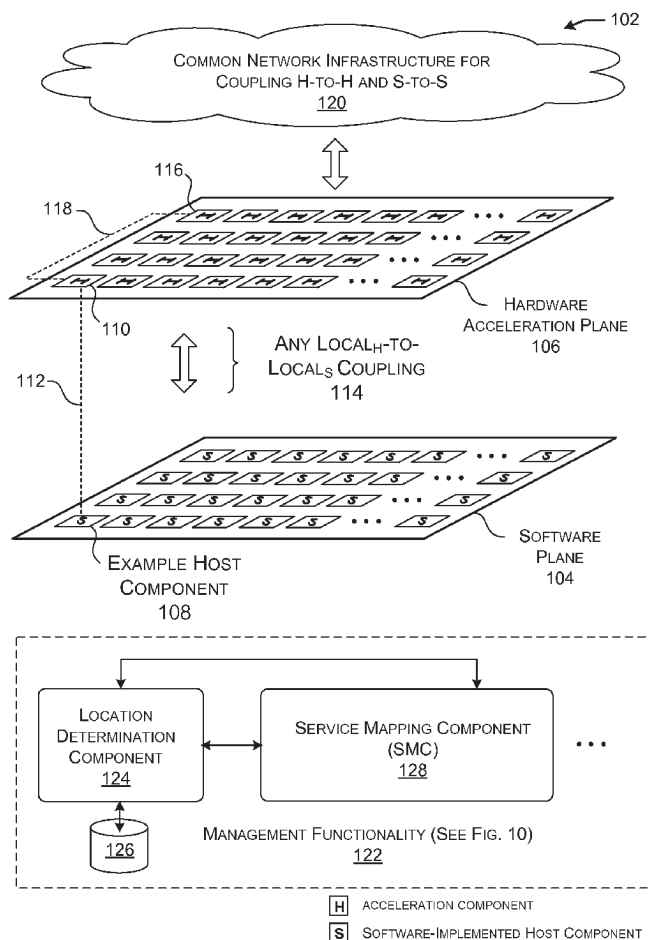


(19) **United States**(12) **Patent Application Publication** (10) **Pub. No.: US 2016/0306668 A1**
Heil et al. (43) **Pub. Date: Oct. 20, 2016**(54) **IMPLEMENTING A SERVICE USING
PLURAL ACCELERATION COMPONENTS**(52) **U.S. CL.**
CPC **G06F 9/5027** (2013.01)(71) Applicant: **Microsoft Technology Licensing, LLC**,
Redmond, WA (US)(57) **ABSTRACT**(72) Inventors: **Stephen F. Heil**, Sammamish, WA
(US); **Adrian M. Caulfield**,
Woodinville, WA (US); **Douglas C.**
Burger, Bellevue, WA (US); **Andrew**
R. Putnam, Seattle, WA (US); **Eric S.**
Chung, Woodinville, WA (US)

A data processing system is described herein that includes two or more software-driven host components that collectively provide a software plane. The data processing system further includes two or more hardware acceleration components that collectively provide a hardware acceleration plane. The hardware acceleration plane implements one or more services, including at least one multi-component service. The multi-component service has plural parts, and is implemented on a collection of two or more hardware acceleration components, where each hardware acceleration component in the collection implements a corresponding part of the multi-component service. Each hardware acceleration component in the collection is configured to interact with other hardware acceleration components in the collection without involvement from any host component. A function parsing component is also described herein that determines a manner of parsing a function into the plural parts of the multi-component service.

(21) Appl. No.: **14/717,788**(22) Filed: **May 20, 2015****Related U.S. Application Data**(60) Provisional application No. 62/149,488, filed on Apr.
17, 2015.**Publication Classification**(51) **Int. Cl.**
G06F 9/50 (2006.01)

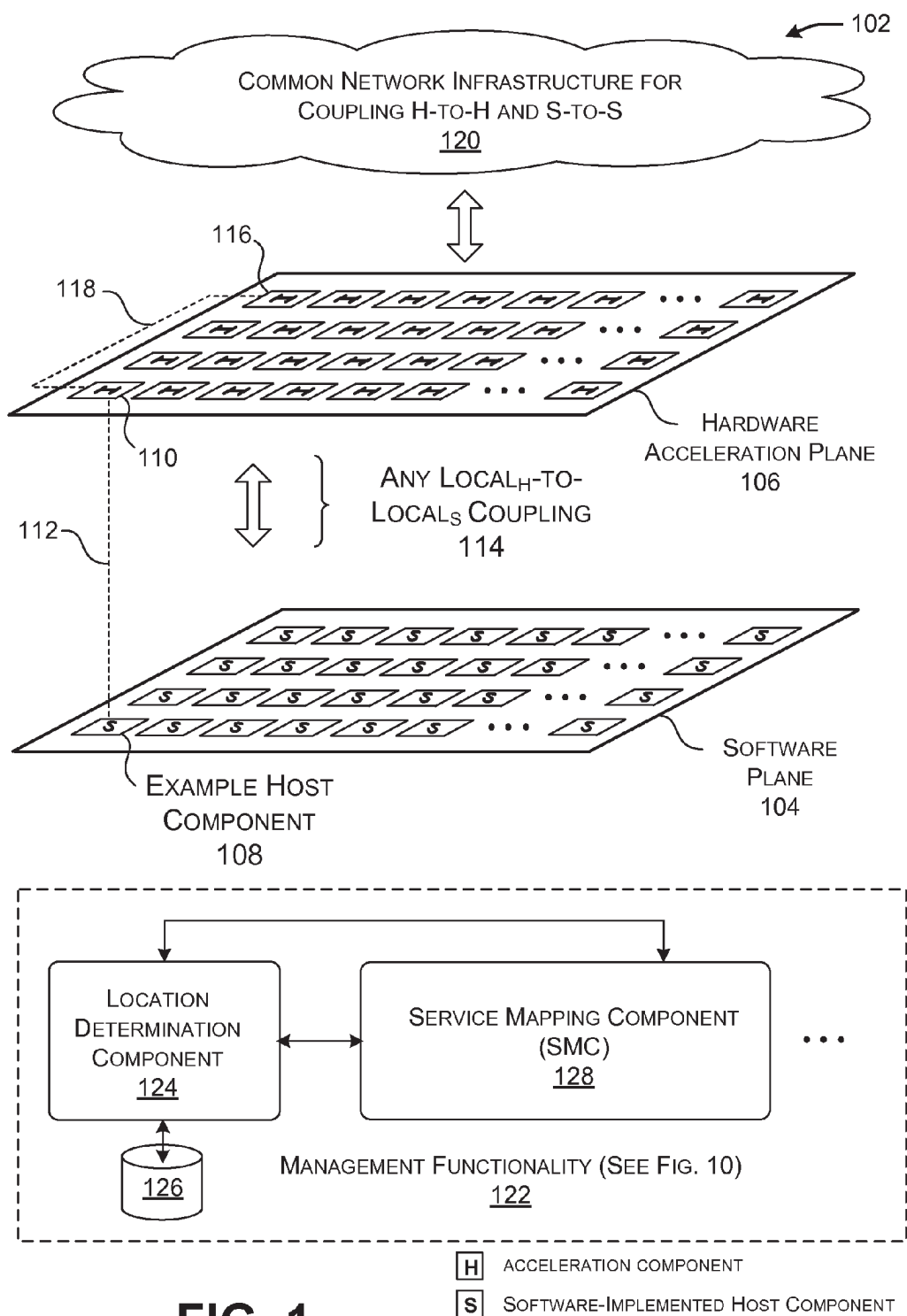


FIG. 1

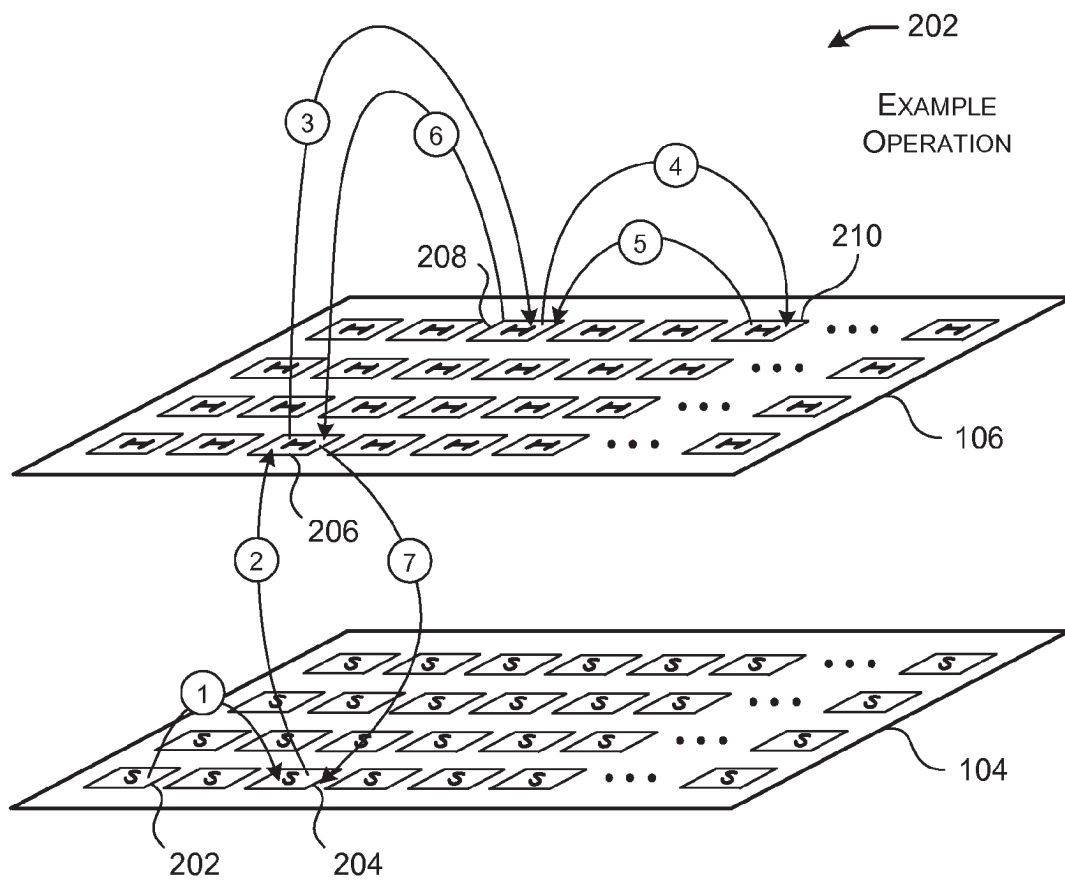


FIG. 2

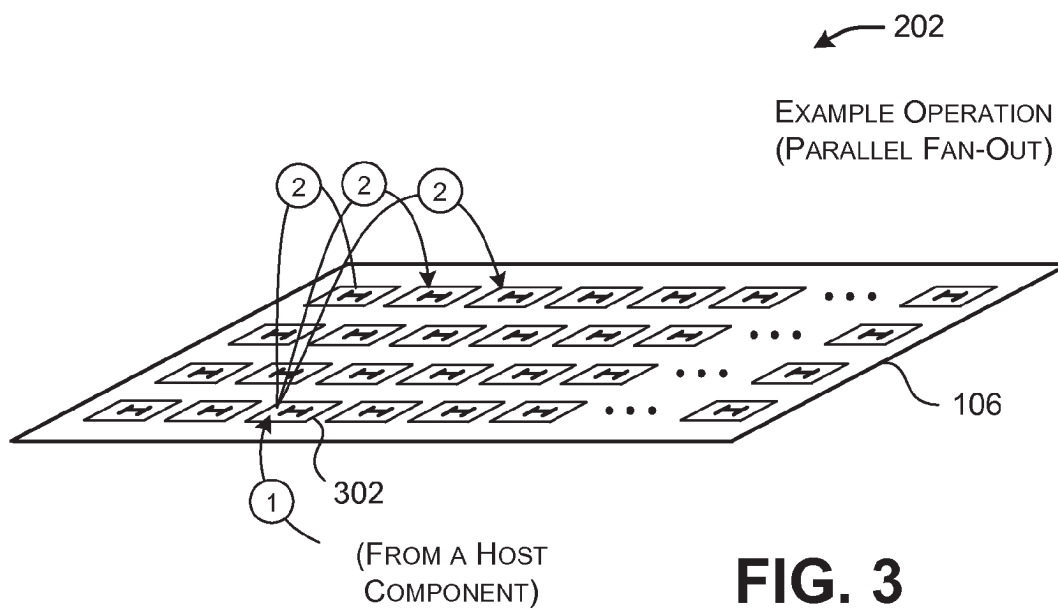


FIG. 3

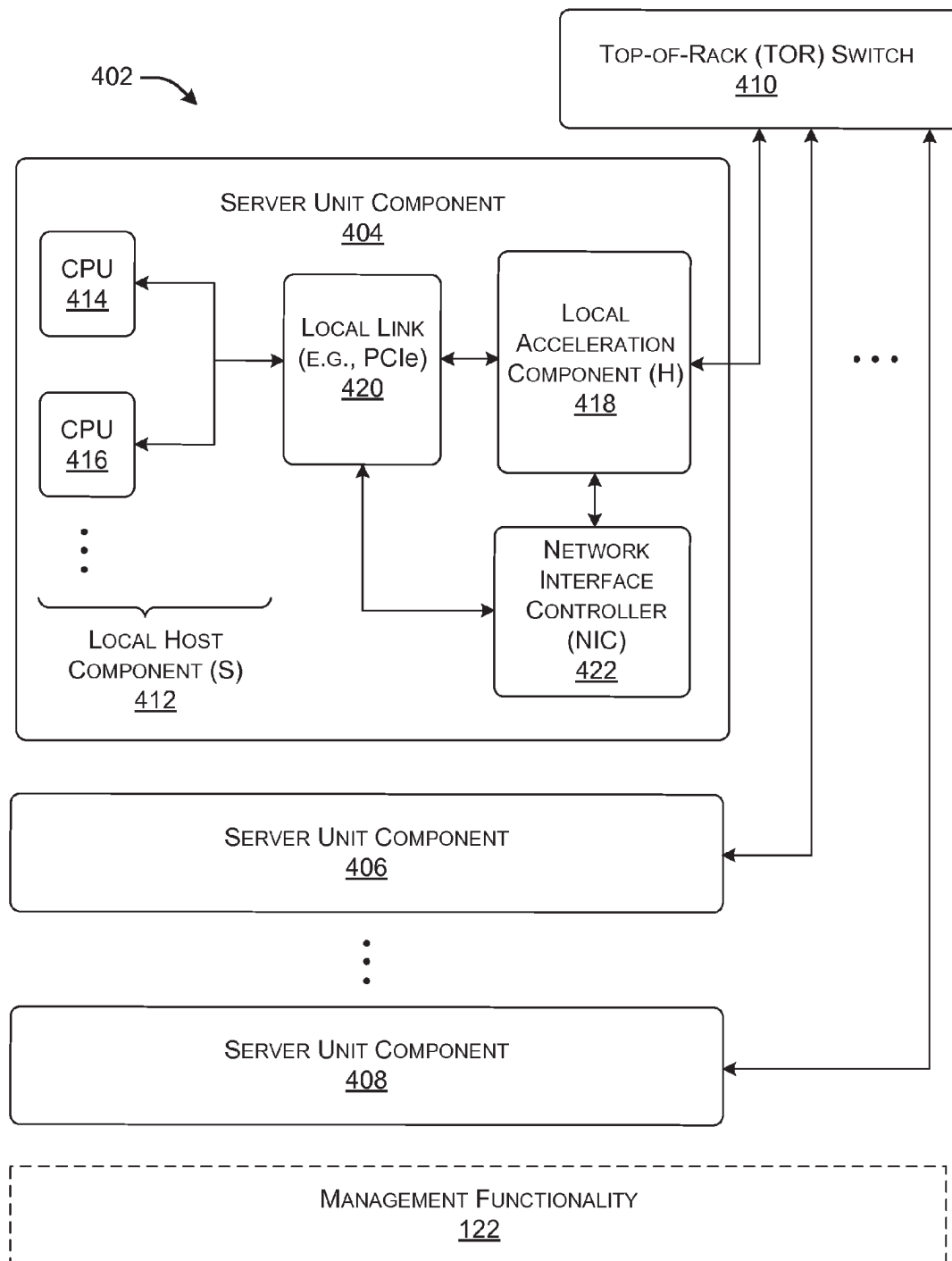


FIG. 4

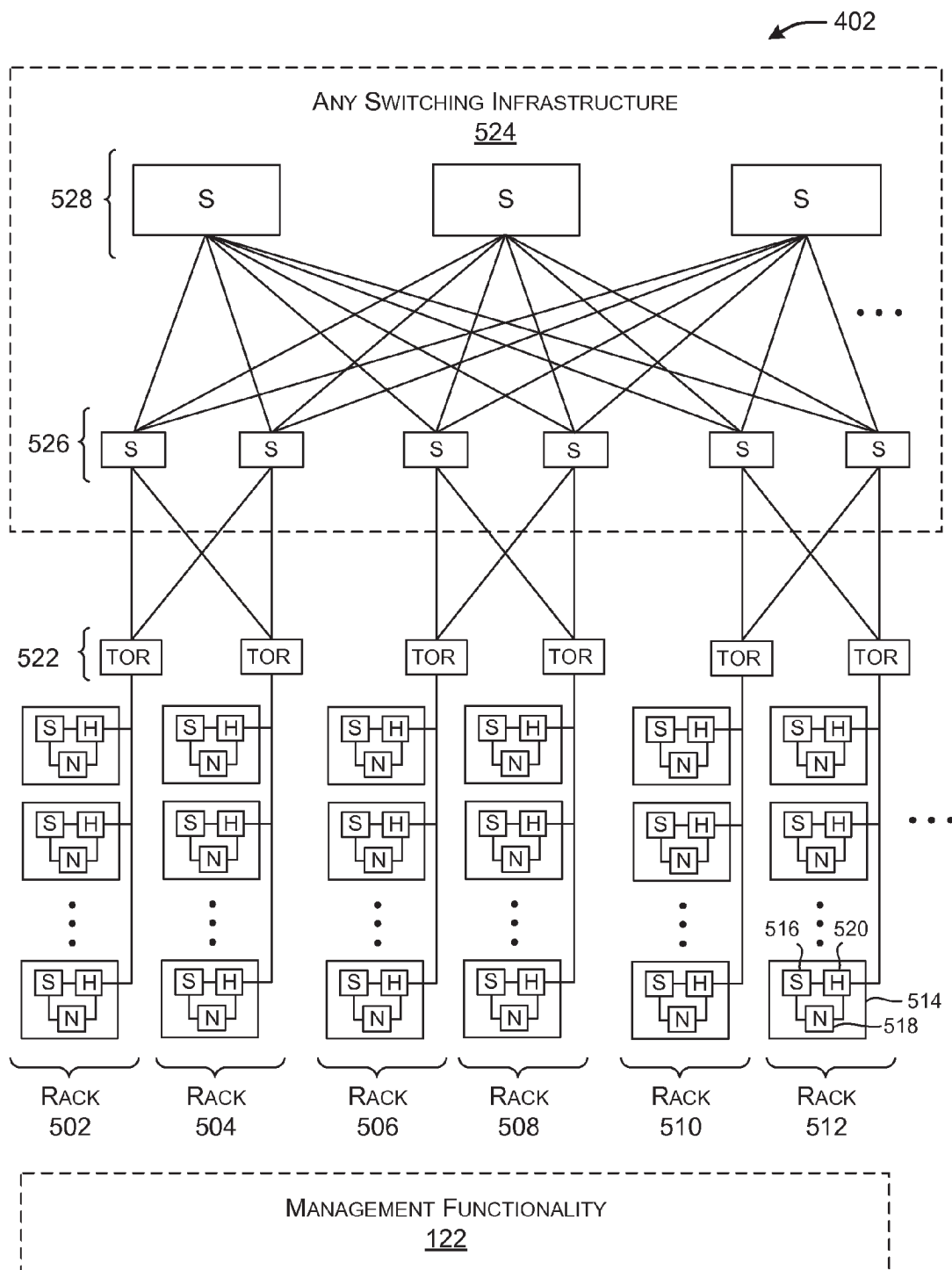


FIG. 5

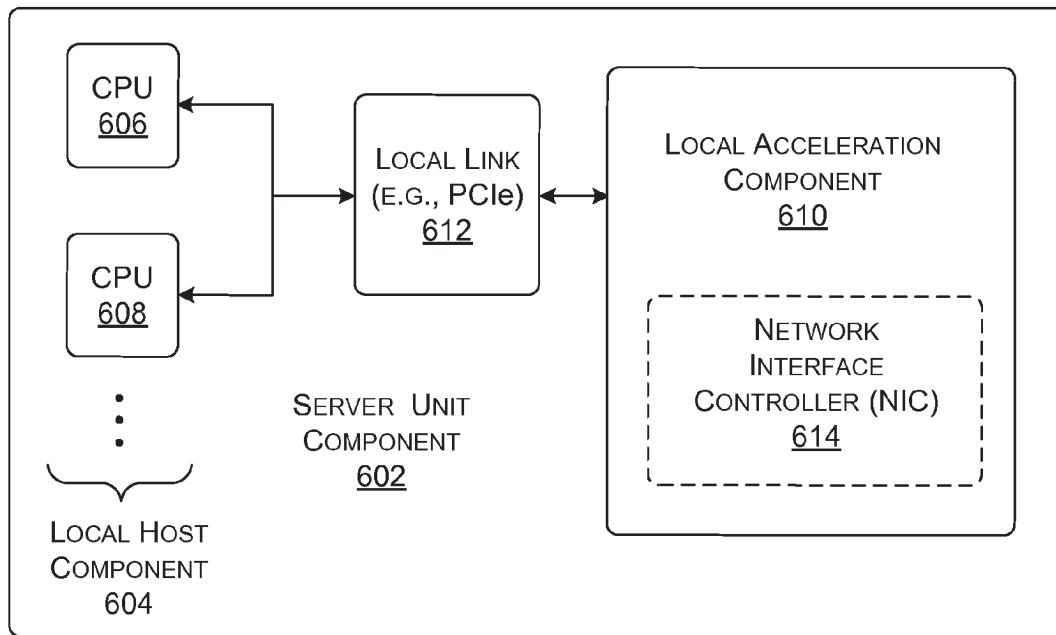


FIG. 6

SERVER UNIT COMPONENT WITH ANY n-TO-m ASSOCIATION OF
HOST COMPONENTS AND ACCELERATION COMPONENTS
702

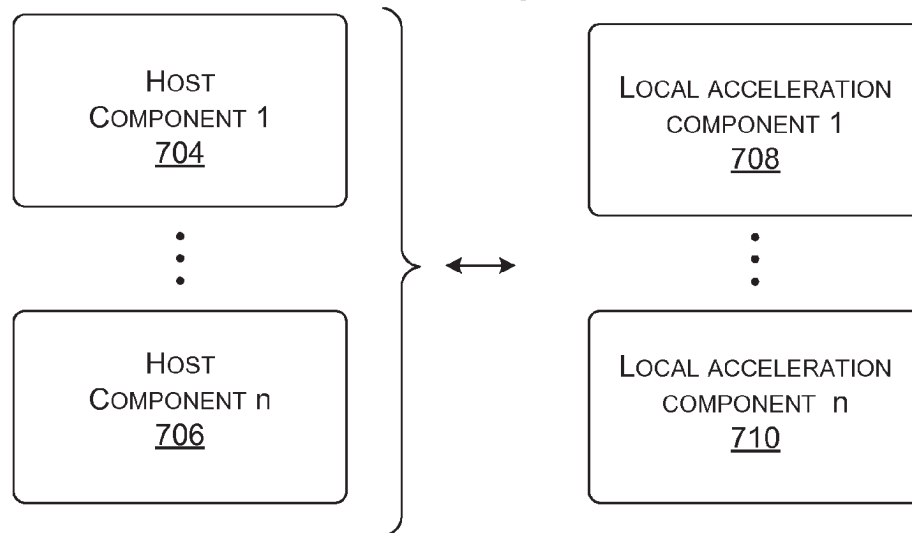


FIG. 7

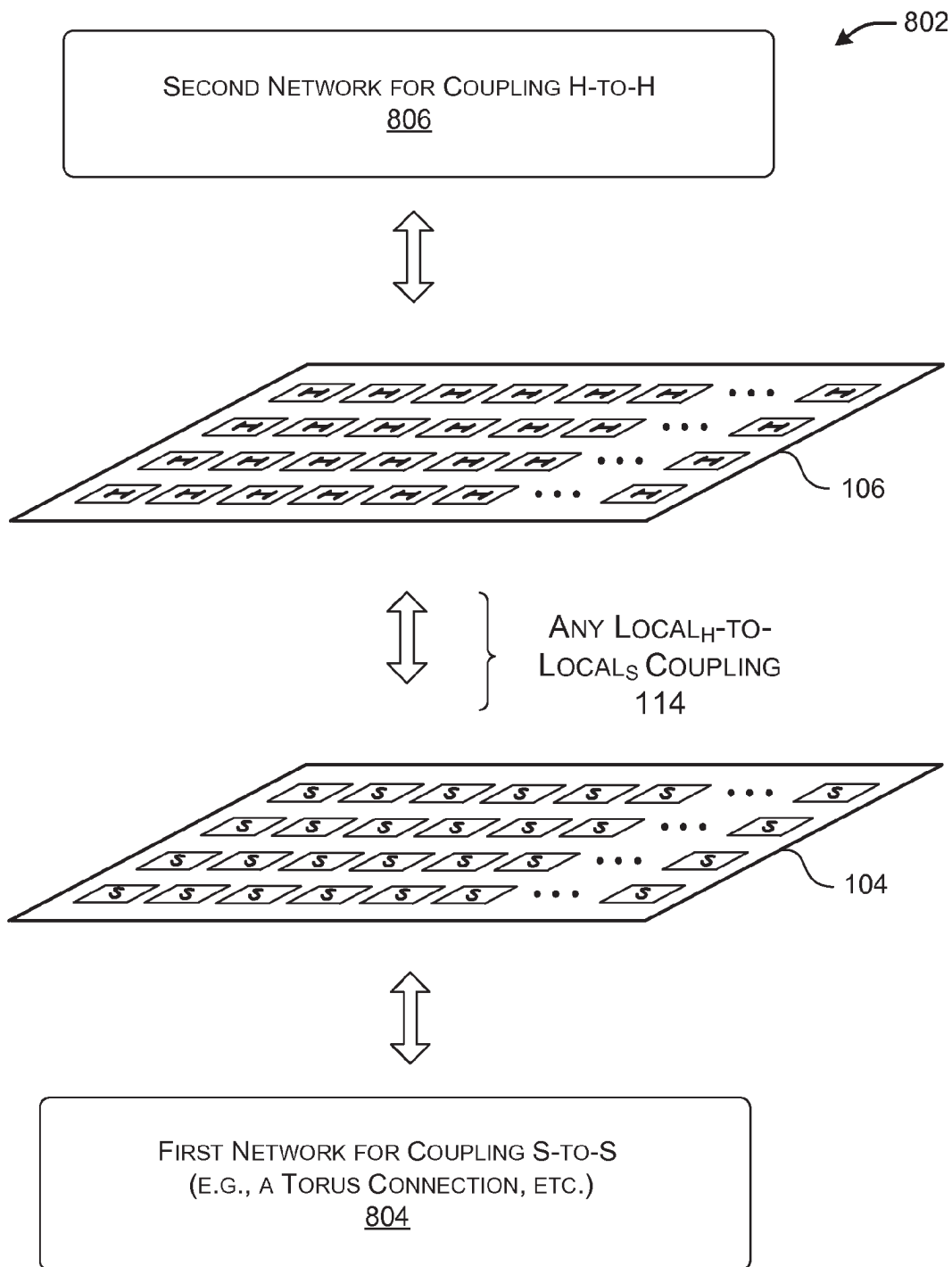


FIG. 8

GENERAL OPERATION OF THE DATA PROCESSING SYSTEM
902

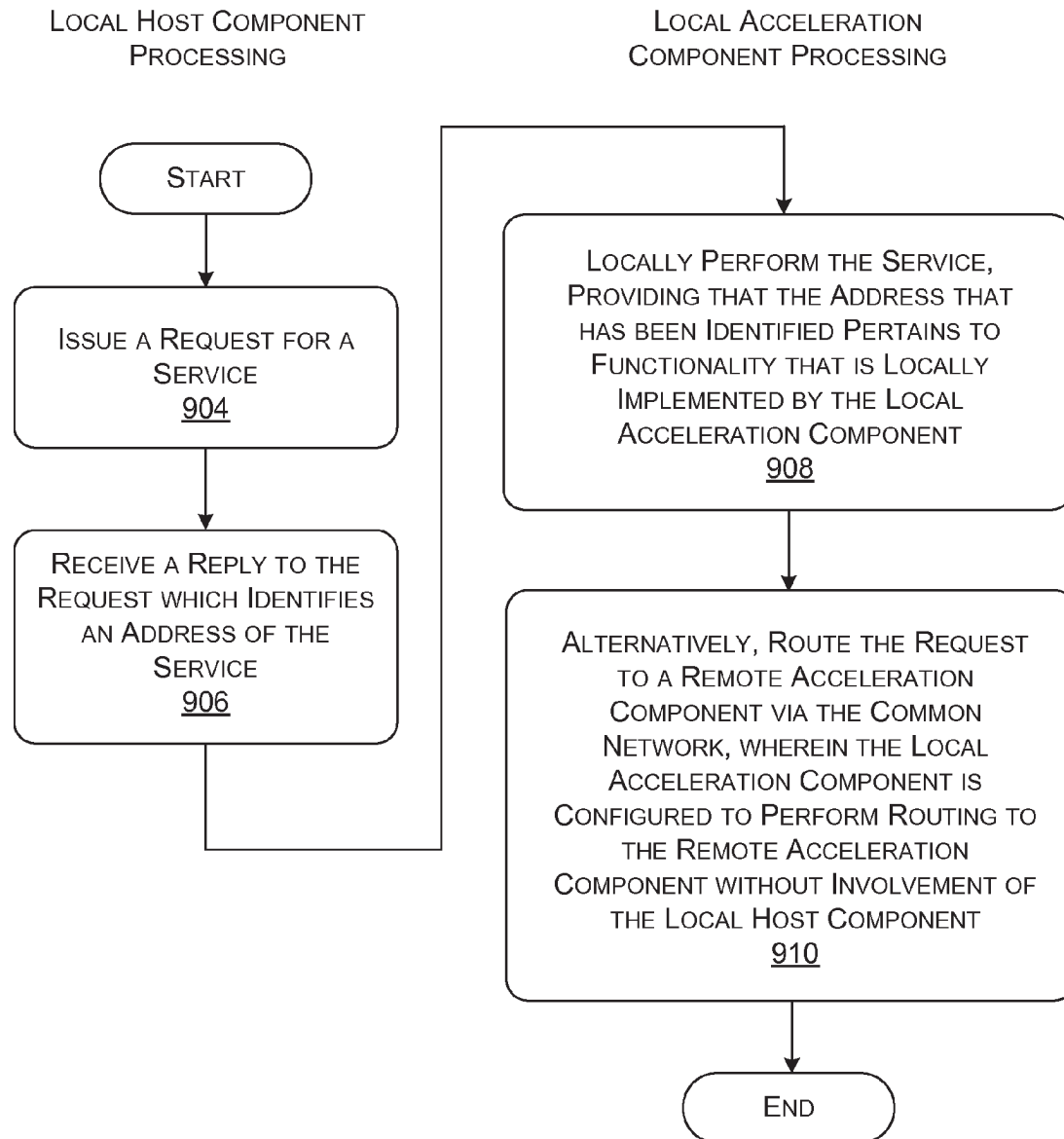


FIG. 9

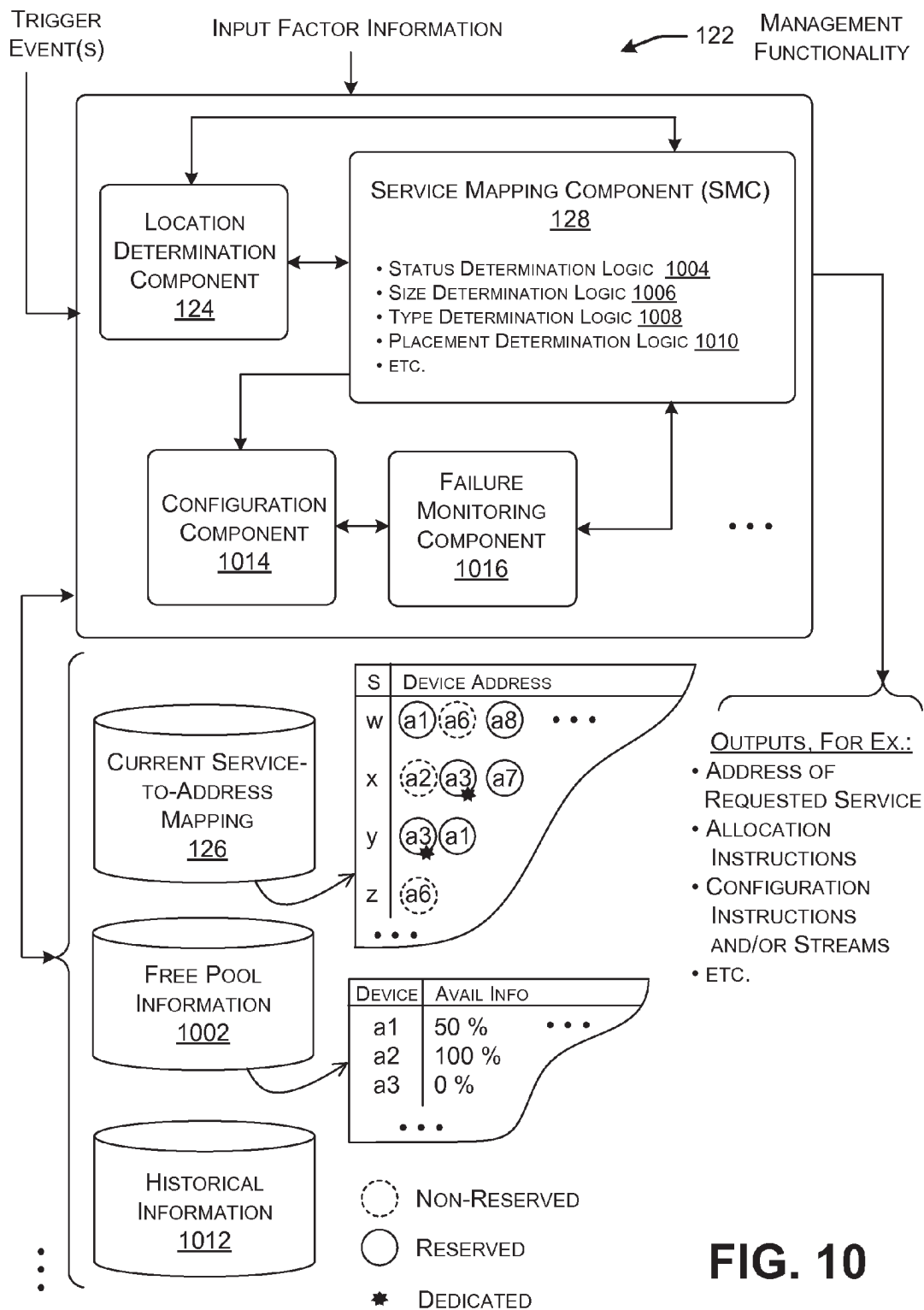


FIG. 10

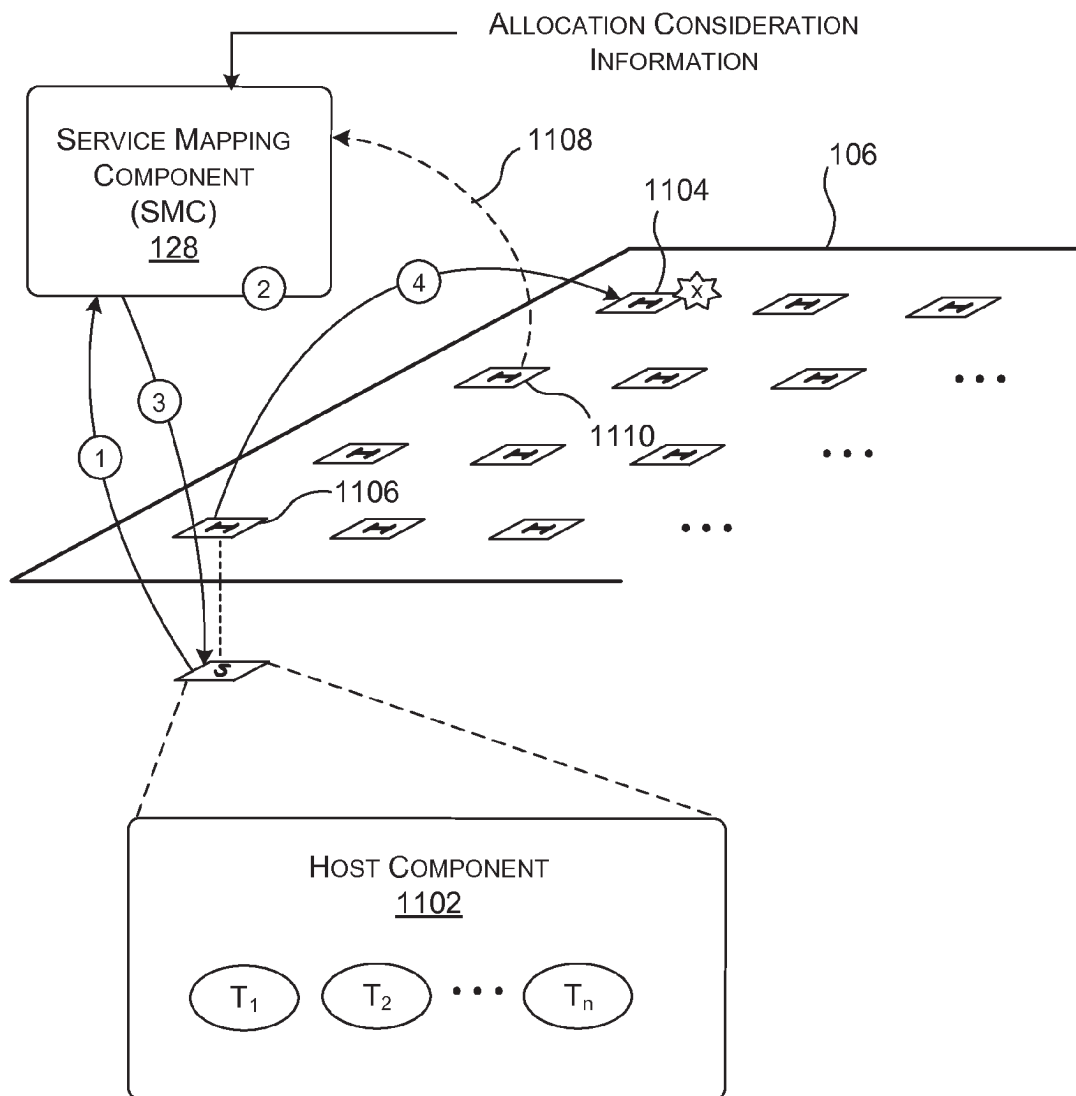


FIG. 11

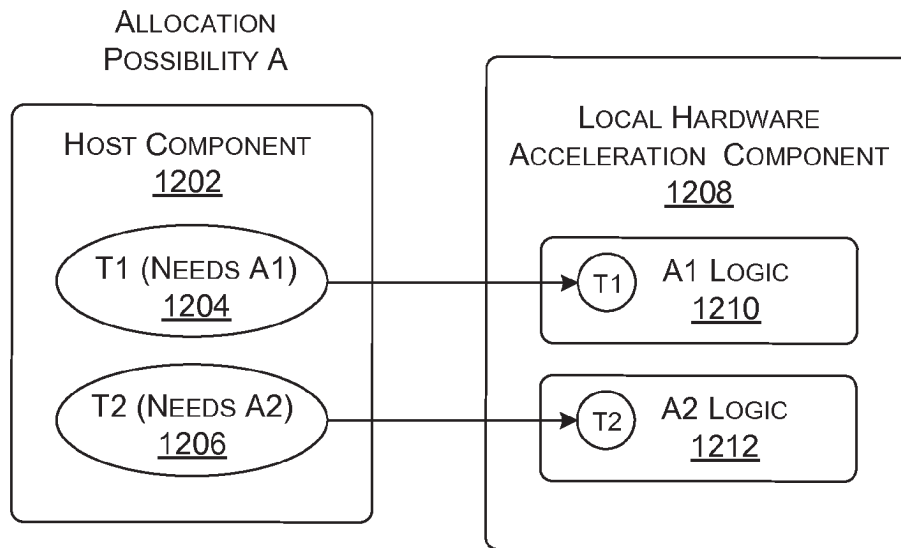


FIG.12

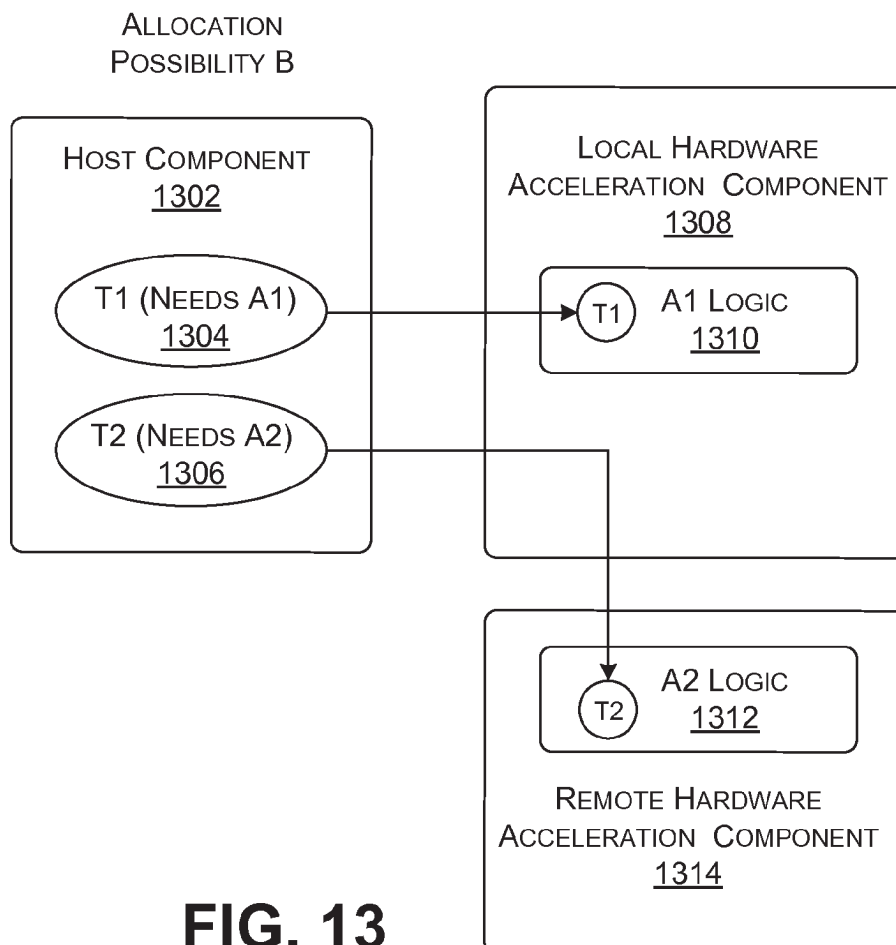
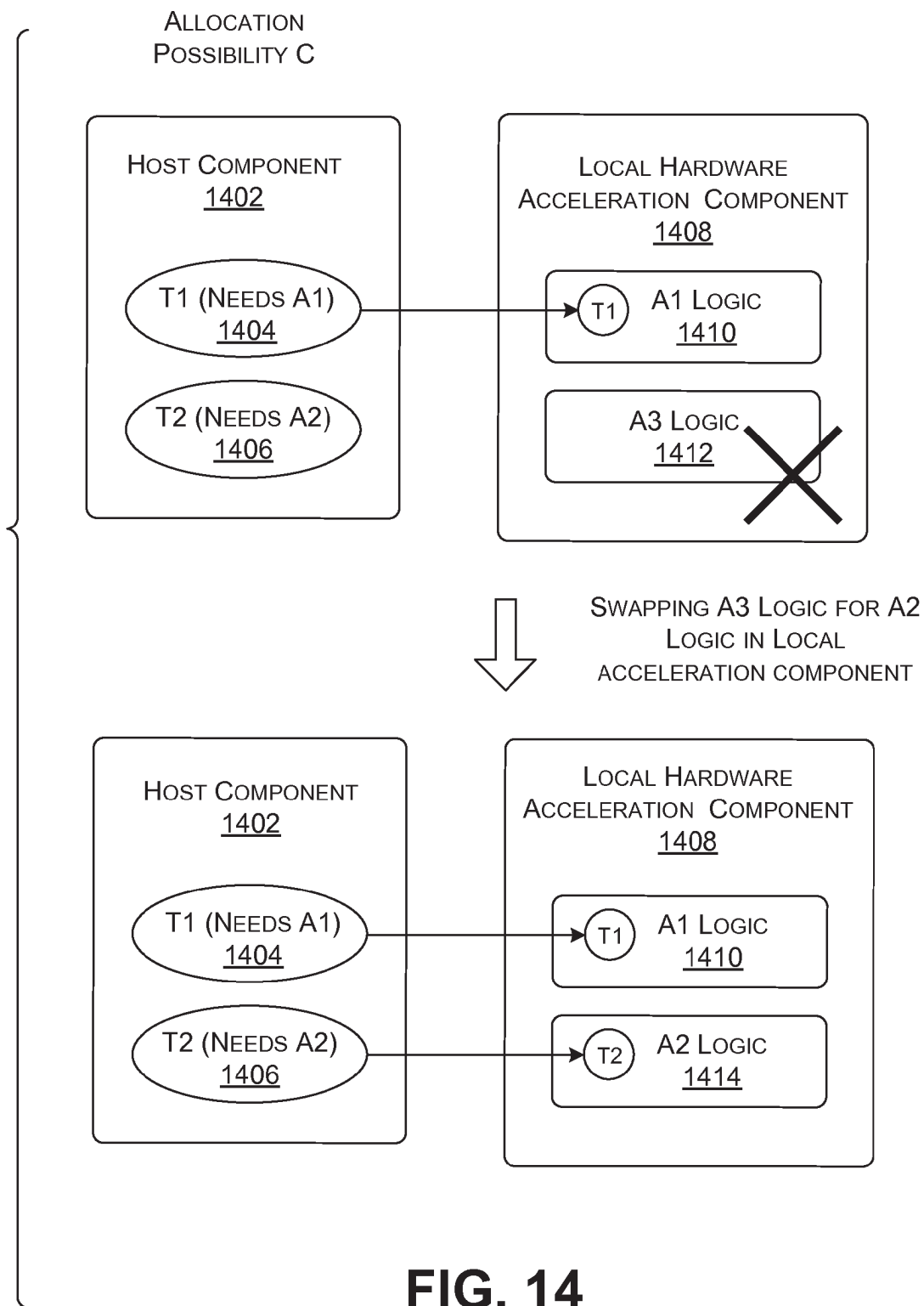
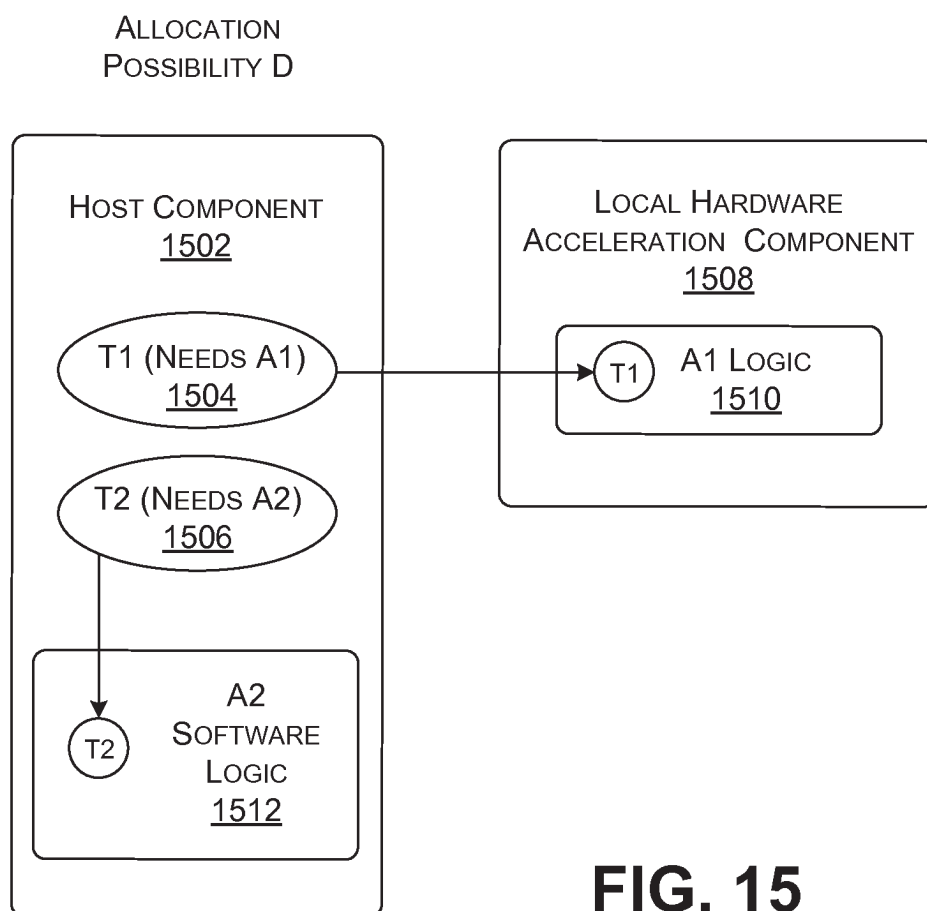


FIG. 13





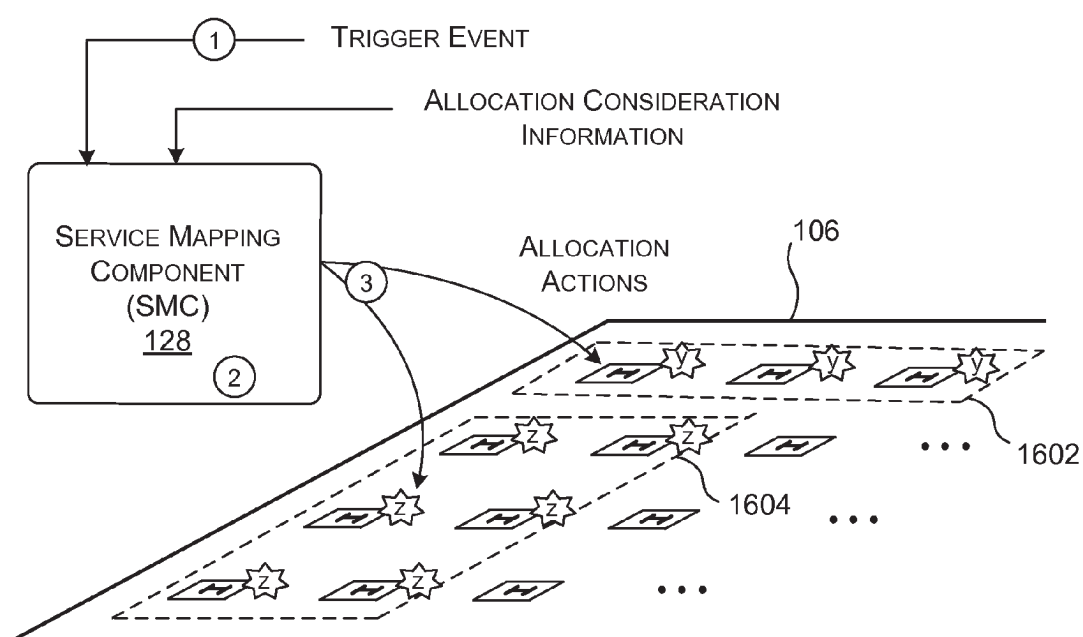


FIG. 16

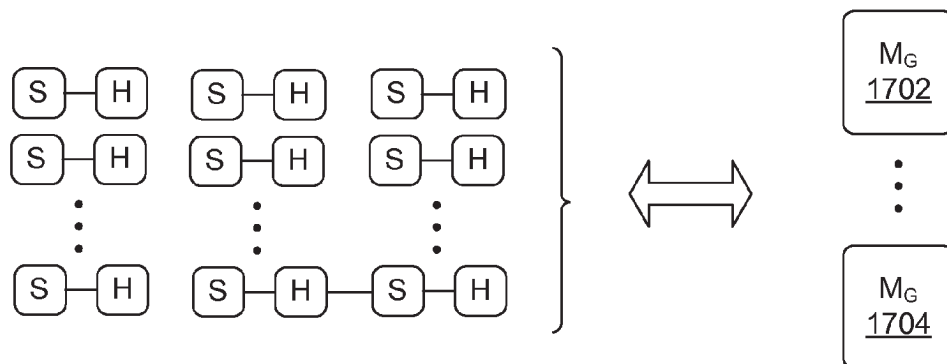


FIG. 17

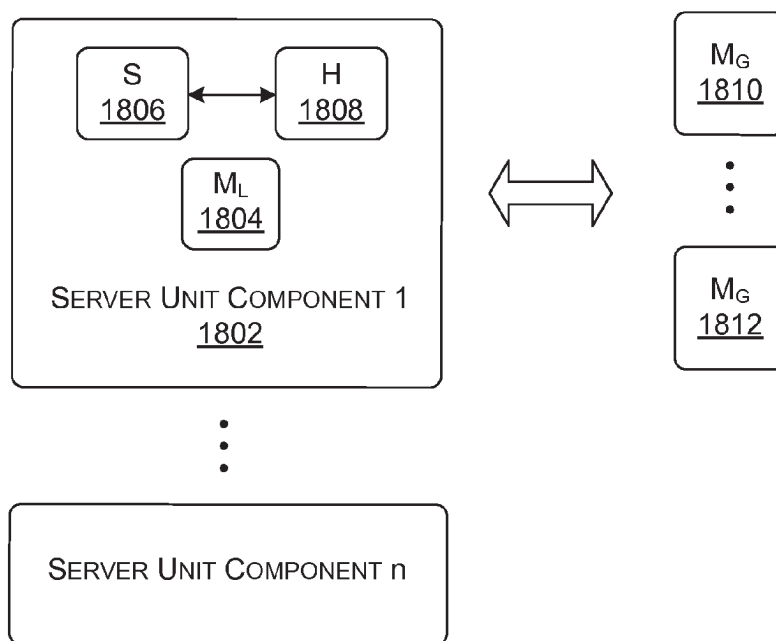


FIG. 18

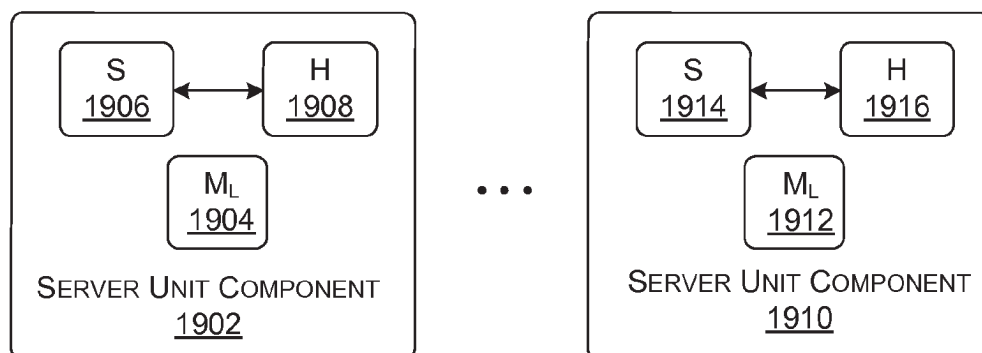


FIG. 19

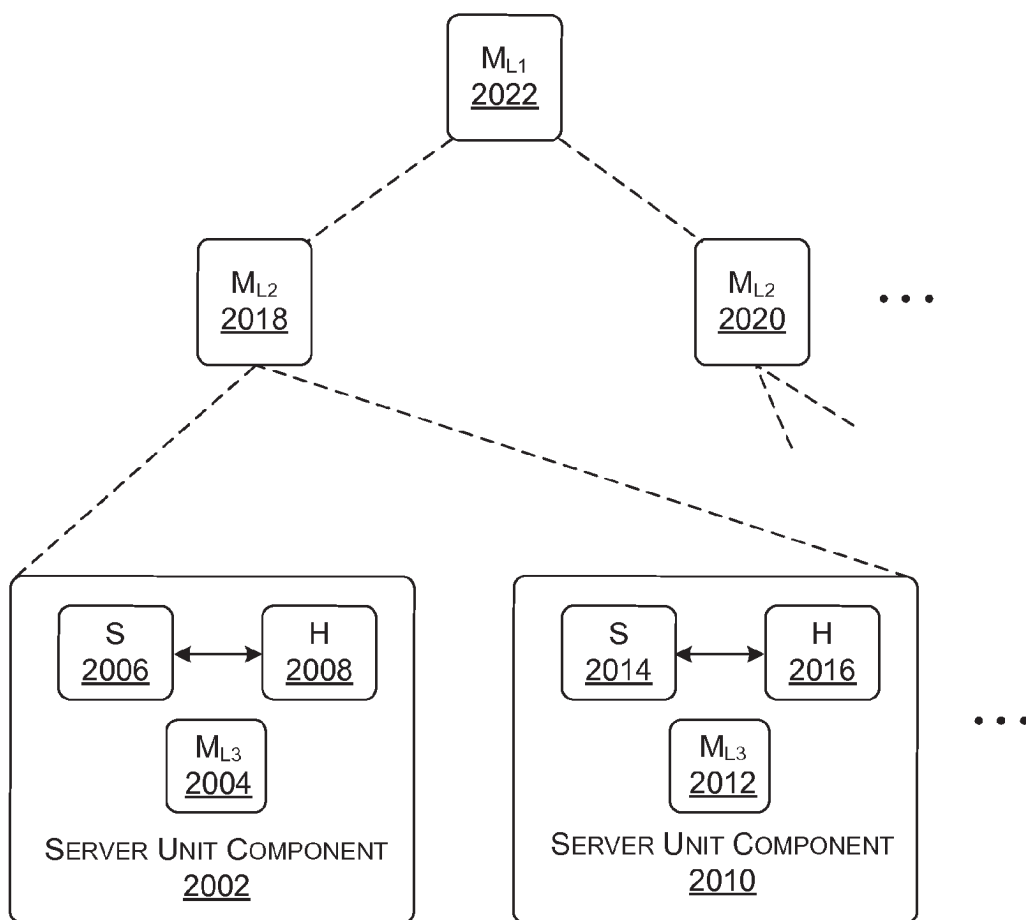


FIG. 20

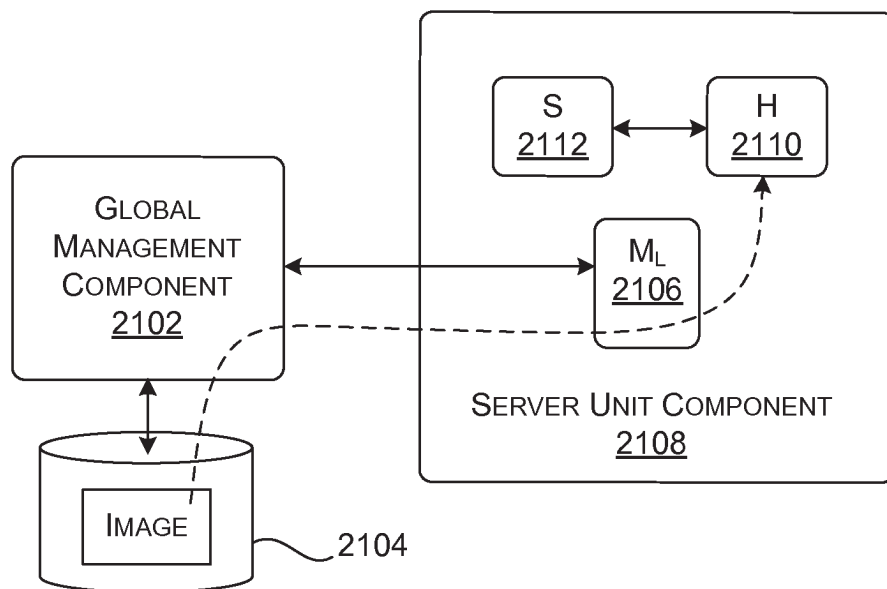


FIG. 21

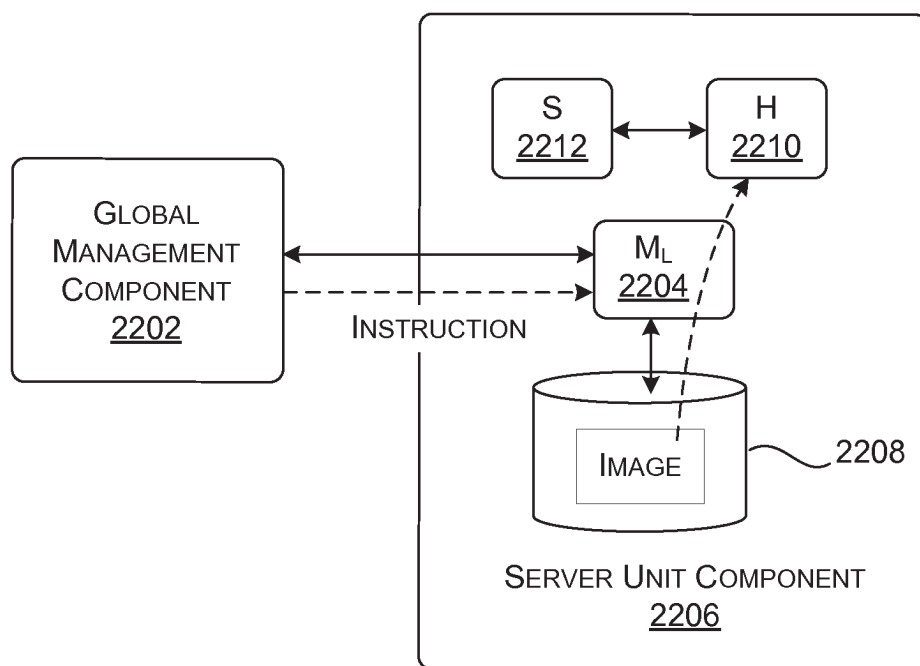


FIG. 22

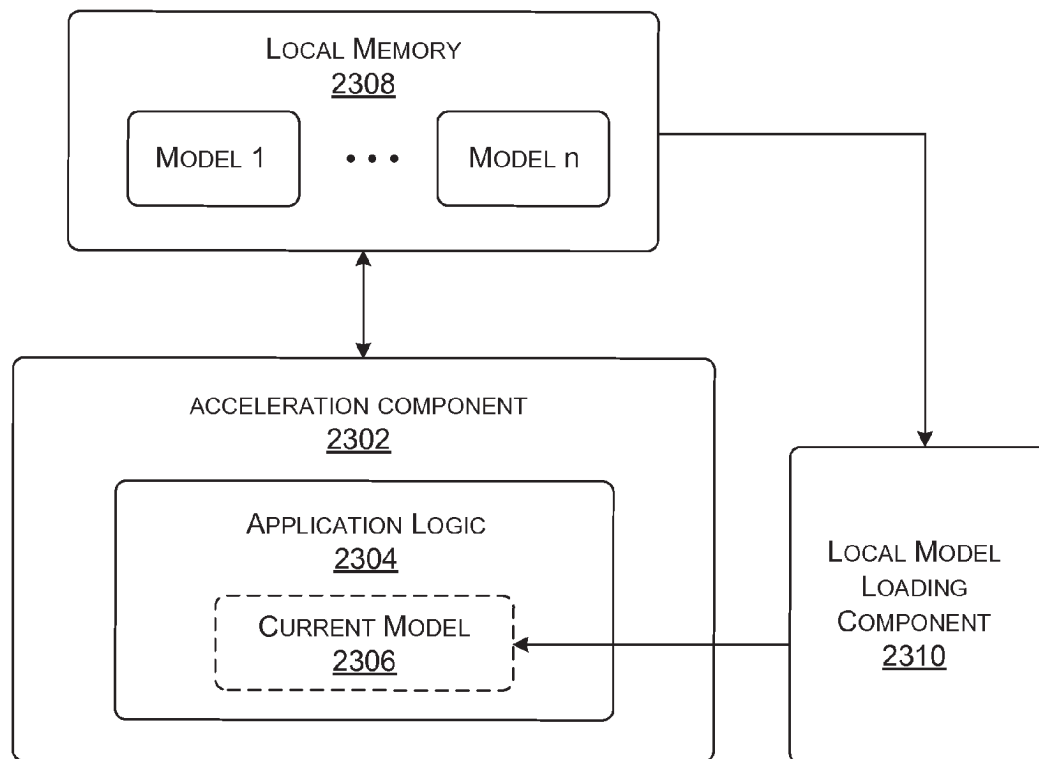


FIG. 23

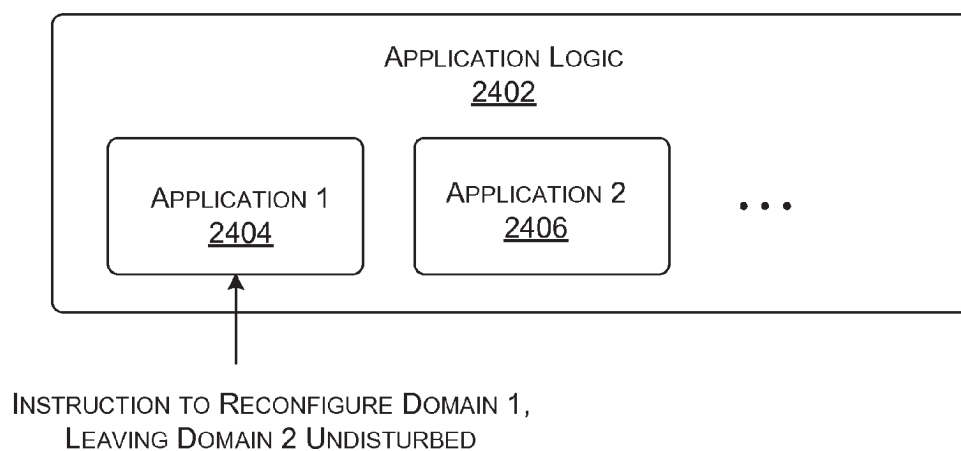


FIG. 24

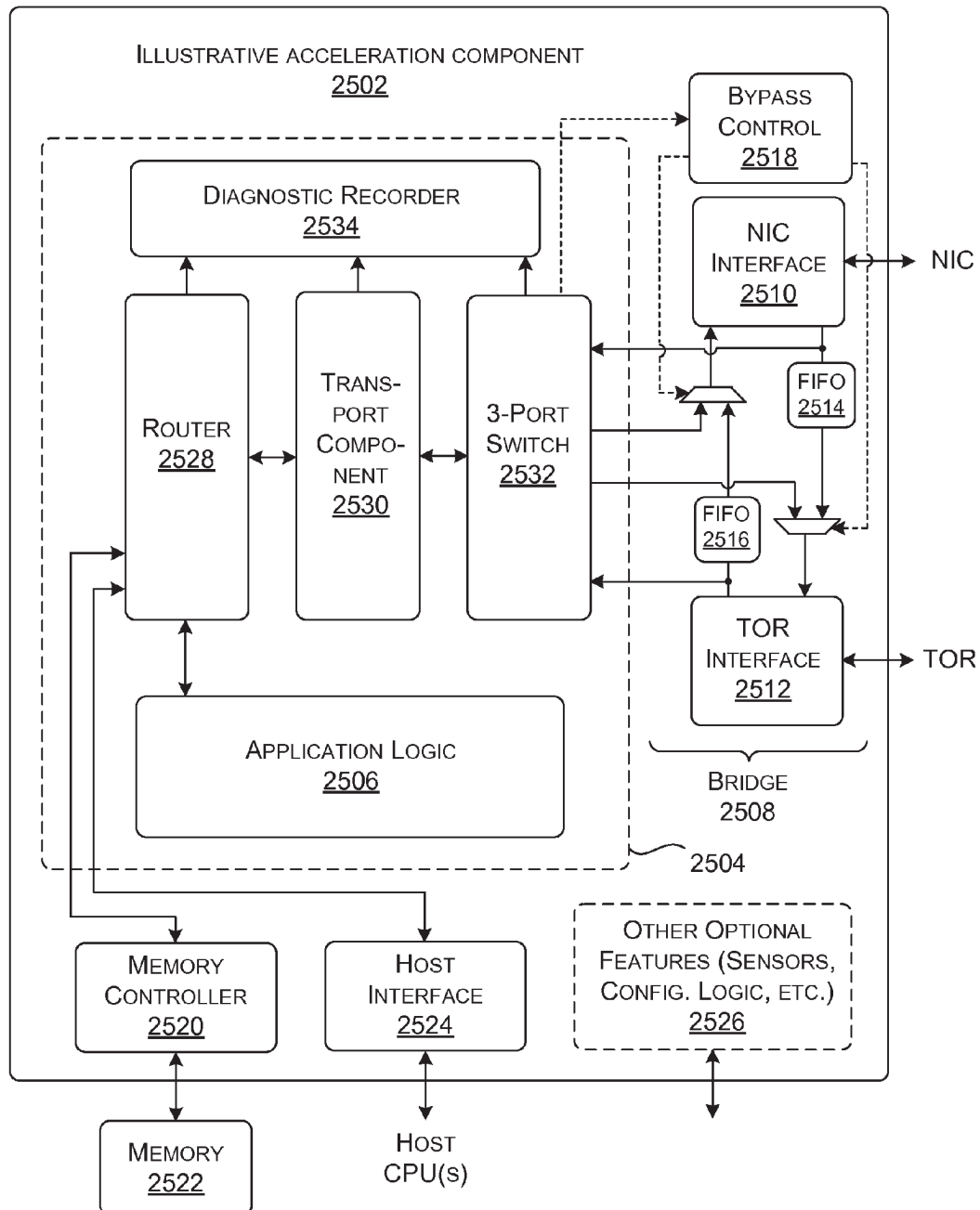


FIG. 25

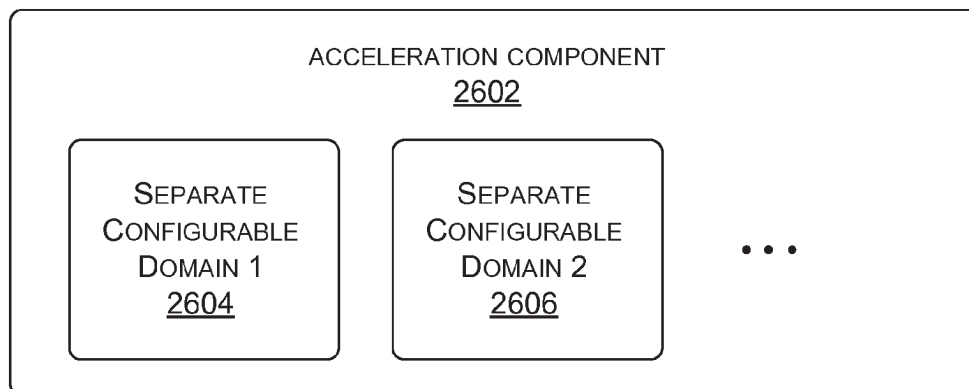


FIG. 26

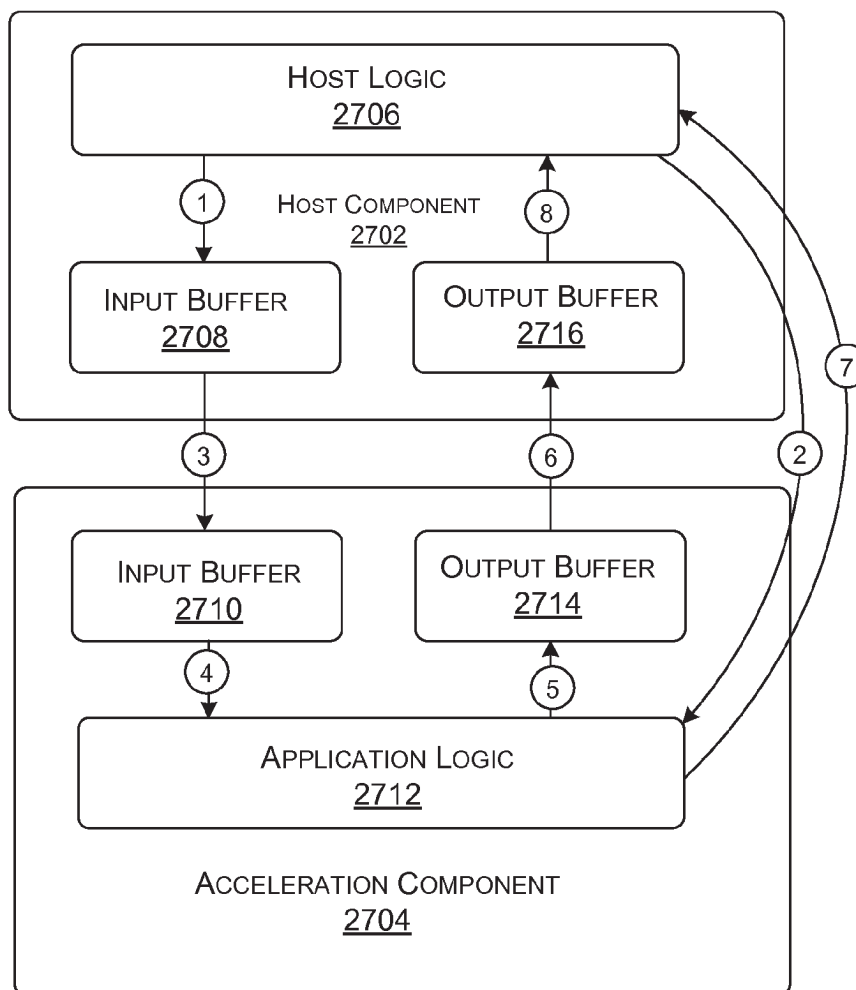


FIG. 27

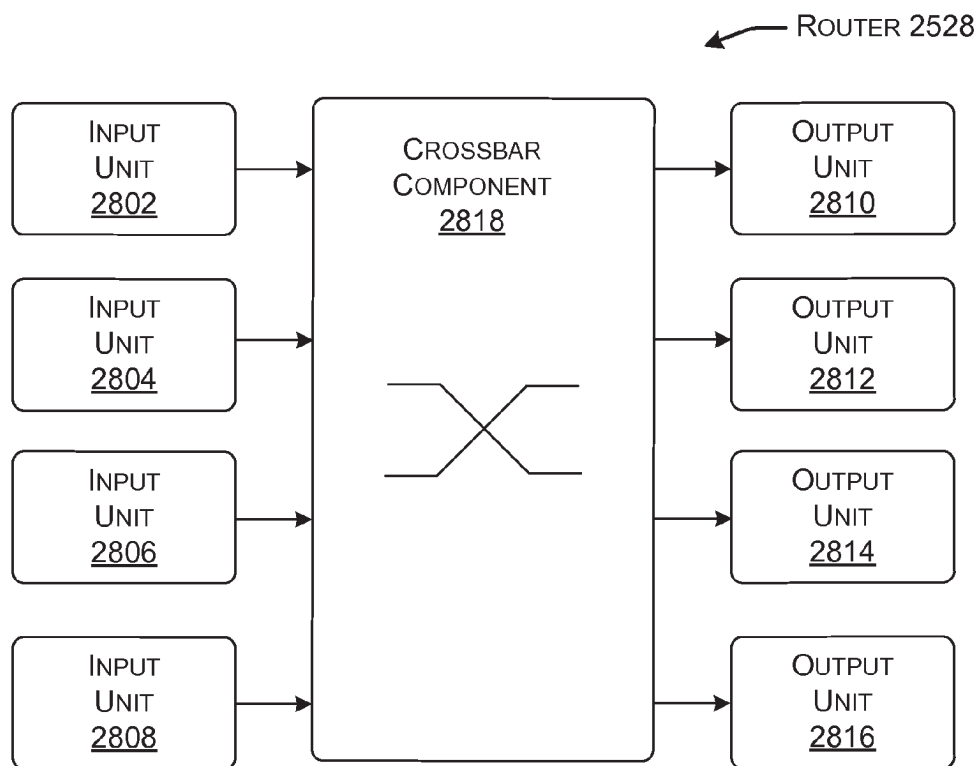


FIG. 28

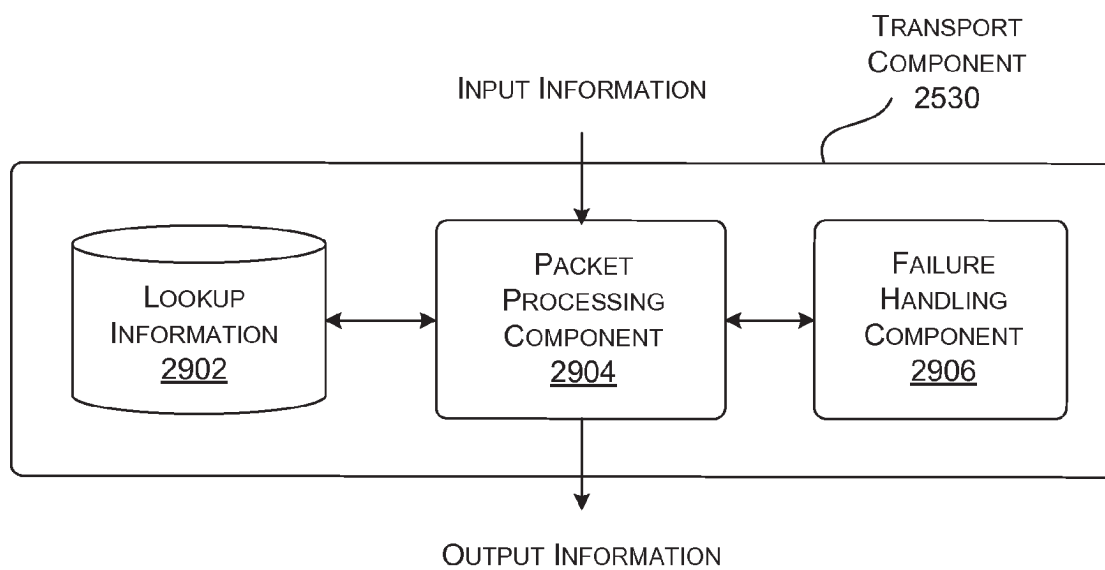


FIG. 29

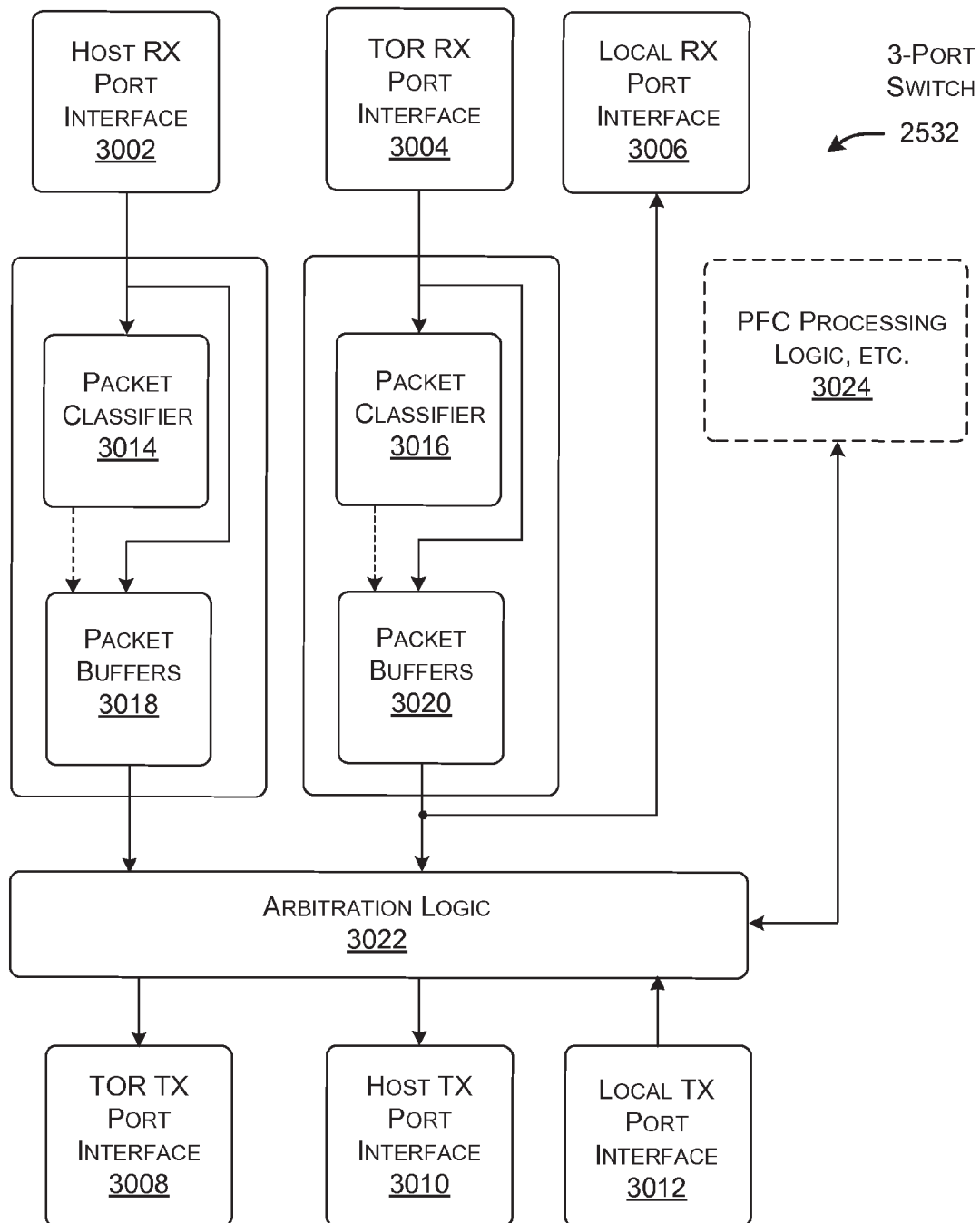


FIG. 30

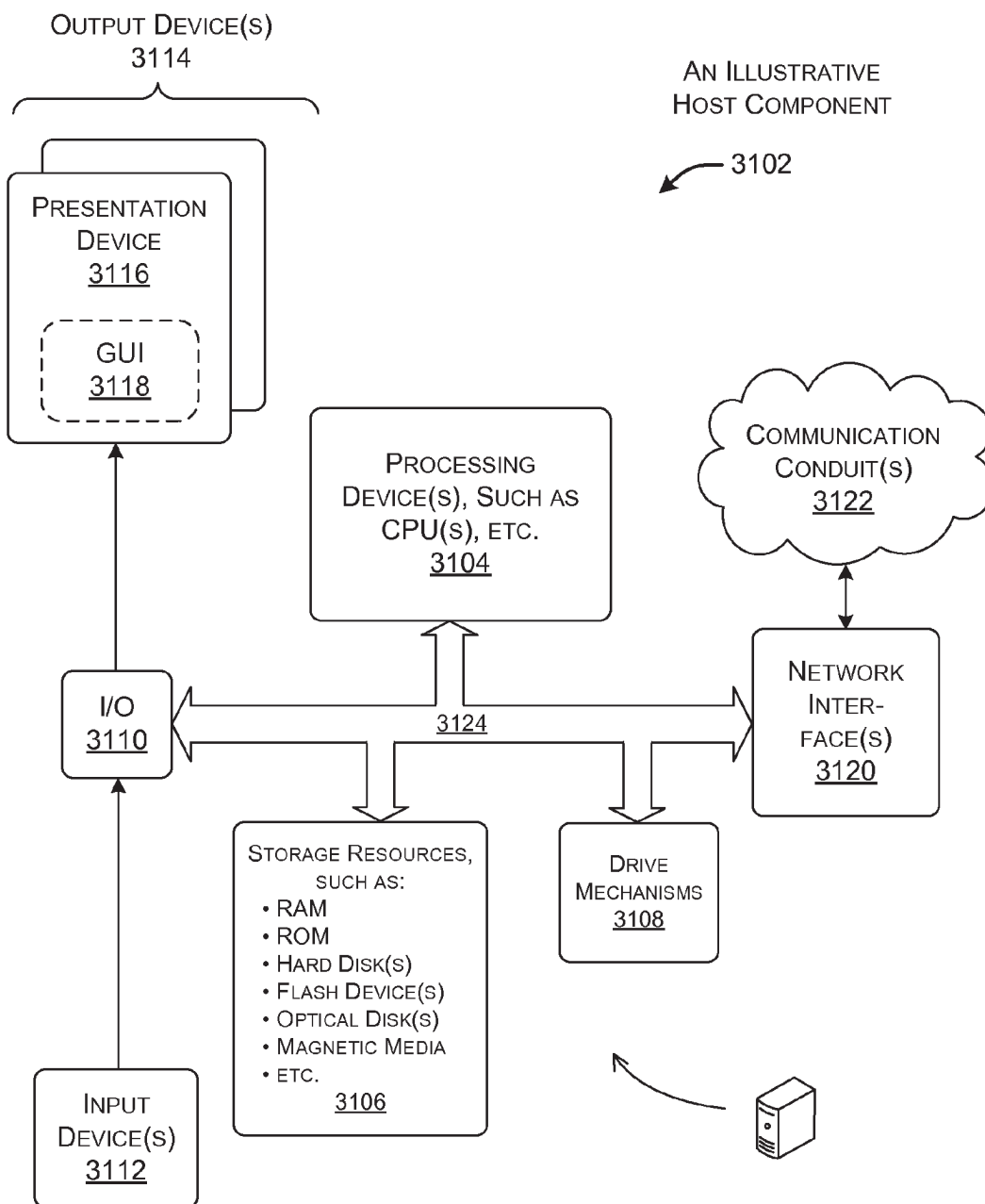


FIG. 31

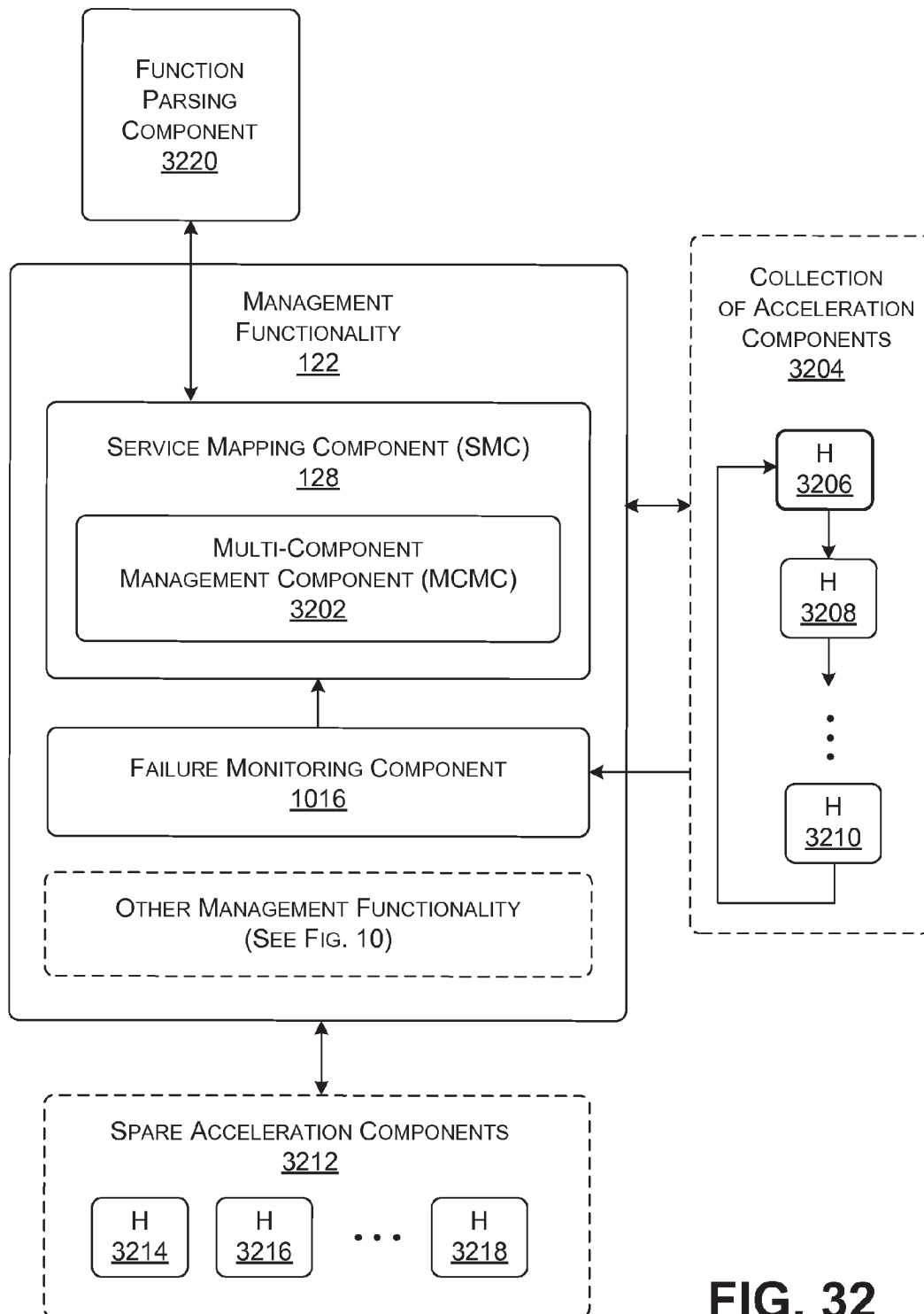


FIG. 32

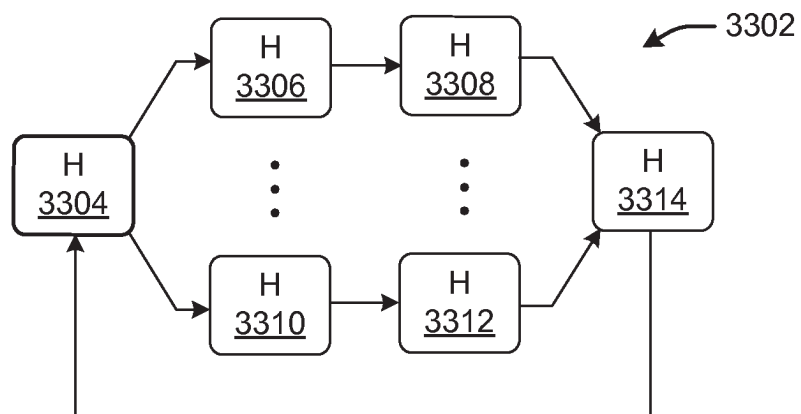


FIG. 33

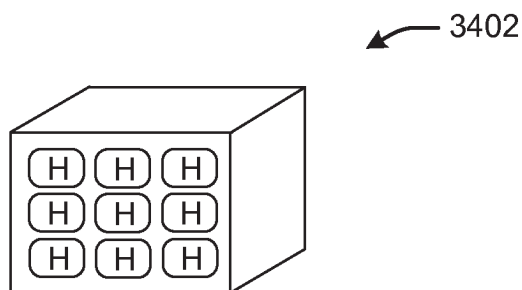


FIG. 34

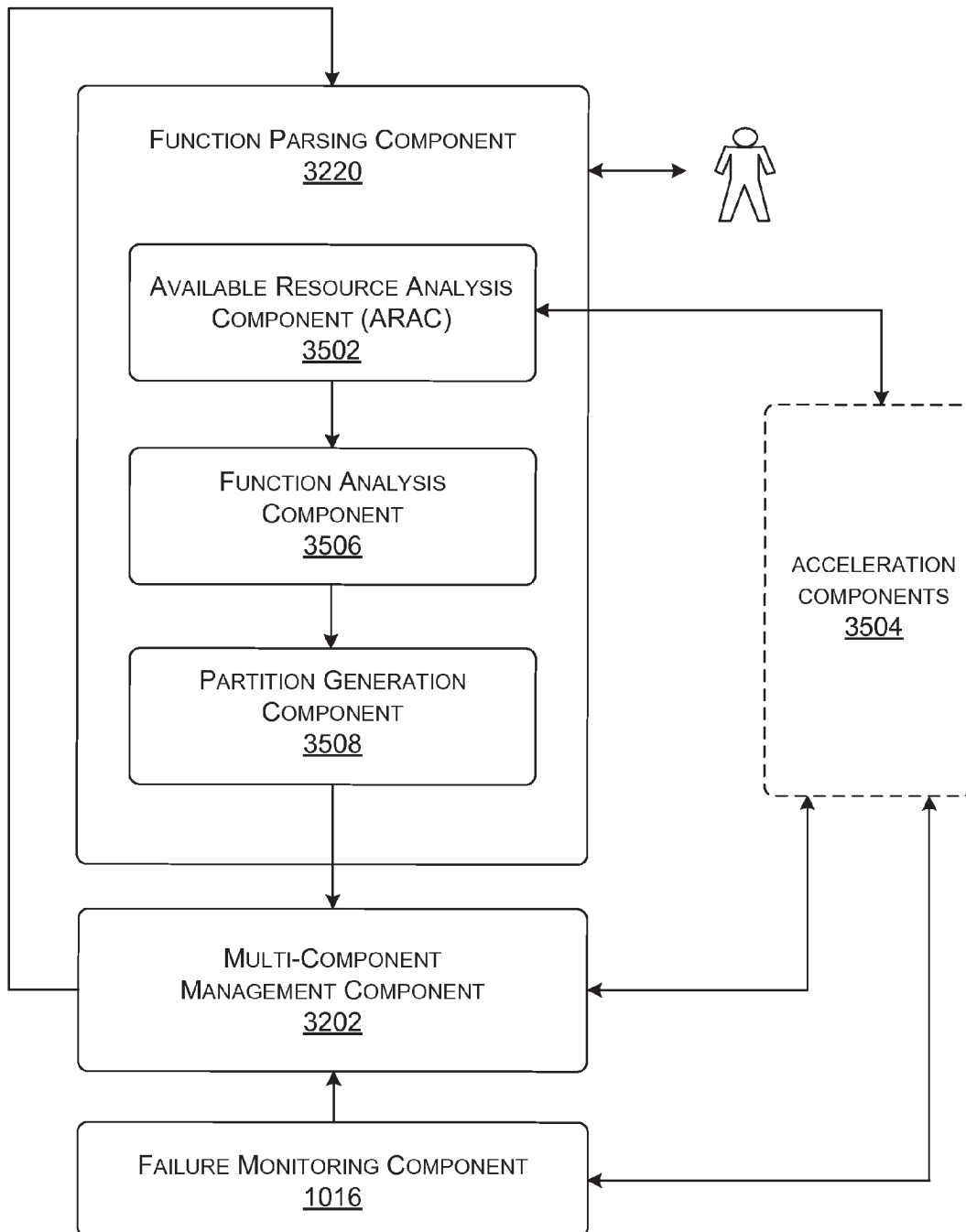


FIG. 35

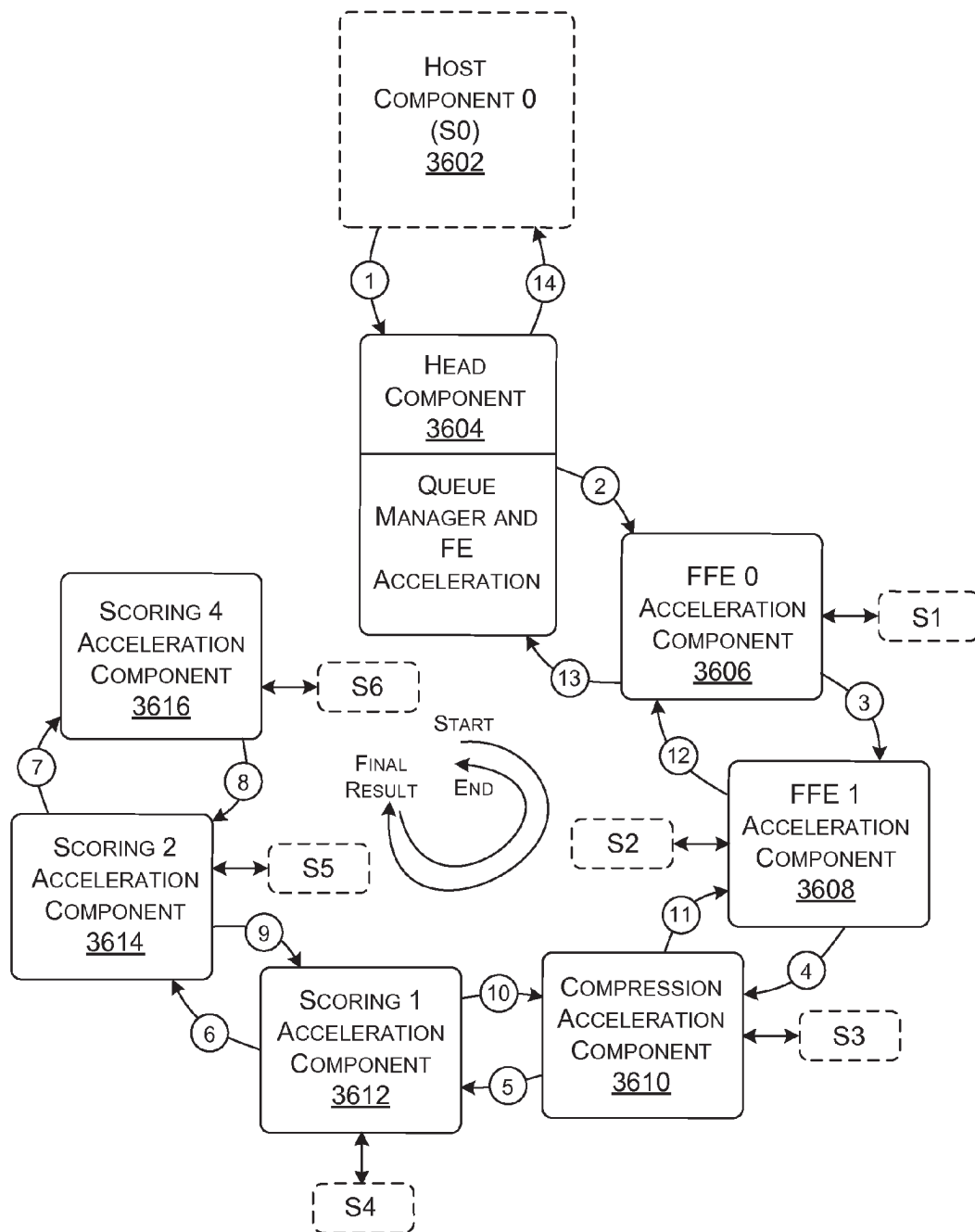


FIG. 36

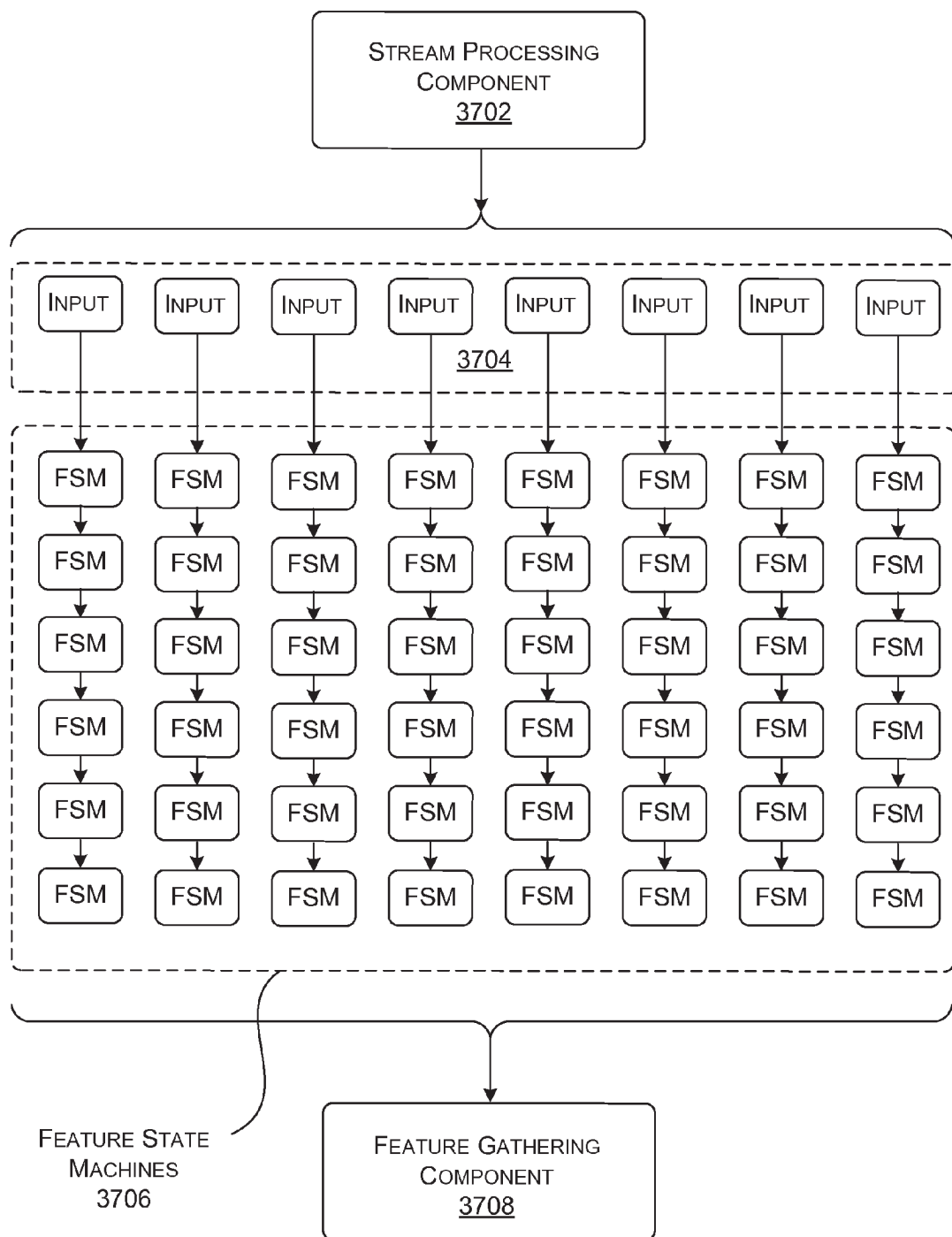


FIG. 37

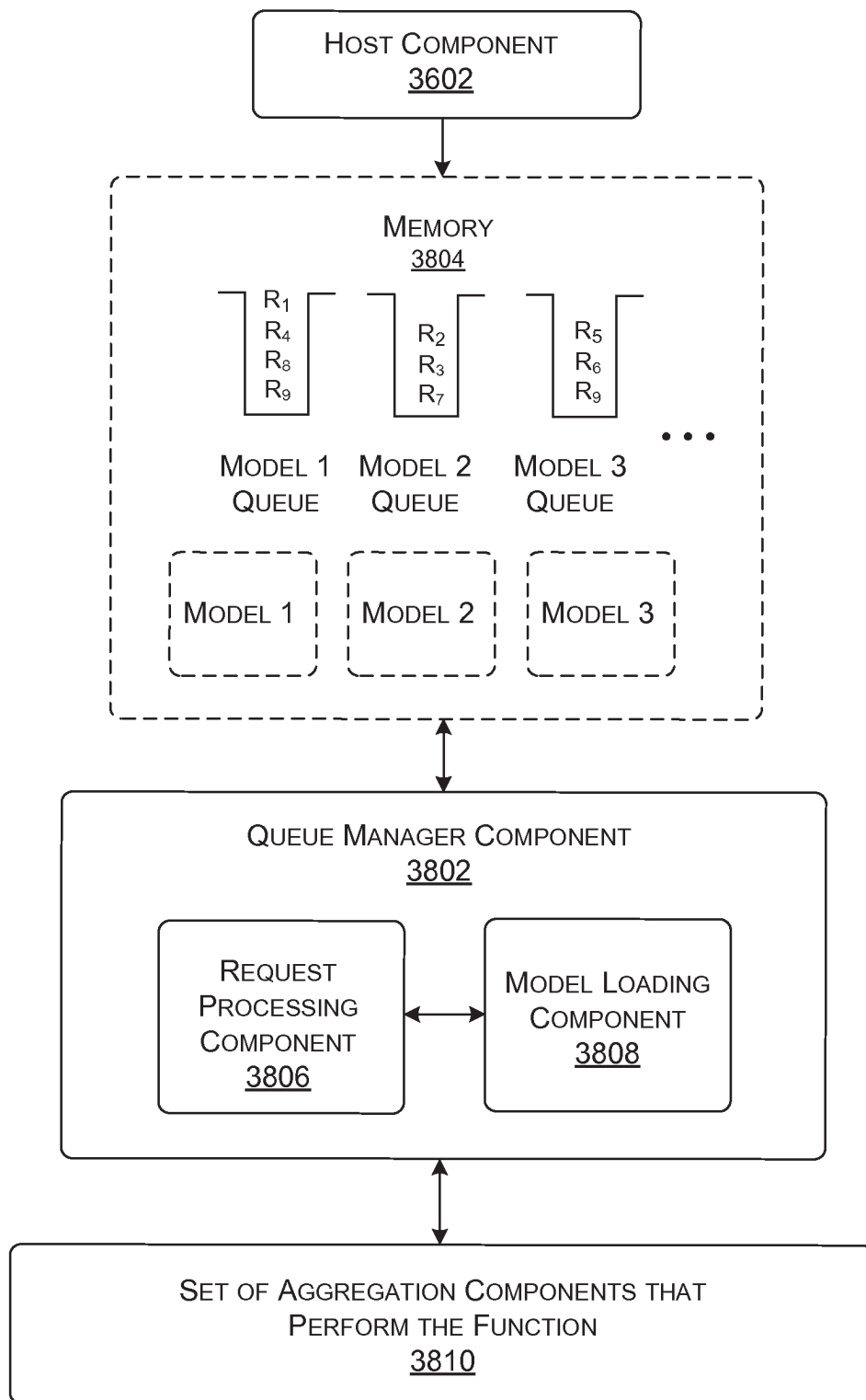


FIG. 38

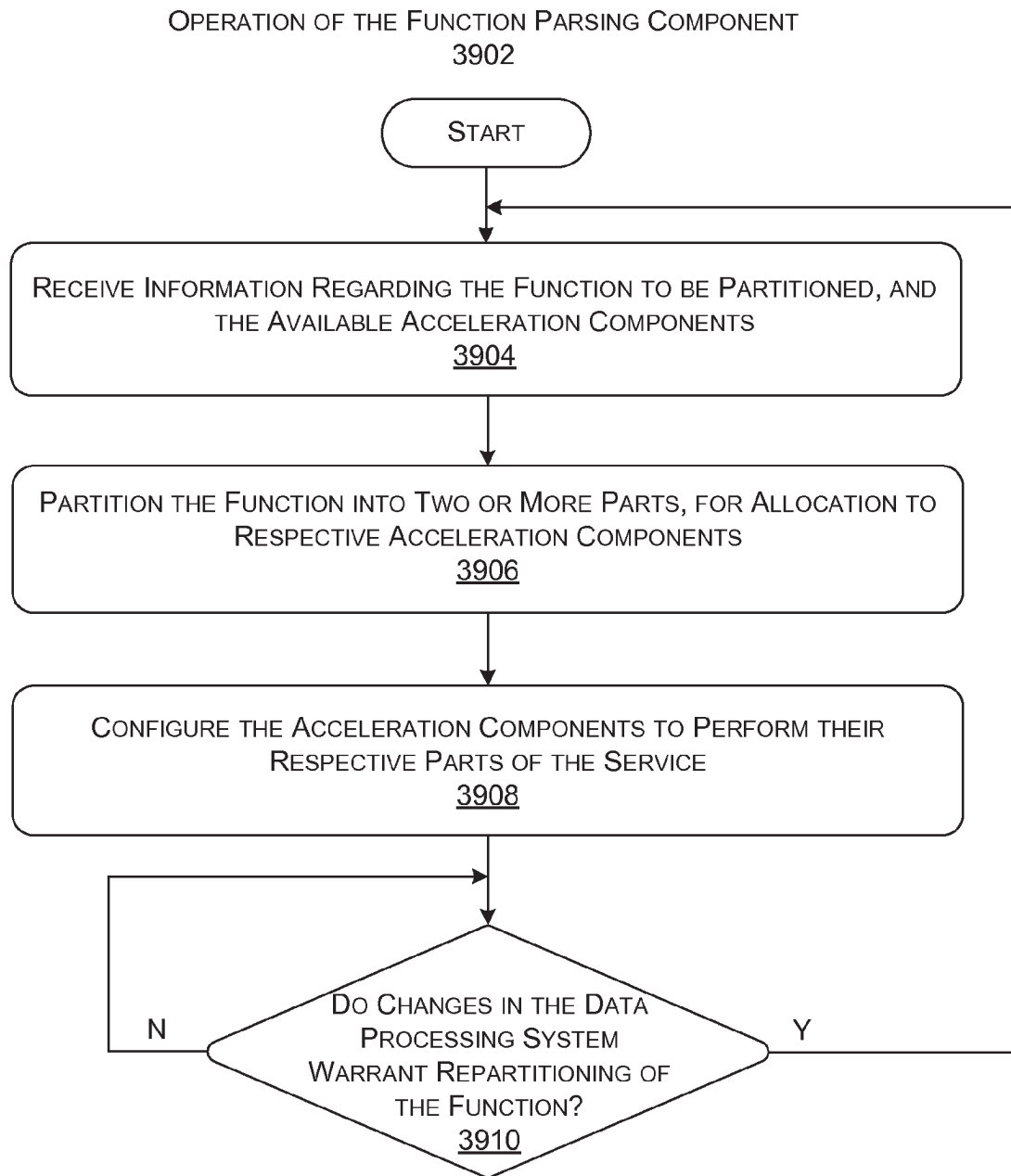


FIG. 39

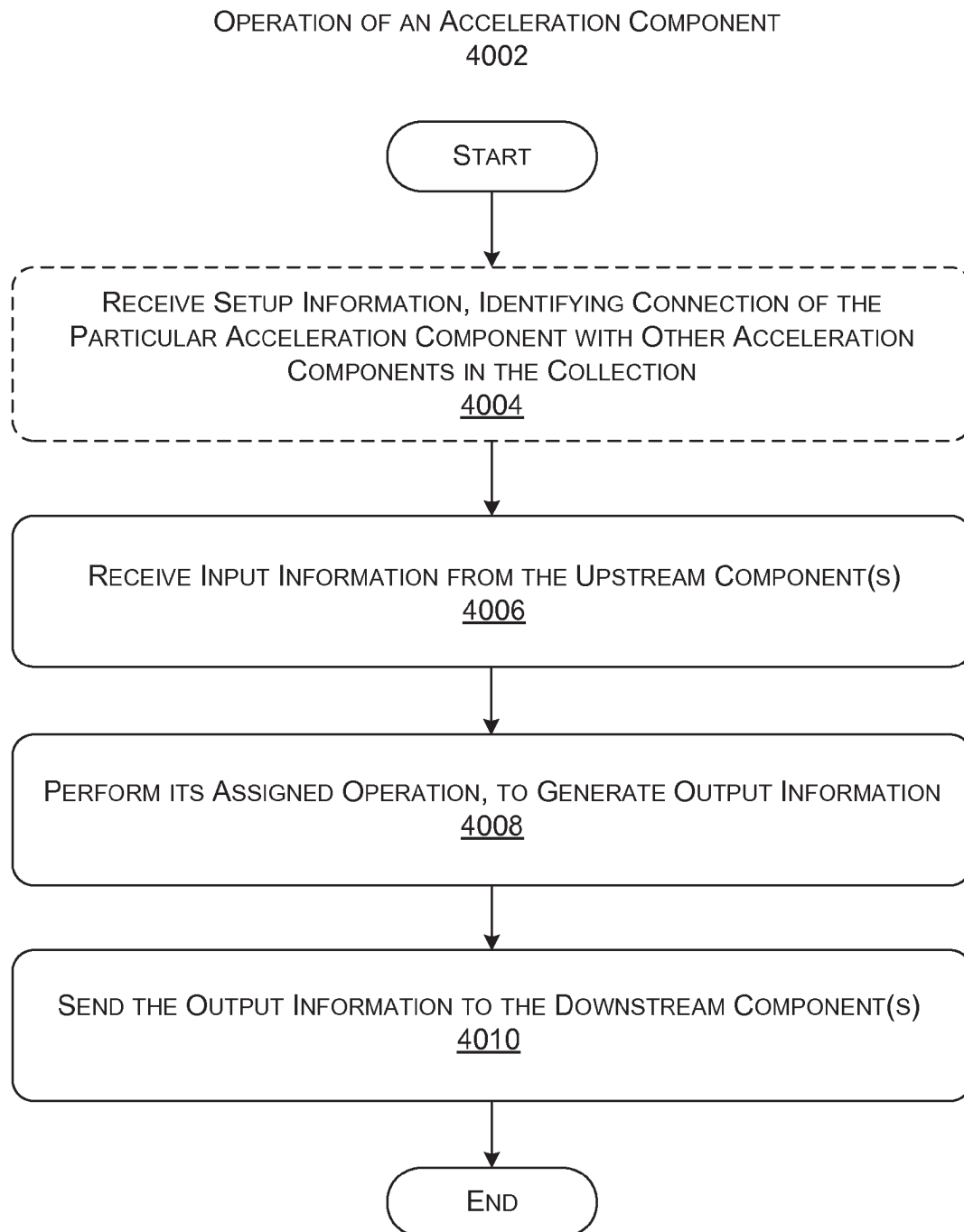


FIG. 40

OPERATION OF THE MANAGEMENT FUNCTIONALITY UPON FAILURES
4102

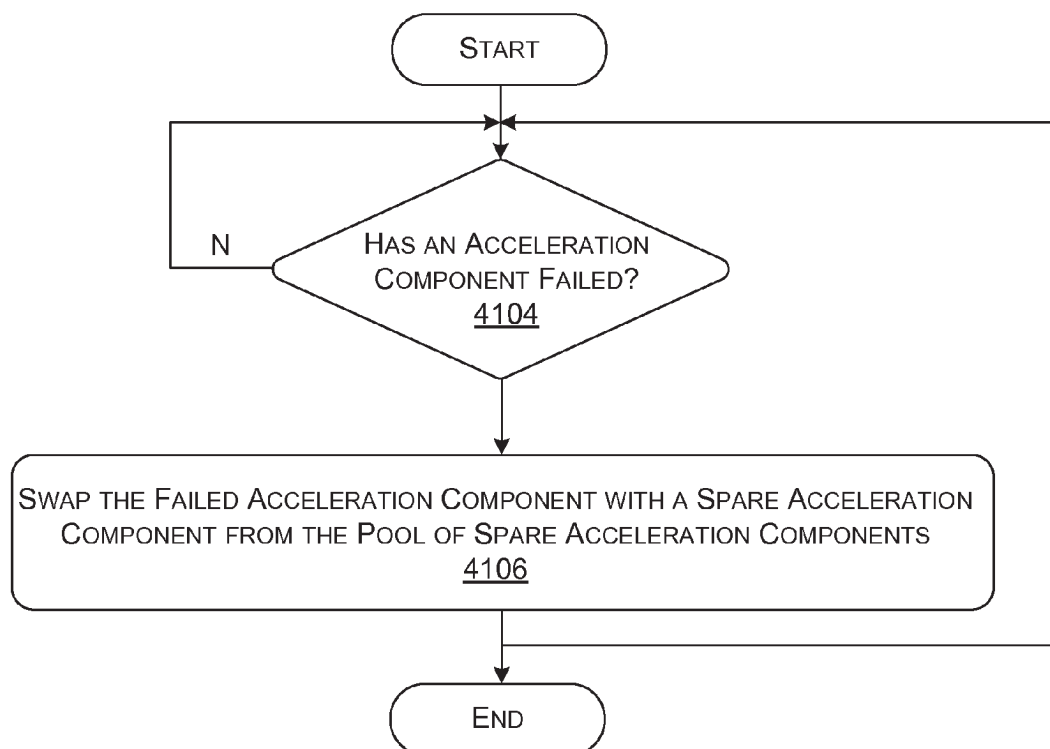


FIG. 41

IMPLEMENTING A SERVICE USING PLURAL ACCELERATION COMPONENTS

[0001] This application claims the benefit of U.S. Provisional Application No. 62/149,488 (the '488 application), filed Apr. 17, 2015. The '488 application is incorporated by reference herein in its entirety.

BACKGROUND

[0002] The computing industry faces increasing challenges in its efforts to improve the speed and efficiency of software-driven computing devices, e.g., due to power limitations and other factors. Software-driven computing devices employ one or more central processing units (CPUs) that process machine-readable instructions in a conventional temporal manner. To address this issue, the computing industry has proposed using hardware acceleration components (such as field-programmable gate arrays (FPGAs)) to supplement the processing performed by software-driven computing devices. However, software-driven computing devices and hardware acceleration components are dissimilar types of devices having fundamentally different architectures, performance characteristics, power requirements, program configuration paradigms, interface features, and so on. It is thus a challenging task to integrate these two types of devices together in a manner that satisfies the various design requirements of a particular data processing environment.

SUMMARY

[0003] A data processing system is described herein that includes two or more software-driven host components that collectively provide a software plane. The data processing system further includes two or more hardware acceleration components that collectively provide a hardware acceleration plane. The hardware acceleration plane implements one or more services, including at least one multi-component service. The multi-component service has plural parts, and is implemented on a collection of two or more hardware acceleration components, where each hardware acceleration component in the collection implements a corresponding part of the multi-component service. Each hardware acceleration component in the collection is configured to interact with other hardware acceleration components in the collection without involvement from any host component.

[0004] A function parsing component is also described herein that determines a manner of parsing a function into the plural parts of the multi-component service. The function parsing component can operate in an automated or semi-automated manner.

[0005] The above-summarized functionality can be manifested in various types of systems, devices, components, methods, computer readable storage media, data structures, graphical user interface presentations, articles of manufacture, and so on.

[0006] This Summary is provided to introduce a selection of concepts in a simplified form; these concepts are further described below in the Detailed Description. This Summary is not intended to identify key features or essential features of the claimed subject matter, nor is it intended to be used to limit the scope of the claimed subject matter.

BRIEF DESCRIPTION OF THE DRAWINGS

[0007] FIG. 1 shows an overview of a data processing system that includes a software plane and a hardware acceleration plane.

[0008] FIG. 2 shows a first example of the operation of the data processing system of FIG. 1.

[0009] FIG. 3 shows a second example of the operation of the data processing system of FIG. 1.

[0010] FIG. 4 shows one implementation of the data processing system of FIG. 1, corresponding to a data center.

[0011] FIG. 5 is a more encompassing depiction of the data center implementation of FIG. 4.

[0012] FIG. 6 shows an alternative way of implementing a server unit component, compared to that shown in FIG. 4.

[0013] FIG. 7 shows yet another way of implementing a server unit component compared to that shown in FIG. 4.

[0014] FIG. 8 shows an alternative data processing system compared to that shown in FIG. 1, e.g., which uses a different network infrastructure compared to that shown in FIG. 1.

[0015] FIG. 9 is a flowchart that shows one manner of operation of the data processing system of FIG. 1.

[0016] FIG. 10 shows an overview of one implementation of management functionality that is used to manage the data processing system of FIG. 1.

[0017] FIG. 11 provides an overview of one request-driven manner of operation of a service mapping component (SMC), which is a component of the management functionality of FIG. 10.

[0018] FIGS. 12-15 show different respective options for handling requests for services made by instances of tenant functionality that reside on a host component.

[0019] FIG. 16 provides an overview of another, background-related, manner of operation of the SMC of FIG. 10.

[0020] FIGS. 17-20 show different respective architectures for physically implementing the management functionality of FIG. 10.

[0021] FIGS. 21-24 show different respective strategies for configuring a hardware acceleration component in the data processing system of FIG. 1.

[0022] FIG. 25 shows one manner of implementing a hardware acceleration component of FIG. 1.

[0023] FIG. 26 shows a hardware acceleration component including separate configurable domains.

[0024] FIG. 27 shows functionality for performing data transfer between a local host component and an associated local hardware acceleration component.

[0025] FIG. 28 shows one implementation of a router introduced in FIG. 25.

[0026] FIG. 29 shows one implementation of a transport component introduced in FIG. 25.

[0027] FIG. 30 shows one implementation of a 3-port switch introduced in FIG. 25.

[0028] FIG. 31 shows one implementation of a host component shown in FIG. 1.

[0029] FIG. 32 provides an overview of functionality for generating and applying a multi-component service; that functionality, in turn, includes a multi-component management component (MCMC).

[0030] FIG. 33 shows one type of collection of hardware acceleration components that may be produced and applied by the functionality of FIG. 32.

[0031] FIG. 34 shows another type of collection of hardware acceleration components that may be produced and applied by the functionality of FIG. 32.

[0032] FIG. 35 shows one implementation of a function parsing component that produces a multi-component service.

[0033] FIG. 36 shows a more detailed example of an illustrative multi-component service, implemented using a collection of hardware acceleration components.

[0034] FIG. 37 shows functionality that performs processing in one of the stages of the multi-component service of FIG. 36.

[0035] FIG. 38 shows functionality for swapping models in the collection of hardware acceleration components of FIG. 36, to accommodate requests that are associated with different models.

[0036] FIG. 39 is a flowchart that shows one manner of operation of the function parsing component of FIG. 35.

[0037] FIG. 40 is a flowchart that shows the operation of one hardware acceleration component within a collection of hardware acceleration components that implements a multi-component service.

[0038] FIG. 41 is a flowchart that shows one way of handling a failure in a collection of hardware acceleration components that implements a multi-component service.

[0039] The same numbers are used throughout the disclosure and figures to reference like components and features. Series 100 numbers refer to features originally found in FIG. 1, series 200 numbers refer to features originally found in FIG. 2, series 300 numbers refer to features originally found in FIG. 3, and so on.

DETAILED DESCRIPTION

[0040] This disclosure is organized as follows. Section A describes an illustrative data processing system that includes a hardware acceleration plane and a software plane. Section B describes management functionality that is used to manage the data processing system of Section A. Section C sets forth one implementation of an illustrative hardware acceleration component in the hardware acceleration plane. And Section D describes the application of the data processing system to implement a multi-component service; a multi-component service refers to a service that is built using plural hardware acceleration components.

[0041] As a preliminary matter, some of the figures describe concepts in the context of one or more structural components, variously referred to as functionality, modules, features, elements, etc. The various components shown in the figures can be implemented in any manner by any physical and tangible mechanisms, for instance, by software running on computer equipment, hardware (e.g., chip-implemented logic functionality), etc., and/or any combination thereof. In one case, the illustrated separation of various components in the figures into distinct units may reflect the use of corresponding distinct physical and tangible components in an actual implementation. Alternatively, or in addition, any single component illustrated in the figures may be implemented by plural actual physical components. Alternatively, or in addition, the depiction of any two or more separate components in the figures may reflect different functions performed by a single actual physical component.

[0042] Other figures describe the concepts in flowchart form. In this form, certain operations are described as constituting distinct blocks performed in a certain order.

Such implementations are illustrative and non-limiting. Certain blocks described herein can be grouped together and performed in a single operation, certain blocks can be broken apart into plural component blocks, and certain blocks can be performed in an order that differs from that which is illustrated herein (including a parallel manner of performing the blocks). The blocks shown in the flowcharts can be implemented in any manner by any physical and tangible mechanisms, for instance, by software running on computer equipment, hardware (e.g., chip-implemented logic functionality), etc., and/or any combination thereof.

[0043] As to terminology, the phrase “configured to” encompasses any way that any kind of physical and tangible functionality can be constructed to perform an identified operation. The functionality can be configured to perform an operation using, for instance, software running on computer equipment, hardware (e.g., chip-implemented logic functionality), etc., and/or any combination thereof.

[0044] The term “logic” encompasses any physical and tangible functionality for performing a task. For instance, each operation illustrated in the flowcharts corresponds to a logic component for performing that operation. An operation can be performed using, for instance, software running on computer equipment, hardware (e.g., chip-implemented logic functionality), etc., and/or any combination thereof. When implemented by computing equipment, a logic component represents an electrical component that is a physical part of the computing system, however implemented.

[0045] Any of the storage resources described herein, or any combination of the storage resources, may be regarded as a computer readable medium. In many cases, a computer readable medium represents some form of physical and tangible entity. The term computer readable medium also encompasses propagated signals, e.g., transmitted or received via physical conduit and/or air or other wireless medium, etc. However, the specific terms “computer readable storage medium” and “computer readable medium device” expressly exclude propagated signals per se, while including all other forms of computer readable media.

[0046] The following explanation may identify one or more features as “optional.” This type of statement is not to be interpreted as an exhaustive indication of features that may be considered optional; that is, other features can be considered as optional, although not explicitly identified in the text. Further, any description of a single entity is not intended to preclude the use of plural such entities; similarly, a description of plural entities is not intended to preclude the use of a single entity. Further, while the description may explain certain features as alternative ways of carrying out identified functions or implementing identified mechanisms, the features can also be combined together in any combination. Finally, the terms “exemplary” or “illustrative” refer to one implementation among potentially many implementations.

[0047] A. Overview

[0048] FIG. 1 shows an overview of a data processing system 102 that includes a software plane 104 and a hardware acceleration plane 106. The software plane 104 includes a collection of software-driven components (each denoted by the symbol “S” in FIG. 1), while the hardware plane includes a collection of hardware acceleration components (each denoted by the symbol “H” in FIG. 1). For instance, each host component may correspond to a server computer that executes machine-readable instructions using

one or more central processing units (CPUs). Each CPU, in turn, may execute the instructions on one or more hardware threads. Each hardware acceleration component, on the other hand, may correspond to hardware logic for implementing functions, such as a field-programmable gate array (FPGA) device, a massively parallel processor array (MPPA) device, a graphics processing unit (GPU), an application-specific integrated circuit (ASIC), a multiprocessor System-on-Chip (MPSoC), and so on.

[0049] The term “hardware” acceleration component is also intended to broadly encompass different ways of leveraging a hardware device to perform a function, including, for instance, at least: a) a case in which at least some tasks are implemented in hard ASIC logic or the like; b) a case in which at least some tasks are implemented in soft (configurable) FPGA logic or the like; c) a case in which at least some tasks run as software on FPGA software processor overlays or the like; d) a case in which at least some tasks run on MPPAs of soft processors or the like; e) a case in which at least some tasks run as software on hard ASIC processors or the like, and so on, or any combination thereof. Likewise, the data processing system 102 can accommodate different manifestations of software-driven devices in the software plane 104.

[0050] To simplify repeated reference to hardware acceleration components, the following explanation will henceforth refer to these devices as simply “acceleration components.” Further, the following explanation will present a primary example in which the acceleration components correspond to FPGA devices, although, as noted, the data processing system 102 may be constructed using other types of acceleration components. Further, the hardware acceleration plane 106 may be constructed using a heterogeneous collection of acceleration components, including different types of FPGA devices having different respective processing capabilities and architectures, a mixture of FPGA devices and other devices, and so on.

[0051] A host component generally performs operations using a temporal execution paradigm, e.g., by using each of its CPU hardware threads to execute machine-readable instructions, one after the other. In contrast, an acceleration component may perform operations using a spatial paradigm, e.g., by using a large number of parallel logic elements to perform computational tasks. Thus, an acceleration component can perform some operations in less time compared to a software-driven host component. In the context of the data processing system 102, the “acceleration” qualifier associated with the term “acceleration component” reflects its potential for accelerating the functions that are performed by the host components.

[0052] In one example, the data processing system 102 corresponds to a data center environment that includes a plurality of computer servers. The computer servers correspond to the host components in the software plane 104 shown in FIG. 1. In other cases, the data processing system 102 corresponds to an enterprise system. In other cases, the data processing system 102 corresponds to a user device or appliance which uses at least one host component that has access to two or more acceleration components, etc. These examples are cited by way of example, not limitation; still other applications are possible.

[0053] In one implementation, each host component in the data processing system 102 is coupled to at least one acceleration component through a local link. That funda-

mental unit of processing equipment is referred to herein as a “server unit component” because that equipment may be grouped together and maintained as a single serviceable unit within the data processing system 102 (although not necessarily so). The host component in the server unit component is referred to as the “local” host component to distinguish it from other host components that are associated with other server unit components. Likewise, the acceleration component(s) of the server unit component are referred to as the “local” acceleration component(s) to distinguish them from other acceleration components that are associated with other server unit components.

[0054] For example, FIG. 1 shows an illustrative local host component 108 that is coupled to a local acceleration component 110 through a local link 112 (such as, as will be described below, a Peripheral Component Interconnect Express (PCIe) link). That pairing of the local host component 108 and the local acceleration component 110 forms at least part of a single server unit component. More generally, FIG. 1 shows that the software plane 104 is coupled to the hardware acceleration plane through many individual local links, which FIG. 1 collectively refers to as a local_H-to-local_S coupling 114.

[0055] The local host component 108 may further indirectly communicate with any other remote acceleration component in the hardware acceleration plane 106. For example, the local host component 108 has access to a remote acceleration component 116 via the local acceleration component 110. More specifically, the local acceleration component 110 communicates with the remote acceleration component 116 via a link 118.

[0056] In one implementation, a common network 120 is used to couple host components in the software plane 104 to other host components, and to couple acceleration components in the hardware acceleration plane 106 to other acceleration components. That is, two host components may use the same network 120 to communicate with each other as do two acceleration components. As another feature, the interaction among host components in the software plane 104 is independent of the interaction among acceleration components in the hardware acceleration plane 106. This means, for instance, that two or more acceleration components may communicate with each other in a transparent manner from the perspective of host components in the software plane 104, outside the direction of the host components, and without the host components being “aware” of the particular interactions that are taking place in the hardware acceleration plane 106. A host component may nevertheless initiate interactions that take place in the hardware acceleration plane 106 by issuing a request for a service that is hosted by the hardware acceleration plane 106.

[0057] According to one non-limiting implementation, the data processing system 102 uses the Ethernet protocol to transmit IP packets over the common network 120. In one implementation, each local host component in a server unit component is given a single physical IP address. The local acceleration component in the same server unit component may adopt the same IP address. The server unit component can determine whether an incoming packet is destined for the local host component as opposed to the local acceleration component in different ways. For example, packets that are destined for the local acceleration component can be formulated as user datagram protocol (UDP) packets specifying a specific port; host-directed packets, on the other

hand, are not formulated in this way. In another case, packets belonging to the acceleration plane **106** can be distinguished from packets belonging to the software plane **104** based on the value of a status flag in each of the packets (e.g., in the header or body of a packet).

[0058] In view of the above characteristic, the data processing system **102** may be conceptualized as forming two logical networks that share the same physical communication links. The packets associated with the two logical networks may be distinguished from each other by their respective traffic classes in the manner described above. But in other implementations (e.g., as described below with respect to FIG. 8), the data processing system **102** may use two distinct physical networks to handle host-to-host traffic and hardware-to-hardware traffic, respectively. Further, in implementations that do use the common network **120**, the host-to-host network infrastructure need not be entirely identical to the hardware-to-hardware network infrastructure; that is, these two infrastructures are common in the sense that most of their network resources are shared, but not necessarily all of their network resources are shared.

[0059] Finally, management functionality **122** serves to manage the operations of the data processing system **102**. As will be set forth in greater detail in Section B (below), the management functionality **122** can be physically implemented using different control architectures. For example, in one control architecture, the management functionality **122** may include plural local management components that are coupled to one or more global management components.

[0060] By way of introduction to Section B, the management functionality **122** can include a number of sub-components that perform different respective logical functions (which can be physically implemented in different ways). A location determination component **124**, for instance, identifies the current locations of services within the data processing system **102**, based on current allocation information stored in a data store **126**. As used herein, a service refers to any function that is performed by the data processing system **102**. For example, one service may correspond to an encryption function. Another service may correspond to a document ranking function. Another service may correspond to a data compression function, and so on.

[0061] In operation, the location determination component **124** may receive a request for a service. In response, the location determination component **124** returns an address associated with the service, if that address is present in the data store **126**. The address may identify a particular acceleration component that hosts the requested service.

[0062] A service mapping component (SMC) **128** maps services to particular acceleration components. The SMC **128** may operate in at least two modes depending on the type of triggering event that it receives which invokes its operation. In a first case, the SMC **128** processes requests for services made by instances of tenant functionality. An instance of tenant functionality may correspond to a software program running on a particular local host component, or, more specifically, a program executing on a virtual machine that, in turn, is associated with the particular local host component. That software program may request a service in the course of its execution. The SMC **128** handles the request by determining an appropriate component (or components) in the data processing system **102** to provide the service. Possible components for consideration include: a local acceleration component (associated with the local

host component from which the request originated); a remote acceleration component; and/or the local host component itself (whereupon the local host component will implement the service in software). The SMC **128** makes its determinations based on one or more mapping considerations, such as whether the requested service pertains to a line-rate service.

[0063] In another manner of operation, the SMC **128** generally operates in a background and global mode, allocating services to acceleration components based on global conditions in the data processing system **102** (rather than, or in addition to, handling individual requests from instances of tenant functionality). For example, the SMC **128** may invoke its allocation function in response to a change in demand that affects one or more services. In this mode, the SMC **128** again makes its determinations based on one or more mapping considerations, such as the historical demand associated with the services, etc.

[0064] The SMC **128** may interact with the location determination component **124** in performing its functions. For instance, the SMC **128** may consult the data store **126** when it seeks to determine the address of an already allocated service provided by an acceleration component. The SMC **128** can also update the data store **126** when it maps a service to one or more acceleration components, e.g., by storing the addresses of those acceleration components in relation to the service.

[0065] Although not shown in FIG. 1, a sub-component of the SMC **128** also manages multi-component services. A multi-component service is a service that is composed of plural parts. Plural respective acceleration components perform the respective parts.

[0066] Note that FIG. 1 illustrates, as a matter of convenience, that the management functionality **122** is separate from the components in the software plane **104** and the hardware plane **106**. But as will be clarified in Section B, any aspect of the management functionality **122** can be implemented using the resources of the software plane **104** and/or the hardware plane **106**. When implemented by the hardware plane **106**, the management functions can be accelerated like any service.

[0067] FIG. 2 shows a first example of the operation of the data processing system **102** of FIG. 1, corresponding to a single transaction, or part of a single transaction. In operation (1), a first host component **202** communicates with a second host component **204** in the course of performing a single computational task. The second host component **204** then requests the use of a service that is implemented in the hardware acceleration plane **106** (although the second host component **204** may not be “aware” of where the service is implemented, beyond that the service can be accessed at a specified address).

[0068] In many cases, a requested service is implemented on a single acceleration component (although there may be plural redundant such acceleration components to choose from among). But in the particular example of FIG. 2, the requested service corresponds to a multi-component service that is spread out over a collection (or cluster) of acceleration components, each of which performs an allocated part of the service. A graph structure may specify the manner by which the individual acceleration components are coupled together in the collection. In some implementations, the graph structure also identifies at least one head component. The head component corresponds to a point of contact by

which entities in the data processing system 102 may interact with the multi-component service in the hardware acceleration plane 106. The head component may also serve as an initial processing stage in a processing pipeline defined by the graph structure.

[0069] In the particular case of FIG. 2, assume that acceleration component 206 corresponds to the local acceleration component that is locally linked to the local host component 204, and that an acceleration component 208 is the head component of the multi-component service. In operations (2) and (3), the requesting host component 204 accesses the acceleration component 208 via its local acceleration component 206. The acceleration component 208 then performs its part of the multi-component service to generate an intermediate output result. In operation (4), the acceleration component 208 then invokes another acceleration component 210, which performs another respective part of the multi-component service, to generate a final result. In operations (5), (6), and (7), the hardware acceleration plane 106 successively forwards the final result back to the requesting host component 204, through the same chain of components set forth above but in the opposite direction. Note that the data flow operations described above, including the flow operations that define the return path, are cited by way of example, not limitation; other multi-component services may use other graph structures that specify any other flow paths. For example, the acceleration component 210 can forward the final result directly to the local acceleration component 206. Section D provides additional information regarding the construction and operation of multi-component services.

[0070] First, note that the operations that take place in the hardware acceleration plane 106 are performed in an independent manner of operations performed in the software plane 104. In other words, the host components in the software plane 104 do not manage the operations in the hardware acceleration plane 106. However, the host components may invoke the operations in the hardware acceleration plane 106 by issuing requests for services that are hosted by the hardware acceleration plane 106.

[0071] Second, note that the hardware acceleration plane 106 performs its transactions in a manner that is transparent to a requesting host component. For example, the local host component 204 may be “unaware” of how its request is being processed in the hardware acceleration plane, including the fact that the service corresponds to a multi-component service.

[0072] Third, note that, in this implementation, the communication in the software plane 104 (e.g., corresponding to operation (1)) takes place using the same common network 120 as communication in the hardware acceleration plane 106 (e.g., corresponding to operations (3)-(6)). Operations (2) and (7) may take place over a local link, corresponding to the local_H-to-local_S coupling 114 shown in FIG. 1.

[0073] The multi-component service shown in FIG. 2 resembles a ring in that a series of acceleration components are traversed in a first direction to arrive at a final result; that final result is then propagated back through the same series of acceleration components in the opposite direction to the head component. But as noted above, other multi-component services may use different collections of acceleration components having different respective flow structures.

[0074] For example, FIG. 3 shows a second example of the operation of the data processing system 102 of FIG. 1

that employs a different flow structure compared to the example of FIG. 1. More specifically, in operation (1), a local host component (not shown) sends a request to its local acceleration component 302. In this case, assume that the local acceleration component is also the head component of the service. In operation (2), the head component may then forward plural messages to plural respective acceleration components. Each acceleration component that receives the message may perform a part of the multi-component service in parallel with the other acceleration components. (Note that FIG. 3 may represent only a portion of a more complete transaction.)

[0075] Moreover, a multi-component service does not necessarily need to employ a single head component, or any head component. For example, a multi-component service can employ a cluster of acceleration components which all perform the same function. The data processing system 102 can be configured to invoke this kind of multi-component service by contacting any arbitrary member in the cluster. That acceleration component may be referred to as a head component because it is the first component to be accessed, but it otherwise has no special status. In yet other cases, a host component may initially distribute plural requests to plural members of a collection of acceleration components.

[0076] FIG. 4 shows a portion of a data center 402 which represents one implementation of the data processing system 102 of FIG. 1. In particular, FIG. 4 shows one rack in that data center 402. The rack includes plural server unit components (404, 406, . . . , 408), each of which is coupled to a top-of-rack (TOR) switch 410. A top-of-rack switch refers to a switch which couples the components in a rack to other parts of a data center. Other racks, although not shown, may exhibit a similar architecture. A rack is a physical structure for housing or otherwise grouping plural processing components.

[0077] FIG. 4 also shows the illustrative composition of one representative server unit component 404. It includes a local host component 412 that includes one or more central processing units (CPUs) (414, 416, . . .), together with a local acceleration component 418. The local acceleration component 418 is directly coupled to the host component 412 via a local link 420. The local link 420, for example, may be implemented as a PCIe link. The local acceleration component 418 is also indirectly coupled to the host component 412 by way of a network interface controller (NIC) 422.

[0078] Finally, note that the local acceleration component 418 is coupled to the TOR switch 410. Hence, in this particular implementation, the local acceleration component 418 represents the sole path through which the host component 412 interacts with other components in the data center 402 (including other host components and other acceleration components). Among other effects, the architecture of FIG. 4 allows the local acceleration component 418 to perform processing on packets that are received from (and/or sent to) the TOR switch 410 (e.g., by performing encryption, compression, etc.), without burdening the CPU-based operations performed by the host component 412.

[0079] Note that the local host component 412 may communicate with the local acceleration component 418 through the local link 420 or via the NIC 422. Different entities may leverage these two paths in different respective circumstances. For example, assume that a program running on the host component 412 requests a service. In one implemen-

tation, assume that the host component 412 provides a local instantiation of the location determination component 124 and the data store 126. Or a global management component may provide the location determination component 124 and its data store 126. In either case, the host component 412 may consult the data store 126 to determine the address of the service. The host component 412 may then access the service via the NIC 422 and the TOR switch 410, using the identified address.

[0080] In another implementation, assume that local acceleration component 418 provides a local instantiation of the location determination component 124 and the data store 126. The host component 412 may access the local acceleration component 418 via the local link 420. The local acceleration component 418 can then consult the local data store 126 to determine the address of the service, upon which it accesses the service via the TOR switch 410. Still other ways of accessing the service are possible.

[0081] FIG. 5 is a more encompassing depiction of the data center 402 shown in FIG. 4. The data center 402 includes a plurality of racks (502-512, . . .). Each rack includes a plurality of server unit components. Each server unit component, in turn, may have the architecture described above in FIG. 4. For example, a representative server unit component 514 includes a local host component (S) 516, a network interface controller (N) 518, and a local acceleration component (S) 520.

[0082] The routing infrastructure shown in FIG. 5 corresponds to one implementation of the common network 120, described above with reference to FIG. 1. The routing infrastructure includes a plurality of top-of-rack (TOR) switches 522 and higher-level switching infrastructure 524. The higher-level switching architecture 524 connects the TOR switches 522 together. The higher-level switching infrastructure 524 can have any architecture, and may be driven by any routing protocol(s). In the illustrated example of FIG. 5, the higher-level switching infrastructure 524 includes at least a collection of aggregation switches 526, core switches 528, etc. The traffic routed through the illustrated infrastructure may correspond to Ethernet IP packets.

[0083] The data center 402 shown in FIG. 5 may correspond to a set of resources provided at a single geographic location, or a distributed collection of resources that are distributed over plural geographic locations (e.g., over plural individual contributing data centers located in different parts of the world). In a distributed context, the management functionality 122 can send work from a first contributing data center to a second contributing data center based on any mapping consideration(s), such as: (1) a determination that acceleration components are available at the second contributing data center; (2) a determination that acceleration components are configured to perform a desired service or services at the second contributing data center; and/or (3) a determination that the acceleration components are not only configured to performed a desired service or services, but they are immediately available (e.g., “online”) to perform those services, and so on. As used herein, the term “global” generally refers to any scope that is more encompassing than the local domain associated with an individual server unit component.

[0084] Generally note that, while FIGS. 4 and 5 focus on the use of a relatively expansive data processing system (corresponding to a data center), some of the principles set forth herein can be applied to smaller systems, including a

case in which a single local host component (or other type of component) is coupled to plural acceleration components, including a local acceleration component and one or more remote acceleration components. Such a smaller system may even be embodied in a user device or appliance, etc. The user device may have the option of using local acceleration resources and/or remote acceleration resources.

[0085] FIG. 6 shows an alternative way of implementing a server unit component 602, compared to the architecture that shown in FIG. 4. Like the case of FIG. 4, the server unit component 602 of FIG. 6 includes a local host component 604 made up of one or more CPUs (606, 608, . . .), a local acceleration component 610, and a local link 612 for coupling the local host component 604 with the local acceleration component 610. Unlike the case of FIG. 4, the server unit component 602 implements a network interface controller (NIC) 614 as an internal component of the local acceleration component 610, rather than as a separate component.

[0086] FIG. 7 shows yet another alternative way of implementing a server unit component 702 compared to the architecture shown in FIG. 4. In the case of FIG. 7, the server unit component 702 includes any number n of local host components (704, . . . , 706) together with any number m of local acceleration components (708, . . . , 710). (Other components of the server unit component 702 are omitted from the figure to facilitate explanation.) For example, the server unit component 702 may include a single host component coupled to two local acceleration components. The two acceleration components can perform different respective tasks. For example, one acceleration component can be used to process outgoing traffic to its local TOR switch, while the other acceleration component can be used to process incoming traffic from the TOR switch. In addition, the server unit component 702 can load any services on any of the local acceleration components (708, . . . , 710).

[0087] Also note that, in the examples set forth above, a server unit component may refer to a physical grouping of components, e.g., by forming a single serviceable unit within a rack of a data center. In other cases, a server unit component may include one or more host components and one or more acceleration components that are not necessarily housed together in a single physical unit. In that case, a local acceleration component may be considered logically, rather than physically, associated with its respective local host component.

[0088] Alternatively, or in addition, a local host component and one or more remote acceleration components can be implemented on a single physical component, such as a single MPSoC-FPGA die. The network switch may also be incorporated into that single component.

[0089] FIG. 8 shows an alternative data processing system 802 compared to that shown in FIG. 1. Like the data processing system 102 of FIG. 1, the data processing system 802 includes a software plane 104 and a hardware acceleration plane 106, and a local_H-to-local_S coupling 114 for connecting local host components to respective local acceleration components. But unlike the data processing system 102 of FIG. 1, the data processing system 802 includes a first network 804 for coupling host components together, and a second network 806 for coupling hardware components together, wherein the first network 804 differs from the second network 806, at least in part. For example, the first network 804 may correspond to the type of data center

switching infrastructure shown in FIG. 5. The second network 806 may correspond to dedicated links for connecting the acceleration components together having any network topology. For example, the second network 806 may correspond to a p_{xr} torus network. Each acceleration component in the torus network is coupled to east, west, north, and south neighboring acceleration components via appropriate cable links or the like. Other types of torus networks can alternatively be used having any respective sizes and dimensions.

[0090] In other cases, local hard CPUs, and/or soft CPUs, and/or acceleration logic provided by a single processing component (e.g., as implemented on a single die) may be coupled via diverse networks to other elements on other processing components (e.g., as implemented on other dies, boards, racks, etc.). An individual service may itself utilize one or more recursively local interconnection networks.

[0091] Further note that the above description was framed in the context of host components which issue service requests that are satisfied by acceleration components. But alternatively, or in addition, any acceleration component can also make a request for a service which can be satisfied by any other component, e.g., another acceleration component and/or even a host component. The SMC 102 can address such a request in a similar manner to that described above. Indeed, certain features described herein can be implemented on a hardware acceleration plane by itself, without a software plane.

[0092] More generally stated, certain features can be implemented by any first component which requests a service, which may be satisfied by the first component, and/or by one or more local components relative to the first component, and/or by one or more remote components relative to the first component. To facilitate explanation, however, the description below will continue to be framed mainly in the context in which the entity making the request corresponds to a local host component.

[0093] Finally, other implementations can adopt different strategies for coupling the host components to the hardware components, e.g., other than the local_H-to-local_S coupling 114 shown in FIG. 14.

[0094] FIG. 9 shows a process 902 which represents one illustrative manner of operation of the data processing system 102 of FIG. 1. In block 904, a local host component issues a request for a service. In block 906, the local host component receives a reply to the request which may identify an address of the service. In an alternative implementation, an associated local acceleration component may perform blocks 904 and 906 after receiving a request from the local host component. In other words, either the local host component or the local acceleration component can perform the address lookup function.

[0095] In block 908, the associated local acceleration component may locally perform the service, assuming that the address that has been identified pertains to functionality that is locally implemented by the local acceleration component. Alternatively, or in addition, in block 910, the local acceleration component routes the request to a remote acceleration component. As noted above, the local acceleration component is configured to perform routing to the remote acceleration component without involvement of the local host component. Further, plural host components communicate in the data processing system 102 with each other over a same physical network as do plural acceleration components.

[0096] In conclusion to Section A, the data processing system 102 has a number of useful characteristics. First, the data processing system 102 uses a common network 120 (except for the example of FIG. 8) that avoids the expense associated with a custom network for coupling acceleration components together. Second, the common network 120 makes it feasible to add an acceleration plane to an existing data processing environment, such as a data center. And after installment, the resultant data processing system 102 can be efficiently maintained because it leverages existing physical links found in the existing data processing environment. Third, the data processing system 102 integrates the acceleration plane 106 without imposing large additional power requirements, e.g., in view of the above-described manner in which local acceleration components may be integrated with existing server unit components. Fourth, the data processing system 102 provides an efficient and flexible mechanism for allowing host components to access any acceleration resources provided by the hardware acceleration plane 106, e.g., without narrowly pairing host components to specific fixed acceleration resources, and without burdening the host components with managing the hardware acceleration plane 106 itself. Fifth, the data processing system 102 provides an efficient mechanism for managing acceleration resources by intelligently dispersing these resources within the hardware plane 106, thereby: (a) reducing the overutilization and underutilization of resources (e.g., corresponding to the “stranded capacity” problem); (b) facilitating quick access to these services by consumers of these services; (c) accommodating heightened processing requirements specified by some consumers and/or services, and so on. The above effects are illustrative, rather than exhaustive; the data processing system 102 offers yet other useful effects.

[0097] B. Management Functionality

[0098] FIG. 10 shows an overview of one implementation of the management functionality 122 that is used to manage the data processing system 102 of FIG. 1. More specifically, FIG. 10 depicts a logical view of the functions performed by the management functionality 122, including its principal engine, the service mapping component (SMC) 128. Different sub-components correspond to different main functions performed by the management functionality 122. FIGS. 17-20, described below, show various possible physical implementations of the logical functionality.

[0099] As described in the introductory Section A, the location determination component 124 identifies the current location of services within the data processing system 102, based on current allocation information stored in the data store 126. In operation, the location determination component 124 receives a request for a service. In response, it returns an address of the service, if present within the data store 126. The address may identify a particular acceleration component that implements the service.

[0100] The data store 126 may maintain any type of information which maps services to addresses. In the small excerpt shown in FIG. 10, the data store 126 maps a small number of services (service w, service x, service y, and service z) to the acceleration components which are currently configured to provide these services. For example, the data store 126 indicates that a configuration image for service w is currently installed on devices having addresses a1, a6, and a8. The address information may be expressed in any manner. Here, the address information is represented in high-level symbolic form to facilitate explanation.

[0101] In some implementations, the data store 126 may optionally also store status information which characterizes each current service-to-component allocation in any manner. Generally, the status information for a service-to-component allocation specifies the way that the allocated service, as implemented on its assigned component (or components), is to be treated within the data processing system 102, such as by specifying its level of persistence, specifying its access rights (e.g., “ownership rights”), etc. In one non-limiting implementation, for instance, a service-to-component allocation can be designated as either reserved or non-reserved. When performing a configuration operation, the SMC 128 can take into account the reserved/non-reserved status information associated with an allocation in determining whether it is appropriate to change that allocation, e.g., to satisfy a current request for a service, a change in demand for one or more services, etc. For example, the data store 126 indicates that the acceleration components having address a1, a6, and a8 are currently configured to perform service w, but that only the assignments to acceleration components a1 and a8 are considered reserved. Thus, the SMC 128 will view the allocation to acceleration component a6 as a more appropriate candidate for reassignment (reconfiguration), compared to the other two acceleration components.

[0102] In addition, or alternatively, the data store 126 can provide information which indicates whether a service-to-component allocation is to be shared by all instances of tenant functionality, or dedicated to one or more particular instances of tenant functionality (or some other indicated consumer(s) of the service). In the former (fully shared) case, all instances of tenant functionality vie for the same resources provided by an acceleration component. In the latter (dedicated) case, only those clients that are associated with a service allocation are permitted to use the allocated acceleration component. FIG. 10 shows, in high-level fashion, that the services x and y that run on the acceleration component having address a3 are reserved for use by one or more specified instances of tenant functionality, whereas any instance of tenant functionality can use the other service-to-component allocations.

[0103] The SMC 128 may also interact with a data store 1002 that provides availability information. The availability information identifies a pool of acceleration components that have free capacity to implement one or more services. For example, in one manner of use, the SMC 128 may determine that it is appropriate to assign one or more acceleration components as providers of a function. To do so, the SMC 128 draws on the data store 1002 to find acceleration components that have free capacity to implement the function. The SMC 128 will then assign the function to one or more of these free acceleration components. Doing so will change the availability-related status of the chosen acceleration components.

[0104] The SMC 128 also manages and maintains the availability information in the data store 1002. In doing so, the SMC 128 can use different rules to determine whether an acceleration component is available or unavailable. In one approach, the SMC 128 may consider an acceleration component that is currently being used as unavailable, while an acceleration component that is not currently being used as available. In other cases, the acceleration component may have different configurable domains (e.g., tiles), some of which are being currently used and others which are not being currently used. Here, the SMC 128 can specify the

availability of an acceleration component by expressing the fraction of its processing resources that are currently not being used. For example, FIG. 10 indicates that an acceleration component having address a1 has 50% of its processing resources available for use. On the other hand, an acceleration component having address a2 is completely available, while an acceleration component having an address a3 is completely unavailable. Individual acceleration components can notify the SMC 128 of their relative levels of utilization in different ways, as will be described in greater detail below.

[0105] In other cases, the SMC 128 can take into consideration pending requests for an acceleration component in registering whether it is available or not available. For example, the SMC 128 may indicate that an acceleration component is not available because it is scheduled to deliver a service to one or more instances of tenant functionality, even though it may not be engaged in providing that service at the current time.

[0106] In other cases, the SMC 128 can also register the type of each acceleration component that is available. For example, the data processing system 102 may correspond to a heterogeneous environment that supports acceleration components having different physical characteristics. The availability information in this case can indicate not only the identities of processing resources that are available, but also the types of those resources.

[0107] In other cases, the SMC 128 can also take into consideration the status of a service-to-component allocation when registering an acceleration component as available or unavailable. For example, assume that a particular acceleration component is currently configured to perform a certain service, and furthermore, assume that the allocation has been designated as reserved rather than non-reserved. The SMC 128 may designate that acceleration component as unavailable (or some fraction thereof as being unavailable) in view of its reserved status alone, irrespective of whether the service is currently being actively used to perform a function at the present time. In practice, the reserved status of an acceleration component therefore serves as a lock which prevents the SMC 128 from reconfiguring the acceleration component, at least in certain circumstances.

[0108] Now referring to the core mapping operation of the SMC 128 itself, the SMC 128 allocates or maps services to acceleration components in response to triggering events. More specifically, the SMC 128 operates in different modes depending on the type of triggering event that has been received. In a request-driven mode, the SMC 128 handles requests for services by tenant functionality. Here, each triggering event corresponds to a request by an instance of tenant functionality that resides, at least in part, on a particular local host component. In response to each request by a local host component, the SMC 128 determines an appropriate component to implement the service. For example, the SMC 128 may choose from among: a local acceleration component (associated with the local host component that made the request), a remote acceleration component, or the local host component itself (whereupon the local host component will implement the service in software), or some combination thereof.

[0109] In a second background mode, the SMC 128 operates by globally allocating services to acceleration components within the data processing system 102 to meet overall anticipated demand in the data processing system 102 and/or

to satisfy other system-wide objectives and other factors (rather than narrowly focusing on individual requests by host components). Here, each triggering event that is received corresponds to some condition in the data processing system **102** as a whole that warrants allocation (or reallocation) of a service, such as a change in demand for the service.

[0110] Note, however, that the above-described modes are not mutually exclusive domains of analysis. For example, in the request-driven mode, the SMC **128** may attempt to achieve at least two objectives. As a first primary objective, the SMC **128** will attempt to find an acceleration component (or components) that will satisfy an outstanding request for a service, while also meeting one or more performance goals relevant to the data processing system **102** as a whole. As a second objective, the SMC **128** may optionally also consider the long term implications of its allocation of the service with respect to future uses of that service by other instances of tenant functionality. In other words, the second objective pertains to a background consideration that happens to be triggered by a request by a particular instance of tenant functionality.

[0111] For example, consider the following simplified case. An instance of tenant functionality may make a request for a service, where that instance of tenant functionality is associated with a local host component. The SMC **128** may respond to the request by configuring a local acceleration component to perform the service. In making this decision, the SMC **128** may first of all attempt to find an allocation which satisfies the request by the instance of tenant functionality. But the SMC **128** may also make its allocation based on a determination that many other host components have requested the same service, and that these host components are mostly located in the same rack as the instance of tenant functionality which has generated the current request for the service. In other words, this supplemental finding further supports the decision to place the service on an in-rack acceleration component.

[0112] FIG. **10** depicts the SMC **128** as optionally including plural logic components that perform different respective analyses. As a first optional component of analysis, the SMC **128** may use status determination logic **1004** to define the status of an allocation that it is making, e.g., as either reserved or non-reserved, dedicated or fully shared, etc. For example, assume that the SMC **128** receives a request from an instance of tenant functionality for a service. In response, the SMC **128** may decide to configure a local acceleration component to provide the service, and, in the process, designate this allocation as non-reserved, e.g., under the initial assumption that the request may be a “one-off” request for the service. In another situation, assume that the SMC **128** makes the additional determination that the same instance of tenant functionality has repeatedly made a request for the same service in a short period of time. In this situation, the SMC **128** may make the same allocation decision as described above, but this time the SMC **128** may designate it as being reserved. The SMC **128** may also optionally designate the service as being dedicated to just the requesting tenant functionality. By doing so, the SMC **128** may enable the data processing system **102** to more effectively satisfy future requests for this service by the instance of tenant functionality. In other words, the reserved status may reduce the chance that the SMC **128** will later

move the service from the local acceleration component, where it is being heavily used by the local host component.

[0113] In addition, an instance of tenant functionality (or a local host component) may specifically request that it be granted a reserved and dedicated use of a local acceleration component. The status determination logic **1004** can use different environment-specific rules in determining whether to honor this request. For instance, the status determination logic **1004** may decide to honor the request, providing that no other triggering event is received which warrants overriding the request. The status determination logic **1004** may override the request, for instance, when it seeks to fulfill another request that is determined, based on any environment-specific reasons, as having greater urgency than the tenant functionality’s request.

[0114] In some implementations, note that an instance of tenant functionality (or a local host component or some other consumer of a service) may independently control the use of its local resources. For example, a local host component may pass utilization information to the management functionality **122** which indicates that its local acceleration component is not available or not fully available, irrespective of whether the local acceleration component is actually busy at the moment. In doing so, the local host component may prevent the SMC **128** from “stealing” its local resources. Different implementations can use different environment-specific rules to determine whether an entity is permitted to restrict access to its local resources in the above-described manner, and if so, in what circumstances.

[0115] In another example, assume that the SMC **128** determines that there has been a general increase in demand for a particular service. In response, the SMC **128** may find a prescribed number of free acceleration components, corresponding to a “pool” of acceleration components, and then designate that pool of acceleration components as reserved (but fully shared) resources for use in providing the particular service. Later, the SMC **128** may detect a general decrease in demand for the particular service. In response, the SMC **128** can decrease the pool of reserved acceleration components, e.g., by changing the status of one or more acceleration components that were previously registered as “reserved” to “non-reserved.”

[0116] Note that the particular dimensions of status described above (reserved vs. non-reserved, dedicated vs. fully shared) are cited by way of illustration, not limitation. Other implementations can adopt any other status-related dimensions, or may accommodate only a single status designation (and therefore omit use of the status determination logic **1004** functionality).

[0117] As a second component of analysis, the SMC **128** may use size determination logic **1006** to determine a number of acceleration components that are appropriate to provide a service. The SMC **128** can make such a determination based on a consideration of the processing demands associated with the service, together with the resources that are available to meet those processing demands.

[0118] As a third component of analysis, the SMC **128** can use type determination logic **1008** to determine the type(s) of acceleration components that are appropriate to provide a service. For example, consider the case in which the data processing system **102** has a heterogeneous collection of acceleration components having different respective capabilities. The type determination logic **1008** can determine

one or more of a particular kind of acceleration components that are appropriate to provide the service.

[0119] As a fourth component of analysis, the SMC 128 can use placement determination logic 1010 to determine the specific acceleration component (or components) that are appropriate to address a particular triggering event. This determination, in turn, can have one more aspects. For instance, as part of its analysis, the placement determination logic 1010 can determine whether it is appropriate to configure an acceleration component to perform a service, where that component is not currently configured to perform the service.

[0120] The above facets of analysis are cited by way of illustration, not limitation. In other implementations, the SMC 128 can provide additional phases of analyses.

[0121] Generally, the SMC 128 performs its various allocation determinations based on one or more mapping considerations. For example, one mapping consideration may pertain to historical demand information provided in a data store 1012.

[0122] Note, however, that the SMC 128 need not perform multi-factor analysis in all cases. In some cases, for instance, a host component may make a request for a service that is associated with a single fixed location, e.g., corresponding to the local acceleration component or a remote acceleration component. In those cases, the SMC 128 may simply defer to the location determination component 124 to map the service request to the address of the service, rather than assessing the costs and benefits of executing the service in different ways. In other cases, the data store 126 may associate plural addresses with a single service, each address associated with an acceleration component that can perform the service. The SMC 128 can use any mapping consideration(s) in allocating a request for a service to a particular address, such as a load balancing consideration.

[0123] As a result of its operation, the SMC 128 can update the data store 126 with information that maps services to addresses at which those services can be found (assuming that this information has been changed by the SMC 128). The SMC 128 can also store status information that pertains to new service-to-component allocations.

[0124] To configure one or more acceleration components to perform a function (if not already so configured), the SMC 128 can invoke a configuration component 1014. In one implementation, the configuration component 1014 configures acceleration components by sending a configuration stream to the acceleration components. A configuration stream specifies the logic to be “programmed” into a recipient acceleration component. The configuration component 1014 may use different strategies to configure an acceleration component, several of which are set forth below.

[0125] A failure monitoring component 1016 determines whether an acceleration component has failed. The SMC 128 may respond to a failure notification by substituting a spare acceleration component for a failed acceleration component.

[0126] B.1. Operation of the SMC in a Request-Driven Mode

[0127] FIG. 11 provides an overview of one manner of operation of the SMC 128 when applied to the task of processing requests by instances of tenant functionality running on host components. In the illustrated scenario, assume that a host component 1102 implements plural instances of tenant functionality ($T_1, T_2, \dots, T_n$). Each

instance of tenant functionality may correspond to a software program that executes, at least in part, on the host component 1102, e.g., in a virtual machine that runs using the physical resources of the host component 1102 (among other possible host components). Further assume that one instance of tenant functionality initiates the transaction shown in FIG. 11 by generating a request for a particular service. For example, the tenant functionality may perform a photo editing function, and may call on a compression service as part of its overall operation. Or the tenant functionality may perform a search algorithm, and may call on a ranking service as part of its overall operation.

[0128] In operation (1), the local host component 1102 may send its request for the service to the SMC 128. In operation (2), among other analyses, the SMC 128 may determine at least one appropriate component to implement the service. In this case, assume that the SMC 128 determines that a remote acceleration component 1104 is the most appropriate component to implement the service. The SMC 128 can obtain the address of that acceleration component 1104 from the location determination component 124. In operation (3), the SMC 128 may communicate its answer to the local host component 1102, e.g., in the form of the address associated with the service. In operation (4), the local host component 1102 may invoke the remote acceleration component 1104 via its local acceleration component 1106. Other ways of handling a request by tenant functionality are possible. For example, the local acceleration component 1106 can query the SMC 128, rather than, or in addition to, the local host component 102.

[0129] Path 1108 represents an example in which a representative acceleration component 1110 (and/or its associated local host component) communicates utilization information to the SMC 128. The utilization information may identify whether the acceleration component 1110 is available or unavailable for use, in whole or in part. The utilization information may also optionally specify the type of processing resources that the acceleration component 1110 possesses which are available for use. As noted above, the utilization information can also be chosen to purposively prevent the SMC 128 from later utilizing the resources of the acceleration component 1110, e.g., by indicating in whole or in part that the resources are not available.

[0130] Although not shown, any acceleration component can also make directed requests for specific resources to the SMC 128. For example, the host component 1102 may specifically ask to use its local acceleration component 1106 as a reserved and dedicated resource. As noted above, the SMC 128 can use different environment-specific rules in determining whether to honor such a request.

[0131] Further, although not shown, other components besides the host components can make requests. For example, a hardware acceleration component may run an instance of tenant functionality that issues a request for a service that can be satisfied by itself, another hardware acceleration component (or components), a host component (or components), etc., or any combination thereof.

[0132] FIGS. 12-15 show different respective options for handling a request for a service made by tenant functionality that is resident on a host component. Starting with FIG. 12, assume that a local host component 1202 includes at least two instances of tenant functionality, T_1 (1204) and T_2 (1206), both of which are running at the same time (but, in actuality, the local host component 1202 can host many

more instances of tenant functionality). The first instance of tenant functionality T1 requires an acceleration service A1 to perform its operation, while the second instance of tenant functionality T2 requires an acceleration service A2 to perform its operation.

[0133] Further assume that a local acceleration component 1208 is coupled to the local host component 1202, e.g., via a PCIe local link or the like. At the current time, the local acceleration component 1208 hosts A1 logic 1210 for performing the acceleration service A1, and A2 logic 1212 for performing the acceleration service A2.

[0134] According to one management decision, the SMC 128 assigns T1 to the A1 logic 1210, and assigns T2 to the A2 logic 1212. However, this decision by the SMC 128 is not a fixed rule; the SMC 128 may make its decision based on plural factors, some of which may reflect conflicting considerations. As such, based on other factors (not described at this juncture), the SMC 128 may choose to assign jobs to acceleration logic in a different manner from that illustrated in FIG. 12.

[0135] In the scenario of FIG. 13, the host component 1302 has the same instances of tenant functionality (1304, 1306) with the same service needs described above. But in this case, a local acceleration component 1308 includes only A1 logic 1310 for performing service A1. That is, it no longer hosts A2 logic for performing the service A2.

[0136] In response to the above scenario, the SMC 128 may choose to assign T1 to the A1 logic 1310 of the acceleration component 1308. The SMC 128 may then assign T2 to the A2 logic 1312 of a remote acceleration component 1314, which is already configured to perform that service. Again, the illustrated assignment is set forth here in the spirit of illustration, not limitation; the SMC 128 may choose a different allocation based on another combination of input considerations. In one implementation, the local host component 1302 and the remote acceleration component 1314 can optionally compress the information that they send to each other, e.g., to reduce consumption of bandwidth.

[0137] Note that the host component 1302 accesses the A2 logic 1312 via the local acceleration component 1308. But in another case (not illustrated), the host component 1302 may access the A2 logic 1312 via the local host component (not illustrated) that is associated with the acceleration component 1314.

[0138] FIG. 14 presents another scenario in which the host component 1402 has the same instances of tenant functionality (1404, 1406) with the same service needs described above. In this case, a local acceleration component 1408 includes A1 logic 1410 for performing service A1, and A3 logic 1412 for performing service A3. Further assume that the availability information in the data store 1002 indicates that the A3 logic 1412 is not currently being used by any instance of tenant functionality. In response to the above scenario, the SMC 128 may use the configuration component 1014 (of FIG. 10) to reconfigure the acceleration component 1408 so that it includes A2 logic 1414, rather than A3 logic 1412 (as shown at the bottom of FIG. 14). The SMC 128 may then assign T2 to the A2 logic 1414. Although not shown, the SMC 128 can alternatively, or in addition, decide to reconfigure any remote acceleration component to perform the A2 service.

[0139] Generally, the SMC 128 can perform configuration in a full or partial manner to satisfy any request by an

instance of tenant functionality. The SMC performs full configuration by reconfiguring all of the application logic provided by an acceleration component. The SMC 128 can perform partial configuration by reconfiguring part (e.g., one or more tiles) of the application logic provided by an acceleration component, leaving other parts (e.g., one or more other tiles) intact and operational during reconfiguration. The same is true with respect to the operation of the SMC 128 in its background mode of operation, described below. Further note that additional factors may play a role in determining whether the A3 logic 1412 is a valid candidate for reconfiguration, such as whether or not the service is considered reserved, whether or not there are pending requests for this service, etc.

[0140] FIG. 15 presents another scenario in which the host component 1502 has the same instances of tenant functionality (1504, 1506) with the same service needs described above. In this case, a local acceleration component 1508 includes only A1 logic 1510 for performing service A1. In response to the above scenario, the SMC 128 may assign T1 to the A1 logic 1510. Further, assume that the SMC 128 determines that it is not feasible for any acceleration component to perform the A2 service. In response, the SMC 128 may instruct the local host component 1502 to assign T2 to local A2 software logic 1512, if, in fact, that logic is available at the host component 1502. The SMC 128 can make the decision of FIG. 15 on various grounds. For example, the SMC 128 may conclude that hardware acceleration is not possible because a configuration image does not currently exist for this service. Or the configuration image may exist, but the SMC 128 concludes that there is insufficient capacity on any of the acceleration devices to load and/or run such a configuration.

[0141] Finally, the above examples were described in the context of instances of tenant functionality that run on host components. But as already noted above, the instances of tenant functionality may more generally correspond to service requestors, and those service requestors can run on any component(s), including acceleration components. Thus, for example, a requestor that runs on an acceleration component can generate a request for a service to be executed by one or more other acceleration components and/or by itself and/or by one or more host components. The SMC 102 can handle the requestor's request in any of the ways described above.

[0142] B.2. Operation of the SMC in a Background Mode

[0143] FIG. 16 provides an overview of one manner of operation of the SMC 128 when operating in a background mode. In operation (1), the SMC 128 may receive some type of triggering event which initiates the operation of the SMC 128. For example, the triggering event may correspond to a change in demand which affects a service, etc. In operation (2), in response to the triggering event, the SMC 128 determines an allocation of one or more services to acceleration components based on one or more mapping considerations and the availability information in the data store 1002, e.g., by assigning the services to a set of one or more available acceleration components. In operation (3), the SMC 128 carries out its allocation decisions. As part of this process, the SMC 128 may call on the configuration component 1014 to configure the acceleration components that have been allocated to perform the service(s), assuming that these components are not already configured to perform the service(s). The SMC 128 also updates the service location

information in the data store **126**, and, if appropriate, the availability information in the data store **1002**.

[0144] In the particular example of FIG. 16, the SMC **102** allocates a first group **1602** of acceleration components to perform a first service (“service y”), and allocates a second group **1604** of acceleration components to perform a second service (“service z”). In actual practice, an allocated group of acceleration components can have any number of members, and these members may be distributed in any fashion across the hardware acceleration plane **106**. The SMC **128**, however, may attempt to group the acceleration components associated with a service in a particular manner to achieve satisfactory bandwidth and latency performance (among other factors). The SMC **128** may apply further analysis in allocating acceleration components associated with a single multi-component service, as also described in greater detail below.

[0145] The SMC **128** can also operate in the background mode to allocate one or more acceleration components, which implement a particular service, to at least one instance of tenant functionality, without necessarily requiring the tenant functionality to make a request for this particular service each time. For example, assume that an instance of tenant functionality regularly uses a compression function, corresponding to “service z” in FIG. 16. The SMC **128** can proactively allocate one or more dedicated acceleration components **1604** to at least this instance of tenant functionality. When the tenant functionality requires use of the service, it may draw from the pool of available addresses associated with the acceleration components **1604** that have been assigned to it. The same dedicated mapping operation can be performed with respect to a group of instances of tenant functionality (instead of a single instance).

[0146] B.3. Physical Implementations of the Management Functionality

[0147] FIG. 17 shows a first physical implementation of the management functionality **122** of FIG. 10. In this case, the management functionality **122** is provided on a single global management component (M_G) **1702**, or on plural global management components (**1702**, . . . , **1704**). The plural global management components (**1702**, . . . , **1704**), if used, may provide redundant logic and information to achieve desired load balancing and failure management performance. In one case, each global management component may be implemented on a computer server device, which may correspond to one of the host components, or a dedicated management computing device. In operation, any individual host component (S) or acceleration component (H) may interact with a global management component via the common network **120** shown in FIG. 1.

[0148] FIG. 18 shows a second physical implementation of the management functionality **122** of FIG. 10. In this case, each server unit component (such as representative server unit component **1802**) provides at least one local management component (M_L) **1804**. For example, a local host component **1806** may implement the local management component **1804** (e.g., as part of its hypervisor functionality), or a local acceleration component **1808** may implement the local management component **1804**, or some other component within the server unit component **1802** may implement the local management component **1804** (or some combination thereof). The data processing system **102** also includes one or more global management components (**1810**, . . . , **1812**). Each global management component may

provide redundant logic and information in the manner described above with respect to FIG. 17. The management functionality **122** collectively presents all of the local and global management components in the data processing system **102**, as set forth above.

[0149] The architecture of FIG. 18 can implement the request-driven aspects of the SMC **128**, for instance, in the following manner. The local management component **1804** may first determine whether the local acceleration component **1808** can perform a service requested by tenant functionality. A global management component (M_G) can perform other decisions, such as identifying a remote acceleration component to perform a service, in the event that the local acceleration component **1808** cannot perform this task. On the other hand, in the architecture of FIG. 17, a single global management component can perform all decisions pertaining to the mapping of a request to an acceleration component.

[0150] Further, the local management component **1804** can send utilization information to a global management component on any basis, such as periodic basis and/or an event-driven basis (e.g., in response to a change in utilization). The global management component can use the utilization information to update its master record of availability information in the data store **1002**.

[0151] FIG. 19 shows a third physical implementation of the management functionality **122** of FIG. 10. In this case, each server unit component stores its own dedicated local management component (M_L) (which can be implemented by a local host component as part of its hypervisor functionality, a local acceleration component, some other local component, or some combination thereof). For instance, a server unit component **1902** provides a local management component **1904**, along with a local host component **1906** and a local acceleration component **1908**. Likewise, a server unit component **1910** provides a local management component **1912**, along with a local host component **1914** and a local acceleration component **1916**. Each instance of a local management component stores redundant logic and information with respect to other instances of the same component. Known distributed system tools can be used to ensure that all distributed versions of this component contain the same logic and information, such as the ZOOKEEPER tool provided by Apache Software Foundation of Forest Hill, Md. (As an aside, note that the same technology can be used to maintain the redundant logic and information in the other examples described in this subsection.) The management functionality **122** collectively presents all of the local management components in the data processing system **102**, as set forth above. That is, there is no central global management component(s) in this implementation.

[0152] FIG. 20 shows a fourth physical implementation of the management functionality **122** of FIG. 10. In this case, the management functionality **122** embodies a hierarchical structure of individual management components. For example, in one merely representative structure, each server unit component includes a low-level local management component (M_{L3}) (which can be implemented by a local host component, a local acceleration component, some other local component, or some combination thereof). For example, a server unit component **2002** provides a low-level local management component **2004**, along with a local host component **2006** and a local acceleration component **2008**. Likewise, a server unit component **2010** provides a low-

level local management component **2012**, along with a local host component **2014** and an acceleration component **2016**. A next management tier of the structure includes at least a mid-level management component **2018** and a mid-level management component **2020**. A top level of the structure includes a single global management component **2022** (or plural redundant such global management components). The illustrated control architecture thus forms a structure having three levels, but the architecture can have any number of levels.

[0153] In operation, the low-level management components (**2004**, **2012**, . . .) handle certain low-level management decisions that directly affect the resources associated with individual server unit components. The mid-level management components (**2018**, **2020**) can make decisions which affect a relevant section of the data processing system **102**, such as an individual rack or a group of racks. The top-level management component (**2022**) can make global decisions which broadly apply to the entire data processing system **102**.

[0154] B.4. The Configuration Component

[0155] FIGS. **21-24** show different respective strategies for configuring an acceleration component, corresponding to different ways of implementing the configuration component **1014** of FIG. **10**. Starting with FIG. **21**, a global management component **2102** has access to a data store **2104** that provides one or more configuration images. Each configuration image contains logic that can be used to implement a corresponding service. The global management component **2102** can configure an acceleration component by forwarding a configuration stream (corresponding to a configuration image) to the acceleration component. For example, in one approach, the global management component **2102** can send the configuration stream to a local management component **2106** associated with a particular server unit component **2108**. The local management component **2106** can then coordinate the configuration of a local acceleration component **2110** based on the received configuration stream. Alternatively, the local host component **2112** can perform the above-described operation, instead of, or in addition to, the local management component **2106**.

[0156] FIG. **22** shows another strategy for configuring an acceleration component. In this case, a global management component **2202** sends an instruction to a local management component **2204** of a server unit component **2206**. In response, the local management component **2204** accesses a configuration image in a local data store **2208** and then uses it to configure a local acceleration component **2210**. Alternatively, a local host component **2212** can perform the above-described operation, instead of, or in addition to, the local management component **2204**.

[0157] FIG. **23** shows another technique for configuring a local acceleration component **2302**. In this approach, assume that the acceleration component **2302** includes application logic **2304**, which, in turn, is governed by a current model **2306** (where a model corresponds to logic that performs a function in a particular manner). Further assume that the acceleration component **2302** has access to local memory **2308**. The local memory **2308** stores configuration images associated with one or more other models (model **1**, . . . , model **n**). When triggered, a local model loading component **2310** can swap out the configuration associated with the current model **2306** with the configuration associated with another model in the local memory **2308**. The model loading

component **2310** may be implemented by the acceleration component **2302** itself, a local host component, a local management component, etc., or some combination thereof. In one implementation, the configuration operation shown in FIG. **23** can be performed in less time than the overall reconfiguration of the application logic **2304** as a whole, as it entails replacing some of the logic used by the application logic **2304**, not the entire application logic **2304** in whole-sale fashion.

[0158] Finally, FIG. **24** shows an acceleration component having application logic **2402** that supports partial configuration. The management functionality **122** can leverage this capability by configuring application **1** (**2404**) separately from application **2** (**2406**), and vice versa.

[0159] C. Illustrative Implementation of a Hardware Acceleration Component

[0160] FIG. **25** shows one manner of implementing an acceleration component **2502** in the data processing system of FIG. **1**, which may be physically implemented as an FPGA device. Note that the detail presented below is set forth in the spirit of illustration, not limitation; other data processing systems may use acceleration components having architectures which vary in one or more ways compared to that shown in FIG. **25**. Further, other data processing systems may employ a heterogeneous design that includes acceleration components having different types.

[0161] From a high-level standpoint, the acceleration component **2502** may be implemented as a hierarchy having different layers of functionality. At a lowest level, the acceleration component **2502** provides an “outer shell” which provides basic interface-related components that generally remain the same across most application scenarios. A core component **2504**, which lies inside the outer shell, may include an “inner shell” and application logic **2506**. The inner shell corresponds to all the resources in the core component **2504** other than the application logic **2506**, and represents a second level of resources that remain the same within a certain set of application scenarios. The application logic **2506** itself represents a highest level of resources which are most readily subject to change. Note however that any component of the acceleration component **2502** can technically be reconfigured.

[0162] In operation, the application logic **2506** interacts with the outer shell resources and inner shell resources in a manner analogous to the way a software-implemented application interacts with its underlying operating system resources. From an application development standpoint, the use of common outer shell resources and inner shell resources frees a developer from having to recreate these common components for each application that he or she creates. This strategy also reduces the risk that a developer may alter core inner or outer shell functions in a manner that causes problems within the data processing system **102** as a whole.

[0163] Referring first to the outer shell, the acceleration component **2502** includes a bridge **2508** for coupling the acceleration component **2502** to the network interface controller (via a NIC interface **2510**) and a local top-of-rack switch (via a TOR interface **2512**). The bridge **2508** supports two modes. In a first mode, the bridge **2508** provides a data path that allows traffic from the NIC or TOR to flow into the acceleration component **2502**, and traffic from the acceleration component **2502** to flow out to the NIC or TOR. The acceleration component **2502** can perform any processing on

the traffic that it “intercepts,” such as compression, encryption, etc. In a second mode, the bridge 2508 supports a data path that allows traffic to flow between the NIC and the TOR without being further processed by the acceleration component 2502. Internally, the bridge may be composed of various FIFOs (2514, 2516) which buffer received packets, and various selectors and arbitration logic which route packets to their desired destinations. A bypass control component 2518 controls whether the bridge 2508 operates in the first mode or the second mode.

[0164] A memory controller 2520 governs interaction between the acceleration component 2502 and local memory 2522 (such as DRAM memory). The memory controller 2520 may perform error correction as part of its services.

[0165] A host interface 2524 provides functionality that enables the acceleration component to interact with a local host component (not shown in FIG. 25). In one implementation, the host interface 2524 may use Peripheral Component Interconnect Express (PCIe), in conjunction with direct memory access (DMA), to exchange information with the local host component.

[0166] Finally, the shell may include various other features 2526, such as clock signal generators, status LEDs, error correction functionality, and so on.

[0167] In one implementation, the inner shell may include a router 2528 for routing messages between various internal components of the acceleration component 2502, and between the acceleration component 2502 and external entities (via a transport component 2530). Each such endpoint is associated with a respective port. For example, the router 2528 is coupled to the memory controller 2520, host interface 1120, application logic 2506, and transport component 2530.

[0168] The transport component 2530 formulates packets for transmission to remote entities (such as remote acceleration components), and receives packets from the remote acceleration components (such as remote acceleration components).

[0169] A 3-port switch 2532, when activated, takes over the function of the bridge 2508 by routing packets between the NIC and TOR, and between the NIC or TOR and a local port associated with the acceleration component 2502 itself.

[0170] Finally, an optional diagnostic recorder 2534 stores transaction information regarding operations performed by the router 2528, transport component 2530, and S-port switch 2532 in a circular buffer. For example, the transaction information may include data about a packet's origin and destination IP addresses, host-specific data, timestamps, etc. A technician may study a log of the transaction information in an attempt to diagnose causes of failure or sub-optimal performance in the acceleration component 2502.

[0171] FIG. 26 shows an acceleration component 3202 that includes separate configurable domains (2604, 2606, . . .). A configuration component (e.g., configuration component 1014 of FIG. 10) can configure each configurable domain without affecting other configurable domains. Hence, the configuration component 1014 can configure one or more configurable domains while the other configurable domains are executing operations based on their respective configurations, which are not disturbed.

[0172] In some implementations, the data processing system 102 of FIG. 1 may dynamically reconfigure its acceleration components to address any mapping considerations. That reconfiguration can be performed on a partial and/or

whole-service basis, and may be performed on a periodic and/or event-driven basis. Indeed, in some cases, the data processing system 102 may appear to be continually in the process of adapting itself to changing conditions in the data processing system 102 by reconfiguring its acceleration logic.

[0173] C.1. The Local Link

[0174] FIG. 27 shows functionality by which a local host component 2702 may forward information to its local acceleration component 2704 via the host interface 2524 shown in FIG. 25 (e.g., using PCIe in conjunction with DMA memory transfer). In one non-limiting protocol, in operation (1), the host logic 2706 places data to be processed into a kernel-pinned input buffer 2708 in main memory associated with the host logic 2706. In operation (2), the host logic 2706 instructs the acceleration component 2704 to retrieve the data and begin processing it. The host logic's thread is then either put to sleep until it receives a notification event from the acceleration component 2704, or it continues processing other data asynchronously. In operation (3), the acceleration component 2704 transfers the data from the host logic's memory and places it in an acceleration component input buffer 2710.

[0175] In operations (4) and (5), the application logic 2712 retrieves the data from the input buffer 2710, processes it to generate an output result, and places the output result in an output buffer 2714. In operation (6), the acceleration component 2704 copies the contents of the output buffer 2714 into an output buffer in the host logic's memory. In operation (7), the acceleration component notifies the host logic 2706 that the data is ready for it to retrieve. In operation (8), the host logic thread wakes up and consumes the data in the output buffer 2716. The host logic 2706 may then discard the contents of the output buffer 2716, which allows the acceleration component 2704 to reuse it in the next transaction.

[0176] C.2. The Router

[0177] FIG. 28 shows one implementation of the router 2528 introduced in FIG. 25. The router includes any number of input units (here four, 2802, 2804, 2806, 2808) for receiving messages from respective ports, and output units (here four, 2810, 2812, 2814, 2816) for forwarding messages to respective ports. As described above, the endpoints associated with the ports include the memory controller 2520, the host interface 2524, the application logic 2506, and the transport component 2530. A crossbar component 2818 forwards a message from an input port to an output port based on address information associated with the message. More specifically, a message is composed of multiple “flits,” and the router 2528 sends messages on a flit-by-flit basis.

[0178] In one non-limiting implementation, the router 2528 supports a number of virtual channels (such as eight) for transmitting different classes of traffic over a same physical link. That is, the router 2528 may support multiple traffic classes for those scenarios in which multiple services are implemented by the application logic 2506, and those services need to communicate on separate classes of traffic.

[0179] The router 2528 may govern access to the router's resources (e.g., its available buffer space) using a credit-based flow technique. In that technique, the input units (2802-2808) provide upstream entities with credits, which correspond to the exact number of flits available in their buffers. The credits grant the upstream entities the right to transmit their data to the input units (2802-2808). More specifically, in one implementation, the router 2528 supports

“elastic” input buffers that can be shared among multiple virtual channels. The output units (2810-2816) are responsible for tracking available credits in their downstream receivers, and provide grants to any input units (2802-2808) that are requesting to send a flit to a given output port.

[0180] C.3. The Transport Component

[0181] FIG. 29 shows one implementation of the transport component 2530 introduced in FIG. 25. The transport component 2530 may provide a register interface to establish connections between nodes. That is, each such connection is one-way and links a send queue on a source component to a receive queue on a destination component. A software process may set up the connections by statically allocating them before the transport component 2530 can transmit or receive data. A data store 2902 stores two tables that control the state of connections, a Send Connection Table and a Receive Connection Table.

[0182] A packet processing component 2904 processes messages arriving from the router 2528 which are destined for a remote endpoint (e.g., another acceleration component). It does so by buffering and packetizing the messages. The packet processing component 2904 also processes packets that are received from some remote endpoint and are destined for the router 2528.

[0183] For messages arriving from the router 2528, the packet processing component 2904 matches each message request to a Send Connection Table entry in the Send Connection Table, e.g., using header information and virtual channel (VC) information associated with the message as a lookup item, as provided by router 2528. The packet processing component 2904 uses the information retrieved from the Send Connection Table entry (such as a sequence number, address information, etc.) to construct packets that it sends out to the remote entity.

[0184] More specifically, in one non-limiting approach, the packet processing component 2904 encapsulates packets in UDP/IP Ethernet frames, and sends them to a remote acceleration component. In one implementation the packets may include an Ethernet header, followed by an IPv4 header, followed by a UDP header, followed by transport header (specifically associated with the transport component 2530), followed by a payload.

[0185] For packets arriving from the network (e.g., as received on a local port of the 3-port switch 2532), the packet processing component 2904 matches each packet to a Receive Connectable Table entry provided in the packet header. If there is a match, the packet processing component retrieves a virtual channel field of the entry, and uses that information to forward the received message to the router 2528 (in accordance with the credit-flow technique used by the router 2528).

[0186] A failure handling component 2906 buffers all sent packets until it receives an acknowledgement (ACK) from the receiving node (e.g., the remote acceleration component). If an ACK for a connection does not arrive within a specified time-out period, the failure handling component 2906 can retransmit the packet. The failure handling component 2906 will repeat such retransmission for a prescribed number times (e.g., 128 times). If the packet remains unacknowledged after all such attempts, the failure handling component 2906 can discard it and free its buffer.

[0187] C.4. The 3-Port Switch

[0188] FIG. 30 shows one implementation of the 3-port switch 2532. The 3-port switch 2532 operates to safely insert

(and remove) acceleration component-generated network packets onto a data center network without compromising host-to-TOR network traffic.

[0189] The 3-port switch 2532 connects to the NIC interface 2510 (corresponding to a host interface), the TOR interface 2512, and a local interface associated with the local acceleration component 2502 itself. The 3-port switch 2532 may be conceptualized as including receiving interfaces (3002, 3004, 3006) for respectively receiving packets from the host component and TOR switch, and for receiving packets at the local acceleration component. The 3-port switch 2532 also includes transmitting interfaces (3008, 3010, 3012) for respectively providing packets to the TOR switch and host component, and receiving packets transmitted by the local acceleration component.

[0190] Packet classifiers (3014, 3016) determine the class of packets received from the host component or the TOR switch, e.g., based on status information specified by the packets. In one implementation, each packet is either classified as belonging to a lossless flow (e.g., remote direct memory access (RDMA) traffic) or a lossy flow (e.g., transmission control protocol/Internet Protocol (TCP/IP) traffic). Traffic that belongs to a lossless flow is intolerant to packet loss, while traffic that belongs to a lossy flow can tolerate some packet loss.

[0191] Packet buffers (3018, 3020) store the incoming packets in different respective buffers, depending on the class of traffic to which they pertain. If there is no space available in the buffer, the packet will be dropped. (In one implementation, the 3-port switch 2532 does not provide packet buffering for packets provided by the local acceleration component (via the local port) because the application logic 2506 can regulate the flow of packets through the use of “back pressuring.”) Arbitration logic 3022 selects among the available packets and transmits the selected packets.

[0192] As described above, traffic that is destined for the local acceleration component is encapsulated in UDP/IP packets on a fixed port number. The 3-port switch 2532 inspects incoming packets (e.g., as received from the TOR) to determine if they are UDP packets on the correct port number. If so, the 3-port switch 2532 outputs the packet on the local RX port interface 3006. In one implementation, all traffic arriving on the local TX port interface 3012 is sent out of the TOR TX port interface 3008, but it could also be sent to the host TX port interface 3010. Further note that FIG. 30 indicates that the acceleration component 2502 intercepts traffic from the TOR, but not from the host component; but it could be configured to intercept traffic from the host component as well.

[0193] PFC processing logic 3024 allows the 3-port switch 2532 to insert Priority Flow Control frames into either the flow of traffic transmitted to the TOR or host component. That is, for lossless traffic classes, if a packet buffer fills up, the PFC processing logic 3024 sends a PFC message to the link partner, requesting that traffic on that class be paused. If a PFC control frame is received for a lossless traffic class on either the host RX port interface 3002 or the TOR RX port interface 3004, the 3-port switch 2532 will cease sending packets on the port that received the control message.

[0194] C.5. An Illustrative Host Component

[0195] FIG. 31 shows one implementation of a host component 3102, corresponding to any of the host components (S) shown in FIG. 1. The host component 3102 can include

one or more processing devices **3104**, such as one or more central processing units (CPUs), each of which may implement one or more hardware threads. The host component **3102** can also include any storage resources **3106** for storing any kind of information, such as code, settings, data, etc. Without limitation, for instance, the storage resources **3106** may include any of RAM of any type(s), ROM of any type(s), flash devices, hard disks, optical disks, and so on. More generally, any storage resource can use any technology for storing information. Further, any storage resource may provide volatile or non-volatile retention of information. Further, any storage resource may represent a fixed or removable component of the host component **3102**. In one case, the host component **3102** may perform any of the operations associated with local tenant functionality when the processing devices **3104** carry out associated instructions stored in any storage resource or combination of storage resources. The host component **3102** also includes one or more drive mechanisms **3108** for interacting with any storage resource, such as a hard disk drive mechanism, an optical disk drive mechanism, and so on.

[0196] The host component **3102** also includes an input/output module **3110** for receiving various inputs (via input devices **3112**), and for providing various outputs (via output devices **3114**). One particular output mechanism may include a presentation device **3116** and an associated graphical user interface (GUI) **3118**. The host component **3102** can also include one or more network interfaces **3120** for exchanging data with other devices via one or more communication conduits **3122**. One or more communication buses **3124** communicatively couple the above-described components together.

[0197] The communication conduit(s) **3122** can be implemented in any manner, e.g., by a local area network, a wide area network (e.g., the Internet), point-to-point connections, etc., or any combination thereof. The communication conduit(s) **3122** can include any combination of hardwired links, wireless links, routers, gateway functionality, name servers, etc., governed by any protocol or combination of protocols.

[0198] D. Multi-Component Service Functionality

[0199] FIG. 32 provides an overview of functionality for generating and applying a multi-component service. As shown there, the SMC **128** includes a multi-component management component (MCMC) **3202**. The MCMC **3202** interacts with the configuration component **1014** (of FIG. 10) to configure a collection (cluster) of acceleration components. Thereafter, the MCMC **3202** manages the collection. The MCMC **3202** also stores information regarding the connection between the acceleration components in the collection. For example, for each member of the collection, the MCMC **3202** can store its upstream component(s) (if any) and its downstream component(s) (if any).

[0200] FIG. 32 also shows one illustrative collection **3204** of acceleration components (**3206**, **3208**, . . . , **3210**) that perform a multi-component service. That particular collection **3204** of acceleration components (**3206**, **3208**, . . . , **3210**) is structured as a ring. But other collections may exhibit other flow structures. Within the collection **3204**, an acceleration component **3206** represents the head component of the multi-component service.

[0201] The MCMC **3202** also receives failure reports from the failure monitoring component **1016**, introduced above in the context of FIG. 10. The failure reports indicate whether

a link between two acceleration components has failed, e.g., which may manifest itself in the inability to exchange messages over the link. If a failure is confirmed, the MCMC **3202** may heal the multi-component service by swapping out one or more failed acceleration components with suitably configured spare acceleration components. To perform this task, the MCMC **3202** may draw from a pool of spare acceleration components **3212**, e.g., including acceleration components (**3214**, **3216**, . . . , **3218**). For example, assume that the acceleration component **3208** in the collection **3204** fails, which performs a particular part of the multi-component service. The MCMC **3202** can replace this component **3208** with another component from the pool that is already configured to perform the same function. Alternatively, the MCMC **3202** can interact with the configuration component **1014** to configure a spare component in a dynamic manner just prior to its assignment to the collection **3204**. The MCMC **3202** stores information regarding the spare components **3212** that are available at any given time, e.g., as part of the availability information in the data store **1002**.

[0202] A function parsing component **3220** may parse a function into plural parts to create the multi-component service. The function parsing component **3220** may then forward instructions to the MCMC **3202** which describe the manner in which the function has been parsed. The MCMC **3202** uses these instructions to configure the acceleration components (**3206**, **3208**, . . . , **3210**) in the multi-component service.

[0203] FIG. 33 shows another simplified collection **3302** of acceleration components (**3304**-**3314**) that may be created and applied using the functionality of FIG. 32. The collection **3302** includes a head component **3304** that branches out to two parallel paths, including a first path made up of acceleration components **3306** and **3308**, and a second path made up of acceleration components **3310** and **3312** (although the branches can have any number of components). FIG. 34 shows another type of collection **3402** of acceleration components that may be produced and applied by the functionality of FIG. 32. Here, the figure generally shows that the acceleration components can be arranged to form a three-dimensional flow structure.

[0204] More generally, in some cases, a multi-component service may be based on a graph structure which defines a fixed interconnection among its acceleration components. That type of multi-component service will use the same set of acceleration components whenever it is called, and pass information among those components in the fixed manner defined by its graph structure. In yet other cases, a multi-component service may dynamically vary its graph structure at runtime based on one or more factors. In doing so, the multi-component service may use different acceleration components for different invocations, and/or may employ different flows among acceleration components for different invocations.

[0205] For example, consider a multi-component service that performs image recognition. The multi-component service may invoke a first collection of acceleration components for processing a first type of input data. The multi-component service may invoke a second collection of acceleration components for processing a second type of input data. Alternatively, or in addition, the multi-component service may dynamically invoke different acceleration components and/or flows based on real-time performance

conditions, such as experienced congestion of an acceleration component and/or a link.

[0206] FIG. 35 shows an illustrative implementation of the function parsing component 3220, introduced above. The function parsing component 3220 can include an available resource analysis component (ARAC) 3502 for determining the capabilities of the acceleration components 3504 that are available for use in constructing a multi-component service. For example, the ARAC 3502 can query the availability information in the data store 1002 (of FIG. 10) to determine a number of acceleration components that are available, the total processing capacity of each acceleration component, and the amount (and type(s)) of processing capacity in each acceleration component that is currently available for use in implementing a part of a multi-component service.

[0207] A function analysis component 3506 can investigate the function itself that is to be partitioned into plural parts. The function analysis component 3506 can perform this operation in different ways depending on the nature of the function. Consider a function that involves repeating the same basic operation a relatively large number of times. The function analysis component 3506 can determine a total number of times that the operation is performed and the computational load associated with each iteration of the operation.

[0208] The function analysis component 3506 can also identify natural transitions within the flow of a function, if any. For example, a function may be characterized by plural stages, and those stages may map to respective parts of a multi-component service, with transitions in between the parts. In addition, or alternatively, a function may invoke a collection of subroutines, and those subroutines may map to respective parts of a multi-component service, with transitions to and from the subroutines. More generally, a function may include programmatic calls of any type, e.g., where one code module calls on another code module. The function parsing component 3220 can consider those calls as natural points at which to divide a function, effectively replacing internal programmatic calls with calls from one acceleration component to the next. Further, when the function has plural parts or stages, the functionality analysis component 3506 can also determine the amount of processing work associated with each part.

[0209] A partition generation component 3508 uses the results of the ARAC 3502 and the function analysis component 3506 to generate an allocation of the function into multiple parts to be allocated to respective allocation components. For example, consider the case in which the function analysis component 3506 has determined that a function involves repeating a particular task a certain number of times (such as by performing ten million iterations of the Monte Carlo simulation algorithm). The partition generation component 3508 can divvy the entire number of tasks into appropriately sized chunks for allocation to individual acceleration components that are available. More specifically, the partition generating component 3508 can choose a number of acceleration components that is sufficient to perform the total number of tasks, and then allocate appropriate portions to each such acceleration component depending on the particular available capacity of each acceleration component. The partition generation component 3508 can also leverage the natural transition information identified by the function analysis component 3506 in selecting specific par-

tition points, e.g., such that an internal programmatic call is repurposed as a call from one acceleration component to another.

[0210] In other cases, the function parsing component 3220 can identify a finite number of possibly ways of partitioning a function into plural parts, and can identify the resultant characteristics of each option that impact its overall desirability (e.g., in terms of speed of computation, cost, power consumption, thermal profile, and/or any other factors). The function parsing component 3220 can then choose the most favorable partitioning option. For example, the function parsing component 3220 can assign a score to each option that reflects a weighted combination of its characteristic features, and then choose the option with the most favorable score. The weights may be chosen based on environment-specific considerations. In other cases, the function parsing component 3220 can apply known search algorithms (such as best-first) to find a suitable solution within a space of options.

[0211] In another scenario, assume that the function has plural stages that the function parsing component 3220 maps to different acceleration components. But assume that one stage is more labor intensive than the others. To avoid a bottleneck in processing associated with this stage, the function parsing component 3220 can allocate two or more acceleration components that operate in parallel for this stage.

[0212] The function parsing component 3220 can be applied in different use contexts. In one use context, the function parsing component 3220 provides a tool with which a developer may interact to manually explore different partition options.

[0213] In another use context, the function parsing component 3220 operates in at least a partially automated manner. For instance, assume that the data processing system 102 provides at least one multi-component service. Further assume that, at any given time, the multi-component service employs a collection of acceleration components that is structured on the basis of a current partitioning strategy. Upon a triggering event, the data processing system 102 can dynamically invoke the function parsing component 3220 to determine whether a the current partitioning strategy continues to be appropriate in view of prevailing conditions in the data processing system 102. For example, the function parsing component 3220 can perform this operation on a periodic basis and/or on an event-driven basis in the course of the operation of the data processing system 102. If the strategy is no longer appropriate, the function parsing component 3220 dynamically updates the allocation of parts associated with the multi-component service, and then deploys the resultant new multi-component service.

[0214] A previously-chosen partitioning strategy may no longer be appropriate for one or more reasons. For example, the data processing system 102 may contain a different set of available resources than originally encountered, which may warrant a repartitioning of the multi-component service. In addition, or alternatively, the data processing system 102 may encounter real-time performance constraints that may differ upon each invocation of the multi-component service.

[0215] In addition, or alternatively, the nature of the task to be performed itself may change based on various factors. For example, as noted above, the multi-component service may have different processing requirements depending on

the nature of the input information that is fed to it, and/or the nature of the customer's requirements, and so on.

[0216] To cite a particular scenario, assume that the multi-component service corresponds to an iterative algorithm that invokes a different number of repetitive tasks depending on the nature of the input data that is fed to it and/or based on a confidence metric specified by a consumer. To address this scenario, the function parsing component 3220 can dynamically repartition the algorithm based on the real-time data processing needs that it encounters when processing a data set. For example, consider the type of collection 3302 shown in FIG. 33. The function parsing component 3220 may dynamically reduce or increase the number of branches in the collection 3302 in response to the real-time processing needs that it encounters to respectively decrease or increase its processing capacity.

[0217] In one case, the function parsing component 3220 corresponds to a program that runs on one or more software-driven computing devices, e.g., one of the host components shown in FIG. 1, or a dedicated computer server. Alternatively, or in addition, an acceleration component (or components) can implement some aspects of the function parsing component 3220, even without assistance from the software plane 104. For example, an acceleration component can automatically detect congestion in its local processing, or in the processing of other acceleration components in the collection. For instance, the acceleration component may determine that it has failed to produce its output result within a specified amount of time, for whatever reason(s). In response, the acceleration component can automatically generate duplicate versions of itself, which thereupon operate in parallel to alleviate the congestion. For example, once again with reference to FIG. 33, an acceleration component can automatically increase or decrease the number of parallel branches or single nodes upon detecting that the processing in one or more branches or nodes is becoming congested or otherwise underperforming.

[0218] FIG. 36 shows a more detailed example of an illustrative multi-component service, implemented using a collection of acceleration components. Overall, the multi-component service assigns a ranking score to a pairing of a query and a document. The ranking score defines the relevance of the document to the query. Traditionally, such a task is performed entirely in software by applying a model produced by machine-learning.

[0219] In the present implementation, a host component 3602 may invoke the service by sending a request to a first acceleration component of the multi-component service, corresponding to a head component 3604. More specifically, FIG. 36 simplifies the first operation by showing, in operation (1), that the host component 3602 directly interacts with the head component 3604. More generally, the host component 3602 directly interacts with its local acceleration component, which may or may not correspond to the head component 3604. If the local acceleration component is not the head component 3604, the local acceleration component will forward the request to the head component 3604.

[0220] The request itself may include various items of information, such as one or more hit vectors which describe the locations of the query terms within the document under consideration, etc. The request may also specify a collection of software-generated feature values. These software-generated features are computed in software (and not hardware)

for any environment-specific reason (e.g., because such computations do not map well to hardware resources).

[0221] The head component 3604 performs two roles. First, the head component calculates various individual feature values to be used as input information for downstream acceleration components. For example, one such feature value may identify the number of times a query word occurs in the document under consideration. The head component 3604 may also perform a queue management role, to be described in greater detail below.

[0222] The next two acceleration components (3606, 3608) perform more complex feature computations, compared to the feature computations performed by the head component 3604. For example, the more complex feature computations may involve mathematically combining feature values computed by the head component 3604. The next acceleration component 3610 in the sequence compresses the feature values computed thus far.

[0223] The last three acceleration components (3612, 3614, 3616) generate a final ranking score using all of the feature values computed thus far as input information to the scoring calculation, and using any environment-specific score-calculation equation or algorithm. The above-described series of acceleration components then routes the final score back to the head component 3604, and thereafter to the host component 3602. Generally, the numbered arrows in FIG. 36 reflect the sequence of operations that are performed by the acceleration components that make up the multi-component service. Here, the flow structure assumes the form of a ring, in which computation flows in a first direction of the ring and a final result flows in the opposite direction of the ring. But again, many other flow structure as possible. The labels S1, S2, . . . , S6 denote local host components that are associated with the respective acceleration components shown in FIG. 36.

[0224] FIG. 37 shows functionality in the head component 3604 for use in calculating feature values. The functionality includes a stream processing component 3702 for splitting an input stream (e.g., associated with the hit vectors) into plural sub-streams 3704. A collection of feature state machines 3706 then operates on the sub-streams in parallel to generate feature values. A feature gathering component 3708 collects the feature values from the feature state machines and makes them available to downstream acceleration components. Although not shown, the acceleration components (3606, 3608) that perform more advanced feature computations can also leverage parallel computational resources.

[0225] FIG. 38 shows a queue manager component 3802, which may be implemented by the head component 3604 of FIG. 36. The queue manager component 3802 operates by changing a model used by the collection of acceleration components of FIG. 36 to accommodate requests that demand different models. More specifically, different requests received from the host component 102 correspond to queries that are expressed in different natural languages. The different languages, in turn, prompt the queue manager component 3802 to load different respective models to process the different requests. For instance, if a received query is expressed in French, then the queue manager component 3802 will seek to load a French-related model in the multi-component service to act on the query (if that model is not already loaded).

[0226] More specifically, the queue manager component 3802 may maintain plural queues in local memory 3804. Each queue is associated with a different respective model. For example, queue 1 is associated with model 1, queue 2 is associated with model 2, queue 3 is associated with model 3, and so on.

[0227] The queue manager component 3802 includes a request processing component 3806 and a model loading component 3808. In operation, the request processing component 3806 adds each incoming request to an appropriate queue, e.g., by adding the above-noted French query to a French queue. The request processing component 3806 also selects among the queues to process based on any policy, such as by selecting among queues on a round-robin basis, queue-fullness basis, priority basis, etc., or any combination thereof; such a policy may generally seek to fairly arbitrate among queues and requests, while also reducing the frequency at which new queues are selected (and consequently, the frequency at which new models are loaded). Upon switching to a new queue (e.g., having *z* unprocessed requests therein), the model loading component 3808 loads the model associated with that queue into the acceleration components 3810, and then submits the *z* requests in the queue to the acceleration components 3810 for processing based on the loaded new model.

[0228] FIG. 39 is a process 3902 that shows one manner of operation of the function parsing component 3820 of FIG. 35. In block 3904, the function parsing component 3220 receives information regarding a function to be partitioned, together with information regarding the available acceleration components. In block 3906, the function parsing component 3220 partitions the function into two or more parts, for allocation to respective acceleration components. In block 3908, the function parsing component 3220 uses the configuration component 1014 to configure the acceleration components with the parts identified in block 3306 (if not already configured). In block 3910, the function parsing component 3220 determines whether conditions in the data processing system 102 (and/or the nature of function itself) warrant repartitioning the function. If so, the process 3902 returns to block 3904.

[0229] FIG. 40 shows a process 4002 that describes the operation of a particular acceleration component within a collection of acceleration components that implements a multi-component service. In block 4004, the acceleration component receives setup information that identifies the connection of the particular acceleration component with other acceleration components in the collection, e.g., by identifying the upstream component(s) (if any) and the downstream component(s) (if any). In block 4006, the particular acceleration component receives input information from the upstream component(s) or from the requesting host component that has requested use of the multi-component service. In block 4008, the particular acceleration component performs its assigned operation to generate output information. In block 4010, the particular acceleration component sends the output information to the downstream component(s) or to the requesting host component.

[0230] FIG. 41 is a process 4102 that describes one way of handling a failure in a collection of acceleration components that implements a multi-component service. In block 4104, the management functionality 122 determines if an acceleration component has failed. In block 4106, the management functionality 122 swaps the failed acceleration com-

ponent with a spare acceleration component, selected from the pool of spare acceleration components. The spare acceleration component may be already configured to perform the role of the failed acceleration component, or it may be configured to perform that role in an on-demand manner.

[0231] The following summary provides a non-exhaustive list of illustrative aspects of the technology set forth herein.

[0232] According to a first aspect, a data processing system is described that includes two or more host components, each of which uses one or more central processing units to execute machine-readable instructions, the two or more host components collectively providing a software plane. The data processing system also includes two or more hardware acceleration components that collectively provide a hardware acceleration plane, the hardware acceleration plane implementing one or more services. Each host component in the software plane is configured to access a service on one or more of the hardware acceleration components via an associated local hardware acceleration component. Further, at least one service corresponds to a multi-component service having plural parts, and which is implemented on a collection of two or more hardware acceleration components. Each hardware acceleration component in the collection implements a corresponding part of the multi-component service, and each hardware acceleration component in the collection is configured to interact with other hardware acceleration components in the collection without involvement from any host component.

[0233] According to a second aspect, the collection of hardware acceleration components is structured as a ring.

[0234] According to a third aspect, the collection of hardware acceleration components is structured to include parallel paths to accommodate parallel processing.

[0235] According to a fourth aspect, the collection of hardware acceleration components has a three-dimensional structure.

[0236] According to a fifth aspect, the collection of hardware acceleration components has a head component, and wherein a requesting host component that requests use of the multi-component service is configured to access the multi-component service by accessing the head component of the multi-component service via a local hardware acceleration component that is associated with the requesting host component.

[0237] According to a sixth aspect, the collection of hardware acceleration components produces a final output result that is returned to the requesting host component via the head component.

[0238] According to a seventh aspect, the data processing system further includes a multi-component management component that maintains information regarding connections associated with the collection of hardware acceleration components.

[0239] According to an eighth aspect, the data processing system further includes a failure monitoring component for detecting a failure in the multi-component service. The data processing system further includes a multi-component management component that is configured to heal the multi-component service by replacing one or more failed hardware acceleration components in the collection with respective one or more replacement hardware acceleration components, selected from a pool of spare hardware acceleration components.

[0240] According to a ninth aspect, the data processing system further includes a queue manager component that is configured to automatically change a configuration of one or more hardware acceleration components in the collection based on at least one triggering event.

[0241] According to a tenth aspect, the above-referenced triggering event corresponds to a receipt of a request that is associated with a new model, compared to a current model that is currently being used by the collection of hardware acceleration components.

[0242] According to an eleventh aspect, the data processing system further includes a function parsing component that is configured to determine a manner of parsing a function into the plural parts of the multi-component service.

[0243] According to a twelfth aspect, a method is described for generating a multi-component service. The method includes receiving information regarding a function to be implemented in a data processing system, or currently implemented in the data processing system, wherein the data processing system includes two or more host components, each of which uses one or more central processing units to execute machine-readable instructions, the two or more host components collectively providing a software plane. The data processing system further includes two or more hardware acceleration components that collectively provide a hardware acceleration plane, the hardware acceleration plane implementing one or more services. Further, each host component in the software plane is configured to access a service on one or more of the hardware acceleration components via an associated local hardware acceleration component.

[0244] The method further includes: receiving information regarding available hardware acceleration components in the data processing system that can be used to implement the function; partitioning the function into two or more parts to form a multi-component service based on the information regarding the function and the information regarding available hardware acceleration components, each part to be executed by an associated hardware acceleration component in a collection of hardware acceleration components; and configuring the collection of hardware acceleration components, if not already configured, to enable the collection of hardware acceleration components to perform their respective parts.

[0245] According to a thirteenth aspect, the method further includes: determining whether a change has occurred in the data processing system that warrants repartitioning of the function; and repartitioning the function in response to a determination that repartitioning is warranted.

[0246] According to a fourteenth aspect, the above-reference partitioning comprises selecting a number of hardware acceleration components in the collection that is sufficient to perform a total amount of repetitive tasks entailed by the function, each hardware acceleration component in the collection being allocated a respective portion of the total amount of repetitive tasks.

[0247] According to a fifteenth aspect, the method further includes: detecting a failure in the multi-component service; and automatically healing the multi-component service by replacing one or more failed hardware acceleration components in the collection with respective one or more replacement hardware acceleration components, selected from a pool of spare hardware acceleration components.

[0248] According to a sixteenth aspect, a hardware acceleration component is described that includes access to a local link for communicating with a local host component associated with the hardware acceleration component, the local host component using one or more central processing units to execute machine-readable instructions. The hardware acceleration component further includes logic configured to receive setup information that identifies a connection of the hardware acceleration component to other hardware acceleration components in a collection of hardware acceleration components that perform a multi-component service. The hardware acceleration component further includes logic configured to receive input information from one or more upstream hardware acceleration components in the collection, or from a requesting host component that has requested use of the multi-component service. The hardware acceleration component further includes logic configured to perform a part of the multi-component service to provide output information, based on the input information that is received. The hardware acceleration component further includes logic configured to provide the output information to one or more downstream components in the collection, or to the requesting host component. The hardware acceleration component performs its processing without involvement by a requesting host component, and without the requesting host component being aware that it has requested a multi-component service.

[0249] According to a seventeenth aspect, the collection of hardware acceleration components is structured as a ring.

[0250] According to an eighteenth aspect, the collection of hardware acceleration components is structured to include parallel paths to accommodate parallel processing.

[0251] According to a nineteenth aspect, the requesting host component accesses the collection of hardware acceleration components via a head component.

[0252] According to a twentieth aspect, the hardware acceleration component is a field-programmable gate array (FPGA) device.

[0253] A twenty-first aspect corresponds to any combination (e.g., any permutation or subset) of the above-referenced first through twentieth aspects.

[0254] A twenty-second aspect corresponds to any method counterpart, device counterpart, system counterpart, means counterpart, computer readable storage medium counterpart, data structure counterpart, article of manufacture counterpart, graphical user interface presentation counterpart, etc. associated with the first through twenty-first aspects.

[0255] In closing, although the subject matter has been described in language specific to structural features and/or methodological acts, it is to be understood that the subject matter defined in the appended claims is not necessarily limited to the specific features or acts described above. Rather, the specific features and acts described above are disclosed as example forms of implementing the claims.

What is claimed is:

1. A data processing system comprising:

two or more host components, each of which uses one or more central processing units to execute machine-readable instructions, the two or more host components collectively providing a software plane; and

two or more hardware acceleration components that collectively provide a hardware acceleration plane, the hardware acceleration plane implementing one or more services,

each host component in the software plane being configured to access a service on one or more of the hardware acceleration components via an associated local hardware acceleration component,

wherein at least one service corresponds to a multi-component service having plural parts, and which is implemented on a collection of two or more hardware acceleration components,

wherein each hardware acceleration component in the collection implements a corresponding part of the multi-component service,

and wherein each hardware acceleration component in the collection is configured to interact with other hardware acceleration components in the collection without involvement from any host component.

2. The data processing system of claim 1, wherein the collection of hardware acceleration components is structured as a ring.

3. The data processing system of claim 1, wherein the collection of hardware acceleration components is structured to include parallel paths to accommodate parallel processing.

4. The data processing system of claim 1, wherein the collection of hardware acceleration components has a three-dimensional structure.

5. The data processing system of claim 1, wherein the collection of hardware acceleration components has a head component, and wherein a requesting host component that requests use of the multi-component service is configured to access the multi-component service by accessing the head component of the multi-component service via a local hardware acceleration component that is associated with the requesting host component.

6. The data processing system of claim 5, wherein the collection of hardware acceleration components produces a final output result that is returned to the requesting host component via the head component.

7. The data processing system of claim 1, further including a multi-component management component that maintains information regarding connections associated with the collection of hardware acceleration components.

8. The data processing system of claim 1, further comprising:

a failure monitoring component for detecting a failure in the multi-component service; and

a multi-component management component configured to heal the multi-component service by replacing one or more failed hardware acceleration components in the collection with respective one or more replacement hardware acceleration components, selected from a pool of spare hardware acceleration components.

9. The data processing system of claim 1, further including a queue manager component that is configured to automatically change a configuration of one or more hardware acceleration components in the collection based on at least one triggering event.

10. The data processing system of claim 9, wherein the triggering event corresponds to a receipt of a request that is associated with a new model, compared to a current model that is currently being used by the collection of hardware acceleration components.

11. The data processing system of claim 1, further including a function parsing component that is configured to

determine a manner of parsing a function into the plural parts of the multi-component service.

12. A method for generating a multi-component service, comprising:

receiving information regarding a function to be implemented in a data processing system, or currently implemented in the data processing system,

the data processing system including two or more host components, each of which uses one or more central processing units to execute machine-readable instructions, the two or more host components collectively providing a software plane; and

two or more hardware acceleration components that collectively provide a hardware acceleration plane, the hardware acceleration plane implementing one or more services,

each host component in the software plane being configured to access a service on one or more of the hardware acceleration components via an associated local hardware acceleration component,

receiving information regarding available hardware acceleration components in the data processing system that can be used to implement the function;

partitioning the function into two or more parts to form a multi-component service based on the information regarding the function and the information regarding available hardware acceleration components, each part to be executed by an associated hardware acceleration component in a collection of hardware acceleration components; and

configuring the collection of hardware acceleration components, if not already configured, to enable the collection of hardware acceleration components to perform their respective parts.

13. The method of claim 12, further comprising:

determining whether a change has occurred in the data processing system that warrants repartitioning of the function; and

repartitioning the function in response to a determination that repartitioning is warranted.

14. The method of claim 12, wherein the partitioning comprises selecting a number of hardware acceleration components in the collection that is sufficient to perform a total amount of repetitive tasks entailed by the function, each hardware acceleration component in the collection being allocated a respective portion of the total amount of repetitive tasks.

15. The method of claim 12, further comprising:

detecting a failure in the multi-component service; and automatically healing the multi-component service by replacing one or more failed hardware acceleration components in the collection with respective one or more replacement hardware acceleration components, selected from a pool of spare hardware acceleration components.

16. A hardware acceleration component, comprising:

access to a local link for communicating with a local host component associated with the hardware acceleration component, the local host component using one or more central processing units to execute machine-readable instructions;

logic configured to receive setup information that identifies a connection of the hardware acceleration component to other hardware acceleration components in a

collection of hardware acceleration components that perform a multi-component service;
logic configured to receive input information from one or more upstream hardware acceleration components in the collection, or from a requesting host component that has requested use of the multi-component service;
logic configured to perform a part of the multi-component service to provide output information, based on the input information that is received; and
logic configured to provide the output information to one or more downstream components in the collection, or to the requesting host component,
the hardware acceleration component performing its processing without involvement by a requesting host component, and without the requesting host component being aware that it has requested a multi-component service.

17. The hardware acceleration component of claim **16**, wherein the collection of hardware acceleration components is structured as a ring.

18. The hardware acceleration component of claim **16**, wherein the collection of hardware acceleration components is structured to include parallel paths to accommodate parallel processing.

19. The hardware acceleration component of claim **16**, wherein the requesting host component accesses the collection of hardware acceleration components via a head component.

20. The hardware acceleration component of claim **16**, wherein the hardware acceleration component is a field-programmable gate array (FPGA) device.

* * * * *