Arquitectura de Computadores

Notas de Clase

Elisardo Antelo Suárez

Tema 2: Subsistema de Memoria e Interconexión

1- Sistema de Memoria: Chips de memoria.

Para estudar o sistema de memoria e comprender as súas limitacións, comezamos estudiando os chips de memoria DRAM. Como vemos na Figura 1, os chips de memoria están compostos por N bancos independentes. Cada banco contén filas de datos chamadas ñpáxinasò, e cada páxina está constituída por varias palabras de b bytes, que é o ancho da interface da memoria co exterior. Polo tanto os accesos fanse en grupos de b bytes, e o tempo de acceso depende de se os enderezos emitidos de xeito consecutivo son á mesma páxina ou a unha páxina diferente dentro do mesmo banco.

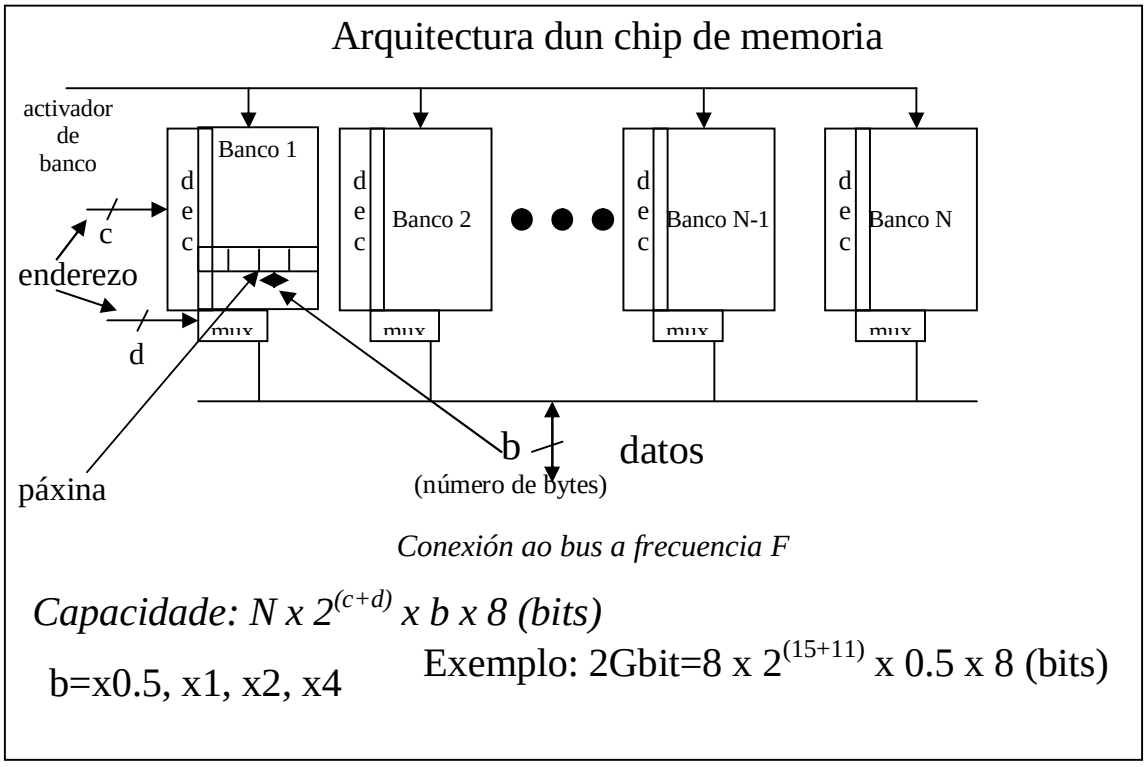

Figura 1: Arquitectura dun chip de memoria.

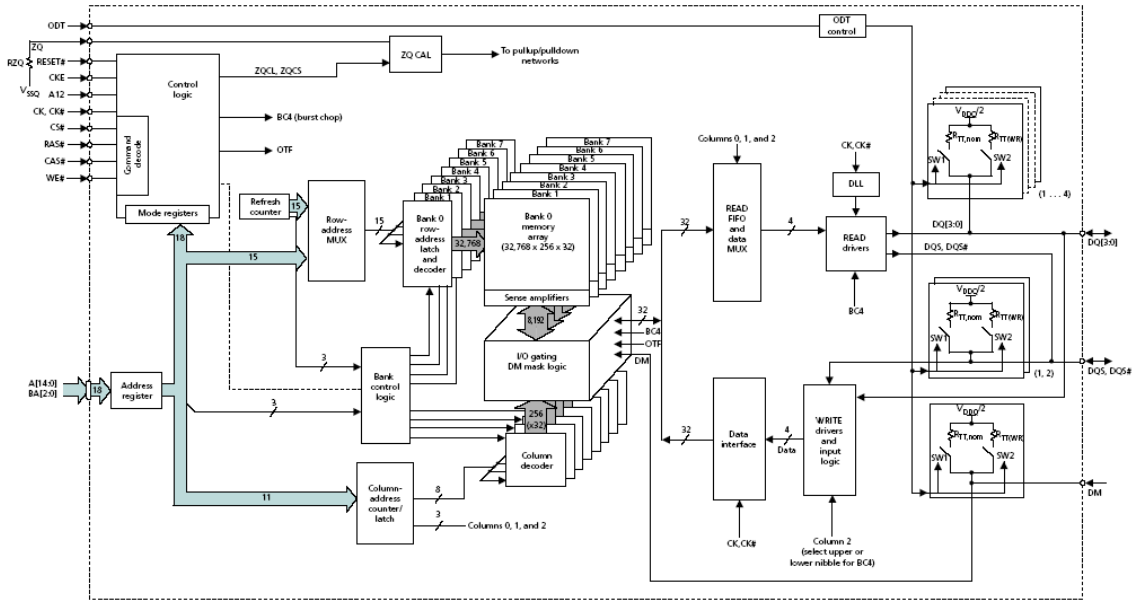

Figura 2: Exemplo de arquitectura detallada dun chip de memoria.

A Figura 2 amosa unha descripción máis en detalle para un chip de memoria de 2Gbit. O chip está organizado en 8 bancos de 32K páxinas de 1024 bytes cada un. A conexión de datos co exterior é de 4 bits (DQ[3:0] na parte dereita). Os enderezos son de 29 bits (0.5 Giga elementos de 4 bits para completar 2Gbit) en total, e chegan en dous grupos: un primeiro grupo de 18 bits con BA[2:0] para seleccionar o banco e A[14:0] para selección da páxina dentro do banco, e logo un grupo de 11 bits para a selección dos bits correspondentes dentro da páxina. Como vemos, en cada lectura seleccionanse 32 bits da páxina correspondente (para isto utilínzase só 8 dos 11 bits). Logo mediante un multiplexo faise a selección dos 4 bits que se pretende leer. Polo tanto faise unha prelectura de 8 feixes de 4 bits. Neste caso dise que a lonxitude de ráfaga é de 8. É dicir podense facer 8 lecturas de feixes de 4 bits á máxima velocidade (debido á prelectura). Para as escrituras ocorre algo semellante.

[Capacidade de memoria]: A capacidade de memoria está determinada polo número de bancos multiplicado polo número de páxinas, e o número de bytes por páxina. Nas implementacións actuais a interface co exterior faise con b=x0.25 (2 bits), x0.5 (4 bits), x1, x2 ou x4 bytes. Deste xeito limítase o número de pads de entrada/saída dos chips de memoria para que sexan de baixo custo. Por exemplo un chip de memoria de 2Gbits pode estar organizado como 8 bancos, con 32K páxinas de 1024 bytes cada unha con entrada/saída de 4 bits.

O escalamento da capacidade de memoria dos chips está dado por varios factores: a área por cela de memoria, o número de celas e os circuítos lóxicos para acceder a información almacenada nas celas e comunicala co exterior. A área por cela está dada polo producto dun factor de área da cela (relativo á forma específica da cela de memoria) e o cadrado do factor xeométrico que caracteriza o nodo tecnolóxico (L^2). O array de celas ocupa aproximadamente o 60% do chip de memoria, pero parece que a tendencia é a que esta porcentaxe se reduza debido ao mal escalamento dos circuítos lóxicos. Actualmente a capacidade dos chips de memoria dóbrase cada tres anos, con cambio de nodo tecnolóxico cada dous anos, o que permite ter dúas versións (en dous

procesos diferentes) para o mesmo tamaño de chip. Debido ao mal escalamento dos circuítos lóxicos, e a que estes representan unha porcentaxe cada vez maior do chip de memoria, os fabricantes afrontan problemas de coste (aumento de área) ao incrementar a capacidade de almacenamento. Para paliar este problema xa se está utilizando integración 3D, é dicir máis dun chip de memoria no mesmo encapsulado.

Actualmente están en producción os chips de 4Gbit, e coexisten no mercado coas xeracións de 512Mbits, 1Gbits e 2Gbit (e comenzan a aparecer chips de 8 Gbit utilizando configuracións 3D, como por exemplo, 4 chips de 2 Gbit cada un no mesmo encapsulado). Este escalamento da capacidade por chip responde ás necesidades de escalamento da memoria principal que demandan as progresivas xeracións de aplicacións software.

Ademais da capacidade da memoria, a latencia da memoria e o seu ancho de banda son dous aspectos fundamentais do sistema de memoria. A latencia da memoria incide directamente (como xa vimos) na latencia por cada fío da aplicación. O ancho de banda ten unha incidencia menor na latencia do fío (aínda que pode aumentar a latencia se o ancho de banda demandado satura o ancho de banda dispoñible), pero é moi importante no que atinxe á productividade en termos do número de fíos (fíos/segundo que se completan) da aplicación.

[Latencia dos chips de memoria]. A latencia da memoria varía dependendo do patrón de acceso. O maior custo temporal na latencia de memoria ven dado pola activación do banco e posterior activación da páxina correspondente. Logo do acceso, é necesario realizar unha operación (precarga) que restaura as condicións eléctricas, necesarias para poder acceder novamente a outra páxina dentro do mesmo banco.

Polo tanto, o patrón máis favorable para o acceso (lectura ou escritura) é cando todos os bytes corresponden á mesma páxina dentro dun banco. Se o número de bytes a ler supera o tamaño da páxina entón accédese a outros bancos, pero a activación do banco e a activación da páxina nos outros bancos xa estará feita cando se teña que ler ou escribir a información.

Podemos en primeira orde estimar a latencia de acceso **Dk** (lecturas ou escrituras) a k x b bytes consecutivos de memoria como

$$Dk = Tst + k\ Tck$$

que corresponde á suma dun termo constante **Tst**, correspondente á primeira activación do banco e páxina máis o tempo de precarga final, máis un tempo proporcional ao número de bytes que se acceden e ao ritmo de lectura ou escritura no sistema de interconexión (ciclo do de transferencia de datos do sistema de interconexión da memoria **Tck=1/F** ï inverso da súa taxa de transferencia de datos).

Para xeracións futuras de memoria DRAM podemos asumir que o termo constante non escala e que toma un valor duns 25-30 ns (isto é debido a que dentro do chip de memoria, algunhas conexións críticas non escalan a súa lonxitude, polo que o aumento de velocidade propio da tecnoloxía vese compensado polo efecto das conexións longas). En termos do nodo tecnolóxico T, podemos expresar esta parte constante da latencia de

acceso en unidades FO4 como Tst=12x 10^4/T(nm) (FO4). Como T escala por x0.7 por nodo tecnolóxico, obtemos que Tst escala por x1/0.7 approx. 1.5 (en unidades de FO4).

Así, por exemplo para o nodo de 90nm, supoñendo un procesador con ciclo de reloxo de 10FO4, o termo constante da latencia da memoria é duns 1333 FO4, é dicir, 133 ciclos de reloxo do procesador. Para a tecnoloxía de 65 nm este termo pasa a ser de 185 ciclos de procesador.

A latencia para o acceso de k x b bytes (secuenciados en accesos de b bytes, correspondentes por exemplo a lectura/escritura de liñas cache) está dado por (120+5k/F(GHz)) x 10^3/T (FO4), donde F(GHz) é a taxa de transferencia do sistema de interconexión da memoria en GHz.

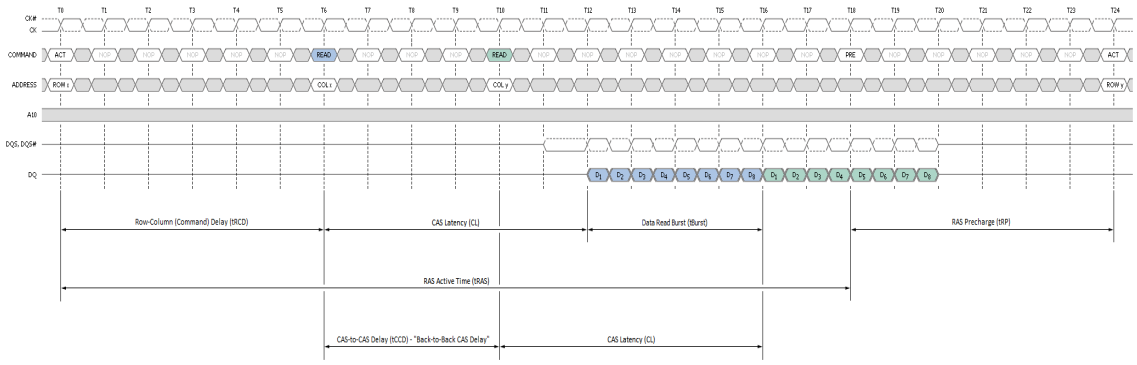

Figura 3: Exemplo de restricións temporais no control dos chips DRAM.

O control temporal dos chips de memoria é bastante complexo (iso incide na complexidade do controlador de memoria). A Figura 3 ilustra un exemplo no que se activa unha páxina e fanse dúas lecturas consecutivas nesa mesma páxina (con k=8), para logo realizar unha precarga que restaura os valores que se leron e permite preparar os circuítos para abrir outra páxina. A figura ilustras algunhas restricións temporais:

tRCD: o número mínimo de ciclos entre o comando de apertura da páxina dun banco determinado e o comando de lectura da columna dentro da páxina.

CL: o número de ciclos entre o comando de lectura da columna e a aparición na saída dos datos (existe unha restrición semellante para escritura).

tRP: o número de ciclos entre a chegada do comando de precarga e a chegada dun novo comando para apertura de páxina. É dicir a latencia de precarga.

tRAS: o mínimo número de ciclos entre un comando para apertura de páxina e o comando para precarga.

tCCD: latencia entre dous comandos consecutivos de lectura na mesma páxina.

tRTP: latencia mínima entre un comando de lectura da columna e un comando de precarga da páxina (este non aparece na figura xa que a precarga está limitada neste exemplo por tRAS).

Os fabricantes habitualmente indican estas latencias coma unha secuencia de catro números: **CL-tRCD-tRP-tRAS**. Todas estas latencias **refírense a ciclos de reloxo da interface que utiliza o chip DRAM para comunicarse co controlador de memoria**. Por exemplo un código 8-8-8-24 indica que **CL**=8 ciclos, **tRCD**=8 ciclos, **tRP**=8 ciclos e **tRAS**=24 ciclos. Notar que para unha determinado nodo tecnolóxico e familia de chips DRAM é posibe ter diferentes valores da frecuencia do sistema de interconexión, polo que en función de cal sexa esta frecuencia, os valores de **CL**, **tRCD**, **tRP** e **tRAS** variarán de tal xeito que en termos de tempo absoluto (ns), os retardos serán case iguais.

[Ancho de banda]: Estimemos agora o ancho de banda por cada chip de memoria (o ancho de banda total do sistema de memoria será a suma do ancho de banda de todos os chips de memoria que se acceden simultaneamente). O mellor caso corresponde ao acceso de **k x b** bytes dentro da mesma páxina para cada un dos **N** bancos sen conflicto de banco (é dicir, accedese aos **N-1** bancos restantes antes de repetir o acceso a un banco). Para simplificar vamos a asumir que o acceso aos **k x b** bytes ten un coste temporal de **Tst+ k Tck= Tst+ k /F**. Isto equivale a considerar o coste do acceso coma a suma de **tRCD+CL+a x tRP** (convertidas a unidades de tempo absolutas), donde a é un valor entreo 0 e 1, para indicar a fracción do retardo de precarga que non solapa coa saída de datos. Polo tanto para facer unha estimación do ancho de banda é necesario considerar dous casos extremos:

a) *Taxa de transferencia do sistema de conexión moi baixa*: tal e como se amosa na parte superior da Figura 4 (exemplo para catro bancos e k=4), o tempo Tst asociado ao acceso de cada banco e páxina queda enmascarado pola lenta utilización do sistema de conexión de memoria para transferir os datos que se van accedendo dende cada un dos bancos. Polo tanto o ancho de banda está dado polo ancho de banda do sistema de conexión (b x F).

b) *Frecuencia do sistema de conexión moi alta*: neste caso, tal e como se amosa na parte inferior da Figura 4, os datos accedidos transfírense moi rápido e non enmascaran o tempo **Tst** para cada banco, polo que incluso o sistema de conexión queda sen utilizar durante algúns ciclos, esperando polo acceso aos datos. O número total de datos accedidos é de **k x b x N** nun tempo **Tst + k /F**. O cociente destas cantidades é o ancho de banda.

$$Bw = \frac{k \times b \times N}{T_{st} + \frac{k}{F}}$$

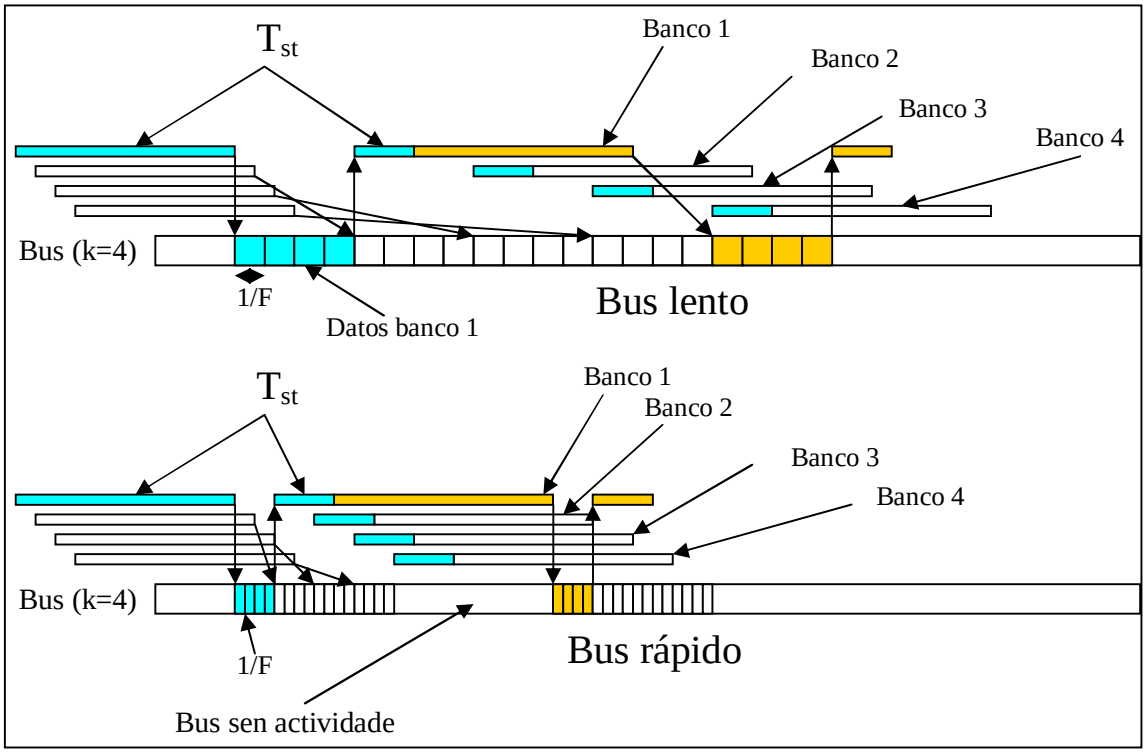

Figura 4: Patróns de acceso para estimar o ancho de banda do chip de memoria.

O obxectivo de deseño dos chips de memoria debe ser o de aproveitar o máximo ancho de banda da canle (sistema de conexión da memoria): **b x F** (bytes/sec por chip de memoria). Polo tanto a ecuación para o deseño obtense igualando o ancho de banda no mellor caso que pode dar o chip de memoria co máximo ancho de banda da canle. Neste caso equilíbrase o ancho de banda que poden dar os bancos e o ancho de banda da canle. Esta ecuación relaciona todos os parámetros relevantes no deseño do sistema de memoria: taxa de transferencia e ancho (por chip) do sistema de conexión, número de bancos, e todo isto coa restrición dunha capacidade máxima do chip de memoria. A conclusión é que neste punto a taxa de transferencia é proporcional ao número de bancos dentro dos chips de memoria.

$$F \times b = \frac{k \times b \times N}{T_{st} + \frac{k}{F}} \quad \Rightarrow \quad F = \frac{k \times (N-1)}{T_{st}}$$

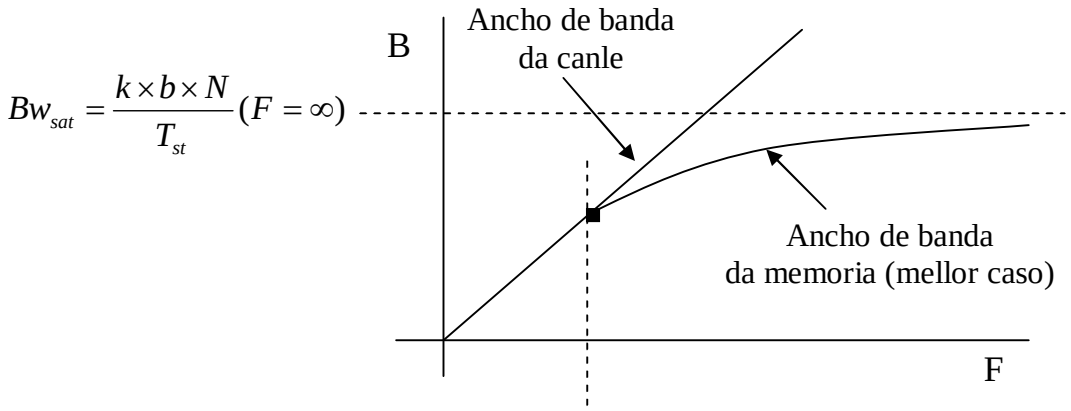

Figura 5: Ancho de banda (B) por chip de memoria vs taxa de transferencia do sistema de interconexión do sistema de memoria.

A Figura 5 amosa o ancho de banda do chip de memoria fronte á taxa de transferencia do sistema de conexión, para o patrón de acceso máis favorable. Como xa vimos, para taxas de transferencia baixas con accesos consecutivos co mellor patrón de acceso, o efecto limitante é o ancho de banda do sistema de conexión, e non o acceso aos bancos e as páxinas (quedan enmascarados pola lentitude de transferencia dos datos). O caso oposto corresponde a taxas de transferencia moi elevada, no que o ancho de banda está determinado polo acceso aos bancos. Este último é o ancho de banda de saturación da memoria (o ancho de banda de saturación obtense poñendo F igual a infinito na expresión do ancho de banda para taxas de transferencia elevadas).

Posto que o consumo de potencia e custe de deseño e implementación do sistema de conexión son proporcionais á taxa de transferencia, esta debe limitarse a un punto próximo ao punto no que o ancho de banda comeza a saturarse (cóbado da curva).

O peor patrón de acceso para o ancho de banda corresponde ao acceso repetido dentro do mesmo banco a diferentes páxinas. Supoñendo unha granularidade de acceso k<P (é dicir, cada acceso corresponde a k x b bytes, e esta cantidade é inferior a unha páxina), o ancho de banda está dado polo cociente dos bytes accedidos e a suma da parte constante (acceso a banco+ páxina+ precarga) e o tempo de transferencia polo sistema de conexións en grupos de b bytes (por chip de memoria). Notar como neste caso non aparece na expresión o número de bancos no chip de memoria.

$$Bw = \frac{k \times b}{T_{st} + \frac{k}{F}}$$

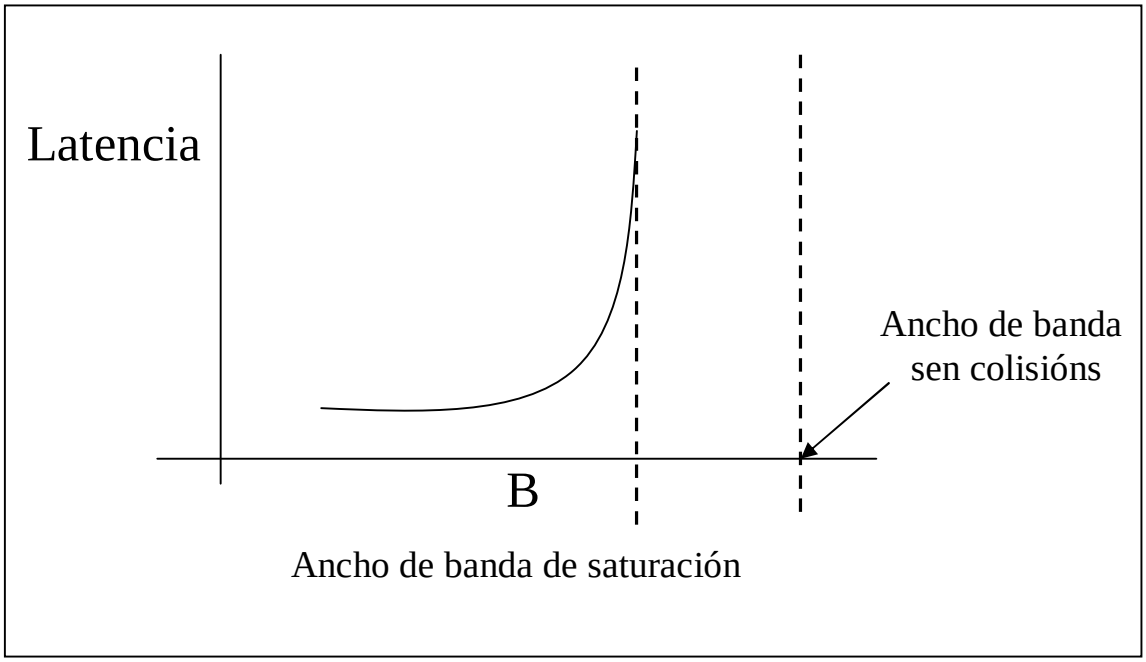

Figura 6: Ancho de banda vs latencia.

[Relación latencia-ancho de banda]. Unha característica común a todos os sistemas de memoria é a relación entre ancho de banda e latencia. Esta relación ten a forma típica que se amosa na Figura 6. A demanda dun elevado ancho de banda (por parte de moitos fíos) pode saturar o acceso á memoria xa que moitos bancos estarán bloqueados con accesos de diferentes fíos da aplicación. O controlador de memoria tamén constitúe un punto de conxestión. Neste caso cada acceso individual resulta nunha elevada latencia xa que se produce unha serialización no acceso á memoria (colisións entre accesos de diferentes fíos con efecto cola). Este pode ser un dos maiores problemas para os procesadores ñmultinúcleoò con moitos fíos en paralelo. A latencia típica que observa un núcleo é de 50-60ns na actualidade, en condicións de baixa demanda de ancho de banda. Este valor pode chegar a 100-150ns cando a demanda de ancho de banda é elevada.

[**Tipos de chips de memoria**]. Na actualidade a familia de chips de memoria máis utilizada son as de tipo ñDual Data Rateò (DDR), nas súas diferentes xeracións. Estas memorias son de tipo síncrono, onde as operacións estas sincronizadas por un reloxo. Inicialmente as memorias síncronas eran do tipo ñSingle Data Rateò, onde para unha frecuencia F no sistema de conexión exterior, o chip de memoria fai lecturas/escrituras internas a unha frecuencia F, e transferencias de datos ao sistema de conexión exterior tamén a un ritmo de F transferencias por segundo.

A seguinte xeración incorporou o mecanismo DDR, no que se fan transferencias ao sistema de interconexión no flanco positivo e no flanco negativo de reloxo, polo que se acada unha taxa de transferencia de datos dobre a da frecuencia de reloxo. Na xeración DDR1, non se variou a frecuencia F interna do chip de memoria respecto á SDR (para o mesmo tipo de memoria), pero ao dobrarse a taxa de transferencia co exterior, foi necesario que internamente o chip lera ou escribira datos a un ritmo de 2b bytes, sendo

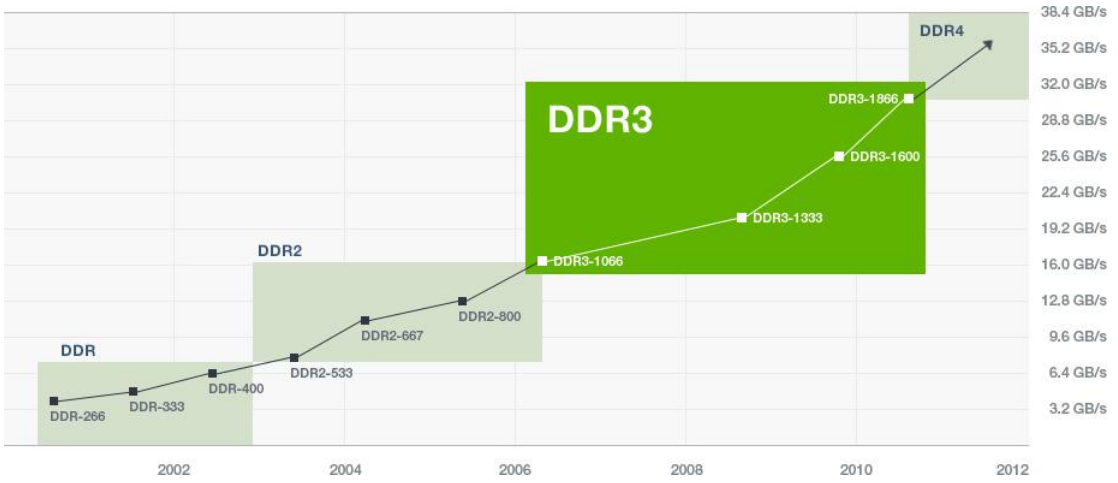

Figura 7: Evolución das memorias DDR (o ancho de banda é para dúas canles de 8 bytes de ancho cada unha).

b o número de bytes que se intercambian co sistema de interconexión exterior.

A seguinte xeración DDR2, dobrou a frecuencia do sistema de conexión exterior con respecto da frecuencia de operación da lóxica no interior do chip. Isto require ler ou escribir datos interiormente a un ritmo de 4b bytes, xa que a taxa de transferencia de datos exterior é 4 veces superior ao ritmo de lectura/escritura interior no chip. Os chips DDR2 están identificados pola taxa de transferencia de datos: DDR2-533 para taxas de 533 MT/s até DDR2-1066 con taxas de 1066 MT/s, a incrementos de 133 MT/s. Por exemplo para DDR2-800, a frecuencia do sistema de interconexión é de 400 MHz, e a frecuencia interior do chip de memoria é de 200 MHz. Os chips tamén se identifican polo ancho de banda que resulta para unha canle de 8 bytes de datos: por exemplo co nome PC2-6400 refírese a chips DDR2-800.

A xeración DDR3, cuadriplicou a frecuencia do sistema de conexión exterior con respecto da frecuencia de operación interior no chip. Isto require ler ou escribir datos interiormente a un ritmo de 8b bytes. Os tipos de chips DDR3 son: DDR3-800, DDR3-1066, DDR3-1333, DDR3-1600, DDR3-1866 e DDR3-2133, é dicir, con frecuencias do sistema de interconexión exterior dende 800 MT/s até 2133 MT/s, a incrementos de 266 MT/s.

A Figura 7 ilustra a evolución das memorias DDR. Como vemos o seguinte paso nesta evolución serán os chips DDR4. É probable que esta xeración sexa a última neste tipo de memorias xa que as taxas de transferencia estarán limitadas a 3.5-4 GT/s.

Existen outras alternativas de chips de memoria para aplicacións máis específicas. Por exemplo está a familia de memorias GDDR, que son utilizadas para o sistema de memoria das tarxetas gráficas. A Figura 8 ilustra a evolución deste tipo de memorias. Na actualidade temos a xeración GDDR5 que pode chegar até unhas taxas de transferencia de 7 GT/s. Para acadar estas taxas de transferencia significativamente máis elevadas que as DDR3, estes sistemas de memoria sacrifican capacidade de memoria por ancho de banda (por exemplo, utilizan un maior número de bancos internos, que implica replicar máis lóxica, o que fai disminuir a cantidade de celas de almacenamento). Utilizan ademais un sistema de transferencia de datos ñquad data

rateò, con catro transferencias por ciclo de reloxo do sistema de interconexión. Isto significa que para transferencia de 7 GT/s é necesario utilizar un reloxo de 1.75Ghz, o que require un deseño bastante sofisticado. Estes chips de memoria transfiren

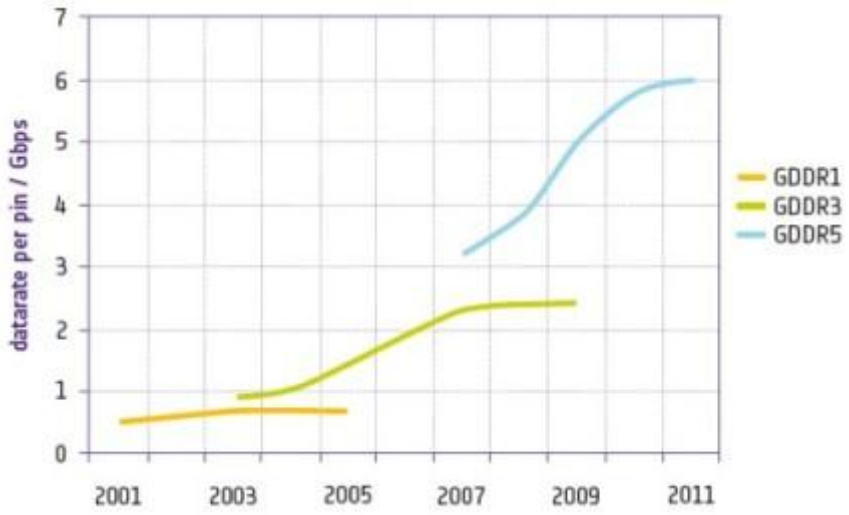

Figura 8: Evolución da familia de chips de memoria GDDR.

tipicamente 4 bytes de datos ao sistema de interconexión. Polo tanto cada chip acada 28GBytes/s de ancho de banda.

As memorias da familia XDR (propiedade intelectual da empresa Rambus) compiten directamente coas memorias GDDR. Na súa versión actual, XDR2, a taxa de transferencia chega até 20 GT/s, con 32 transferencias de datos por ciclo de reloxo do sistema de interconexión. Isto permite uitlizar un reloxo relativamente baixo (comparado con GDDR5) de 625Mhz. Para chips con interface de 4 bytes, acádanse 80 GBytes/s de ancho de banda por chip. De todos os xeitos, este tipo de sistema de memoria é moi sofisticado, con tecnoloxías exóticas para acadar estas elevadas taxas de transferencia.

2- Sistemas de Interconexión

Un aspecto importante para o ancho de banda do sistema de memoria e o ancho de banda de comunicación entre nodos de procesamento é o dos sistemas de conexións. Neste tema trataremos só o nivel físico dunha canle de interconexión para coñecer o máximo ancho de banda do que se pode dispoñer por canle. Noutras materias xa se tratan outros aspectos como a topoloxía das redes de conexión, o encamiñamento, ou o control de fluxo que finalmente delimitan o ancho de banda efectivo.

Tal e como se amosa na Figura 9, unha canle de interconexión consta dun transceiver (circuíto que prepara os datos para ser enviados pola canle, e que pode ter funcións como codificar os datos, cambios eléctricos no sinal, transformación de datos en paralelo a serie, etc) e un receiver (circuíto que recupera os datos que chegan pola canle, e que realiza procesos inversos aos que realiza o transceiver). A canle é un medio físico cun certo ancho de banda (taxa de transmisión da información) limitado debido as súas características físicas, a presenza de ruído eléctrico, etc.

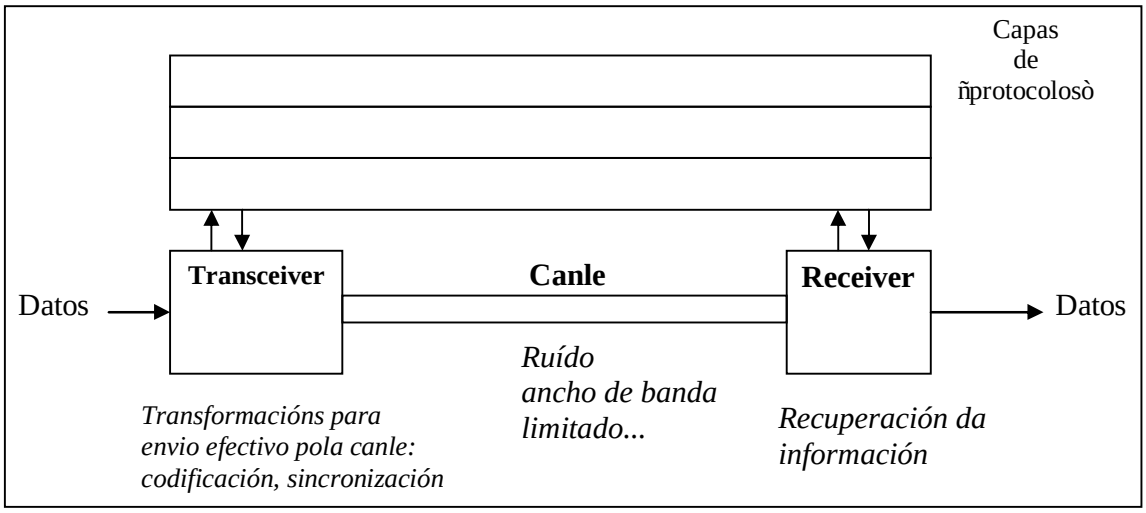

Figura 9: Esquema xenérico dun sistema de interconexión.

Para redes de computadores de altas prestacións a canle física está constituída por cables de cobre (ben illados, con velocidades típicas de propagación de 4ns/m até uns 50 metros de lonxitude), ou por fibra óptica (velocidades de 3.3ns/m, con lonxitudes moi elevadas). Para distancias cortas (dende logo dentro da placa) o cobre adoita ser a opción preferida xa que o ancho de banda máximo que admite non é unha limitación (polo momento) e a fibra óptica presenta unha maior probabilidade de sufrir danos cando conecta nodos moi próximos debido ao trazado irregular que se ten que levar a cabo.

A canle pode ter unha disposición física para transmisión en paralelo ou en serie (ver Figura 10). En paralelo implica que todos os bits da palabra de datos se envían simultaneamente. En serie implica que a palabra de datos transmítese en grupos de bits máis reducidos co ancho de palabra, serializando o envío.

As canles de conexión están caracterizadas polos seguintes parámetros: l (número de liñas de conexión), t (número de transferencias/segundo), h (lonxitude total de cada conexión da canle), m (número de módulos conectados). Tipicamente se l é inferior ao número de bits da palabra a transferir por unha canle, a transmisión será en **serie**. Por outra banda para conexións en paralelo, se m=2, a conexión é **punto a punto**, e se m>2, será unha conexión de **medio compartido**.

Tipicamente as conexións paralelas requiren pouca preparación dos datos para envialos pola canle. Sen embargo requiren que a lonxitude de cada liña de conexión sexa a mesma para axustar ben os tempos de propagación. Esto obriga a conexionados complicados nas placas, sobre todo cando se trata de buses moi anchos.

Tipicamente, as conexións serie requiren unha maior preparación dos datos para envialas pola canle debido a que realizan unha taxa de envíos por segundo moito máis elevada que as conexións paralelas.

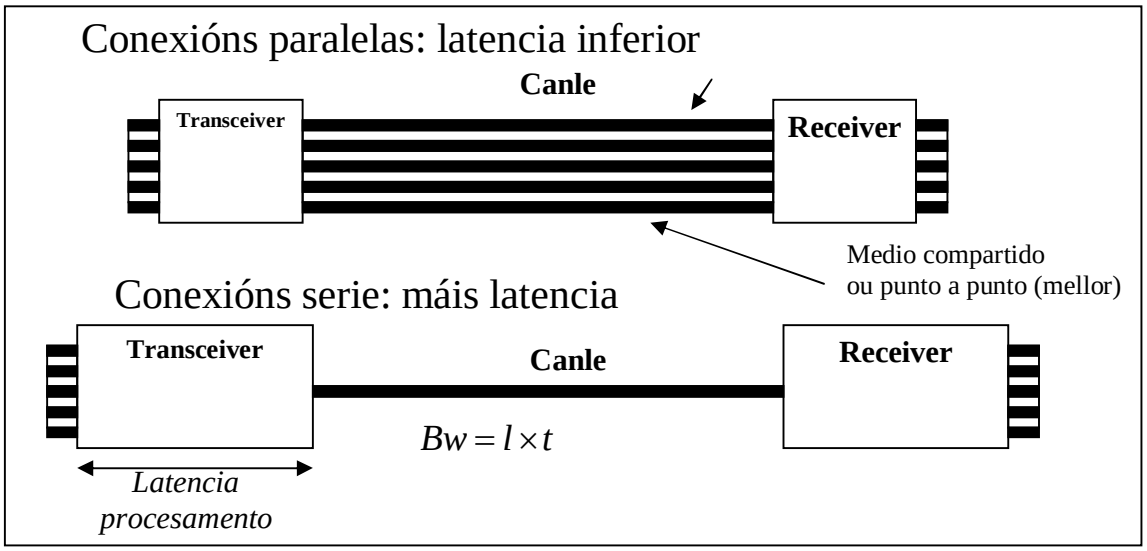

Figura 10: Conexión serie vs paralela.

Actualmente estanse a utilizar técnicas de conexións serie tamén no deseño de conexións paralelas para poder facer conexionados na placa menos complexos con elevadas taxas de transferencia de datos (admiten unha certa diferenza entre as lonxitudes das diferentes liñas de conexión). Isto incrementa a latencia de ditas conexións paralelas.

Para comprender as limitacións das diferentes configuracións, en primeira orde, podemos pensar nunha ecuación do seguinte tipo para relacionar todos os parámetros de deseño:

$$(l \times h)^{\alpha} \times t^{\beta} \times (m-1)^{\gamma} = cte$$

Os coeficientes nos expoñentes modulan a contribución de cada parámetro e o seu valor concreto parece que non se ten estudado. En todo caso con esta expresión podemos sacar algunhas conclusión de interese: se h é elevado (conexións longas, tipicamente conectando placas nun armario ou entre diferentes armarios), unha opción é baixar l (reducido número de liñas) e manter t elevado (a opción de ter l alto non resulta práctica polas dificultades para cablear unha conexión paralela moi longa; isto queda reflectido na ecuación co feito de que l está multiplicado por h co mesmo expoñente). Vemos tamén que para unha lonxitude e número de liñas dada (lxh), o número de transferencias por segundo pode ser máis elevada en conexións punto a punto (m=2) que en medio compartido (m>2, empeorando ao aumentar m; en medio compartido o comportamento dos sinais eléctricos é peor debido as descontinuidades das impedancias, etc). Se h é pequeno, unha opción é utilizar un l elevado (conexión paralela para ter menos latencia) mantendo un valor aceptable de t.

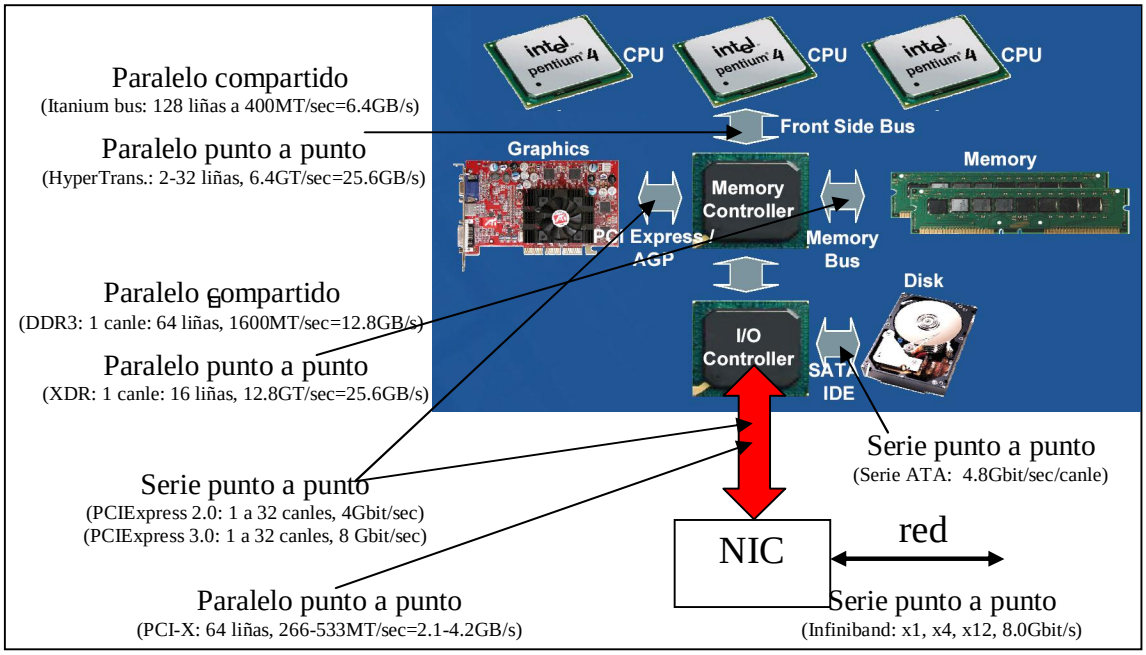

Figura 11: Dominios de aplicación dos diferentes tipos de interconexións.

A Figura 11 amosa un servidor típico fai algúns anos, con diferentes dominios de aplicación para cada un dos tipo de conexións que acabamos de ver.

Para a conexión entre as CPUs e o controlador de memoria as opcións típicas son paralelo compartido (exemplo: bus compartido nos sistemas Itanium de 8 bytes con 400 Megatransferencias/segundo até un máximo de 4 CPUs), ou paralelo punto a punto (exemplo: sistema hypertransport, que pode ter de 2 a 32 liñas, con unha taxa de transferencia de até 6.4 GT/sec por conexión).

O extremo oposto ao anterior estaría dado pola conexión entre nodos de procesamento conectados pola rede. Neste caso utilízanse conexións serie punto a punto (por exemplo Infiniband con ancho de banda por canle (x1) de 8.0Gbit/seg para a versión ñquad data rateò).

A nivel intermedio está PCIExpress, que se utiliza para conexión con diferentes subsistemas dentro da placa: por exemplo para a conexión á tarxeta gráfica ou a conexión as tarxetas de rede. Na actualidade coexisten as especificacións 2.0, na que cada canle bidireccional (x1) ten un ancho de banda de 4 Gbit/sec por dirección (5 GT/sec de ancho de banda de sinal, pero debido ao tipo de codificación, o ancho de banda de datos é de 4 Gbit/sec), e a 3.0, na que cada canle ten un ancho de banda de 8 Gbit/sec por dirección. Poden definirse diferentes anchos: x1, x4, x8, x16 ou x32 canles bidireccionais.

En contraposición á configuración anterior, na Figura 12 amósamos un servidor que representa un nova xeración en canto a súa organización. Os procesadores están comunicados con conexións punto a punto xestionadas por un controlador, que en si

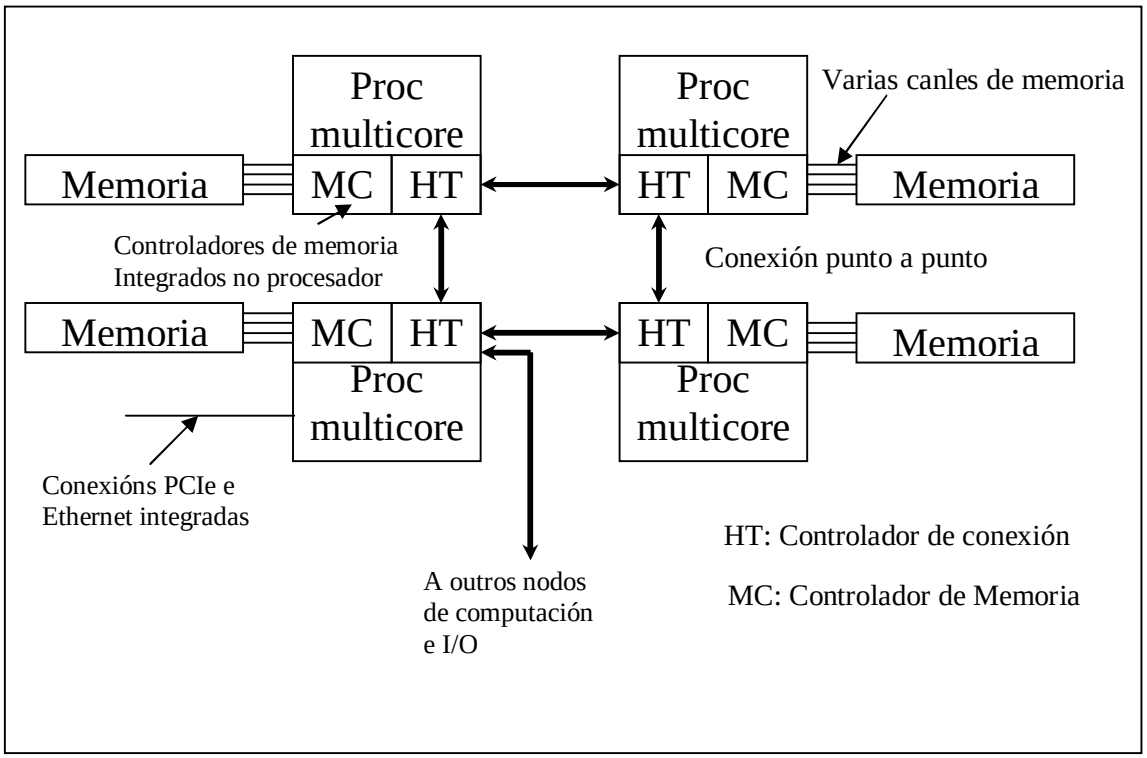

Figura 12: Configuración de elevado rendemento con conexións punto a punto (Hypertransport para o caso de procesadores AMD, é QPI no caso de Intel).

mesmo é un procesador dedicado á comunicación. Cada procesador ten asociado bancos de memoria locais que se acceden a través do controlador de memoria integrado no propio chip co procesador (o procesador pode integrar un ou varios controladores que operan concorrentemente sobre diferentes bancos de memoria). Isto reduce considerablemente a latencia de acceso á memoria local de cada procesador (entre o 40-60%), xa que a transacción entre o procesador e o controlador de memoria non ten que pasar por un bus compartido (que non está sempre dispoñíbel). Os procesadores poden acceder á memoria non local a través dos HT e as conexións punto a punto. Isto dá lugar a un sistema NUMA (Non Uniform Memory Access).

A Figura 13 presenta estes mesmos conceptos pero para un sistema con un só procesador. Como parte das interconexión PICExpress do sistema están directamente integradas dentro do procesador, o que permite unha comunicación de elevado ancho de banda e baixa latencia entre o exterior e a memoria.

Parece que a tendencia nos sistemas de interconexión é a de utilizar canles paralelos punto a punto para interconexión local de baixa latencia, e canles serie punto a punto para interconexións máis longas (sistema de memoria, entrada/saída, e nivel de sistema multiprocesador).

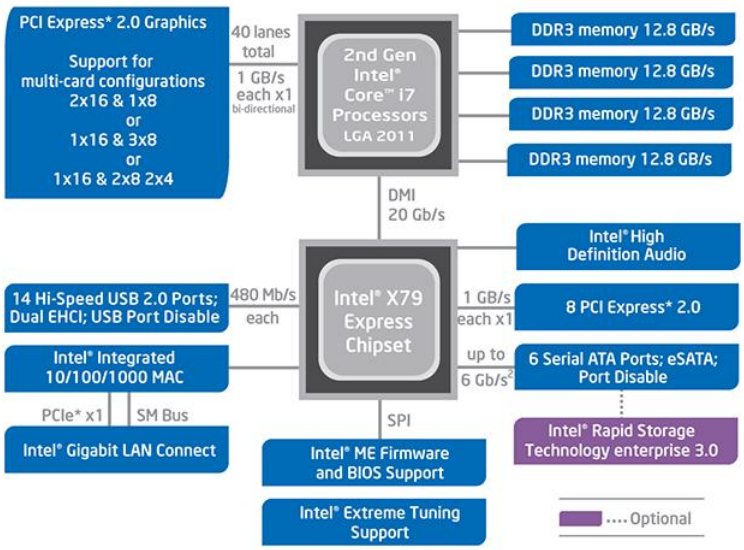

Figura 13: Configuración do sistema de interconexións para un sistema actual cun só procesador.

Na gráfica da Figura 14 amósase o escalamento no tempo do ancho de banda para unha canle serie. Vemos dúas liñas, a liña continua marca a tendencia real de implementacións xa realizadas. A liña descontinua marca o límite de ancho de banda que se pode acadar con circuítos CMOS. Como vemos até fai pouco tempo non se acadara ese límite, e o ancho de banda estaba limitado por outros efectos (control do ruído, etc). A partires do 2008-2010 o efecto limitante pasa a ser a velocidade dos circuítos CMOS. A partir dese momento espérase que o máximo ancho de banda que se poida acadar sexa de entre 2 e 4 FO4/bit. Posto que 1 FO4 escala coa tecnoloxía CMOS a razón de x0.7 por xeración, podemos esperar que no futuro próximo o ancho de banda dos links serie escale por un factor x1.43 de xeración en xeración. Para cables de cobre de alta calidade e distancias típicas a nivel de grandes servidores, o límite de ancho de banda que pode dar o propio medio físico está arredor dos 40Gbit/segundo.

Con respecto as canles paralelas, podemos dicir que con técnicas agresivas e conexións punto a punto tamén se poden acadar os 2-4FO4/bit para conexións entre chips dentro dunha mesma placa.

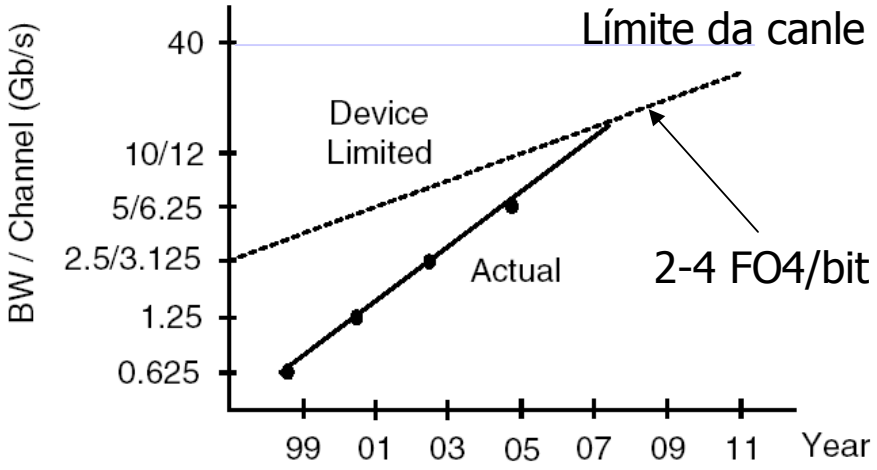

Figura 14: Escalamento do ancho de banda por canle serie.

A modo de exemplo podemos citar as especificacións dalgúns dos sistemas de interconexión punto a punto serie máis populares actualmente:

QPI (Quick Path Interconnect): tecnoloxía de Intel para a interconexión coa memoria e entre procesadores para os vindeiros anos. Na Figura 15 podemos ver que para a interconexión existe unha canle dedicada en cada dirección. Cada unha das canles consta de 21 canles serie (cada canle serie corresponde a 2 cables físicos), para un total de 84 cables físicos. Os datos que se transmiten en cada dirección son 2 bytes. O resto de liñas utilizanse para o reloxo e os códigos de redundancia para garantizar a fiabilidade da transmisión (códigos CRC neste caso). Na súa primeira versión, cada canle utiliza un reloxo de referencia de 3.2GHz (tamén de 2.4GHz; isto irá evolucionando co tempo), facendo a transmisión de datos en cada un dos flancos de reloxo (DDR), dando lugar a unha taxa de transferencia de 6.4 GT/s. Isto da lugar a anchos de banda de 6.4 GT/s x 2 bytes/T=12.8 Gbytes/s por cada dirección (25.6 Gbytes/s en total).

Podemos comparar (ver Táboa 1) as características deste sistema de interconexión punto a punto, con un bus compartido de altas prestacións, en concreto o Front Side Bus (FSB) que utilizaba Intel para interconectar os procesadores entre si e o controlador de memoria.

Tabla 1: Comparación FSB e QPI de Intel.

	FSB	QPI
Número de sinais	150	84
Transferencias/s	1.6 GT/s	6.4 GT/s
Anchura de datos	8 bytes	2 x 2 bytes=4 bytes
Ancho de banda	12.8 GBytes/s	25.6 Gbytes/s

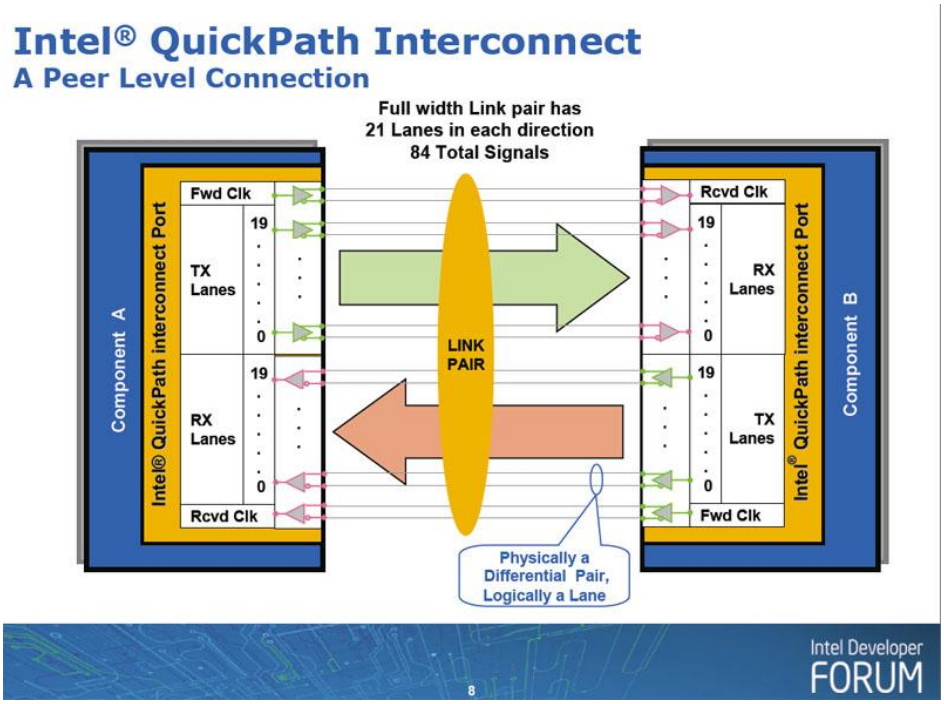

Figura 15: Arquitectura de interconexión QPI de Intel.

Como podemos observar, o sistema QPI permite acadar o dobre de ancho de banda utilizando menos do 60% de sinais. Además a arquitectura QPI incorpora algunhas características interesantes, como modos de operación para baixo consumo de potencia cando o sistema de interconexión non se utiliza. Además permite aillar conxuntos de 5 canles serie por se se produce un fallo de operación. Deste xeito o sistema de interconexión segue funcionando, aínda que con menos canles (menor ancho de banda).

Hypertransport: esta tecnoloxía pertenece a un consorcio creado por varias empresas. É a tecnoloxía de interconexión empregada polos sistemas de AMD, e é semellante a QPI (áinda que Hypertransport foi introducida bastante antes nos sistemas comerciais). Dende o seu comenzo (ano 2001) existen varias xeracións. Por exemplo para a versión 3.1 as frecuencias de reloxio das canles son 2.8GHz, 3.0GHz ou 3.2GHz, utilizando transmisión de datos DDR, polo que as taxas de transferencia de datos poden chegar a 6.4GT/s. As canles constan de 2, 4, 8 16 ou 32 canles serie para datos, segundo a configuración. Polo tanto as prestacións son semellantes ao caso de QPI.

3- Sistema de Memoria

Unha vez revisadas as características das conexións físicas e dos chips de memoria, abordamos agora nesta parte final do tema o sistema de memoria completo cun pouco máis de detalle.

O tipo convencional (con moitas limitacións de escalabilidade cara o futuro) de organización de memoria é o que se amosa na Figura 16. Utilizar un bus compartido para conectar os chips de memoria co controlador de memoria. Os chips de memoria

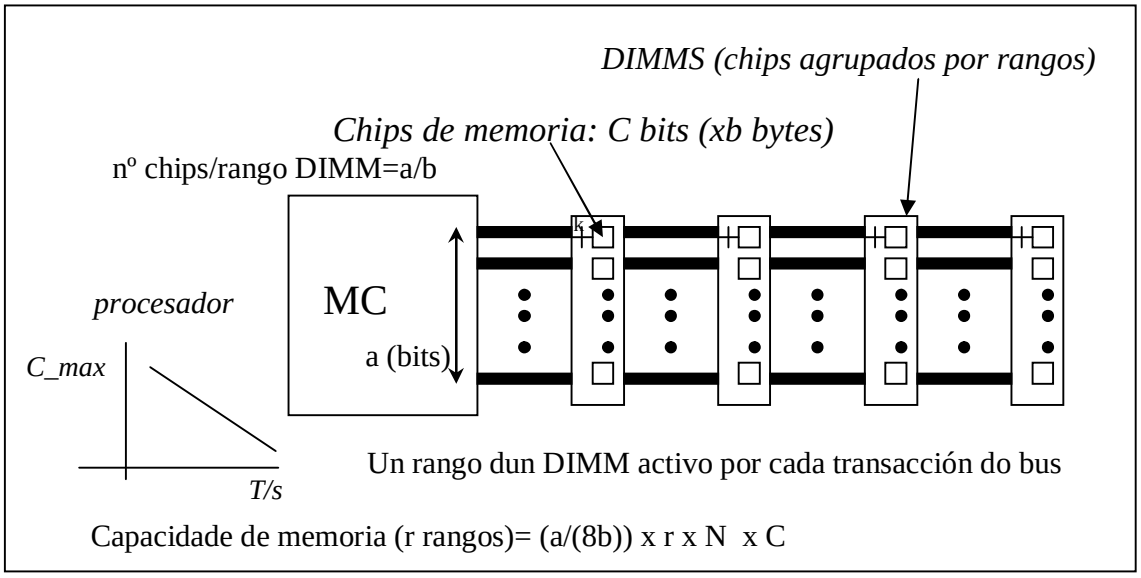

Figura 16: Subsistema de memoria con bus compartido con DIMMS DDR-x.

están montados en placas (DIMMS) dispostas ortogonalmente ao bus. Dentro de cada DIMM, os chips de memoria están organizados en rangos. Cada rango corresponde ao conxunto de chips de memoria que se acceden de xeito simultáneo nunha lectura/escritura por parte do controlador de memoria. Existen DIMMS con un, dous e catro rangos. Por outra banda existen os UDIMM e os RDIMM. Os RDIMM incorporan un rexistro entre o controlador de memoria e os chips de DRAM, o que permite acadar maiores frecuencias no sistema de interconexións, ou incoporar un maior número de DIMMs ao medio compartido, para a mesma frecuencia. A desventaxa, además dun maior coste e consumo de potencia, é que incrementa lixeiramente a latencia de memoria. Esta opción adoita estar dispoñible para servidores, nos que se requiren elevadas capacidades de memoria. Os UDIMM non incorporan este rexistro, polo que están máis limitados en canto ao número que se poden poñer en medio compartido e ancho de banda máximo.

Na Figura 16 utilizamos chips do tipo C bits x b, é dicir, de capacidade de C bits e con b bytes de conexión co medio compartido exterior (8b liñas de conexión). Para un ancho de bus ñaò, o número de chips que temos por rango dun DIMM está dado por a/(8b) para completar todas as liñas do bus. Actualmente os DIMM admiten típicamente uns 16 chips por rango (para datos; en realidade son 18 se considerásemos os bits de paridade para tolerancia de fallos), e poden ter un, dous ou catro rangos (r), multiplexando os chips para o acceso ao bus (só un rango dun DIMM está activo en cada transacción do bus). O ancho de datos para unha canle é habitualmente de 64 bits (72 incluíndo os bits de redundancia). O número de sinais totais que se requiren por canle para conectar co controlador de memoria é de 240. Para unha canle de N DIMMS teremos polo tanto unha capacidade de memoria total de C bits/chip x (a/(8b)) chips x r rangos/DIMM x N DIMMS. A Figura 17 ilustra a conexión dun DIMM co controlador de memoria. Cada chip de memoria aporta un byte ao sistema de conexión, completando unha canle de 8 bytes de datos.

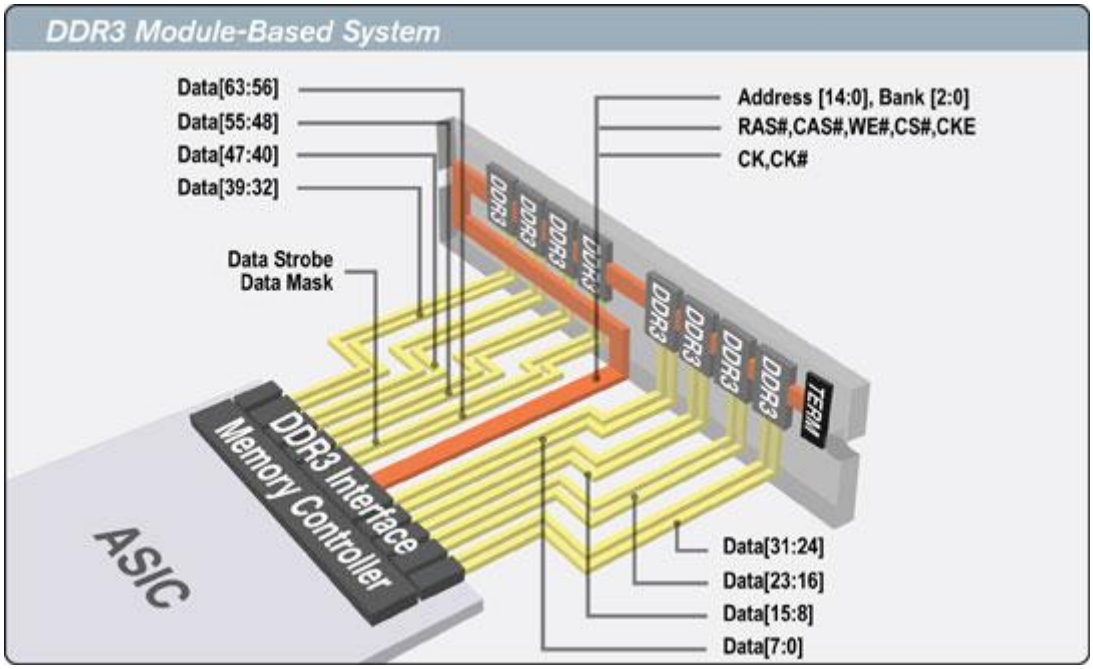

Figura 17: Conexión co controlador de memoria dun DIMM con chips DDR-3 cunha interface de 8 bits por chip (b=1).

Como xa dixemos para aumentar o número de transaccións por segundo no bus compartido é necesario reducir o número de elementos conectados (neste caso reducir o número de DIMMS). Podemos ilustrar estas restriccións cun exemplo específico para o servidor que aparece na Figura 18.

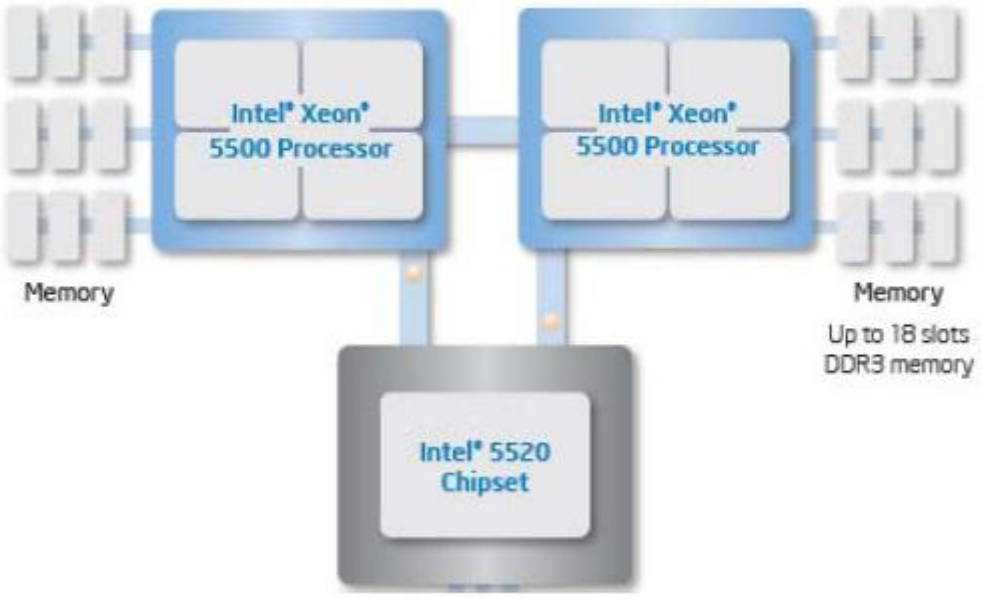

Figura 18: Configuración típica do sistema de memoria nun servidor medio actual.

Tabla 2: Configuracións posibles para o sistema da Figura 42.

Usage	Workloads	Memory Frequency	Max B/W	Recommended # of DIMMs per Channel	Max DIMMs[2]	Max Capacity[3]	CPU
Maximum bandwidth	HPC	1333 MHz	32 GB/s	1	6	48GB	X5570 X5560 X5550
General purpose	Various apps	1066 MHz	25.5 GB/s	2	12	96GB	X5570 E5540 X5560 E5530 X5550 E5520
Maximum capacity	Virtualized platforms	800 MHz	19.2 GB/s	2 QR[O] RDIMM 16GB	12	192GB	All
				3 SR/DR RDIMM 8GB	18	144GB	

[O]SR=Single rank DR=Dual rank QR=Quad rank

O servidor conta con dous procesadores (quadcore), donde cada un deles consta de tres canles de memoria de medio compartido (un controlador de memoria integrado no chip con tres canles de memoria). En total permite até 18 DIMMs de memoria tipo DDR3. Este sistema exemplo soporta varias configuracións, dependendo do balance entre tamaño da memoria e ancho de banda. Na Táboa 2 amósanse as diferentes configuracións posibles. Por exemplo, para máximo ancho de banda utilizaríamos DIMMs con chips DDR3-1330, con un máximo de un DIMM por canle, co que para 6 canles permitiría unha capacidade de memoria de até 48 Gbytes utilizando DIMMs de 8Gbytes (por exemplo dous rangos con chips de 2Gbit: 2 x 16 x 2Gbit/8=8Gbytes). Con esta configuración, cada canle ten un ancho de banda pico de 10.66 Gbytes/s, co que cada CPU ten un máximo de 32 Gbytes/s de ancho de banda coa memoria. Na configuración de capacidade máxima, pode chegar até 192 Gbypes, pero cun ancho de banda por CPU de até 19.2 Gbytes/s.

Unha alternativa para aumentar a capacidade de memoria mantendo elevadas taxas de transferencia é a utilización de chips búfer para expansión de DIMMs na placa base. A Figura 19 ilustra esta técnica, na que se introducen chips BoB (Buffer-on-board) que permiten que cada canle de memoria soporte catro DIMMS no canto de dous. Para isto o chip BoB multiplexa os requirimentos do controlador de memoria entre dúas ramas con dous DIMMs cada unha. O servidor da Figura 19 ten dous controladores de memoria, cada un dos cales soporta dúas canles, e cada canle desdóbrase en dúas ramas nos chips BoB, con dous DIMMs por rama, polo que en total admite até 16 DIMMS (128 Gbytes de memoria con DIMMS de 8 Gbytes).

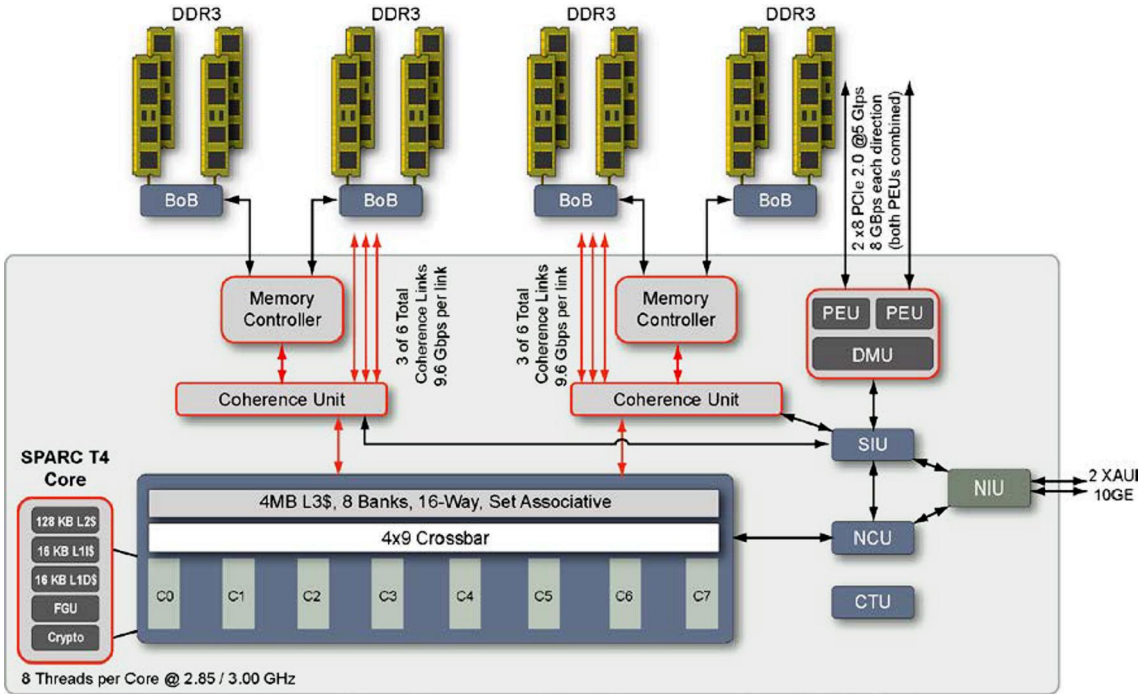

Figura 19: Sistema de Memoria do procesador Oracle Sparc T4, con BoB (Buffer-on-Boardò para aumentar a capacidade de memoria.

A estratexia de utilizar chips búfer ten o inconveniente de ter que incorporar máis chips na placa base, polo que se require máis espazo, e además un maior consumo de potencia. Esta solución non é escalable, e representa simplemente unha extensión do modelo convencional de medio compartido sen demasiadas posibilidades de continuidade no futuro.

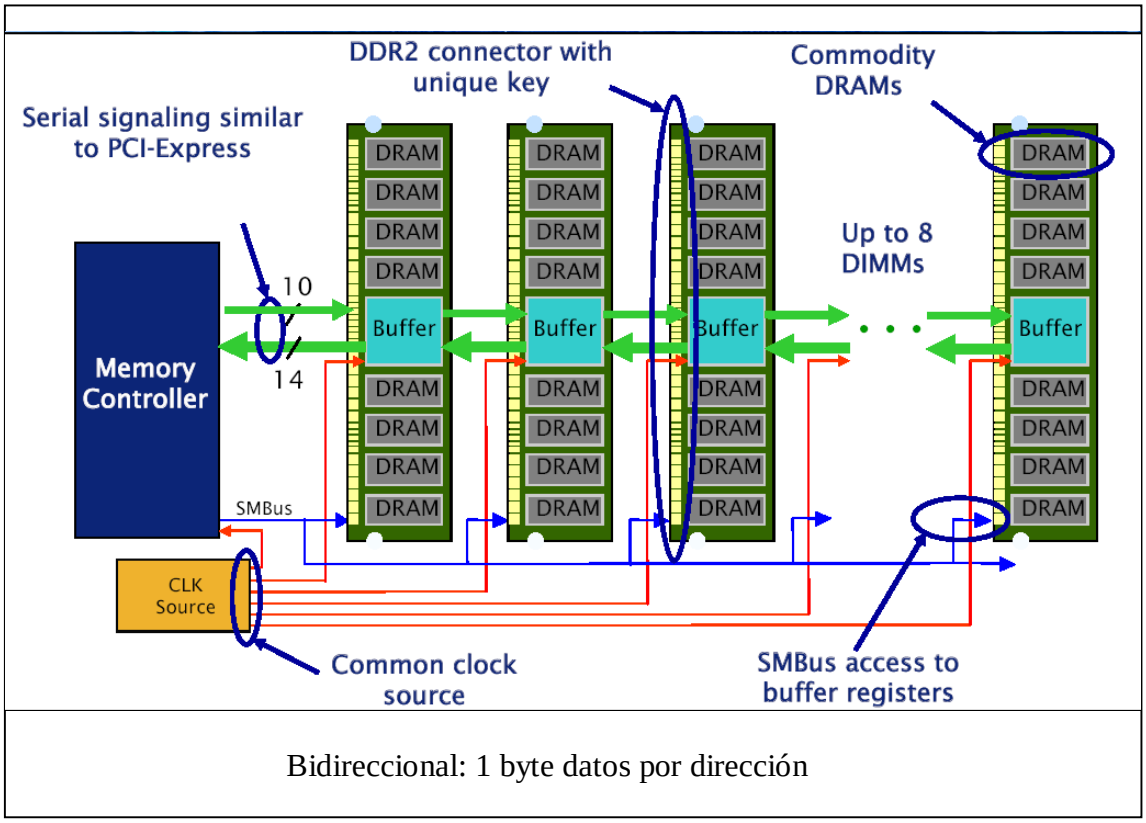

Figura 20: Configuración FB-DIMM.

Outra alternativa para compatibilizar unha memoria de elevada capacidade e un elevado ancho de banda é a organización de memoria **FB-DIMM** (DIMMS con buffering). Este sistema, que se ilustra na Figura 20, utiliza conexións punto a punto serie (sistema semellante a PCI Express, con velocidades do link de 3.2Gb/s, 4.0 e 4.8 Gb/s na especificación actual) con buffering (este é un chip que se denomina AMB: Advanced Memory Buffer) entre os DIMMS. A conexión é bidireccional con unha canle para lectura de 14 conexións serie punto a punto, e outra para escritura de 10 (cada conexión require dous cables físicos). O número total de conexións que se requiren co controlador de memoria é de 69. O sistema soporta até 8 DIMMS por canle. O aumento simultáneo de ancho de banda e capacidade prodúcese a costa dun moderado incremento na latencia da memoria debido ao buffering (chip AMB). Un problema máis significativo é o elevado consumo de potencia, sobre todo debido a introducción do sistema de buffering e as conexións serie con elevada taxa de transferencia. Unha característica interesante deste sistema é que permite utilizar chips de memoria DDR convencionais (DDR2, DDR3).

Vexamos o xeito en que se realizan as lecturas e as escrituras, e as prestacións que se obteñen en termos de ancho de banda. En canto as lecturas, ao longo de 12 ciclos do sistema de interconexión faise a lectura do que se chama un ñframeò (que poderíamos traducilo por paquete), que corresponde ao equivalente de dúas lecturas nun DIMM tipo DDR convencional con bus compartido. Específicamente con 14 liñas de conexión para lectura, ao longo de 12 ciclos temos un total de 14 x 12=168 bits. Destes 168 bits, 24 corresponden a códigos de redundancia CRC para garantizar a fiabilidade das

comunicacións, co que 144 bits son de datos (72 bits é o que corresponde a cada lectura nun DIMM DDR convencional, con 64 bits de datos e 8 de paridade).

Para estimar o ancho de banda de lectura e a relación que hai entre o tipo de chips de memoria e a taxa de transferencia das conexións serie, vamos a supoñer un caso concreto. Especificamente supoñamos canles serie de 4.8 GT/s. Os chips de memoria deberían proporcionar dúas lectura de rango en 12 ciclos. Polo tanto para chips DDR o ciclo de reloxio que rixe a transferencia de datos dende o chip de memoria ao exterior (con dúas transferencias por ciclo) debe ser de 4.8 GHz/12=400 MHz. É dicir, os chips DRAM deben ser por exemplo do tipo DDR2-800 (transferencias as 800 MT/s). O ancho de banda de lectura é de 16 bytes de datos cada 12 ciclos, é dicir 16 x 4.8/12=16 x 0.4 GT/s=6.4 GBytes/s por cada canle.

Para escrituras o paquete (ñframeò) corresponde tamén a 12 ciclos, no que se transmite un total de 10 x 12 =120 bits. Neste caso 22 bits son de código CRC e 98 de datos ou comandos. Os 98 bits distribúense do seguinte xeito: utilizanse 2 bits para indentificar o tipo de paquete, 24 bits para un comando (3 bits dos cales son para identificar o DIMM), e unha das seguintes alternativas: i) 72 bits para datos (con paridade) ou ii) dous comandos adicionais de 24 bits ou iii) un comando de 24 bits e 36 bits de datos (a metade do ancho dun rango). Para o caso no que se envía un comando por paquete e 8 bytes de datos (sen contar os de paridade), o ancho de banda de escritura por canle é de 8 x 4.8/12=3.2 GBytes/s.

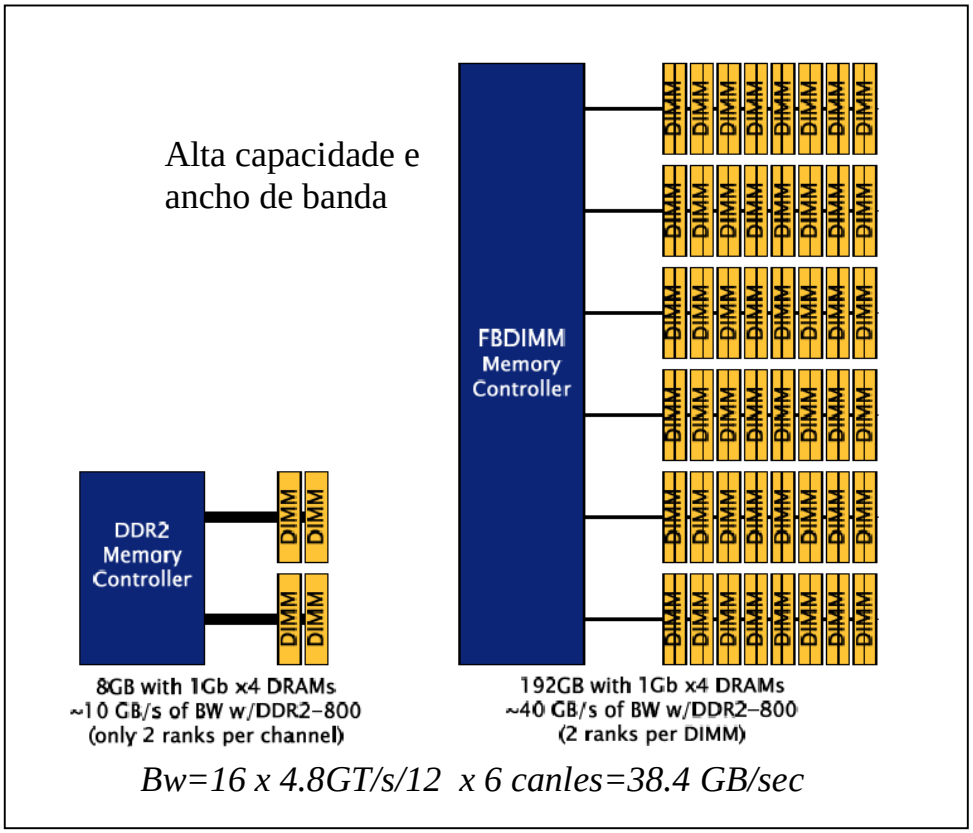

Figura 21: Comparación FB-DIMM vs DDR.

A Figura 21 ilustra a comparación do sistema de bus compartido (con chips DDR2-800) co sistema con conexións punto a punto e buffering. A configuración con medio

compartido presenta dúas canles de 8 bytes de datos cada unha. Cada canle soporta 2 DIMMS, con chips só en un rango. A capacidade total está dada por 4 DIMMS x 1 rango x 16 chips/rango x 1Gbit/8= 8GBytes (de datos sen ter en conta os bits de paridade). No que atinxe ao ancho de banda temos 16 bytes a un ritmo de 800 MT/segundo=12.8GBytes/segundo. O número de conexións que se require por cada bus DDR é de 240, polo que para as dúas canles son necesarias 480 conexións cos dous controladores de memoria.

Con un número semellante de pads de conexión cos controladores de memoria (6 x 69=414), a configuración con FBDIMM que se mostra ten 6 canles cunha capacidade de até 192 GBytes con chips de 1Gbit x4 (6 canles x 8 FBDIMM x 2 rangos x 16 chips/rango x 1Gbit/8). O ancho de banda de lectura para links de 4.8Gb/s é de 16 bytes/frame x 4.8 GT/s/12 T/frame x 6 canles=6.4 GBytes/s x 6=38.4 GBytes/sec con chips de memoria DDR2-800. Este ancho de banda irá medrando a medida que escale a a taxa de transferencia dos links serie.

As memorias FBDIMM non teñen penetrado moito no mercado, e polo momento o seu elevado consumo de potencia parece limitar o seu uso xeneralizado. Por outra banda, os microprocesadores actuais tenden a integrar varios controladores de memoria dentro do chip do procesador. Isto fai que, de momento, os sistemas convencionais de medio compartido dominan o mercado, xa que satisfacen os requirimentos de ancho de banda e capacidade de memoria.

Memorias orientadas a alto ancho de banda: xa mencionamos cando falamos dos chips de memoria, que existen variantes como as memorias GDDR5 ou XDR2 que están orientadas a obter un elevado ancho de banda, sacrificando para iso capacidade de memoria. A aplicación típica é a do sistema de memoria dos procesadores gráficos. O ancho de bus de interconexión co controlador de memoria por canle é bastante máis elevado que nos sistemas DDR. Actualmente os achos están entre 256 a 512 bits. O bus paralelo é realidade un sistema de conexión punto a punto, polo que as liñas só está conectadas a un chip de memoria. Deste xeito para expandir a memoria é necesario contar con buses máis anchos e chips de maior capacidade. Os chips teñen unha interface típica de 32 bits, polo que para un ancho de 256, o número de chips está limitado a 8 (16 para anchos de 512). Con chips de 2Gbit (x32), o sistema de memoria ten unha capacidade de 2 GBytes para anchos de 256, e 4 GBytes para anchos de 512 bits. Con respecto aos anchos de banda GDDR5 chega até 7GT/seg, polo que acada entre 224GBytes/seg para anchos de 256 (256/8 x 7) e 448 GBytes/seg para anchos de 512. Con XDR2, con taxas de transferencia extremas de 20 GT/seg, pode sobrepasarse 1 TByte/seg (20 GT/seg x 512/8=1,280TBytes/seg).

4- Futuro do Sistema de Memoria

O sistema de memoria presente moitas cuestións abertas cara o futuro. Como xa mencionamos resulta complexo escalar a capacidade de memoria e ao mesmo tempo escalar o ancho de banda. Por outra parte, o escalamento de ancho de banda fai que o sistema de memoria aporte un consumo de potencia ao sistema cada vez máis significativo.

[**Escalamento da capacidade de memoria**]: a capacidade de memoria está determinada polo número de canles de memoria, o número de módulos DIMM por canle, o número de chips de memoria por DIMM e a capcidade de almacenamento dos chips de memoria. Coas tecnoloxías actuais, escalar en ancho de banda implica reducir o número de DIMMs a un por canle e diminuir o número de chips por DIMM a un rango. Polo tanto para escalar a capacidade de memoria é necesario escalar a capacidade dos chips e o número de canles. Escalar o número de canles está limitado polo número de sinais de entrada/saída que poden ter os encapsulados para os microprocesadores. O escalamento da capacidade de almacenamento dos chips depende do escalamento da tecnoloxía, pero como xa dixemos a tecnoloxía actual DRAM non escala ben, polo que é previsible que non se poida manter a tendencia de dobrar a capacidade cada tres anos.

[**Consumo de potencia]**: A disipación de potencia nos chips de memoria ten varios compoñentes: potencia consumida no array de almacenamento, potencia consumida polos circuítos de entrada/saída e potencia polo camiño de datos entre a celas de almacenamento e os circuítos de entrada/saída. Actualmente un 50% da potencia consúmeo o camiño de datos, e o outro 50% está divido a metade entre o array de almacenamento e os circuítos de entrada saída. O consumo de potencia dos chips de memoria de tipo DDR-x está no rango **40-200 mW/Gbps** (miliwatts/Gigabit/seg). Con respecto aos circuítos de entrada/saída, os consumos son de **10-15 mW/Gbps** para memorias DDR e 20-25 para memorias GDDR5. Deste xeito un módulo DIMM DDR3 a 1.6 GT/seg, tería un consumo total de 40 mW/Gps x 1.6GT/seg x 64 bits/transferencia = 4,1 Watts. Para dar unha idea da dimensión deste valor, se un sistema multinúcleo pretendese acadar 1Tbyte/seg de ancho de banda de memoria, o consumo do sistema de memoria sería 40 mW/Gbps x 8000 Gbps =320 Watts. Evidentemente a tensión de alimentación pode seguir escalando, pero actualmente está no rango 1.5-1.35 V para DDR3, e no parece doado que baixe de 1.2V, polo que a reducción por escalamento da tensión de alimentación non pode ser considerada a única solución.

[**Solucións**]: parece que todas das solucións aos problemas do sistema de memoria pasan pola tecnoloxía de integración 3D. De feito varias multinacionais da industria de semiconductores está colaborando en desenvolver o que denominan ñHybrid Memory Cubeò. A Figura 22 ilustra este concepto. Consite en integrar varios chips de memoria nun mesmo encapsulado 3D, e con un chip de lóxica para facer de interface co exterior. A comunicación entres os chips de DRAM e o chips de lóxica de interface é a través de vías (TSV na figura) que conectan un chip con outro en dirección vertical. Coa integración 3D preténdese conseguir moita máis capacidade de memoria nun espazo reducido e cunha reducción do consumo de potencia moi importante. Por outra banda estase a traballar moi activamente en obter interfaces co sistema de interconexión que acaden eficiencias de 1mW/Gps a velocidades de transferencia de 10GT/seg, para o que se investiga tanto en novas técnicas a nivel de circuíto coma nos materiais e topoloxía con conforma a propia canle de interconexión. Por exemplo a utilización da tecnoloxía 3D permitirá ocupar un espazo moi inferior o que facilita conexións de lonxitude reducida entre o sistema de memoria e o procesador.

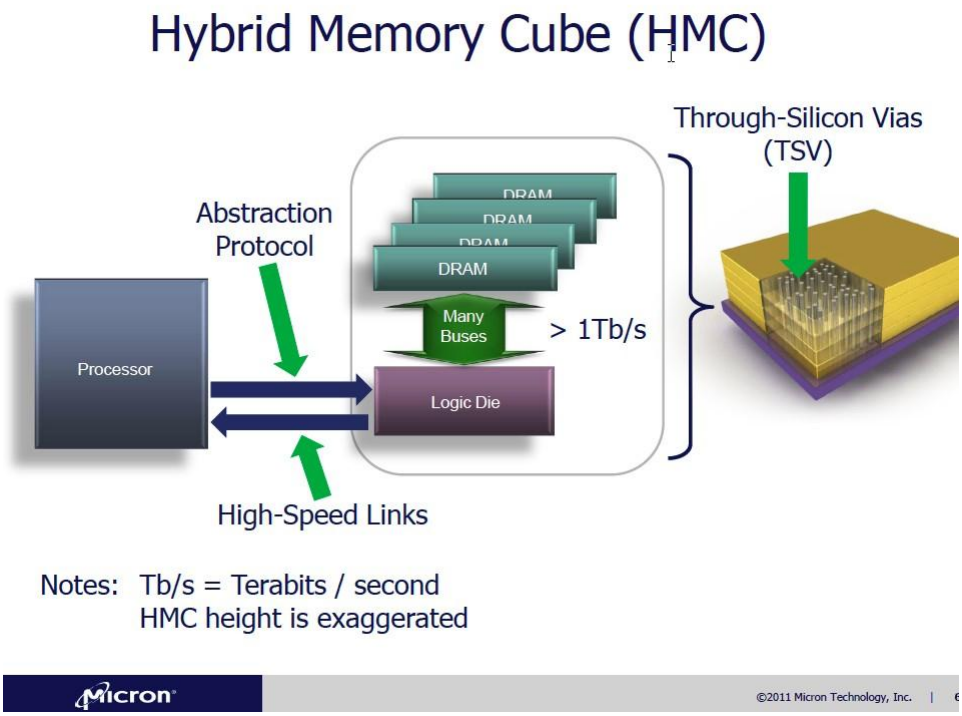

Figura 22: Concepto de ñHybrid Memory Cubeò: integración de varios chips en tecnoloxía 3D.