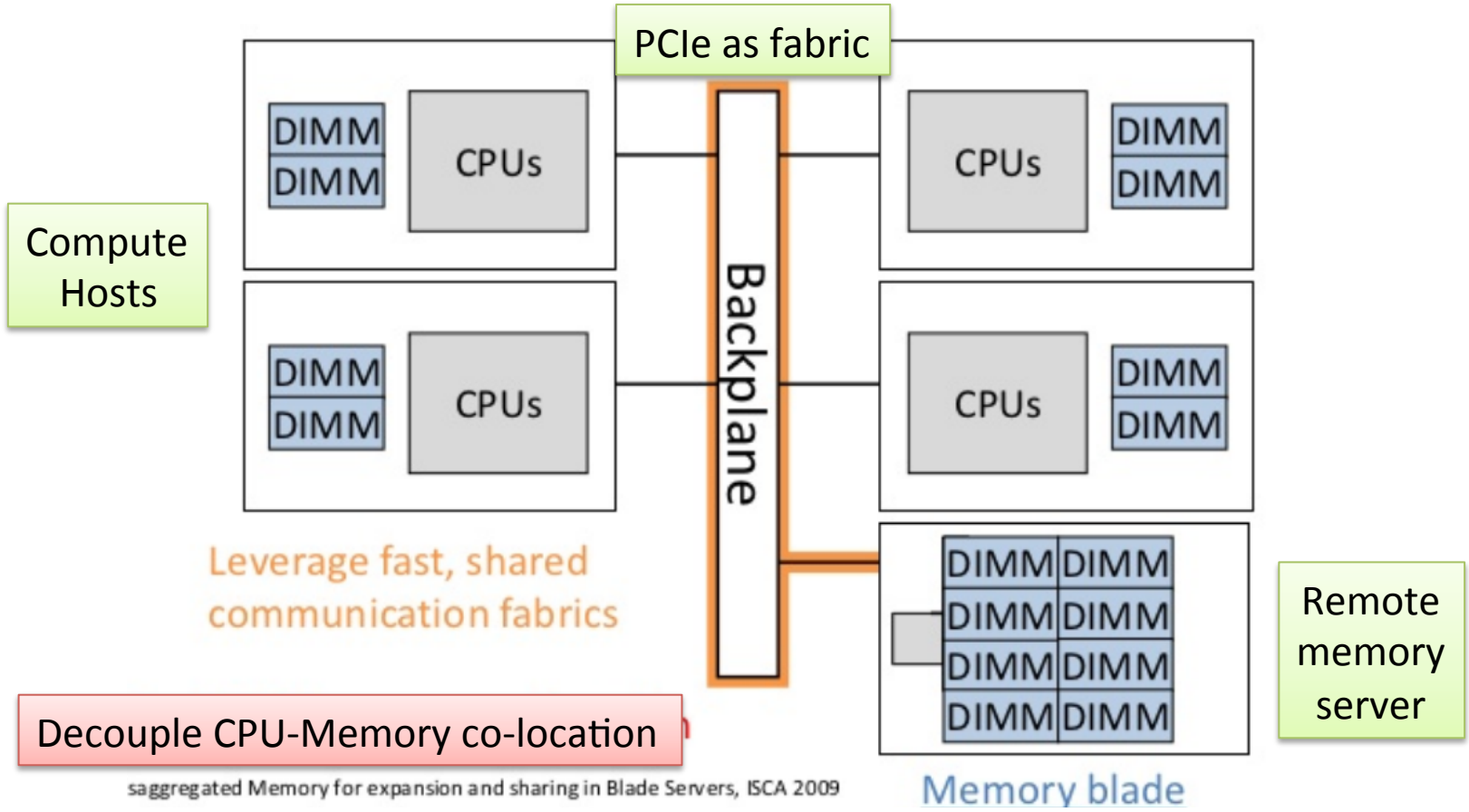
Remote Memory Server

4/16/2013

Cheng-Chun Tu

Advisor: Tzi-cker Chiueh

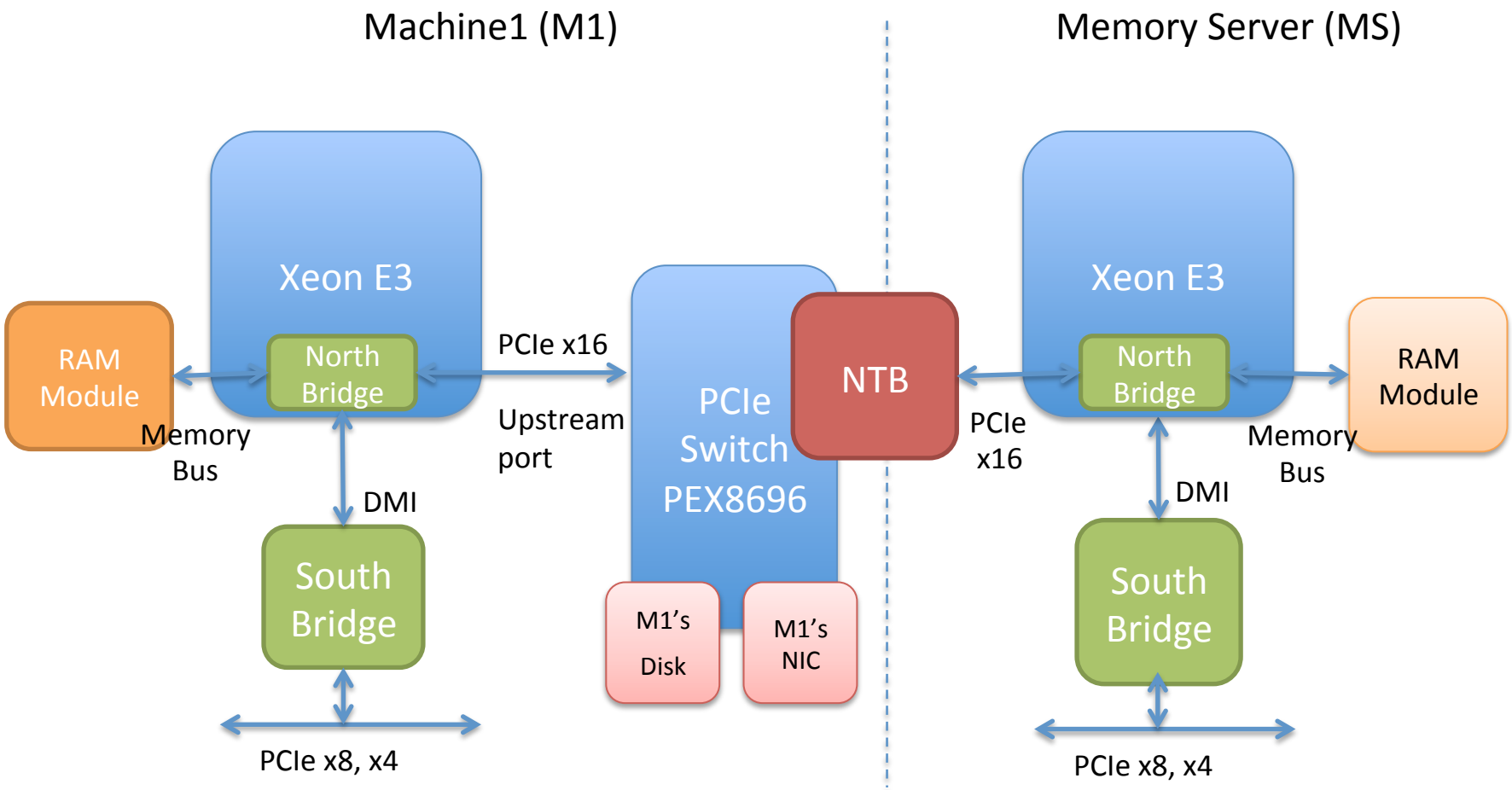
Decoupling CPU-Memory Co-location



Reference: Disaggregated Memory for Expansion and Sharing in Blade Servers [ISCA'09]

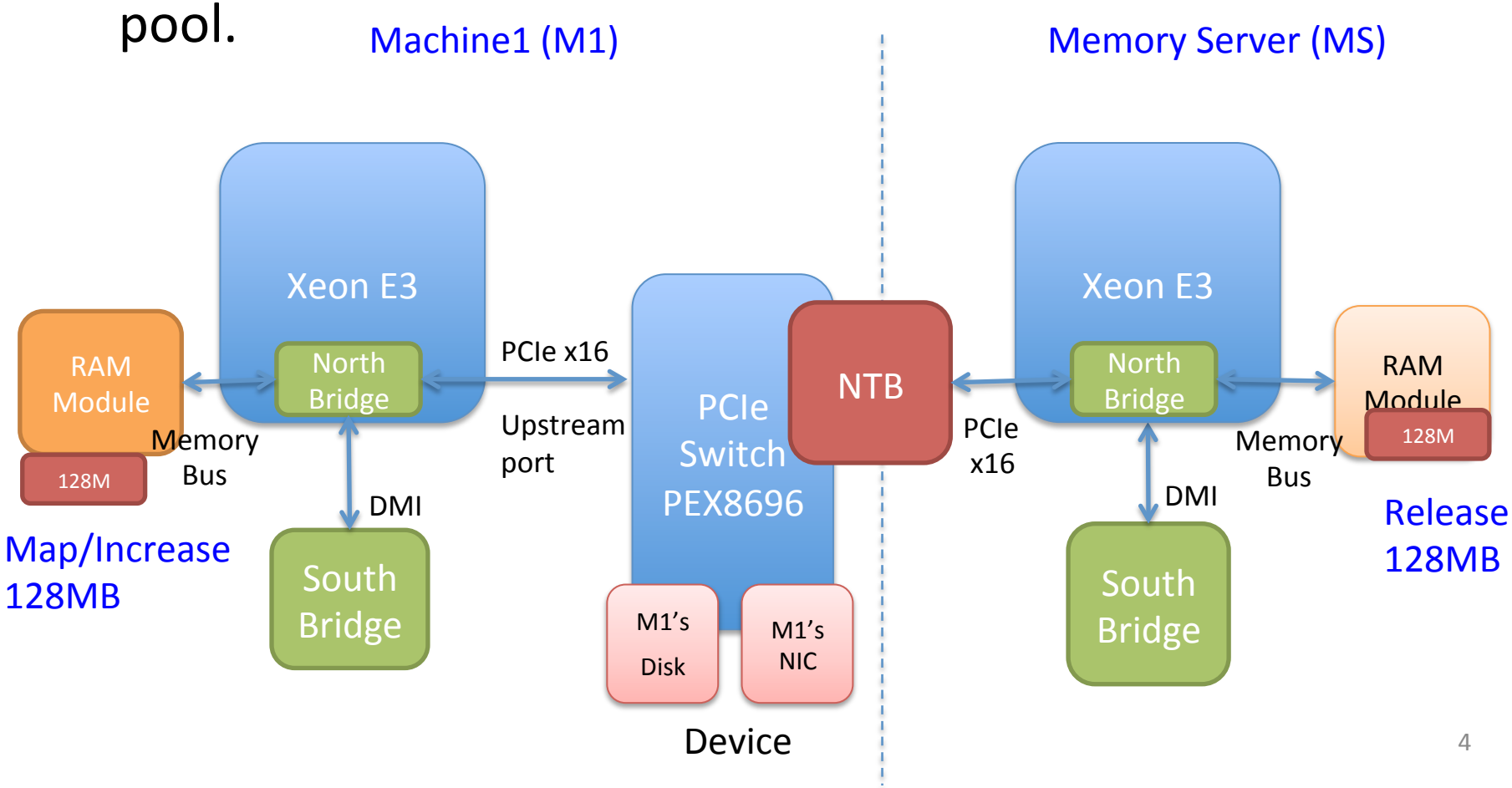
NTB-Based Solution

Case: MS assigns/allocates 128MB RAM to M1



Memory Release / Assignment

- MS releases a portion of its memory.
- M1 maps this portion via NTB and adds to its memory pool.



Configuration

- NTB maps M1: 0x4,0800,0000 to MS: 0x8000000, size: 128MB (0x8000000)
  - M1 read/write 0x408000000 redirects to MS's 0x8000000

- Offline 0x800,0000 at MS
  - 0x8000000 == memory1, see CONFIG_SPARE_MEM

  > echo "offline" > /sys/devices/system/memory/memory1/state
- Online 0x4,0800,0000 at M1

  > echo 0x40800000 > probe

  > echo online > /sys/devices/system/memory/memory129/state

  -- 0x408000000 == memory129

Reference: https://www.kernel.org/doc/Documentation/memory-hotplug.txt

M1 after Assignment

/proc/zoneinfo

Before online

```
Node 0, zone      DMA
            present  3914
Node 0, zone    DMA32
            present  826572
Node 0, zone   Normal
            present  1228160
```

After online

```
Node 0, zone      DMA
            present  3914
Node 0, zone    DMA32
            present  826572
Node 0, zone   Normal
            present  1260928
```

/proc/meminfo

```
MemTotal:       8130996 kB
```

```
MemTotal:       8262068 kB
```

/proc/iomem

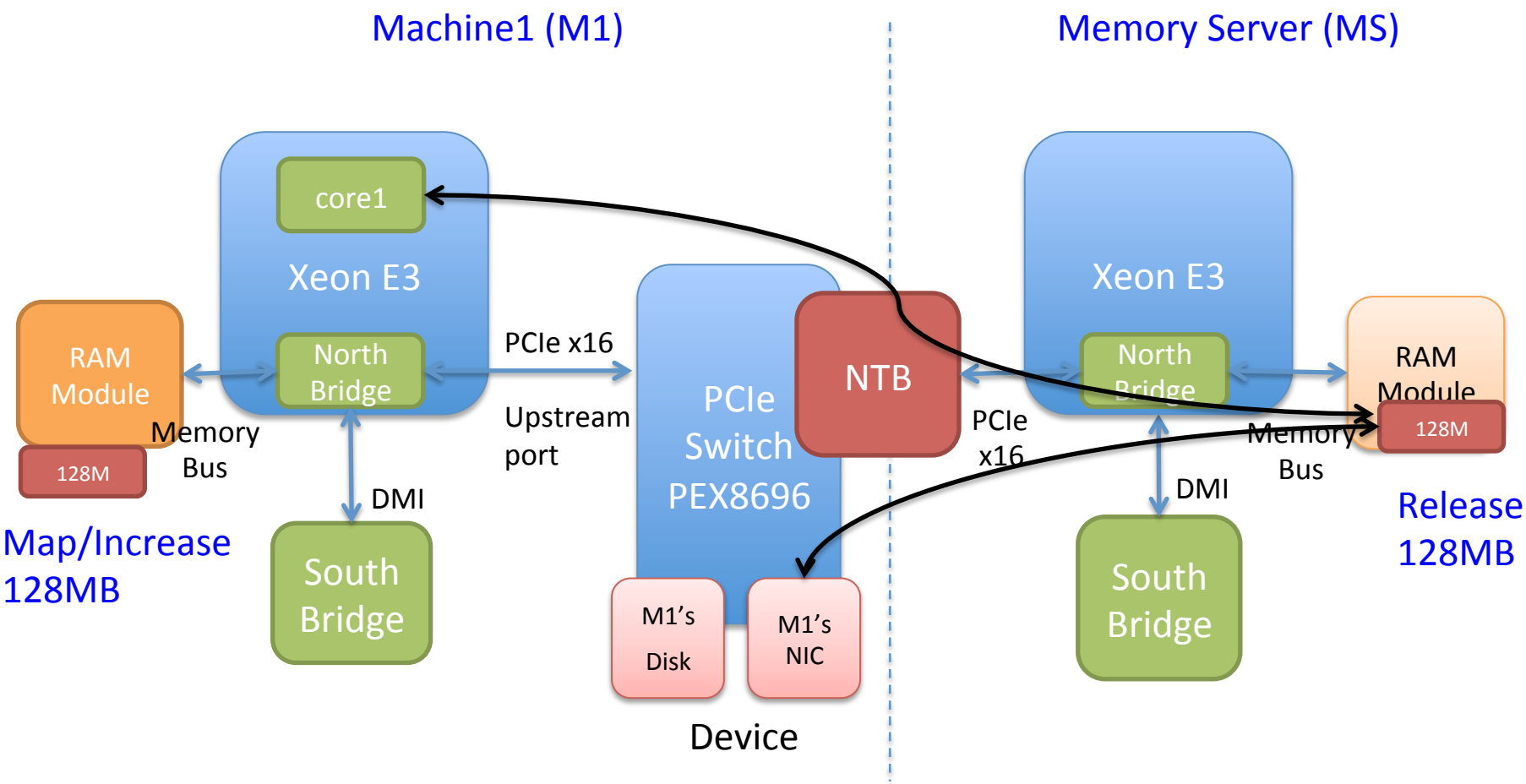
```
fee00000-fee00fff : Local APIC
  fee00000-fee00fff : reserved
ff000000-ffffffff : reserved
  ff000000-ffffffff : pnp 00:10
100000000-22fffffff : System RAM
```

```
fee00000-fee00fff : Local APIC
  fee00000-fee00fff : reserved
ff000000-ffffffff : reserved
  ff000000-ffffffff : pnp 00:10
100000000-22fffffff : System RAM
408000000-40fffffff : System RAM
```

Cacheable?

- Each core snoops memory bus to guarantee cache coherence
- Cache coherent scenarios
  - CPU read/write local RAM: Good
  - Device DMA in/out local RAM: Good
  - CPU read/write remote RAM: Bad, because the read/write does not pass through memory bus
  - Device DMA in/out local RAM: Bad
- How worse can it get using uncachable memory?

Uncachable 2 Cases



Configuration: uncachable

```
> # Make this remote  memory uncachable
> echo "base=0x408000000 size=0x8000000 type=uncachable" >/proc/mtrr
```

```
reg00: base=0x000000000 (    0MB), size= 2048MB, count=1: write-back
reg01: base=0x080000000 ( 2048MB), size= 1024MB, count=1: write-back
reg02: base=0x0c0000000 ( 3072MB), size=  256MB, count=1: write-back
reg03: base=0x100000000 ( 4096MB), size= 4096MB, count=1: write-back
reg04: base=0x200000000 ( 8192MB), size=  512MB, count=1: write-back
reg05: base=0x220000000 ( 8704MB), size=  256MB, count=1: write-back
reg06: base=0x408000000 (16512MB), size=  128MB, count=1: uncachable
```

Reference: https://www.kernel.org/doc/Documentation/x86/mtrr.txt

Setup: No DMA, Only for CPU

- Linux divide memory into three zones:
  - DMA: for 16MB ISA device
  - DMA32: for < 4GB dma memory
  - Normal: the rest of memory
- To prevent the remote memory being used by any DMA controller
  - Put the remote memory into Normal zone
  - Trap all dma_set_mask to make sure only 32bit DMA is supported. (So only pages from DMA32 is allocated for DMA)

Reference: DMA32 zone: http://lwn.net/Articles/152337/

memset performance

- Experiment: memset 128MB region locally and remotely
- Local: phys_to_virt + MTRR settings to noncachable
  - 1.0 sec → 128MB/sec
- Local: phys_to_virt + MTRR settings to cachable
  - 10 ms → 12800MB/sec
- Remote: ioremap_nocache:
  - 5.1 sec → 25MB/sec
- Remote: ioremap_wc (write-combining):
  - 3.7 sec → 34.6MB/sec
- Remote: ioremap_cachable + MTRR → crash!

Remote non-cacheable = x500 slower than local cacheable
= x5 slower than local non-cacheable

Problems / Discussion

- The system hangs when putting the remote memory into the generic memory pool
  - If the backing memory is DRAM, the L3 cache controller interacts with the DRAM controller in such a way to fetch the cache line in a burst mode.
  - However, if the backing memory is a set of registers on a PCIe device, the L3 cache controller may confuses and completely screwed up.
- Remote memory applications:
  - As a buffer/staging memory managed by the OS through special API, rather than normal memory read/write instructions.
  - Remote swapping disk using DMA
  - Install customized hardware in compute servers to support cache coherent

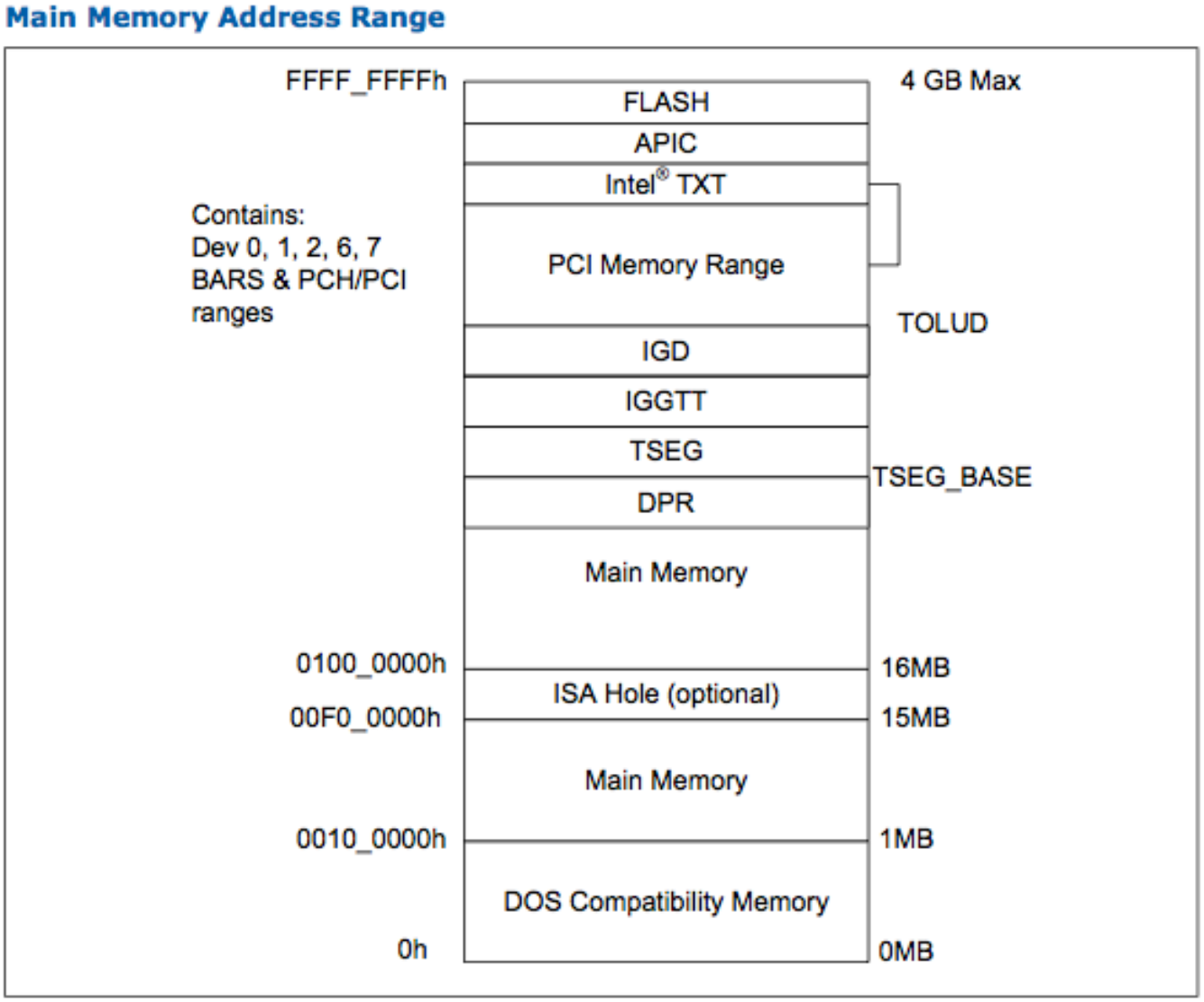
Reference: http://forum.osdev.org/viewtopic.php?f=1&t=26509

References

- Intel's Optimization for memset/memcpy
  - http://software.intel.com/en-us/articles/memcpy-performance
- SIMD (Single Instruction, Multiple Data) SSE. Memcpy optimization
  - http://software.intel.com/en-us/articles/memcpy-performance
  - http://en.wikipedia.org/wiki/SIMD
  - movdqa is suitable for 16-byte aligned operands.
  - movdqu is suitable for fetching byte-aligned groups of 16 bytes from memory, but not useful for storing them.
- ScaleMP. Versatile SMP (vSMP) architecture
- Texas Memory System
  - http://en.wikipedia.org/wiki/Texas_Memory_Systems

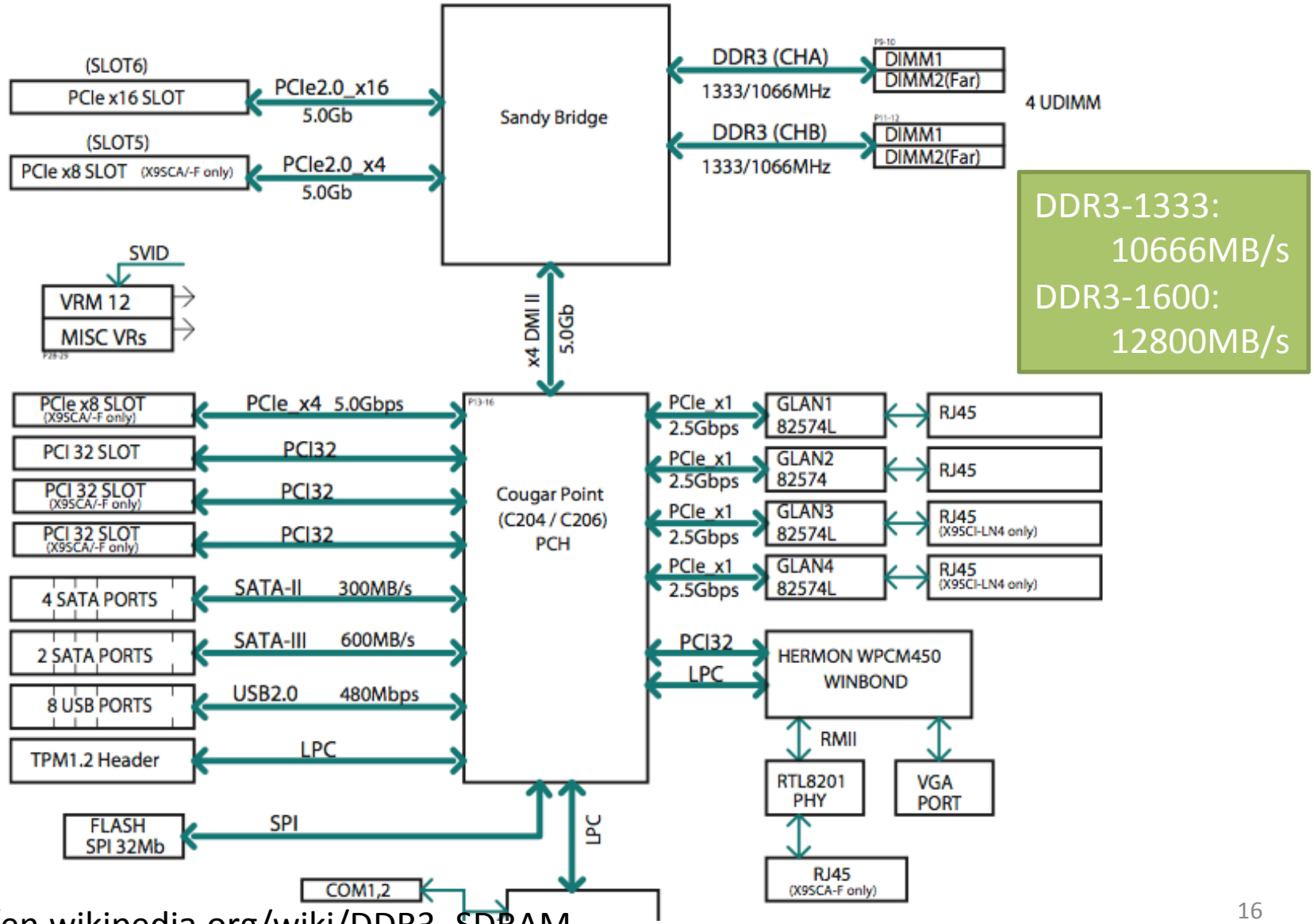
END

Intel Xeon Memory Map



Reference: Intel® Xeon® Processor E3-1200 v2 Product Family

SuperMicro X9SCi-LN4/X9SCi-LN4F/X9SCA/X9SCA Block Diagram



http://en.wikipedia.org/wiki/DDR3_SDRAM

SuperMicro X9SC/X9SCA lspci

00:00.0 Host bridge: Intel Corporation Device 0158 (rev 09)
00:01.0 PCI bridge: Intel Corporation Device 0151 (rev 09)
00:16.0 Communication controller: Intel Corporation Cougar Point HECI Controller #1 (rev 04)
00:16.1 Communication controller: Intel Corporation Cougar Point HECI Controller #2 (rev 04)
00:1a.0 USB Controller: Intel Corporation Cougar Point USB Enhanced Host Controller #2 (rev 05)
00:1c.0 PCI bridge: Intel Corporation Cougar Point PCI Express Root Port 1 (rev b5)
00:1c.4 PCI bridge: Intel Corporation Cougar Point PCI Express Root Port 5 (rev b5)
00:1c.5 PCI bridge: Intel Corporation Cougar Point PCI Express Root Port 6 (rev b5)
00:1d.0 USB Controller: Intel Corporation Cougar Point USB Enhanced Host Controller #1 (rev 05)
00:1e.0 PCI bridge: Intel Corporation 82801 PCI Bridge (rev a5)
00:1f.0 ISA bridge: Intel Corporation Cougar Point LPC Controller (rev 05)
00:1f.2 SATA controller: Intel Corporation Cougar Point 6 port SATA AHCI Controller (rev 05)
00:1f.3 SMBus: Intel Corporation Cougar Point SMBus Controller (rev 05)
01:00.0 PCI bridge: PLX Technology, Inc. PEX 8696 96-lane, 24-Port PCI Express Gen 2 (5.0 GT/s) Multi-Root Switch (rev aa
02:04.0 PCI bridge: PLX Technology, Inc. PEX 8696 96-lane, 24-Port PCI Express Gen 2 (5.0 GT/s) Multi-Root Switch (rev aa
02:05.0 PCI bridge: PLX Technology, Inc. PEX 8696 96-lane, 24-Port PCI Express Gen 2 (5.0 GT/s) Multi-Root Switch (rev aa
02:08.0 PCI bridge: PLX Technology, Inc. PEX 8696 96-lane, 24-Port PCI Express Gen 2 (5.0 GT/s) Multi-Root Switch (rev aa
02:09.0 PCI bridge: PLX Technology, Inc. PEX 8696 96-lane, 24-Port PCI Express Gen 2 (5.0 GT/s) Multi-Root Switch (rev aa
02:0d.0 PCI bridge: PLX Technology, Inc. PEX 8696 96-lane, 24-Port PCI Express Gen 2 (5.0 GT/s) Multi-Root Switch (rev aa
02:10.0 PCI bridge: PLX Technology, Inc. PEX 8696 96-lane, 24-Port PCI Express Gen 2 (5.0 GT/s) Multi-Root Switch (rev aa
02:11.0 PCI bridge: PLX Technology, Inc. PEX 8696 96-lane, 24-Port PCI Express Gen 2 (5.0 GT/s) Multi-Root Switch (rev aa
02:14.0 PCI bridge: PLX Technology, Inc. PEX 8696 96-lane, 24-Port PCI Express Gen 2 (5.0 GT/s) Multi-Root Switch (rev aa
02:15.0 PCI bridge: PLX Technology, Inc. PEX 8696 96-lane, 24-Port PCI Express Gen 2 (5.0 GT/s) Multi-Root Switch (rev aa
03:00.0 Serial Attached SCSI controller: LSI Logic / Symbios Logic SAS2008 PCI-Express Fusion-MPT SAS-2 [Falcon] (rev 03)
07:00.0 Bridge: PLX Technology, Inc. PEX 8619 16-lane, 16-Port PCI Express Gen 2 (5.0 GT/s) Switch with DMA (rev ba)
10:03.0 VGA compatible controller: Matrox Graphics, Inc. MGA G200eW WPCM450 (rev 0a)

Five PCI devices in Xeon (Northbridge)

- Device 0, Function 0 (Memory Controller)
  - 00:00.0 Host bridge: Intel Corporation Device 0158 (rev 09)
- Device 1, Function 0: (PCIe x16 Controller)
  - 00:01.0 PCI bridge: Intel Corporation Device 0151 (rev 09)
- Device 1, Function 1: (PCIe x8 Controller)
- Device 1, Function 2: (PCIe x4 Controller)
- Device 6, Function 0: (PCIe x4 Controller)
- Device 2, Function 0: (Integrated Graphics Device (IGD))

System Memory Map [0:0.0]

```
[root@66-0A00202-17 ~]# lspci -xxx -s0:0.0
00:00.0 Host bridge: Intel Corporation Device 0158 (rev 09)
00: 86 80 58 01 46 01 90 20 09 00 00 06 00 00 00 00
10: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
20: 00 00 00 00 00 00 00 00 00 00 00 00 86 80 58 01
30: 00 00 00 00 e0 00 00 00 00 00 00 00 00 00 00 00
40: 01 90 d1 fe 00 00 00 00 01 00 d1 fe 00 00 00 00 -->48–4Fh MCHBAR Host Memory Mapped
50: 03 00 00 00 09 00 00 00 00 00 00 00 01 00 80 df
60: 05 00 00 f8 00 00 00 00 01 80 d1 fe 00 00 00 00 -->60–67h PCIEXBAR PCI Express Register Ra
70: 00 00 f0 ff 7f 00 00 00 00 04 00 00 00 00 00 00
80: 10 11 00 00 00 00 11 00 1a 00 00 00 00 00 00 00
90: 01 00 00 00 02 00 00 00 01 00 f0 1f 02 00 00 00
a0: 01 00 00 00 02 00 00 00 01 00 00 20 02 00 00 00 --> Top of Memory
b0: 01 00 00 e0 01 00 00 e0 01 00 80 df 01 00 00 e0
c0: 00 00 00 00 00 00 00 00 00 00 00 00 03 00 00 00
d0: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
e0: 09 00 0c 01 92 aa 00 61 d0 08 40 16 00 00 00 00
f0: 00 00 00 00 00 00 00 00 c8 0f 09 00 00 00 00 00
```

90–97h REMAPBASE Remap Base Address Register
98–9Fh REMAPLIMIT Remap Limit Address Register
A0–A7h TOM Top of Memory
A8–AFh TOUUD Top of Upper Usable DRAM
B0–B3h BDSM Base Data of Stolen Memory
B4–B7h BGSM Base of GTT stolen Memory
B8–BBh TSEGMB TSEG Memory Base
BC–BFh TOLUD Top of Low Usable DRAM
50–51h GGC GMCH Graphics Control Register

memset performance

- Experiment
- Remote: ioremap_nocache:
  - 5.1 sec → memset(), total 128MB
  - 25MB/s = 200Mbps
- Remote: ioremap_cachable + mtrr → crash
- Remote: ioremap_wc:
  - 3.7 sec
- Local: phys_to_virt + mtrr settings to noncachable
  - 1.0 sec
- Local: phys_to_virt + mtrr settings to cachable
  - 10 ms