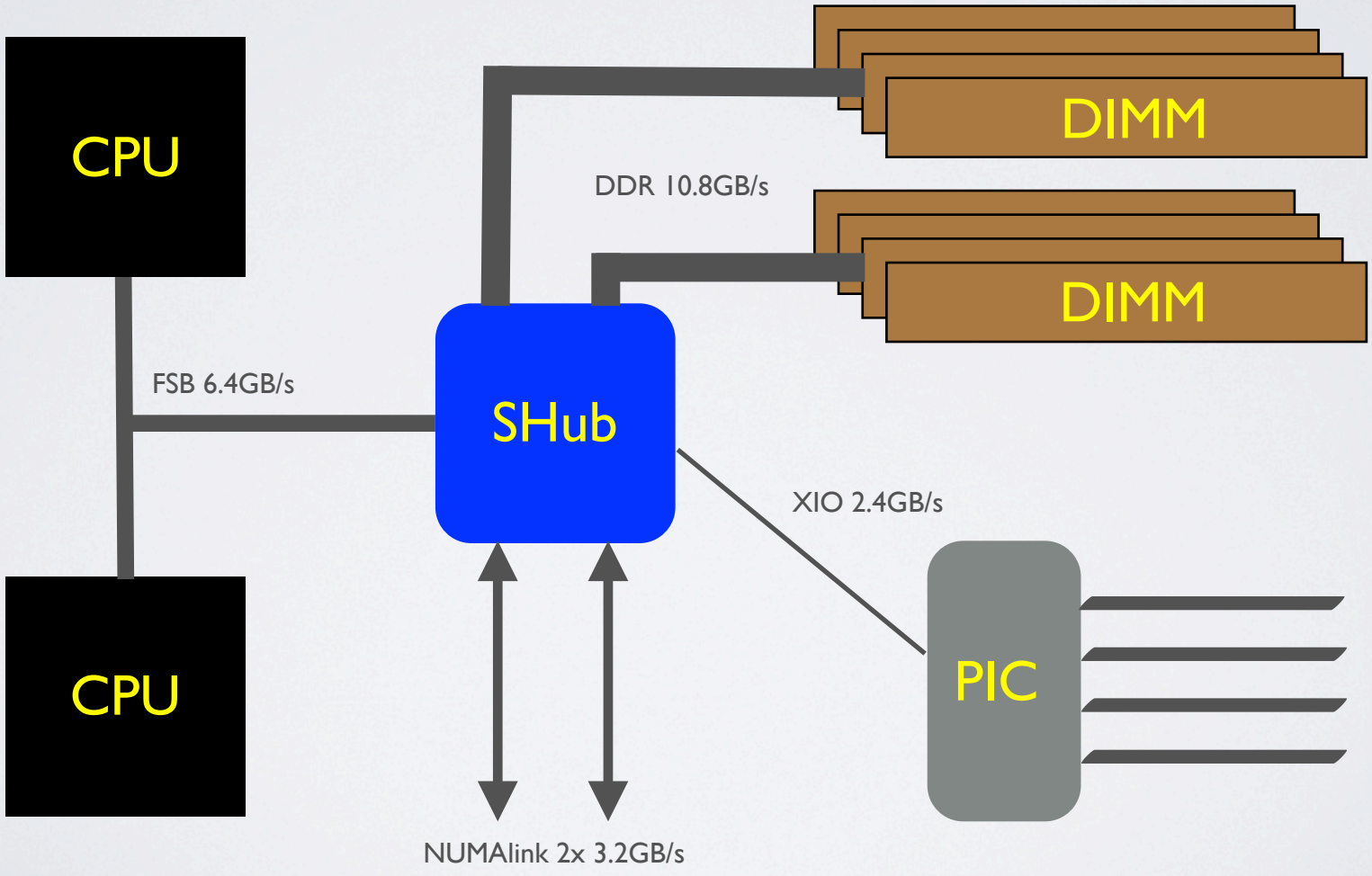
CC-NUMA SUPPORT

Marcel Moolenaar

COMPUTE NODE



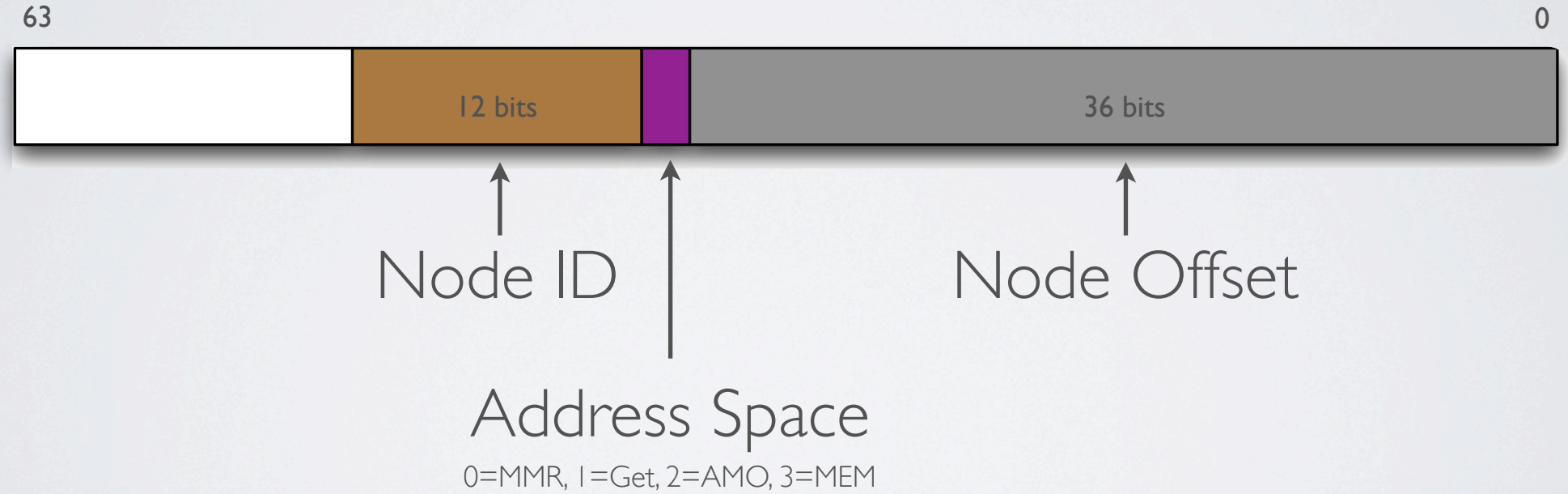
ARCHITECTURE

- Same architecture as Origin 3000
- Compare with:
  - AMD's Direct Connect Architecture
  - Intel's QuickPath Interconnect

THE NUMA PROBLEM

- Maximize locality of reference:
  - Text/data replication
  - Data migration
  - Eliminate false sharing

GLOBAL SHARED MEMORY



BOOTING FREEBSD

- Use EFI's memory allocator:
  - Non-deterministic physical placement
- Add paged virtual addressing (aka PBVM):
  - arch_loadaddr() - handle alignment based on object loaded
  - arch_loadseg() - inform MD code about loaded segments
- Relink kernel against Pre-Boot Virtual Memory.

DEVICE ENUMERATION

- BSP enumerates devices:
  - Devices are enumerated from within all nodes
  - Memory is allocated without affinity
- Driver memory ideally allocated close to device

ALLOCATION PRINCIPLE

- Given maximum locality of reference...
- ... allocating memory close to current CPU is optimal
- Add M_NUMA_LOCAL malloc flag:
  - Set -- fail allocation if no free local memory exists
  - Unset -- allocate “distant” memory if needed

DEVICE ENUM -- REVISITED

- Enumerate local (to current CPU) device only
- Bootstrap "monarch" CPU in each node first
- Consequence: concurrent enumeration across nodes
  - Pros: scales well with large number of nodes
  - Cons: non-deterministic unit assignment without extra effort
- Monarchs can start APs within their nodes

TO DO (ALTIX 350)

- Atomic operations (need to go through AMO address space)
- DMA: unified busdma implementation with I/O MMU support
- Replication of text and R/O data before monarch bootstrap
- Memory allocation control
- Data migration support

TO DISCUSS

- Should we start monarchs early and enumerate I/O locally?
- Can we parallelize device enumeration?
  - what to do about the order of devices attached?
- What are good interfaces for allocating physical memory?
  - should we take advantage of newbus?
  - Special interface for remote memory?

TO DISCUSS

- When and how to replicate text and/or R/O data?
  - should we always replicate?
- Allow binding IRQs to remote CPUs?
- Policies for scheduling processes across nodes?
  - should struct proc & struct thread move with the process?
- What's the impact on synchronization primitives?

TO DISCUSS

- What is the impact on having non-uniform nodes?
  - Bus clock frequency differences
  - CPU and time counter frequency differences
  - SHub node ID bit position differences
- What does it mean to support MPI?
- Is soft partitioning support highly desirable or a must?