(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 324 577**

(21) Número de solicitud: 200701403

(51) Int. Cl.:
H04L 12/56 (2006.01)

(12) PATENTE DE INVENCIÓN CON EXAMEN PREVIO B2

(22) Fecha de presentación: **10.05.2007**

(43) Fecha de publicación de la solicitud: **10.08.2009**

Fecha de la concesión: **19.01.2010**

(45) Fecha de anuncio de la concesión: **01.02.2010**

(45) Fecha de publicación del folleto de la patente: **01.02.2010**

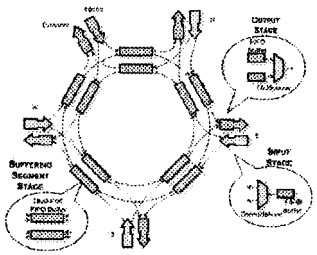
(73) Titular/es: **Universidad de Cantabria**
Pabellón de Gobierno
Avda. de los Castros, s/n
39005 Santander, Cantabria, ES

(72) Inventor/es: **Abad Fidalgo, Pablo;**
Puente Varona, Valentín;
Prieto Torralbo, Pablo y
Gregorio Monasterio, José Ángel

(74) Agente: **No consta**

(54) Título: **Encaminador de mensajes para redes de interconexión de sistemas multiprocesador.**

(57) Resumen:
Encaminador de mensajes para redes de interconexión de sistemas multiprocesador caracterizado por estar especialmente adaptado para el intercambio de información de forma adaptativa e independiente de la topología entre los elementos de proceso integrados en un solo chip.
El encaminador resuelve importantes problemas técnicos que se presentan en la interconexión de un número elevado de procesadores en un único chip cuando actualmente únicamente se dispone de encaminadores especialmente diseñados para la interconexión de dispositivos localizados en chips separados.
El encaminador se caracteriza por los siguientes elementos básicos:
- Dos anillos concéntricos, cada uno esta formado por un grupo de buffers de doble puerto.
- Un conjunto de etapas de entrada y de salida en número igual al grado del encaminador, a través de las cuales entran o salen los paquetes provenientes de los encaminadores vecinos.
- Una etapa de inyección y consumo para su comunicación con el elemento de procesoasociado.



ES 2 324 577 B2

DESCRIPCIÓN

Encaminador de mensajes para redes de interconexión de sistemas multiprocesador.

Sector de la técnica

La invención está relacionada con las redes de interconexión de altas prestaciones que puedan ser empleadas en sistemas multiprocesadores a nivel de sistema o a nivel de un solo chip (sistemas conocidos como CMPs) y en general cualquier sistema que emplee un mecanismo de comunicación descentralizado basado en redes punto a punto, conmutación de paquetes y empleando tecnología electrónica. Dada la elevada aplicabilidad de la idea, ejemplificaremos su uso en el caso específico de los sistemas CMP.

Estado de la técnica

La integración de varios procesadores en un solo chip aparece como la vía más efectiva para manejar el enorme incremento de complejidad de las actuales y futuras microarquitecturas. La organización de su jerarquía de memoria es una característica de diseño de primer orden y el subsistema de comunicación para su soporte una parte esencial del mismo. Aunque actualmente en los multiprocesadores en chip (CMP) es común el uso de estructuras de interconexión centralizadas [7][11] el incremento del número de bloques funcionales [2] obligará a la descentralización en el sistema de comunicación y las redes punto a punto se postulan como el mejor candidato [3][4] y en este contexto es en el que se centra la presentación de esta invención. En el caso particular de los CMPs, dentro del chip, la red y los niveles superiores del sistema están mucho más próximos que en las redes fuera del chip. Esto significa menor latencia y mayor anchura de banda, pero a costa de un incremento del coste de implementación en área de silicio y mayores restricciones de potencia y estos son factores decisivos de diseño [1][8][10] por lo que representa un campo de aplicación más exigente que otros fuera del chip.

La invención aquí presentada cumple los principales requerimientos mencionados, manteniendo un coste hardware adecuado. La arquitectura está basada en un encaminador, denominado Rotary Router, que toma ventaja del pequeño tamaño de los paquetes de información en estos entornos, permite el uso de encaminamiento adaptativo en topologías arbitrarias y los paquetes apenas sufren el efecto del bloqueo de cabeza (*HOL, head of line blocking*).

Actualmente, ninguno de los sistemas multiprocesadores en chip comerciales posee una estructura de interconexión como la que se presenta en esta invención y la bibliografía existente sobre este tipo de sistemas, aunque es profusa, son escasos los trabajos focalizados sobre las redes de interconexión de este tipo de sistemas. Un trabajo reciente en esta línea es [8], pero únicamente se analizan estructuras centralizadas y las redes punto a punto se mencionan como futura línea. Uno los trabajos seminales sobre el empleo de estas redes en los sistemas multiprocesadores en chip con argumentaciones consistentes es [3] y posteriormente ampliado en [1]. No obstante, solo fueron consideradas arquitecturas tradicionales de encaminador. Recientemente Mullins y otros en [9] introducen un nuevo encaminador con un tiempo de paso de un ciclo sin contención, pero con esquemas convencionales de almacenamiento de entrada-salida por lo que no suministra ningún mecanismo especial para el aprovechamiento de los recursos de almacenamiento y mejorar el retraso producido por la congestión, como en el caso de la invención presentada. En [6][2] se propone una organización seccionada para el crossbar, aunque requiere encaminamiento adelantado (*look-ahead*). Del mismo modo, no existe ninguna propuesta similar en el campo de las redes fuera del chip.

Nuestra invención se basa en el empleo de un encaminador de mensajes que se aparta de las estructuras mencionadas en los trabajos anteriores y habitualmente empleadas en los sistemas multiprocesadores dentro y fuera del chip. Su estructura se asemeja a la de una rotonda de tráfico de automóviles. Los paquetes de información entran en una estructura de almacenamiento circular y avanzan buscando el puerto de salida que les acerque a destino. La principal novedad estriba en que si el puerto de salida deseado no está disponible, el paquete continuará circulando por el interior de la estructura circular hasta encontrar una salida valida. El principal efecto positivo conseguido es que evita el bloqueo de los paquetes que circulan tras él y por consiguiente elimina el efecto tan pernicioso que ello tiene sobre el rendimiento del sistema.

[1] J. **Balfour**, W. **Dally**, "Design Tradeoffs for Tiled CMP On-Chip Networks", *International Conference on Supercomputing* (ICS) 2006.

[2] S. **Borkar**, *et al*. "Platform 2015: Intel Platform and Evolution for the Next Decade", *Technology@Intel Magazine*, March 2005.

[3] W. **Dally**, B. **Towles**, "Route Packets, Not Wires: On-Chip Interconnection Networks", *Design Automation Conference* (DAC) 2001.

[4] W. **Daily**, B. **Towles**, "Principles and Practices of Interconnection Networks". *Morgan Kaufmann*, 2004.

[5] P. **Kerman**, L. **Kleinrock**, "Virtual Cut-Through: A New Computer Communication Switching Technique". *Computer Networks*, Vol. 3, pp. 267-286, September 1979.

[6] J. **Kim**, D. **Park**, T. **Theocharides**, N. **Vijaykrishnan**, C. It Das. "A low latency router supporting adaptivity for on-chip interconnects". *Design Automation Conference* (DAC) 2005.

[7] P. **Kongetira**, K. **Aingaran**, K. **Olukotun**, "Niagara: A 32-way Multithreaded SPARC Processor", *IEEE Micro*. Vol. 25, No. 2, pp. 21-29, March/April 2005.

[8] IL **Kumar**, V. **Zyuban**, D. **Tullsen**, "Interconnections in Multi-Core Architectures: Understanding Mechanisms, Overheads and Scaling", *International Symposium on Computer Architecture* (ISCA), 2005.

[9] R **Mullins**, A. **West**, S. **Moore** "Low-Latency Virtual-Channel Routers for On-Chip Networks", *International Symposium on Computer Architecture* (ISCA), 2004.

[10] K. **Olukotun**, L. **Hammond**, "The future of Microprocessors" *ACM Queue*, Vol. 3, No. 7, September 2005.

[11] H. **Hofstee**, "Power Efficient Processor Architecture and The Cell Processor", *International Symposium on High-Performance Computer Architecture* (HPCA), 2005.

[12] V. **Puente**, J.A. **Gregorio**, J. M. **Prellezo**, R. **Beivide**, J. **Duato**, C. **Izu**, "Adaptive Bubble Router. a Design to Improve Performance in Torus Networks", *International Conference of Parallel Processing* (ICPP) 1999.

Descripción de la invención

Breve descripción de la invención

La invención, denominada Rotary Router, es un encaminador de mensajes con una nueva arquitectura especialmente diseñada para sistemas multiprocesadores en chip. El encaminador emplea una estructura de almacenamiento organizada en dos anillos concéntricos y que establecen sentido de giro opuesto en el flujo de los paquetes de información, permitiendo así un control completamente descentralizado y da lugar a una estructura altamente escalable con el número de puertos de comunicación del dispositivo. Los paquetes de información que entran en el encaminador son obligados a circular por los anillos hasta que encuentran un puerto de salida que les acerque a su destino. Ello evita el bloqueo por parada de la cabeza puesto que cuando un puerto de salida se encuentra disponible en un lapso de tiempo breve será utilizado por uno de los paquetes en circulación por cualquiera de los dos anillos. Adicionalmente, permite evitar los problemas de interbloqueo entre paquetes sin utilizar canales virtuales e independientemente de la topología empleada para la interconexión. Esta estructura puede ser implementada con un costo razonable en términos de área y consumo de potencia adecuados para su empleo fundamentalmente en los sistemas multiprocesadores en chip. No obstante, sus especiales características de independencia topológica pueden hacerle útil para su empleo también en sistemas multiprocesadores fuera del chip.

Breve descripción del contenido de las figuras

Figura 1. Diagrama de bloques del encaminador Rotary Router.

Figura 2. Mecanismos de control de flujo empleados.

Figura 3. Estructura de los buffers FIFO de doble puerto.

Figura 4. Ejemplo de multiprocesador en chip (CMP) y detalle del encaminador de mensajes.

Descripción detallada de la invención

El encaminador evita gran parte de los efectos negativos de las estructuras con bufferes en la entrada. No hace uso de mecanismos de arbitrio ni *crossbar* centralizados sino que el arbitrio se lleva a cabo de manera independiente en cada puerto de salida e independientemente del número de puertos. Por otro lado, introduce un nuevo mecanismo para evitar el efecto negativo de parada de los paquetes por bloqueo de la cabeza (*HOL blocking*) y permite adaptar el camino seguido por los paquetes hacia su destino al estado de la red.

La figura 1 muestra un bosquejo del encaminador para una red de grado 2 con un elemento de proceso (*host*) adosado. La estructura esta basada en dos anillos independientes que fuerzan a los paquetes a circular girando en ambas direcciones, yendo de puerto en puerto hasta que encuentren el deseado que les acerque a destino. Cada anillo esta formado por un grupo de buffers de doble puerto (*Dual-port FIFO Buffers, DBF*).

La operación del Rotary Router es simple, cuando un paquete llega a un puerto de entrada del encaminador es enviado a uno de los anillos que forman el encaminador. El paquete comienza a moverse hacia un puerto de salida adecuado empleando los DFBs del anillo. Una vez que el paquete alcanza un puerto de salida adecuado, si está libre dejará el anillo y avanzará al próximo encaminador, en caso contrario, el paquete seguirá circulando por el anillo hasta que alcance otro (o el mismo) puerto de salida adecuado. En principio, el paquete es capaz de dar tantas vueltas como sea necesario antes de dejar el encaminador. Ello evita el arbitrio centralizado, la adaptatividad no añade complejidad y la reducción del bloqueo por parada de la cabeza es inherente a la estructura propuesta.

El encaminador Rotary Router lo componen un número de bloques proporcional a su número de puertos, pero con una estructura y complejidad independientes del grado del nodo y por lo tanto hacen al encaminador altamente escalable. Los tres tipos de bloques son el de entrada (*input*), salida (*output*) y segmento de almacenamiento (*buffer segment*). Los dos primeros constituyen la entrada y salida del encaminador y el tercero es con el que se construyen los anillos.

Bloques básicos

La etapa de entrada, como puede verse en la Figura 1, esta compuesta básicamente por un buffer FIFO y un demultiplexor. Dependiendo de la topología de la red y de los nodos origen y destino, esta etapa es responsable del encaminamiento de cualquier paquete que entre en el encaminador. Esta etapa también es la encargada de seleccionar la dirección por la que se moverá el paquete en el interior, de forma que el tiempo empleado en atravesar el encaminador sea mínimo.

La etapa de salida es la responsable de sacar los paquetes de los anillos y enviarlos al encaminador vecino. Dado que los paquetes provenientes de ambos anillos pueden intentar acceder al mismo puerto de salida simultáneamente, esta etapa tiene dos bufferes y un multiplexor para compartir el único canal físico disponible de comunicación entre encaminadores vecinos. Este espacio de almacenamiento en la salida mejora considerablemente la utilización de los enlaces. Junto con la etapa de entrada, esta etapa es responsable de aplicar los mecanismos de control de flujo entre encaminadores contiguos.

Por último, el bloque más importante del encaminador es el "segmento de almacenamiento" (*buffer segment*). Esta etapa es la que proporciona al encaminador la mayor parte de su funcionalidad y es el origen de sus numerosas ventajas. Esta construida con dos DFBs conectando cada par de puertos contiguos del encaminador. El DFB tiene dos pares de puertos de lectura-escritura, un par es utilizado para construir un anillo (se conecta al DFB previo y al próximo) y el otro par se conecta a las etapas de entrada y salida. Esta etapa debe decodificar la información de encaminamiento generada por la etapa de entrada, e incluida en la cabecera de cada paquete, con el fin de comprobar si debe sacar el paquete por el puerto de salida al que está conectada (si está libre) o por el contrario debe dirigirlo hacia el próximo DFB. En la figura 3 se muestra una implementación de la funcionalidad mencionada.

Control de flujo y algoritmo de encaminamiento

En el encaminador cuya invención se esta describiendo, el control de flujo y los mecanismos de encaminamiento están, en gran medida, ligados al método de evitación de bloqueos. Como se muestra esquemáticamente en la Figura 2, coexisten tres mecanismos de control de flujo dentro del encaminador: uno controla el acceso de los paquetes al encaminador, otro se emplea para manejar su avance por el interior y un tercer mecanismo controla el tránsito hacia otros encaminadores.

El control de flujo empleado para controlar el avance entre encaminadores es Virtual Cut Through [5], es decir, para avanzar un paquete debe haber espacio para su almacenamiento completo en destino. Sin embargo, la entrada de paquetes en los anillos del encaminador esta regulado mediante diversas variantes del mecanismo de control de flujo Burbuja [12] que permite una cierto balanceo de carga y evita anomalías como el interbloqueo de paquetes y/o la inanición (*starvation*). Si el paquete proviene de una elemento de proceso (*host* u otro dispositivo capaz de generar tráfico y que se encuentre conectado al encaminador) se exigirán al menos tres huecos para almacenar tres paquetes (dos más de los estrictamente necesarios). Por el contrario, si el paquete proviene de otro encaminador vecino se exigirán como mínimo dos huecos, pero se incrementará la exigencia de espacio si se detecta un fuerte desbalanceo respecto a la entrada de paquetes por los restantes puertos. Esto evita el efecto de inundación del encaminador por la entrada de paquetes por este puerto y por lo tanto la inanición de los demás.

Por último, para manejar el avance de los paquetes en el interior de los anillos se emplea un mecanismo de control de flujo basado en la ocupación. Cada DFB mantiene información sobre la ocupación de los DFBs vecinos. Así, un paquete avanzará al próximo DFB solo si el nivel de ocupación del buffer destino es menor o igual que el actual. Este control de flujo balancea la ocupación de todos los DFBs en el anillo y ecualiza la probabilidad de inyección en cada puerto de entrada.

Con respecto al algoritmo de encaminamiento, el paquete siempre trata de aproximarse a destino y por lo tanto puede calificarse como adaptativo mínimo. No obstante, bajo ciertas condiciones el paquete puede ser obligado a salirse de la ruta más corta (*misrouting*). Si un paquete lleva a cabo un número predeterminado (y suficientemente elevado) de vueltas a un anillo sin conseguir ningún puerto que le acerque a destino, será marcado para salirse de la ruta más corta (hacer *misrouting*) y saldrá del encaminador por el primer puerto de salida libre que encuentre en su avance por el anillo. Una vez el paquete ha abandonado el encaminador, la marca será eliminada y el paquete intentará seguir ruta mínima de nuevo. Por tanto, mediante el empleo en la estructura descrita de los algoritmos de encaminamiento y control de flujo mencionados, la invención evita la aparición de anomalías típicas de las redes de interconexión: interbloqueo, inanición y deambulado infinito (*livelock*).

Ejemplo de realización de la invención

Para la mejor comprensión del invento, a continuación se propone un ejemplo sobre dónde y cómo aplicar la invención. El encaminador Rotary Router puede ser empleado, en principio, en cualquier red de interconexión entre procesadores, pero las principales novedades de su desarrollo suministran mucha ventaja en los sistemas multiprocesadores en chip. En la figura 4 se muestra un sistema CMP compuesto de 16 procesadores, cada uno de los cuales posee una memoria cache privada de nivel uno (L1) y comparte la memoria cache de nivel dos (L2) con el resto de los procesadores en una estructura no uniforme con direccionamiento estático, conocida habitualmente como S-NUCA. El sistema, obviamente, también dispondrá de una memoria global, en este ejemplo, de 4 gigabytes.

El empleo natural del encaminador es la implementación de la red de interconexión para la comunicación entre los 256 bancos de la S-NUCA, los procesadores con sus correspondientes cache privadas y la jerarquía de memoria de nivel superior (memoria principal). La topología elegida en el ejemplo ha sido una red toroidal 8x8, de manera que los bancos del nivel 2 de cache se conectan de 4 en 4 en cada encaminador formando una matriz de 16x16. Los 16 procesadores se conectan en la periferia de la matriz y los accesos a memoria principal se distribuyen de manera uniforme en la red. Por ello, fácilmente se deduce que la función básica del encaminador es el intercambio de paquetes de información correspondientes a, por una parte, líneas de cache intercambiadas entre los distintos niveles de la jerarquía, y por otra, los comandos necesarios para su requerimiento así como el mantenimiento de la coherencia entre los mencionados niveles. En el ejemplo que nos ocupa se esta suponiendo que el sistema tiene líneas de cache de 64 bytes y comandos de 16 bytes. Considerando enlaces entre encaminadores de anchura, es decir lo que se denomina el *phit*, es de 128 bits, se concluye que el ancho de los paquetes que circularán por la red-ejemplo es de 5 y 1 *phit* respectivamente.

REIVINDICACIONES

1. Encaminador de mensajes para redes de interconexión de sistemas multiprocesador. Está **caracterizado** por estar compuesto de los elementos siguientes:

- Una etapa de entrada por cada uno de los enlaces de interconexión, incluido el del elemento de proceso al que esté asociado.
- Una etapa de salida por cada uno de los enlaces de interconexión, incluido el del elemento de proceso al que esté asociado.
- Una estructura de búferes de almacenamiento interconectados formando dos anillos concéntricos de direcciones opuestas a través de los cuales son obligados a circular los paquetes de información hasta que alcancen un puerto de salida libre que los acerque a su destino.
- Mecanismo para la limitación del número máximo de paquetes que pueden introducirse en cada anillo.
- Un mecanismo para la cuenta y limitación del número de giros que podrán dar los paquetes de información en el interior de cada anillo.

2. Encaminador de mensajes para redes de interconexión de, según reivindicación 1ª, **caracterizado** porque cualquiera de sus etapas de entrada está formada por un buffer de almacenamiento de tipo FIFO, conectado a un multiplexor que permite dirigir los paquetes a cualquiera de los dos anillos de conforman el encaminador.

3. Encaminador de mensajes para redes de interconexión, según reivindicación 1ª, **caracterizado** porque cualquiera de sus etapas de salida está formada por dos buffer de almacenamiento de tipo FIFO que recoge los paquetes de cada uno de los anillos que conforman el encaminador que deseen utilizar esta salida y los multiplexa en el único canal de salida.

4. Encaminador de mensajes para redes de interconexión, según reivindicación 1ª, **caracterizado** porque cualquiera de sus etapas intermedias de almacenamiento está formada por dos búferes de doble puerta, cada uno de los cuales forman parte de uno de los anillos y se encargan de almacenar los paquetes provenientes tanto del puerto de entrada asociado, como de los que circulan por el mismo anillo.

5. Encaminador de mensajes para redes de interconexión, según reivindicación 1ª, **caracterizado** porque el número de paquetes que puede ser introducido en cada anillo nunca será superior a su capacidad máxima menos dos por lo que evitará el interbloqueo de paquetes en el interior de cada anillo.

6. Encaminador de mensajes para redes de interconexión de sistemas multiprocesador en chip, según reivindicación 1ª, **caracterizado** porque contabilizará el número de giros que vaya dando cualquier paquete de información en el interior de cualquiera de sus anillos. Si transcurrido un número máximo de vueltas a cualquiera de los anillos, un paquete no ha encontrado libre uno de sus puertos deseados, será obligado a abandonar el encaminador por el primer puerto de salida libre que se encuentre en su camino de giro.

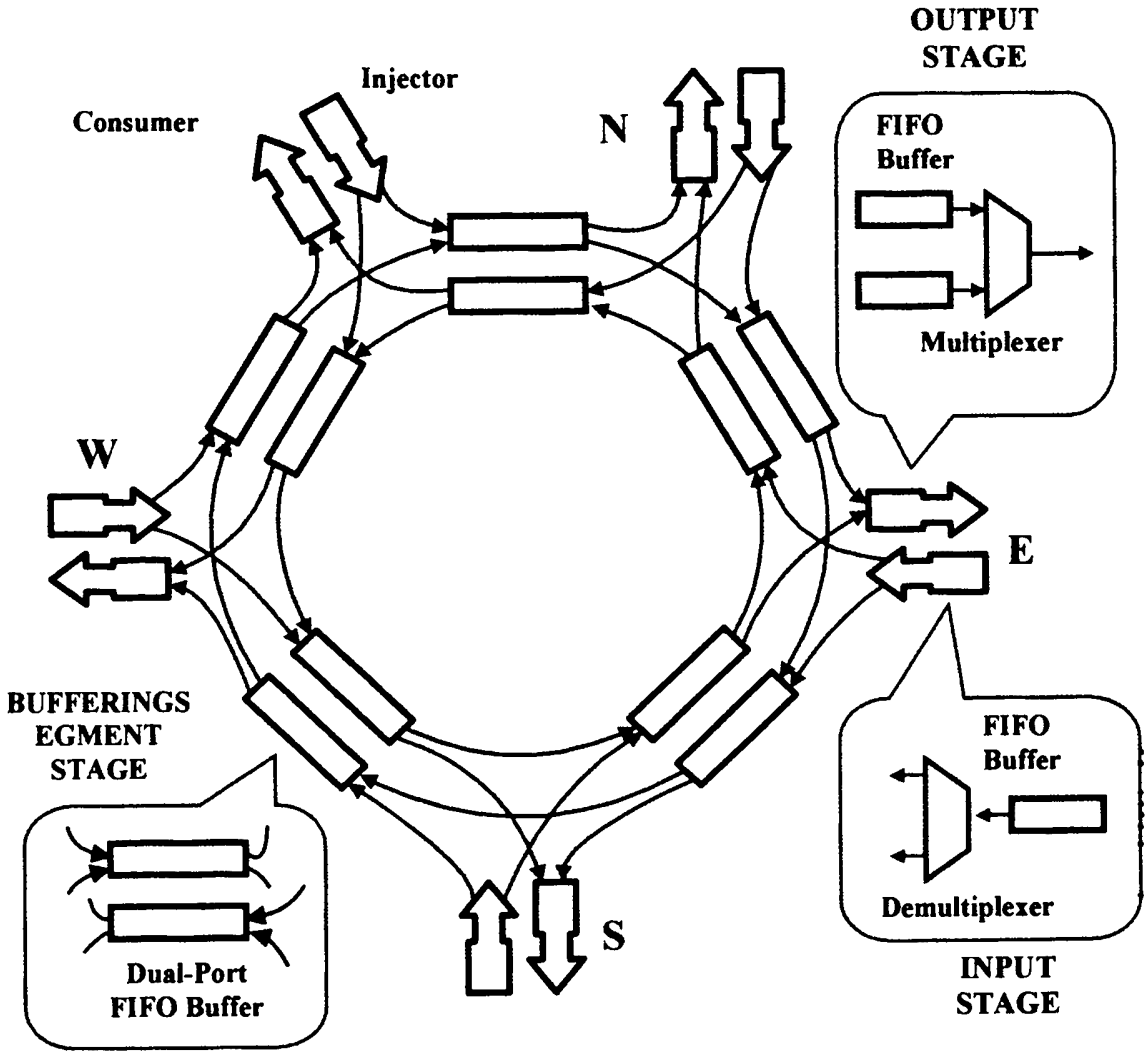

OUTPUT
STAGE
Injector
Consumer
N
FIFO
Buffer
Multiplexer
W
E
BUFFERINGS
EGMENT
STAGE
FIFO
Buffer
Demultiplexer
Dual-Port
FIFO Buffer
S
INPUT
STAGE

Figura 1.

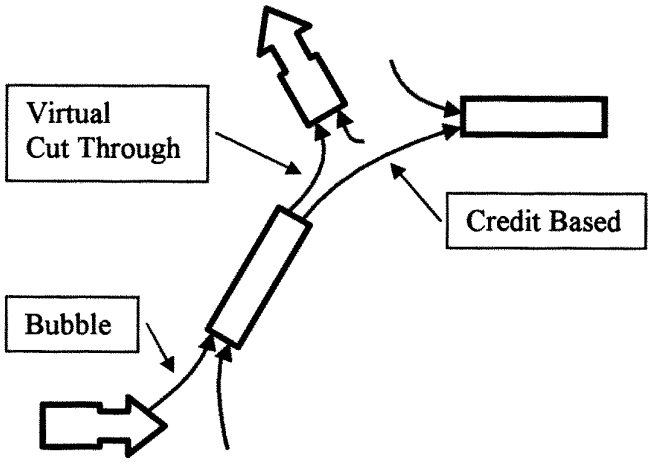


Figura 2

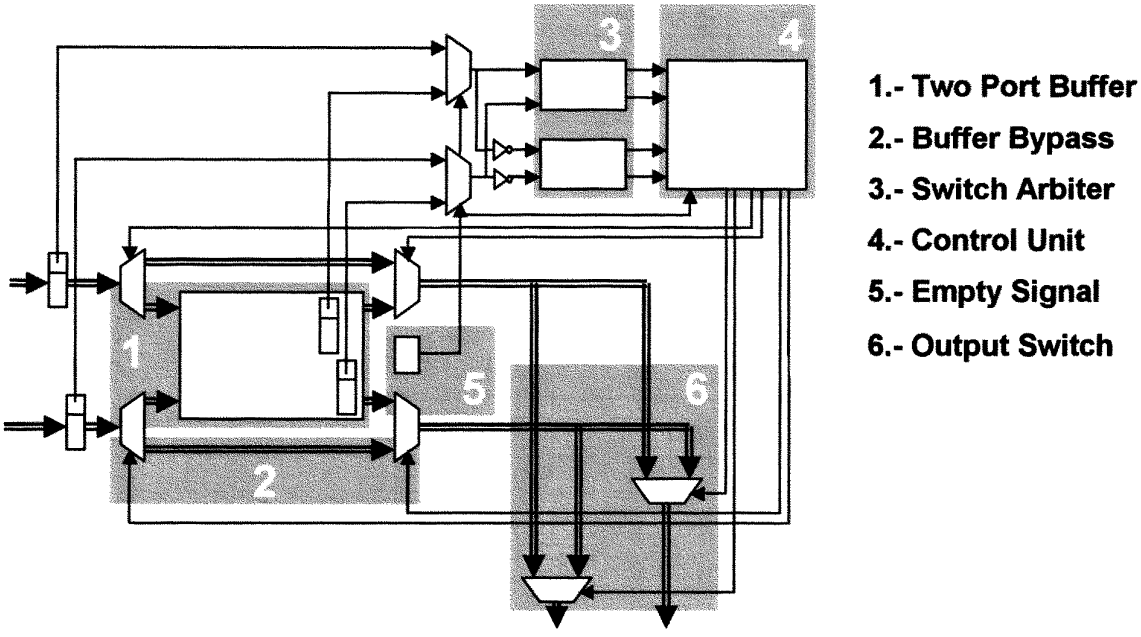


Figura 3

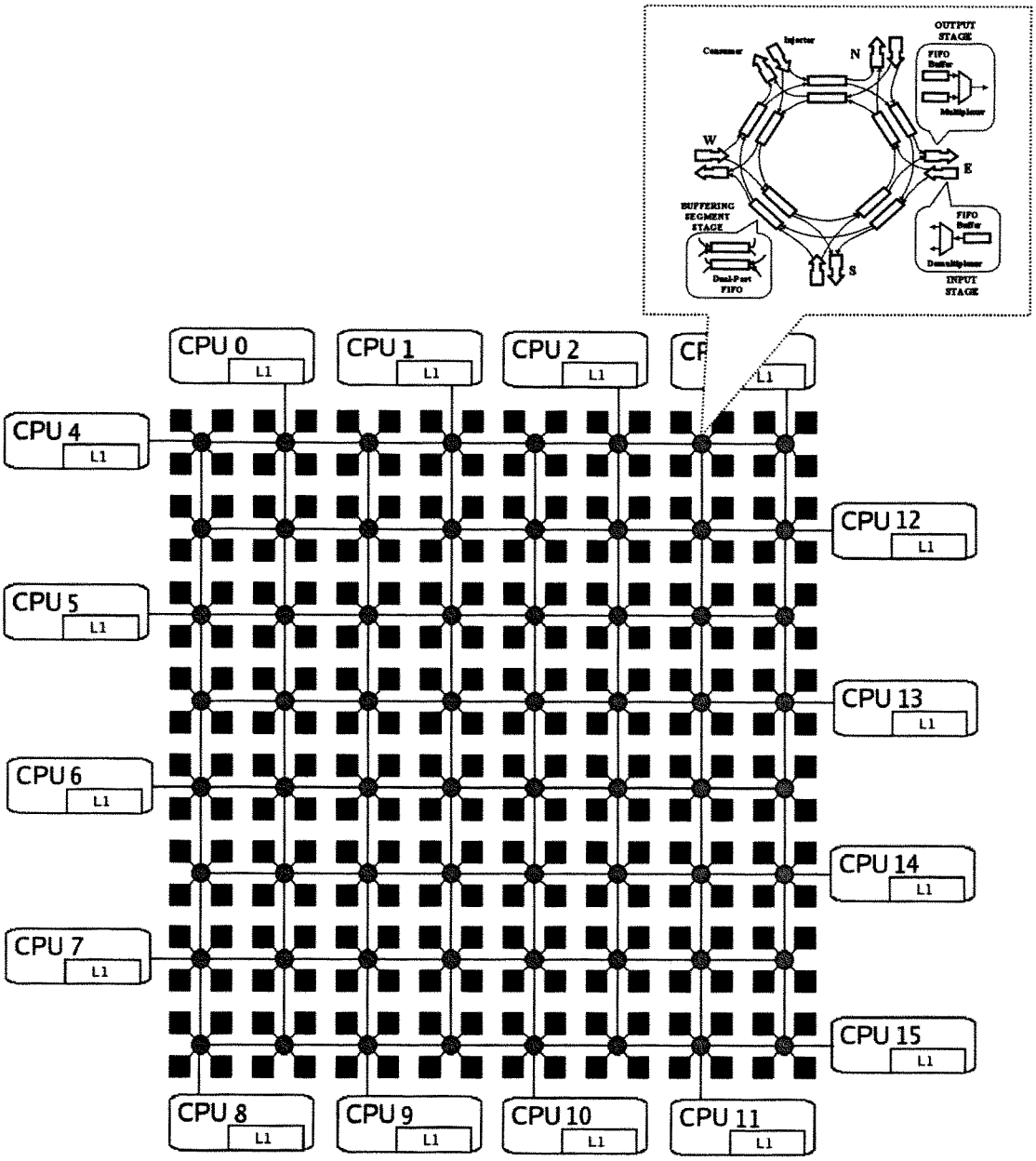

CPU 0
L1
CPU 1
L1
CPU 2
L1
L1
CPU 4
L1
CPU 12
L1
CPU 5
L1
CPU 13
L1
CPU 6
L1
CPU 14
L1
CPU 7
L1
CPU 15
L1
CPU 8
L1
CPU 9
L1
CPU 10
L1
CPU 11
L1
OUTPUT STAGE
FIFO Buffer
Multiplexer
N
W
E
S
BUFFERING SEGMENT STAGE
Dual-Port FIFO
FIFO Buffer
Demultiplexer
INPUT STAGE

Figura 4.

OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) ES 2 324 577

(21) Nº de solicitud: 200701403

(22) Fecha de presentación de la solicitud: **10.05.2007**

(32) Fecha de prioridad:

INFORME SOBRE EL ESTADO DE LA TÉCNICA

(51) **Int. Cl.:** ***H04L 12/56*** (2006.01)

DOCUMENTOS RELEVANTES

Categoría	(56) Documentos citados	Reivindicaciones afectadas
A	US 2004230751 A1 (BLAKE et al.) 18.11.2004, figura 1; párrafos [17,43-49].	1,4-6
A	US 2004230726 A1 (BLAKE et al.) 18.11.2004, figura 12; reivindicaciones 1-13.	1
A	DUATO, LYSNE, PANG y PINKSTON, "Part 1: Theory of Deadlock-Free Dynamic Network Reconfiguration", Mayo 2005. <http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=01411730>	1-3
A	MOUSSA, MULLER, BAGHDADI y JÉZÉQUEL - Electronics Department ENST Bretagne (Francia), "Butterfly and Benes-Based on Chip Communication Networks for Multiprocessor Turbo Decoding" <http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=04211873>	1-3

Categoría de los documentos citados

X: de particular relevancia
Y: de particular relevancia combinado con otro/s de la misma categoría
A: refleja el estado de la técnica
O: referido a divulgación no escrita
P: publicado entre la fecha de prioridad y la de presentación de la solicitud
E: documento anterior, pero publicado después de la fecha de presentación de la solicitud

El presente informe ha sido realizado

☒ para todas las reivindicaciones ☐ para las reivindicaciones nº:

Fecha de realización del informe	Examinador	Página
27.07.2009	B. Pérez García	1/1