The SGI Altix 3000 Global Shared Memory Architecture

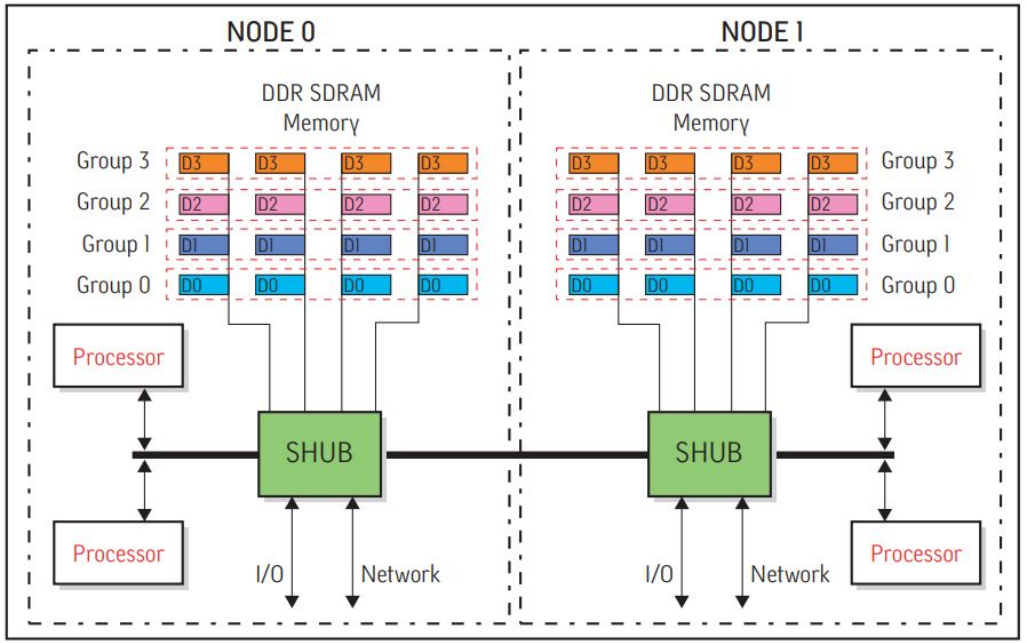
Paper: Michael Woodacre, Derek Robb, Dean Roe, and Karl Feind
Presentation: Daniel Nichols

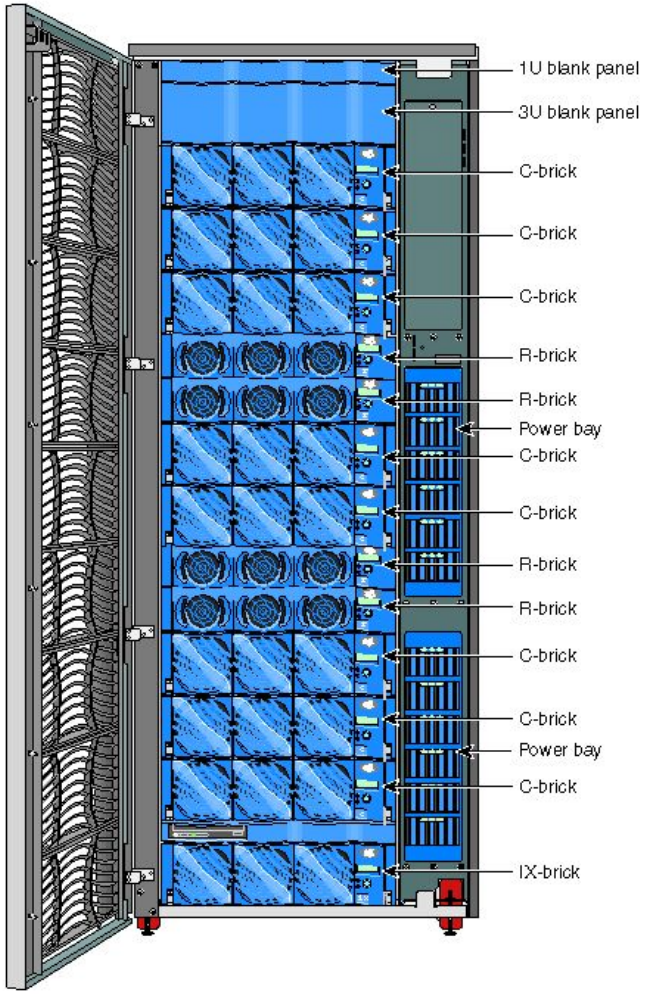
Paper Info

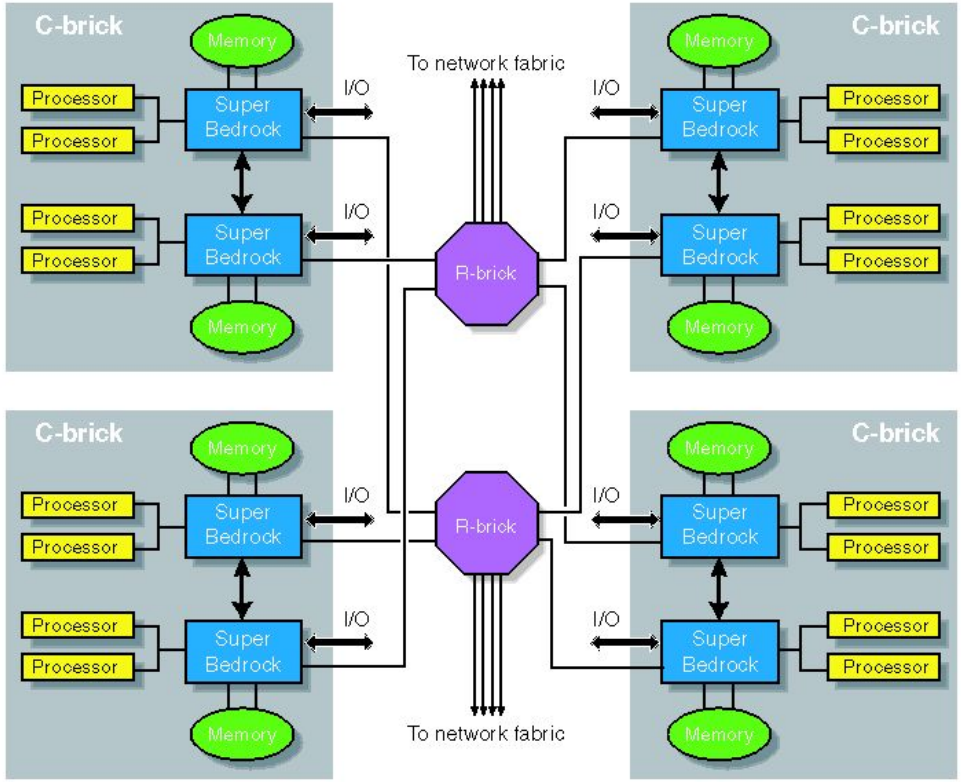
- Technical report from Silicon Graphics International
- Released in 2003
- Authors
  - Engineers and researchers at SGI
- Introduces the Altix 3000 system and highlights its architecture
- Outline
  - Components (NUMAflex, processors, memory, network)
  - RAS
  - OS and Software

NUMAflex and Shared Memory

- Protocol which offers cache coherence and global shared memory
- “Bricks” as building blocks
  - C-Brick, M-Brick, R-Brick, IX-Brick, PX-brick, D-Brick
- Driven by SHUB ASIC controller

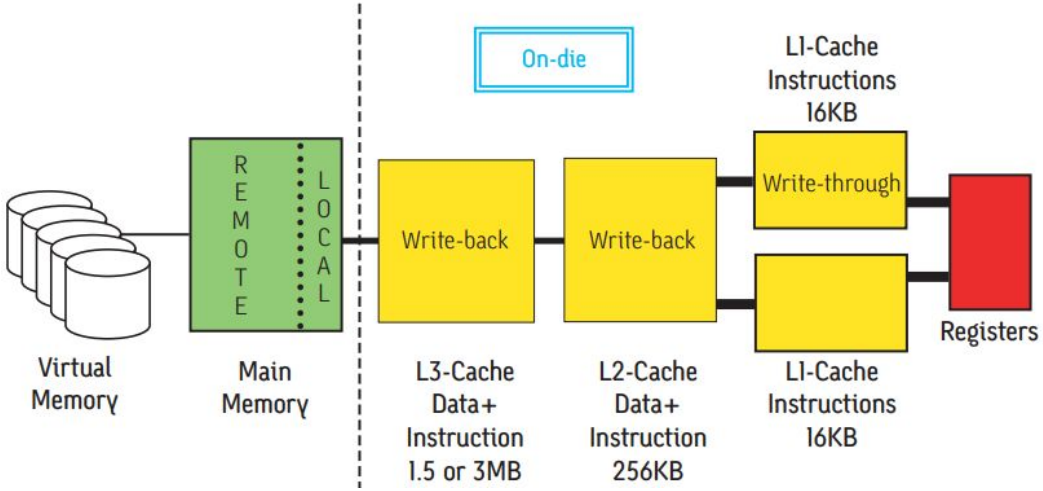



1U blank panel
3U blank panel
C-brick
C-brick
C-brick
R-brick
R-brick
Power bay
C-brick
C-brick
R-brick
R-brick
C-brick
C-brick
Power bay
C-brick
IX-brick


C-brick
Memory
Processor
Processor
Super Bedrock
I/O
Processor
Processor
Super Bedrock
I/O
Memory
To network fabric
R-brick
C-brick
Memory
Processor
Processor
Super Bedrock
I/O
Processor
Processor
Super Bedrock
I/O
Memory
C-brick
Memory
Processor
Processor
Super Bedrock
I/O
Processor
Processor
Super Bedrock
I/O
Memory
R-brick
To network fabric
C-brick
Memory
Processor
Processor
Super Bedrock
I/O
Processor
Processor
Super Bedrock
I/O
Memory
Memory
Processor
Processor
Super Bedrock
I/O
Processor
Processor
Super Bedrock
I/O
Memory

Processors

- Intel Itanium 2 Microprocessor
- 1 C-Brick
  - 4 processors
  - 32 GB memory
  - 2 SHUBs
- Up to 512 on a single Altix (later 1024)

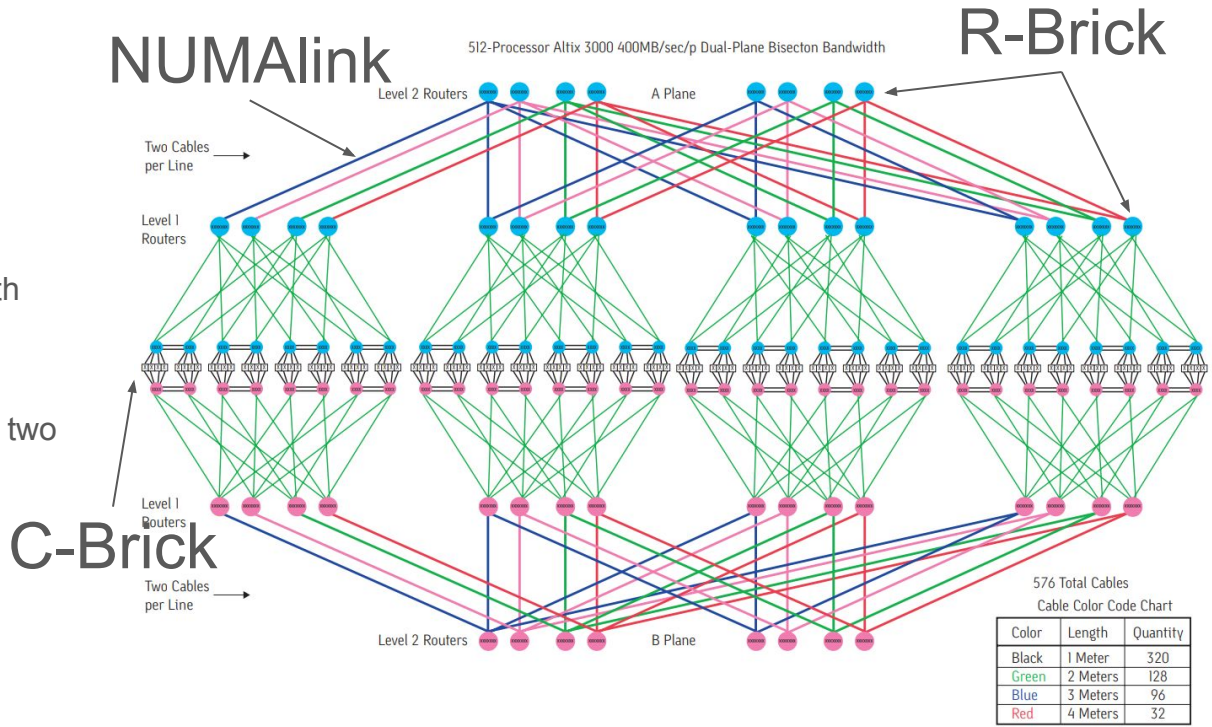


Cache Coherency

- Copies of data can exist
- Uses snooping of CPU in addition to directory-based cache-coherence
- Transparent to software/user
- Directory-Based Cache-Coherence
  - Keep “directory” of what processes own which cache lines and their current state
  - More scalable than snooping
  - SGI uses combination of course and full bit vector format
  - Adds memory overhead

Network

- Dual-plane fat-tree
  - Increases bisection bandwidth
  - Adds reliability
- NUMAlink 4 interconnect
  - 6.4 GB/s peak bandwidth via two 3.2 GB/s unidirectional links
- Can also use other vendors



Color	Length	Quantity
Black	1 Meter	320
Green	2 Meters	128
Blue	3 Meters	96
Red	4 Meters	32

Reliability, Availability, and Serviceability

- Redundant fans and power supplies
- Memory and cache error coding
  - Correct 1 bit errors
  - Identify >1 bit errors
- Messages use cyclic redundancy check to re-send bad messages
- Half of network can go down
- Bricks are easy to be added, removed, and/or swapped

OS and Software

- SUSE Linux Enterprise Server
- OS can run across 64 processors
  - Later increased to 512
- SGI Message Passing Toolkit
  - Tuned implementations of MPI and SHMEM APIs
  - Uses XPMEM API for underlying system calls
- Also works with other MPI implementations

Altix 3000 in the Wild

- NASA's Columbia supercomputer was Altix 3000 based
- Ran from 2004-2013
- Consisted of 20 Altix 3000s each with 512 microprocessors
- Connected by Infiniband

Rank	System	Cores	Rmax (GFlop/s)	Rpeak (GFlop/s)	Power (kW)
1	**BlueGene/L beta-System** - BlueGene/L DD2 beta-System (0.7 GHz PowerPC 440), IBM IBM/DOE United States	32,768	70,720.0	91,750.0	
2	**Columbia** - SGI Altix 1.5 GHz, Voltaire Infiniband, HPE NASA/Ames Research Center/NAS United States	10,160	51,870.0	60,960.0	
3	Earth-Simulator, NEC Japan Agency for Marine-Earth Science and Technology Japan	5,120	35,860.0	40,960.0	3,200