

(19) World Intellectual Property Organization
International Bureau



PCT



(43) International Publication Date
13 November 2008 (13.11.2008)

(10) International Publication Number
WO 2008/137334 A1

(51) International Patent Classification:
H04Q 7/20 (2006.01)

(21) International Application Number:
PCT/US2008/061401

(22) International Filing Date: 24 April 2008 (24.04.2008)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
11/744,394 4 May 2007 (04.05.2007) US

(71) Applicant (for all designated States except US): **BLACK-WAVE INC.** [US/US]; 289 Great Road, Acton, MA 01720 (US).

(72) Inventors; and

(75) Inventors/Applicants (for US only): **GEROVAC, Branko J.** [US/US]; 15 Lake Street, Lexington, MA 02421 (US). **CARVER, David C.** [US/US]; 6 Independence Avenue, Lexington, MA 02421 (US).

(74) Agents: **ROHLICEK, J. Robin** et al.; Occhiuti Rohlicek & Tsao LLP, 10 Fawcett Street, Cambridge, Massachusetts 02138 (US).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MT, NL, NO, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— with international search report

(54) Title: SYSTEM AND METHOD FOR RENDEZVOUS IN A COMMUNICATIONS NETWORK

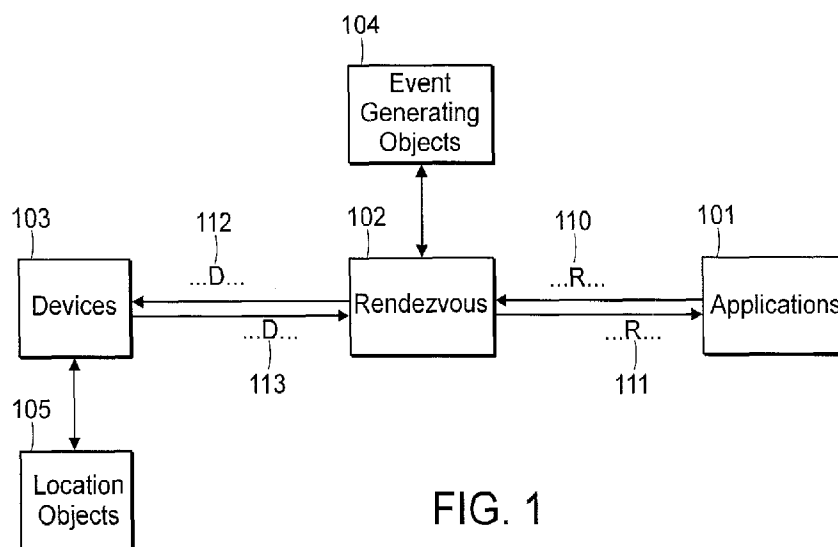


FIG. 1

(57) Abstract: The present invention introduces the notions of a rendezvous component and rendezvous functionality into the communications network environment. Using the invention, an application can express information regarding when an operation requested of a device should complete and at which location, and it enables the device to perform its operations respecting this information while also improving the device's overall behavior. In an embodiment, one or more data objects are distributed across one or more collections of storage devices using a dispersal technique. When access to a data object is desired, a rendezvous component issues a set of constituent requests to the collections of storage devices. These requests typically include location and timing rendezvous parameters specifying a destination location where and a given time when a given data object is to be reconstituted.

WO 2008/137334 A1

SYSTEM AND METHOD FOR RENDEZVOUS IN A COMMUNICATIONS NETWORK

BACKGROUND OF THE INVENTION

5 **Technical Field**

The present invention relates generally to optimizing and harmonizing the use of shared devices by applications in a communications network. Optimization includes performance and scalability; harmonization includes enabling heretofore disparate applications to coexist and share network resources.

10 **Background of the Related Art**

The term “rendezvous” is well-established in the computing literature in the context of programming language support for concurrent computing and real-time operating systems. In particular, computer languages, such as Ada and Distributed Logic Programming (DLP), use the concept of “rendezvous” to refer, respectively, to the
15 process of synchronizing concurrent tasks and objects. Real-time operating systems commonly provide interfaces allowing applications to express deadlines, thereby facilitating deadline scheduling (which is a form of rendezvous scheduling). These uses of the term “rendezvous” are oriented toward managing or scheduling the use of processing resources in closed systems.

20 The term “rendezvous” is also used in literature relating to communications networks. In particular, so-called rendezvous protocols have been defined, but typically in the context of some human or network organizational task. For example, rendezvous scheduling is used to manage meeting rooms, A/V equipment, video conference resources, and other resources for in-person and/or on-line meetings. Or, for example,
25 on-line dating services may arrange for a rendezvous between its clientele. Computer and network applications and/or systems of communications may be employed, but the rendezvous is between humans.

In network environments, “rendezvous” may be used by nodes or services of a cluster in establishing a quorum. “Rendezvous” also is used in network routing to mean
30 a place, a rendezvous point, through which network traffic flows. Further, “rendezvous” is also used in the mobile network context, such as with Bluetooth and other wireless

network technologies, to mean a place where access through surrounding networks is established. These uses are principally concerned with network organization, namely, management, configuration, authorization, connection and session establishment, and the like.

5 Thus, while literature in computing theory and practice describes many potential benefits associated with rendezvous techniques (rendezvous or deadline scheduling, rendezvous synchronization, and the like), such techniques have been limited. Devices in a network may or may not use rendezvous functionality internally – embedded operating systems often include rendezvous interfaces – but rendezvous functionality,
10 even if it exists, is rarely exposed to applications across a communications network and no general method is known to exist. At best, some protocols provide a means for setting priority. Although priority is a useful attribute, it alone severely limits the ability to optimize the use of shared resources in a communications network, especially when heterogeneous usage patterns are involved.

15 Matters concerning performance in a communications network are generally addressed from a Quality of Service (QoS) perspective. QoS mechanisms operate on the service (e.g., Differentiated Services) or session (e.g., Resource Reservation Protocol) level, and are not typically applied to or invoked with individual messages. Indeed, doing so would be at variance with the design intent of such mechanisms.

BRIEF SUMMARY OF THE INVENTION

It is a general object of the invention to create a new notion of *rendezvous* with respect to shared resources across communications networks.

The present invention introduces the notions of a *rendezvous component* and *rendezvous functionality* into the communications network environment. Using the invention, an application can express information regarding when an operation requested of a device should complete and at which location, and it enables the device to perform its operations respecting this information while also improving the device's overall behavior. In one embodiment, one or more data objects are distributed across one or more collections of storage devices using a dispersal technique. When access to a data object is desired, a rendezvous component issues a set of constituent requests to the collections of storage devices. These requests typically include location and timing rendezvous parameters specifying a destination location where and a given time when a given data object is to be reconstituted. The data object is then reconstituted at the destination location and at the given time, while respecting a given operating or performance characteristic of a storage device. In this embodiment, the collection of storage devices may comprise disk devices and the requests are issued by a given application, such as a media server.

In another embodiment, a method is operative in a distributed system comprising a set of storage devices. The method begins upon receipt of a rendezvous access request that comprises a set of zero or more action parameters, a set of zero or more event parameters, and a set of zero or more location parameters. For a given access operation, an action parameter describes an access to be performed, an associated event parameter specifies a time with respect to which the access is to be performed, and the location parameter describes a source or destination object associated with the access. In response to receipt of a rendezvous access request, the request is transformed into a set of one or more device-specific requests. The one or more device-specific requests are then provided to a subset of the storage devices in the set of storage devices such that the given access operation is performed in a manner that satisfies the action parameters, event parameter and the location parameters.

The foregoing has outlined some of the more pertinent features of the invention. These features should be construed to be merely illustrative. Many other beneficial results can be attained by applying the disclosed invention in a different manner or by modifying the invention as will be described.

BRIEF DESCRIPTION OF THE DRAWINGS

For a more complete understanding of the present invention and the
5 advantages thereof, reference is now made to the following descriptions taken in
conjunction with the accompanying drawings, in which:

Fig. 1 shows the general case of rendezvous in a communications network,
including a rendezvous component, its relationship to devices and applications in a
system, and the flow of messages involved with setting up and performing a rendezvous;

10 **Fig. 2** shows the application of rendezvous in a communications network to
access data objects in storage devices of a system; and

Fig. 3 shows the use of rendezvous transformations in enabling multiple
rendezvous components to cooperate.

DETAILED DESCRIPTION OF AN ILLUSTRATIVE EMBODIMENT

Rendezvous in a Communications Network

Referring to **Fig. 1**, one or more rendezvous components **102** extend one or more
5 devices **103** with rendezvous functionality, and provide an interface to this functionality
to one or more applications **101**. Rendezvous interface functionality refers to any setup,
monitoring, and management of rendezvous and rendezvous components **102**. Thus,
some components of the system use the rendezvous interface to obtain rendezvous
functionality whereas other components of the system use the rendezvous interface to
10 manage the operation, health, and performance of one or more rendezvous components
and/or of other components in the system. A device **103** also may be *rendezvous aware*,
meaning the device supports rendezvous functionality and provides an interface to that
functionality (e.g., the equivalent of a rendezvous component has been integrated into
the device). It is envisioned that over time more and more devices will become
15 rendezvous aware.

Although **Fig. 1** shows rendezvous components **102** used with applications **101**
and devices **103** suggesting a client/server context, it is envisioned that rendezvous
components **102** will also be used with objects that communicate as peers (i.e., in a peer-
to-peer context). For example, it is envisioned that grid or peer-to-peer computing
20 systems will benefit from rendezvous. More generally, the distinction between
applications and devices is meant to be instructional, not constraining. Though storage,
communications, and computational objects are of particular interest, applications **101**
and devices **103** may be, without restriction, any objects in the system where one object
wants to request a service from another object; rendezvous can be used to augment the
25 request and/or service. Though **Fig. 1** shows a set of independent components, this also
is intended to be instructional, not constraining. Implementations may, for example,
employ one or more of the following constructs: *Hybrid components*: components with
a relationship to some set of objects in the system and with one or more different
relationships to one or more other sets of objects in the system. For example, a device in
30 one part of a system could be an application to another part of the system; one or more
rendezvous components communicating with sets of devices could in turn relate as

devices to one or more other rendezvous components, etc. *Integrated components*: components that bind together two or more components, or the functionality of two or more components. *Composite components*: components that bind two or more components together such that they appear as components that have private communication links among them. Integrated or composite components that present rendezvous functionality are referred to as being *rendezvous aware* (e.g., rendezvous aware devices, rendezvous aware storages, rendezvous aware processing elements, rendezvous aware network elements, rendezvous aware protocols, etc.). *Distributed components*: components composed of multiple instances, whether geographically localized or disbursed. Multiple instances of a particular component (or collection of components) can use any of various distributed system techniques. For example, an implementation can use multiple rendezvous components in parallel (e.g., to divide load, to service particular subnets, to service particular application classes, to service particular device classes, etc.) and/or in succession (e.g., to provide different forms of rendezvous functionality, to aggregate requests, to disaggregate requests, etc.).

A rendezvous is a requested condition whereby a specified action is to be performed with respect to a specified event and/or with respect to (at) a specified location, e.g., a request for a response by a specified time, a request to send data to a specified object by a specified time, etc. A principal function of rendezvous components is to schedule the performance of actions so as to meet the objectives of their respective rendezvous (this functionality is also called *rendezvous scheduling* in this system description). *Performed with respect to an event* refers to how an action is initiated or completed relative to an event, e.g., an action is to complete before an event occurs, an action is to complete after an event occurs, an action is to start after an event occurs, an action is to start before an event occurs, an action is to start within a probability distribution near an event, an action is to occur with some probability in response to an event, etc. Rendezvous can be applied to any style of scheduling semantics (including alternate event models and exception handling). Different implementations may benefit from different primitives, such as scheduled start semantics, or from a less primitive semantics, such as “complete by time or abort-and-continue without error”, or “best-effort complete by time or complete late”. Many other

such semantics as described in the literature can be applied in this novel concept of rendezvous.

An event is generated by an event generating object **104** in the system, examples of which include periodic time events from a system or network clock, events generated by one or more devices, events generated by one or more applications, etc. Events can be explicitly ordered (e.g., time, enumeration) or otherwise sequenced (e.g., an object's state transitions) or can be non-sequenced (e.g., stochastic). It is expected that rendezvous components **102** will communicate with event generating objects **104** to exchange information about events. Examples include requests for and notifications of events, information exchanged in advance of events that enable rendezvous components to simulate or model events as a proxy to receiving event notification messages, etc. A location **105** is an object in the system where the effects of a specified action are made manifest. The location **105** may be, without restriction, any object in the system, including but not limited to the device **103** performing the action, other devices, applications, rendezvous components in the system, etc. How an effect of an action is made manifest depends on the requirements of the particular implementation and though **Fig. 1** shows devices **103** communicating with locations **105**, this communication is not strictly required (e.g., the means by which the effect of an action is made manifest could be external to the system).

Resulting rendezvous primitives can be used to create additional rendezvous functionality. For example, to initiate an action after a first event and finish it before a second event could be expressed with completion semantics primitives as two consecutive rendezvous statements 'complete wait before first event' followed by 'complete action before second event'. Thus, it is envisioned that implementations often will employ one or more suitable means of organizing rendezvous primitives. Examples include, but are not limited to, sequencing requests, nesting one request within another, grouping requests into tasks, issuing concurrent requests, defining conditional requests, issuing interdependent requests, defining compound requests, etc. This principle of creating more complex functionality through the use of such organizations is intended to apply to all functionality described in this system description thereby enabling more

clear and concise description of primitive functionality without loss of generality or claims to invention of more complex functionality.

A rendezvous protocol is a set of rules governing access to rendezvous interface functionality through the transmission of messages to and from rendezvous components

5 **102**. Similarly, a device protocol is a set of rules governing access to device interface functionality through the transmission of messages to and from devices **103**. Different implementations may employ different rendezvous and device protocols, and a given implementation may employ one or more rendezvous and device protocols. Rendezvous protocols and device protocols also may be integrated to produce rendezvous aware

10 protocols. Further, a rendezvous protocol may encapsulate or otherwise augment one or more device protocols thereby allowing rendezvous functionality to be added to a system in a manner that is transparent to certain components and subsystems. This may be done, for example, in situations where devices **103** are not rendezvous aware, or where devices **103** are rendezvous aware, but the protocol(s) being used do not support

15 rendezvous functionality. A rendezvous protocol may also operate out-of-band from a device protocol it is augmenting, for example, applications **101** may communicate separately to devices **103** and to rendezvous components **102** which in turn communicate with devices **103** to produce desired rendezvous behaviors.

Here and throughout this system description, communications between

20 components of the system are described in terms of representative messages moving from one component to another for select pairs of components. This is intended to be instructive, not constraining. Depending on the implementation, components may communicate even if a line of communications is not shown in the figures. Zero or more messages are expected to be sent and/or received in accordance to the rules of one or

25 more protocols as described above. Different messages between a pair of components need not travel the same communications path. The description of a single representative message between a pair of components is not intended to imply a requirement to use one-to-one communication: the use of one-to-one (e.g., unicast), one-to-many (e.g., multicast), and/or many-to-one (e.g., multi-homing) methods of

30 communication may be advantageous depending on the situation and the requirements of the implementation.

R 110 represents a message sent by an application 101 to a rendezvous component 102. The message may contain a request for the rendezvous component to perform some function or may contain a description of some state or change of state in the application. Requests include, but are not limited to requests to setup and manage rendezvous (these requests are also called *rendezvous requests* in this system description). Rendezvous requests may include a set of action parameters, a set of event parameters, and a set of location parameters, though these sets may overlap (e.g., event and/or location information may be included in the set of action parameters) and one or more of these sets may be null. Action parameters describe an action to be performed. Event parameters describe an event with respect to which the action is to be performed. Location parameters describe a location with respect to which the action is to be performed. Rendezvous request parameters may also include contingency or failure response conditions. For example, with some types of rendezvous it may be important to abort execution and report an error to an application; with other types of rendezvous it may be important to proceed with execution; with yet other types of rendezvous it may be important to abort execution according to some timeout condition even though the rendezvous conditions have not been met.

R 111 represents a message transmitted by a rendezvous component 102 to an application 101. The message may contain a description of some state or change of state in the rendezvous component or may contain a request for the application to perform some function. A description may include the status of outstanding or completed requests, the status of one or more rendezvous, the state of one or more rendezvous components, the state of one or more other components in the system (for example, information pertaining to the operation, health, and performance of one or more devices). Further, it may be advantageous to include information concerning the conditions under which requested actions were completed. It may also be advantageous to transmit multiple messages in response to a request: one to acknowledge receipt of a request; one to notify the application about the progress and completion of a requested action, one to indicate when the associated rendezvous event occurs, etc. Such information may be used by applications to regulate admissions to the system. For example, if actions are consistently completed well in advance of their rendezvous

requirements, applications may infer that it is safe to increase admissions. If, however, actions complete near to or exceeding their rendezvous requirement, applications may choose to deny future admissions or even to shed or to shift loads to protect the viability of the system as a whole.

5 D **112** represents a message transmitted by a rendezvous component **102** to a device **103**. The message may contain a request for the device to perform some function (these requests are also called *device requests* in this system description) or contain a description of some state or change in state in the rendezvous component. A requested function may, among other things, relate to the target action of a rendezvous or to
10 monitoring and managing the operation, health, and performance of one or more devices **103** or other components in the system. Typically, though this is not strictly required, a rendezvous component **102** will transmit one or more messages D **112** to a device **103** in response to receiving one or more messages R **110** from an application **101**. A principal function of rendezvous components **102** is to use available information to transform one
15 or more rendezvous requests into one or more device requests so as to optimize the use of resources and to thereby optimize some aspect of system performance and/or behavior. Examples of such transformations include, but are not limited to reordering, commingling, sequencing, merging, dividing, etc. requested actions into a set of device requests. Examples of available information include, but are not limited to, the current
20 request parameters, information about other past and present requests, information about one or more devices, information about the state of one or more rendezvous components **102**, information about applications, information about the communications network, information about other components in the system, etc.

 D **113** represents a message transmitted by a device **103** to a rendezvous
25 component **102**. The message may contain a description of state or changes of state in the device or contain requests for the rendezvous component to perform some function. A description may include, for example, the status of outstanding or completed requests, the state of a device, the state of one or more other objects in the system, etc. For example, message D **113** may be sent to acknowledge receipt of message D **112**, and
30 subsequent messages may be sent to notify a rendezvous component **102** about the progress and completion of the requested function. A message may also include

information pertaining to the operation, health, or performance of a device **103** or of other objects in the system. Note that a device **103** may transmit messages to other objects in the system without intervention of a rendezvous component **102**. Further, in some implementations, some communications between applications **101** and devices **103** may involve rendezvous components **102**, and some may not (e.g., applications **101** and devices **103** may communicate directly for some operations).

It is anticipated that rendezvous components **102** will collect information relating to the performance of rendezvous requests and device requests as compared to the goals of the requests and that this information could be queried through a management interface. Some implementations will benefit from having system components exchange management information; management components, not shown, may also exist that monitor and/or control various aspects of the system. For example, a device health monitor could be used to inform an application whether a particular device is operating normally, operating in some degraded mode, or not operating at all. Or, for example, data generated by a rendezvous component could be collected by a process that acts to balance loads and to relieve congestion in the system.

Some implementations will benefit from creating an abstract device model with individual device models for each type or class of device. For example, device models could include any access characteristics that determine access behavior and performance, e.g., (a) disk drive track, cylinder, and sector access relationships, (b) disk zones, (c) logical block, (d) memory bank, page, block access relationships, etc. Thereby, a single rendezvous scheduler (or collection of schedulers) can be used to access many different types of devices or to optimize different aspects of devices by using the appropriate device model. Device models may be very rich, and may describe many aspects of device behavior under various operating conditions, such as incorporating an aging model of the device's behavior. Device models may be static or dynamic. For example, a dynamic model might adapt based on past device behavior. A detailed device model can be developed for a device prior to incorporating it into a particular implementation, or a device model can be derived when a device is incorporated into a particular implementation through a discovery process that tests and probes device behavior with modeling workloads.

Rendezvous may be adaptive, and may determine scheduling based on changes occurring in the types of requests being made and in the manner in which actions are performed within or across devices and components in the system. For example, rendezvous may adapt scheduling decisions due to factors and behaviors such as load
5 and/or degradation.

Rendezvous may be provisional. A provisional rendezvous associates a potential future action with an event, but the action is not performed unless a subsequent request is received. Generally a provisional rendezvous indicates a probable, but not committed action. A provisional rendezvous allows a rendezvous component to prepare for the
10 request in advance of actually receiving the request. It may be advantageous for a provisional rendezvous request to include information related to the probability of a subsequent request.

Rendezvous functionality may be normative and/or informational. Normative functionality is functionality that is required for an application to behave correctly.
15 Informational functionality is functionality that permits a rendezvous component to optimize some aspect of the system, but is not necessarily required by an application.

Access Rendezvous in a Communications Network

Referring to **Fig. 2**, rendezvous scheduling is applied to accessing data objects in a communications network. One or more rendezvous components **202** extend one or
20 more storage devices **203** with rendezvous functionality, and provide an interface to this functionality to one or more applications **201**. Applications **201** can be, for example, direct producers or consumers of stored objects in the system, or for example, they can be components that mediate access to stored objects to one or more other components in the system. Storage devices **203** are examples of devices **103**, time sources **204** are
25 examples of event generating objects **104**, and applications **201** are examples of applications **101**. Applications **201** and rendezvous components **202**, or objects associated with these components, are also examples of rendezvous locations **105**. Throughout this section of the system description the term *access* is used broadly, meaning for example, to create, delete, read, write, modify, monitor, manage, etc.
30 storage devices **203** and the objects the storage devices **203** handle. Storage devices **203** can be, without restriction, any kind and combination of storage devices. For example,

storage devices **203** can be individual physical storage devices (e.g., disk drives, memory stores, etc.), aggregated physical storage devices (e.g., disk arrays, clustered storage, storage area networks, etc.), logical or virtual storage devices (e.g., partitions, volumes, or other storage abstractions and organizations), etc. More generally, it will be
5 evident to one studying the details of this invention that the broad definition of component and the flexibility in component organizations and topologies described for **Fig. 1** can be applied to implement many useful systems and sub-systems incorporating access rendezvous functionality. In this system description, integrated or composite components that present rendezvous functionality together with storage functionality are
10 referred to as *rendezvous aware storage devices*.

A rendezvous-access protocol is a set of rules governing access to rendezvous aware storage functionality through the transmission of messages to and from rendezvous components **202**. Similarly, an access protocol is a set of rules governing access to storage through the transmission of messages to and from storage devices **203**.
15 It is anticipated that rendezvous aware access protocols will be defined, that rendezvous awareness will be added to existing access protocols, and that access aware rendezvous protocols will be defined for use in conjunction with access protocols. Examples of existing access protocols that would benefit from being made rendezvous aware range from lower layer protocols such as the Small Computer Systems Interconnect (SCSI),
20 protocols for alternative SCSI transport (e.g., iSCSI), FibreChannel, protocols for Direct Data Placement (DDP), and Remote Direct Memory Access (RDMA), etc. to upper layer protocols such as HyperText Transport Protocol (HTTP), File Transport Protocol (FTP), peer-to-peer protocols (e.g., BitTorrent), etc. Rendezvous enabling HTTP, for example, would offer tremendous benefits: it would enable improved performance of traditional
25 World Wide Web (WWW/Web) services by providing a means to discriminate, organize, and optimize high volume handling of disparate and unassociated messages and message flows; further, it would enable the application of HTTP to efficiently handle data types, such as audio and video, that require something more advanced than best effort access and delivery semantics.

30 **R 210** represents a message sent from an application **201** to a rendezvous component **202**; message **R 210** is an example of message **R 110**. **R 210** may be, for

example, a rendezvous request with a set of action parameters that specify an access to be performed, a set of event parameters that specify a time with respect to which the access is to be performed, and a set of location parameters that specify one or more source and/or destination objects associated with the access (these sets of parameters
5 may overlap and one or more of the sets may be null). For example, the request could be to read a data object from storage into an application buffer by a specified time, or to write a data object to storage from an application buffer by a specified time, or to read data object from storage into a sequence of application buffers by a sequence of specified times, etc. Message R **210** is also referred to as a *rendezvous-access request*
10 in this system description.

R **211** represents a message sent from a rendezvous component **202** to an application **201**: message R **211** is an example of message R **111**. R **111** may describe a state or change in state in the rendezvous component; zero or more such messages may be sent in response to a rendezvous-access request, for example, to acknowledge the
15 request, to describe the disposition (e.g., scheduling) of the requested rendezvous-access, to indicate progress, success, or failure of the requested rendezvous-access, etc. It may or may not be advantageous, depending on the implementation, for message R **211** to transfer data associated with the requested rendezvous-access.

A **212** represents a message sent from a rendezvous component **202** to a storage device **203**; message A **212** is an example of message D **112**. Message A **212** may be, for example, an access request. A principal function of rendezvous components **202** is to transform rendezvous-access requests into access requests such that (1) accesses are performed in a manner that satisfy the parameters of their associated rendezvous, and (2)
20 the performance of one or more components (e.g., storage, processing, network components, etc.) in the system is improved thereby increasing the scalability of the system. For example, by using rendezvous request information a rendezvous component **202** can transform a random work load generated by unassociated applications into an organized work load that minimizes head movement in disk storage devices. As discussed below, access requests generated by rendezvous components can be
25 rendezvous-access requests as would be the case, for example, when the storage devices are themselves rendezvous aware.
30

A **213** represents a message sent from a storage device **203** to a rendezvous component **202**; message A **213** is an example of message D **113**. Zero or more such messages may be sent in response to an access request, for example, to acknowledge the request, to indicate progress, success, or failure of the requested access, etc. It may or
5 may not be advantageous, depending on the implementation, for message A **213** to transfer data associated with the requested access.

X **215** represents a message sent from an application **201** or a rendezvous component **202** to a storage device **203**, or a message sent to an application **201** or a rendezvous component **202** from a storage device **203**. Depending on the
10 implementation and the circumstances, it may or may not be advantageous for data associated with accesses to flow through one or more rendezvous components **202**. Fig. 2 highlights the option of having data transfers such as X **215** flow directly between applications **201** and storage devices **203** thereby relieving rendezvous components **202** from the burden of handling some or all such data in addition to other tasks. Note that
15 depending on the implementation one or more data transfers such as X **215** may be performed in response to a rendezvous-access request. Further, depending on the implementation and the circumstances, the order that data is transmitted may or may not be the same as the order in which the data is subsequently used. These alternatives highlight the flexibility in system design and optimization that rendezvous bring to
20 accessing objects in a communications network (e.g., rendezvous ensures that applications get what they need when and where they need it, while the system is given greater freedom to optimize its behavior in responding to such needs).

It is anticipated that implementations of rendezvous components **202** will vary in sophistication depending on multiple factors: sophistication in rendezvous component
25 and protocol design, knowledge about storage devices **203** and the access methods used, information exchanged with applications **201**, network components, and other components of the system, etc. It is further anticipated that depending on the implementation it may be advantageous for rendezvous components to incorporate support for mirrored storage devices (e.g., to perform load balancing) as well as for other
30 storage device abstractions and virtualizations.

The ability to probe a data object may be supported directly in the access protocol or it may be supported by using provisional rendezvous. A probe is generally used to indicate probable, but not committed future accesses as is the case, for example, when an interactive application presents a user with a set of choices. A web page, DVD menu, or virtually any user interface containing links to other stored data objects could probe some or all of the data objects associated with choices thereby giving rendezvous components **202** information about potential future requests. Rendezvous components **202** may use such information to prepare for prompt response once a choice is made, for example, by bringing probed data objects into a cache. A programmatic (versus an interactive) application could also use probes. For example, if the outcome of one access determines which of a set of accesses is performed next, the one access could be accompanied with one or more probes depending on the likely next accesses.

The following scenarios describe how access rendezvous in a communications network is important to optimizing aspects of a system. These are a few representative examples; further scenarios will be evident to one studying the details of this invention.

In one scenario, one or more applications **201** generate rendezvous-access requests according to its performance requirements, and one or more rendezvous components **202** interface to one or more disk devices (examples of storage devices **203**). Rendezvous components **202** use rendezvous-access request information as input to a heuristic that optimizes disk performance – the literature describes a number of heuristic approaches (e.g., EDF, SCAN, LOOK, CSCAN, PSCAN, FDSCAN, SSTF, SSED0/V, etc.) and other variations are possible depending on the requirements and objectives of a particular implementation. In this scenario, rendezvous components **202** model one or more disk devices with which they communicate and maintain a queue of outstanding accesses for each. As rendezvous-access requests arrive at a rendezvous component, the rendezvous and access parameters are examined, and using one or more device models each request is placed in one or more access queues so as to minimize disk head movement and rotational delays while also giving priority to completing accesses by their rendezvous time. In some implementations, rendezvous components **202** might divide a large request into multiple smaller requests (e.g., to balance access with other requests) or might consolidate or combine separate small requests into a

larger request (e.g., to reduce message overhead) in order to better manage device operation and access objectives.

In a second scenario, the first scenario is extended such that rendezvous components **202** either detect or are informed explicitly that a data object is being
5 accessed in a particular temporal/spatial pattern or distribution (e.g., periodic sequential, interactive statistical, best effort quasi-random, etc.) such that probability and other information parameters can be associated with potential future rendezvous-access requests (i.e., predictions can be made). Marshalling a streaming media object to a record or playback application is one example: rendezvous components **202** may
10 transform a request to access one block of the object into a request to access many blocks of the object, thereby moving or staging blocks that are likely soon to be accessed into more readily accessible memory (e.g., into a device cache).

In a third scenario, streaming applications, interactive applications, and search applications are commingled. Each exhibit a different access pattern: sequential
15 periodic, statistical interactive, and best-effort quasi-random respectively. Here, rendezvous-access requests provide crucial information that lets rendezvous components **202** optimize the performance of storage devices **203** across heretofore disparate types of access. In particular, the rendezvous parameters enable the accesses generated by query applications to proceed when the accesses generated by streaming and interactive
20 applications with distant rendezvous times can be deferred. Further, the rendezvous information may call for the accesses generated by interactive applications to proceed even if the rendezvous time is exceeded, thus should contention arise, the interactive application avoids an outright denial of service by allowing the accesses generated by streaming applications to take precedence. Optimizations, especially opportunistic
25 optimizations, can be applied across all accesses. For example, a sequential access with a distant rendezvous time (or a portion thereof) may be grouped or sequenced with a best effort random access that is in close spatial proximity (e.g., on disk). Or the converse, a best effort random access may be grouped or sequenced with a sequential access with a near rendezvous time.

30 These scenarios describe a few ways that rendezvous in a communications network (rendezvous components **202** and the exchange of rendezvous information) can

be used to optimize aspects of a system. Though disk drives are the common examples given in these scenarios, it is important to recognize (a) that optimizations may involve other resources in the system (e.g., network bandwidth, network congestion, link utilization, power consumption), and (b) that all systems, subsystems, and components, not only disk drives, have resource constraints that can be managed with rendezvous. Whereas in these scenarios the rendezvous information is used to increase disk performance by reducing head seek and rotational delays (significant factors in disk performance), rendezvous information is equally applicable to increasing performance in situations where a different set of constraints dominate (e.g., switching, processing, or power resources in a RAM based storage).

Further, the above scenarios focus on the use and benefits of rendezvous access to storage with a specified rendezvous time. In so doing, multiple access requests can be managed to optimize overall performance, while honoring rendezvous times of the requests. Rendezvous concepts can be applied in many other innovative ways. Rendezvous can be applied to optimize one or more access parameters instead of or in addition to time and performance, such as power saving, device wear and failure conditions, latency, graceful degradation, buffer capacity, channel bandwidth, network traffic shaping, etc. One or more of these parameters would be incorporated into alternate device models similar to those described above.

A solid-state memory storage device, for example, might contain several banks of memory. When not being accessed, a bank of memory can be placed into a standby powered state to reduce power consumption and heat generation. When being accessed, a bank of memory would be placed into an active powered state. The device model describes how the memory is divided into banks and the operational behavior of power switching and accessing banks of memory. Rendezvous scheduling can be used to order multiple accesses so as to group accesses to a bank of memory, such that the bank of memory transitions from standby to active powered state for the group of accesses then back to standby. Thereby, power switching and consumption costs (e.g., power, latency, etc.) can be optimized across the group of accesses. Also, accesses can be grouped to overlap active access to one bank of memory, while the next bank to access is being prepared to switch from standby to active power, and so forth.

In addition to storage, rendezvous can be used to manage the full breadth of resources in the network and the resources of the network – servers, services, connections, routers, routes, etc. For example, rendezvous can be used to manage a shared compute service on a network, so that multiple requests for service can be reordered to optimize overall service performance. Moreover, as (micro)computer architectures, increasingly use more networking concepts, a rendezvous protocol can be used for intra computer accesses. For example, memory access by one or more CPUs can be augmented similarly to the scenarios described above. Rendezvous provides more flexibility and control of access than existing cache management and memory probe mechanisms. Rendezvous can be applied to access other computer components beside memory.

Rendezvous Transformations in a Communications Network

Referring to **Fig. 3**, one or more additional rendezvous components are added to the arrangement illustrated in **Fig. 2** such that (1) components **302a**, examples of rendezvous components **202**, relate to applications **301** as a set of rendezvous aware storage devices and relate to rendezvous components **302b** as a set of rendezvous applications, and (2) components **302b**, also examples of rendezvous components **202**, relate to rendezvous components **302a** as a set of rendezvous aware storage devices and relate to storage devices **303** as a set of rendezvous components. Rendezvous components **302a** could, for example, present a rendezvous aware storage abstraction (e.g., rendezvous aware storage aggregation or virtualization) to applications **301**, or could, for example, be an assembly component that assembles and/or disassembles (or controls the assembly and/or disassembly) of objects in a communications network. **Fig. 3** highlights for instructional/illustrative purposes just one of many useful arrangements envisioned with respect to the present invention. Though **Fig. 3** shows two sets of rendezvous components (**302a** and **302b**), any number or topology of rendezvous components could be employed depending on the requirements of an implementation. Applications **301** are examples of applications **201**. Storage devices **303** are examples of storage devices **203**.

R 310a represents a message sent by an application **301** to a rendezvous component **302a**; message **R 310a** is an example of message **R 210**. **R 311a** represents a

message sent by a rendezvous component **302a** to an application **301**; message **R 311a** is an example of message **R 211**. **R' 310b** represents a message sent by a rendezvous component **302a** to a rendezvous component **302b**; message **R' 310b** is an example of message **R 210**. **R' 311b** represents a message sent by a rendezvous component **302b** to a rendezvous component **302a**; message **R' 311b** is an example of message **R 211**. A principal function of rendezvous component **302a** is to transform zero or more rendezvous-access requests such as **R 310a** into zero or more rendezvous-access requests such as **R' 310b** and to transform zero or more messages such as **R' 311b** into zero or more messages such as **R 311a**. **A 312** represents a message sent by a rendezvous component **302b** to a storage device **303**; message **A 312** is an example of message **A 212**. **A 313** represents a message sent to a rendezvous component **302b** from a storage device **303**; message **A 313** is an example of message **A 213**. **X 315a-c** represent messages sent from one or more applications **301**, one or more rendezvous components **302a**, or one or more rendezvous components **302b** to one or more storage devices **303**, or messages sent to one or more applications **301**, one or more rendezvous components **302a**, or one or more rendezvous components **302b** from one or more storage devices **303**; messages **X 315a-c** are examples of message **X 215**. As in **Fig. 2**, **Fig. 3** highlights the option for data transfers to be handled separately or even out-of-band using separate communications network resources. **Fig. 3** further highlights that in some implementations rendezvous components may not be involved in the data transfers (e.g., **X 315a-c**) or may only be involved under certain circumstances.

Consider an example application for this kind of arrangement: One or more data objects are striped or otherwise distributed across one or more rendezvous aware storage devices in a communications network. A rendezvous aware assembly component uses information about the placement of the data objects to transform an incoming rendezvous-access request such as **R 310a** into one or more constituent rendezvous-access requests such as **R 310b** sent to the appropriate rendezvous aware storage device(s). The assembly component could, for example, take a request to read a single object into an application buffer by a specified time and transform the request into two or more constituent requests each to read a piece of the object from an appropriate rendezvous aware storage device into an appropriate portion of the application buffer.

The access and rendezvous parameters for the constituent requests would be transformed from those in the original request such that all the pieces of the object rendezvous at the right places in the application buffer by the appointed time. The assembly component may or may not be involved in the actual data transfers (e.g., X **315a-c**) depending on the data placement and reconstruction strategies used. The assembly component could also issue probes or provisional rendezvous-access requests to storage devices not affected directly by the current request in preparation for subsequent accesses.

Consider a second example: A disk drive with an integrated rendezvous component (a rendezvous-aware disk device) is built to support tens or hundreds of outstanding requests. One or more such devices could be linked to one or more other rendezvous components that handle thousands or tens of thousands or more of outstanding rendezvous requests by transforming them into a sequence of requests appropriate to the rendezvous capabilities of each rendezvous aware device.

One should expect an interactive application to be given higher priority than a data mining application, as responsiveness is important to interactive applications. Yet, perception of responsiveness in many situations can be achieved with actual response times in the range of $\frac{1}{4}$ to $\frac{3}{4}$ second. High priorities would dictate that requests from interactive applications be pushed in front of lower priority requests without regard to effects on device performance and system efficiency. With the addition of a rendezvous component as described here, the request can be processed at a time that yields better overall or aggregate performance and efficiency while still achieving desired responsiveness.

One should further consider that in a large scale system where many applications or services are issuing requests to a shared device, the resulting work load will likely appear random. To many devices, random work loads result in worst case operation. The rendezvous component introduces functionality that enables the order of fulfilling requests to be modified to remove at least some of the randomness, thereby enabling improved performance (such as increased aggregate service bandwidth) and/or device behavior (such as improved MTBF, e.g., as a result of more ordered mechanical movement of the disk head arm or less power cycling of semiconductor memory (RAM, flash, etc.). More generally, the rendezvous component introduces functionality that

modifies how the request is fulfilled in order to meet operating objectives; in addition to changing the order of requests, a request could be delayed, or redirected to an alternate device, or rejected, and the like.

5 Though several examples here present the rendezvous component as a separate entity, some implementations will benefit from having affinity between a rendezvous component and a device or set of devices. More generally, the rendezvous component may have affinity with other components in a system (storages, network switches and routers, applications, and the like), and the current invention envisions and enables such affinities. Further, some implementations will benefit from devices with integrated
10 rendezvous components. Again, more generally, the rendezvous component may be integrated with other components in a system (storages, network switches and routers, applications, components, etc.), and the current invention envisions and enables such integrations. Thus, the concept of a rendezvous-aware component is introduced wherein a component has integrated rendezvous capabilities. A rendezvous component as
15 described herein is inherently rendezvous-aware.

As noted above, matters concerning performance in a communications network are generally addressed from a Quality of Service (QoS) perspective. QoS mechanisms operate on the service (e.g., Differentiated Services) or session (e.g., Resource Reservation Protocol) level, and are not typically applied to or invoked with individual
20 messages. The current invention may coexist with QoS mechanisms, for example, where the communications network is engineered, provisioned, and/or equipped with appropriate QoS attributes. Further, the current invention envisions and enables network components (switches, routers, and such) to handle data differentially based on information contained in rendezvous protocols (i.e., network components can be
25 rendezvous-aware).

The present invention extends and augments rendezvous scheduling to operate in a communications network, thereby overcoming many present limitations (e.g., scale, scalability, cost, applicability, interoperation, manageability, and the like) of networked systems.

30 Moreover, the basic rendezvous concepts described above are useful in many application environments. As described above, particular attention has been given to

scenarios where devices are storage devices containing data objects and support some kind of access (read, write, probe, and the like), although this is not a limitation of the invention. Consequently, this invention is envisioned as useful to (a) building systems that scale from local to enterprise, metropolitan, regional, to national uses, and beyond, and (b) building systems that unify the handling of radically different data objects without compromising on the benefits that separate and specialized systems traditionally offered.

While given components of the system have been described separately, one of ordinary skill will appreciate that some of the functions may be combined or shared in given instructions, program sequences, code portions, and the like.

While the above describes a particular order of operations performed by certain embodiments of the invention, it should be understood that such order is exemplary, as alternative embodiments may perform the operations in a different order, combine certain operations, overlap certain operations, or the like. References in the specification to a given embodiment indicate that the embodiment described may include a particular feature, structure, or characteristic, but every embodiment may not necessarily include the particular feature, structure, or characteristic.

The invention can take the form of an entirely hardware embodiment, an entirely software embodiment, or an embodiment containing both hardware and software elements. In one preferred embodiment, the initial layout and reflow algorithms are implemented in software executing in one or more server machines. The invention (or portions thereof) may take the form of a computer program product accessible from a computer-usable or computer-readable medium providing program code for use by or in connection with a computer or any instruction execution system. A computer-usable or computer readable medium can be any device or apparatus that can include, store or communicate the program for use by or in connection with the instruction execution system, apparatus, or device. The medium can be an electronic, magnetic, optical, or the like. Examples of a computer-readable medium include a semiconductor or solid state memory, magnetic tape, a removable computer diskette, a random access memory (RAM), a read-only memory (ROM), a rigid magnetic disk and an optical disk. Current

examples of optical disks include compact disk – read only memory (CD-ROM), compact disk – read/write (CD-R/W) and DVD.

The principals of this invention described here are meant to be illustrative examples, and are not meant to limit the scope of the invention. One skilled in the art
5 will recognize that further combinations, permutations, and equivalents of the invention are possible. Thus, the present invention is intended to embrace all such embodiments.

Having described our invention, what is claimed is as follows:

CLAIMS

1. A method operative in a distributed system comprising a set of storage
5 devices, comprising:
receiving a rendezvous access request that comprises a set of zero or more action
parameters, a set of zero or more event parameters, and a set of zero or more location
parameters, wherein, for a given access operation, an action parameter describes an
access to be performed, an associated event parameter specifies a time with respect to
10 which the access is to be performed, and the location parameter describes a source or
destination object associated with the access;
in response to receipt of a rendezvous access request, transforming the
rendezvous access request into a set of one or more device-specific requests;
providing the one or more device-specific requests to a subset of the storage
15 devices in the set of storage devices such that the given access operation is performed in
a manner that satisfies the action parameters, event parameter and the location
parameters.
2. The method as described in claim 1 wherein the set of storage devices
20 comprise disk devices and the rendezvous access request is generated by a given
application.
3. The method as described in claim 2 wherein the given application is one
of: a media server, and an http server.
25
4. The method as described in claim 3 wherein the media server application
is a streaming application.
5. The method as described in claim 1 wherein a given storage device has an
30 associated device model that describes a given access characteristic of the device, and
wherein the device-specific request provided to the given storage device is a function of
the given access characteristic in the device model.

6. The method as described in claim 1 wherein a given device-specific request provided to a given storage device includes information associated with one of: a current access, and a current access and an expected future access.

5

7. The method as described in claim 1 wherein the given access operation has an associated access pattern characteristic, wherein the characteristic is one of: sequential periodic, statistical interactive, and best-effort quasi-random.

10 8. The method as described in claim 7 wherein the given access operation is prioritized with respect to one or more other access operations as a function of the associated access pattern characteristic.

15 9. A method operative in a distributed system comprising a set of devices, wherein a given device has an associated device model that describes a given device characteristic, comprising:

receiving a rendezvous request that comprises a set of parameters;
in response to receipt of a rendezvous access request, transforming the rendezvous access request into a set of one or more device-specific requests, wherein a
20 given device-specific request is a function of a given device characteristic;
providing the one or more device-specific requests to a subset of the devices in the set of devices such that a given operation is performed in a manner that satisfies the set of parameters in the rendezvous request.

25 10. The method as described in claim 9 wherein the devices are storage devices.

11. The method as described in claim 9 wherein the set of parameters include one of: time, performance, power saving, device wear and failure conditions, latency,
30 graceful degradation, buffer capacity, channel bandwidth, network traffic shaping, and combinations thereof.

12. The method in claim 1 wherein one or more data objects are distributed across one or more collections of storage devices using a dispersal technique, and wherein access to the data objects includes a set of constituent requests to the collections
5 of storage devices, the requests including location and timing rendezvous parameters specifying a destination location where and a time when a given data object is to be reconstituted.

13. A method operative in a distributed system comprising a set of storage
10 devices, comprising:

distributing data objects across a collection of storage devices using a dispersal technique;

upon a given occurrence, issuing a set of constituent requests to the collection of storage devices, the requests including location and timing rendezvous parameters
15 specifying a destination location where and a given time when a given data object is to be reconstituted;

reconstituting the given data object at the destination location and at the given time.

14. The method as described in claim 13 wherein the collection of storage
20 devices comprise disk devices and the given occurrence is receipt of a rendezvous access request generated by a given application.

15. The method as described in claim 14 wherein the given application is one
25 of: a media server, and http server.

16. The method as described in claim 15 wherein the media server application is a streaming application.

17. The method as described in claim 13 wherein a given storage device has
30 an associated device model that describes a given access characteristic of the device, and

wherein the constituent request provided to the given storage device is a function of the given access characteristic in the device model.

1/2

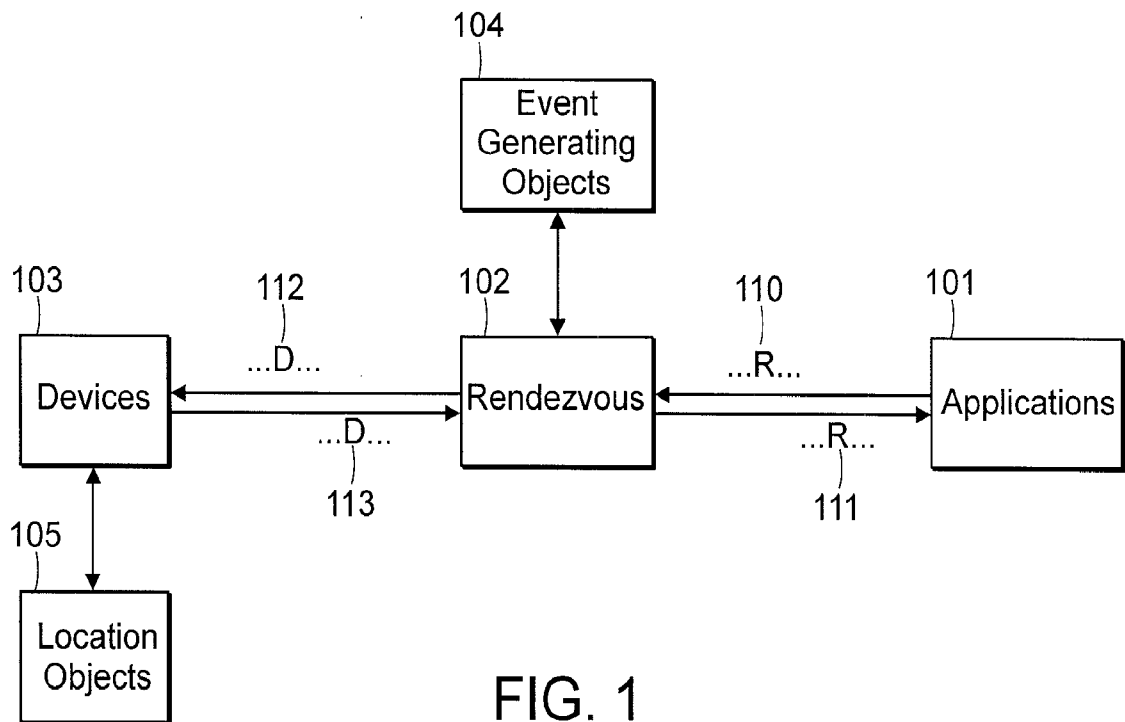


FIG. 1

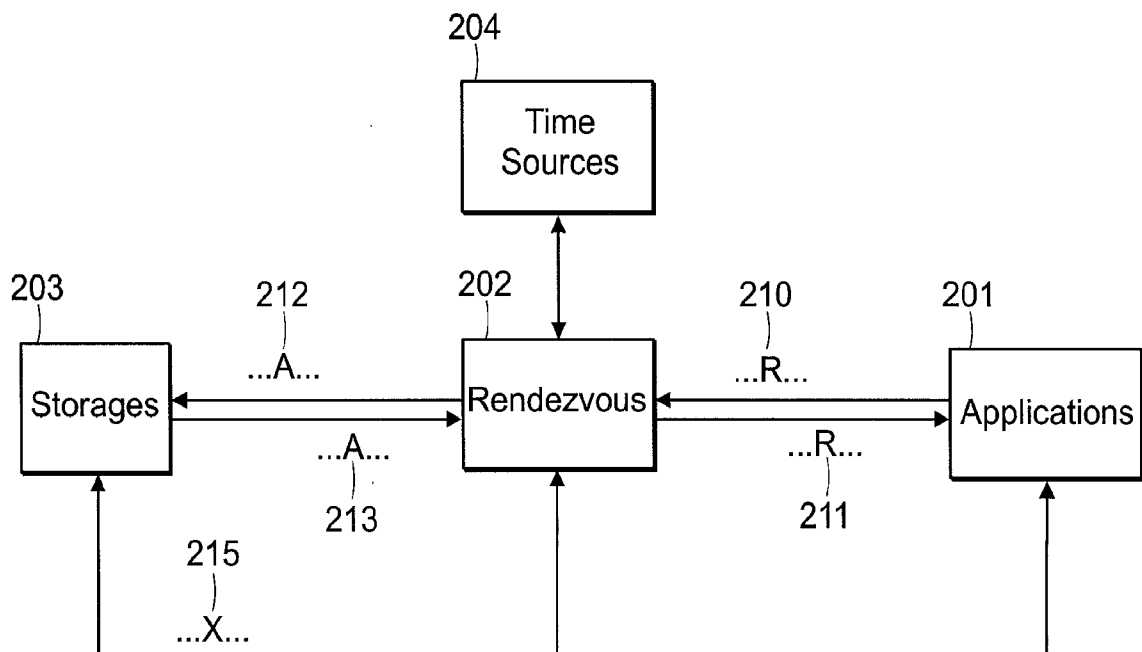


FIG. 2

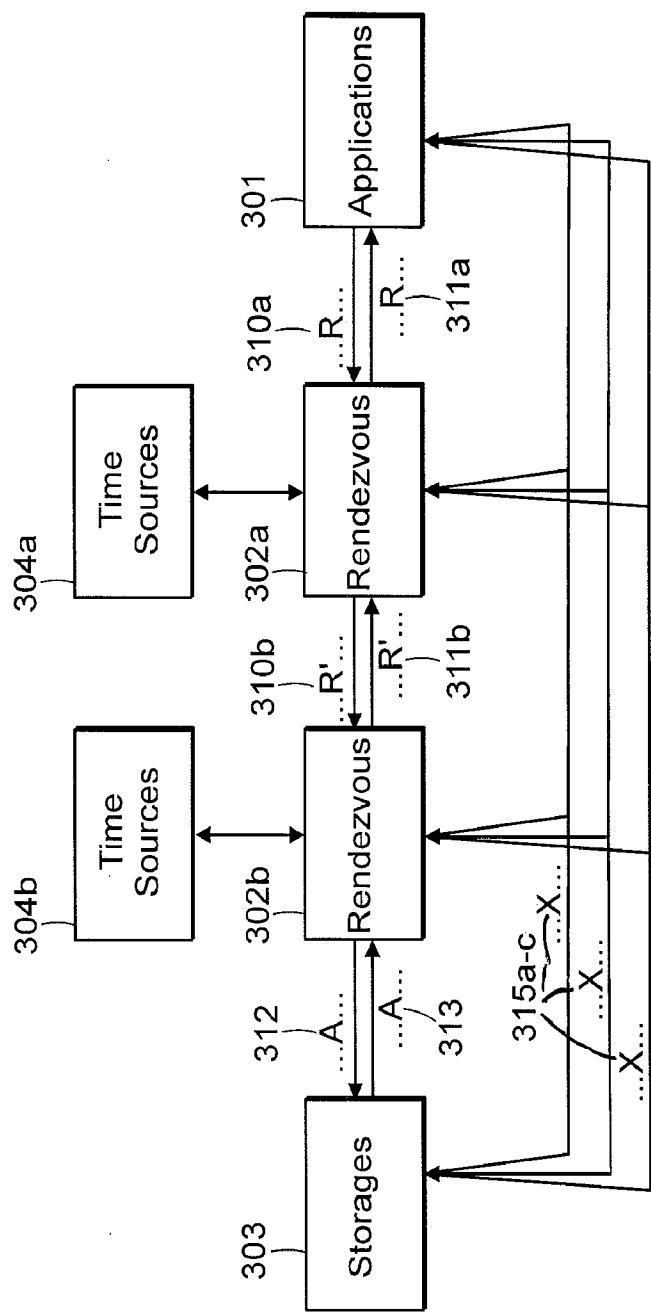


FIG. 3

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US2008/061401

A. CLASSIFICATION OF SUBJECT MATTER

IPC(8) - H04Q 7/20 (2008.04)

USPC - 455/456.1

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC(8) - H04Q 7/20 (2008.04)

USPC - 455/456.1

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

MicroPatent

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2005/0181803 A1 (WEAVER et al) 18 August 2005 (18.08.2005) Entire Document.	1-17
A	US 2004/0198386 A1 (DUPRAY) 07 October 2004 (07.10.2004) Entire Document.	1-17

☐ Further documents are listed in the continuation of Box C.


* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

12 June 2008

Date of mailing of the international search report

01 JUL 2008

Name and mailing address of the ISA/US

Mail Stop PCT, Attn: ISA/US, Commissioner for Patents
P.O. Box 1450, Alexandria, Virginia 22313-1450

Facsimile No. 571-273-3201

Authorized officer:

Blaine R. Copenheaver

PCT Helpdesk: 571-272-4300

PCT OSP: 571-272-7774