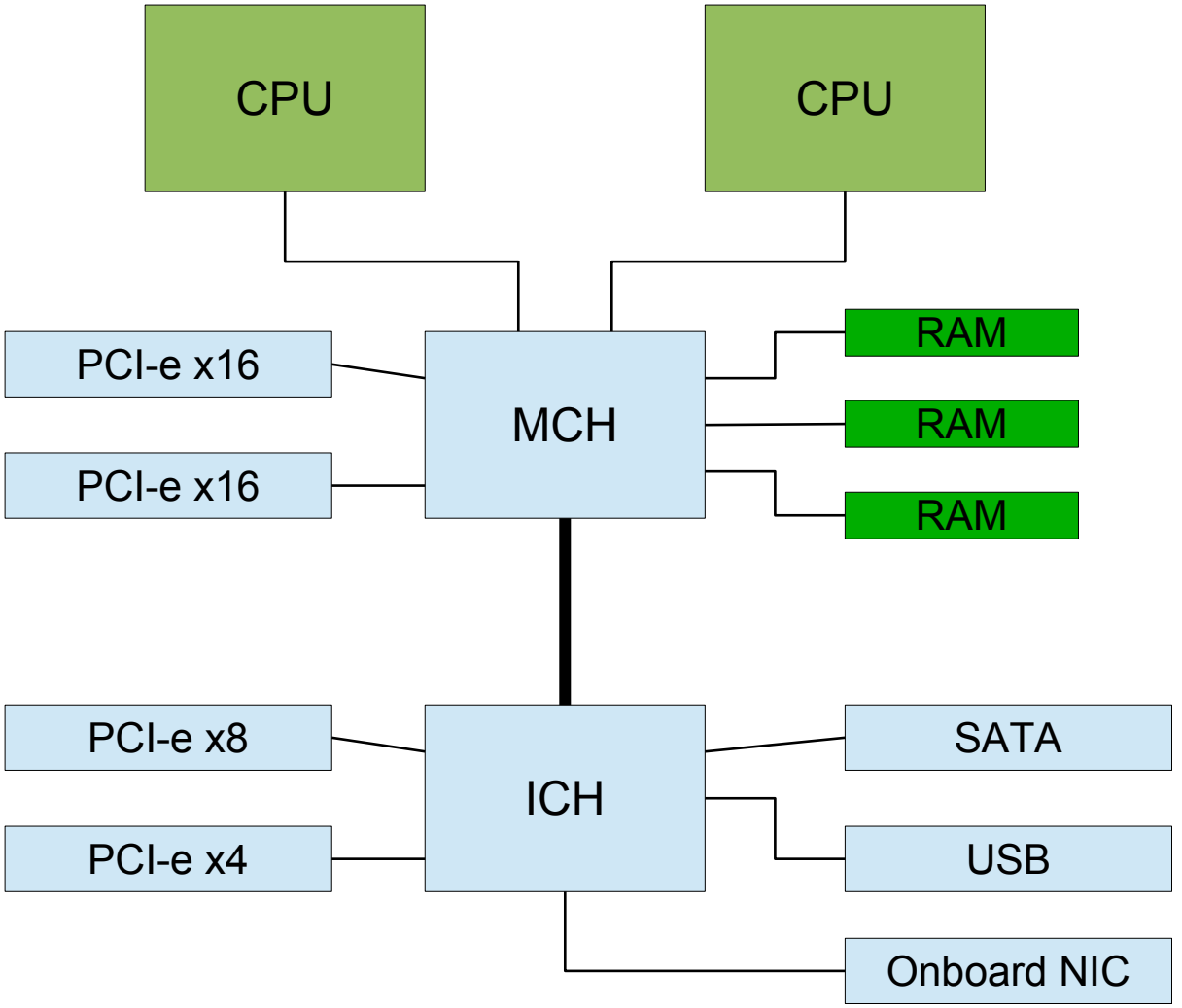
FreeBSD and NUMA

John Baldwin
NYC*BUG
June 3, 2015

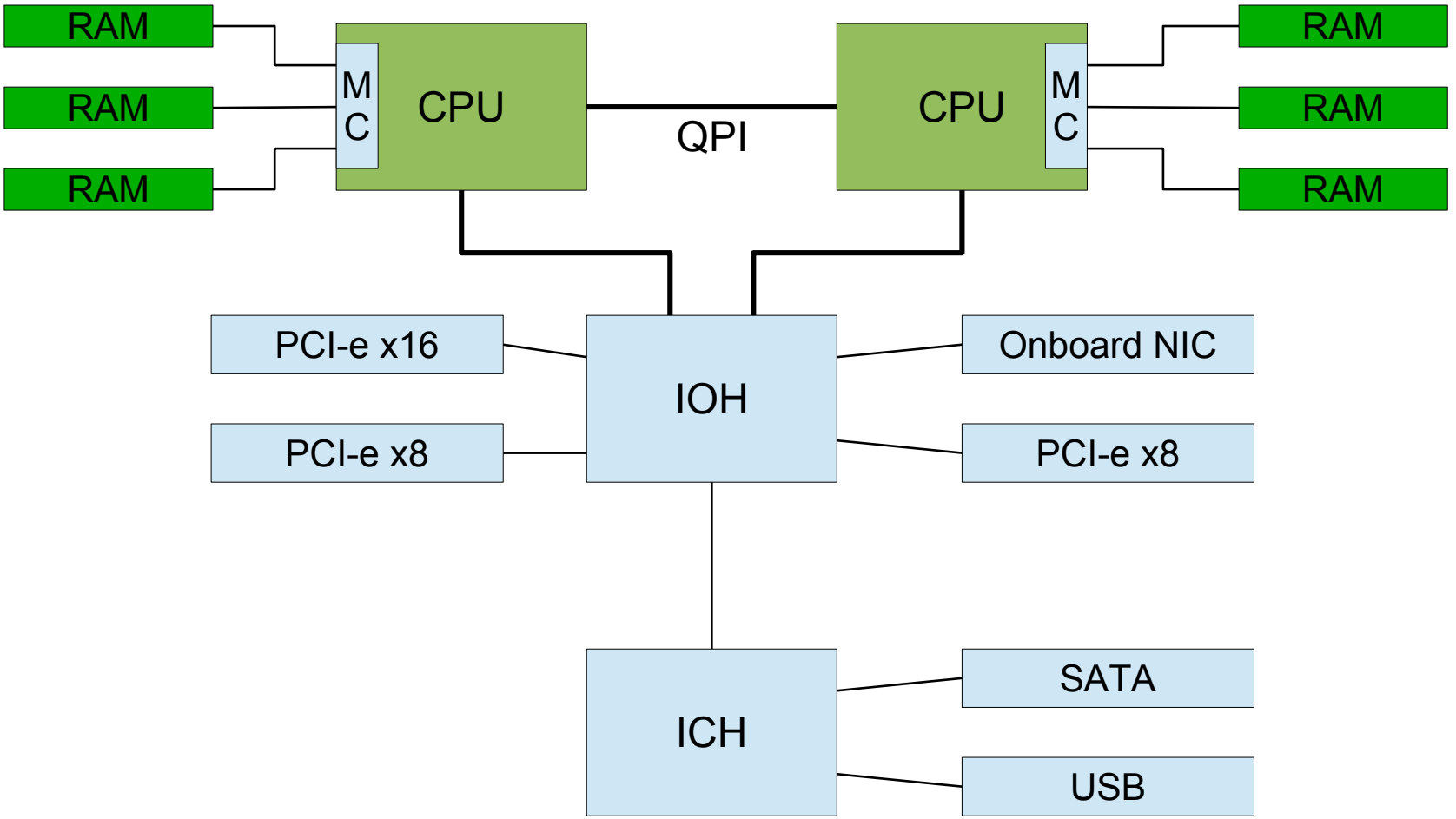
What is NUMA

- Non-Uniform Memory Architecture
- “Slow” vs “Fast” Memory
  - From CPUs
  - From I/O Devices
- Present on x86 starting with AMD Opterons (HyperTransport) and Intel Nehalem (QPI)

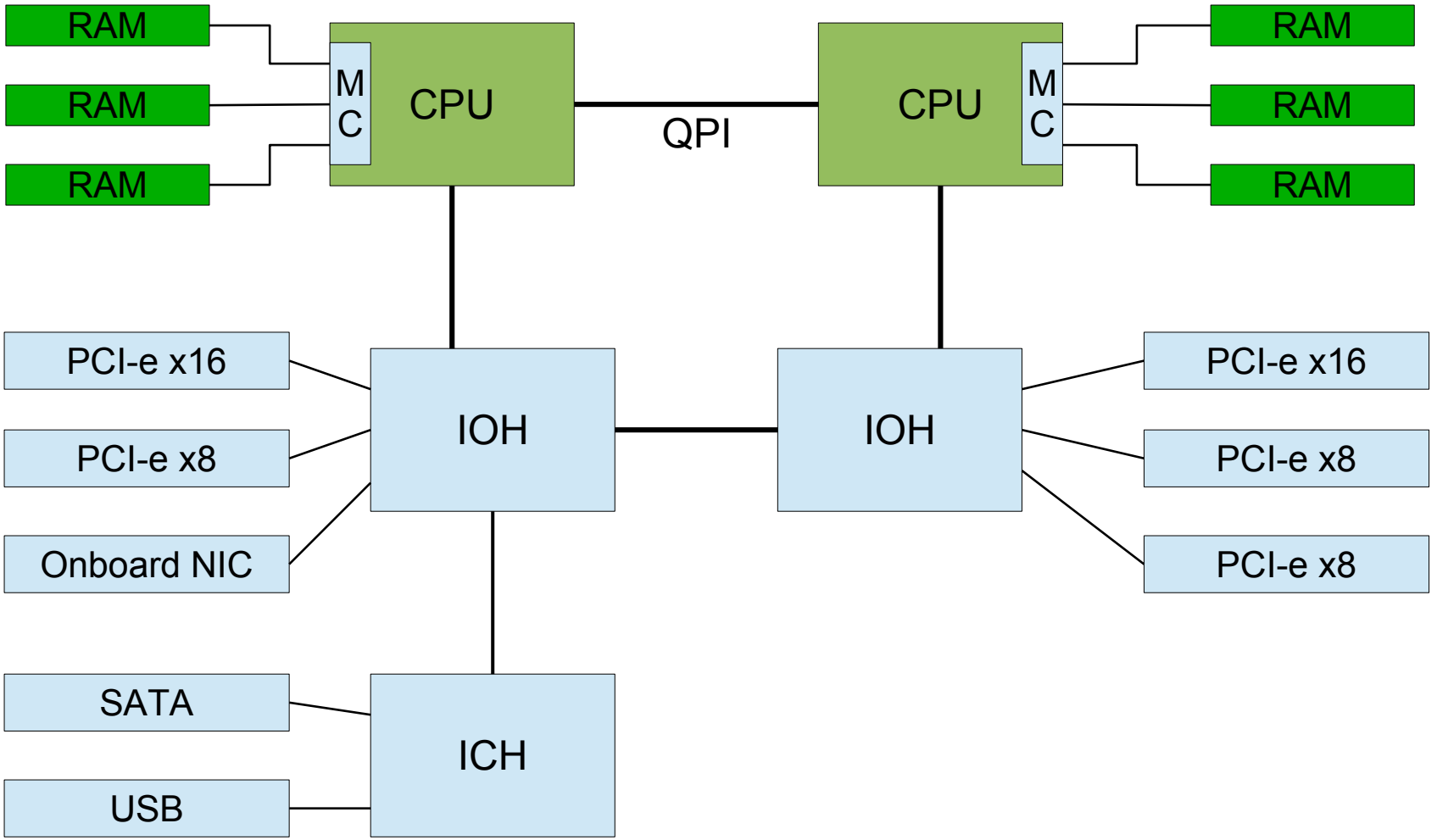
Front Side Bus (FSB)



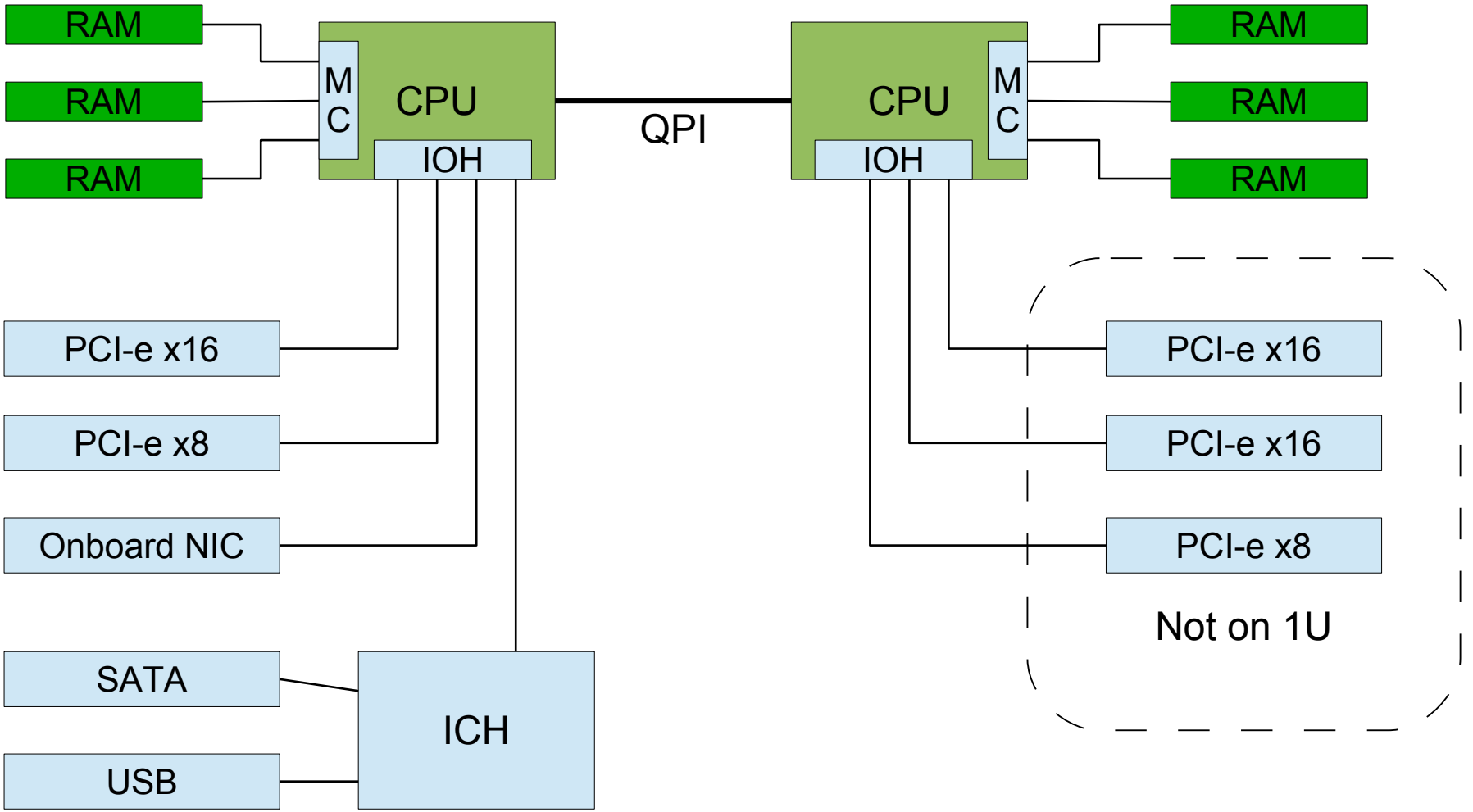
Nehalem 1U


RAM
RAM
RAM
MC
CPU
QPI
CPU
MC
RAM
RAM
RAM
PCI-e x16
PCI-e x8
IOH
Onboard NIC
PCI-e x8
ICH
SATA
USB

Nehalem 2U


RAM
RAM
RAM
MC
CPU
QPI
CPU
MC
RAM
RAM
RAM
PCI-e x16
PCI-e x8
Onboard NIC
IOH
IOH
PCI-e x16
PCI-e x8
PCI-e x8
SATA
USB
ICH

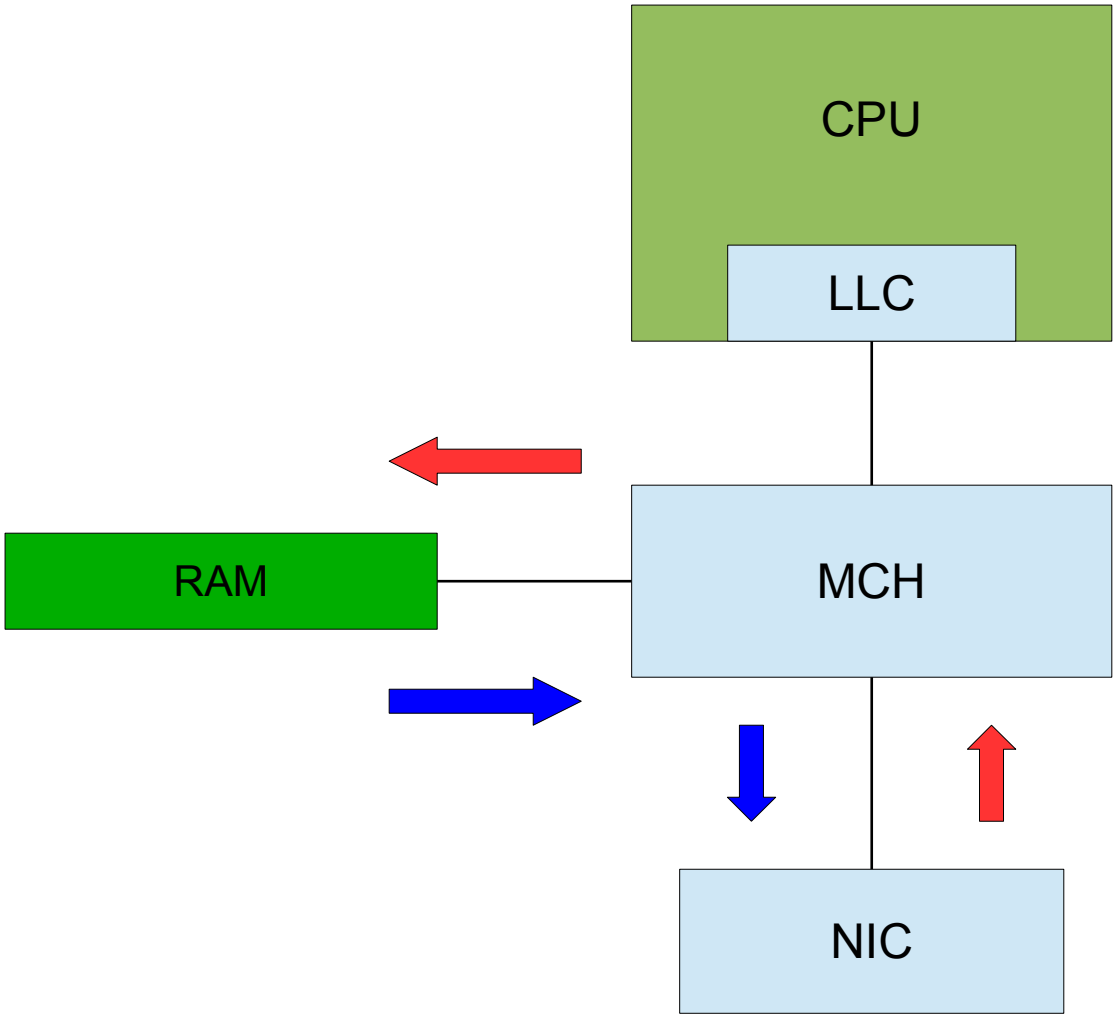
Sandy Bridge (Romley)



PCI-e Transactions

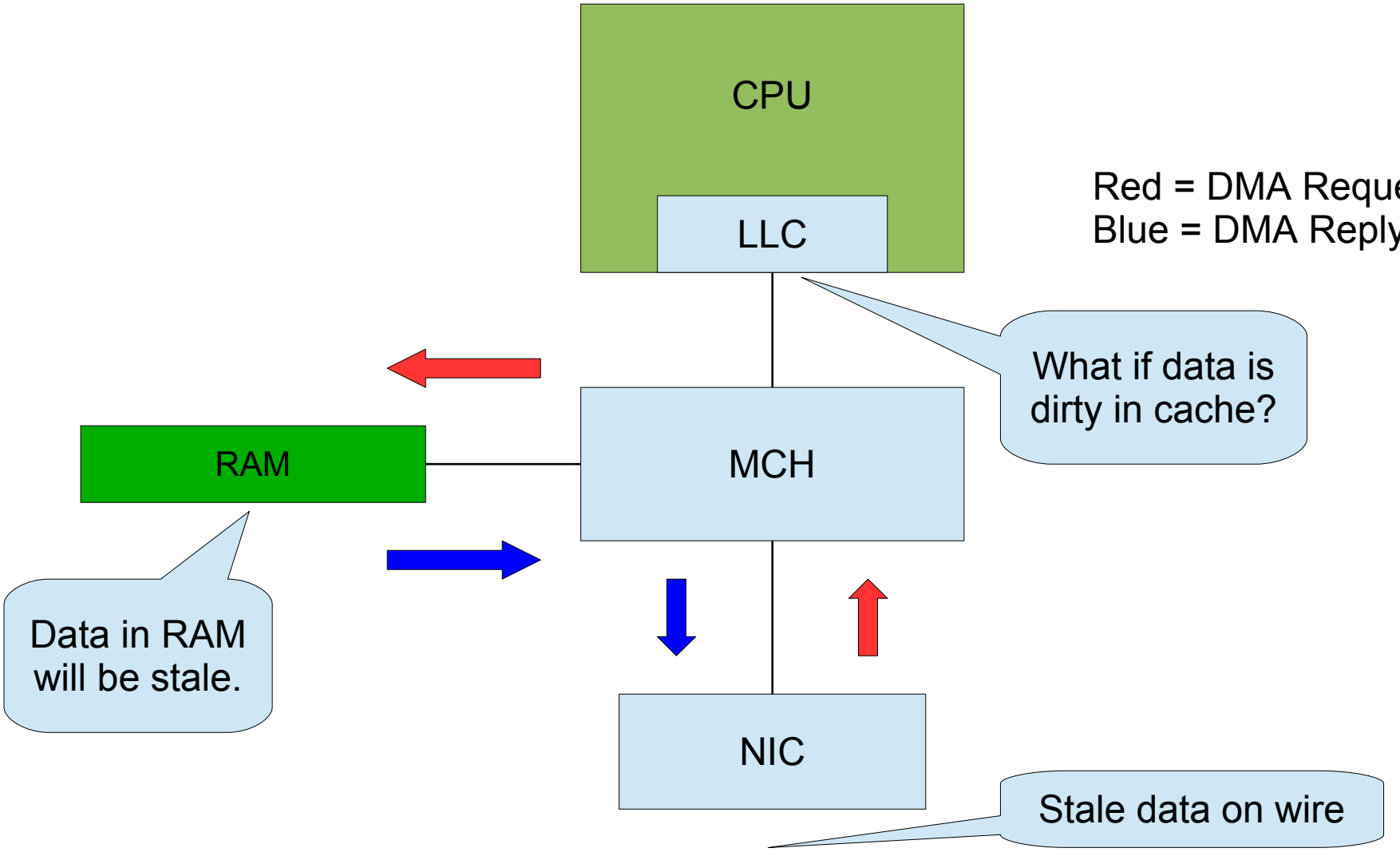
- Memory Read / Write Initiated by Device (DMA)
- Memory Read / Write Initiated by CPU (PIO)
  - Managed by the I/O hub / MCH
- Memory Address Space
  - RAM (via MC)
  - Device Registers (via I/O Hub)

DMA & Cache Snooping

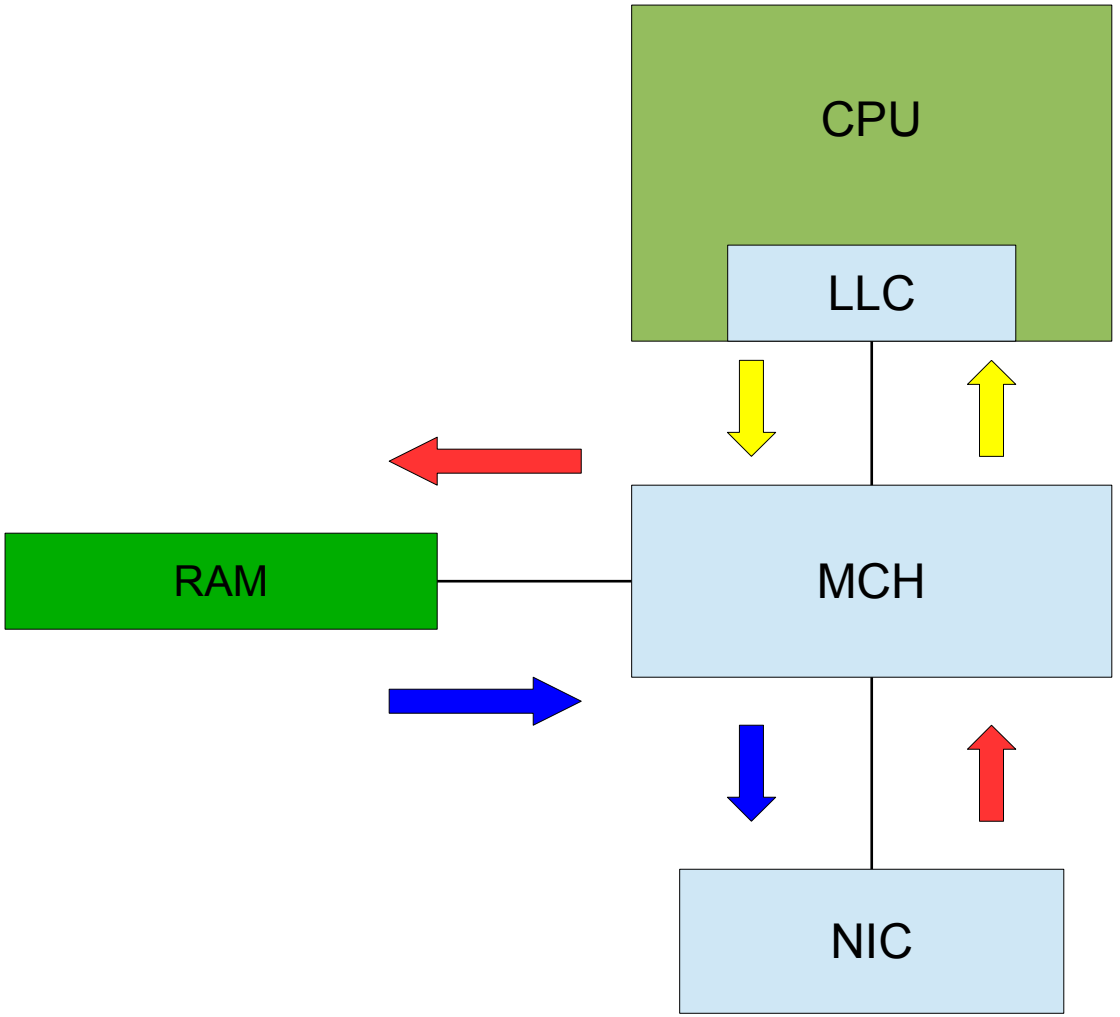


Red = DMA Request
Blue = DMA Reply

DMA & Cache Snooping

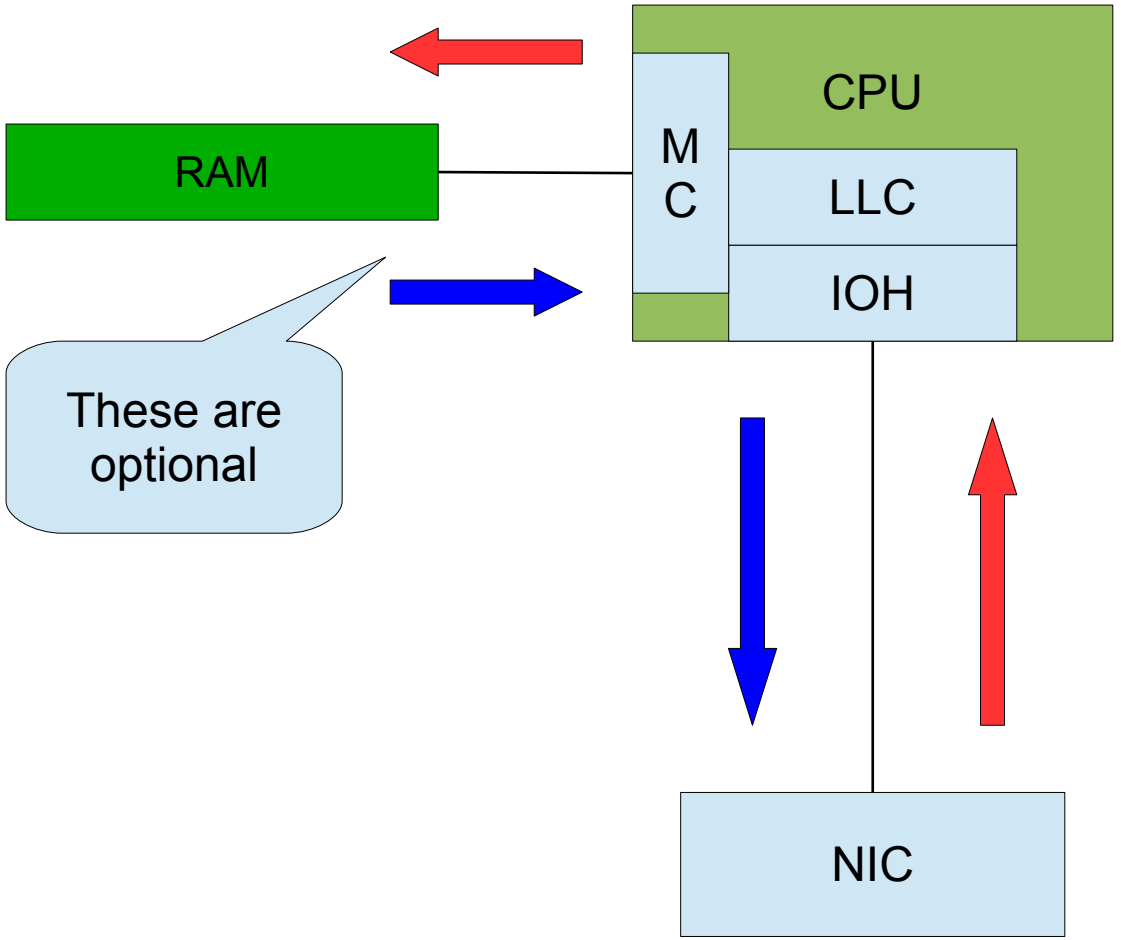


DMA & Cache Snooping



Red = DMA Request
Blue = DMA Reply
Yellow = Snooping

DDIO (Romley)

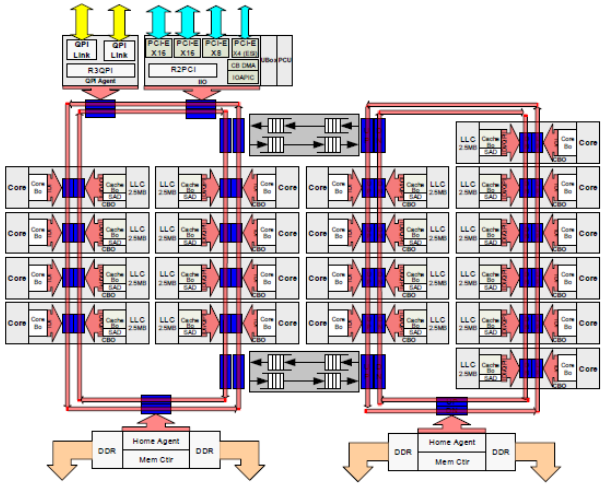


Red = DMA Request
Blue = DMA Reply

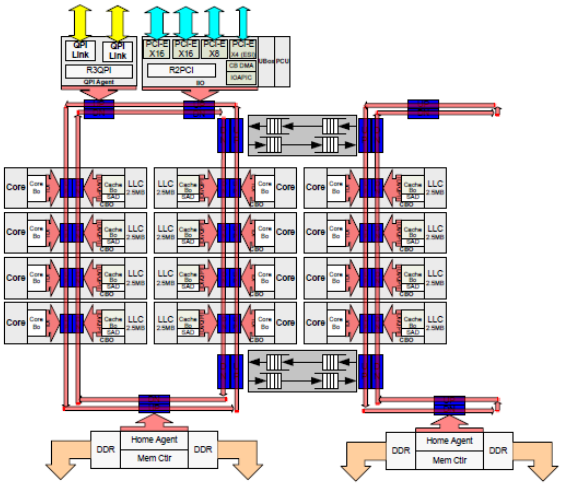
Haswell EP

Haswell EP Die Configurations

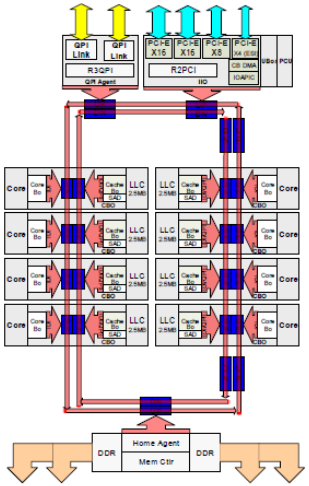
14-18 Core (HCC)



10-12 Core (MCC)



4-8 Core (LCC)



Not representative of actual die-sizes, orientation and layouts – for informational use only.

Chop	Columns	Home Agents	Cores	Power (W)	Transitors (B)	Die Area (mm²)
HCC	4	2	14-18	110-145	5.69	662
MCC	3	2	6-12	65-160	3.84	492
LCC	2	1	4-8	55-140	2.60	354

Source:
http://www.anandtech.com/show/8423/intel-xeon-e5-version-3-up-to-18-haswell-ep-cores-/4

NUMA Implications / Tradeoffs

- Local vs Remote CPU Accesses
- Local vs Remote I/O Accesses
  - Maximize DDIO
  - Except When You Don't?
- Problems are Akin to SMP Scaling
  - (We Know How Well That's Working Out)
- “Soft” Partitioning

NUMA Support in FreeBSD 9

- Hackish “first-touch” Policy
- Not Enabled by Default
- Not Very General Purpose
- No I/O Awareness

NUMA Support in FreeBSD 10

- Start on a More Mature Framework...
- … But Mostly Out of Tree
  - At Least Three Variants
- Stock Tree Only Has “round-robin”
- Not Enabled By Default
- No I/O Awareness

NUMA Support in FreeBSD 11+

- More Work from More Folks
- Goal is to Permit Tuning
  - Not Trying to be Automagical
- Will Include (Some) I/O Awareness
  - Interrupts
- http://wiki.freebsd.org/NUMA
  - Not Set in Stone
- Merge to 10?
- Enabled in GENERIC?