



- (51) **International Patent Classification:**  
*G06F 11/14* (2006.01)
- (21) **International Application Number:**  
PCT/EP2013/066040
- (22) **International Filing Date:**  
30 July 2013 (30.07.2013)
- (25) **Filing Language:** English
- (26) **Publication Language:** English
- (71) **Applicant:** NEC EUROPE LTD. [DE/DE]; Kurfürsten-Anlage 36, 69115 Heidelberg (DE).
- (72) **Inventors:** DUSI, Maurizio; Dossenheimer Landstraße 42, 69121 Heidelberg (DE). FIORI, Luca; Rottmannstraße 42, 69121 Heidelberg (DE). GRINGOLI, Francesco; Via G. Zadei, 36, I-25123 Brescia (IT).
- (74) **Agents:** HAGMANN, F. et al.; Ullrich & Naumann, Schneidmühlstraße 21, 69115 Heidelberg (DE).
- (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, BN, BR, BW, BY, BZ, CA, CH, CL, 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, 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, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, 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, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

**Published:**

— with international search report (Art. 21(3))

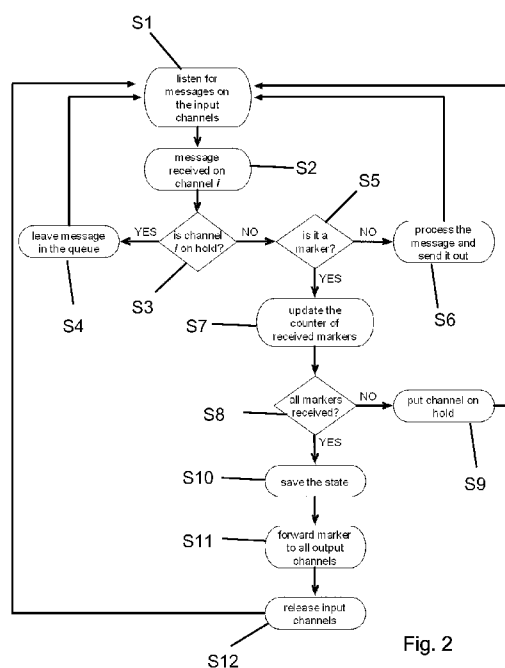
(54) **Title:** METHOD AND SYSTEM FOR CHECKPOINTING A GLOBAL STATE OF A DISTRIBUTED SYSTEM

Fig. 2

(57) **Abstract:** The present invention relates to a method for check pointing a global state of a distributed system with one or more distributed applications, wherein the one or more distributed applications are organized in a directed acyclic graph topology, and wherein sources providing data to one or more tasks each having one or more input channels and one or more output channels for exchanging processed data between tasks and wherein a task processes data received on its input channels and wherein processed data is sent out on one or more of its output channels to other tasks, and wherein one or more destinations collect processed data, comprising the steps of • a) Upon receiving of a marker in an active input channel of a task the active input channel is put on hold, • b) Perform check pointing by saving the internal state of the task when all input channels are on hold, • c) Forward the marker message via all output channels of the task to other tasks, and • d) Reactive all the input channels of the task, wherein the global state is the union of all internal states of the tasks after each task has been check pointed.

## METHOD AND SYSTEM FOR CHECKPOINTING A GLOBAL STATE OF A DISTRIBUTED SYSTEM

5 The present invention relates to a method for check pointing a global state of a distributed system with one or more distributed applications, wherein the one or more distributed applications are organized in a directed acyclic graph topology, and wherein sources providing data to one or more tasks each having one or more input channels and one or more output channels for exchanging processed data between tasks and wherein a task processes data received on its input channels and wherein processed data is sent out on one or more of its output channels to other tasks, and wherein one or more destinations collect processed data.

15 The present invention further relates to a distributed system with one or more distributed applications on a plurality of nodes wherein the nodes are operable to one or more distributed applications which are organized in a directed acyclic graph topology, and wherein sources providing data to one or more tasks each having one or more input channels and one or more output channels for exchanging processed data between tasks and wherein a task processes data received on its input channels and wherein processed data is sent out on one or more of its output channels to other tasks, and wherein one or more destinations collect processed data, preferably for performing with a method according to one of the claims 1-5.

25 Check pointing techniques are used in distributed computing systems for recording a consistent global state of an asynchronous system. When an application running on such a system is composed of several processes or tasks each of them running in parallel and exchanging messages between each other through connecting channels check pointing takes a snap shot of each process at a given point in time in terms of which messages the process is elaborating and the states of its internal variables, for example the value of some counters and takes a snap shot of each channel in terms of the messages sent but not yet received. The global state is given by the union of the internal state of each process and of all the channels. The execution of the application can be resumed from the latest snap shot in case of a system failure.

The time needed to complete the check pointing operation is in particular important in real-time applications having usually a high message rate. Messages coming into the system cannot be controlled but must be processed.

5

Conventional check pointing techniques have to serialize the state of each channel resulting in a high message rate. One of the drawbacks is therefore that the operation of serializing the state of each channel heavily slows down the entire execution which might also lead to violate real-time requirements.

10

One of the conventional techniques for check pointing of applications running on distributed computing systems is based on the so-called Chandy-Lamport algorithm based on the non-patent literature of K. Mani Chandy and Leslie Lamport. 1985, "Distributed snapshots: determining global states of distributed systems", *ACM Trans. Comput. Syst.* 3, 1 (February 1985), 63-75. DOI=10.1145/214451.214456, <http://doi.acm.org/10.1145/214451.214456>. The algorithm uses marker messages and ensures that a consistent global state of a distributed computing system can be saved under the following assumptions:

15

1. There are no failures and all messages arrive intact and only once.
2. The communication channels are unidirectional and First-In-First-Out (FIFO) ordered.
3. There is a communication path between any two processes in the system
4. Any process may initiate the snapshot algorithm.
5. Each process in the system records its local state and the state of its incoming channels.

20

25

The Chandy-Lamport algorithm has inter alia the drawback, that it requires to save for each process both its internal state and the state of all its input channels. Saving the state of the channels requires to (de)serialize messages slowing down the execution and systems with high message rates, which is in particular typical in stream applications.

30

It is therefore an objective of the present invention to provide a method for check

pointing a global state of a system and a system which enables a fast and reliable execution even with a high message rate.

5 It is a further objective of the present invention to provide a method for check pointing a global state of a system and a system which saves computing and memory resources.

10 It is an even further objective of the present invention to provide a method for check pointing a global state of a system and a system which is more flexible in terms of parallel execution of applications when check pointing.

The aforementioned objectives are accomplished by a method of claim 1 and a system of claim 6.

15 In claim 1 a method for check pointing a global state of a distributed system with one or more distributed applications is defined with one or more distributed applications, wherein the one or more distributed applications are organized in a directed acyclic graph topology, and wherein sources providing data to one or more tasks each having one or more input channels and one or more output  
20 channels for exchanging processed data between tasks and wherein a task processes data received on its input channels and wherein processed data is sent out on one or more of its output channels to other tasks, and wherein one or more destinations collect processed data.

25 According to claim 1 the method is characterized by the steps of

- a) Upon receiving of a marker in an active input channel of a task the active input channel is put on hold,
- b) Perform check pointing by saving the internal state of the task when all input channels are on hold,
- 30 c) Forward the marker via all output channels of the task to other tasks, and
- d) Reactive all the input channels of the task,

wherein the global state is the union of all internal states of the tasks after each task has been check pointed.

In claim 6 a distributed system with one or more distributed applications on a plurality of nodes is defined, wherein a system with one or more distributed applications on a plurality of nodes wherein the nodes are operable to execute one or more distributed applications which are organized in a directed acyclic graph topology, and wherein sources providing data to one or more tasks each having one or more input channels and one or more output channels for exchanging processed data between tasks and wherein a task processes data received on its input channels and wherein processed data is sent out on one or more of its output channels to other tasks, and wherein one or more destinations collect processed data, preferably for performing with a method according to one of the claims 1-5.

According to claim 6 the system is characterized in that

- a) put an active input channel of a task running on the node upon receiving of a marker in the active input channel on hold,
- b) perform check pointing by saving the internal state of the task when all input channels are on hold,
- c) forward the marker via all output channels of the task to other tasks, and
- d) to reactivate all the input channels of the task,

wherein the global state is the union of all internal states of the tasks after each task has been check pointed.

According to the invention it has been recognized that serializing and saving messages on the channels is not required to obtain a consistent global state.

According to the invention it has been further recognized that offloading the process from serializing the messages on its channels saves computing and memory resources.

According to the invention it has been further recognized that a continued execution of the applications is enabled while check pointing takes place on a task.

- 5 According to the invention it has been further recognized that there is no need to save the state of the channels.

Further features, advantages and preferred embodiments are described in the following subclaims.

10

According to a preferred embodiment a marker is provided by the one or more sources downstream along the processes of the directed acyclic graph topology. This enables an easy implementation with the need to generate further markers by intermediate nodes.

15

According to a further preferred embodiment the input and output channels are unidirectional and/or messages in these channels are ordered according to the first-in-first-out principle. This allows an easy handling of messages in channels.

20

According to a further preferred embodiment upon receiving further messages in input channels on hold, these further messages are queued until the input channel is reactivated. By queuing the messages reliable snap shot of the state of the channel is provided without losing messages to be processed in the future upon reactivation.

25

According to a further preferred embodiment step b) and c) are swapped. This enables to forward the marker to downstream tasks without having to wait for the checkpointing operation to complete. Thus, parallelization is enabled.

30

There are several ways how to design and further develop the teaching of the present invention in an advantageous way. To this end it is to be referred to the patent claims subordinate to patent claim 1 on the one hand and to the following explanation of preferred embodiments of the invention by way of example, illustrated by the figure on the other hand. In connection with the explanation of the

preferred embodiments of the invention by the aid of the figure, generally preferred embodiments and further developments of the teaching will be explained. In the drawings

5      Fig. 1      shows a system according to a first embodiment of the present invention;  
and

Fig. 2 shows a flow diagram for a method according to a second embodiment of the present invention.

Fig. 1 shows a system according to a first embodiment of the present invention.

15 In Fig. 1 distributed applications with a direct acyclic graph DAG topology are shown. Sources source 1, source 2 inject data into the system and intermediate tasks a, b, c, d, e, f process the data. A destination destination collects the process data by the tasks task a-f and eventually exports it.

Fig. 2 shows a flow diagram for a method according to a second embodiment of  
20 the present invention.

In Fig. 2 a flow chart of an embodiment of the present invention is shown.

A process or task listens for messages on its input channels in a first step S1.

When a message is received on a channel  $i$  in a second step S2 it is determined in a third step S3 if this channel  $i$  is on hold.

30 If yes then in a fourth step S4 the message is queued in the input channel i and the steps S1-S3 are performed again.

If the channel  $i$  is not on hold then in a fifth step S5 it is checked whether the message received on channel  $i$  is a marker message.

If the received message is not a marker message then in a sixth step S6 the message is processed and sent out via one or more output channels and steps S1-S3 are performed again.

- 5      If the message received on channel *i* is a marker message then in a seventh step S7 a counter of received markers is updated, i.e. incremented by +1.

Then in a eighth step S8 it is checked if all markers from all input channels have been received.

10

If the counter of received markers is smaller than the number of input channels, then in a ninth step S9 the channel *i* is put on hold and steps S1-S3 are performed again.

- 15      If all markers have been received then in a tenth step S10 the state of the process is saved. After that the markers are forwarded to all output channels of a process in an eleventh step S11. In a final step S12 the input channels are released, i.e. are reactivated and steps S1-S3 are performed again. To parallelize operations, step S10 and S11 can be swapped.

20

In other words the action of saving the state of each process is postponed until the process receives a marker from all its input channels.

- 25      In summary the present invention enables to check point the state of an application which is composed of distributed processes exchanging messages in a DAG topology, wherein only the internal state of the processes is check pointed but not their channels. Further the present invention does not have to serialize and save messages on the channels.

- 30      The present invention has inter alia the following advantages: The present invention enables a fast check pointing even when the message rate for the messages exchange is high. A further advantage is that computing and memory resources are saved due to the offload of processes from serializing the messages on its channels. An even further advantage is that an execution of applications is

enabled to continue while the snap shot respectively the check pointing takes place on a process or task.

5 Many modifications and other embodiments of the invention set forth herein will come to mind the one skilled in the art to which the invention pertains having the benefit of the teachings presented in the foregoing description and the associated drawings. Therefore, it is to be understood that the invention is not to be limited to the specific embodiments disclosed and that modifications and other embodiments are intended to be included within the scope of the appended claims. Although  
10 specific terms are employed herein, they are used in a generic and descriptive sense only and not for purposes of limitation.

## Claims

- 5 1. A method for check pointing a global state of a distributed system with one or more distributed applications (source 1, source 2, task a-f, destination), wherein the one or more distributed applications (source 1, source 2, task a-f, destination) are organized in a directed acyclic graph (DAC) topology, and wherein sources (source 1, source 2) providing data to one or more tasks (task a-f) each having  
10 one or more input channels and one or more output channels for exchanging processed data between tasks (task a-f) and wherein a task (task a-f) processes data received on its input channels and wherein processed data is sent out on one or more of its output channels to other tasks (task a-f), and wherein one or more destinations (destination) collect processed data,  
15 characterized by the steps of
- a) Upon receiving of a marker in an active input channel of a task (task a-f) the active input channel is put on hold,
  - b) Perform check pointing by saving the internal state of the task (task a-f) when all input channels are on hold,
  - 20 c) Forward the marker via all output channels of the task (task a-f) to other tasks (task a-f), and
  - d) Reactive all the input channels of the task (task a-f),
- wherein the global state is the union of all internal states of the tasks after each task has been check pointed.
- 25 2. The method according to claim 1, characterized in that a marker is provided by the one or more sources (source 1, source 2) downstream along the tasks (tasks a-f) of the directed acyclic graph topology (DAG).
- 30 3. The method according to one of the claims 1-2, characterized in that the input and output channels are unidirectional and/or messages in these channels are ordered according to the first-in-first-out principle.

4. The method according to one of the claims 1-3, characterized in that upon receiving messages in input channels on hold, these messages are queued until the input channel is reactivated.

5. The method according to one of the claims 1-4, characterized in that steps b) and c) are swapped.

6. A distributed system with one or more distributed applications (source 1, source 2, task a-f, destination) on a plurality of nodes wherein the nodes are operable to execute one or more distributed applications which are organized in a directed acyclic graph (DAG) topology, and wherein sources (source 1, source 2) providing data to one or more tasks (task a-f) each having one or more input channels and one or more output channels for exchanging processed data between tasks (task a-f) and wherein a task (task a-f) processes data received on its input channels and wherein processed data is sent out on one or more of its output channels to other tasks (task a-f), and wherein one or more destinations (destination) collect processed data, preferably for performing with a method according to one of the claims 1-5

characterized in that a node is operable to

- a) put an active input channel of a task (task a-f) running on the node upon receiving of a marker in the active input channel on hold,
- b) perform check pointing by saving the internal state of the task (task a-f) when all input channels are on hold,
- c) forward the marker via all output channels of the task (task a-f) to other tasks (task a-f), and
- d) to reactivate all the input channels of the task (task a-f),

wherein the global state is the union of all internal states of the tasks (task a-f) after each task (task a-f) has been check pointed.

1/2

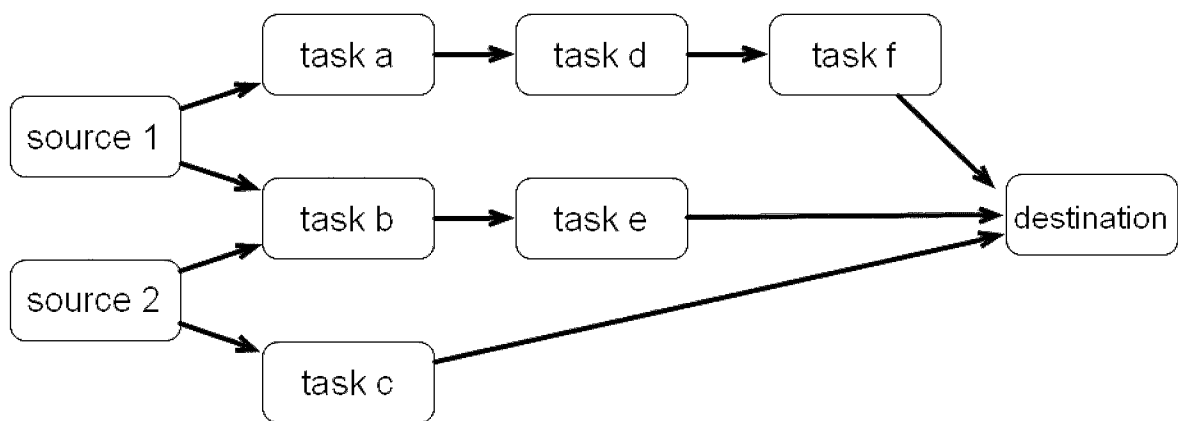


Fig. 1

2/2

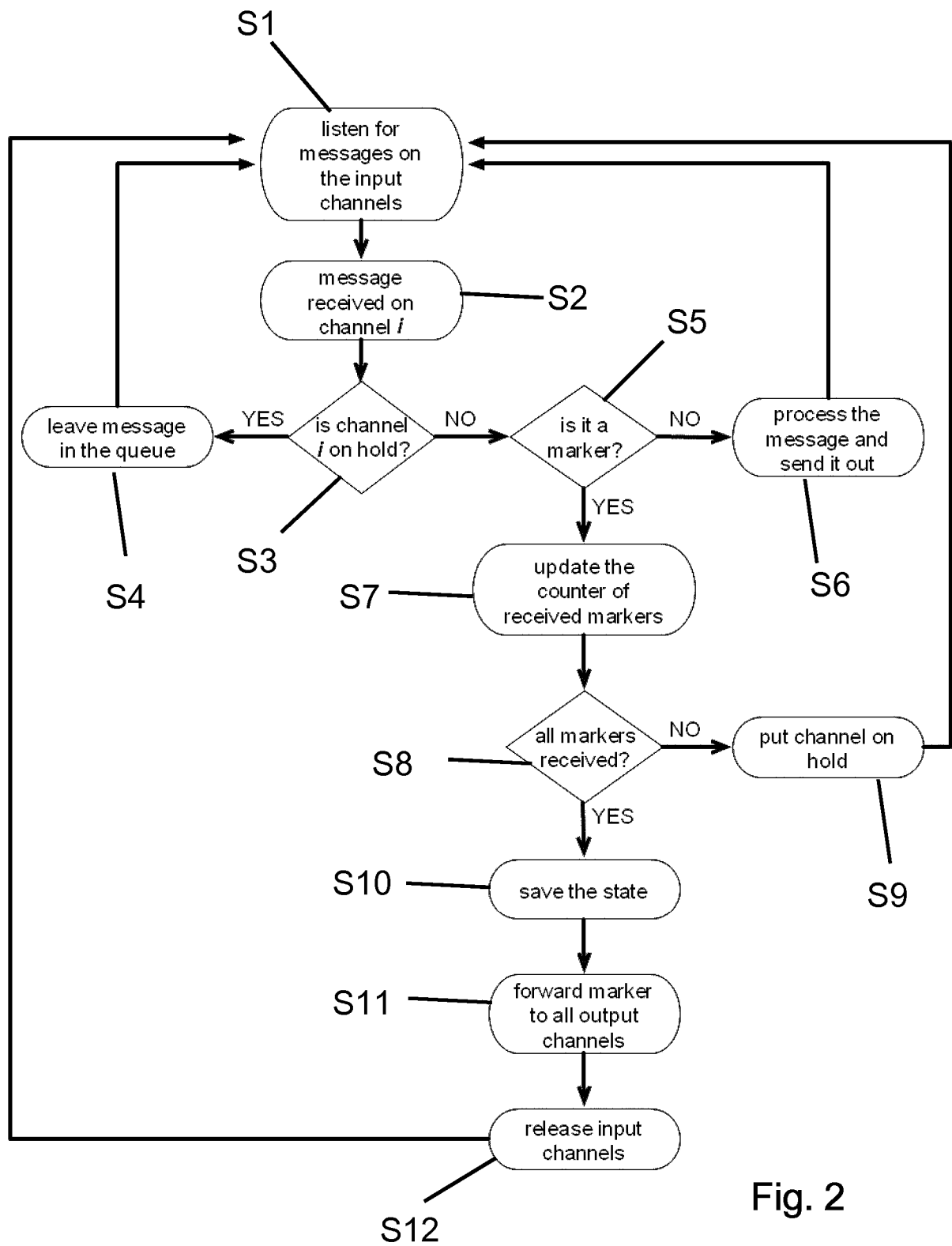


Fig. 2

## INTERNATIONAL SEARCH REPORT

International application No  
PCT/EP2013/066040

A. CLASSIFICATION OF SUBJECT MATTER  
INV. G06F11/14  
ADD.

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)  
G06F

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)

EPO-Internal, WPI Data

## C. DOCUMENTS CONSIDERED TO BE RELEVANT

| Category* | Citation of document, with indication, where appropriate, of the relevant passages                                                                                                                                            | Relevant to claim No. |
|-----------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------|
| X         | <p>WO 01/42920 A1 (INITIO SOFTWARE CORP AB [US]) 14 June 2001 (2001-06-14)<br/>abstract; figures 2,3<br/>page 3, lines 4-22<br/>page 6, line 3 - page 7, line 15<br/>page 8, line 2 - page 10, line 15<br/>-----<br/>-/--</p> | 1-6                   |



Further documents are listed in the continuation of Box C.



See patent family annex.

\* 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 March 2014

Date of mailing of the international search report

20/03/2014

Name and mailing address of the ISA/

European Patent Office, P.B. 5818 Patentlaan 2  
NL - 2280 HV Rijswijk  
Tel. (+31-70) 340-2040,  
Fax: (+31-70) 340-3016

Authorized officer

Weber, Vincent

## INTERNATIONAL SEARCH REPORT

International application No  
PCT/EP2013/066040

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

| Category* | Citation of document, with indication, where appropriate, of the relevant passages                                                                                                                                                                                                                                                                                            | Relevant to claim No. |
|-----------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------|
| A         | <p>MANI CHANDY K ET AL: "DISTRIBUTED SNAPSHOTS: DETERMINING GLOBAL STATES OF DISTRIBUTED SYSTEMS", ACM TRANSACTIONS ON COMPUTER SYSTEMS (TOCS), ASSOCIATION FOR COMPUTING MACHINERY, INC, US, vol. 3, no. 1, 1 February 1985 (1985-02-01), pages 63-75, XP002342863, ISSN: 0734-2071, DOI: 10.1145/214451.214456 cited in the application the whole document</p> <p>-----</p> | 1-6                   |
| A         | <p>US 5 802 267 A (SHIRAKIHARA TOSHIO [JP] ET AL) 1 September 1998 (1998-09-01) abstract figures 2,3,5,7 page 1, lines 27-40 page 3, lines 8-20 page 3, line 64 - page 4, line 22 page 7, line 23 - page 8, line 44</p> <p>-----</p>                                                                                                                                          | 1-6                   |
| A         | <p>US 2009/182782 A1 (KARVE ALEXEI A [US] ET AL) 16 July 2009 (2009-07-16) abstract paragraph [0018] claim 1</p> <p>-----</p>                                                                                                                                                                                                                                                 | 1-6                   |

# INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/EP2013/066040

| Patent document<br>cited in search report | Publication<br>date | Patent family<br>member(s)  | Publication<br>date         |
|-------------------------------------------|---------------------|-----------------------------|-----------------------------|
| WO 0142920                                | A1                  | 14-06-2001                  | AT 525691 T 15-10-2011      |
|                                           |                     | AU 4309801 A 18-06-2001     |                             |
|                                           |                     | CA 2392300 A1 14-06-2001    |                             |
|                                           |                     | EP 1244965 A1 02-10-2002    |                             |
|                                           |                     | HK 1046976 A1 09-03-2012    |                             |
|                                           |                     | JP 3696832 B2 21-09-2005    |                             |
|                                           |                     | JP 2003516581 A 13-05-2003  |                             |
|                                           |                     | US 6584581 B1 24-06-2003    |                             |
|                                           |                     | US 2001042224 A1 15-11-2001 |                             |
|                                           |                     | WO 0142920 A1 14-06-2001    |                             |
| US 5802267                                | A                   | 01-09-1998                  | JP 3154942 B2 09-04-2001    |
|                                           |                     |                             | JP H09138754 A 27-05-1997   |
|                                           |                     |                             | US 5802267 A 01-09-1998     |
| US 2009182782                             | A1                  | 16-07-2009                  | CN 101910971 A 08-12-2010   |
|                                           |                     |                             | EP 2247996 A1 10-11-2010    |
|                                           |                     |                             | JP 2011510381 A 31-03-2011  |
|                                           |                     |                             | KR 20100120131 A 12-11-2010 |
|                                           |                     |                             | TW 200937218 A 01-09-2009   |
|                                           |                     |                             | US 2009182782 A1 16-07-2009 |
|                                           |                     |                             | WO 2009089408 A1 16-07-2009 |