IBM BlueGene

Cap al Petaflop

Jordi Fornés

May 15, 2003

1 IBM BlueGene

- Al 1999, IBM va anunciar un projecte de 5 anys i 100 M$
- L'objectiu era construir un supercomputador que arribés al petaflop.
- Guanyaria de llarg al *NEC Earth Simulator*.
- En principi, orientat a problemes de bioinformàtica.

- Al maig del 2003!
  - Ja s'han sobrepassat els 100 M$
  - S'espera que un BlueGene de 180/360 Tflops estigui llest a finals del 2004.
  - S'han expandit els problemes a tractar: simulació climàtica i anàlisi de riscs financers.
  - BlueGene/L esdevindrà la pròxima generació de sistemes de computació massivament paral.lels.

1.1 La família

1.1 La família

- BlueGene/L
  - 65000 nodes.
  - 2 processadors per node.
  - 180/360 Tflops.

1.1 La família

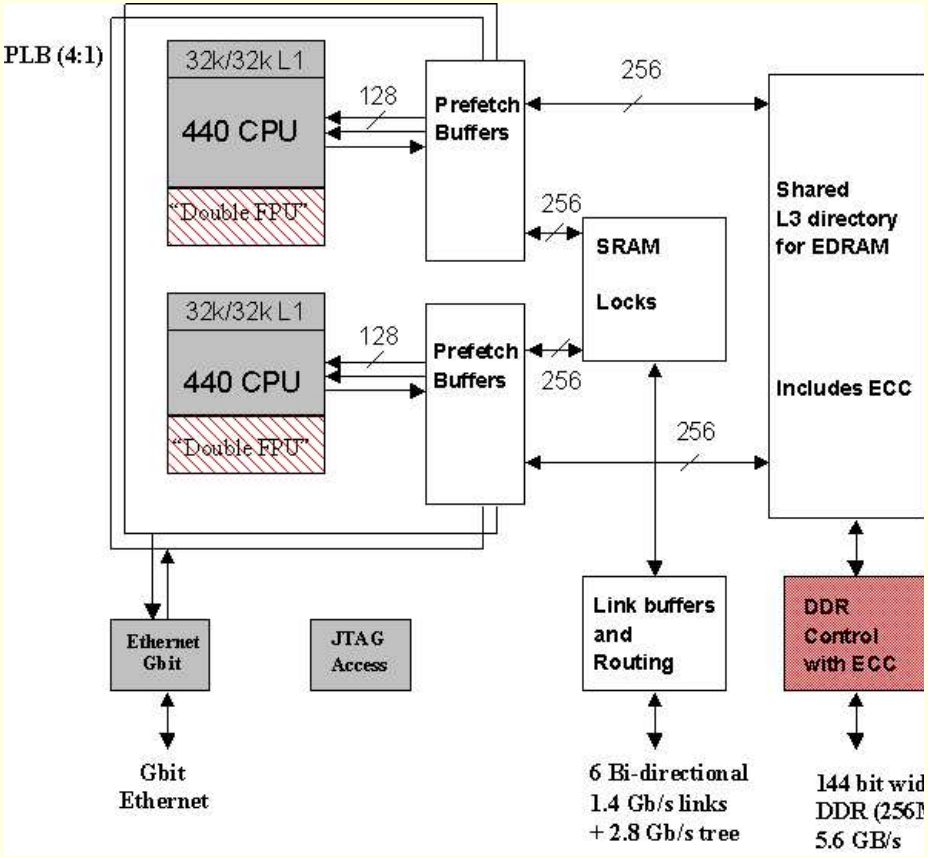
- BlueGene/L
  - 65000 nodes.
  - 2 processadors per node.
  - 180/360 Tflops.
- BlueGene/C
  - 64 processadors per node.

1.1 La família

- BlueGene/L
  - 65000 nodes.
  - 2 processadors per node.
  - 180/360 Tflops.
- BlueGene/C
  - 64 processadors per node.
- BlueGene/P
  - La posada en marxa dels dos anteriors (ara en fase de fabricació) marcarà molts aspectes d'aquest BlueGene.
  - L'objectiu és de 1000 Tflops.

2 El maquinari del BlueGene/L

Diagrama d'un node del BlueGene/L



2.1 PowerPC 440 Avenger

- Processador superescalar de 32 bits.
- Conjunt de instruccions de l'arquitectura del PowerPC.
  - Accés a memòria amb `Load/Store`.
  - Operacions `registre/registre`.
  - 32 registres de propòsit general.
- 2 instruccions per cicle issue/fi.
  - 1 instrucció `load/store`.
  - 1 instrucció de `float` per cicle.

2.2 Unitat de Gestió de memòria

(MMU)

- TLB gestionada per software.
- 64 entrades, totalment associatives.
- Pàgines de tamany variable (1kB-256MB) simultàneament residint a la TLB.
  - Linux als I/O nodes fa servir 4kB.
  - BLRTS als compute nodes fa servir tamany variable per mapejar totalment la memòria al procés.

2.3 Jerarquia de Memòria

2.3.1 L1 - caches

- Cada CPU té una cache de 32 kB.
- 32 bytes per línia.
- Interfície de 128 bits cap al processador/fpu.
- No coherent entre les CPUs.
- Latència de 3 cicles.

2.3.2 L2 - caches

- Cada CPU té 2kB de cache de nivell 2.
- Coherent.
- Totalment associativa.
- Programable com a pre-fetch buffer.
- Latència de 11 cicles.

2.3.3 L3 - caches

- 4 MB EDRAM de cache de nivell 3.
- 128 bytes per linia.
- Part de l'EDRAM pot ser configurada com a memòria direccionable (scratchpad).
- 23 - 31 cicles de latència.

2.3.4 SRAM

- 16 KB de SRAM compartida per les dues CPUs.
- Accessible pels nodes de servei externs via la xarxa JTAG.
- Mecanisme principal per bootar i inicialitzar la màquina.

2.3.5 External DRAM

- L'arquitectura suporta adreçament de 2GB.
  - Compute nodes configurats amb 256 MB de DRAM.
  - I/O nodes configurats amb 512 MB de DRAM.
- Línia de transferència de 128 bytes entre la DRAM externa i la L3.

2.4 $Hummer^2$ fpu

- 2 unitats aritmètiques de coma flotant.
- 2 grups de de registres 32x64 bits.
- Cada unitat pot agafar dades de tots dos grups:
  - En paral.lel.
  - Replicada.
  - Creuada.
- Les operacions a cada unitat són gaire bé independents.
- SIOMD: Single Instructions Multiple Operation Multiple Data.

2.5 Lockbox

- 256 locks d'un bit.
- Memòria mapejada (cadascún dels locks té una adreça diferent).
  - El `read` intenta apropiar-se del lock.
  - `write` 0 allibera el lock.
- Es pot simplement testejar el lock.
- Es podem fer servir per `barriers`.

2.6 La xarxa

- El torus

  La xarxa torus es fa servir per comunicacions de missatges d'usuari punt a punt i multicast entre grups de nodes.

  Cada node té sis connexions per connexió directe pels nodes més propers.

  Els 64K nodes es poden organizar en 64x32x32 3D torus.

  Només pels compute nodes

- L'arbre.

  Comunicació entre compute nodes i I/O nodes.

  Per punt a punt, broadcast i reduccions de packets.

- L'Ethernet

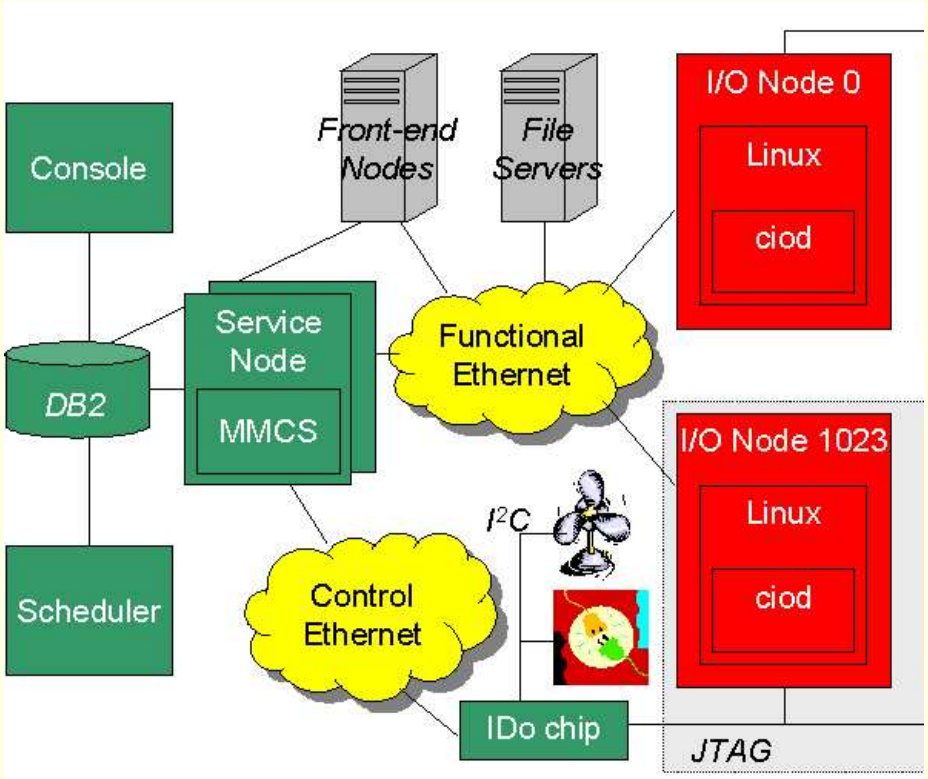
  Una Gbit/s ethernet serveix per a la comunicació amb l'exterior.

  Només accessible pels I/O nodes.

- JTAG

  Per bootar, control i monitorització externa.

3 El programari del BlueGene/L



3.1 Compute nodes

- Les aplicacions d'usuari s'executen exclusivament als compute nodes.
- La màquina s'ha de centrar en aquesta execució i no pas en gestionar el sistema.
  - No events asíncrons.
  - No operacions complexes als compute nodes.
- Solució adhoc: BLRTS

3.1.1 BLRTS

- Blue Gene/L Runtime Supervisor.
- Kernel minimalista pels compute nodes.
  - Protegeix el hardware de errors d'aplicació: espai d'usuari/espai de kernel.
  - Monousuari, monoprogramat.
  - Com a molt dos threads [a].
  - Tamany fixe de 256 MB d'espai d'adreces.
  - Sense paginació.
  - API POSIX (GLIBC 2.2.5 llibreria de runtime).
  - Simple: unes 5000 linies de codi en C++.

[a] en realitat 1,5

- Ús de la segona CPU
  - Tasques de comunicació: enviar missatges MPI.
  - El thread principal pot engegar un thread a la segona CPU.
  - Comunicació via scratchpad.
  - EL kernel NO és multithread: el codi d'usuari NO pot fer crides al sistema.

3.2 I/O nodes

- Hardware semblant als compute nodes.
  - Més memòria i Gbit Ehernet.
- La resta del món veu només els I/O nodes.
  - El sistema es veu com un cluster de 1024 I/O nodes.
- Solució comode: Linux

3.2.1 Linux

- Modificacions:
  - Afegides utilitats i llibreries.
  - Afegits scripts d'inicialització.
  - Alguns moduls de kernel.
  - Utilitats del simulador de BG/L.
  - Només usuari `root`, per ara.
- Ús de la segona CPU
  - Només fa servir una per problemes de cohrencia de la cache.

4 Suport MPI al BG/L

- Comunicacions punt a punt.
- Gestió de processos.
- Suport al coneixement de topologies als programes MPI

  `MPI_cart_map()`
- Optimizaciós de certes crides.
- Un processador dedicat a la comunicació de missatges MPI.

5 El simulador de BG/L

- Entorn de desenvolupament de software per BG/L abans de que el hardware estigui disponible.
- Simula tant un sol xip com tot un cluster.
- Simula 2 milions de instruccions per segon sobre un Linux 2GHz.

6 Bibliografia

http://www.research.ibm.com/bluegene

http://www.llnl.gov/asci/platforms/bluegene

http://www.cepba.upc.es/ciri