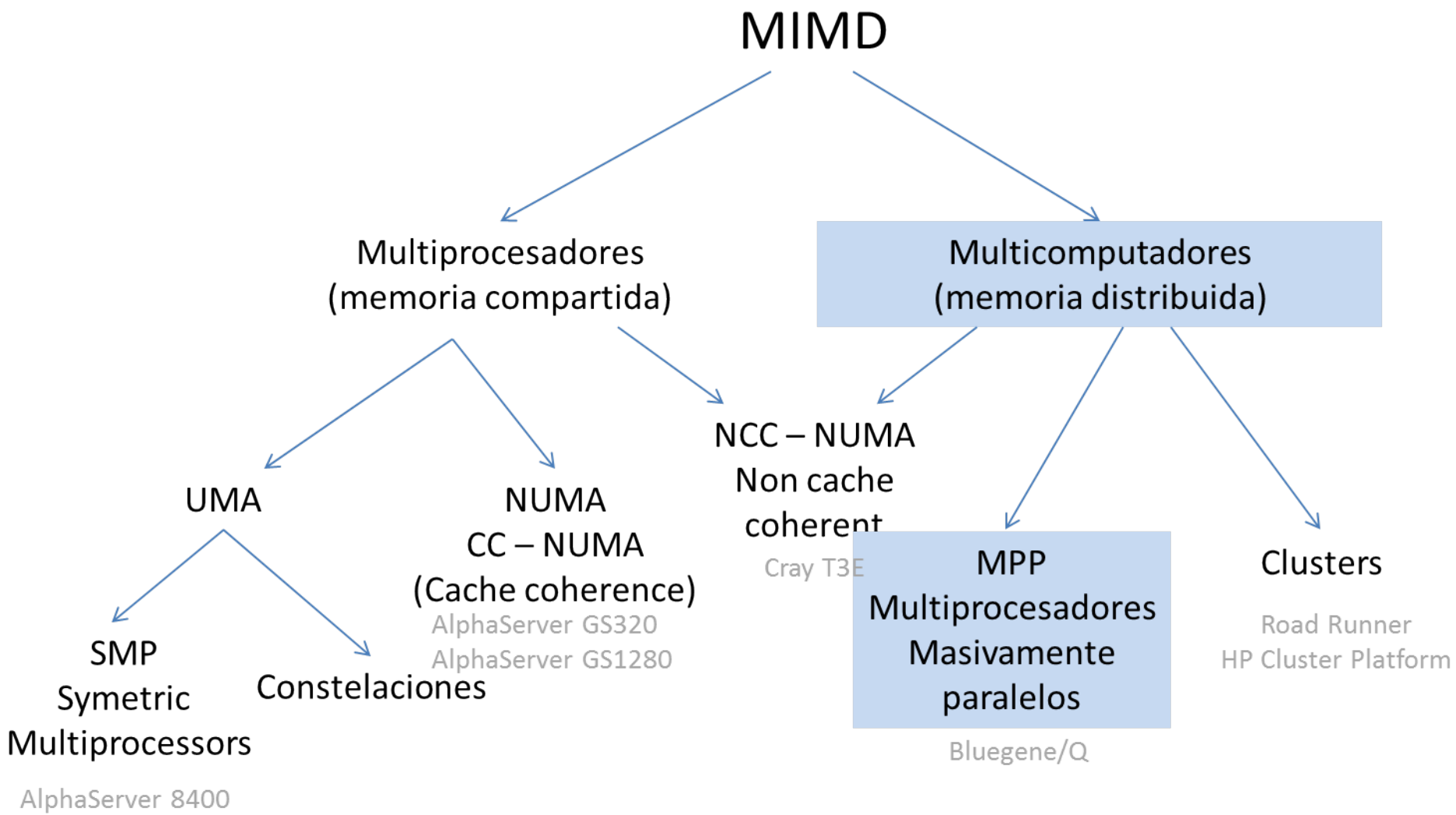
UNIVERSIDAD DE BURGOS

MPP

MIMD

Computador Masivamente Paralelo


MIMD
Multiprocesadores
(memoria compartida)
Multicomputadores
(memoria distribuida)
UMA
NUMA
CC – NUMA
(Cache coherence)
AlphaServer GS320
AlphaServer GS1280
NCC – NUMA
Non cache
coherent
Cray T3E
MPP
Multiprocesadores
Masivamente
paralelos
Bluegene/Q
Clusters
Road Runner
HP Cluster Platform
SMP
Symetric
Multiprocessors
AlphaServer 8400
Constelaciones

BLUE GENE/Q

Introducción

- Se trata del tercer representante de una serie que comenzó con el Blue Gene/L y continuó con el Blue Gene/P.
- Son máquinas desarrolladas por IBM con el propósito de encontrarse al frente del ranking mundial.
- Versiones de tamaño intermedio llegan a alcanzar un cierta difusión dentro del ámbito de la supercomputación.
- La implementación Sequoia, que ha llegado a alcanzar el nº1 mundial, se encuentra instalada en el laboratorio Lawrence Livemore de Estados Unidos e incluye un total de 1.572.864 núcleos de procesamiento.

Componentes Hardware[1]

- Tarjeta de computación: formada por un ASIC de computación y 72 chips de memoria SDRAM DDR3.
- Tarjeta nodo: incluye 32 tarjetas de computación conectadas en un toro 5D (2x2x2x2x2).
- Midplane: permite conectar hasta 512 nodos de procesamiento (16 tarjetas nodo), creando un toro 5D (4x4x4x4x2).
- Rack: incluye 2 midplanes (4x4x4x8x2).

Componentes Hardware[1]

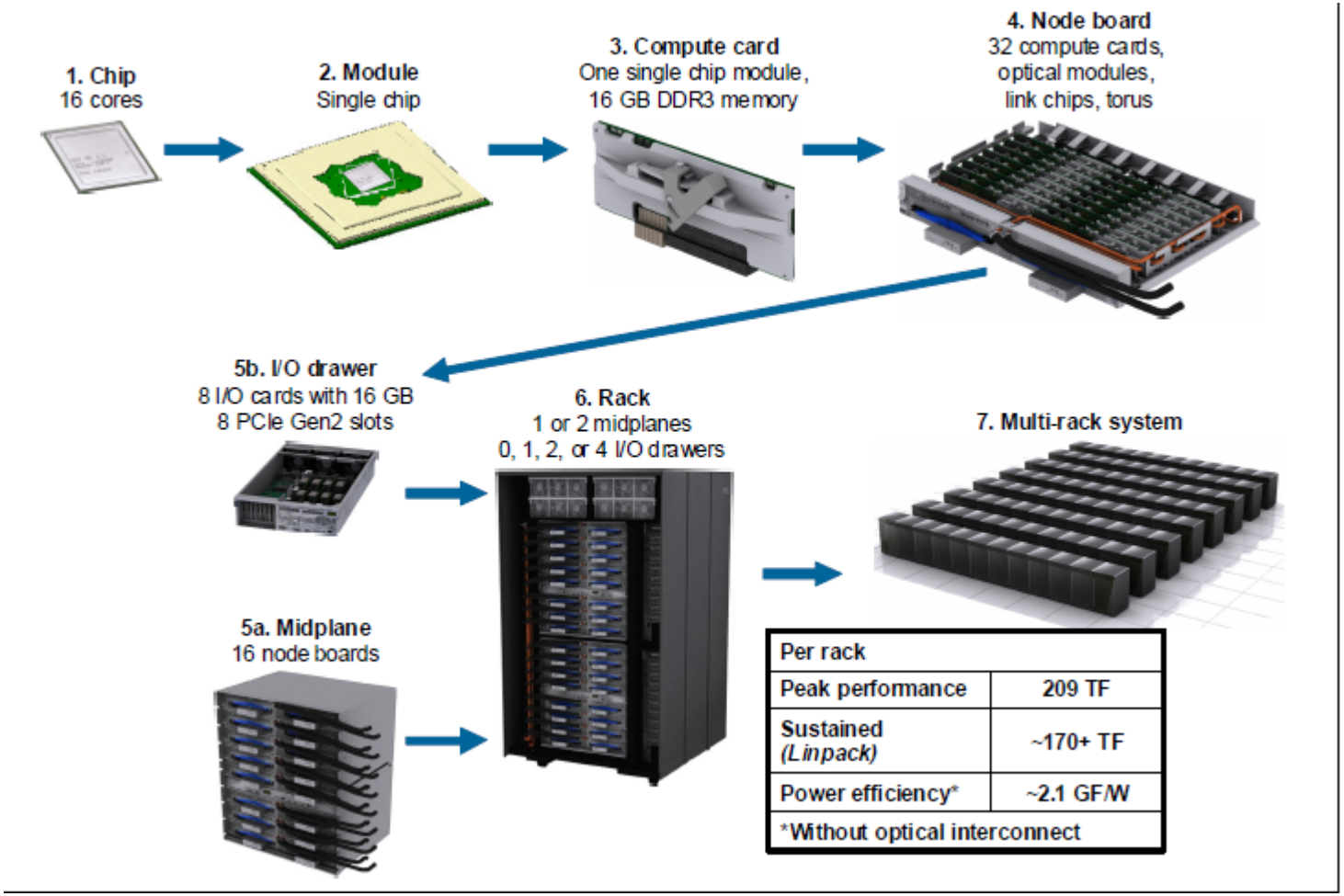


Figure credit: IBM Red Books

ASIC DE COMPUTACIÓN[2]

Dispositivo multicore de 16 núcleos de computación + 1 de servicio.
Basado en la tecnología PowerEN™.

- Reloj a 1,6 GHz.
- Arquitectura de 64 bits.
- Estructura multihilo de 4 vías en cada núcleo.
- Predicción dinámica de saltos.

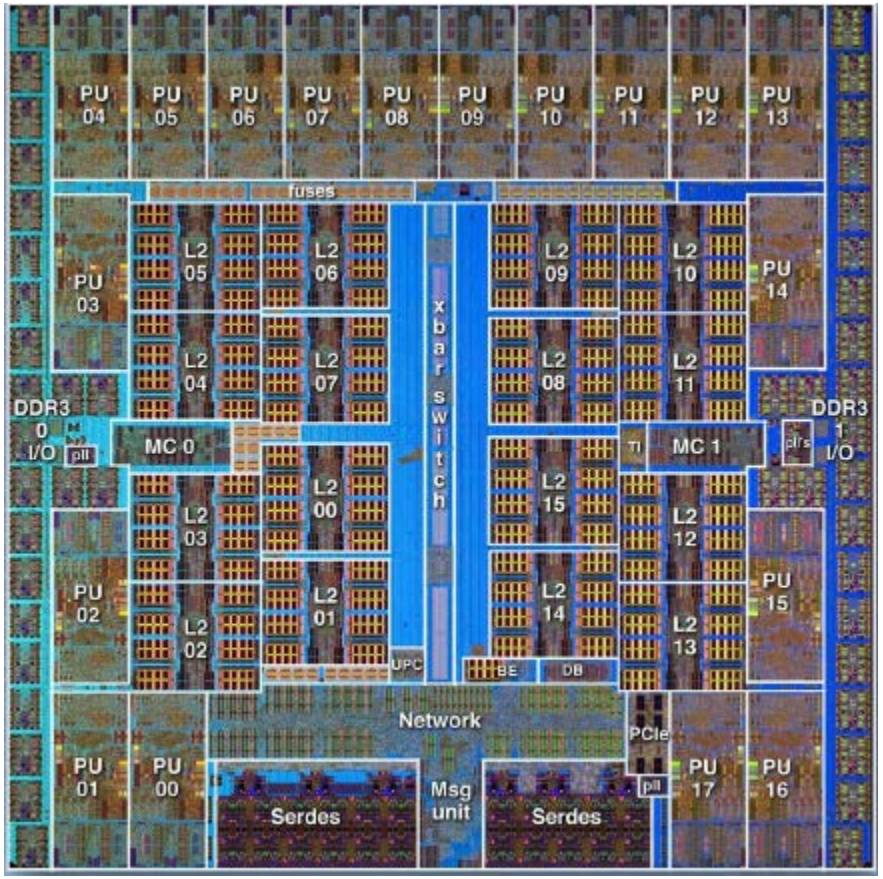


Figure credit: IBM Red Books

- Cache L1, 16 + 16 kB
- Cache L2 compartida de 32MB.
- Cuádruple unidad de coma flotante (posible uso como SIMD).
- Interconexión interna vía crossbar.

Jerarquía de memoria

- Cache L1, 64 bytes por línea de cache:
  - 16kB datos parcialmente asociativa de 8 vías.
  - 16kB inst. Parcialmente asociativa de 4 vías.
- Cache L2 compartida de 32MB. Parcialmente asociativa de 16 vías. Líneas de cache de 128 bytes.
- Soporta detección de condiciones de carrera entre núcleos, pero deben ser resueltas por software.
- Líneas de cache L2 de 128 bytes. Líneas de cache L1 de 64 bytes.
- DRAM externa en tarjeta.
- ECC

Interconexión I[3]

- La red de interconexión para computación es un toro 5D con una configuración máxima de 16x16x16x16x2 nodos de procesamiento. Sequoia es (16x16x16x12x2).
- El toro 5D crea redes colectivas embebidas para operaciones específicas. Por ejemplo, árboles binarios para operaciones colectivas.
- Los enlaces bidireccionales proporcionan 2 GBps en cada sentido.
- Las conexiones dentro de cada midplane son eléctricas, fuera de él son ópticas.

Interconexión II (encaminamiento)

- La lógica de control de red (router y unidad de mensajes) se encuentra integrada en el ASIC.
- El router dispone de 11 unidades de entrada y 11 de salida (5 dimensiones x 2 sentidos + I/O).
- Se implementan 4 canales virtuales: adaptativo, determinista, alta prioridad y de sistema.
- Se implementan varias colas por canal virtual para reducir el bloqueo de principio de línea.
- Tanto para la inyección como para el consumo, existen colas suficientes para mantener ocupados todos los canales del toro.
- Las distintas colas de inyección no están asignadas a un canal de salida en particular, mientras que las de consumo sí lo están.
- Control de flujo virtual cut through.

Interconexión III (unidad de mensajes)

- Proporciona comunicación entre la interconexión y la memoria.
- Conecta con el router mediante una serie de colas FIFO de inyección y consumo.
- Compone y descompone los mensajes en paquetes.

Interconexión IV (arbitraje)

- El arbitraje es distribuido:
  - Las unidades de envío (emisores) informan de la disponibilidad de sus canales de salida y el espacio en las colas de destino en los nodos vecinos a las unidades de recepción (receptores) y a las FIFOs de inyección.
  - El receptor de cada canal virtual selecciona con esa información un paquete al que dar servicio y envía una petición de arbitraje al árbitro de su receptor.
  - El árbitro del receptor selecciona un ganador entre los canales virtuales demandantes y traslada su petición al emisor requerido.
  - El árbitro del emisor asigna prioridades primero de acuerdo con la naturaleza de los paquetes: colectivos y de sistema, de usuario de alta prioridad, de usuario con prioridad normal. Dentro de cada categoría es posible configurar la prioridad relativa entre las colas de inyección y los canales virtuales.
  - Las FIFOs de inyección se arbitran de una manera similar.
  - La política de decisión final no está documentada.
  - Las comunicaciones colectivas son arbitradas primero por una lógica central que tienen en cuenta las disponibilidades de los múltiples canales de salida requeridos.

Mapeo de nodos [4]

- Qué es? Decidir cómo se mapean los procesos en los nodos asignados.
- Solución por defecto: los rangos MPI se incrementan en orden ABCDET. ABCDE son las coordenadas del toro 5D. T es la ID de un procesador dentro de un nodo de procesamiento.
- Si se asume un procesador por nodo, T es 0.
- El comando *runjob* se puede usar para modificar esta opción:
  - Se pueden permutar las coordenadas del toro: TBCDAE p.e.
  - Se puede especificar un fichero de mapeo personalizado.
- Si se usa un planificador de lotes, éste puede contener un equivalente al comando *runjob* para lograr el mismo objetivo.

Software [4]

- Kernel Linux completo en los nodos de e/s.
- Compute Node Kernel en los nodos de computación.
- Soporte para librerías MPI.
- OpenMP API.
- Compiladores y soporte para aplicaciones y depuradores.
- Herramientas de administración y gestión (IBM LoadLeveler scheduler).

Referencias

- [1] IBM System Blue Gene Solution: Blue Gene/Q Hardware Overview and Installation Planning. J. Milano & P. Lembke. IBM redbooks, March 2012
- [2] The Blue Gene/Q Compute Chip. R. Jaring / IBM BleGene Team. IBM Corporation, July 2011
- [3] The IBM Blue Gene/Q Interconnection Network and Message Unit. D. Chen et al. Proc. Of 2011 International Conference for High Performance Computing, Network, Storage and Analysis. Article No. 26.
- [4] IBM System Blue Gene Solution. Blue Gene/Q Application Development. M. Gilge. IBM redbooks, December 2012