



- (51) **International Patent Classification:**
H04L 12/771 (2013.01) *H04L 12/721* (2013.01)
- (21) **International Application Number:**
PCT/CN2014/080973
- (22) **International Filing Date:**
27 June 2014 (27.06.2014)
- (25) **Filing Language:** English
- (26) **Publication Language:** English
- (71) **Applicant:** **TSINGHUA UNIVERSITY** [CN/CN];
Tsinghua University, Haidian District, Beijing 100084 (CN).
- (72) **Inventors:** **XIANG, Dong**; Tsinghua East Side Building,
Haidian District, Beijing 100084 (CN). **LIU, Xiaowei**;
Tsinghua East Side Building, Haidian District, Beijing
100084 (CN). **YU, Zhigang**; Tsinghua East Side Building,
Haidian District, Beijing 100084 (CN).
- (74) **Agent:** **CN-KNOWHOW INTELLECTUAL PROP-
ERTY AGENT LIMITED**; 18th Floor, Tower B, CEC
Plaza, No. 3 Dan Ling Street, Haidian District, Beijing
100080 (CN).

(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, IR, 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, SA, 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:** DEADLOCK-FREE ADAPTIVE ROUTING OF INTERCONNECT NETWORK

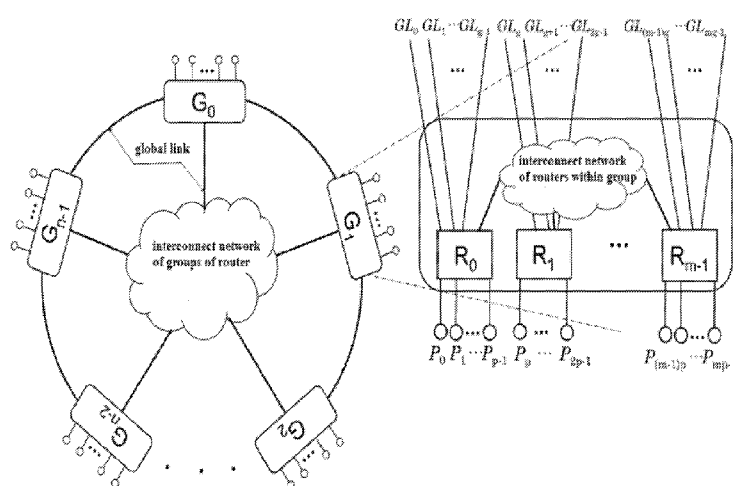


FIG. 1

(57) **Abstract:** The present invention discloses a novel interconnect network system. The system divides the routers in the system into n groups, and n groups of routers achieve full connectivity. Each group contains one or more routers; and routers within a group achieve full connectivity. The present invention also provides an adaptive routing method based on the new interconnect network, by labeling the routers in the system, and using label minus hops first or label plus hops algorithm, deadlock-free adaptive routing can be achieved without using any virtual channel, and hops between any two routers can be ensured no more than five steps. The invention also proposes a fully adaptive routing scheme. With the deadlock free characteristic of partially adaptive routing algorithm of the invention, in combination with a new flow control mechanism, deadlock free full adaptive routing can be achieved.

Deadlock-free adaptive routing of interconnect network

Field of the Invention

The present invention relates to distributed computing, and more particularly, to an interconnect network and adaptive routing algorithm.

Background of the Invention

An interconnection network is used to implement synchronizing and communicating among different process nodes, and is an important component in multiprocessor systems for connecting the processors, memory modules and I/O devices. The interconnection network is a critical aspect that may significantly impact the overall performance and scalability of multi-processor systems. In view of the prior art, it is a key requirement for interconnection networks to provide low latency, high efficiency communication among process nodes.

Development of microprocessor technology has made computing capacity of a single processing node increasing rapidly, which also brings about higher requirements for the performance of the interconnection network. In fact, bandwidth increase of the interconnection network has a 30% disparity from performance increase of microprocessors. Therefore, the latency and bandwidth of interconnection network have become the bottleneck of performance improvement of multiprocessor systems.

As the development of semiconductor industry and the improvement of electrical circuits, serial channel becomes an effective selection of signal transmission. High-speed serial channel may highly increase the bandwidth of a single pin, and thus the number of pin of a router chip may be cut down. In the early 1990s, the bandwidth of pins of a router is limited within 10Gbps, which reaches to 10-20Tbps in 21st century. These improvements enable the implementing of high-radix routers.

The use of high-radix routers is becoming a tendency of interconnection networks. The interconnect networks constructed by high-radix routers may connect tens of thousands of processors with several hops, and usually have shorter network diameter as well as smaller message latency

and lower cost, and therefore can realize high efficiency communication among process nodes and improve the system performance.

In 2006, the Black Widow system from the Cray Corporation firstly employs radix-64 router chips and provides interconnection among 32000 processors with Clos topology, which guarantees that any two nodes can communicate within 7 hops. In 2012, the Cascade system from the Cray Corporation utilizes radix-48 router chips and provides interconnection among 370216 processors with dragonfly architectures, which guarantees that any two nodes can communicate within 5 hops. However, Cray Cascade requires 4 virtual channels to provide deadlock-free adaptive routing.

According to the existing chip technology and high-radix router technology, how to design a highly efficient network interconnection topology and efficient routing algorithms has become a new urgent research topic. Blackwidow Clos network is the first network using high-radix routers. But it has a high cost and a long network diameter, which has a bad impact on the system performance. Cascade Dragonfly network also uses high-radix routers, with high scalability and low cost features, and has a short network diameter. But it requires 4 virtual channels to avoid deadlocks, which increases the router design complexity and message latency, and limits the system performance.

Summary of the Invention

Technical problems to be solved

Technical problems solved by the present invention are: providing an interconnection network based on router label and its corresponding adaptive routing algorithm. The interconnection network and adaptive routing algorithm not only maintain low cost, high scalability, and short network diameter feature of the existing dragonfly network, but require no virtual channels to achieve fully deadlock-free adaptive routing algorithm.

The technical solutions

To solve the technical problem above, the present invention provides an

interconnect network, wherein the routers in the system are divided into n groups of routers, each group of routers contains one or more routers, with achieving full connectivity among the n group of routers, so that the communication between any two groups of routers only requires one hop, wherein n is an integer greater than or equal to 1; the connection algorithm among router groups complies the following principles:

P1: label the routers in the interconnection network, wherein n groups of routers are sequentially labeled $G_0, G_1, \dots, G_{n-1}$, and m routers in the group are sequentially labeled $R_0, R_1, \dots, R_{m-1}$;

P2: for i from 0 to $n-2$, connect the router R_{m-1} in group G_i to the router R_0 in group G_{i+1} ;

P3: for i from 0 to $n-3$, for j from $i+2$ to $n-1$, connect the router R_v in group G_i with the router $R_{v'}$ in group G_j , wherein v is the highest available slot in group G_i and v' is the lowest available slot in group G_j for the global connection.

Preferably, each group of routers comprises m routers, wherein m is an integer greater than or equal to one.

Preferably, each router in the network comprises a local ports, p processor ports and g global ports; wherein the a local ports are to achieve the interconnection among the routers within a router group, the corresponding channels are local interconnected channels; wherein the p processors ports interconnects processor nodes with routers; wherein the g global ports interconnects among the groups of routers, and the corresponding interconnected channels are global channels.

Preferably, each group of routers includes $m=a+1$ routers, the number of groups of routers is $n = mg + 1$, which means that each group contains $k' = m(p + g)$ ports.

In another aspect of the invention, a partially adaptive routing algorithm of interconnect network is provided, which comprises the steps of:

S1: label the interconnection network routers, wherein n groups of routers are sequentially labeled $G_0, G_1, \dots, G_{n-1}$, and m routers in the

groups are sequentially labeled $R_0, R_1, \dots, R_{m-1}$; interconnect the highest labeled router $G_s R_{m-1}$ in any group of routers G_s with the lowest labeled router $G_t R_0$ in group of routers G_t during labeling, wherein $t = (s+1) \bmod n$, i.e., t is the remainder of dividing $s+1$ by n ; interconnect the highest labeled router $G_u R_{m-1}$ in group of routers G_u with the lowest labeled router $G_s R_0$ in group of routers G_s during labeling, wherein $u = (s-1) \bmod n$, i.e., u is the remainder of dividing $s-1$ by n ;

S2: according to the label of routers at each end of the channel, a unidirectional channel can be classified as minus and plus directions;

S3: when routing a message packet, all the paths should comply with the principle of minus-first routing that all routing paths from the source router to the destination router comprise w minus direction paths followed by v plus direction paths; or all the paths should comply with the principle of plus-first routing that all routing paths from the source router to the destination router comprise w plus direction paths followed by v minus direction paths.

Preferably, the partially adaptive routing algorithm, when classified channels to label minus direction or label plus direction in step S2, should comply with the following principles:

For any channel $(G_{i1}R_{j1}, G_{x1}R_{y1})$,

If $i1 > x1$, the channel is defined as label minus direction;

If $i1 < x1$, the channel is defined as label plus direction;

If $i1 = x1$ and $j1 > y1$, the channel is defined as label minus direction;

If $i1 = x1$ and $j1 < y1$, the channel is defined as label plus direction.

Preferably, in the partially adaptive routing algorithm complies with the principle of label minus direction-first routing when routing message package in step S3.

Preferably, step S2 comprises the following steps:

S31: When routing a message packet, set the router of the current node as G_iR_j , and the router of destination node as G_xR_y , Ch is an available set of message packet output channels, which has an initial value of empty set $\emptyset$, compare the group label i of the router of the current node with the group label x of the router of the destination node, if $i = x$, go to step S32 if $i < x$, go to step S33, otherwise go to step S34;

S32: check whether the router of the current node is the router of the destination node, if the router of the current node is the router of the destination node, i.e., $i = x$ and $j = y$, the message packet has reached the destination node by the router, and the algorithm ends; otherwise, if $i = x$ and $j \neq y$, providing local channel (G_iR_j, G_xR_y) so that the message packet can be delivered to the router of destination node, $Ch = Ch + \{(G_iR_j, G_xR_y)\}$, return Ch;

S33: assume router G_iR_a in the group of routers of the current node and router G_xR_b in the group of routers of the destination node are connected;

If $b \geq y$ and $a \neq j$, then provide routing channel from G_iR_j to G_iR_a , $Ch = Ch + \{(G_iR_j, G_iR_a)\}$;

If $b < y$, set router G_iR_c in the group of routers of the current node connecting with the router G_mR_d and the router G_mR_e connecting with the router G_xR_f ;

For $m < i$, only when $j \leq c$ and $f \geq y$, misroute the packet to G_m ;

For $i < m < x$, only when $j \leq c$ and $d \leq e$, misroute the packet to G_m ;

If $c \neq j$, provide router channel from G_iR_j to G_iR_c , $Ch = Ch + \{(G_iR_j, G_iR_c)\}$;

If $c = j$, provide router channel from G_iR_j to G_mR_d , $Ch = Ch + \{(G_iR_j, G_mR_d)\}$, return Ch;

S34: assume router G_iR_a in the group of routers of the current node

and router G_xR_b in the group of routers of the destination node are connected;

If $a < j$, provide routing channel from G_iR_j to G_iR_a , $Ch = Ch + \{(G_iR_j, G_iR_a)\}$;

If $a = j$, provide routing channel from G_iR_j to G_xR_b , $Ch = Ch + \{(G_iR_j, G_xR_b)\}$;

If $a > j$, set router G_iR_c in the group of routers of the current node connecting with router G_mR_d and router G_mR_e connecting with router G_xR_f ;

for $m < x$, only when $j \leq c$ and $f \geq y$, misroute the packet to G_m ;

for $x < m < i$, only when $j \leq c$ and $d \leq e$, misroute the packet to G_m ;

If $c \neq j$, provide router channel from G_iR_j to G_iR_c , $Ch = Ch + \{(G_iR_j, G_iR_c)\}$;

If $c = j$, provide router channel from G_iR_j to G_mR_d , $Ch = Ch + \{(G_iR_j, G_mR_d)\}$, return Ch .

In another aspect of the invention, a fully adaptive routing algorithm on interconnect network is provided, which comprises the steps of:

D1: according to the channel type used by the next hop of message packet, classify the message package into safe package or unsafe package; if routing channel used by the message packet to the next hop complies with the principle of minus direction-first routing, the packet is a safe package for the router of the next hop; otherwise, it is an unsafe package;

D2: set channel buffer with a size of *buff* packets, which can store *buff* message packets at most, wherein *buff* is greater than or equal to 2;

D3: set the router of the current node as G_iR_j , and the router of the destination node as G_xR_y , $Ch1$ is a set including all the routing channels for message packets which comply the principle of minus direction-first; if $i = x$ and $j = y$, the packet has reached the destination node, so the routing process ends; otherwise, invoke the partially

adaptive routing algorithm of interconnect network stated above to provide set of channels Ch1 which comply the principle of minus direction-first for message packets;

D4: set the router of the current node as G_iR_j , and the router of the destination node as G_xR_y , Ch2 is a set including all the routing channels for message packets which comply minimal routing; if $i = x$ and $j = y$, the message packet has arrived at the destination node, so the routing process ends; if $i = x$ and $j \neq y$, $Ch2 = \{(G_iR_j, G_xR_y)\}$; if $i \neq x$, set router G_xR_b in the current group of routers and the router G_iR_a in the destination group are connected; if $a \neq j$, provide routing channel from router G_iR_j to router G_iR_a , $Ch2 = \{(G_iR_j, G_iR_a)\}$, if $a = j$, provide routing channel from router G_iR_j to router G_xR_b , $Ch2 = \{(G_iR_j, G_xR_b)\}$;

D5: for the set of channels Ch2 in which all the routing channels comply the minimal routing path, the channels should be considered separately. Assume the channel by which the message packages arrives the next router node according to the minimal routing path provided by Ch2 has f empty buffers, and s buffer safe packages, the message packets comply the following flow control mechanisms:

if $f > 1$, the message packet can be delivered to the next node with the channel included in Ch2;

if $f = 1$ and $s \geq 1$, the message packet can be delivered to the next node with the channel included in Ch2;

if $f = 1$ and $s = 0$, the message packet is the safe package for the channel to the next hop, which is included in Ch2, and the package can be delivered to the next node with the channel included in Ch2;

in other cases, the packet cannot be delivered via channels included in Ch2, and then empty Ch2, i.e. $Ch2 = \emptyset$;

D6: according to the routing channel set Ch1 provided by step D3 and the routing channel set Ch2 provided by step D5, get all the routing channels that provided by fully adaptive routing algorithm Ch3, Ch3

= Ch1 + Ch2, wherein packets will be routed by Claim5 if they do not comply the flow control principle; or let Ch3 = Ch2 and route them with the minimal routing path after they are converted into complying with the flow control in the next routing, return Ch3.

Preferably, message packages are divided into safe packages and unsafe packages in step D1. Unsafe packets which do not comply with the flow control principle can be routed with the partial algorithm, and then the unsafe packets can be converted into safe packets. Otherwise, the unsafe packets should wait route process until they comply with the flow control principle so that they can route according to minimal path algorithm.

Beneficial effects

The present invention provides a novel interconnection network; the interconnection network not only keeps the low cost, high scalability, short network diameter of the original dragonfly network, but also is conducive to implementing the adaptive routing algorithm proposed by the present invention. The present invention proposes an adaptive routing algorithm based on the novel interconnection network, through labeling the interconnection network routers, and using specific label minus direction first or label plus direction first principle, this algorithm does not require virtual channel to achieving deadlock-free adaptive routing, and guarantees that communication steps between any processor nodes no more than five hops. Because the interconnection network and its adaptive routing algorithm can achieve deadlock-free without using virtual channel, which avoids the virtual channel distribution mechanism, the interconnection network and routing algorithm of the present invention can greatly simplify router design, reduce message latency and improve network performance. The invention further proposes fully adaptive routing method, with using of free deadlock characteristics of the part adaptive routing method of the invention, supplemented by a new flow control mechanism, fully adaptive deadlock-free routing can be obtained. The fully adaptive routing algorithm allows all messages reach the destination node router through the minimal path that effectively improve channel utilization, and network throughput. On the other hand, the flow

control mechanism can apply to all virtual channels switching network, which can implement fully adaptive routing without virtual channel.

Brief Description of the Drawings

FIG.1 is a block diagram of an example embodiment of the interconnection network.

FIG.2 is a block diagram of an example embodiment of connection of the interconnection network.

FIG.3 is a flow diagram of an example embodiment of the step S3 in the partially adaptive routing algorithm.

FIG.4 is a schematic diagram of an example embodiment of hopping in partially adaptive deadlock free routing algorithm.

FIG.5 is a schematic diagram illustrating the process of the partially deadlock free adaptive routing algorithm.

FIG.6 (a) is a block diagram illustrating a routing algorithm which generate deadlock.

FIG.6 (b) is a block diagram illustrating a deadlock caused by cyclic dependency.

FIG.6 (c) is a block diagram illustrating eliminating deadlock according to the label minus first principle.

FIG.7 is a flow diagram of an example embodiment of the partially adaptive routing algorithm.

FIG.8 is a flow diagram of an example embodiment of the fully adaptive routing algorithm.

FIG.9 is a block diagram illustrating the router label of the partially adaptive routing algorithm.

Detailed Description of Preferred Embodiments

In the following detailed description of example embodiments of the invention, reference is made to specific examples by way of drawings and illustrations.

As shown in Figure 1, the invention provides an interconnect network system, wherein the routers in the system are divided into n groups of routers, with achieving full connectivity among the n groups of routers, so that the communication between any two groups of routers only requires one hop, wherein n is an integer greater than or equal to 1; each router group contains one or more routers, with full connecting among the routers in the group, so that the distance between any two routers in the group is only one hop; the following are steps of the connection algorithm among groups of routers :

P1: label the interconnection network routers, wherein n groups of routers are sequentially labeled $G_0, G_1, \dots, G_{n-1}$, m routers in a group are sequentially labeled $R_0, R_1, \dots, R_{m-1}$;

P2: for i from 0 to $n-2$, connect the router R_{m-1} in router group G_i to router R_0 in group G_{i+1} ;

P3: for i from 0 to $n-3$, and j from $i+2$ to $n-1$, connect R_v within G_i with $R_{v'}$ within G_j , wherein v is the highest available slot in G_i and v' is the lowest available slot in G_j .

In order to unify the interconnect network and facilitate the algorithm, each group of routers may be set including m routers, wherein m is greater than or equal to one.

The routers in the interconnect network may be k -radix routers; and each k -radix router within the system comprises a local ports, p processor ports and g global ports, a , p and g are integer greater than or equal to 1; wherein a local ports are used to achieve the interconnection among the routes within a group of routes, and the corresponding channels are local interconnected channels; wherein p processors ports are to used to achieve interconnection between the processor nodes and the route nodes, the corresponding channels are processor interconnected channels; wherein g global ports are to used to achieve interconnection among the groups, the corresponding channels are global interconnected channels, and $k = p + a + g$. Because m routers compose a group of routers, for implementing fully connection of routers within a router group, the number of local ports $a = m$

– 1; for implementing fully connection of groups of routers, the number of global ports $g = (n - 1) / m$; while the number of process nodes is nmp , and each router group acts like a virtual router with $k' = m(p + g)$ ports.

FIG2 show an example embodiment of the interconnect network proposed by the present invention, in this interconnect network, $m = 4$, $p = 2$, $g = 2$, $n = 9$. As shown in FIG2, routers within a group of routers achieve fully connection, ensuring that distance of any two routers within a group is one hop. Groups of router also achieve fully connection, ensuring that there must be a global channel between any two groups of routers.

Based on the interconnect network proposed by the present invention, as shown in FIG7, the present invention proposes a partially adaptive routing algorithm of interconnect network, comprising the steps of:

S1: label the interconnection network routers, wherein n group of routers are sequentially labeled $G_0, G_1, \dots, G_{n-1}$, m routers in the group are sequentially labeled $R_0, R_1, \dots, R_{m-1}$. Thereby, each router can be determined with its group label and intra-group label.

Meanwhile, interconnect the highest labeled router $G_s R_{m-1}$ in any group of routers G_s with the lowest labeled router $G_t R_0$ in group of routers G_t during labeling, wherein $t = (s+1) \bmod n$, i.e., t is the remainder of dividing $s+1$ by n ; interconnect the highest labeled router $G_u R_{m-1}$ in group of routers G_u with the lowest labeled router $G_s R_0$ in group of routers G_s during labeling, wherein $u = (s-1) \bmod n$, i.e., u is the remainder of dividing $s-1$ by n ;

S2: according to the labels of routers at two ends of the channel, a unidirectional channel can be classified as minus direction and plus direction;

Since each channel connects two routers, for a channel with a leading end router $G_{i1} R_{j1}$ and a rear end router $G_{x1} R_{y1}$, the following principles should be followed when classifying a channel as a minus direction or plus direction channel:

for any channel $(G_{i1} R_{j1}, G_{x1} R_{y1})$,

if $i_1 > x_1$, the channel is defined as minus direction;

if $i_1 < x_1$, the channel is defined as plus direction;

if $i_1 = x_1$ and $j_1 > y_1$, the channel is defined as minus direction;

if $i_1 = x_1$ and $j_1 < y_1$, the channel is defined as plus direction.

S3: when routing a message packet, all the paths should comply with the principle of minus-first routing that all routing paths from the source node to the destination node comprise w minus direction paths followed by v plus direction paths; or all the paths should comply with the principle of plus-first routing that all routing paths from the source router to the destination router comprise w plus direction paths followed by v minus direction paths.. Therein, v and w are integers greater than or equal to 1.

For convenience aspect of showing the present invention, the message packet comply the label minus first direction when forward message package in step 3 of the embodiment of the algorithm invention.

As shown in FIG3, an example embodiment of the partially adaptive routing which comply the label minus first direction should comprise the following steps:

S31: When routing a message packet, set the router of the current node as G_iR_j , and the router of destination node as G_xR_y , Ch is an available set of message packet output channels, which has an initial value of empty set $\emptyset$, compare the group label i of the router of the current node with the group label x of the router of the destination node, if $i = x$, go to step S32 if $i < x$, go to step S33, otherwise go to step S34;

S32: check whether the router of the current node is the router of the destination node, if the router of the current node is the router of the destination node, i.e., $i = x$ and $j = y$, the message packet has reached the destination node by the router, and the algorithm ends; otherwise, if $i = x$ and $j \neq y$, providing local channel (G_iR_j, G_xR_y) so that the message packet can be delivered to the router of destination node, Ch

$= Ch + \{(G_iR_j, G_xR_y)\}$, return Ch;

S33: assume router G_iR_a in the group of routers of the current node and router G_xR_b in the group of routers of the destination node are connected;

If $b \geq y$ and $a \neq j$, then provide routing channel from G_iR_j to G_iR_a , $Ch = Ch + \{(G_iR_j, G_iR_a)\}$;

If $b < y$, set router G_iR_c in the group of routers of the current node connecting with the router G_mR_d and the router G_mR_e connecting with the router G_xR_f ;

For $m < i$, only when $j \leq c$ and $f \geq y$, misroute the packet to G_m ;

For $i < m < x$, only when $j \leq c$ and $d \leq e$, misroute the packet to G_m ;

If $c \neq j$, provide router channel from G_iR_j to G_iR_c , $Ch = Ch + \{(G_iR_j, G_iR_c)\}$;

If $c = j$, provide router channel from G_iR_j to G_mR_d , $Ch = Ch + \{(G_iR_j, G_mR_d)\}$, return Ch;

S34: assume router G_iR_a in the group of routers of the current node and router G_xR_b in the group of routers of the destination node are connected;

If $a < j$, provide routing channel from G_iR_j to G_iR_a , $Ch = Ch + \{(G_iR_j, G_iR_a)\}$;

If $a = j$, provide routing channel from G_iR_j to G_xR_b , $Ch = Ch + \{(G_iR_j, G_xR_b)\}$;

If $a > j$, set router G_iR_c in the group of routers of the current node connecting with router G_mR_d and router G_mR_e connecting with router G_xR_f ;

for $m < x$, only when $j \leq c$ and $f \geq y$, misroute the packet to G_m ;

for $x < m < i$, only when $j \leq c$ and $d \leq e$, misroute the packet to G_m ;

If $c \neq j$, provide router channel from G_iR_j to G_iR_c , $Ch = Ch + \{(G_iR_j, G_iR_c)\}$;

If $c = j$, provide router channel from G_iR_j to G_mR_d , $Ch = Ch + \{(G_iR_j, G_mR_d)\}$, return Ch .

As shown in FIG 8, based on the partially routing algorithm above proposed by the present invention, the present invention proposed a fully adaptive routing algorithm of interconnect network, comprising the following steps:

D1: according to the channel type used by the next hop of message packet, classify the message package into safe package or unsafe package according to the following principles:

if routing channel used by the message packet to the next hop complies with the principle of minus direction-first routing, the packet is a safe package for the router of the next hop; otherwise, it is an unsafe package;

D2: set channel buffer with a size of *buff* packets, which can store *buff* message packets at most, wherein *buff* is greater than or equal to 2;

D3: set the router of the current node as G_iR_j , and the router of the destination node as G_xR_y , $Ch1$ is a set including all the routing channels for message packets which comply the principle of minus direction-first; if $i = x$ and $j = y$, the packet has reached the destination node, so the routing process ends; otherwise, invoke the partially adaptive routing algorithm of interconnect network stated above to provide set of channels $Ch1$ which comply the principle of minus direction-first for message packets;

D4: set the router of the current node as G_iR_j , and the router of the destination node as G_xR_y , $Ch2$ is a set including all the routing channels for message packets which comply minimal routing; if $i = x$ and $j = y$, the message packet has arrived at the destination node, so the routing process ends; if $i = x$ and $j \neq y$, $Ch2 = \{(G_iR_j, G_xR_y)\}$; if $i \neq x$, set router G_xR_b in the current group of routers and the router G_iR_a in the destination group are connected; if $a \neq j$, provide routing channel from router G_iR_j to router G_iR_a , $Ch2 = \{(G_iR_j, G_iR_a)\}$, if $a = j$, provide routing channel from router G_iR_j to router G_xR_b , $Ch2 = \{(G_iR_j, G_xR_b)\}$.

$G_x R_b\} \}$;

D5: for the set of channels Ch2 in which all the routing channels comply the minimal routing path, the channels should be considered separately. Assume the channel by which the message packages arrives the next router node according to the minimal routing path provided by Ch2 has f empty buffers, and s buffer safe packages, the message packets comply the following flow control mechanisms:

if $f > 1$, the message packet can be delivered to the next node with the channel included in Ch2;

if $f = 1$ and $s \geq 1$, the message packet can be delivered to the next node with the channel included in Ch2;

if $f = 1$ and $s = 0$, the message packet is the safe package for the channel to the next hop, which is included in Ch2, and the package can be delivered to the next node with the channel included in Ch2;

in other cases, the packet cannot be delivered via channels included in Ch2, and then empty Ch2, i.e. $Ch2 = \emptyset$.

With the process above, message packages will not be deadlocked with the channel provided by Ch2.

D6: according to the routing channel set Ch1 provided by step D3 and the routing channel set Ch2 provided by step D5, get all the routing channels that provided by fully adaptive routing algorithm Ch3, $Ch3 = Ch1 + Ch2$, wherein packets will be routed by the partially adaptive routing algorithm if they do not comply the flow control principle; or let $Ch3 = Ch2$ and route them with the minimal routing path after they are converted into complying with the flow control in the next routing, return Ch3.

Figure 4 shows two example embodiments of the partially routing algorithm of the present application, which respectively are (source router node $G_0 R_1$, destination router node $G_3 R_1$) and (source router node $G_3 R_1$, destination router node $G_0 R_1$), based on the label minus direction first partially adaptive deadlock free routing algorithm, providing the following

routing:

For (source node G_0R_1 , the destination node G_3R_1), the routing path is:
 $(G_0R_1, G_2R_1) \rightarrow (G_2R_1, G_2R_2) \rightarrow (G_2R_2, G_3R_0) \rightarrow (G_3R_0, G_3R_1)$;

For (source node G_3R_1 , the destination node G_0R_1), the routing path is:
 $(G_3R_1, G_3R_0) \rightarrow (G_3R_0, G_2R_2) \rightarrow (G_2R_2, G_2R_1) \rightarrow (G_2R_1, G_0R_1)$.

Figure 5 shows the label minus first direction always ensures that message packet can arrive at the destination router node in five hops for any two router nodes according to the interconnection network system and the partially adaptive routing algorithm of the present invention. In figure 5, G_iR_j is source node router, G_xR_y is destination node router. Based on the label minus first direction partially adaptive deadlock free routing algorithm, it provides the following routing path:

For $i < x$, the deadlock free partially routing algorithm always provides routing path: $(G_iR_j, G_iR_c) \rightarrow (G_iR_c, G_{x-1}R_d) \rightarrow (G_{x-1}R_d, G_{x-1}R_{max}) \rightarrow (G_{x-1}R_{max}, G_xR_0) \rightarrow (G_xR_0, G_xR_y)$;

For $i > x$, the deadlock free partially routing algorithm always provides routing path: $(G_iR_j, G_iR_0) \rightarrow (G_iR_0, G_{i-1}R_{max}) \rightarrow (G_{i-1}R_{max}, G_{i-1}R_d) \rightarrow (G_{i-1}R_d, G_xR_c) \rightarrow (G_xR_c, G_xR_y)$.

The max subscript means the *max* label router within a router group.

The reason that the partially routing algorithm proposed by the present invention can avoid deadlock is: consider the four necessary conditions for deadlock: resource mutex, request and hold, non-preemptive, and circular wait. Therefore, using the label minus first direction, by eliminating the circular wait, deadlock can be avoided. For example, a routing algorithm in figure 6 (a) generates a deadlock is showed in figure 6 (b). After restricting some waiting paths, the circular dependency is eliminated and deadlock is avoided as shown in figure 6 (c). Therefore, the embodiment of the partially adaptive routing algorithm proposed by the present invention can implement an adaptive routing in interconnection network without using virtual channels.

Although specific embodiments have been illustrated and described herein,

it will be appreciated by those of ordinary skill in the art that any arrangement which is calculated to achieve the same purpose may be substituted for the specific embodiments shown. This application is intended to cover any adaptations or variations of the example embodiments of the invention described herein. It is intended that this invention be limited only by the claims, and the full scope of equivalents thereof.

What is claimed is:

1. An interconnect network system, wherein routers in the system are divided into n groups of routers with full connecting among the n router groups, so that communication between any two groups of routers only requires one hop, wherein n is an integer greater than or equal to 1; each group of routers contains one or more routers with full connecting among the routers in the group, so that a distance between any two routers in the group is only one hop; a connection algorithm among the groups of routers complies the following principles:

P1: label the routers in the interconnection network, wherein n groups of routers are sequentially labeled $G_0, G_1, \dots, G_{n-1}$, and m routers in the group are sequentially labeled $R_0, R_1, \dots, R_{m-1}$;

P2: for i from 0 to $n-2$, connect the router R_{m-1} in group G_i to the router R_0 in group G_{i+1} ;

P3: for i from 0 to $n-3$, for j from $i+2$ to $n-1$, connect the router R_v in group G_i with the router $R_{v'}$ in group G_j , wherein v is the highest available slot in group G_i and v' is the lowest available slot in group G_j for the global connection. 2. The interconnect network of Claim 1, wherein each group of routers including m routers, wherein m is greater than or equal to one.

3. The interconnect network of Claim 1 or 2, each router in the network comprises a local ports, p processor ports and g global ports; wherein the a local ports are to achieve the interconnection among the routers within a router group, the corresponding channels are local interconnected channels; wherein the p processors ports interconnects processor nodes with routers; wherein the g global ports interconnects among the groups of routers, and the corresponding interconnected channels are global channels.

4. The interconnect network of Claim 3, wherein each group of routers includes $m=a+1$ routers, the number of groups of routers is $n = mg + 1$, which means that each group contains $k' = m(p + g)$ ports.

5. The partially adaptive routing algorithm for interconnect network, comprising the steps of:

S1: label the interconnection network routers, wherein n groups of routers are sequentially labeled $G_0, G_1, \dots, G_{n-1}$, and m routers in the groups are sequentially labeled $R_0, R_1, \dots, R_{m-1}$; interconnect the highest labeled router $G_s R_{m-1}$ in any group of routers G_s with the lowest labeled router $G_t R_0$ in group of routers G_t during labeling, wherein $t = (s+1) \bmod n$, i.e., t is the remainder of dividing $s+1$ by n ; interconnect the highest labeled router $G_u R_{m-1}$ in group of routers G_u with the lowest labeled router $G_s R_0$ in group of routers G_s during labeling, wherein $u = (s-1) \bmod n$, i.e., u is the remainder of dividing $s-1$ by n ;

S2: according to the label of routers at each end of the channel, a unidirectional channel can be classified as minus and plus directions;

S3: when routing a message packet, all the paths should comply with the principle of minus-first routing that all routing paths from the source router to the destination router comprise w minus direction paths followed by v plus direction paths; or all the paths should comply with the principle of plus-first routing that all routing paths from the source router to the destination router comprise w plus direction paths followed by v minus direction paths.

6. The partially adaptive routing algorithm for interconnect network of Claim 5, wherein when classified channels as minus direction and plus direction in the step S2, the following principles should be complied with:

For any channel $(G_{i1}R_{j1}, G_{x1}R_{y1})$,

If $i1 > x1$, the channel is defined as label minus direction;

If $i1 < x1$, the channel is defined as label plus direction;

If $i1 = x1$ and $j1 > y1$, the channel is defined as label minus direction;

If $i1 = x1$ and $j1 < y1$, the channel is defined as label plus direction.

7. The partially adaptive routing algorithm for interconnect network of

Claim 5, wherein the step S3 complies with minus direction-first principle when routing a packet.

8. The partially adaptive routing algorithm for interconnect network of Claim 7, wherein the step S3 comprises the following steps:

S31: When routing a message packet, set the router of the current node as G_iR_j , and the router of destination node as G_xR_y , Ch is an available set of message packet output channels, which has an initial value of empty set $\emptyset$, compare the group label i of the router of the current node with the group label x of the router of the destination node, if $i = x$, go to step S32 if $i < x$, go to step S33, otherwise go to step S34;

S32: check whether the router of the current node is the router of the destination node, if the router of the current node is the router of the destination node, i.e., $i = x$ and $j = y$, the message packet has reached the destination node by the router, and the algorithm ends; otherwise, if $i = x$ and $j \neq y$, providing local channel (G_iR_j, G_xR_y) so that the message packet can be delivered to the router of destination node, $Ch = Ch + \{(G_iR_j, G_xR_y)\}$, return Ch;

S33: assume router G_iR_a in the group of routers of the current node and router G_xR_b in the group of routers of the destination node are connected;

If $b \geq y$ and $a \neq j$, then provide routing channel from G_iR_j to G_iR_a , $Ch = Ch + \{(G_iR_j, G_iR_a)\}$;

If $b < y$, set router G_iR_c in the group of routers of the current node connecting with the router G_mR_d and the router G_mR_e connecting with the router G_xR_f ;

For $m < i$, only when $j \leq c$ and $f \geq y$, misroute the packet to G_m ;

For $i < m < x$, only when $j \leq c$ and $d \leq e$, misroute the packet to G_m ;

If $c \neq j$, provide router channel from G_iR_j to G_iR_c , $Ch = Ch + \{(G_iR_j, G_iR_c)\}$;

If $c = j$, provide router channel from G_iR_j to G_mR_d , $Ch = Ch + \{(G_iR_j, G_mR_d)\}$, return Ch ;

S34: assume router G_iR_a in the group of routers of the current node and router G_xR_b in the group of routers of the destination node are connected;

If $a < j$, provide routing channel from G_iR_j to G_iR_a , $Ch = Ch + \{(G_iR_j, G_iR_a)\}$;

If $a = j$, provide routing channel from G_iR_j to G_xR_b , $Ch = Ch + \{(G_iR_j, G_xR_b)\}$;

If $a > j$, set router G_iR_c in the group of routers of the current node connecting with router G_mR_d and router G_mR_c connecting with router G_xR_b ;

for $m < x$, only when $j \leq c$ and $f \geq y$, misroute the packet to G_m ;

for $x < m < i$, only when $j \leq c$ and $d \leq e$, misroute the packet to G_m ;

If $c \neq j$, provide router channel from G_iR_j to G_iR_c , $Ch = Ch + \{(G_iR_j, G_iR_c)\}$;

If $c = j$, provide router channel from G_iR_j to G_mR_d , $Ch = Ch + \{(G_iR_j, G_mR_d)\}$, return Ch . 9. The fully adaptive routing algorithm for the interconnect network, comprising the steps of:

D1: according to the channel type used by the next hop of message packet, classify the message package into safe package or unsafe package; if routing channel used by the message packet to the next hop complies with the principle of minus direction-first routing, the packet is a safe package for the router of the next hop; otherwise, it is an unsafe package;

D2: set channel buffer with a size of *buff* packets, which can store *buff* message packets at most, wherein *buff* is greater than or equal to 2;

D3: set the router of the current node as G_iR_j , and the router of the destination node as G_xR_y , $Ch1$ is a set including all the routing channels for message packets which comply the principle of minus

direction-first; if $i = x$ and $j = y$, the packet has reached the destination node, so the routing process ends; otherwise, invoke the partially adaptive routing algorithm of interconnect network stated above to provide set of channels Ch1 which comply the principle of minus direction-first for message packets;

D4: set the router of the current node as G_iR_j , and the router of the destination node as G_xR_y , Ch2 is a set including all the routing channels for message packets which comply minimal routing; if $i = x$ and $j = y$, the message packet has arrived at the destination node, so the routing process ends; if $i = x$ and $j \neq y$, $\text{Ch2} = \{(G_iR_j, G_xR_y)\}$; if $i \neq x$, set router G_xR_b in the current group of routers and the router G_iR_a in the destination group are connected; if $a \neq j$, provide routing channel from router G_iR_j to router G_iR_a , $\text{Ch2} = \{(G_iR_j, G_iR_a)\}$, if $a = j$, provide routing channel from router G_iR_j to router G_xR_b , $\text{Ch2} = \{(G_iR_j, G_xR_b)\}$;

D5: for the set of channels Ch2 in which all the routing channels comply the minimal routing path, the channels should be considered separately. Assume the channel by which the message packages arrives the next router node according to the minimal routing path provided by Ch2 has f empty buffers, and s buffer safe packages, the message packets comply the following flow control mechanisms:

if $f > 1$, the message packet can be delivered to the next node with the channel included in Ch2;

if $f = 1$ and $s \geq 1$, the message packet can be delivered to the next node with the channel included in Ch2;

if $f = 1$ and $s = 0$, the message packet is the safe package for the channel to the next hop, which is included in Ch2, and the package can be delivered to the next node with the channel included in Ch2;

in other cases, the packet cannot be delivered via channels included in Ch2, and then empty Ch2, i.e. $\text{Ch2} = \emptyset$;

D6: according to the routing channel set Ch1 provided by step D3 and

the routing channel set Ch2 provided by step D5, get all the routing channels that provided by fully adaptive routing algorithm Ch3, $Ch3 = Ch1 + Ch2$, wherein packets will be routed by Claim5 if they do not comply the flow control principle; or let $Ch3 = Ch2$ and route them with the minimal routing path after they are converted into complying with the flow control in the next routing, return Ch3.

10: The fully adaptive routing algorithm of Claim 10, wherein message packages are divided into safe packages and unsafe packages in step D1. Unsafe packets which do not comply with the flow control principle can be routed with the partial algorithm, and then the unsafe packets can be converted into safe packets. Otherwise, the unsafe packets should wait route process until they comply with the flow control principle so that they can route according to minimal path algorithm.

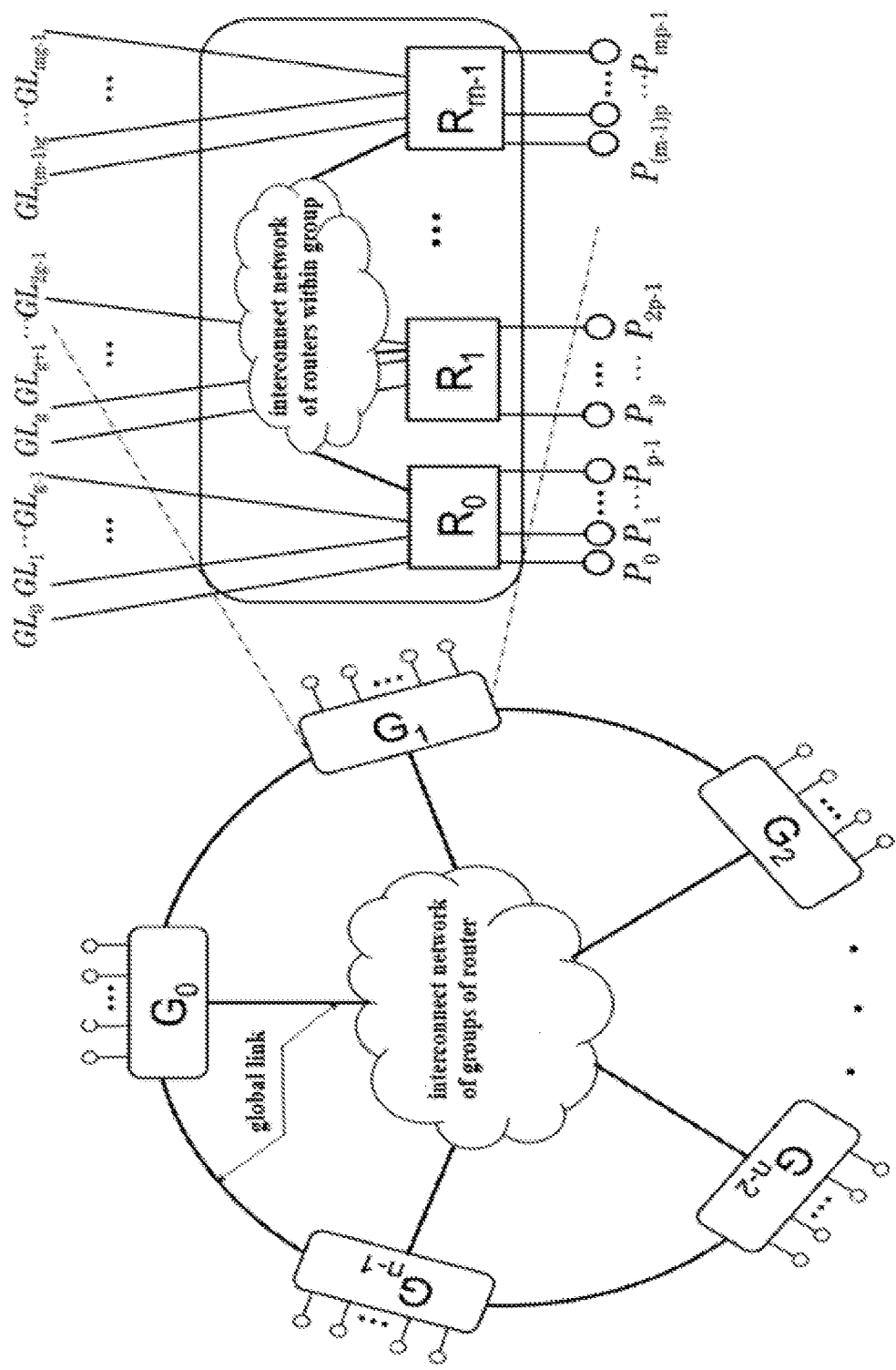


FIG. 1

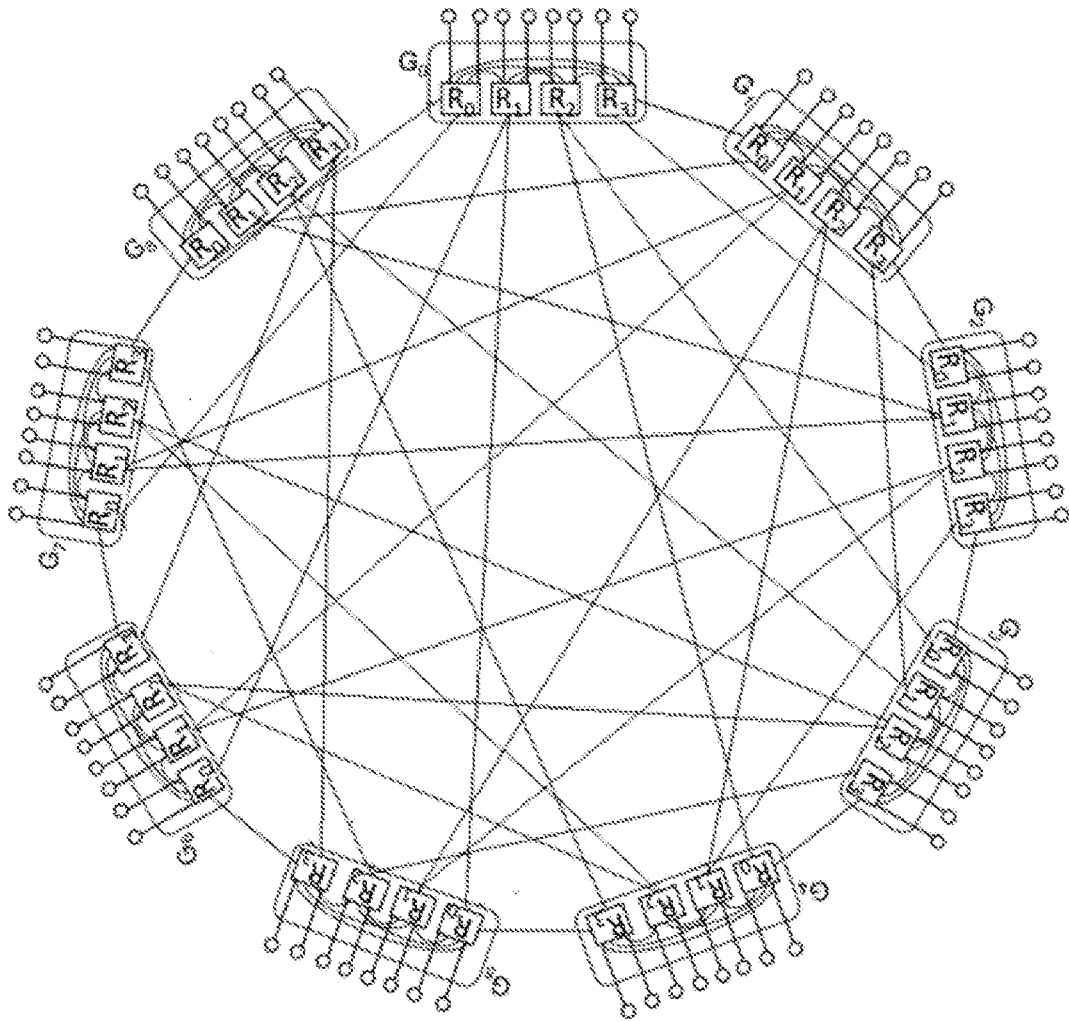


FIG. 2

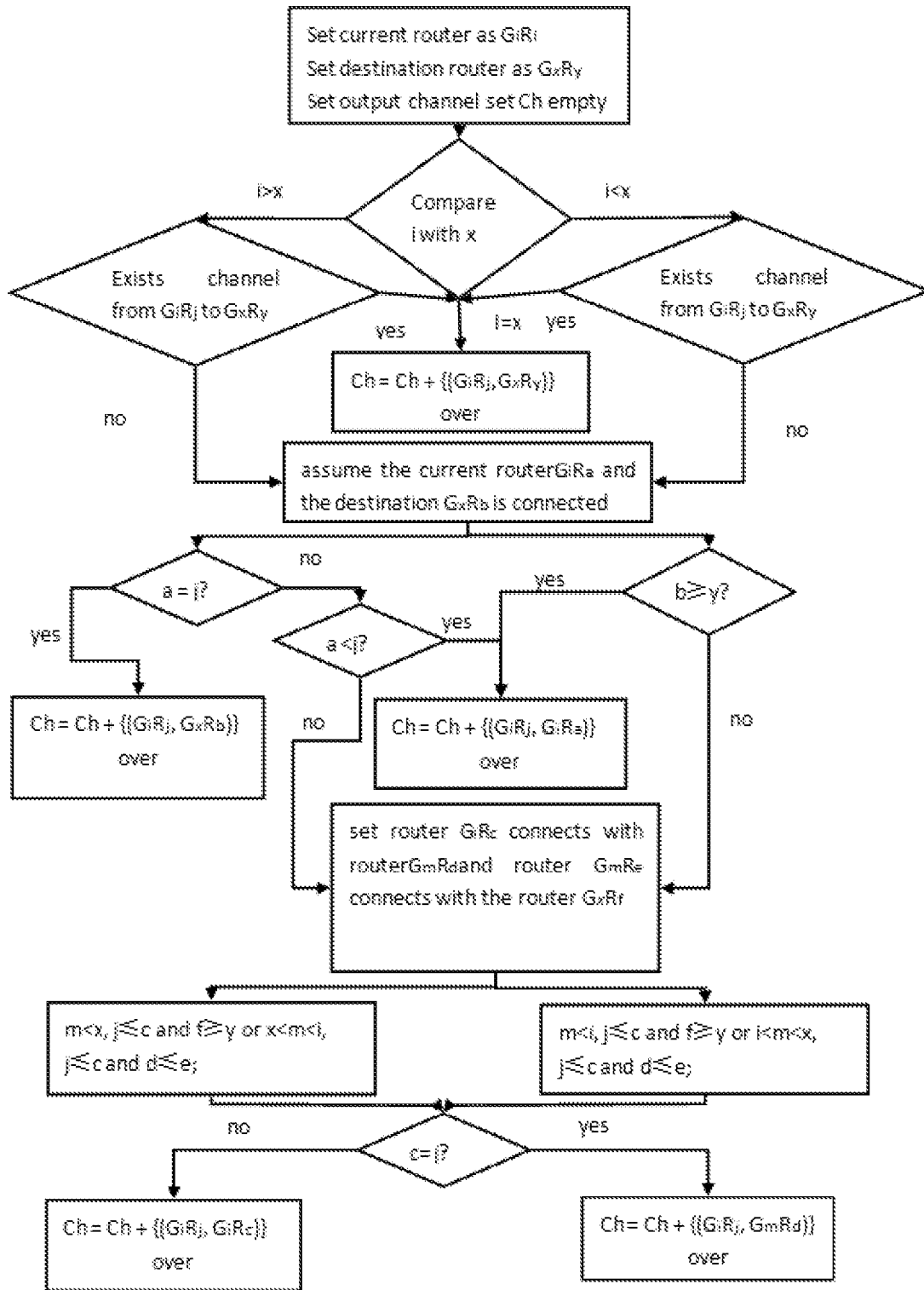


FIG. 3

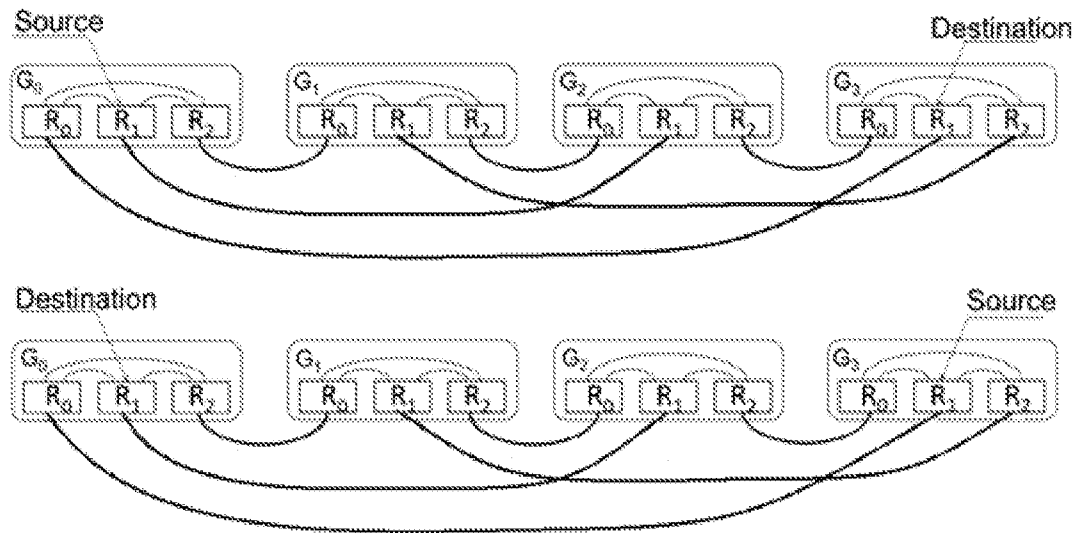


FIG. 4

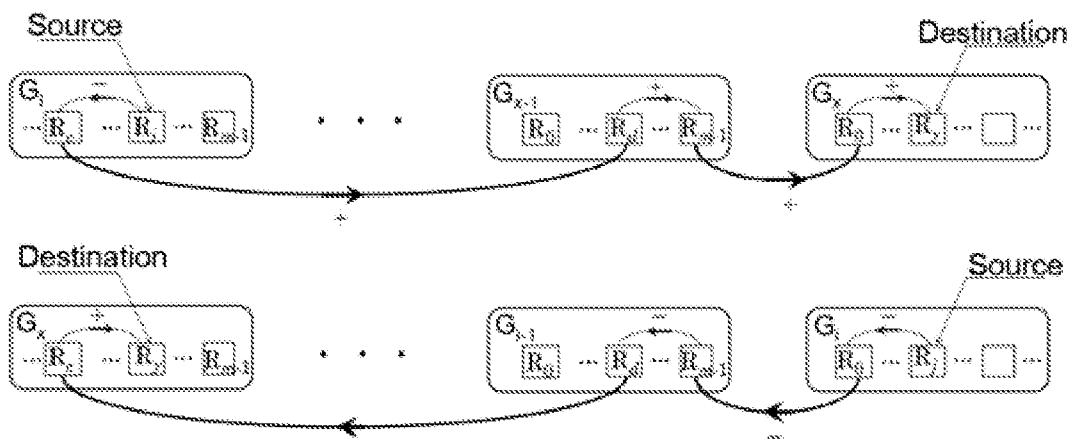


FIG. 5

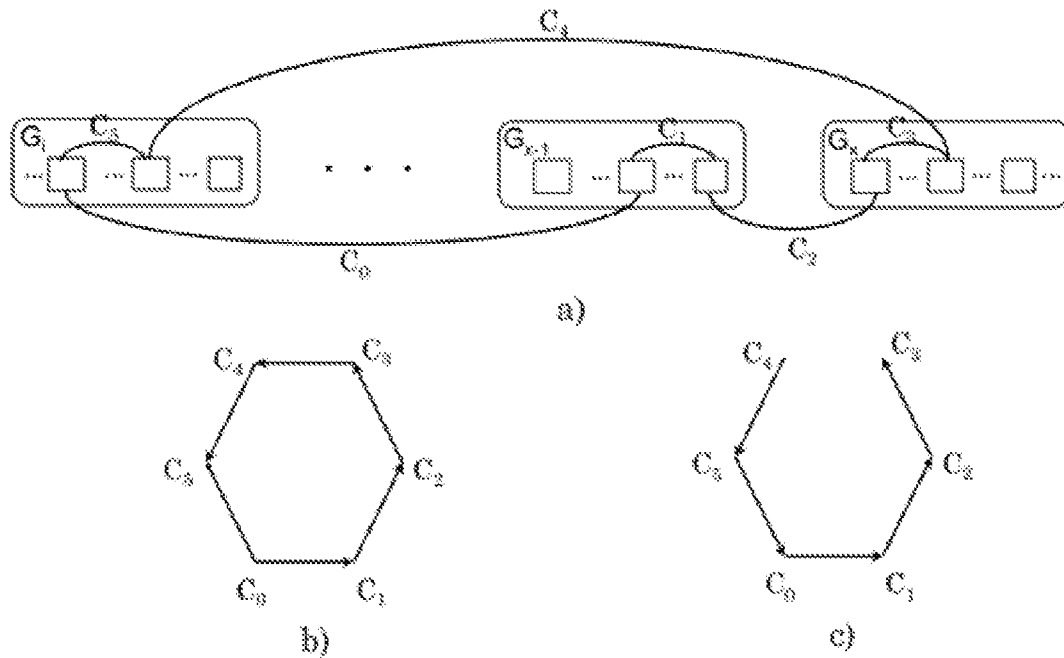


FIG. 6

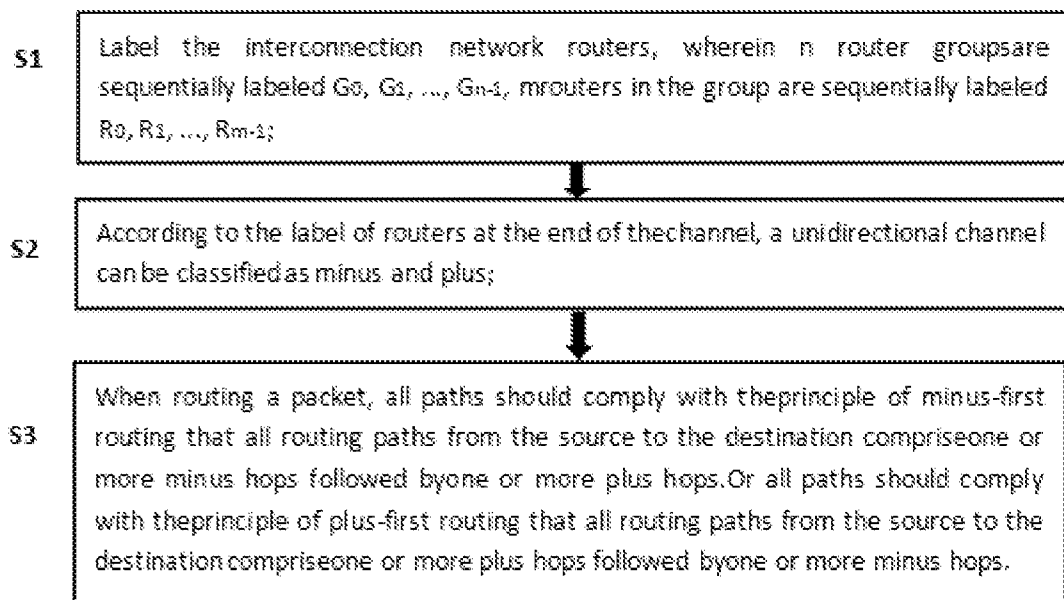


FIG. 7

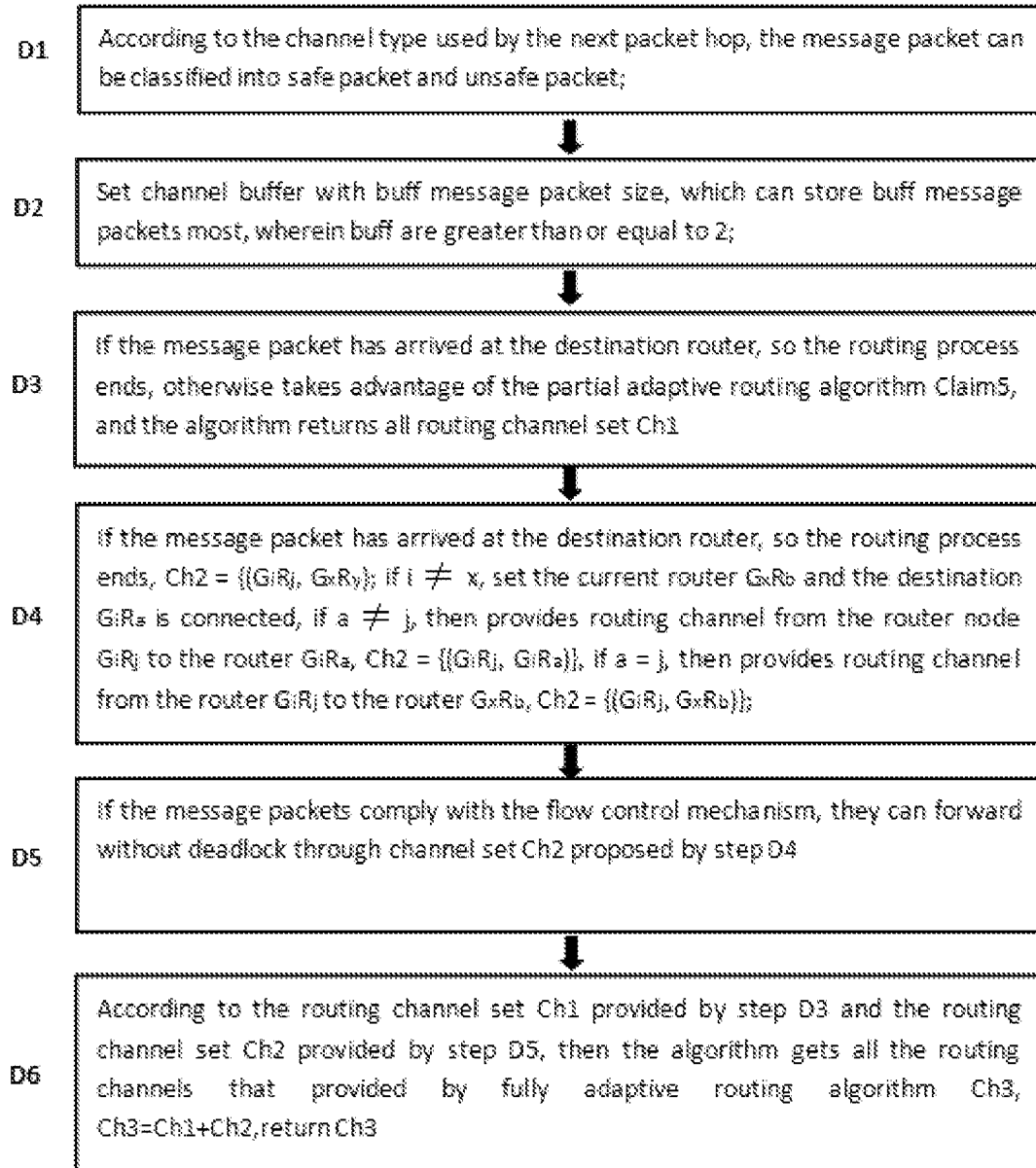


FIG. 8

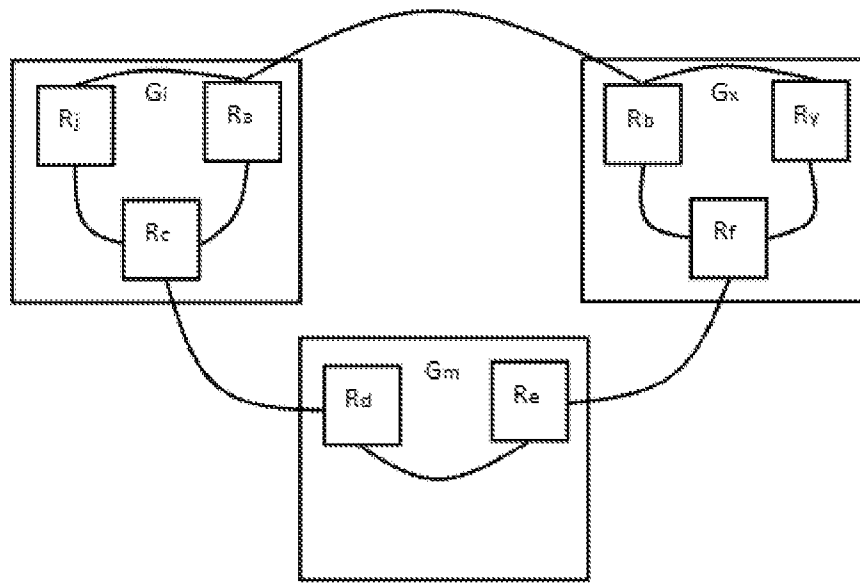


FIG. 9

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2014/080973

A. CLASSIFICATION OF SUBJECT MATTER

H04L 12/771(2013.01)i; H04L 12/721(2013.01)i

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)

H04L; 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)

CNPAT;CNKI;WPI;EPODOC: router?, group?, interconnect, network, inter w group, intra w group, global, channel?, hop?, minus, plus, direction, deadlock 3d free, adaptive, partially, fully, algorithm

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
E	CN 103973564 A (TSINGHUA UNIVERSITY) 06 August 2014 (2014-08-06) claims 1-10 and figures 1-8	5-7
A	US 2010049942 A1 (KIM, JOHN ET AL.) 25 February 2010 (2010-02-25) claims 1, 10 and figure 1	1-10
A	US 2012144065 A1 (CRAY INC.) 07 June 2012 (2012-06-07) the whole document	1-10



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

13 March 2015

Date of mailing of the international search report

03 April 2015

Name and mailing address of the ISA/CN

STATE INTELLECTUAL PROPERTY OFFICE OF THE
P.R.CHINA(ISA/CN)
6,Xitucheng Rd., Jimen Bridge, Haidian District, Beijing
100088, China

Facsimile No. (86-10)62019451

Authorized officer

XIN,Haiming

Telephone No. (86-10)62413285

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2014/080973

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	103973564	A	06 August 2014	Non e			
US	2010049942	A1	25 February 2010	Non e			
US	2012144065	A1	07 June 2012	JP	2012105265	A	31 May 2012
				EP	2461254	A1	06 June 2012