Titán Cray XK7

Titan es un superordenador construido por Cray en el Laboratorio Nacional de Oak Ridge para su uso en una variedad de proyectos de ciencia. Titan es una actualización de Jaguar , un superordenador anterior en Oak Ridge, que utiliza unidades de procesamiento gráfico (GPU), además de convencionales unidades de procesamiento central (CPU). Titán es el primer híbrido para llevar a cabo más de 10 petaFLOPS. [1]

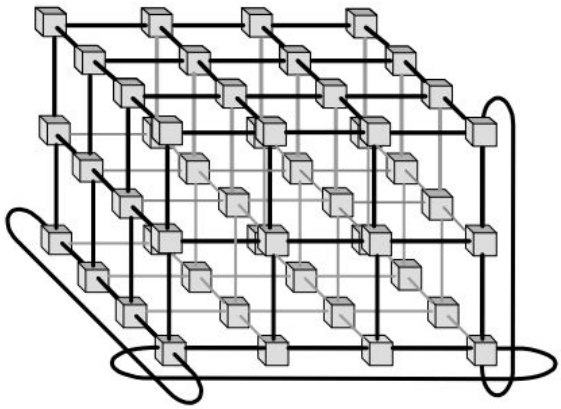
Titan tiene 18.688 nodos (4 nodos por módulo, 24 módulos por armario), conteniendo cada uno un 16-core AMD Opteron 6274 CPU con 32 GB de DDR3 memoria ECC y un K20X GPU Nvidia Tesla con 6 GB GDDR5 memoria ECC. Hay un total de 299,008 núcleos de procesador, y un total de 693,6 TiB de CPU y la RAM GPU. [2]

SOBRE LA RED:

La mayoría de trabajos son ejecutados remotamente, por ello la conectividad es tan importante como la computación. Hay docenas de enlaces de 10GbE [3] a la máquina, y está conectada a la "DoEs Energy Science Network" (ESNET) con 100Gbps.[4]

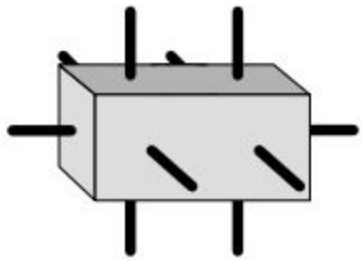
Gemini es la nueva red para los sistemas de supercomputación de Cray. Se desarrolla el diseño altamente escalable Seastar utilizado para entregar el sistema 225.000 núcleos Jaguar, con la mejora de la funcionalidad de red, la latencia y velocidad de emisión. Gemini utiliza un diseño novedoso sistema-en-chip (SoC) para construir redes torus 3D directos que se pueden escalar a más de 100.000 nodos de múltiples núcleos. Géminis está diseñado para ofrecer un alto rendimiento en aplicaciones MPI y el tráfico de sistema de archivos, además de que proporciona soporte de hardware para la programación del espacio de direcciones global. Gemini permite la implementación eficiente de los lenguajes de programación como capilla, UPC y Co-Array Fortan en sistemas masivamente paralelos.

Figura 1 Cray XT Red Torus 3D



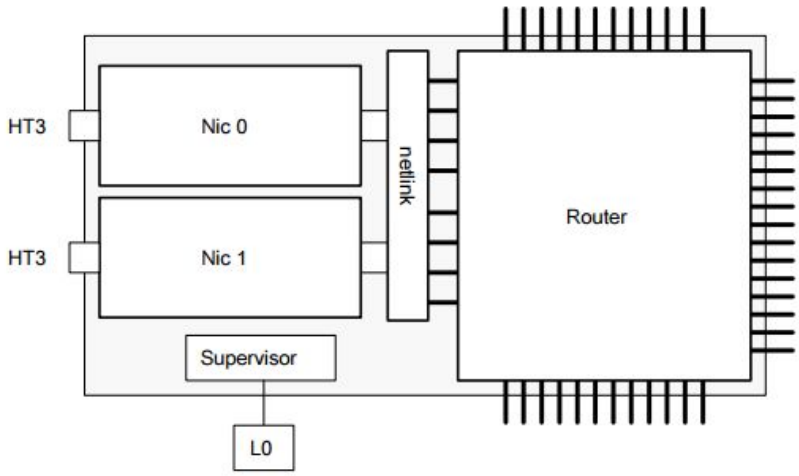
[1] https://en.wikipedia.org/wiki/Titan_(supercomputer)
[2] http://www.webcitation.org/6FPlkwAkf
[3] https://en.wikipedia.org/wiki/10_Gigabit_Ethernet
[4] https://web.archive.org/web/20130125142636/http://www.anandtech.com/show/6421/inside-the-titan-supercomputer-299k-amd-x86-cores-and-186k-nvidia-gpu-cores

Cada ASIC Gemini proporciona dos controladores de interfaz de red (NIC), y un enrutador de 48 puertos “YARC”. Cada
una de las tarjetas de red tiene su propia interfaz de host HyperTransport 3, lo que permite a Gemini conectar dos
nodos Opteron a la red. Este bloque de construcción proporciona 10 enlaces torus, 4 cada uno en dos de la
dimensión ('x' y 'z') y 2 en la tercera dimensión ('y'), como se muestra en la siguiente figura. El tráfico entre los dos nodos conectados a un solo Gemini se enruta internamente. El router utiliza un diseño de baldosas, con 8 azulejos dedicados a los NICs y 40 (10 grupos de 4) dedicados a la red.



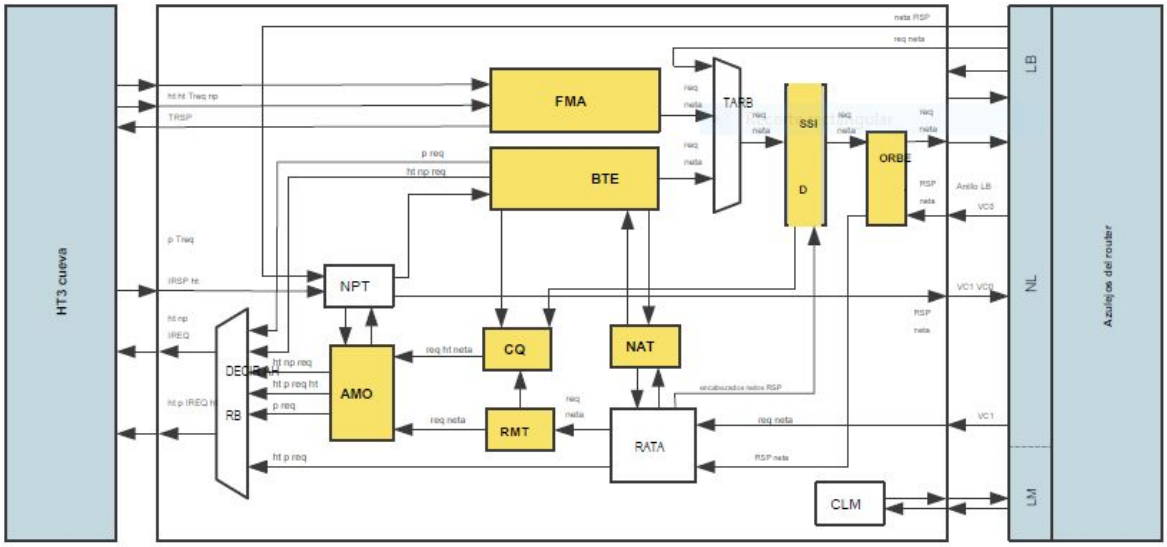
Geminis

La estructura de bloques del diseño Gemini se ilustra en la figura siguiente . El bloque Netlink conecta las tarjetas de red al router. También se ocupa de los cambios en la velocidad de reloj entre la NIC y dominios del router. El bloque de supervisor conecta Gemini a un procesador de control incrustado (L0) por lo tanto, la red Cray supervisora de sistema de hardware (HSS) se utiliza para supervisar el dispositivo y la carga de sus tablas de encaminamiento.



Gemini NIC

Cada ASIC Gemini tiene un par de tarjetas de red, cada uno con su propia interfaz HyperTransport 3.La NIC es una tubería de hardware. El nodo emite comandos, escribiendo ellos a través de la interfaz HyperTransport 3. Los ASIC estan formados por los siguientes módulos, los cuales se muestran en la siguiente imagen[5].



El Gemini NIC funciona a 650 MHz, el router a 800 MHz y los SERDES enlace en la 3.125GHz. La velocidad de la interfaz HyperTransport va de 1600 MHz a 2600MHz dependiendo del tipo de nodo

Fast Memory Access (FMA)

Fast Memory Access (FMA) es un mecanismo mediante el cual los procesos de usuario generan las transacciones de red, realizando operaciones de memoria atómicas (AMO), almacenandolas directamente en el bloque NIC correspondiente. FMA proporciona baja latencia y alta tasa de emisión de transferencias pequeñas por lo tanto el módulo FMA se suele utilizar para transferencias pequeñas.

Block Transfer Engine (BTE)

bloque (BTE) admite transferencia asíncrona entre memoria local y remota. El Software escribe descriptores de transferencia en bloques a una cola y el hardware Gemini se encarga de realizar las transferencias de forma asíncrona. El BTE es compatible con las operaciones de memoria (PUT / GET), donde el usuario especifica una dirección local, una dirección de red y un tamaño de transferencia. En los módulos BTE las transferencias tardan más en empezar, pero una vez están en funcionamiento puede transferir grandes cantidades de datos (hasta 4 GB) sin la intervención de la CPU.

[5] https://wiki.alcf.anl.gov/parts/images/2/2c/Gemini-whitepaper.pdf

Completion Queue (CQ)

Colas de terminación proporcionan un mecanismo ligero de notificación de eventos. La finalización de una transacción BTE o FMA puede generar un evento en una cola específica de usuario (o de hilos de núcleo). eventos de terminación pueden ser generadas en el nodo de origen o el nodo de destino. Se incluyen tanto los datos de usuario y la información de estado de la transacción.

Atomic Memory Operation (AMO)

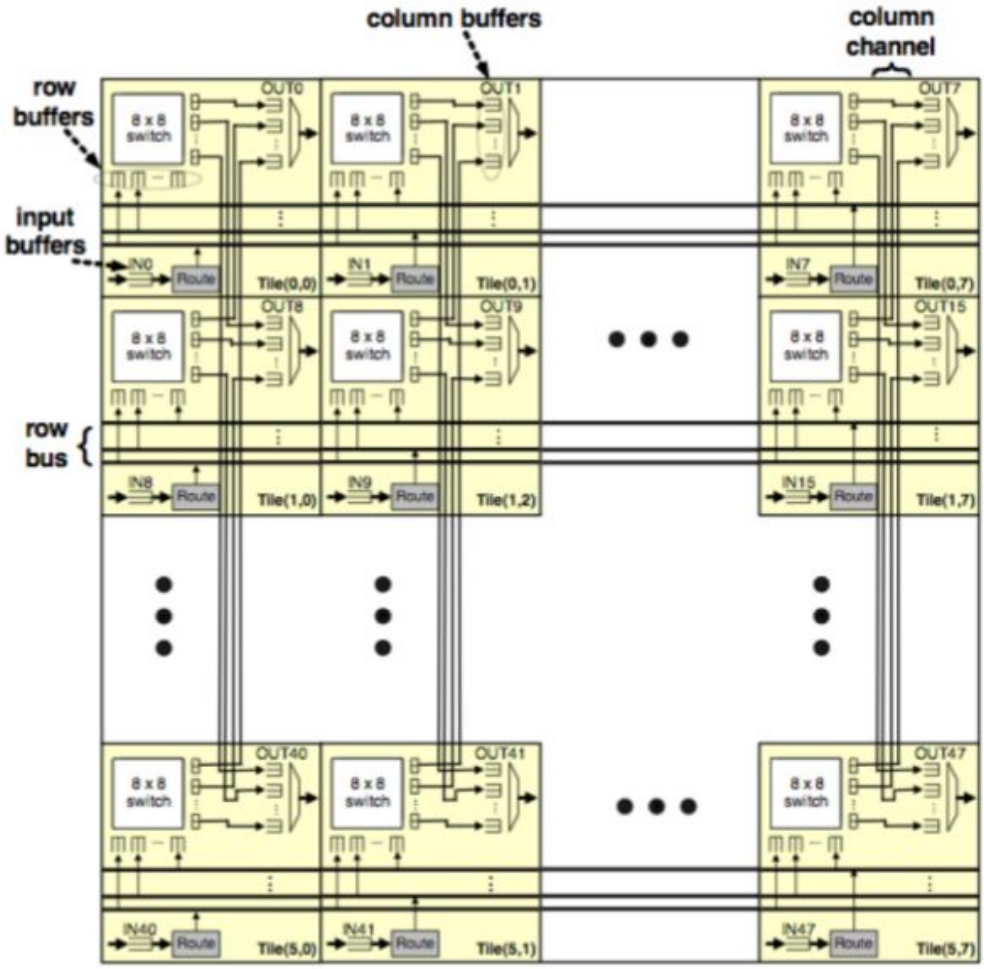
Gemini es compatible con una amplia gama de operaciones atómicas tales como suma atómica o intercambio condicional. Gemini mantiene una caché AMO, reduciendo la necesidad de leer de la memoria del host cuando varios procesos acceden a la misma variable atómica. la memoria del host se actualiza cada vez que se actualiza la variable (también dispone de mecanismos de actualización para reducir la carga en el host), pero dado que la red atómica no es coherentes con respecto a las operaciones de memoria local todos los procesos deben utilizar la interfaz de aplicación Gemini para actualizar una variable atómica.

Synchronization Sequence Identification

Gemini utiliza un mecanismo conocido como identificación de secuencia para seguir el conjunto de paquetes que componen una transacción. Cada paquete de la secuencia contiene la misma Identificación de Secuencia de sincronización (SSID). Los paquetes pueden ser entregados en un orden arbitrario ya que cada uno contiene una dirección de red y puede ser ubicado memoria tan pronto como llega por lo que no hay necesidad de utilizar un almacenamiento temporal para la reordenación de los paquetes recibidos. Este mecanismo se implementa utilizando los bloques de SSID y de Output Request Buffer(ORB) en el lado de salida y el bloque Receive Message Table(RMT) en el lado de entrada. El caché RMT activa el estado SSID evitando un viaje de ida y vuelta en la red para operaciones donde el rendimiento es critico.

ROUTER GEMINI

El bloque de construcción del router Géminis es el siguiente (ver siguiente Figura). Cada mosaico contiene toda la lógica y el almacenamiento en búfer asociado con un puerto de entrada, un puerto de salida, un 8 × 8 switches y sus búferes asociados. En Gemini, cada módulo acepta entradas desde seis hileras de buses que conducen por los puertos de entrada en esas filas y separa los canales de salida a los ocho puertos de salida en su columna. Usando una microarquitectura mosaico facilita la implementación, ya que cada azulejo es idéntico y produce una estructura muy regular para la replicación y la aplicación física en silicio.



El diseño basado en la baldosa se entiende mejor después de ver como circula un paquete a través del router. Un paquete llega al enlace de entrada de una baldosa. Cuando el paquete llega a la cabeza de la memoria intermedia de entrada, se toma una decisión de enrutamiento para seleccionar la columna de salida para el paquete. El paquete es

entonces impulsado en el bus de su fila asociada con el puerto de entrada y se almacena en un buffer de fila en la entrada de la 8 × 8 en la unión de la columna de la fila de entrada y de salida del paquete. En este punto, la decisión de enrutamiento debe ser refinado para seleccionar un puerto de salida particular dentro de la columna de salida. El interruptor entonces encamina el paquete al canal de la columna asociado con el puerto de salida seleccionado. El canal de la columna entrega el paquete a un búfer de salida (asociado con la fila de entrada) en el multiplexor del puerto de salida. Los paquetes en las memorias intermedias de salida se arbitran para el acceso al puerto de salida y, cuando se concede el acceso, se conmutan en el puerto de salida a través del multiplexor.

La tolerancia de fallos

Gemini proporciona un CRC de paquetes de 16 bits, que protege hasta 64 bytes de datos y los encabezados asociados (768 bits max). Dentro de cada Géminis, grandes recuerdos se protegen utilizando ECC. Los enlaces Gemini proporcionan una entrega fiable utilizando un protocolo de ventana deslizante. El enlace de recepción comprueba el CRC como llega un paquete, y devuelve un error si no es correcta. El enlace retransmite el envío a la recepción de un error. El bloque de conexión incluye un buffer de envío de tamaño suficiente para cubrir el viaje redondo. El CRC se comprueba también como un paquete de hojas de cada Gemini y en la transición desde el router a la NIC, permitiendo la detección de errores que se producen dentro del núcleo router.