold voltage (see block **702**D). Such adjustment thereby achieves channel-wide adjustment of corresponding temperatures in the one or more memory cells of the second bundle. In some embodiments, both of the second clock and the second PS voltage of the one or more memory cells in the second channel are adjusted. An example of channel-wide adjustment is the arrangement of DDVFS device **340** in HBM **304**C of FIG. **3**C. In some embodiments, the second clock and/or the second PS voltage is/are adjusted based on the second temperature and the second threshold voltage. In some embodiments, the second clock and/or the second PS voltage is/are adjusted based on only the second temperature. In some embodiments, the second clock and/or the second PS voltage is/are adjusted based on only the second threshold voltage.

[0126] Turning now to FIG. **7**E, flowchart **700**E is a core-die-wide type of differentiated thermal throttling of an HBM. An example of the HBM is HBM **304**D of FIG. **3**D.

[0127] In FIG. **7**E, flowchart **700**E includes blocks **702**E, **704**E and **706**E. Flowchart **700**E assumes a scenario in which a first set of one or more of the memory cells in the HBM is contained within at least a first collection of one or more core-dies of the HBM, and a second set of one or more of the memory cells is contained within at least a second collection of one or more core-dies of the HBM. In terms of membership, in some embodiments, the first and second collections are non-

overlapping of each other.

[0128] Block **702**E is similar to block **702**B, the latter having been discussed above in the context of FIG. **7**D. At block **702**E, from each of first and second sensing units, first and second instances of a temperature signal and/or first and second instances of a threshold voltage signal are received. The first and second instance of the temperature signal and the first and second instances of the threshold voltage signal are associated with first and second transistors in corresponding first and second memory cells of the corresponding first and second collections of corresponding one or more core-dies. From block **702**E, flow proceeds to block **704**E.

[0129] Block **704**E is similar to block **704**B, the latter having been discussed above in the context of FIG. **7**D. At block **704**E, a first clock and/or a first PS voltage of one or more memory cells in the first core-die is/are adjusted based correspondingly on the first temperature (see block **702**E) and/or the first threshold voltage (see block **702**E). Such adjustment thereby achieves core-die-wide adjustment of corresponding temperatures in the one or more memory cells of the first collection. In some embodiments, both of the first clock and the first PS voltage of the first core-die are adjusted. An example of core-die-wide adjustment is the arrangement of DDVFS device **340** in HBM **304**D of FIG. **3**D. In some embodiments, the first clock and/or the first PS voltage is/are adjusted based on the first temperature and the first threshold voltage. In some embodiments, the first clock and/or the first PS voltage is/are adjusted based on only the first temperature. In some embodiments, the first clock and/or the first PS voltage is/are adjusted based on only the first threshold voltage. From block **704**E, flow proceeds to block **706**E.

[0130] Block **706**E is similar to block **706**B, the latter having been discussed above in the context of FIG. **7**D. At block **706**E, a second clock and/or a second PS voltage of one or more memory cells in the second core-die is/are adjusted based correspondingly on the second temperature (see block **702**E) and/or the second threshold voltage (see block **702**E). Such adjustment thereby achieves core-die-wide adjustment of corresponding temperatures in the one or more memory cells of the second collection. In some embodiments, both of the second clock and the second PS voltage of the one or more memory cells in the second core-die are adjusted. An example of core-die-wide adjustment is the arrangement of DDVFS device **340** in HBM **304**D of FIG. **3**D. In some embodiments, the second clock and/or the second PS voltage is/are adjusted based on the second temperature and the second threshold voltage. In some embodiments, the second clock and/or the second PS voltage is/are adjusted based on only the second temperature. In some embodiments, the second clock and/or the second PS voltage is/are adjusted based on only the second threshold voltage.

[0131] FIG. **7**F is a flowchart showing blocks **704**A and **706**A in more detail, in accordance with some embodiments.

[0132] In FIG. **7**F, block **704**A includes blocks **730**-**734** and block **706**A includes blocks **736**-**740**.

[0133] At block **730**, the first temperature is compared against a first temperature-reference which yields first temperature-comparison results. The first temperature-reference represents a maximum temperature. An example of the first temperature-reference is temperature-reference MAX_temp of FIGS. **4**A-**4**C and **4**F, or the like. Examples of comparing the first temperature against the first temperature-reference are the comparisons performed by temperature comparator **444** of FIGS. **4**A-**4**C and **4**F, or the like. From block **730**, flow proceeds to block **732**.

[0134] At block **732**, a frequency of the first clock is altered based on the first temperature-comparison results (see block **730**). An example of such alteration is reducing the frequency of the first clock, e.g., as in the coordinated operation of clock adjuster **450** and CLK generator **454** of FIGS. **4**A-**4**C, or the like. From block **732**, flow proceeds to block **734**.

[0135] At block **734**, a value of the first PS voltage is altered based on the first temperature-comparison results (see block **730**). An example of such alteration is reducing the value/magnitude of the first PS voltage, e.g., as in the coordinated operation of PSV adjuster **452** and PSV generator **456** of FIGS. **4**A-**4**B and **4**F, or the like. From block **734**, flow proceeds to block **736** in block

706A.

[0136] At block **736**, the second temperature is compared against the first temperature-reference which yields second temperature-comparison results. Examples of comparing the second temperature against the second temperature-reference are the comparisons performed by temperature comparator **444** of FIGS. **4**A-**4**C, or the like. From block **736**, flow proceeds to block **738**.

[0137] At block **738**, a frequency of the second clock is altered based on the first temperature-comparison results (see block **736**). An example of such alteration is reducing the frequency of the second clock, e.g., as in the coordinated operation of clock adjuster **450** and CLK generator **454** of FIGS. **4**A-**4**C, or the like. From block **738**, flow proceeds to block **740**.

[0138] At block **740**, a value of the second PS voltage is altered based on the second temperature-comparison results (see block **736**). An example of such alteration is reducing the value/magnitude of the second PS voltage, e.g., as in the coordinated operation of PSV adjuster **452** and PSV generator **456** of FIGS. **4**A-**4**B and **4**F, or the like.

[0139] FIG. **7**G is a flowchart showing blocks **704**A and **706**A in more detail, in accordance with some embodiments.

[0140] In FIG. **7**G, block **704**A includes blocks **750**-**754** and block **706**A includes blocks **756**-**760**.

[0141] At block **750**, the first threshold voltage is compared against a first voltage-reference which yields first voltage-comparison results. The first voltage-reference represents a minimum threshold voltage. An example of the first voltage-reference is voltage-reference MIN_Vt of FIGS. **4**A, **4**D-**4**E and **4**G, or the like. Examples of comparing the first voltage against the first voltage-reference are the comparisons performed by threshold voltage comparator **446** of FIGS. **4**A, **4**C-**4**D and **4**G, or the like. From block **750**, flow proceeds to block **752**.

[0142] At block **752**, a frequency of the first clock is altered based on the first voltage-comparison results (see block **750**). An example of such alteration is reducing the frequency of the first clock, e.g., as in the coordinated operation of clock adjuster **450** and CLK generator **454** of FIGS. **4**A, **4**D and **4**G, or the like. From block **752**, flow proceeds to block **754**.

[0143] At block **754**, a value of the first PS voltage is altered based on the first threshold voltage-comparison results (see block **750**). An example of such alteration is reducing the value/magnitude of the first PS voltage, e.g., as in the coordinated operation of PSV adjuster **452** and PSV generator **456** of FIGS. **4**A and **4**D-**4**E, or the like. From block **754**, flow proceeds to block **756** in block **706**A.

[0144] At block **756**, the second threshold voltage is compared against a first threshold voltage-reference which yields second threshold voltage-comparison results. Examples of comparing the second threshold voltage against the second threshold voltage reference are the comparisons performed by threshold voltage comparator **444** of FIGS. **4**A-**4**C and **4**F, or the like. From block **756**, flow proceeds to block **758**.

[0145] At block **758**, a frequency of the second clock is altered based on the second threshold voltage-comparison results (see block **756**). An example of such alteration is reducing the frequency of the second clock, e.g., as in the coordinated operation of clock adjuster **450** and CLK generator **454** of FIGS. **4**A, **4**D and **4**G, or the like. From block **758**, flow proceeds to block **760**.

[0146] At block **760**, a value of the second PS voltage is altered based on the second threshold voltage-comparison results (see block **756**). An example of such alteration is reducing the value/magnitude of the second PS voltage, e.g., as in the coordinated operation of PSV adjuster **452** and PSV generator **456** of FIGS. **4**A and **4**D-**4**E, or the like.

[0147] FIG. **7**H is a flowchart showing blocks **704**A and **706**A in more detail, in accordance with some embodiments.

[0148] In FIG. **7**H, block **704**A includes blocks **770**-**774** and block **706**A includes blocks **776**-**780**.

[0149] At block **770**, the first temperature is compared against a second temperature-reference which yields third temperature-comparison results. The second temperature-reference represents a

moderate temperature. An example of the second temperature-reference is temperature-reference MOD_temp of FIGS. **4**A-**4**C and **4**F, or the like. Examples of comparing the first temperature against the second temperature-reference are the comparisons performed by temperature comparator **444** of FIGS. **4**A-**4**C and **4**F, or the like. From block **770**, flow proceeds to block **772**.

[0150] At block **772**, a frequency of the first clock is altered based on the third temperature-comparison results (see block **770**). More specifically, the frequency of the first clock is increased at block **772**. An example of such alteration is increasing the frequency of the first clock, e.g., as in the coordinated operation of clock adjuster **450** and CLK generator **454** of FIGS. **4**A-**4**C, or the like. From block **772**, flow proceeds to block **774**.

[0151] At block **774**, a value of the first PS voltage is altered based on the third temperature-comparison results (see block **770**). More specifically, the value of the first PS voltage is increased at block **774**. An example of such alteration is increasing the value/magnitude of the first PS voltage, e.g., as in the coordinated operation of PSV adjuster **452** and PSV generator **456** of FIGS. **4**A-**4**B and **4**F, or the like. From block **774**, flow proceeds to block **776** in block **706**A.

[0152] At block **776**, the second temperature is compared against the second temperature-reference which yields fourth temperature-comparison results. Examples of comparing the second temperature against the second temperature-reference are the comparisons performed by temperature comparator **444** of FIGS. **4**A-**4**C and **4**F, or the like. From block **776**, flow proceeds to block **778**.

[0153] At block **778**, a frequency of the second clock is altered based on the fourth temperature-comparison results (see block **776**). More specifically, the frequency of the second clock is increased at block **778**. An example of such alteration is increasing the frequency of the second clock, e.g., as in the coordinated operation of clock adjuster **450** and CLK generator **454** of FIGS. **4**A-**4**C, or the like. From block **778**, flow proceeds to block **780**.

[0154] At block **780**, a value of the second PS voltage is altered based on the fourth temperature-comparison results (see block **776**). More specifically, the value of the second PS voltage is increased at block **780**. An example of such alteration is increasing the value/magnitude of the second PS voltage, e.g., as in the coordinated operation of PSV adjuster **452** and PSV generator **456** of FIGS. **4**A-**4**B and **4**F, or the like.

[0155] FIG. **7**I is a flowchart showing blocks **704**A and **706**A in more detail, in accordance with some embodiments.

[0156] In FIG. **7**I, block **704**A includes blocks **790-794** and block **706**A includes blocks **796-799**.

[0157] At block **790**, it is determined if all members, e.g., memory cells MC, of the first set are idle. If so(i.e., when all members of the first set are idle/asleep), then flow proceeds to block **792**. In some embodiments, the HBM, e.g., HBM **204** FIGS. **2**A-**2**B, is included in an internet-of-things (IoT) device. It is common for an IoT device to be idle for a majority of the time.

[0158] At block **792**, when all members of the first set are idle, a frequency of the first clock is reduced. An example of such alteration is reducing the frequency of the first clock, e.g., as in the coordinated operation of clock adjuster **450** and CLK generator **454** of FIGS. **4**A-**4**C, or the like. From block **792**, flow proceeds to block **794**.

[0159] At block **794**, a value of the first PS voltage is reduced. An example of such alteration is reducing the value/magnitude of the first PS voltage, e.g., as in the coordinated operation of PSV adjuster **452** and PSV generator **456** of FIGS. **4**A-**4**B and **4**F, or the like. From block **794**, flow proceeds to block **796** in block **706**A.

[0160] At block **796**, it is determined if all members, e.g., memory cells MC, of the second set are idle. If so(i.e., when all members of the second set are idle), then flow proceeds to block **798**.

[0161] At block **798**, a frequency of the second clock is reduced. An example of such alteration is reducing the frequency of the second clock, e.g., as in the coordinated operation of clock adjuster **450** and CLK generator **454** of FIGS. **4**A-**4**C, or the like. From block **798**, flow proceeds to block **799**.

[0162] At block **799**, a value of the second PS voltage is reduced. An example of such alteration is reducing the value/magnitude of the second PS voltage, e.g., as in the coordinated operation of PSV adjuster **452** and PSV generator **456** of FIGS. **4**A-**4**B and **4**F, or the like.

[0163] Alternatively, in some embodiments, block **792** is not included and instead flow proceeds from block **790** to block **794**. Alternatively, in some embodiments, block **794** is not included.

[0164] Alternatively, in some embodiments, block **798** is not included and instead flow proceeds from block **796** to block **799**. Alternatively, in some embodiments, block **799** is not included.

[0165] In some embodiments, the value of the first and second PS voltages is reduced to about zero. In such embodiments, while the first and second PS voltages are set to about zero, the corresponding devices receiving the first and second PS voltages enjoy a reduction in power consumption of about 99%.

[0166] In some embodiments, the value of the first and second PS voltages is reduced to a lowest workable voltage, i.e., a lowest voltage at which the HBM will operate correctly/reliably. In such embodiments, while the first and second PS voltages are set to the lowest workable voltage, the corresponding devices receiving the first and second PS voltages enjoy a reduction in power consumption of about 1%.

[0167] FIG. **8** is a flowchart of a method **800** of manufacturing a semiconductor device, in accordance with some embodiments.

[0168] Method **800** is implementable, for example, using EDA system **900** (FIG. **9**, discussed below) and an integrated circuit (IC) manufacturing system **1000** (FIG. **10**, discussed below), in accordance with some embodiments. Examples of a semiconductor device which can be manufactured according to method **800** include semiconductor device **100** of FIG. **1**, semiconductor devices corresponding to various ones of the layout diagrams disclosed herein, or the like.

[0169] In FIG. **8**, method **800** includes blocks **802-804**. At block **802**, a layout diagram is generated which, among other things, includes one or more of layout diagrams corresponding to semiconductor devices, e.g., FIG. **1**, or the like, where the semiconductor devices are implementations of systems such as the systems disclosed herein, or the like. Block **802** is implementable, for example, using EDA system **900** (FIG. **9**, discussed below), in accordance with some embodiments.

[0170] More particularly, block **802** includes generating shapes corresponding to structures in a semiconductor device resulting in a layout diagram in according to which the semiconductor device is manufacturable. From block **802**, flow proceeds to block **804**.

[0171] At block **804**, based on the layout diagram, perform at least one of (A) one or more photolithographic exposures are made or (B) one or more semiconductor masks are fabricated or (C) one or more components in a layer of a semiconductor device are fabricated. See discussion below of FIG. **10**.

[0172] FIG. **9** is a block diagram of an electronic design automation (EDA) system **900** in accordance with some embodiments.

[0173] In some embodiments, EDA system **900** includes an automatic placement and routing (APR) system. Methods described herein of designing layout diagrams represent wire routing arrangements, in accordance with one or more embodiments, are implementable, for example, using EDA system **900**, in accordance with some embodiments.

[0174] In some embodiments, EDA system **900** is a general purpose computing device including a hardware processor **902** and a non-transitory, computer-readable storage medium **904**. Storage medium **904**, amongst other things, is encoded with, i.e., stores, computer program code **906**, i.e., a set of executable instructions. Execution of instructions **906** by hardware processor **902** represents (at least in part) an EDA tool which implements a portion or all of the methods described herein in accordance with one or more embodiments (hereinafter, the noted processes and/or methods).

[0175] Processor **902** is electrically coupled to computer-readable storage medium **904** via a bus

908. Processor **902** is also electrically coupled to an I/O interface **910** by bus **908**. A network interface **912** is also electrically connected to processor **902** via bus **908**. Network interface **912** is connected to a network **914**, so that processor **902** and computer-readable storage medium **904** are capable of connecting to external elements via network **914**. Processor **902** is configured to execute computer program code **906** encoded in computer-readable storage medium **904** in order to cause system **900** to be usable for performing a portion or all of the noted processes and/or methods. In one or more embodiments, processor **902** is a central processing unit (CPU), a multi-processor, a distributed processing system, an application specific integrated circuit (ASIC), and/or a suitable processing unit.

[0176] In one or more embodiments, computer-readable storage medium **904** is an electronic, magnetic, optical, electromagnetic, infrared, and/or a semiconductor system (or apparatus or device). For example, computer-readable storage medium **904** includes a semiconductor or solid-state memory, a magnetic tape, a removable computer diskette, a random access memory (RAM), a read-only memory (ROM), a rigid magnetic disk, and/or an optical disk. In one or more embodiments using optical disks, computer-readable storage medium **904** includes a compact disk-read only memory (CD-ROM), a compact disk-read/write (CD-R/W), and/or a digital video disc (DVD).

[0177] In one or more embodiments, storage medium **904** stores computer program code **906** configured to cause system **900** (where such execution represents (at least in part) the EDA tool) to be usable for performing a portion or all of the noted processes and/or methods. In one or more embodiments, storage medium **904** also stores information which facilitates performing a portion or all of the noted processes and/or methods. In one or more embodiments, storage medium **904** stores library **907** of standard cells including such standard cells as disclosed herein. In one or more embodiments, storage medium **904** stores one or more layout diagrams **909** corresponding to one or more layout diagrams disclosed herein.

[0178] EDA system **900** includes I/O interface **910**. I/O interface **910** is coupled to external circuitry. In one or more embodiments, I/O interface **910** includes a keyboard, keypad, mouse, trackball, trackpad, touchscreen, and/or cursor direction keys for communicating information and commands to processor **902**.

[0179] EDA system **900** also includes network interface **912** coupled to processor **902**. Network interface **912** allows system **900** to communicate with network **914**, to which one or more other computer systems are connected. Network interface **912** includes wireless network interfaces such as BLUETOOTH, WIFI, WIMAX, GPRS, or WCDMA; or wired network interfaces such as ETHERNET, USB, or IEEE-1364. In one or more embodiments, a portion or all of noted processes and/or methods, is implemented in two or more systems **900**.

[0180] System **900** is configured to receive information through I/O interface **910**. The information received through I/O interface **910** includes one or more of instructions, data, design rules, libraries of standard cells, and/or other parameters for processing by processor **902**. The information is transferred to processor **902** via bus **908**. EDA system **900** is configured to receive information related to a UI through I/O interface **910**. The information is stored in computer-readable medium **904** as user interface (UI) **942**.

[0181] In some embodiments, a portion or all of the noted processes and/or methods is implemented as a standalone software application for execution by a processor. In some embodiments, a portion or all of the noted processes and/or methods is implemented as a software application that is a part of an additional software application. In some embodiments, a portion or all of the noted processes and/or methods is implemented as a plug-in to a software application. In some embodiments, at least one of the noted processes and/or methods is implemented as a software application that is a portion of an EDA tool. In some embodiments, a portion or all of the noted processes and/or methods is implemented as a software application that is used by EDA system **900**. In some embodiments, a layout diagram which includes standard cells is generated

using a tool such as VIRTUOSO® available from CADENCE DESIGN SYSTEMS, Inc., or another suitable layout-diagram-generating tool.

[0182] In some embodiments, the processes are realized as functions of a program stored in a non-transitory computer readable recording medium. Examples of a non-transitory computer readable recording medium include, but are not limited to, external/removable and/or internal/built-in storage or memory unit, e.g., one or more of an optical disk, such as a DVD, a magnetic disk, such as a hard disk, a semiconductor memory, such as a ROM, a RAM, a memory card, and the like.

[0183] FIG. **10** is a block diagram of an integrated circuit (IC) manufacturing system **1000**, and an IC manufacturing flow associated therewith, in accordance with some embodiments. In some embodiments, based on a layout diagram, at least one of (A) one or more semiconductor masks or (B) at least one component in a layer of a semiconductor integrated circuit is fabricated using manufacturing system **1000**.

[0184] In FIG. **10**, IC manufacturing system **1000** includes entities, such as a design house **1020**, a mask house **1030**, and an IC manufacturer/fabricator (fab) **1050**, that interact with one another in the design, development, and manufacturing cycles and/or services related to manufacturing an IC device **1060**. The entities in system **1000** are connected by a communications network. In some embodiments, the communications network is a single network. In some embodiments, the communications network is a variety of different networks, such as an intranet and the Internet. The communications network includes wired and/or wireless communication channels. Each entity interacts with one or more of the other entities and provides services to and/or receives services from one or more of the other entities. In some embodiments, two or more of design house **1020**, mask house **1030**, and IC fab **1050** is owned by a single larger company. In some embodiments, two or more of design house **1020**, mask house **1030**, and IC fab **1050** coexist in a common facility and use common resources.

[0185] Design house (or design team) **1020** generates an IC layout diagram **1022**. IC layout diagram **1022** includes various geometrical patterns designed for an IC device **1060**. The geometrical patterns correspond to patterns of metal, oxide, or semiconductor layers that make up the various components of IC device **1060** to be fabricated. The various layers combine to form various IC features. For example, a portion of IC layout diagram **1022** includes various IC features, such as an active region, gate electrode, source and drain, metal lines or vias of an interlayer interconnection, and openings for bonding pads, to be formed in a semiconductor substrate (such as a silicon wafer) and various material layers disposed on the semiconductor substrate. Design house **1020** implements a proper design procedure to form IC layout diagram **1022**. The design procedure includes one or more of logic design, physical design or place and route. IC layout diagram **1022** is presented in one or more data files having information of the geometrical patterns. For example, IC layout diagram **1022** can be expressed in a GDSII file format or DFII file format.

[0186] Mask house **1030** includes data preparation **1032** and mask fabrication **1044**. Mask house **1030** uses IC layout diagram **1022** to manufacture one or more masks **1045** to be used for fabricating the various layers of IC device **1060** according to IC layout diagram **1022**. Mask house **1030** performs mask data preparation **1032**, where IC layout diagram **1022** is translated into a representative data file (RDF). Mask data preparation **1032** provides the RDF to mask fabrication **1044**. Mask fabrication **1044** includes a mask writer. A mask writer converts the RDF to an image on a substrate, such as a mask (reticle) **1045** or a semiconductor wafer **1053**. The layout diagram **1022** is manipulated by mask data preparation **1032** to comply with particular characteristics of the mask writer and/or requirements of IC fab **1050**. In FIG. **10**, mask data preparation **1032** and mask fabrication **1044** are illustrated as separate elements. In some embodiments, mask data preparation **1032** and mask fabrication **1044** can be collectively referred to as mask data preparation.

[0187] In some embodiments, mask data preparation **1032** includes optical proximity correction (OPC) which uses lithography enhancement techniques to compensate for image errors, such as those that can arise from diffraction, interference, other process effects and the like. OPC adjusts IC

layout diagram **1022**. In some embodiments, mask data preparation **1032** includes further resolution enhancement techniques (RET), such as off-axis illumination, sub-resolution assist features, phase-shifting masks, other suitable techniques, and the like or combinations thereof. In some embodiments, inverse lithography technology (ILT) is also used, which treats OPC as an inverse imaging problem.

[0188] In some embodiments, mask data preparation **1032** includes a mask rule checker (MRC) that checks IC layout diagram **1022** that has undergone processes in OPC with a set of mask creation rules which contain certain geometric and/or connectivity restrictions to ensure sufficient margins, to account for variability in semiconductor manufacturing processes, and the like. In some embodiments, the MRC modifies IC layout diagram **1022** to compensate for photolithographic implementation effects during mask fabrication **1044**, which may undo part of the modifications performed by OPC in order to meet mask creation rules.

[0189] In some embodiments, mask data preparation **1032** includes lithography process checking (LPC) that simulates processing that will be implemented by IC fab **1050** to fabricate IC device **1060**. LPC simulates this processing based on IC layout diagram **1022** to create a simulated manufactured device, such as IC device **1060**. The processing parameters in LPC simulation can include parameters associated with various processes of the IC manufacturing cycle, parameters associated with tools used for manufacturing the IC, and/or other aspects of the manufacturing process. LPC takes into account various factors, such as aerial image contrast, depth of focus (DOF), mask error enhancement factor (MEEF), other suitable factors, and the like or combinations thereof. In some embodiments, after a simulated manufactured device has been created by LPC, if the simulated device is not close enough in shape to satisfy design rules, OPC and/or MRC are be repeated to further refine IC layout diagram **1022**.

[0190] It should be understood that the above description of mask data preparation **1032** has been simplified for the purposes of clarity. In some embodiments, data preparation **1032** includes additional features such as a logic operation (LOP) to modify IC layout diagram **1022** according to manufacturing rules. Additionally, the processes applied to IC layout diagram **1022** during data preparation **1032** may be executed in a variety of different orders.

[0191] After mask data preparation **1032** and during mask fabrication **1044**, a mask **1045** or a group of masks **1045** are fabricated based on the modified IC layout diagram **1022**. In some embodiments, mask fabrication **1044** includes performing one or more lithographic exposures based on IC layout diagram **1022**. In some embodiments, an electron-beam (e-beam) or a mechanism of multiple e-beams is used to form a pattern on a mask (photomask or reticle) **1045** based on the modified IC layout diagram **1022**. Mask **1045** can be formed in various technologies. In some embodiments, mask **1045** is formed using binary technology. In some embodiments, a mask pattern includes opaque regions and transparent regions. A radiation beam, such as an ultraviolet (UV) beam, used to expose the image sensitive material layer (e.g., photoresist) which has been coated on a wafer, is blocked by the opaque region and transmits through the transparent regions. In one example, a binary mask version of mask **1045** includes a transparent substrate (e.g., fused quartz) and an opaque material (e.g., chromium) coated in the opaque regions of the binary mask. In another example, mask **1045** is formed using a phase shift technology. In a phase shift mask (PSM) version of mask **1045**, various features in the pattern formed on the phase shift mask are configured to have proper phase difference to enhance the resolution and imaging quality. In various examples, the phase shift mask can be attenuated PSM or alternating PSM. The mask(s) generated by mask fabrication **1044** is used in a variety of processes. For example, such a mask(s) is used in an ion implantation process to form various doped regions in semiconductor wafer **1053**, in an etching process to form various etching regions in semiconductor wafer **1053**, and/or in other suitable processes.

[0192] IC fab **1050** is an IC fabrication business that includes one or more manufacturing facilities for the fabrication of a variety of different IC products. In some embodiments, IC Fab **1050** is a

semiconductor foundry. For example, there may be a manufacturing facility for the front end fabrication of a plurality of IC products (front-end-of-line (FEOL) fabrication), while a second manufacturing facility may provide the back end fabrication for the interconnection and packaging of the IC products (back-end-of-line (BEOL) fabrication), and a third manufacturing facility may provide other services for the foundry business.

[0193] IC fab **1050** includes fabrication tools **1052** configured to execute various manufacturing operations on semiconductor wafer **1053** such that IC device **1060** is fabricated in accordance with the mask(s), e.g., mask **1045**. In various embodiments, fabrication tools **1052** include one or more of a wafer stepper, an ion implanter, a photoresist coater, a process chamber, e.g., a CVD chamber or LPCVD furnace, a CMP system, a plasma etch system, a wafer cleaning system, or other manufacturing equipment capable of performing one or more suitable manufacturing processes as discussed herein.

[0194] IC fab **1050** uses mask(s) **1045** fabricated by mask house **1030** to fabricate IC device **1060**. Thus, IC fab **1050** at least indirectly uses IC layout diagram **1022** to fabricate IC device **1060**. In some embodiments, semiconductor wafer **1053** is fabricated by IC fab **1050** using mask(s) **1045** to form IC device **1060**. In some embodiments, the IC fabrication includes performing one or more lithographic exposures based at least indirectly on IC layout diagram **1022**. Semiconductor wafer **1053** includes a silicon substrate or other proper substrate having material layers formed thereon. Semiconductor wafer **1053** further includes one or more of various doped regions, dielectric features, multilevel interconnects, and the like (formed at subsequent manufacturing steps).

[0195] Details regarding an integrated circuit (IC) manufacturing system (e.g., system **1000** of FIG. **10**), and an IC manufacturing flow associated therewith are found, e.g., in U.S. Pat. No. 9,256,709, granted Feb. 9, 2016, U.S. Pre-Grant Publication No. 20150278429, published Oct. 1, 2015, U.S. Pre-Grant Publication No. 20140040838, published Feb. 6, 2014, and U.S. Pat. No. 7,260,442, granted Aug. 21, 2007, the entireties of each of which are hereby incorporated by reference.

[0196] In some embodiments, a system (for controlling temperatures in a memory) includes: a high bandwidth memory (HBM) including core dies, the HBM being arranged into portions, each of the portions including memory cells, the HBM further including: a first sensing unit configured to generate one or more first environmental signals corresponding to at least a first transistor in a corresponding at least a first one of the memory cells in a first one of the portions; and a second sensing unit configured to generate one or more second environmental signals corresponding to at least a second transistor in a corresponding at least a second one of the memory cells in a second one of the portions; and a controller configured to perform non-monolithic temperature control of the HBM based on the one or more first environmental signals and the one or more second environmental signals such that temperature of one or more of the portions is differentially controlled.

[0197] In some embodiments, the controller is a differentiated dynamic voltage and frequency scaling (DDVFS) device configured to perform as follows including, for a first set of one or more of the memory cells in the first portion which includes the first memory cell, controlling a temperature of the first set based on the one or more first environmental signals, and for a second set of one or more of the memory cells in the second portion which includes the second memory cell, controlling a temperature of the second set based on the one or more second environmental signals.

[0198] In some embodiments, in terms of membership, the first and second sets are non-overlapping of each other.

[0199] In some embodiments, the first sensing unit is further configured to sense a first temperature or detect a first threshold voltage corresponding to the at least the first transistor in the corresponding at least one of the memory cells of the first set; the second sensing unit is further configured to sense a second temperature or detect a second threshold voltage corresponding to the at least the second transistor in the corresponding at least one of the memory cells of the second set;

the one or more first environmental signals include the first temperature or the first threshold voltage; [0200] the one or more second environmental signals include the second temperature or the second threshold voltage; and the controller is further configured to perform the following: adjusting a first clock or a first power supply (PS) voltage based on the first temperature or the first threshold voltage; and adjusting a second clock or a second PS voltage based on the second temperature or the second threshold voltage.

[0201] In some embodiments, each of the portions includes banks; the first set of one or more of the memory cells is contained within at least a first lot of one or more of the banks; the second set of one or more of the memory cells is contained within at least a second lot of one of the banks; in terms of membership, the first and second lots are non-overlapping of each other; and the controller is further configured to perform the following: adjusting the first clock or the first PS voltage of the first lot resulting in first bank-wide adjustment of corresponding temperatures amongst the memory cells of the banks of the first lot; and adjusting the second clock or the second PS voltage of the second lot resulting in second bank-wide adjustment of corresponding temperatures amongst the memory cells of the banks of the second lot.

[0202] In some embodiments, each of the portions includes banks; the banks of the HBM are arranged in corresponding groups; the first set of one or more of the memory cells is contained within at least a first block of one or more of the groups; the second set of one or more of the memory cells is contained within at least a second block of one or more of the groups; in terms of membership, the first and second blocks are non-overlapping of each other; and the controller is further configured to perform the following: adjusting the first clock or the first PS voltage of the first block resulting in a first group-wide adjustment of temperatures amongst the memory cells of the groups of the first block; and adjusting the second clock or the second PS voltage of the second block resulting in second group-wide adjustment of temperatures amongst the memory cells of the groups of the second block.

[0203] In some embodiments, each of the core dies of the HBM is arranged in channels; the first set of one or more of the memory cells is contained within at least a first bundle of one or more of the channels; the second set of one or more of the memory cells is contained within at least a second bundle of one or more of the channels; in terms of membership, the first and second bundles are non-overlapping of each other; and the controller is further configured to perform the following: adjusting the first clock or the first PS voltage of the first bundle resulting in a first channel-wide adjustment of temperatures amongst the memory cells of the channels of the first bundle; and adjusting the second clock or the second PS voltage of the second bundle resulting in a second channel-wide adjustment of temperatures amongst the memory cells of the channels of the second bundle.

[0204] In some embodiments, for a first set of one or more of the memory cells in the first portion which includes the first memory cell, the first set is contained within a first collection of one or more of the core dies; for a second set of one or more of the memory cells in the second portion which includes the second memory cell, the second set is contained within a second collection of one or more of the core dies; in terms of membership, the first and second collections are non-overlapping of each other; and the controller is further configured to perform the following: adjusting the first clock or the first PS voltage of the first collection resulting in a first core-die-wide adjustment of temperatures amongst the memory cells of the core dies of the first collection; and adjusting the second clock or the second PS voltage of the second collection resulting in a second core-die-wide adjustment of temperatures amongst the memory cells of the core dies of the second collection.

[0205] In some embodiments, the controller includes: an adjuster configured to perform the following: a scenario (A) including: altering a frequency of the first clock based on the first temperature from the first sensing unit; and altering a frequency of the second clock based on the second temperature from the second sensing unit; or a scenario (B) including: altering a value of

the first PS voltage based on the second temperature from the first sensing unit; and altering a value of the second PS voltage based on the second temperature from the second sensing unit.

[0206] In some embodiments, the controller further includes: a temperature comparator configured to perform the following: comparing the first temperature against a first temperature-reference representing a maximum temperature thereby yielding first temperature-comparison results; and comparing the second temperature against the first temperature-reference thereby yielding second temperature-comparison results; and the adjuster is further configured to perform the following: executing a scenario (A) including: altering the frequency of the first clock based on the first temperature-comparison results; and altering the frequency of the second clock based on the second temperature-comparison results; or executing a scenario (B) including: altering the value of the first PS voltage based on the first temperature-comparison results; and altering the value of the second PS voltage based on the second temperature-comparison results.

[0207] In some embodiments, the controller further includes: a threshold-voltage comparator configured to perform the following: comparing the first threshold voltage against a voltage-reference representing a minimum threshold-voltage thereby yielding first voltage-comparison results; and comparing the second threshold voltage against the voltage-reference thereby yielding second voltage-comparison results; and the controller is further configured to perform the following: executing a scenario (A) including: altering the frequency of the first clock based on the first voltage-comparison results; and altering the frequency of the second clock based on the second voltage-comparison results; or executing a scenario (B) including: altering the value of the first PS voltage based on the first voltage-comparison results; and altering the value of the second PS voltage based on the second voltage-comparison results.

[0208] In some embodiments, for a first set of one or more of the memory cells in the first portion which includes the first memory cell, when all members of the first set are idle, the adjuster is configured to perform the following: reducing the frequency of the first clock; or reducing the value of the first PS voltage; and for a second set of one or more of the memory cells in the second portion which includes the second memory cell, when all members of the second set are idle, the adjuster is configured to perform the following: reducing the frequency of the second clock; or reducing the value of the second PS voltage.

[0209] In some embodiments, a system (for controlling temperatures in a memory) includes: a high bandwidth memory (HBM) including core dies, the HBM being arranged into organizational units, each organizational unit including memory cells, the HBM further including: a first sensing unit configured to sense a first temperature corresponding to at least a first transistor in a corresponding at least a first one of the memory cells in a first one of the organizational units; and a second sensing unit configured to sense a second temperature corresponding to at least a second transistor in a corresponding at least a second one of the memory cells in a second one of the organizational units; and a controller configured to perform non-monolithic temperature control of the HBM based on the first temperature and the second temperature such that temperature of one or more of the organizational units is differentially controlled.

[0210] In some embodiments, the controller is a differentiated dynamic voltage and frequency scaling (DDVFS) device configured to perform as follows including, for a first set of one or more of the memory cells which includes the first memory cell, controlling a temperature of the first set based on the first temperature, and for a second set of one or more of the memory cells which includes the second memory cell, controlling a temperature of the second set based on the second temperature.

[0211] In some embodiments, the controller further includes: a temperature comparator configured to perform the following: comparing the first temperature against a first temperature-reference representing a maximum temperature thereby yielding first temperature-comparison results; and comparing the second temperature against the first temperature-reference thereby yielding second temperature-comparison results; and the controller is further configured to perform the following:

executing a scenario (A) including: altering a frequency of a first clock based on the first temperature-comparison results; and altering a frequency of a second clock based on the second temperature-comparison results; or executing a scenario (B) including: altering a value of a first power supply (PS) voltage based on the first temperature-comparison results; and altering the value of a second PS voltage based on the second temperature-comparison results.

[0212] In some embodiments, when the first temperature is below a second temperature-reference representing a moderate temperature, the controller is further configured to perform the following: increasing the frequency of the first clock; or increasing the value of the first PS voltage; and when the second temperature is below the second temperature-reference, the controller is further configured to perform the following: increasing the frequency of the second clock; or increasing the value of the second PS voltage.

[0213] In some embodiments, for a first set of one or more of the memory cells in the first organizational unit which includes the first memory cell, when all members of the first set are idle, the controller is further configured to perform the following: reducing the frequency of the first clock; or reducing the value of the first PS voltage; and for a second set of one or more of the memory cells in the second organizational unit which includes the second memory cell, when all members of the second set are idle, the controller is further configured to perform the following: reducing the frequency of the second clock; or reducing the value of the second PS voltage.

[0214] In some embodiments, a method (of controlling temperatures in a high bandwidth memory (HBM) that includes core dies, the HBM being arranged into portions, each of the portions including memory cells, the HBM further including sensing units arranged correspondingly within the portions) includes: from a first one of the sensing units arranged within the HBM, receiving a first threshold voltage corresponding to at least a first transistor in a corresponding at least a first one of the memory cells in a first one of the portions; from a second one of the sensing units arranged within the HBM, receiving a second threshold voltage corresponding to at least a second transistor in a corresponding at least a second one of the memory cells in a second one of the portions; and performing non-monolithic temperature control of the HBM based on the first threshold voltage and the second threshold voltage such that temperature of one or more of the portions is differentially controlled.

[0215] In some embodiments, the performing non-monolithic temperature control of the HBM includes: for a first one of the portions represented by a first set of one or more of the memory cells which includes the first memory cell, controlling a temperature of the first portion based on the first threshold voltage; and for a second one of the portions represented a second set of one or more of the memory cells which includes the second memory cell, controlling a temperature of the second portion based on the second threshold voltage.

[0216] In some embodiments, the method further includes: comparing the first threshold voltage against a first voltage-reference representing a minimum threshold-voltage thereby yielding first voltage-comparison results; comparing the second threshold voltage against the first voltage-reference thereby yielding second voltage-comparison results; and executing a scenario (A) or a scenario (B); and wherein: the scenario (A) includes: altering a frequency of a first clock based on the first voltage-comparison results; and altering the frequency of a second clock based on the second voltage-comparison results; or the scenario (B) includes: altering a value of a first power supply (PS) voltage based on the first voltage-comparison results; and altering the value of a second PS voltage based on the second voltage-comparison results.

[0217] It will be readily seen by one of ordinary skill in the art that one or more of the disclosed embodiments fulfill one or more of the advantages set forth above. After reading the foregoing specification, one of ordinary skill will be able to affect various changes, substitutions of equivalents and various other embodiments as broadly disclosed herein. It is therefore intended that the protection granted hereon be limited only by the definition contained in the appended claims and equivalents thereof.

Claims

1. A system for controlling temperatures in a memory, the system comprising: a high bandwidth memory (HBM) including core dies, the HBM being arranged into portions, each of the portions including memory cells, the HBM further including: a first sensing unit configured to generate one or more first environmental signals corresponding to at least a first transistor in a corresponding at least a first one of the memory cells in a first one of the portions; and a second sensing unit configured to generate one or more second environmental signals corresponding to at least a second transistor in a corresponding at least a second one of the memory cells in a second one of the portions; and a controller configured to perform non-monolithic temperature control of the HBM based on the one or more first environmental signals and the one or more second environmental signals such that temperature of one or more of the portions is differentially controlled.

2. The system of claim 1, wherein: the controller is a differentiated dynamic voltage and frequency scaling (DDVFS) device configured to perform as follows including, for a first set of one or more of the memory cells in the first portion which includes the first memory cell, controlling a temperature of the first set based on the one or more first environmental signals, and for a second set of one or more of the memory cells in the second portion which includes the second memory cell, controlling a temperature of the second set based on the one or more second environmental signals.

3. The system of claim 2, wherein: in terms of membership, the first and second sets are non-overlapping of each other.

4. The system of claim 1, wherein: the first sensing unit is further configured to sense a first temperature or detect a first threshold voltage corresponding to the at least the first transistor in the corresponding at least one of the memory cells of the first set; the second sensing unit is further configured to sense a second temperature or detect a second threshold voltage corresponding to the at least the second transistor in the corresponding at least one of the memory cells of the second set; the one or more first environmental signals include the first temperature or the first threshold voltage; the one or more second environmental signals include the second temperature or the second threshold voltage; and the controller is further configured to perform the following: adjusting a first clock or a first power supply (PS) voltage based on the first temperature or the first threshold voltage; and adjusting a second clock or a second PS voltage based on the second temperature or the second threshold voltage.

5. The system of claim 4, wherein: each of the portions includes banks; the first set of one or more of the memory cells is contained within at least a first lot of one or more of the banks; the second set of one or more of the memory cells is contained within at least a second lot of one of the banks; in terms of membership, the first and second lots are non-overlapping of each other; and the controller is further configured to perform the following: adjusting the first clock or the first PS voltage of the first lot resulting in first bank-wide adjustment of corresponding temperatures amongst the memory cells of the banks of the first lot; and adjusting the second clock or the second PS voltage of the second lot resulting in second bank-wide adjustment of corresponding temperatures amongst the memory cells of the banks of the second lot.

6. The system of claim 4, wherein: each of the portions includes banks; the banks of the HBM are arranged in corresponding groups; the first set of one or more of the memory cells is contained within at least a first block of one or more of the groups; the second set of one or more of the memory cells is contained within at least a second block of one or more of the groups; in terms of membership, the first and second blocks are non-overlapping of each other; and the controller is further configured to perform the following: adjusting the first clock or the first PS voltage of the first block resulting in a first group-wide adjustment of temperatures amongst the memory cells of the groups of the first block; and adjusting the second clock or the second PS voltage of the second

block resulting in second group-wide adjustment of temperatures amongst the memory cells of the groups of the second block.

7. The system of claim 4, wherein: each of the core dies of the HBM is arranged in channels; the first set of one or more of the memory cells is contained within at least a first bundle of one or more of the channels; the second set of one or more of the memory cells is contained within at least a second bundle of one or more of the channels; in terms of membership, the first and second bundles are non-overlapping of each other; and the controller is further configured to perform the following: adjusting the first clock or the first PS voltage of the first bundle resulting in a first channel-wide adjustment of temperatures amongst the memory cells of the channels of the first bundle; and adjusting the second clock or the second PS voltage of the second bundle resulting in a second channel-wide adjustment of temperatures amongst the memory cells of the channels of the second bundle.

8. The system of claim 4, wherein: for a first set of one or more of the memory cells in the first portion which includes the first memory cell, the first set is contained within a first collection of one or more of the core dies; for a second set of one or more of the memory cells in the second portion which includes the second memory cell, the second set is contained within a second collection of one or more of the core dies; in terms of membership, the first and second collections are non-overlapping of each other; and the controller is further configured to perform the following: adjusting the first clock or the first PS voltage of the first collection resulting in a first core-die-wide adjustment of temperatures amongst the memory cells of the core dies of the first collection; and adjusting the second clock or the second PS voltage of the second collection resulting in a second core-die-wide adjustment of temperatures amongst the memory cells of the core dies of the second collection.

9. The system of claim 4, wherein the controller includes: an adjuster configured to perform the following: a scenario (A) including: altering a frequency of the first clock based on the first temperature from the first sensing unit; and altering a frequency of the second clock based on the second temperature from the second sensing unit; or a scenario (B) including: altering a value of the first PS voltage based on the second temperature from the first sensing unit; and altering a value of the second PS voltage based on the second temperature from the second sensing unit.

10. The system of claim 9, wherein: the controller further includes: a temperature comparator configured to perform the following: comparing the first temperature against a first temperature-reference representing a maximum temperature thereby yielding first temperature-comparison results; and comparing the second temperature against the first temperature-reference thereby yielding second temperature-comparison results; and the adjuster is further configured to perform the following: executing a scenario (A) including: altering the frequency of the first clock based on the first temperature-comparison results; and altering the frequency of the second clock based on the second temperature-comparison results; or executing a scenario (B) including: altering the value of the first PS voltage based on the first temperature-comparison results; and altering the value of the second PS voltage based on the second temperature-comparison results.

11. The system of claim 9, wherein: the controller further includes: a threshold-voltage comparator configured to perform the following: comparing the first threshold voltage against a voltage-reference representing a minimum threshold-voltage thereby yielding first voltage-comparison results; and comparing the second threshold voltage against the voltage-reference thereby yielding second voltage-comparison results; and the controller is further configured to perform the following: executing a scenario (A) including: altering the frequency of the first clock based on the first voltage-comparison results; and altering the frequency of the second clock based on the second voltage-comparison results; or executing a scenario (B) including: altering the value of the first PS voltage based on the first voltage-comparison results; and altering the value of the second PS voltage based on the second voltage-comparison results.

12. The system of claim 9, wherein: for a first set of one or more of the memory cells in the first

portion which includes the first memory cell, when all members of the first set are idle, the adjuster is configured to perform the following: reducing the frequency of the first clock; or reducing the value of the first PS voltage; and for a second set of one or more of the memory cells in the second portion which includes the second memory cell, when all members of the second set are idle, the adjuster is configured to perform the following: reducing the frequency of the second clock; or reducing the value of the second PS voltage.

13. A system for controlling temperatures in a memory, the system comprising: a high bandwidth memory (HBM) including core dies, the HBM being arranged into organizational units, each organizational unit including memory cells, the HBM further including: a first sensing unit configured to sense a first temperature corresponding to at least a first transistor in a corresponding at least a first one of the memory cells in a first one of the organizational units; and a second sensing unit configured to sense a second temperature corresponding to at least a second transistor in a corresponding at least a second one of the memory cells in a second one of the organizational units; and a controller configured to perform non-monolithic temperature control of the HBM based on the first temperature and the second temperature such that temperature of one or more of the organizational units is differentially controlled.

14. The system of claim 13, wherein: the controller is a differentiated dynamic voltage and frequency scaling (DDVFS) device configured to perform as follows including, for a first set of one or more of the memory cells which includes the first memory cell, controlling a temperature of the first set based on the first temperature, and for a second set of one or more of the memory cells which includes the second memory cell, controlling a temperature of the second set based on the second temperature.

15. The system of claim 13, wherein: the controller further includes: a temperature comparator configured to perform the following: comparing the first temperature against a first temperature-reference representing a maximum temperature thereby yielding first temperature-comparison results; and comparing the second temperature against the first temperature-reference thereby yielding second temperature-comparison results; and the controller is further configured to perform the following: executing a scenario (A) including: altering a frequency of a first clock based on the first temperature-comparison results; and altering a frequency of a second clock based on the second temperature-comparison results; or executing a scenario (B) including: altering a value of a first power supply (PS) voltage based on the first temperature-comparison results; and altering the value of a second PS voltage based on the second temperature-comparison results.

16. The system of claim 15, wherein: when the first temperature is below a second temperature-reference representing a moderate temperature, the controller is further configured to perform the following: increasing the frequency of the first clock; or increasing the value of the first PS voltage; and when the second temperature is below the second temperature-reference, the controller is further configured to perform the following: increasing the frequency of the second clock; or increasing the value of the second PS voltage.

17. The system of claim 15, wherein: for a first set of one or more of the memory cells in the first organizational unit which includes the first memory cell, when all members of the first set are idle, the controller is further configured to perform the following: reducing the frequency of the first clock; or reducing the value of the first PS voltage; and for a second set of one or more of the memory cells in the second organizational unit which includes the second memory cell, when all members of the second set are idle, the controller is further configured to perform the following: reducing the frequency of the second clock; or reducing the value of the second PS voltage.

18. A method of controlling temperatures in a high bandwidth memory (HBM) that includes core dies, the HBM being arranged into portions, each of the portions including memory cells, the HBM further including sensing units arranged correspondingly within the portions, the method comprising: from a first one of the sensing units arranged within the HBM, receiving a first threshold voltage corresponding to at least a first transistor in a corresponding at least a first one of

the memory cells in a first one of the portions; from a second one of the sensing units arranged within the HBM, receiving a second threshold voltage corresponding to at least a second transistor in a corresponding at least a second one of the memory cells in a second one of the portions; and performing non-monolithic temperature control of the HBM based on the first threshold voltage and the second threshold voltage such that temperature of one or more of the portions is differentially controlled.

19. The method of claim 18, wherein the performing non-monolithic temperature control of the HBM includes: for a first one of the portions represented by a first set of one or more of the memory cells which includes the first memory cell, controlling a temperature of the first portion based on the first threshold voltage; and for a second one of the portions represented a second set of one or more of the memory cells which includes the second memory cell, controlling a temperature of the second portion based on the second threshold voltage.

20. The method of claim 18, further comprising: comparing the first threshold voltage against a first voltage-reference representing a minimum threshold-voltage thereby yielding first voltage-comparison results; comparing the second threshold voltage against the first voltage-reference thereby yielding second voltage-comparison results; and executing a scenario (A) or a scenario (B); and wherein: the scenario (A) includes: altering a frequency of a first clock based on the first voltage-comparison results; and altering the frequency of a second clock based on the second voltage-comparison results; or the scenario (B) includes: altering a value of a first power supply (PS) voltage based on the first voltage-comparison results; and altering the value of a second PS voltage based on the second voltage-comparison results.