

Execution Environments for Parallel Applications

Master CANS 2007/2008
Departament d'Arquitectura de Computadors
Universitat Politècnica de Catalunya

UPC

Dept. d'Arquitectura de Computadors

Execution Environments for Parallel Applications



Outline

- ❑ Supercomputers
- ❑ OS abstractions
- ❑ Extended OS interfaces
 - ❑ IRIX sysmp
 - ❑ IRIX memory placement
- ❑ Tools for performance analysis
- ❑ Example: the CPU Manager
- ❑ Linux interfaces
- ❑ Conclusions

UPC

Dept. d'Arquitectura de Computadors

Execution Environments for Parallel Applications



Supercomputers

- ❑ Variety of machines...

UPC

Dept. d'Arquitectura de Computadors

Execution Environments for Parallel Applications



Supercomputers

- ❑ From ...
 - ❑ www.top500.org, Cray, SGI, IBM, Earth Simulator Center...

Name	Processor	Number	Ghz	Peak	Performance	% of peak	Interconnect
Cray X1E	Vector	8192	1.13	147000	?	-	2D torus/custom
Cray X1	Vector	504	1.13	6451	5895	91.00%	2D torus/custom
Altix 3700	Itanium	10160	1.5	60960	51870	85.00%	Infiniband
Origin 3900	R16000	512	0.8	?	?	-	?
BG/L	ppc440	65536	0.7	183500	135300	73.00%	3D torus/custom
Marenostrum	ppc970	3564	2.2	31363	20530	65.00%	Myrinet
Earth Simulator	Vector	5120	?	40960	35860	87.00%	Custom
IBM pSeries	Power4+	2944	1.7	20019	10310	51.00%	Federation Switch
ASCI Q	Alpha	8192	1.25	20480	13880	67.00%	Quadrics

UPC

Dept. d'Arquitectura de Computadors

Execution Environments for Parallel Applications



Origin2000 & Origin 3000

Origin2800, 64 processors
max. 512 processors

Origin 3400, 32 processors



8 processors



4 processors

UPC

Dept. d'Arquitectura de Computadors

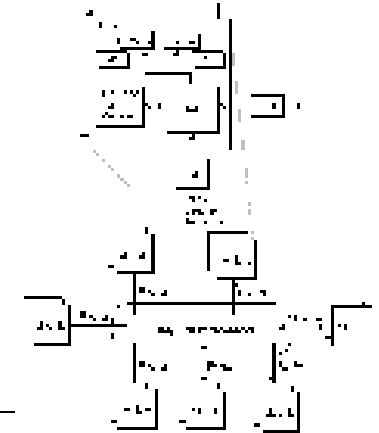
Execution Environments for Parallel Applications

5

Origin2000

Basic node

- Two processors (R10000/L2 cache)
- One Hub
- Directory/main memory

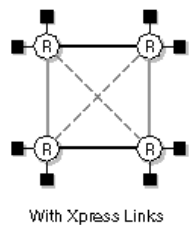


UPC

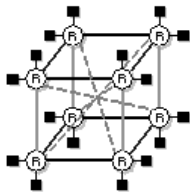
Dept. d'Arquitectura de Computadors

6

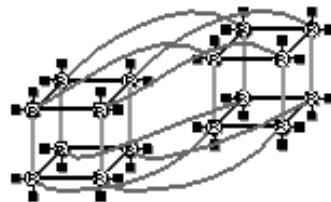
Origin2000 architecture



16 cpus



32 cpus



64 cpus

UPC

Dept. d'Arquitectura de Computadors

Execution Environments for Parallel Applications

7

Origin2000

Network latency

- L1 cache 5.1 ns.
- L2 cache 56.4
- Local memory 310
- 4P remote memory 540
- 8P avg. remote memory 707
- 16P avg. remote memory 726
- 32P avg. remote memory 773
- 64P avg. remote memory 867
- 128P avg. remote memory 945

- <http://www.sgi.com/origin/images/isca.pdf>

UPC

Dept. d'Arquitectura de Computadors

Execution Environments for Parallel Applications

8

Origin 3000

Basic node (C-brick)

- Two/four processors
 - R12000/L2 cache
- Bedrock ASIC
 - CPU/memory/I/O
- Memory modules
 - 512 Mb. to 8 Gb.
- Xtown2
 - Connection to I/O brick



UPC

9

Dept. d'Arquitectura de Computadors

Execution Environments for Parallel Applications



Origin 3000

Components

- C-brick (cpu)
- I-brick (I/O)
- P-brick (PCI I/O)
- X-brick (XIO)
- D-brick (Disk I/O)
- R-brick (router brick, hub)
- G-brick (graphics)



UPC

10

Dept. d'Arquitectura de Computadors

Execution Environments for Parallel Applications



Origin 3000 architecture

128 processors



UP

11



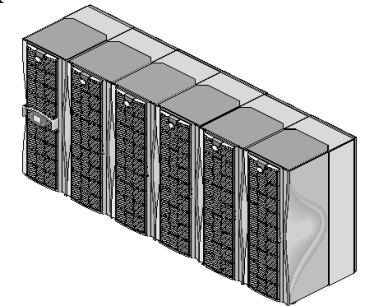
ications

Origin 3900

512 processors, 1 TB. RAM

- 32 Cx-bricks
 - 16 CPU's (R16000) per Cx-brick
 - 32 GB. of memory
- IX-brick - base I/O
- PX-brick - PCI expansion
- R-brick - router interconnect
- ...

IRIX 6.5 OS



UPC

12

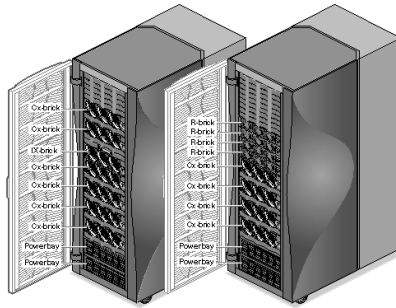
Dept. d'Arquitectura de Computadors

Execution Environments for Parallel Applications



Origin 3900

- ❑ **Two racks system**
 - ❑ $10 \times 16 = 160$ cpus, 320 GB of memory



UPC

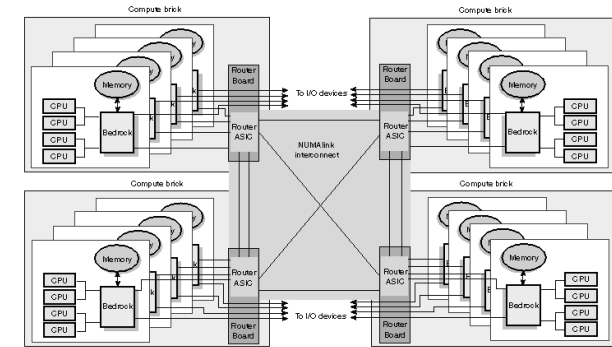
13

Dept. d'Arquitectura de Computadors

Execution Environments for Parallel Applications

Origin 3900

- ❑ **4 Cx-bricks**



UPC

14

Dept. d'Arquitectura de Computadors

Execution Environments for Parallel Applications

Origin

- ❑ **IRIX 6.5**
 - ❑ Processes
 - ❑ Sprocs
 - ❑ Kernel threads
 - ❑ Memory placement
- ❑ **Posix threads based on kernel threads**
- ❑ **OpenMP based on sprocs**



UPC

15

Dept. d'Arquitectura de Computadors

Execution Environments for Parallel Applications

Altix 3000

- ❑ **64 Itanium-2 processors (1600 Mhz)**
 - ❑ Global memory
 - ❑ 129 ns. local access
 - ❑ 559 ns. remote access



UPC

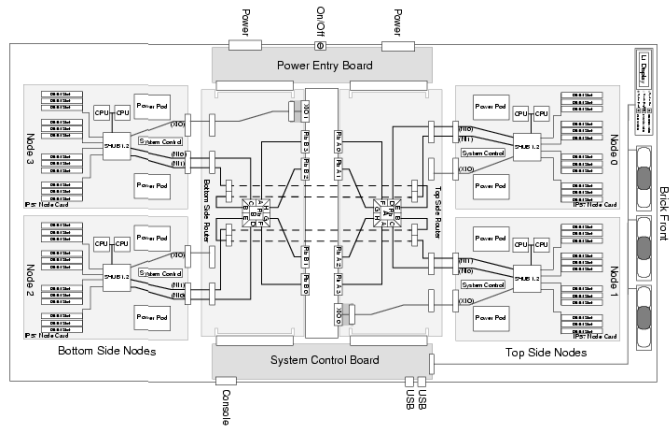
16

Dept. d'Arquitectura de Computadors

Execution Environments for Parallel Applications

Altix 3000 Bx2

- ❑ CR-brick - 8 cpus - 768 GB of memory



UPC

17

Dept. d'Arquitectura de Computadors

Execution Environments for Parallel Applications

Altix

- ❑ 64-bit linux
 - ❑ Processes
 - ❑ Clones
 - ❑ Kernel threads (?)
- ❑ Posix threads based on clones
 - ❑ Will kernel threads be used in the future?
- ❑ OpenMP based on Posix threads

UPC

18

Dept. d'Arquitectura de Computadors

Execution Environments for Parallel Applications

HP-Compaq AlphaServer

- ❑ GS series
 - ❑ up to 32 processors (1Ghz. Alpha)
 - ❑ up to 256 Gb. of shared memory
 - ❑ Advanced crossbar switch
 - ❑ 64 PCI buses, 224 PCI slots



UPC

19

Dept. d'Arquitectura de Computadors

Execution Environments for Parallel Applications

HP-Compaq AlphaServer

- ❑ SC series
 - ❑ based on 4-processor nodes
 - ❑ 1250 Mhz. Alpha EV68
 - ❑ distributed memory
 - ❑ up to 4096 processors
 - ❑ up to 32 Tb. of memory



UPC

20

Dept. d'Arquitectura de Computadors

Execution Environments for Parallel Applications

HP-Compaq AlphaServer

❑ GS 1280

- ❑ 64 EV7 Alpha processors (1300 Mhz)
- ❑ 32 (2-cpus) virtual servers
- ❑ 512 GB. of memory



UPC

21

Dept. d'Arquitectura de Computadors

Execution Environments for Parallel Applications

AlphaServer

❑ Tru64 UNIX

- ❑ Processes
- ❑ Kernel-level threads (mach)

❑ Posix threads based on kernel threads

❑ OpenMP based on Posix threads

❑ OpenVMS



UPC

22

Dept. d'Arquitectura de Computadors

Execution Environments for Parallel Applications

IBM R6000 SP2

- ❑ 375 MHz POWER3 SMP
- ❑ Processors per node: 4/8/12/16
- ❑ Up to 128 nodes,
for a total of 2048 processors
- ❑ On demand, 512 nodes
- ❑ Memory: 1-64 Gb., distributed
- ❑ AIX 4.3



UPC

23

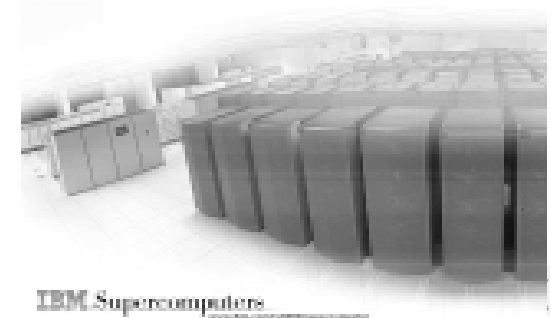
Dept. d'Arquitectura de Computadors

Execution Environments for Parallel Applications

IBM SP2

❑ IBM ASCI White

- ❑ 8192 RS 6000 processors
- ❑ 512 nodes x 16 cpus/node
- ❑ Cluster architecture
- ❑ 12 teraflops/s peak
- ❑ AIX 5.1



IBM Supercomputers

UPC

24

Dept. d'Arquitectura de Comput

Cray X1E

- ❑ **8192 vector processors**
 - ❑ 64-bit
 - ❑ Multistreaming
 - ❑ 1.13 Ghz
- ❑ **4-way SMP nodes**
- ❑ **Custom interconnect providing DSM**
- ❑ **True Single System Image OS**
 - ❑ UNICOS/mp



UPC

25

Dept. d'Arquitectura de Computadors

Execution Environments for Parallel Applications

Marenostrum

- ❑ **4564 Power PC 970 FX processors**
 - ❑ VMX extension
- ❑ **2282 2-way nodes**
- ❑ **9 TB of memory**
- ❑ **Networks**
 - ❑ Myrinet
 - ❑ Gigabit
 - ❑ 10/100 Ethernet
- ❑ **Linux 2.6**



UPC

26

Dept. d'Arquitectura de Computadors

Execution Environments for Parallel Applications

Blue Gene/L

- ❑ **A 5 year project announced by IBM in 1999, to build a petaflop/s scale supercomputer to attack science problems such as protein folding**
 - ❑ Advance the state of the art of scientific simulation
 - ❑ Advance the state of the art in computer design and software for extremely large scale systems
- ❑ **November 2001: Research partnership with Lawrence Livermore National Laboratory**
- ❑ **November 2002: Planned acquisition of BG/L by LLNL as part of ASCI Purple**
- ❑ **June 2003: First chips completed (DD1)**
 - ❑ Summer 2003: half rack built in IBM T.J. Watson
 - ❑ System software & MPI library ported from the simulation environment



UPC

27

Dept. d'Arquitectura de Computadors

Execution Environments for Parallel Applications

Blue Gene/L

- ❑ **June 2003**
 - ❑ Chips arriving
 - ❑ Diagnostics run the same day
 - ❑ System software booted one week later (first day of access)
 - ❑ MPI and Linux running the second day
 - ❑ Some hardware problems make it run slower (500 Mhz) and show memory problems
 - ❑ Memory access: div2 instead of div3



UPC

28

Dept. d'Arquitectura de Computadors

Execution Environments for Parallel Applications

BG/L October 2003

❑ BG/L half rack prototype

- ❑ 500 Mhz - DD1
- ❑ 512 nodes / 1024 processors
- ❑ 2 TFlops/s peak
- ❑ 1.4 TFlops/s sustained



UPC

29

Dept. d'Arquitectura de Computadors

Execution Environments for Parallel Applications

Blue Gene/L

❑ November 2003: BG/L Half rack prototype

- ❑ Ranked #73 on 22nd Top500 List announced at SC2003 (1.4 TFlop/s)
- ❑ Protein folding algorithm running

❑ February 2, 2004: Second pass BG/L chips delivered to Watson

- ❑ 700 Mhz processors, DD2
- ❑ No problems found

❑ March 19, 2004: 2048 nodes of BG/L run first application - DD1

- ❑ Achieving 5.7 Tflop/s (Linpack)

❑ All this on 40 sq. ft. of floor space! (< 4 m²)



UPC

30

Dept. d'Arquitectura de Computadors

Execution Environments for Parallel Applications

Blue Gene/L

❑ December 2004: got pole position on Top500 list

- ❑ 16384 nodes
- ❑ 32768 DD2 processors
- ❑ 70 TFlops (Linpack)

❑ Planned 6144-compute node system at ASTRON (Netherlands)

- ❑ DD2 processors
- ❑ Planned around 25 TFlops

❑ Planned 128-node machine at Argonne National Labs

❑ Current status

- ❑ 32768-node machine installed at LLNL
- ❑ 135.3 Tflops (Linpack)



UPC

31

Dept. d'Arquitectura de Computadors

Execution Environments for Parallel Applications

Blue Gene/L installations

- ❑ LLNL
- ❑ Watson
- ❑ Lausana
- ❑ Astron
- ❑ Argone (?)
- ❑ ...



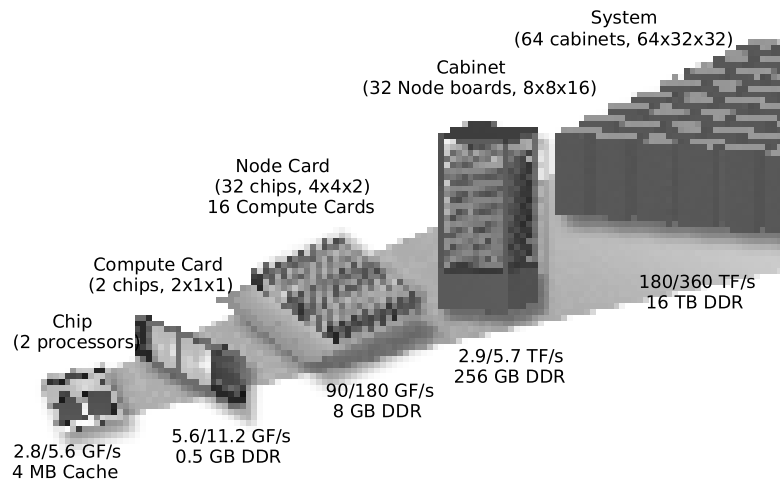
UPC

32

Dept. d'Arquitectura de Computadors

Execution Environments for Parallel Applications

Blue Gene/L



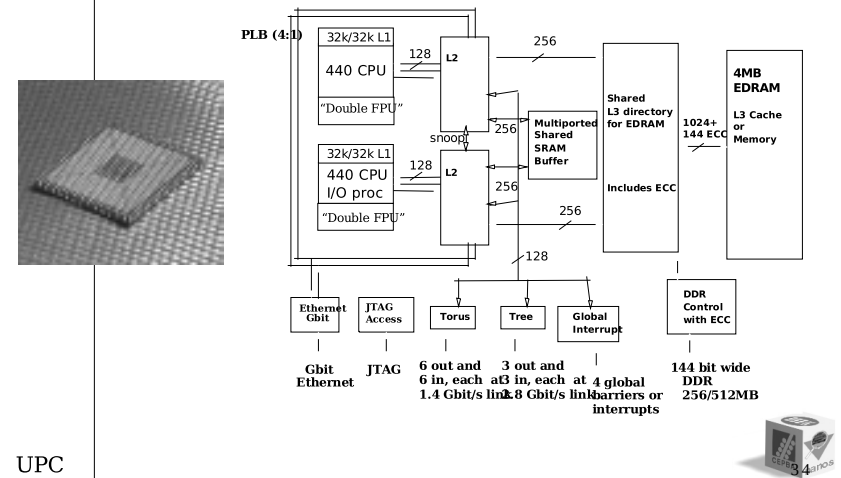
UJ

33

Dept. d'Arquitectura de Computadors

Execution Environments for Parallel Applications

Blue Gene/L Compute ASIC



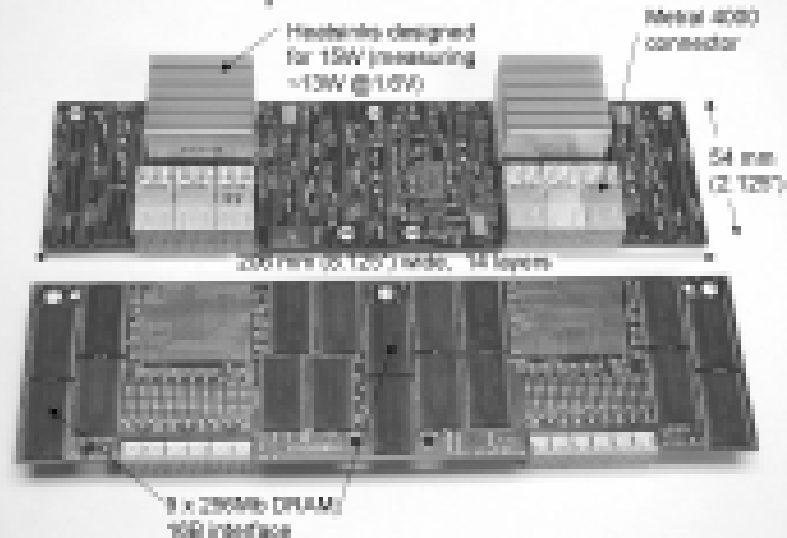
UPC

34

Dept. d'Arquitectura de Computadors

Execution Environments for Parallel Applications

Dual Node Compute Card



16
compute
cards

2 I/O cards

DC-DC Converters:
40V 1.5, 2.5V



BG/L rack	
☐ Cabled... X Cables Y Cables Z Cables	
UPC	
38	Dept. d'Arquitectura de Computadors

Blue Gene/L Interconnection	
☐ 3 Dimensional Torus ☐ Communications backbone for computations ☐ Interconnects all compute nodes (65,536) ☐ Hardware routing ☐ 1.4Gb/s on all 12 node links (2.1 GBytes/s per node) ☐ 350/700 GBytes/s bisection bandwidth	
☐ Global Tree ☐ One-to-all broadcast ☐ Reduction operations ☐ Interconnects all compute and I/O nodes (1024) ☐ 2.8 Gb/s of bandwidth per link ☐ Latency of tree traversal in the order of 5 μs	
UPC	
39	Dept. d'Arquitectura de Computadors Execution Environments for Parallel Applications

Blue Gene/L Interconnection	
☐ Ethernet ☐ Incorporated into every node ASIC ☐ Active in the I/O nodes (1:64) ☐ All external comm. (file I/O, control, user interaction, etc.)	
☐ Low Latency Global Barrier ☐ 8 single wires crossing whole system, touching all nodes	
☐ Control Network (JTAG) ☐ For booting, checkpointing, error logging	
UPC	
40	Dept. d'Arquitectura de Computadors Execution Environments for Parallel Applications



Blue Gene/L cooling

- ❑ **Bottom-top airflow**
 - ❑ Flow rate in cold duct is largest at bottom
 - ❑ Flow rate in hot duct is largest at top
 - ❑ The duct volume is constant regardless of flow rate

Marenosturm

Hot Cold Rack Cold Rack Hot Cold

UPC

42
Dept. d'Arquitectura de Computadors
Execution Environments for Parallel Applications

Blue Gene/L cooling

- ❑ **Left-right airflow, direct from raised floor**
 - ❑ Ducts are larger where flow is greater

BG/L
Alternative Ducting Scheme

Hot Hot Hot Hot
Cold Cold Cold Cold

UPC

43
Dept. d'Arquitectura de Computadors
Execution Environments for Parallel Applications

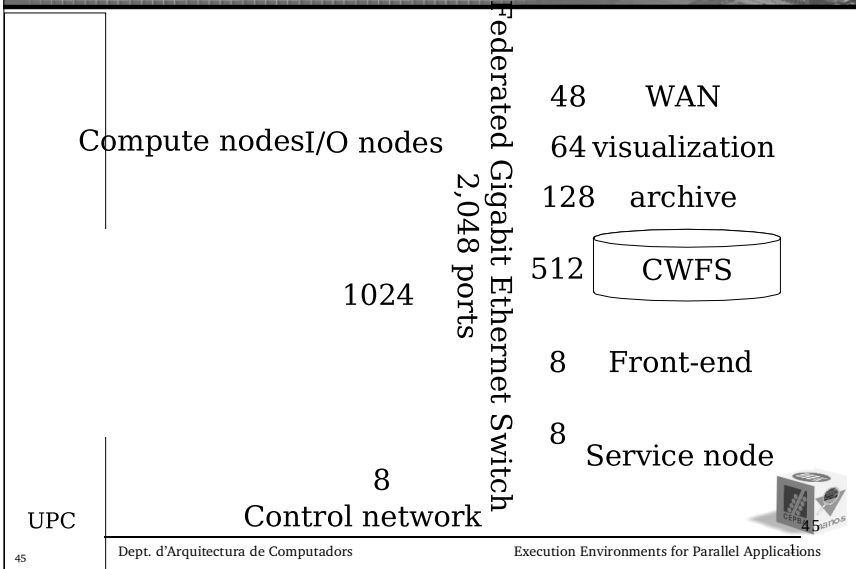
Blue Gene/L cooling

- ❑ **Typical BG/L installation**

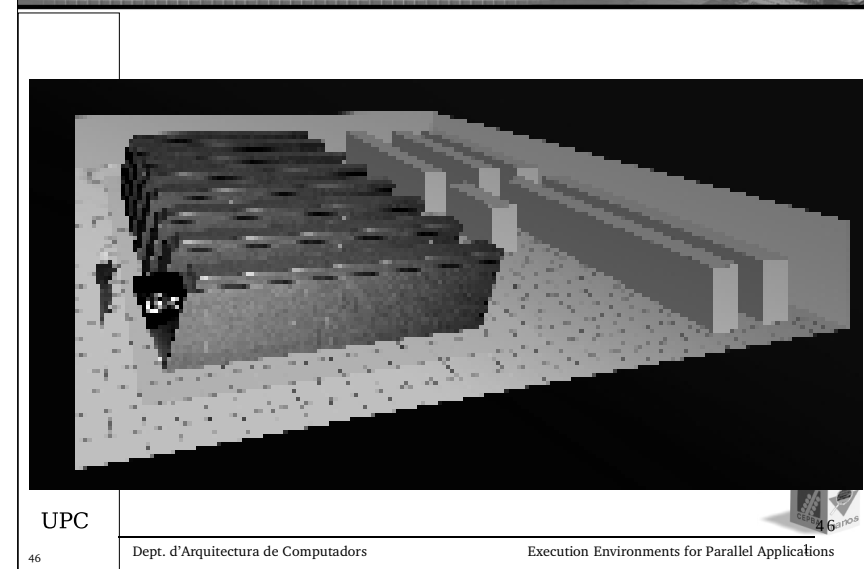
UPC

44
Dept. d'Arquitectura de Computadors
Execution Environments for Parallel Applications

Complete Blue Gene/L System at LLNL



Blue Gene/L System



Software Stack in BG/L

- ❑ CNK
 - ❑ controls all access to hardware, and enables bypass for application use
- ❑ User-space libraries and applications
 - ❑ directly access torus and tree through bypass
 - ❑ As a policy, user-space code should not directly touch hardware, but there is no enforcement of that policy
 - ❑ Application code can use both processors in a compute node

Application code

User-space libraries

CNK

Bypass

BG/L ASIC

UPC

47

Dept. d'Arquitectura de Computadors

Execution Environments for Parallel Applications

Machine characteristics

- ❑ From...
 - ❑ www.top500.org, Cray, SGI, Cray, Earth Simulator Center...

Name	Processor	Number	Ghz	Peak	Performance	% of peak	Interconnect
Cray X1E	Vector	8192	1.13	147000	?	-	2D torus/custom
Cray X1	Vector	504	1.13	6451	5895	91.00%	2D torus/custom
Altix 3700	Itanium	10160	1.5	60960	51870	85.00%	Infiniband
Origin 3900	R16000	512	0.8	?	?	-	?
BG/L	ppc440	65536	0.7	183500	135300	73.00%	3D torus/custom
Marenostrum	ppc970	3564	2.2	31363	20530	65.00%	Myrinet
Earth Simulator	Vector	5120	?	40960	35860	87.00%	Custom
IBM pSeries	Power4+	2944	1.7	20019	10310	51.00%	Federation Switch
ASCI Q	Alpha	8192	1.25	20480	13880	67.00%	Quadrics

UPC

48

Dept. d'Arquitectura de Computadors

Execution Environments for Parallel Applications