Partitioning on Origin 3000 and SNIA

Russ Anderson

RAS/Partitioning Project Tech Lead

rja@sgi.com

Agenda

- Partitioning Overview
- Partitioning Successes
- Current Status
  - Origin 3000 (Mips/Irix)
  - SNIA (Intel/Linux)

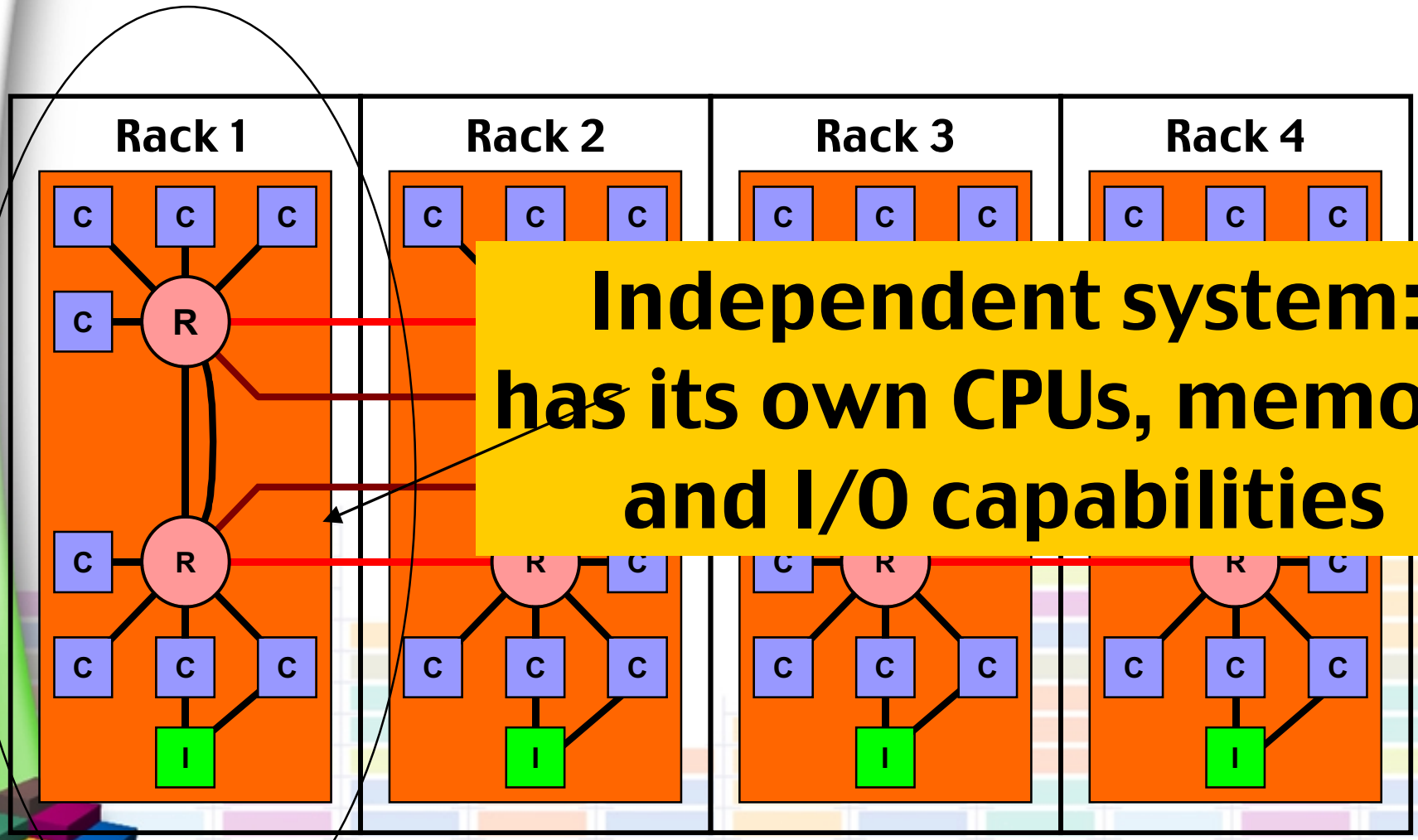
Agenda

- Partitioning Overview
  - What is partitioning?
  - Advantages of Partitioning
  - Partitioning Specifics

Partitioning Overview

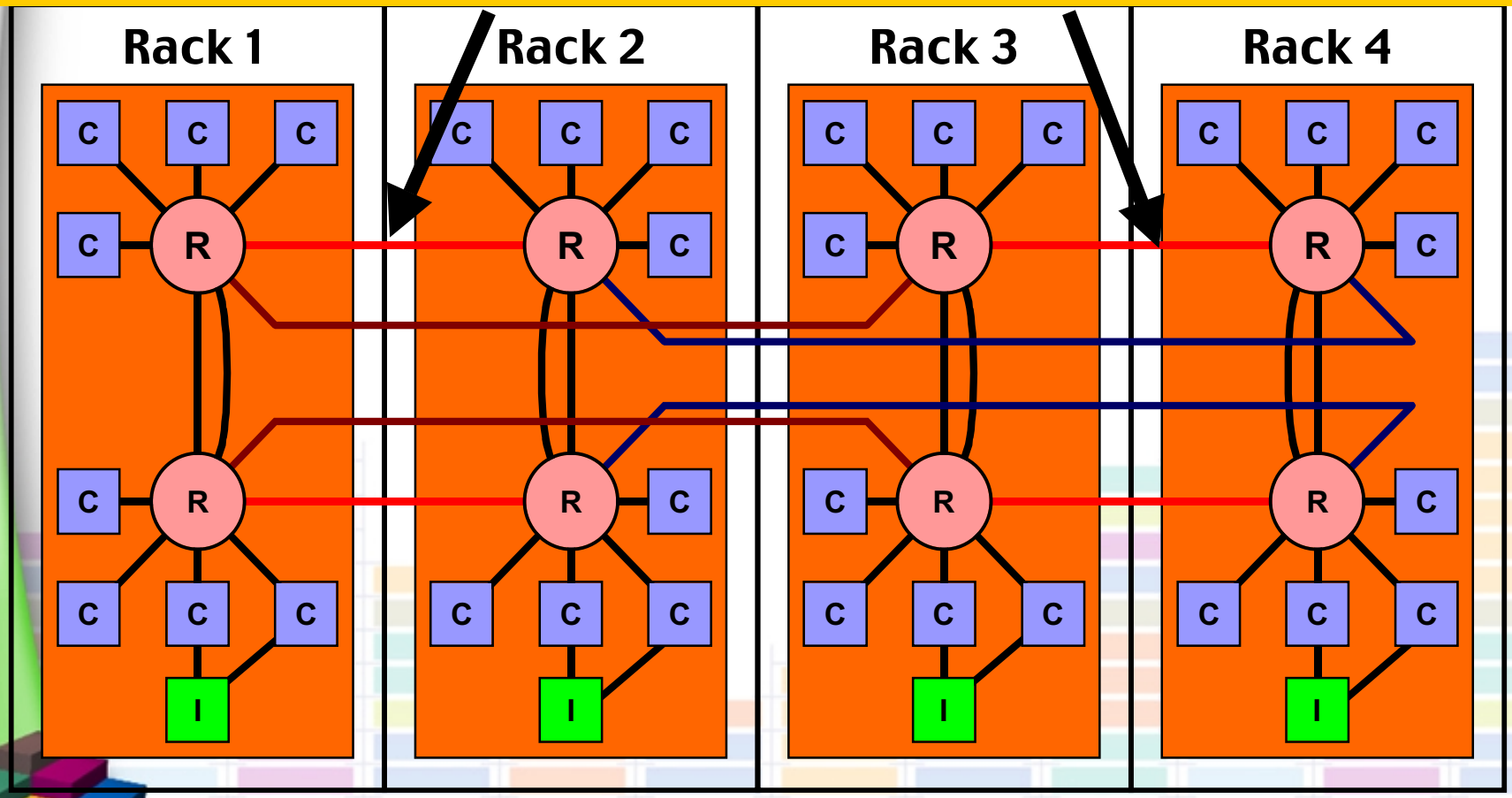
- What is Partitioning?
  - The ability to run a large system as multiple smaller independent systems
  - Use NUMAlink as a fast interconnect between partitions (TCP/IP, MPI)

Partitioning Overview



Partitioning Overview

NUMAlink: Memory to memory communication.



Partitioning Overview

- Advantages of Partitioning
  - Flexible configuration
  - Use NUMAlink as a fast interconnect between partitions
  - Fault containment
  - Enhanced HW serviceability

Partitioning Overview

- Advantages of Partitioning
  - Flexible configuration
    - Change configurations without re-cabling
    - Rolling OS Upgrades
    - Small kernel image & Large kernel image in single system
    - Change partition size (Linux)

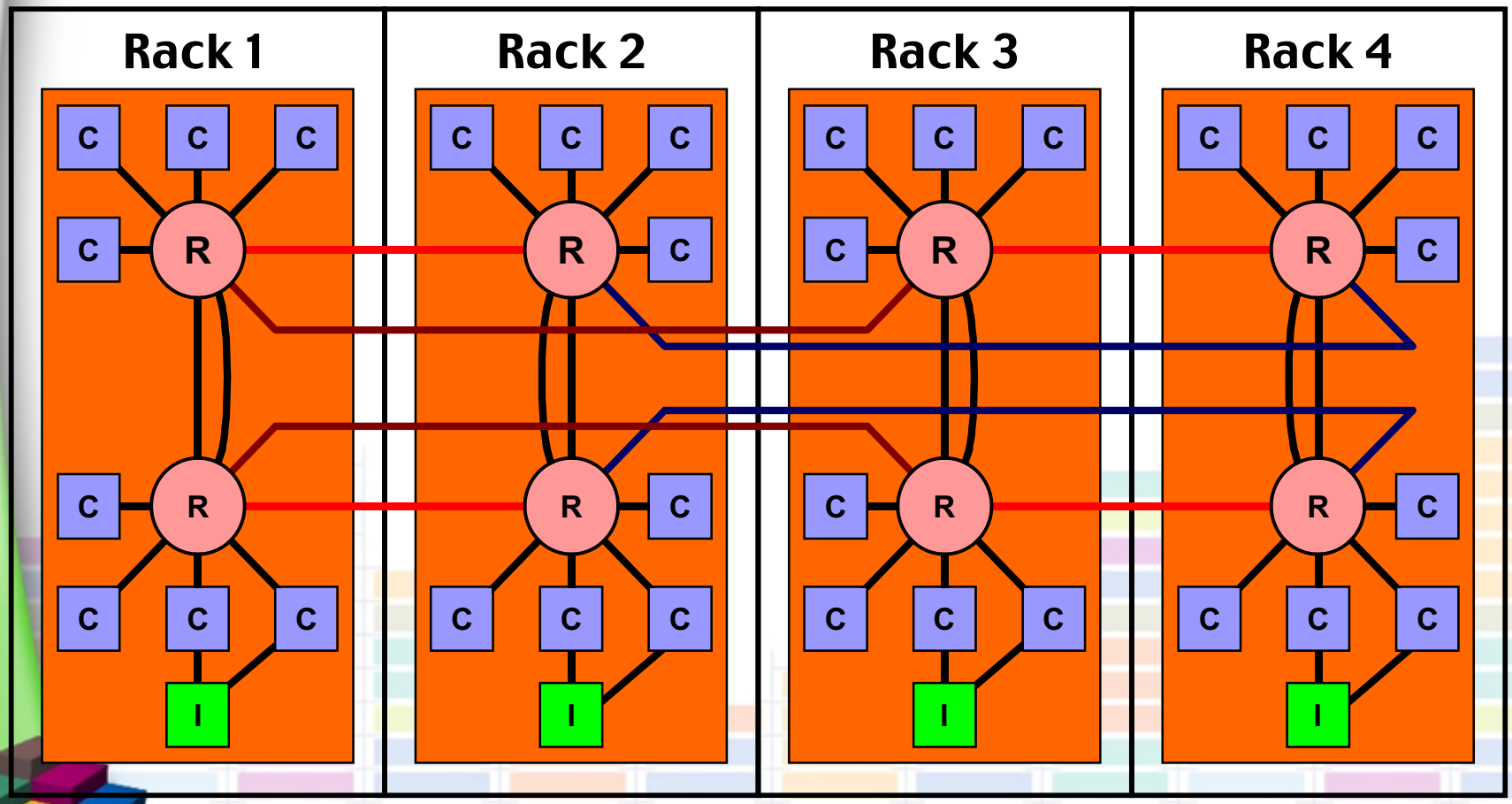
Partitioning Overview

- Advantages of Partitioning
  - Use NUMAlink as a fast interconnect between partitions
    - TCP/IP communication (NFS, CXFS)
    - MPI / XPMEM

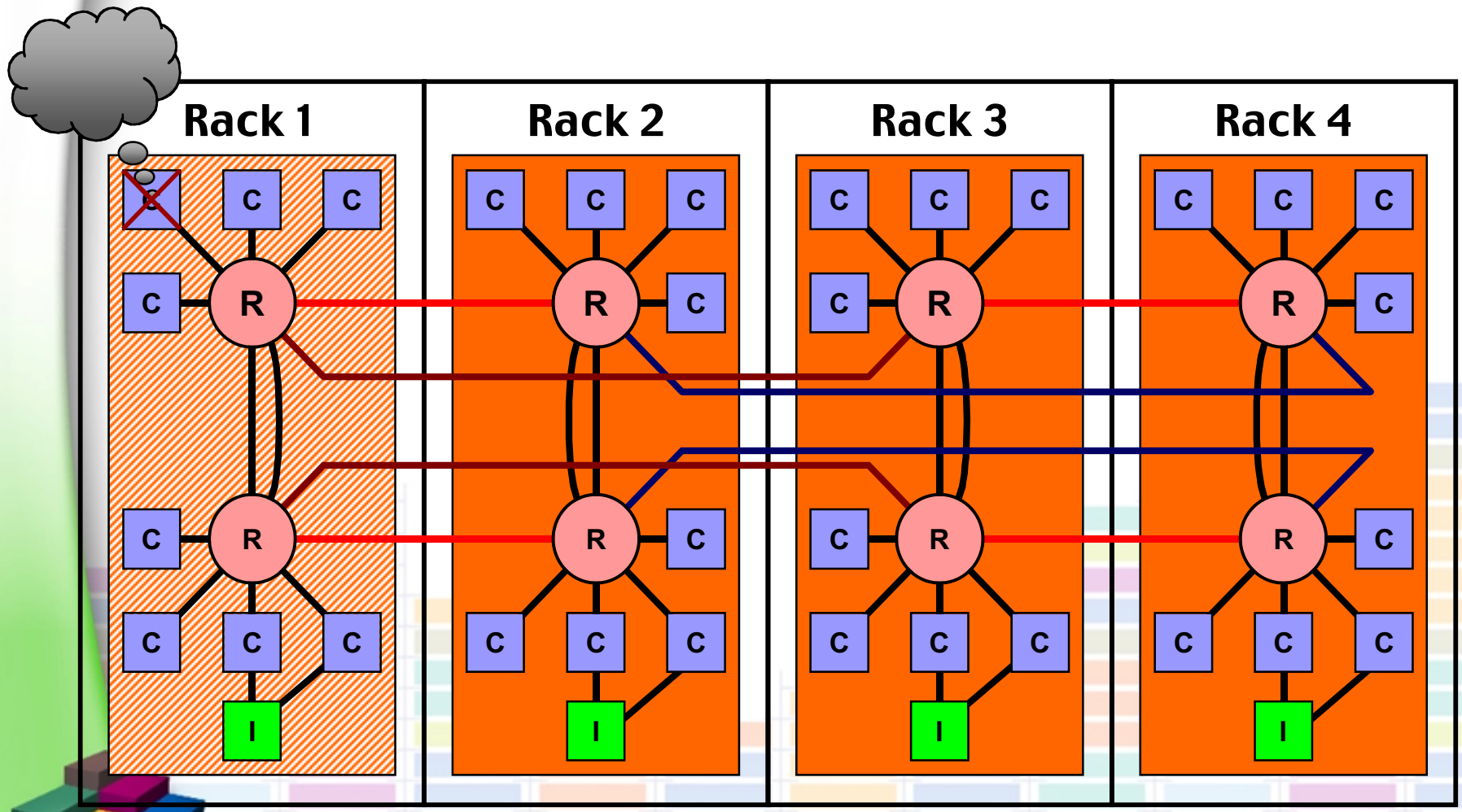
Partitioning Overview

- Advantages of Partitioning
  - Fault containment
    - Increase MTTI
    - Maximize availability
  - Enhanced HW serviceability

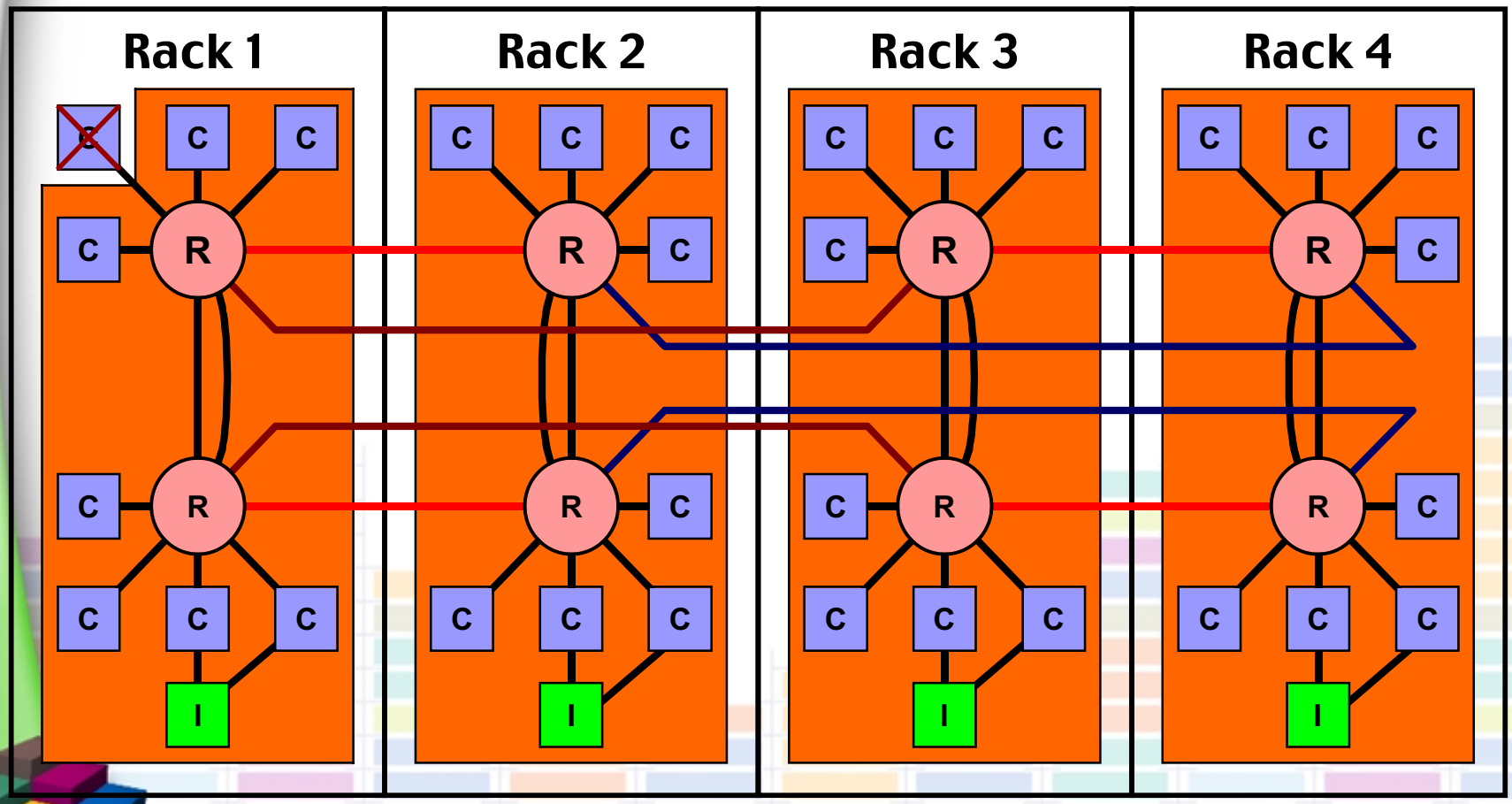
Partitioning Example



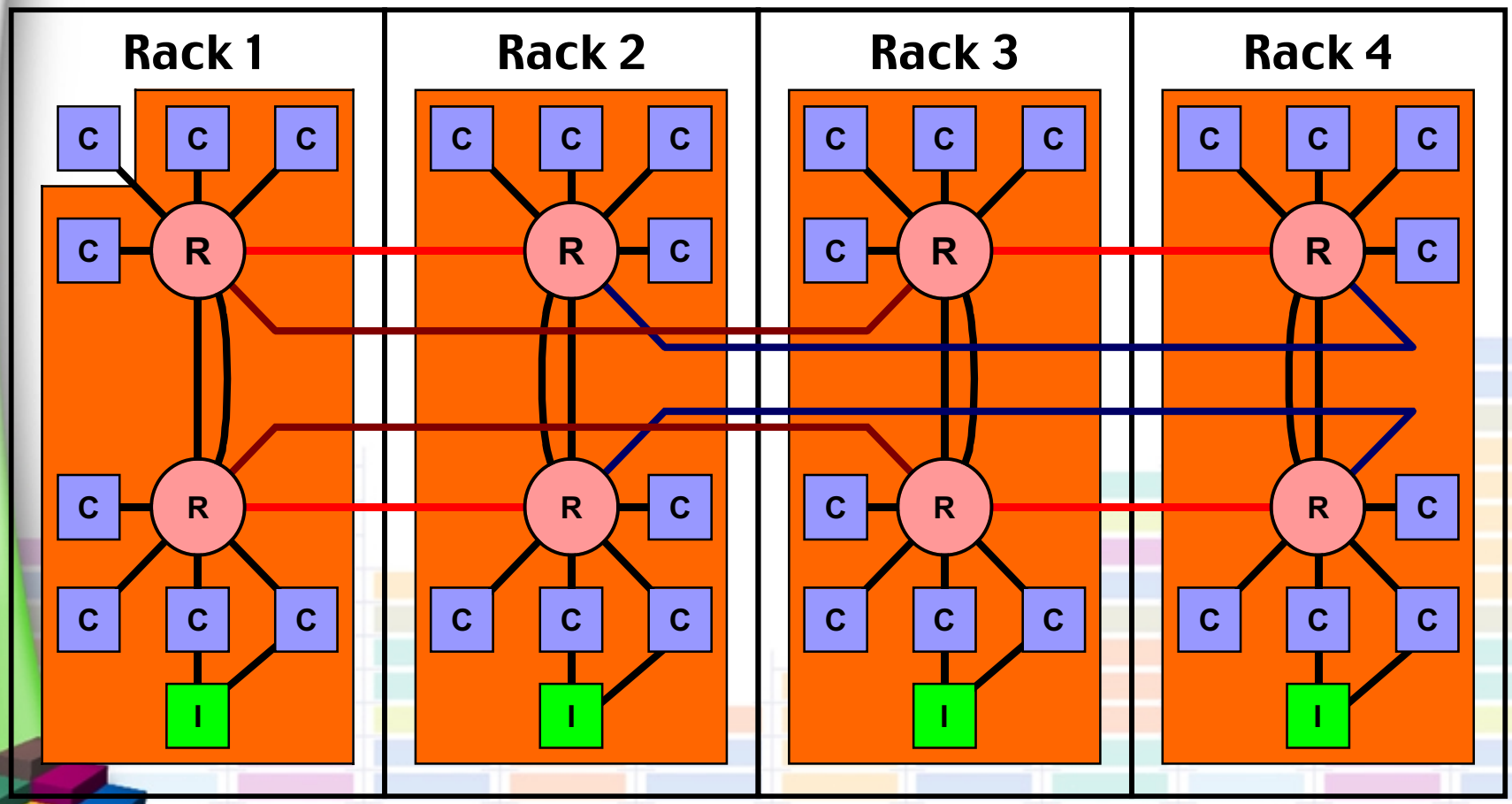
Partitioning Example



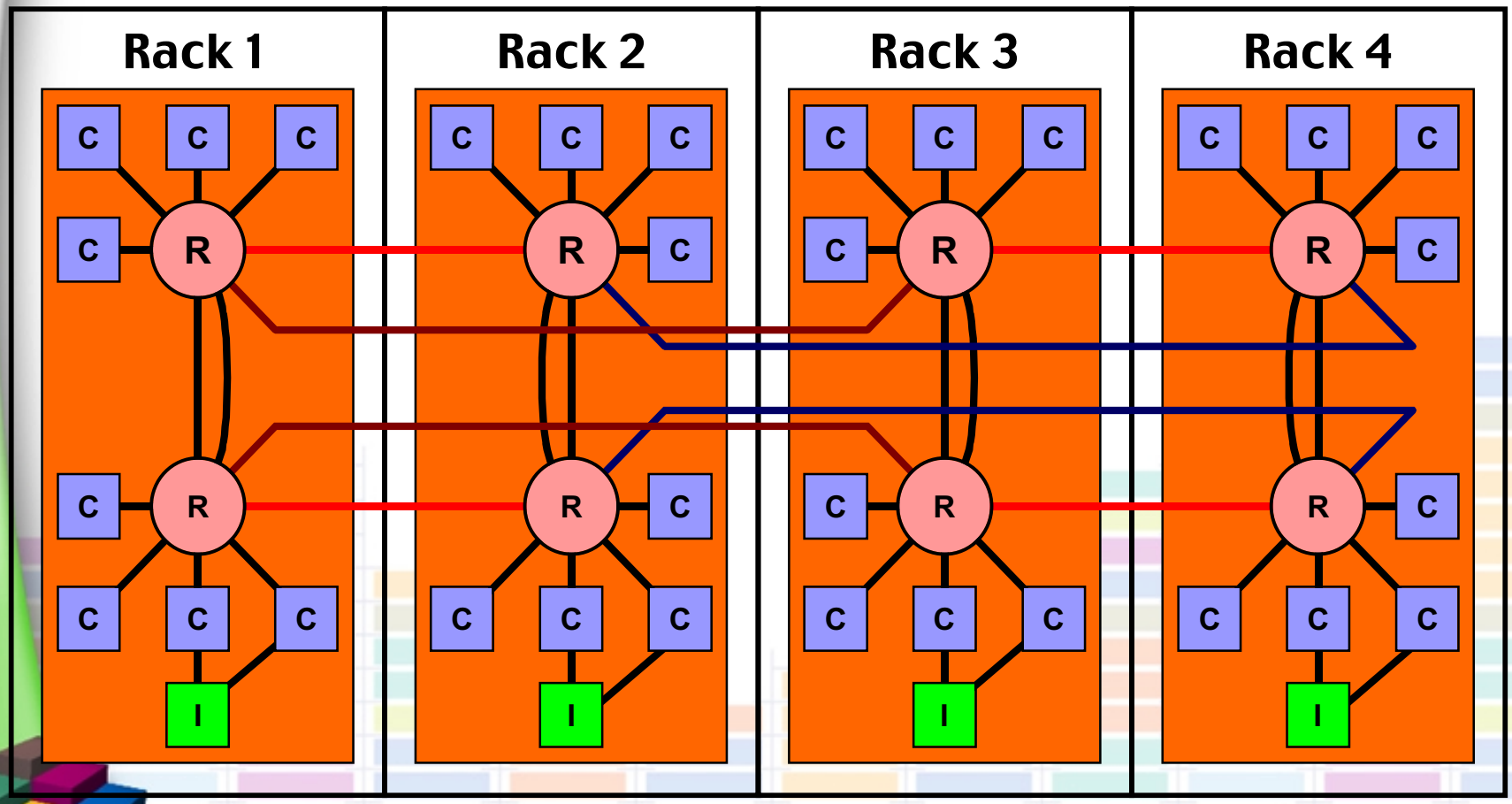
Partitioning Example



Partitioning Example



Partitioning Example



Origin 3000 Partitioning

- Partitioning Specifics
  - Origin 3000 Architecture Overview
  - Hardware support for partitioning
  - Partitioning Configuration Rules
  - Brick replacement

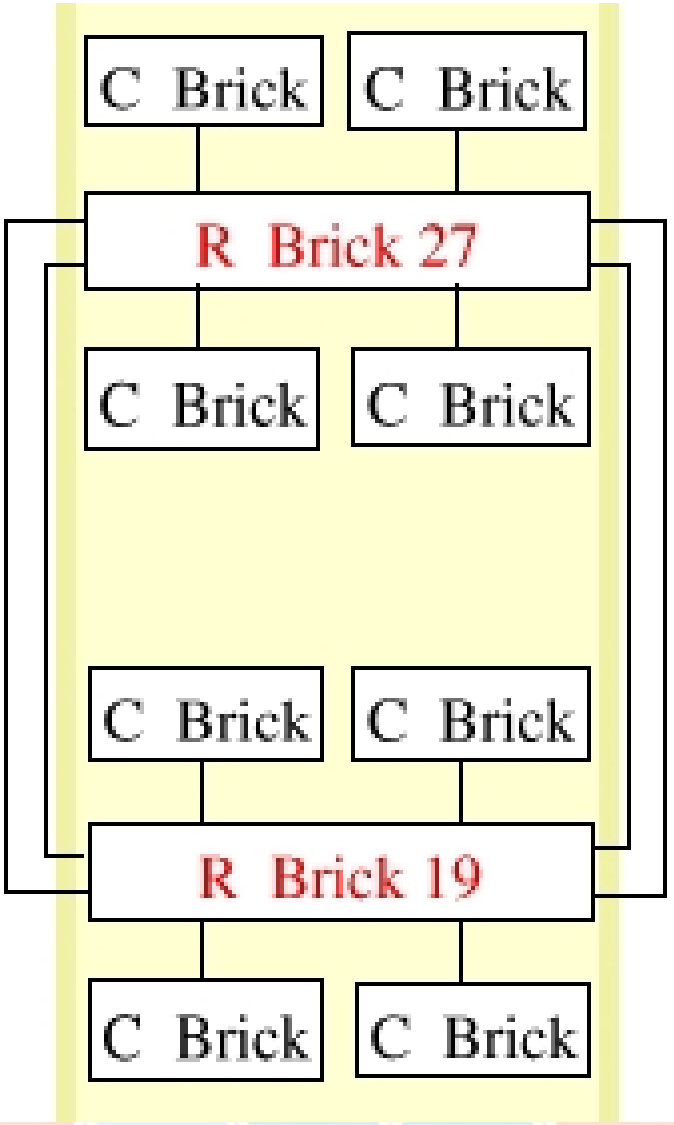
Origin 3000 Partitioning

- NUMAflex technology
  - Bricks

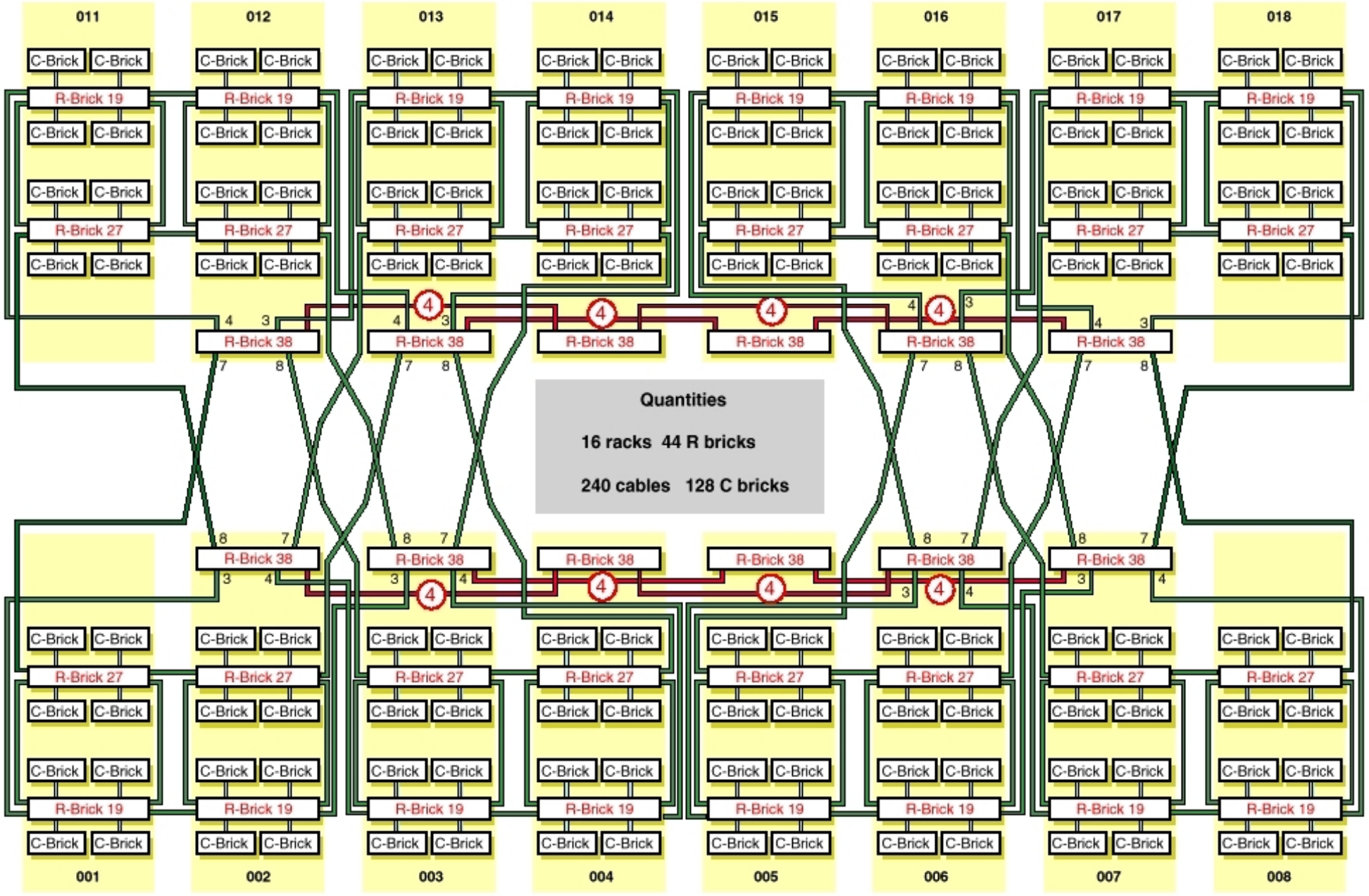
    C-Brick - Compute (MIPS)
    I-Brick - Basic I/O slots/Root disk/Etc.
    X-Brick - XTALK I/O Brick
    P-Brick - PCI I/O Brick
    R-Brick - Router Brick
  - Building Block Approach

Origin 3000 16-32p

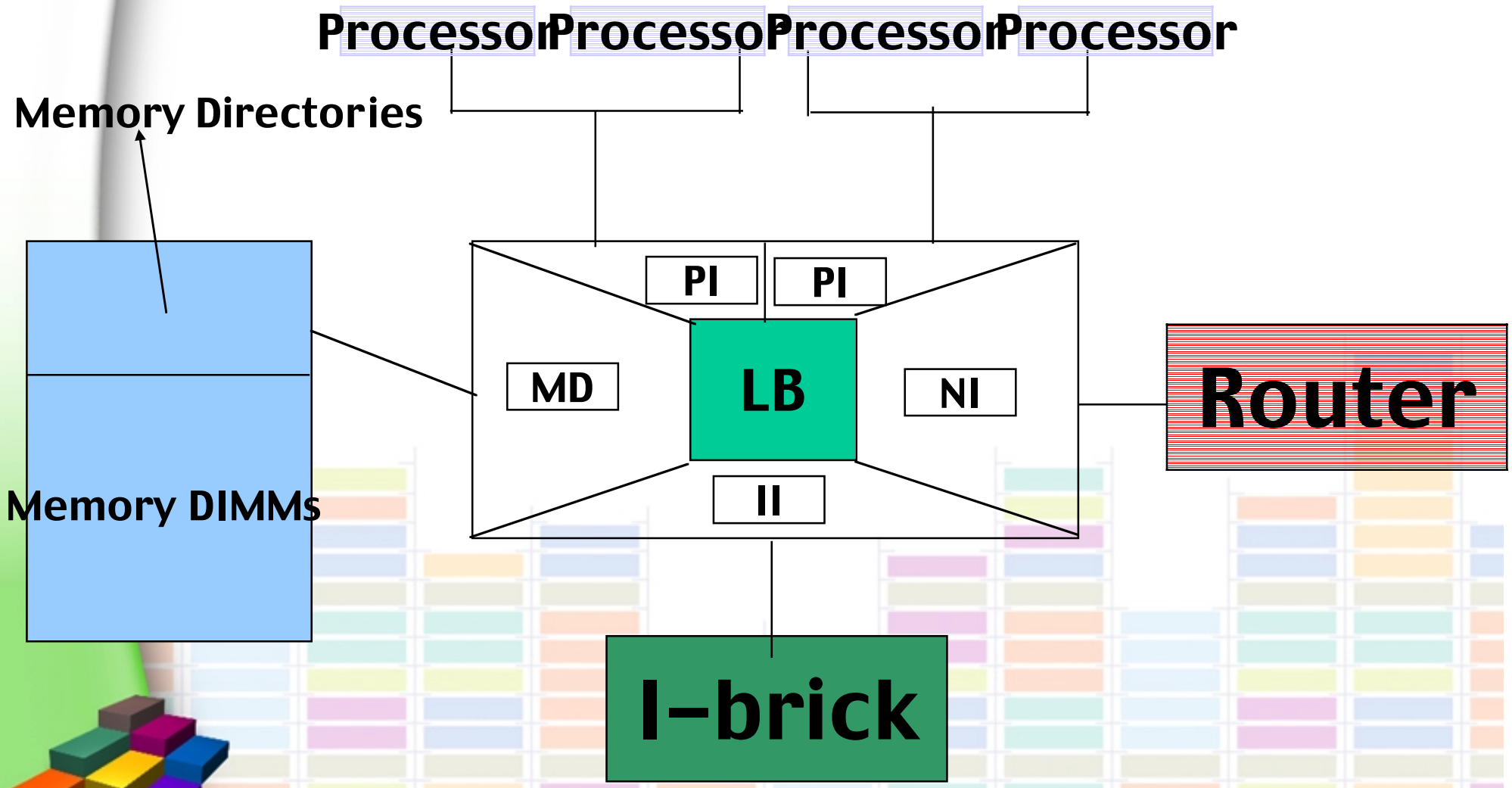
- Rack



Origin 3000 512p



SN1 Architecture



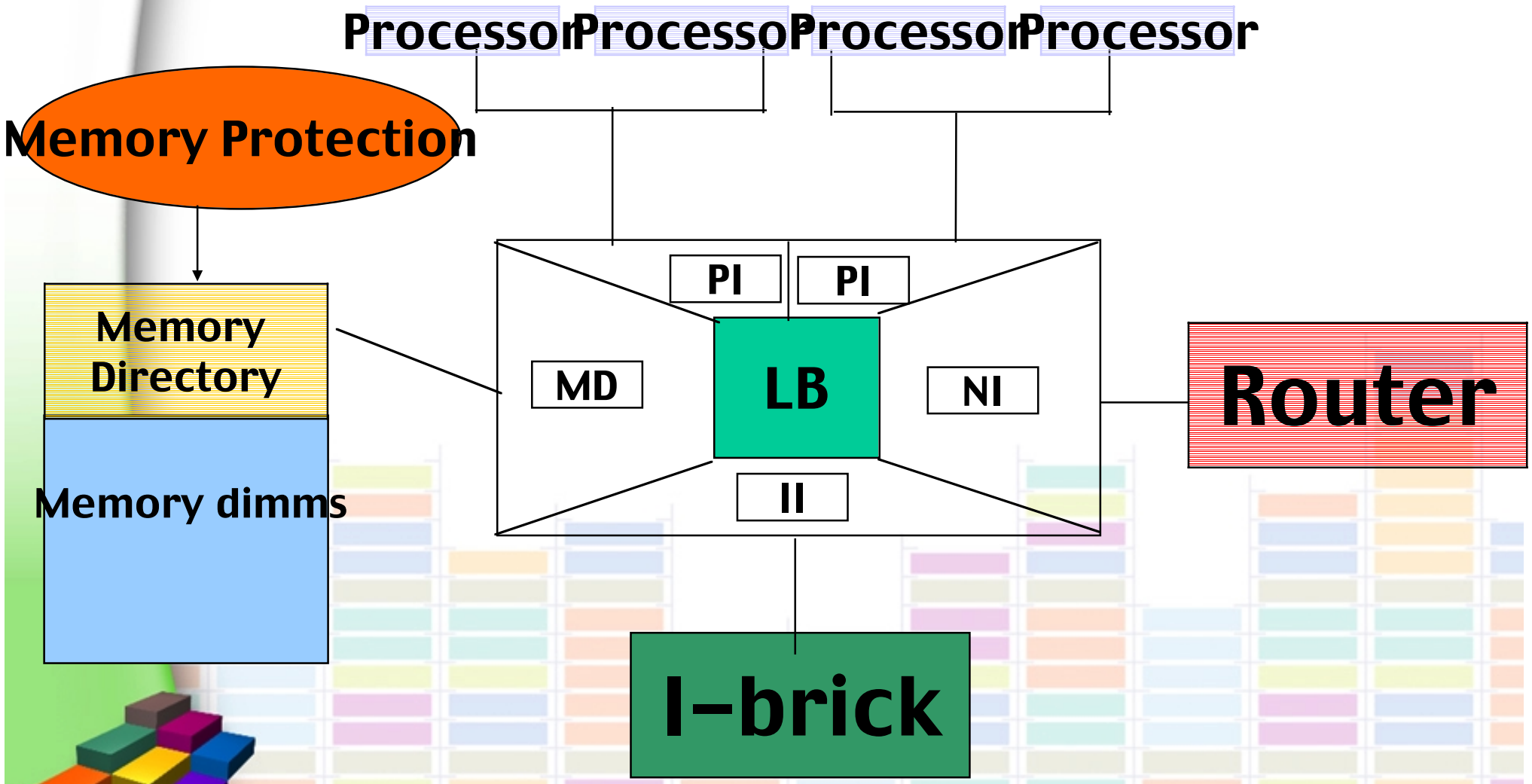
Partitioning Hardware Support

- Partitioning Hardware Support
  - Memory Protection
  - Reset Fences
  - BTE and NUMAlink

Partitioning Hardware Support

- Memory Protection
  - Built into Hub chip hardware
  - Protect a partition from unexpected writes from another partition (Fault Containment)
  - Protection can be modified to allow access to specific memory pages

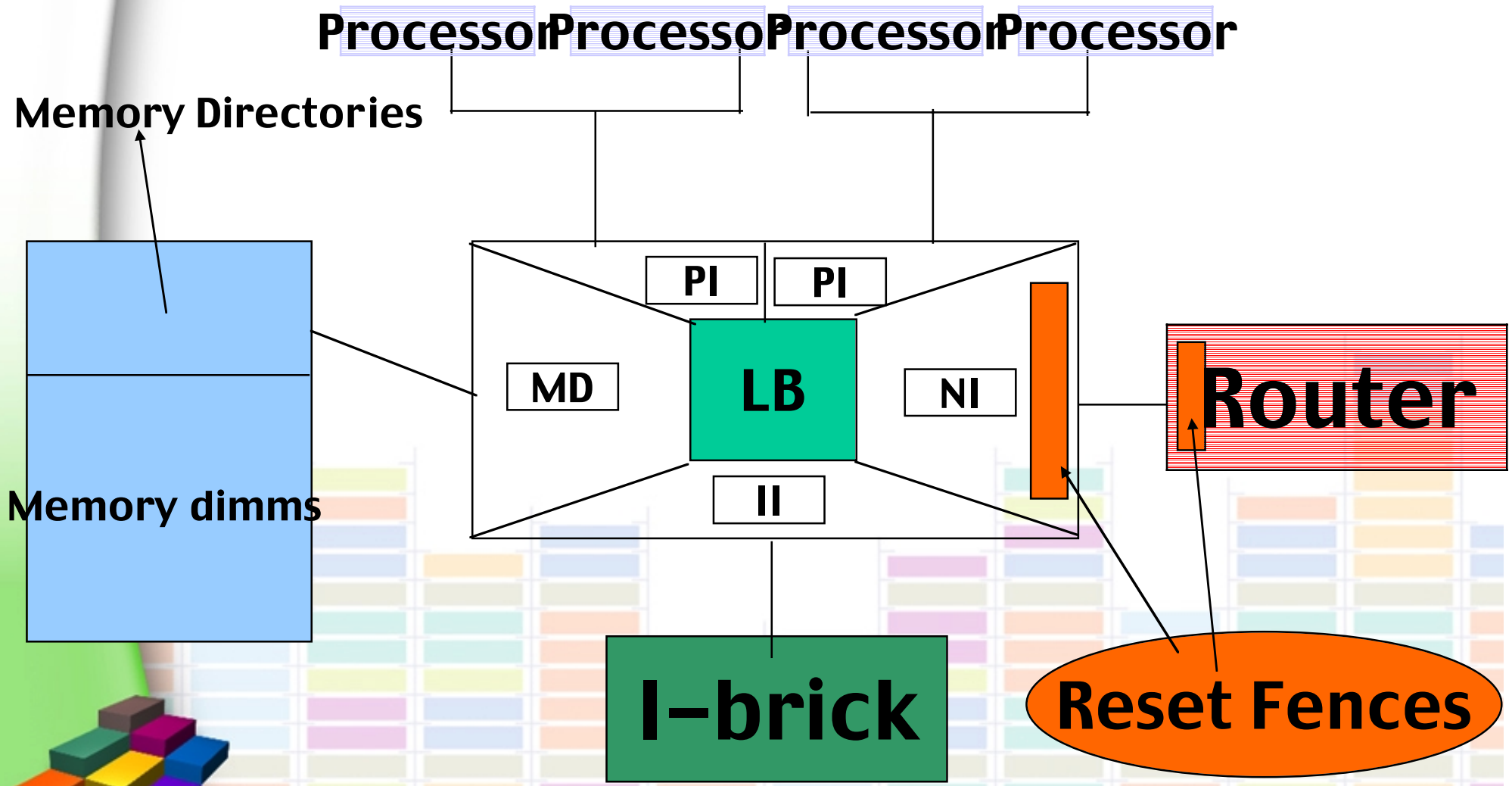
Origin 3000 Architecture Memory Protection



Partitioning Hardware Support

- Reset Fences
  - Built into Hub and Router chip hardware
  - Protect a partition from hardware resets in other partitions
  - Used to support concurrent brick replacement

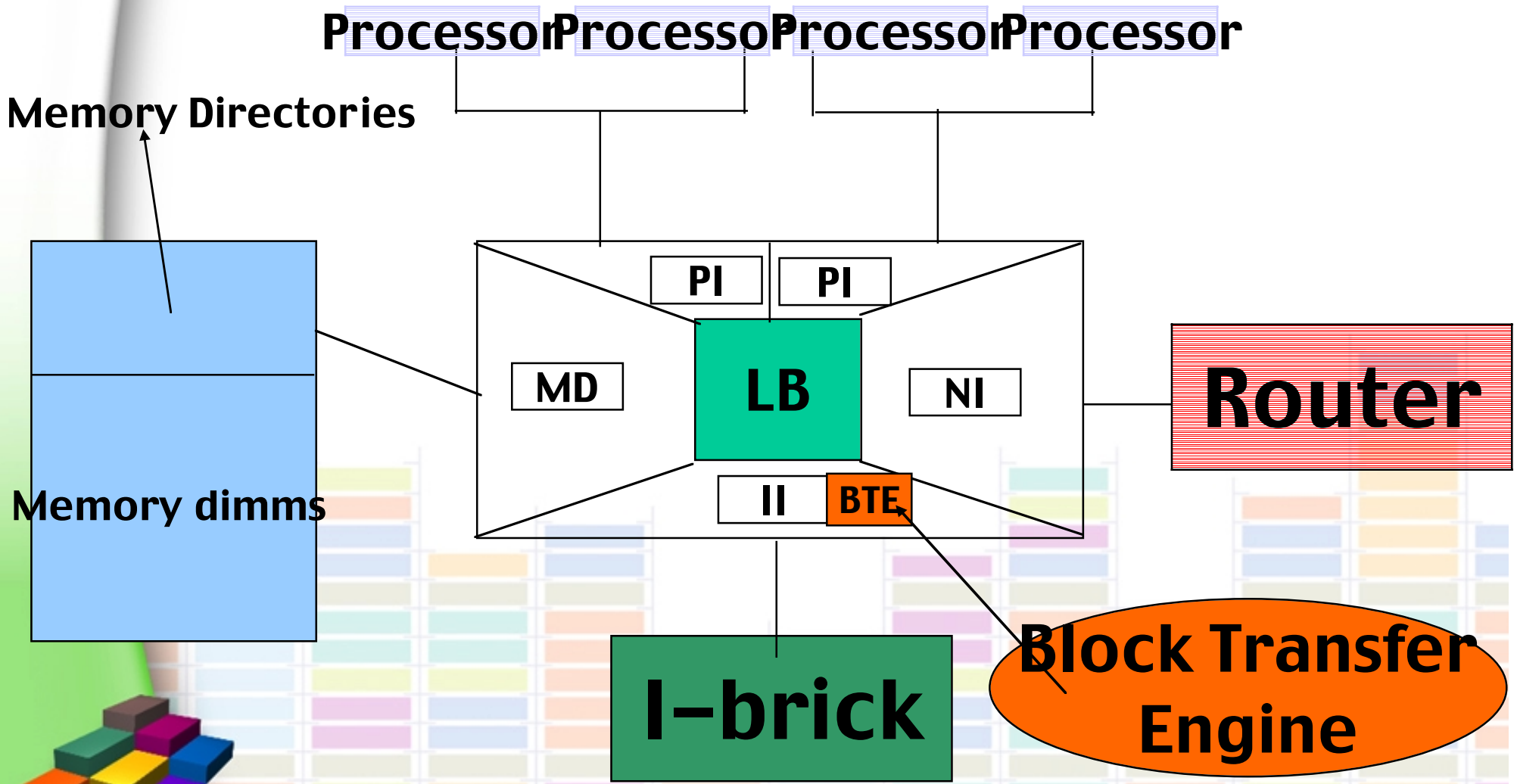
Origin 3000 Architecture Reset Fences



Partitioning Hardware Support

- Block Transfer Engine
  - Built into Hub chip hardware
  - Transfers data between partitions (without changing memory protection)
  - Allows processors to do other work

Origin 3000 Architecture Block Transfer Engine



Configuration Rules

- Current Partitioning Rules
  - Partitions Run like stand alone systems
  - Partition on rack & half-rack boundaries
  - I/O Bricks belong to the partition of the attached C-Brick

Concurrent Maintenance

- C-Brick Replacement
  - Power down independently
  - Diagnostic Testing
  - Re-integrate with partition Reboot

Agenda

- Partitioning Overview
- **Partitioning Successes**
- Current Status
  - Origin 3000 (Mips/Irix)
  - SNIA (Intel/Linux)

Origin 3000 Partitioning Successes

- Partitioned system running at SARA
  - 128p partitioned into 2x64p
- 512p partitioned 2x256p and 4x128p
- 192p (6 rack) system partitioned (4x32p, 1x64p) in production
- Production + development partitioned system
- Partitioned systems with GFX & GSN

Origin 3000 Current Status

- Irix 6.5.9 - Hard wall partitioning supported
- Irix 6.5.10 - Cross Partition communication (TCP/IP)
- Irix 6.5.11-12 - More supported configurations
- Irix 6.5.13 -
  - XPMEM (support for MPI)
  - BTE performance enhancements
  - Improved system administration

SNIA Current Status

- Itanium/Linux SNIA development system running two partitions with TCP/IP over NUMAlink
- SNIA low-level functionality working
  - Reset fences
  - Memory protection

Origin 3000 Partitioning

Conclusion