了解内存

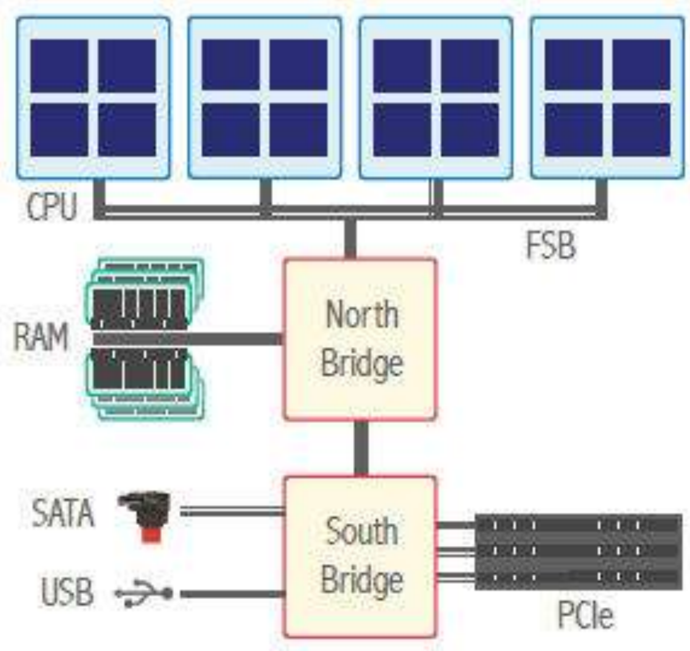
核心系统数据库组　余锋

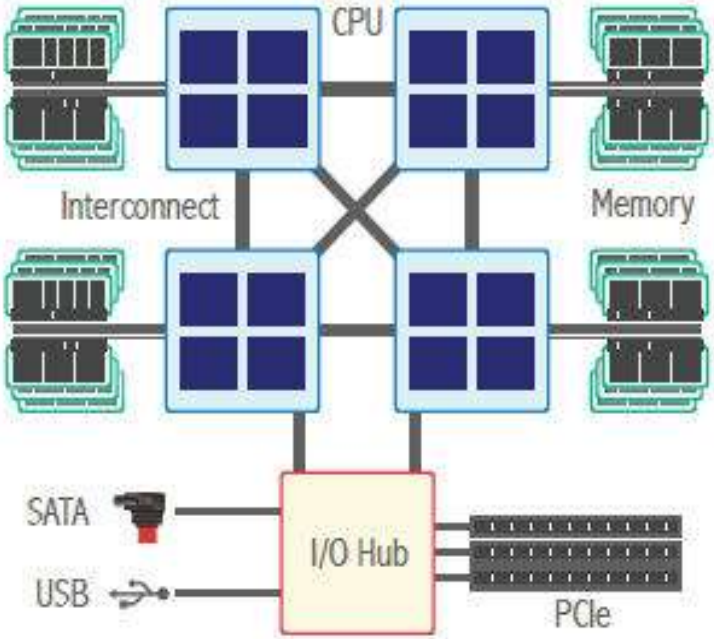
http://yufeng.info

@淘宝褚霸

2012-03-17

内存体系变更


CPU
FSB
RAM
North Bridge
SATA
USB
South Bridge
PCIe


CPU
Interconnect
Memory
SATA
USB
I/O Hub
PCIe

过去

现在

服务器内存条

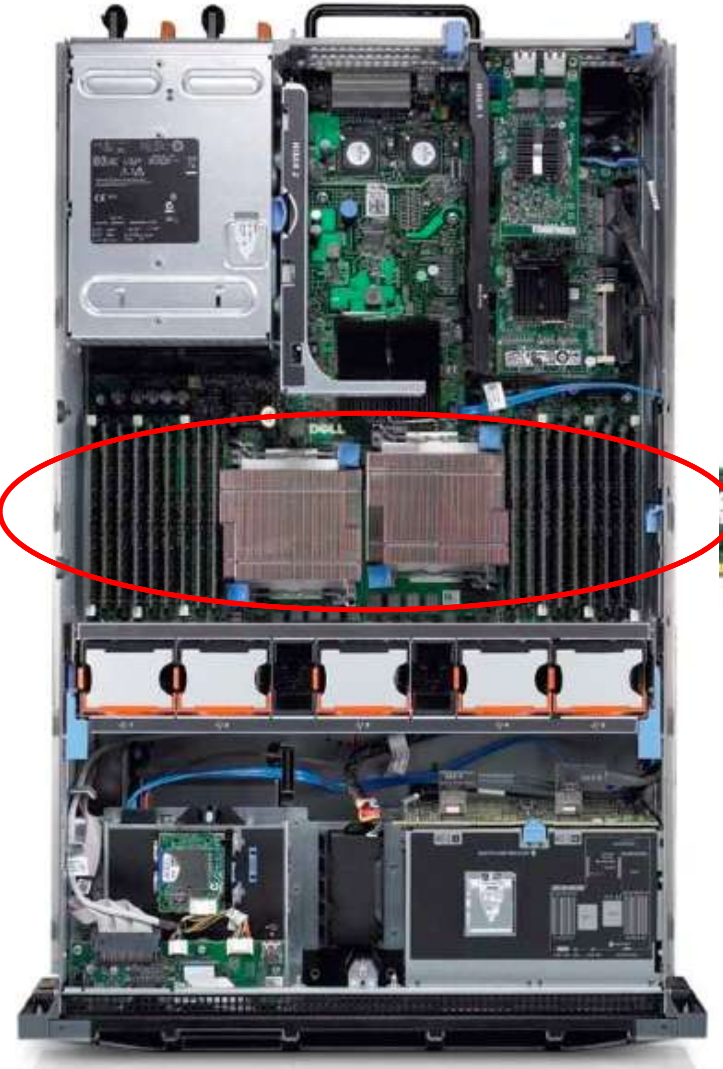


Figure: PowerEdge R710 Main Components

内存总体信息

Memory:

31.5GB / 32GB 1333MHz DDR3 == 8 x 4GB - 4GB PC3-10600 Samsung DDR3-1333 ECC Registered CL9 2Rx8

单条内存信息

asset="02120761"

cas="9"

form="DIMM"

handle="76"

locator="DIMM_A1"

org=“x8”

ranks="2"

serial="87BE9BB9"

size="4096 MB"

speed="1333MHz"

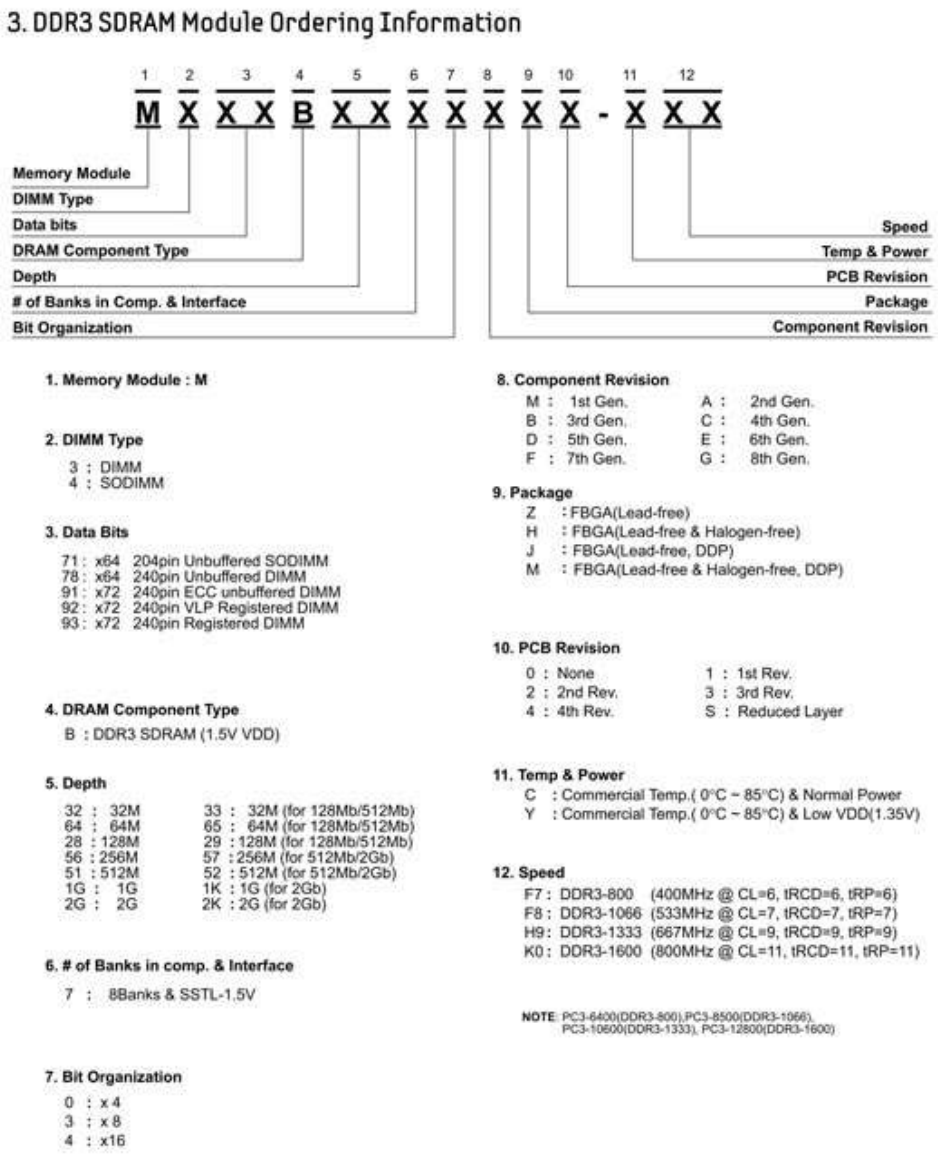
type="DDR3"

width="72 bits“

part_number="M393B5273CH0-YH9“

pretty="4GB PC3-10600 Samsung DDR3-1333 ECC Registered CL9 2Rx8”

模组型号解读



3. DDR3 SDRAM Module Ordering Information

```
 1   2    3    4    5     6   7   8   9   10     11    12
 M   X   X X   B   X X    X   X   X   X   X   -  X    X X
```

Memory Module (1)
DIMM Type (2)
Data bits (3)
DRAM Component Type (4)
Depth (5)
# of Banks in Comp. & Interface (6)
Bit Organization (7)
Component Revision (8)
Package (9)
PCB Revision (10)
Temp & Power (11)
Speed (12)

1. Memory Module : M

2. DIMM Type

3 : DIMM
4 : SODIMM

3. Data Bits

71 :	x64	204pin Unbuffered SODIMM
78 :	x64	240pin Unbuffered DIMM
91 :	x72	240pin ECC unbuffered DIMM
92 :	x72	240pin VLP Registered DIMM
93 :	x72	240pin Registered DIMM

4. DRAM Component Type

B : DDR3 SDRAM (1.5V VDD)

5. Depth

32	32M	33	32M (for 128Mb/512Mb)
64	64M	65	64M (for 128Mb/512Mb)
28	128M	29	128M (for 128Mb/512Mb)
56	256M	57	256M (for 512Mb/2Gb)
51	512M	52	512M (for 512Mb/2Gb)
1G	1G	1K	1G (for 2Gb)
2G	2G	2K	2G (for 2Gb)

6. # of Banks in comp. & Interface

7 : 8Banks & SSTL-1.5V

7. Bit Organization

0 : x 4
3 : x 8
4 : x16

8. Component Revision

M :	1st Gen.	A :	2nd Gen.
B :	3rd Gen.	C :	4th Gen.
D :	5th Gen.	E :	6th Gen.
F :	7th Gen.	G :	8th Gen.

9. Package

Z : FBGA(Lead-free)
H : FBGA(Lead-free & Halogen-free)
J : FBGA(Lead-free, DDP)
M : FBGA(Lead-free & Halogen-free, DDP)

10. PCB Revision

0 :	None	1 :	1st Rev.
2 :	2nd Rev.	3 :	3rd Rev.
4 :	4th Rev.	S :	Reduced Layer

11. Temp & Power

C : Commercial Temp.(0°C ~ 85°C) & Normal Power
Y : Commercial Temp.(0°C ~ 85°C) & Low VDD(1.35V)

12. Speed

F7 :	DDR3-800	(400MHz @ CL=6, tRCD=6, tRP=6)
F8 :	DDR3-1066	(533MHz @ CL=7, tRCD=7, tRP=7)
H9 :	DDR3-1333	(667MHz @ CL=9, tRCD=9, tRP=9)
K0 :	DDR3-1600	(800MHz @ CL=11, tRCD=11, tRP=11)

NOTE: PC3-6400(DDR3-800),PC3-8500(DDR3-1066),
PC3-10600(DDR3-1333), PC3-12800(DDR3-1600)

M393B5273CH0-YH9

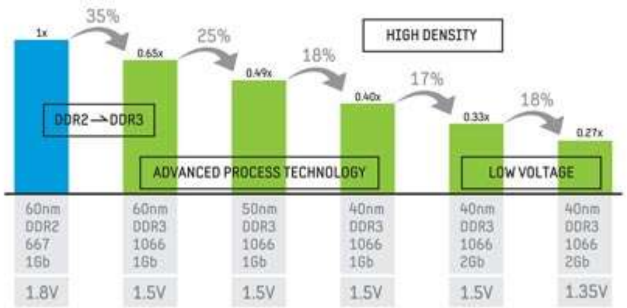
DIMM、x72 240pin Registered、2Gb颗粒、x8位宽、第四代产品、无铅无汞FBGA封装、1.35V低电压、DDR3-1333

功耗

SAMSUNG DDR3 VS. DDR2
POWER SAVINGS FOR SERVER MODULES

* All same density

FIGURE 3. SAMSUNG 48GB DDR3 POWER REDUCTION



Overall, a server built using Samsung's 40nm class, 1.35V, 2Gb Green DDR3 memory uses 73% less power than a 60nm, 1.8V, 1Gb DDR2 chip.
In a 48GB server, this translates into a 38% power savings.
Higher density servers will show even greater savings.

FIGURE 4. 48GB MEMORY POWER SAVINGS COMPARISON

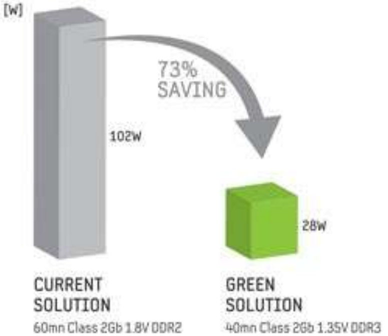
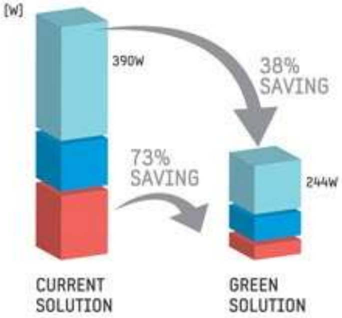
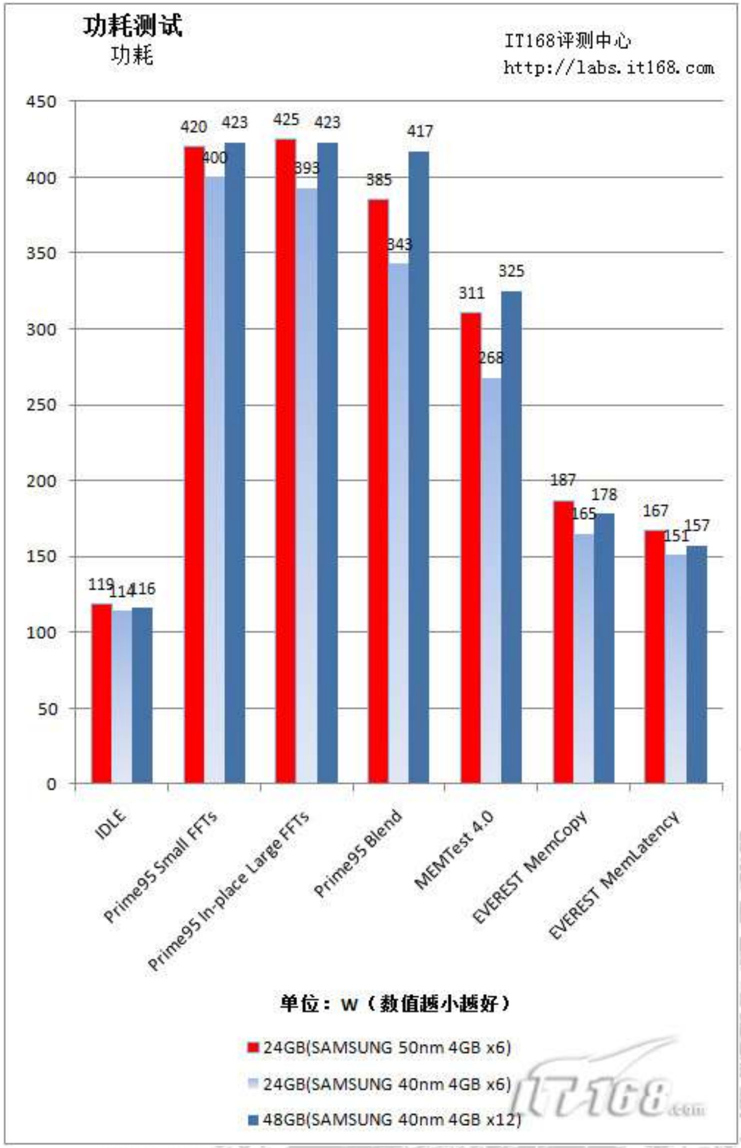


FIGURE 5.
48GB BASED SERVER SYSTEM-LEVEL POWER SAVINGS



*Source : Measured by SAMSUNG Lab



	24GB(SAMSUNG 50nm 4GB x6)	24GB(SAMSUNG 40nm 4GB x6)	48GB(SAMSUNG 40nm 4GB x12)
IDLE	119	114	116
Prime95 Small FFTs	420	400	423
Prime95 In-place Large FFTs	425	393	423
Prime95 Blend	385	343	417
MEMTest 4.0	311	268	325
EVEREST MemCopy	187	165	178
EVEREST MemLatency	167	151	157

内存相关性能数字

L1 cache reference	0.5 ns
Branch mispredict	5 ns
L2 cache reference	7 ns
Mutex lock/unlock	25 ns
Main memory reference	100 ns
Compress 1K bytes with Zippy	3,000 ns

```
Memory latencies in nanoseconds - smaller is better
--------------------------------------------------------------
---------------
Host OS Mhz L1 $ L2 $ Main mem Rand mem Guesses
--------------------------------------------------------------
---
Dr4000 Linux 2.6.32- 2631 1.1590 5.7170  78.0 110.4
```

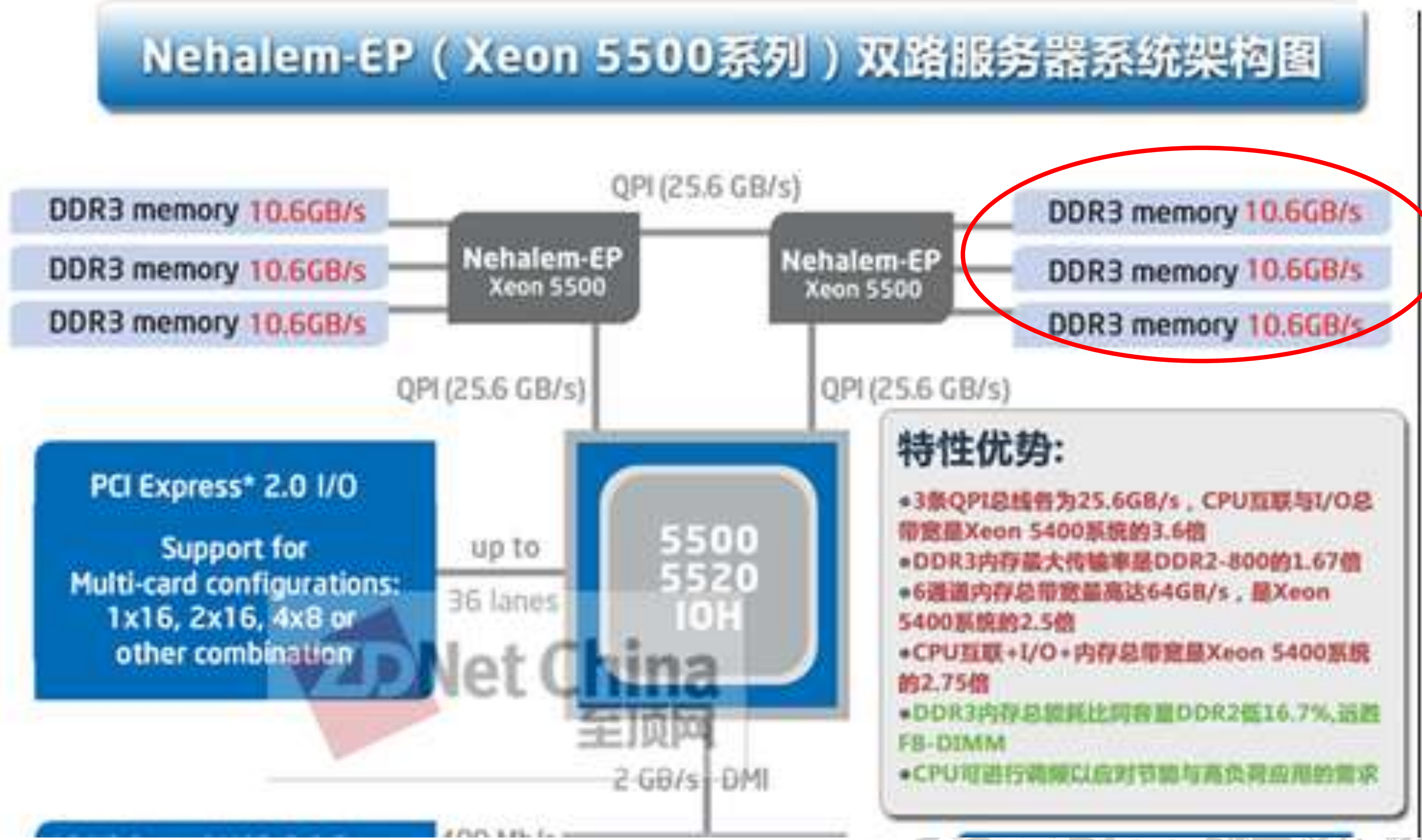
内存**带宽计算**

内存**带宽计算公式：带宽**=内存核心**频率**×内存**总线**位数×倍增系数。

每个通道 (1333/8)*64*8 /8 = 10.6G Byte；

而我**们的CPU**是3个通道的，也就是**说这个CPU**的**总**的内存**带宽是 10.6*3=31.8G**

QPI和内存通道理论带宽

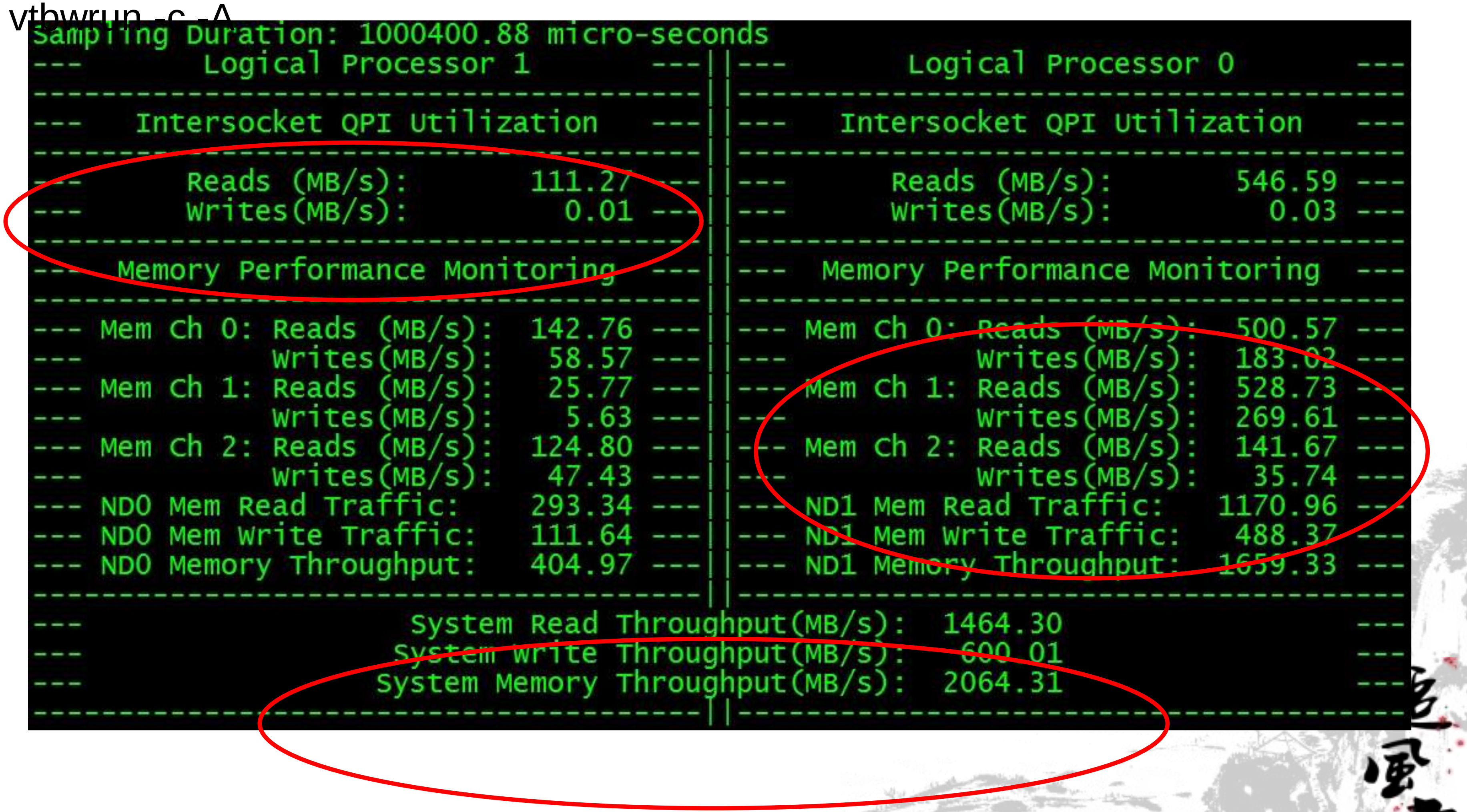


特性优势:

- 3条QPI总线各为25.6GB/s，CPU互联与I/O总带宽是Xeon 5400系统的3.6倍
- DDR3内存最大传输率是DDR2-800的1.67倍
- 6通道内存总带宽最高达64GB/s，是Xeon 5400系统的2.5倍
- CPU互联+I/O+内存总带宽是Xeon 5400系统的2.75倍
- DDR3内存总能耗比同容量DDR2低16.7%,远胜FB-DIMM
- CPU可进行调频以应对节能与高负荷应用的需求

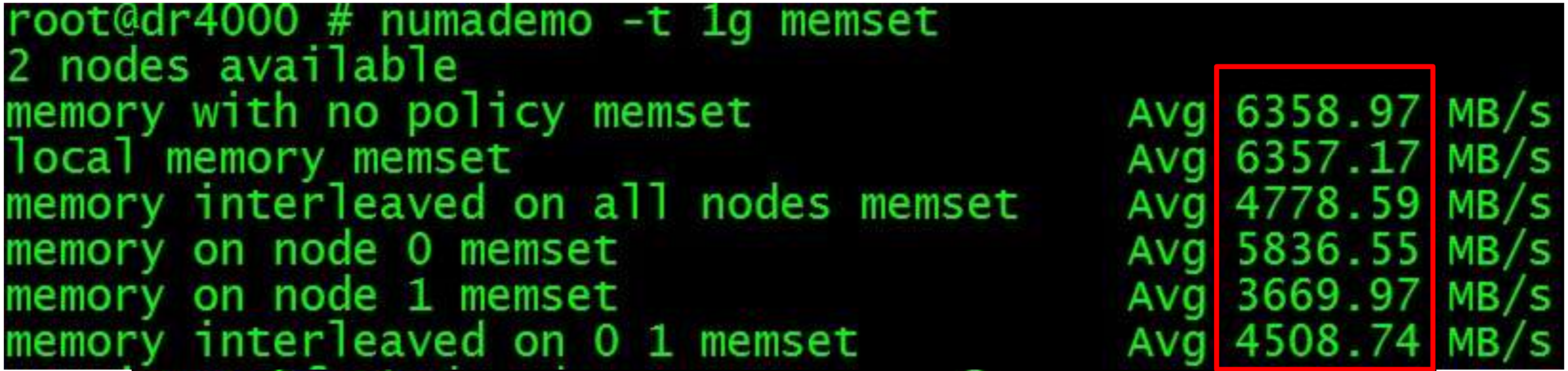
内存通道使用情况

vtbwrun -c -A

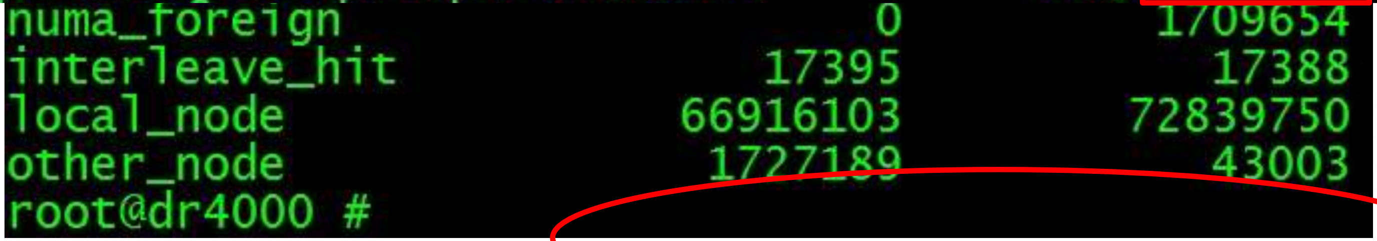


```
Sampling Duration: 1000400.88 micro-seconds
---        Logical Processor 1         ---||---        Logical Processor 0         ---
--------------------------------------------||--------------------------------------------
---    Intersocket QPI Utilization     ---||---    Intersocket QPI Utilization     ---
--------------------------------------------||--------------------------------------------
---        Reads  (MB/s):       111.27 ---||---        Reads  (MB/s):       546.59 ---
---        Writes(MB/s):          0.01 ---||---        Writes(MB/s):          0.03 ---
--------------------------------------------||--------------------------------------------
---    Memory Performance Monitoring   ---||---   Memory Performance Monitoring    ---
--------------------------------------------||--------------------------------------------
--- Mem Ch 0: Reads  (MB/s):   142.76 ---||--- Mem Ch 0: Reads  (MB/s):   500.57 ---
---           Writes(MB/s):     58.57 ---||---           Writes(MB/s):    183.02 ---
--- Mem Ch 1: Reads  (MB/s):    25.77 ---||--- Mem Ch 1: Reads  (MB/s):   528.73 ---
---           Writes(MB/s):      5.63 ---||---           Writes(MB/s):    269.61 ---
--- Mem Ch 2: Reads  (MB/s):   124.80 ---||--- Mem Ch 2: Reads  (MB/s):   141.67 ---
---           Writes(MB/s):     47.43 ---||---           Writes(MB/s):     35.74 ---
--- ND0 Mem Read Traffic:      293.34 ---||--- ND1 Mem Read Traffic:     1170.96 ---
--- ND0 Mem Write Traffic:     111.64 ---||--- ND1 Mem Write Traffic:     488.37 ---
--- ND0 Memory Throughput:     404.97 ---||--- ND1 Memory Throughput:    1659.33 ---
--------------------------------------------||--------------------------------------------
---                    System Read Throughput(MB/s):   1464.30                      ---
---                   System Write Throughput(MB/s):    600.01                      ---
---                  System Memory Throughput(MB/s):   2064.31                      ---
--------------------------------------------||--------------------------------------------
```

NUMA节点内存访问速度差异

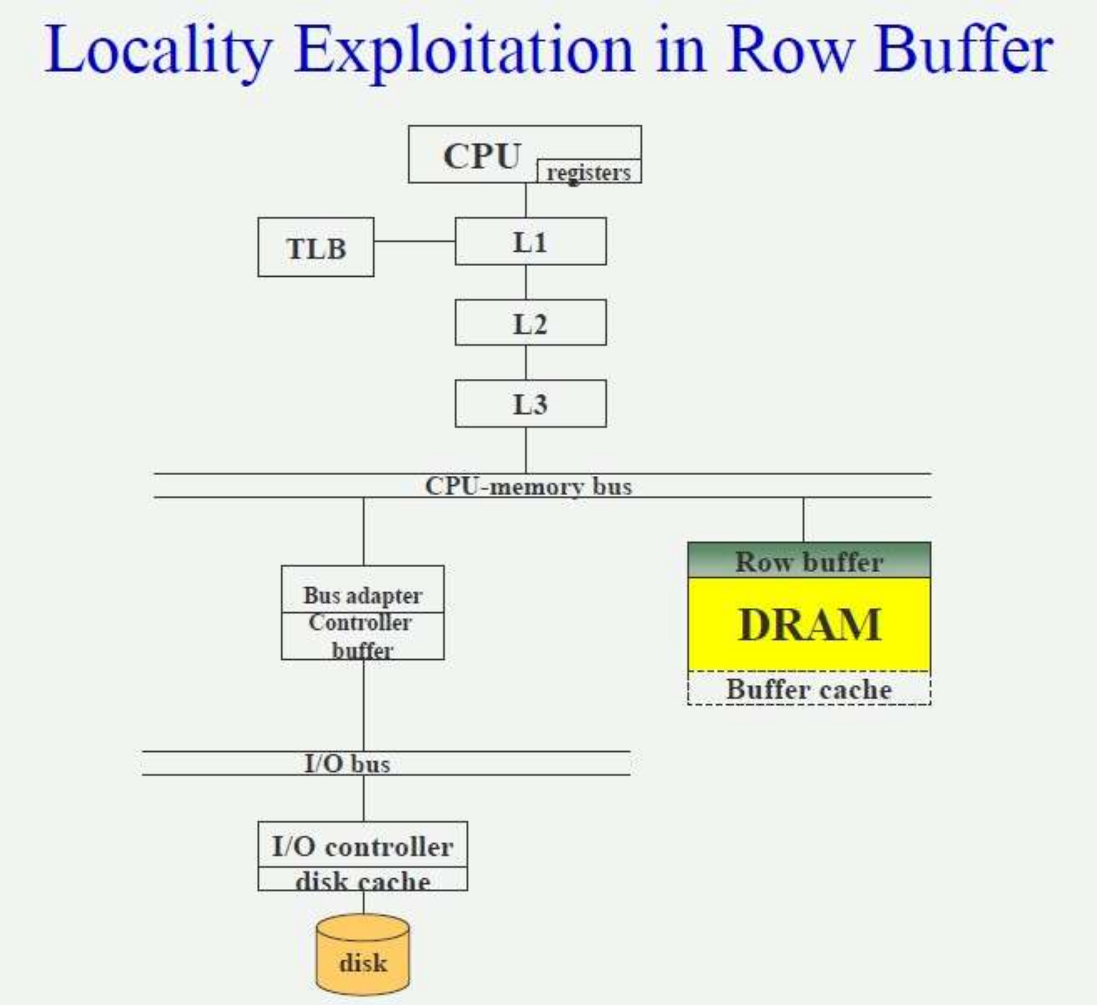


```
root@dr4000 # numademo -t 1g memset
2 nodes available
memory with no policy memset                    Avg 6358.97 MB/s
local memory memset                             Avg 6357.17 MB/s
memory interleaved on all nodes memset          Avg 4778.59 MB/s
memory on node 0 memset                         Avg 5836.55 MB/s
memory on node 1 memset                         Avg 3669.97 MB/s
memory interleaved on 0 1 memset                Avg 4508.74 MB/s
```



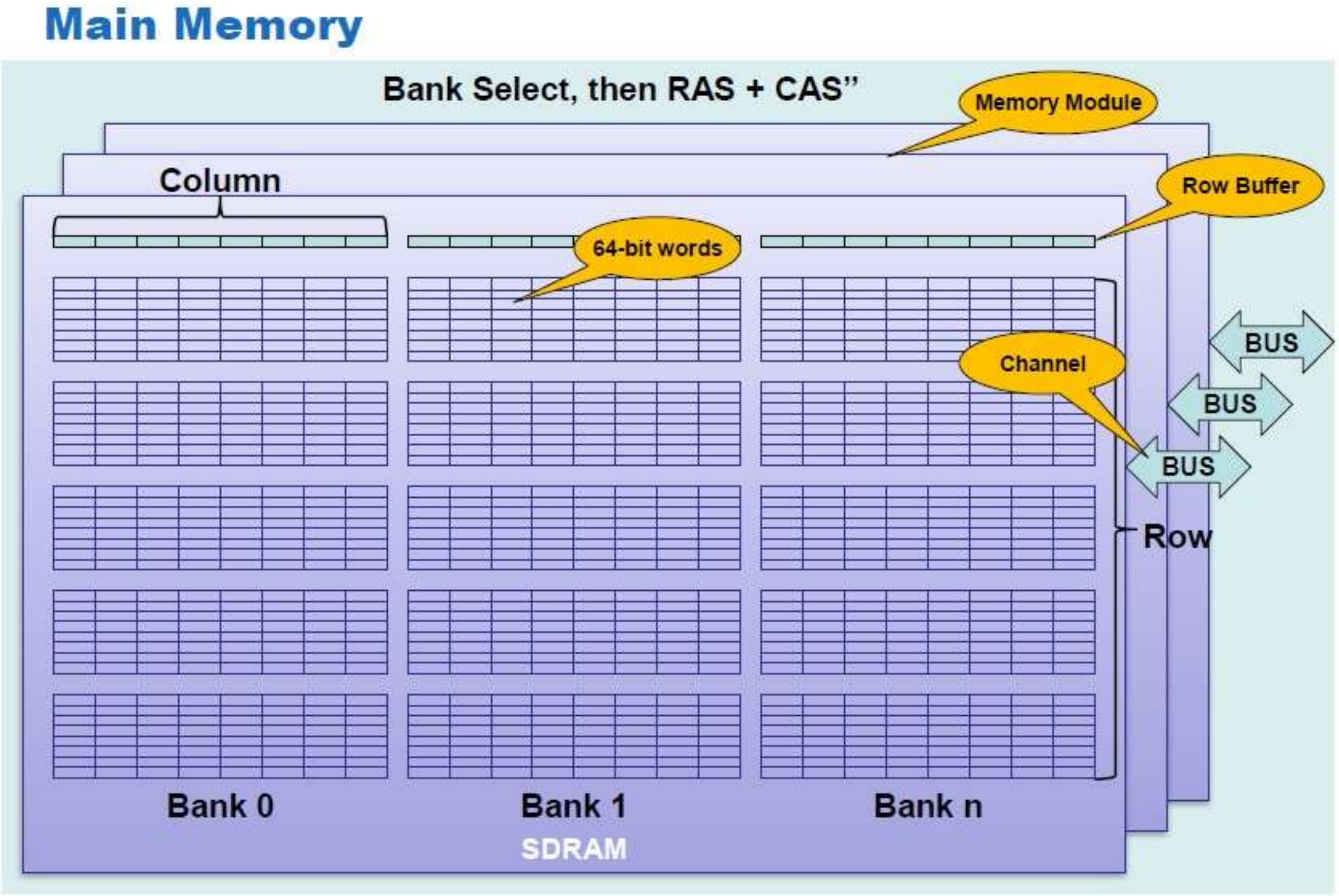
```
numa_foreign                    0          1709654
interleave_hit              17395            17388
local_node               66916103         72839750
other_node                1727189            43003
root@dr4000 #
```

DRAM总体位置



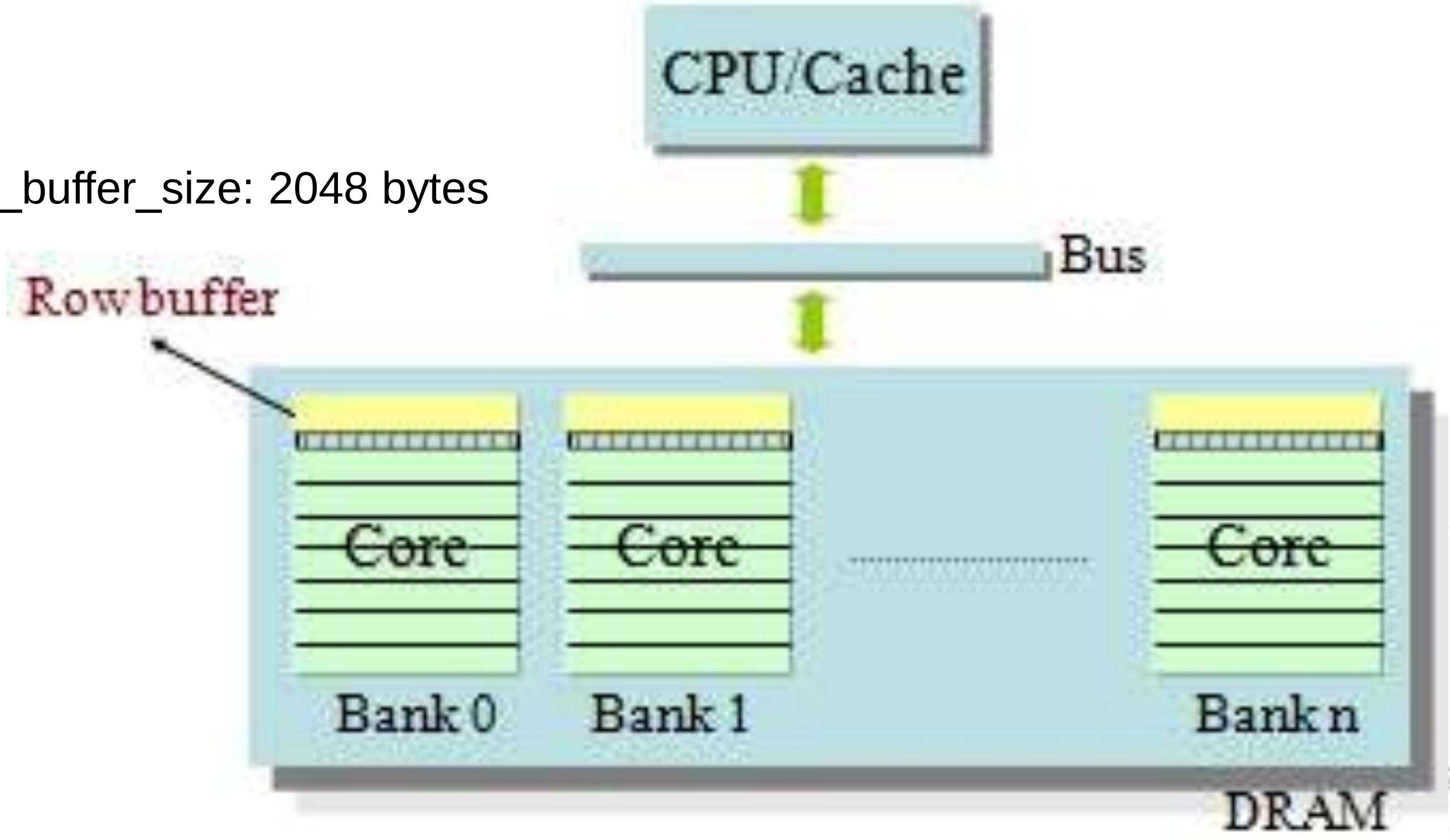
Locality Exploitation in Row Buffer

DRAM结构图


Main Memory
Bank Select, then RAS + CAS”
Memory Module
Column
Row Buffer
64-bit words
BUS
BUS
Channel
BUS
Row
Bank 0
Bank 1
Bank n
SDRAM

DRAM架构图

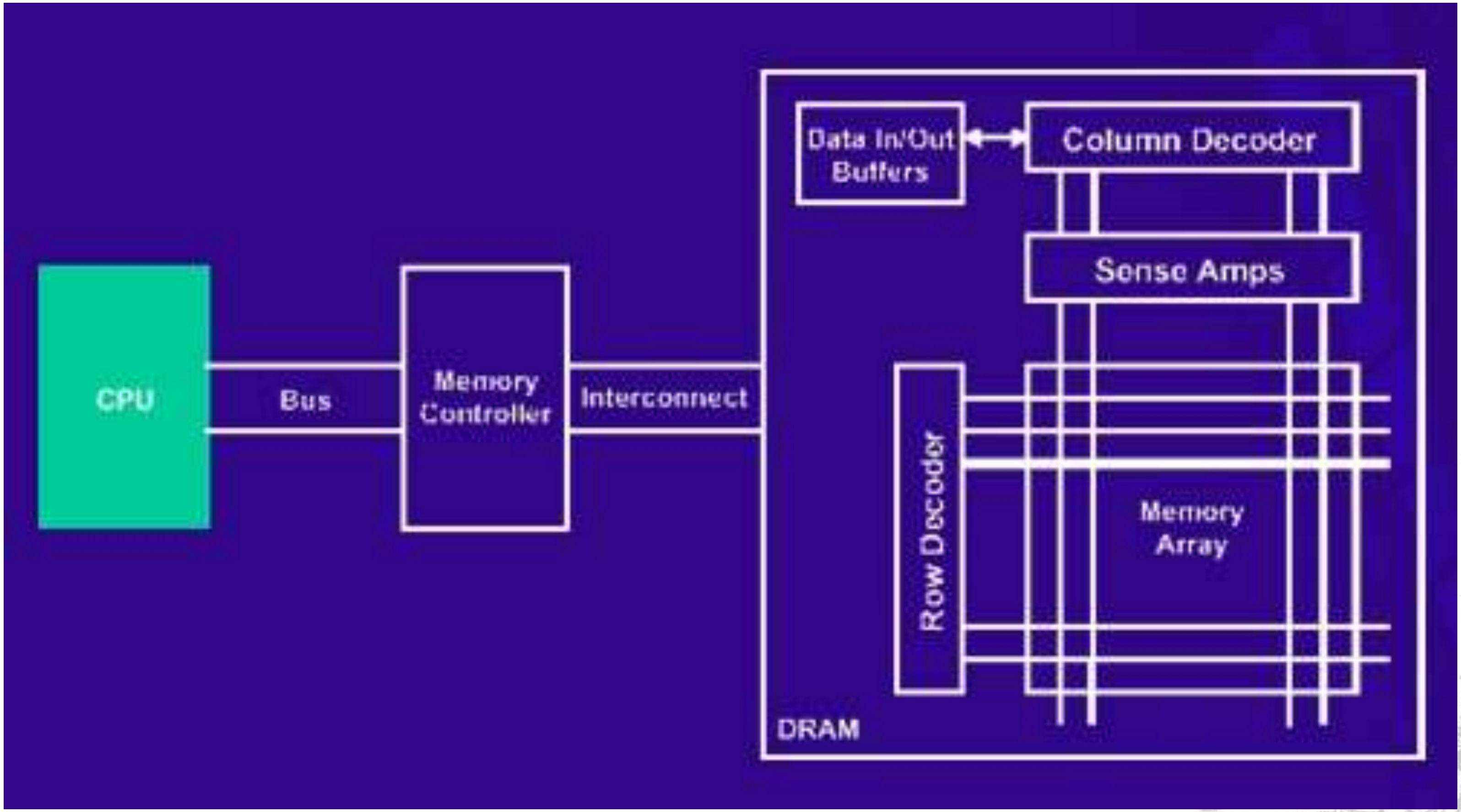
row_buffer_size: 2048 bytes



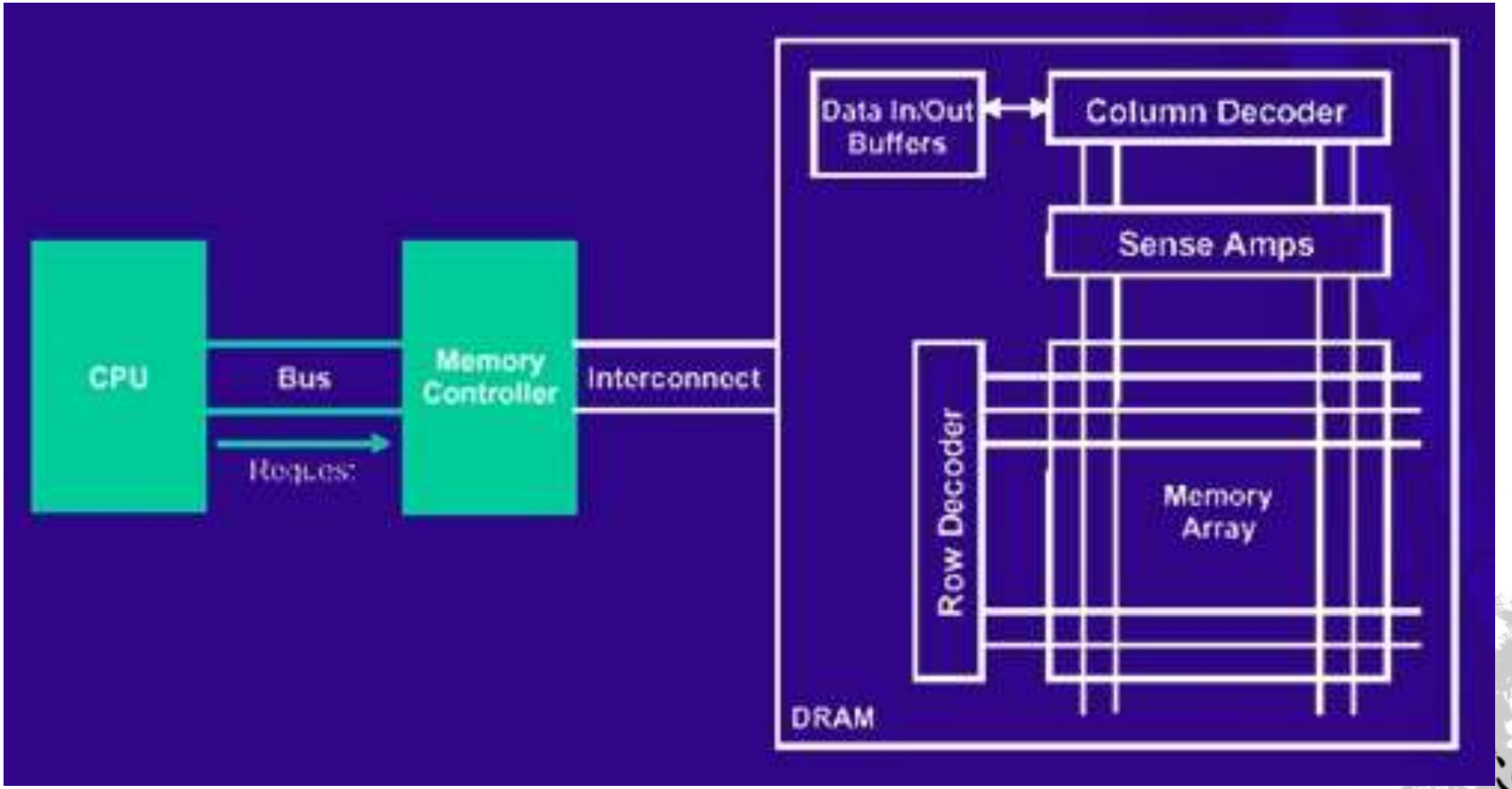
DRAM 访问流程

- Precharge: charge a DRAM bank before arow access
- Row access: activate a row (page) of a DRAM bank
- Column access: select and return a block of data in an activated row
- Refresh: periodically read and write DRAM to keep data

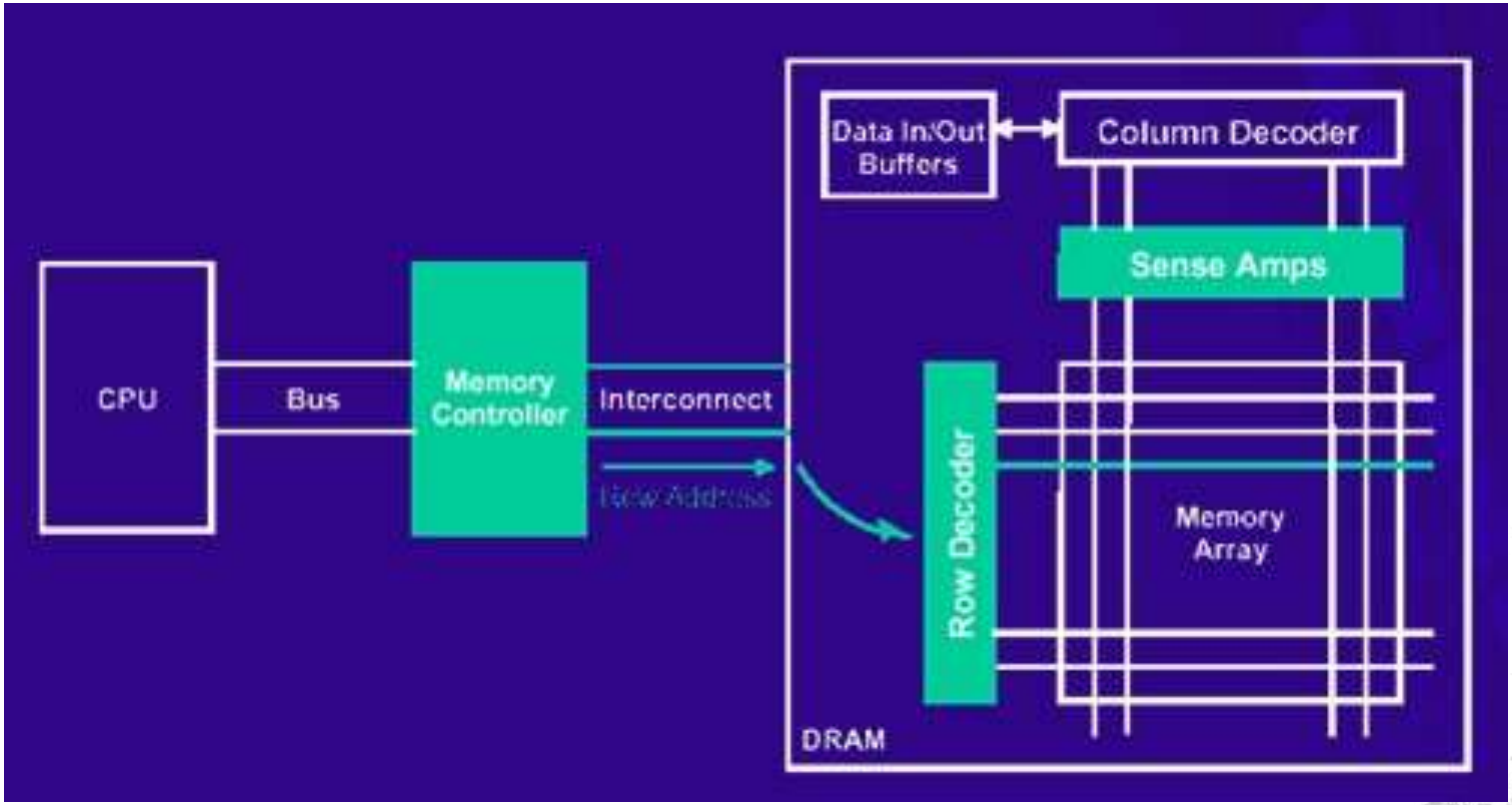
图解访问步骤1


CPU
Bus
Memory Controller
Interconnect
Data In/Out Buffers
Column Decoder
Sense Amps
Row Decoder
Memory Array
DRAM

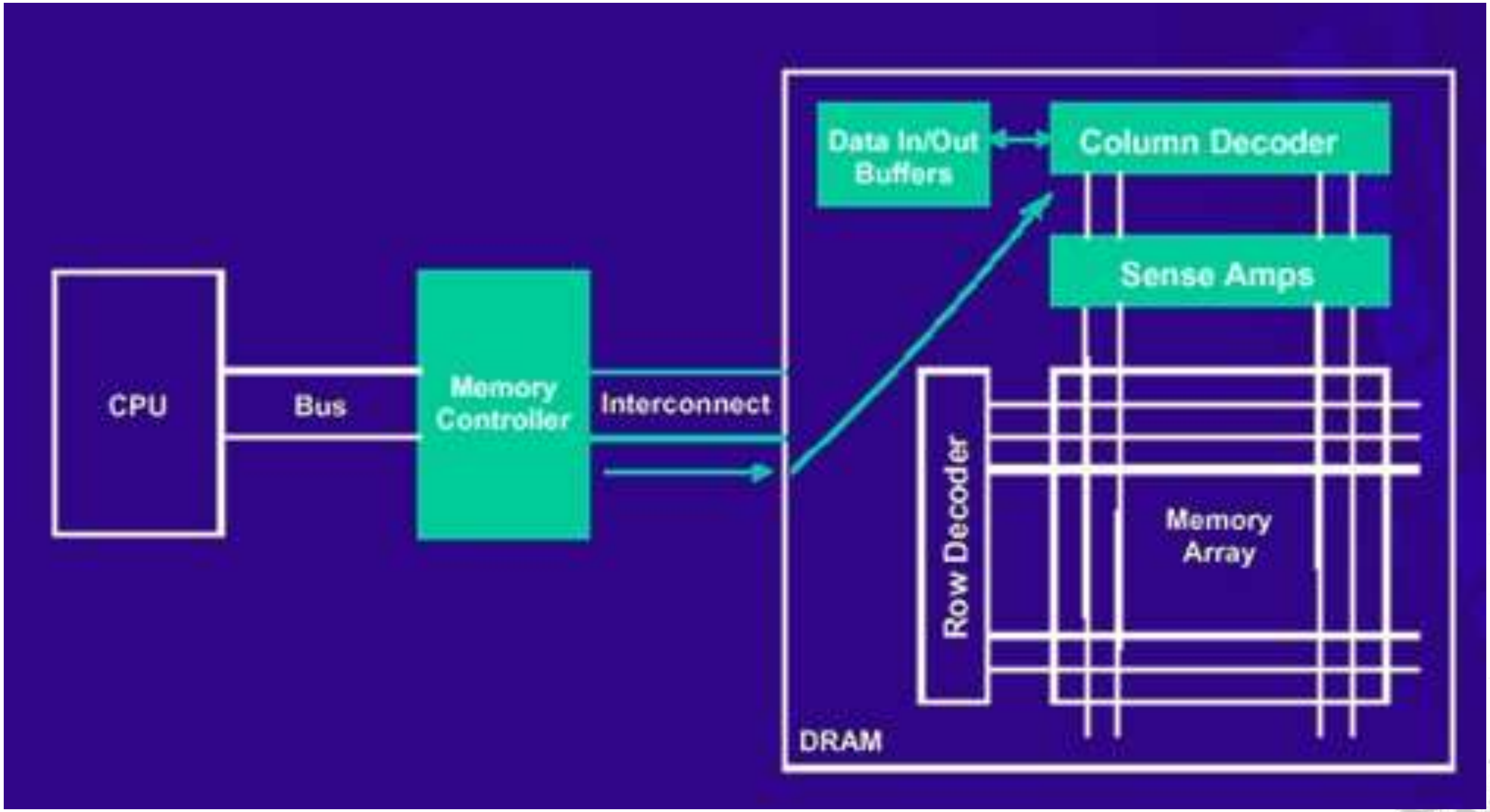
图解访问步骤2


Data In/Out
Buffers
Column Decoder
Sense Amps
CPU
Bus
Memory
Controller
Interconnect
Request
Row Decoder
Memory
Array
DRAM

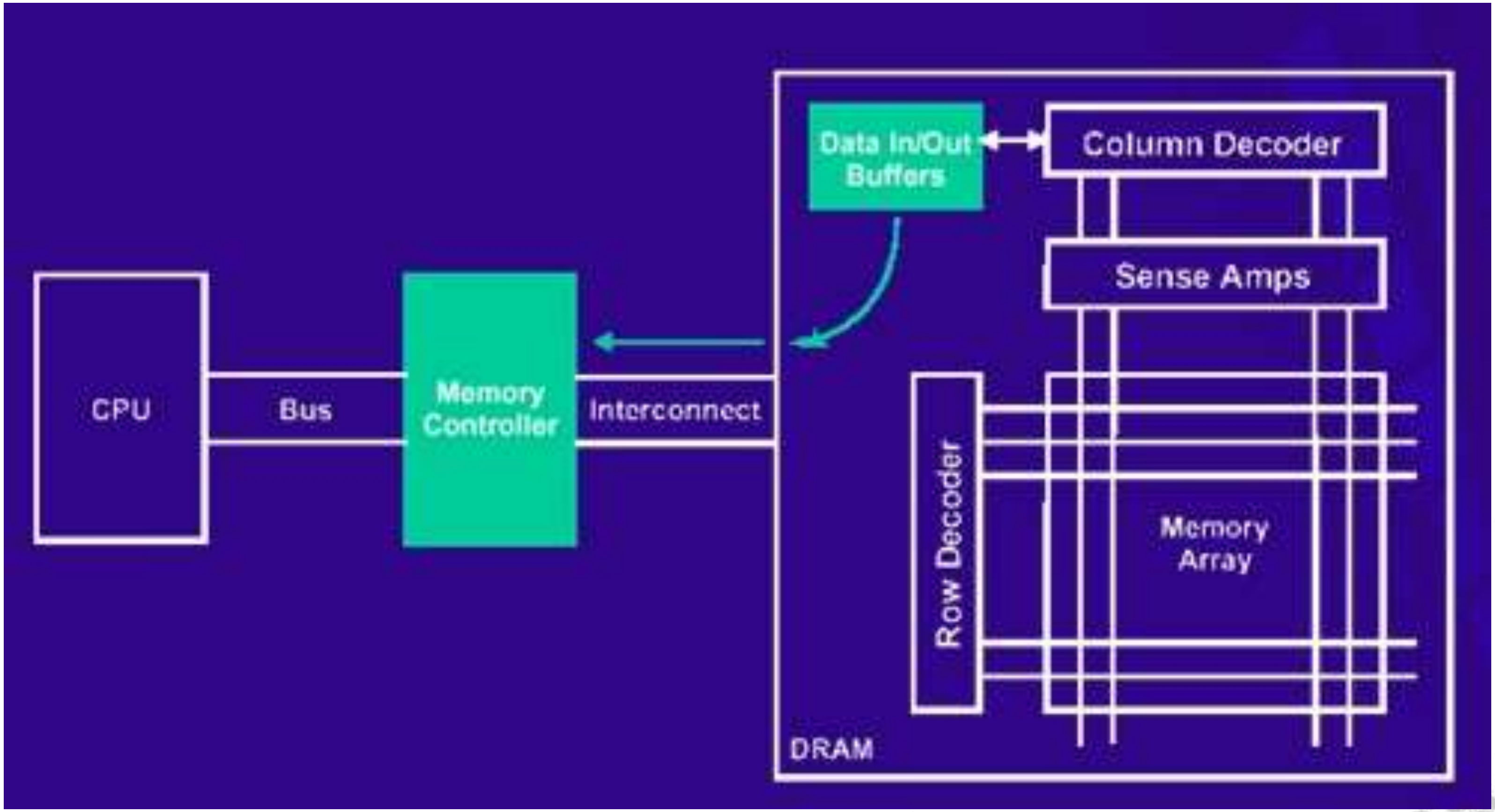
图解访问步骤3


Data In/Out
Buffers
Column Decoder
Sense Amps
CPU
Bus
Memory
Controller
Interconnect
Row Address
Row Decoder
Memory
Array
DRAM

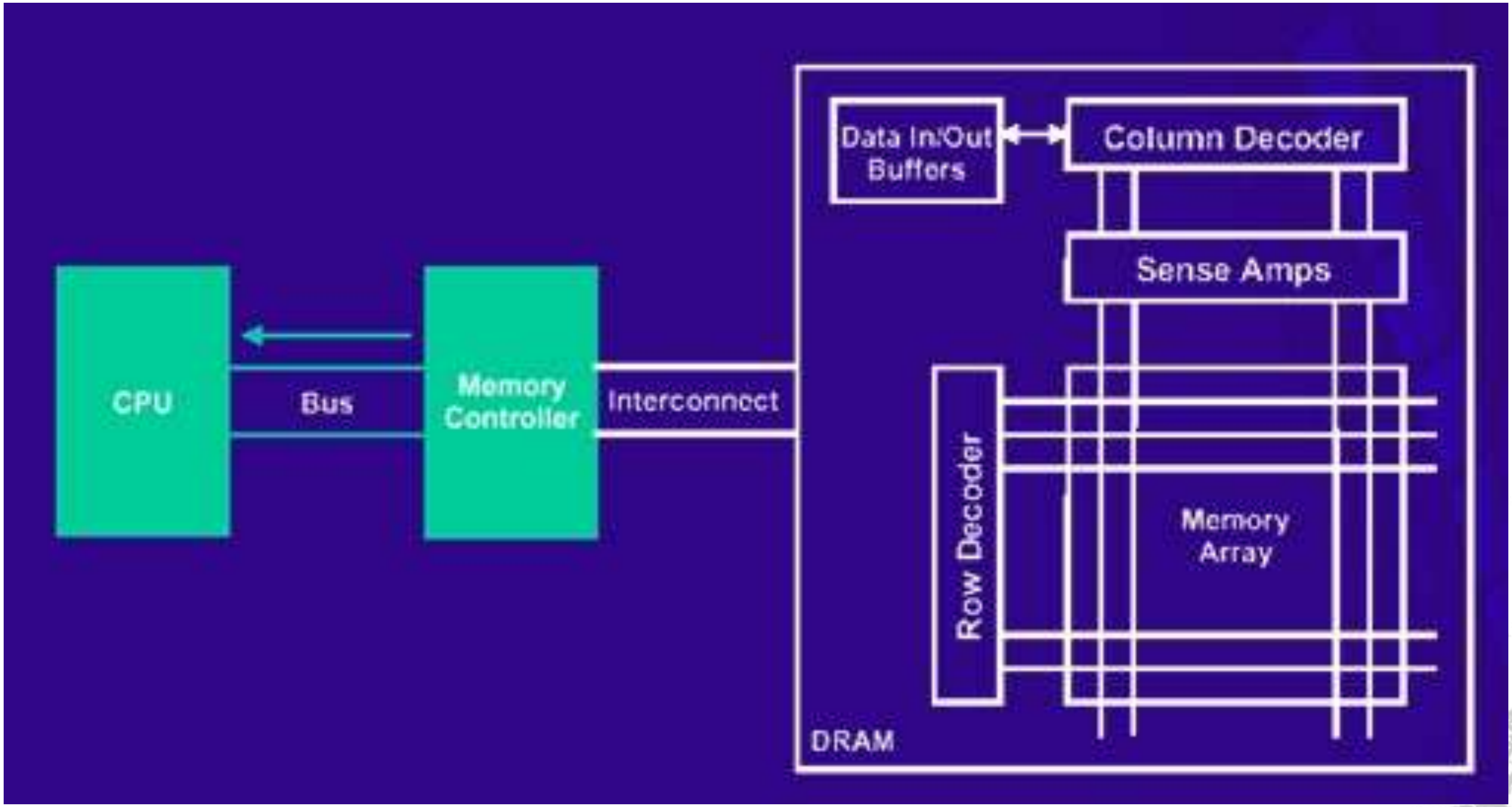
图解访问步骤4


CPU
Bus
Memory Controller
Interconnect
Data In/Out Buffers
Column Decoder
Sense Amps
Row Decoder
Memory Array
DRAM

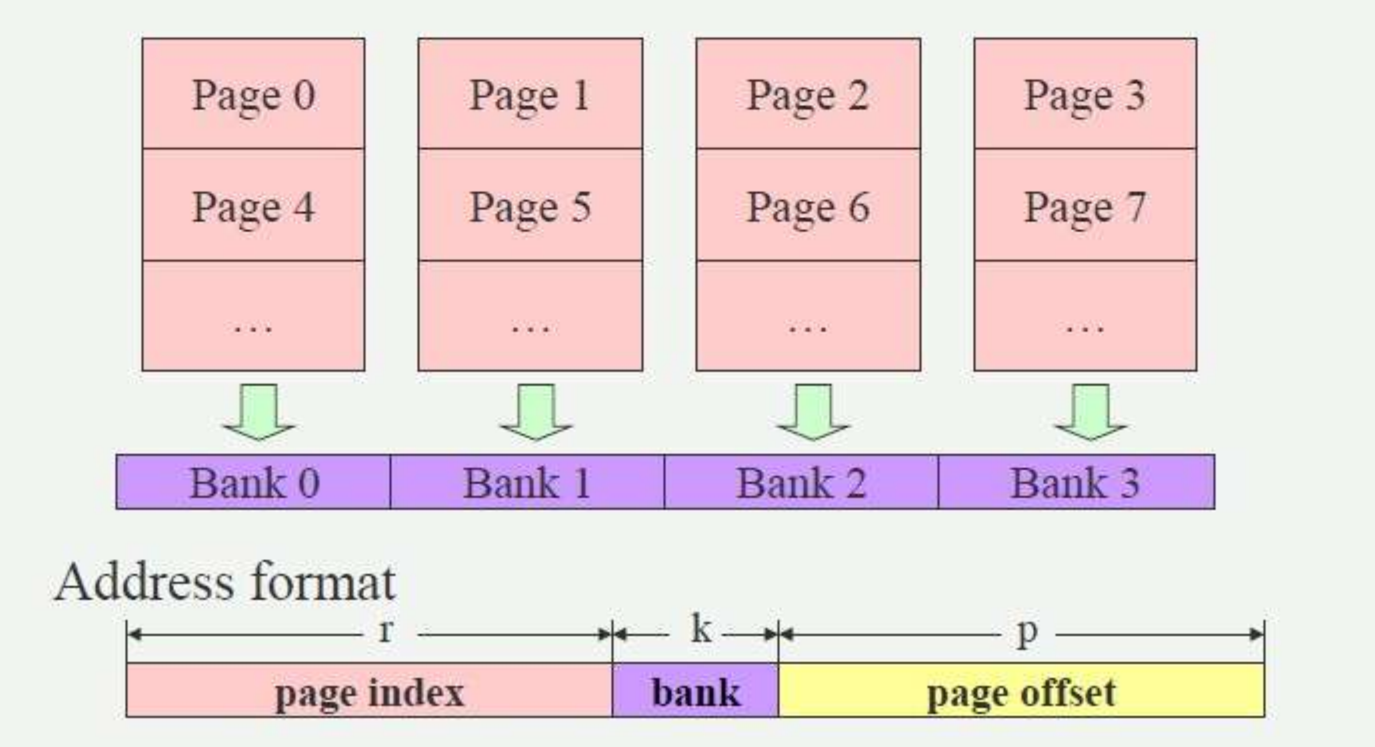
图解访问步骤5


CPU
Bus
Memory Controller
Interconnect
Data In/Out Buffers
Column Decoder
Sense Amps
Row Decoder
Memory Array
DRAM

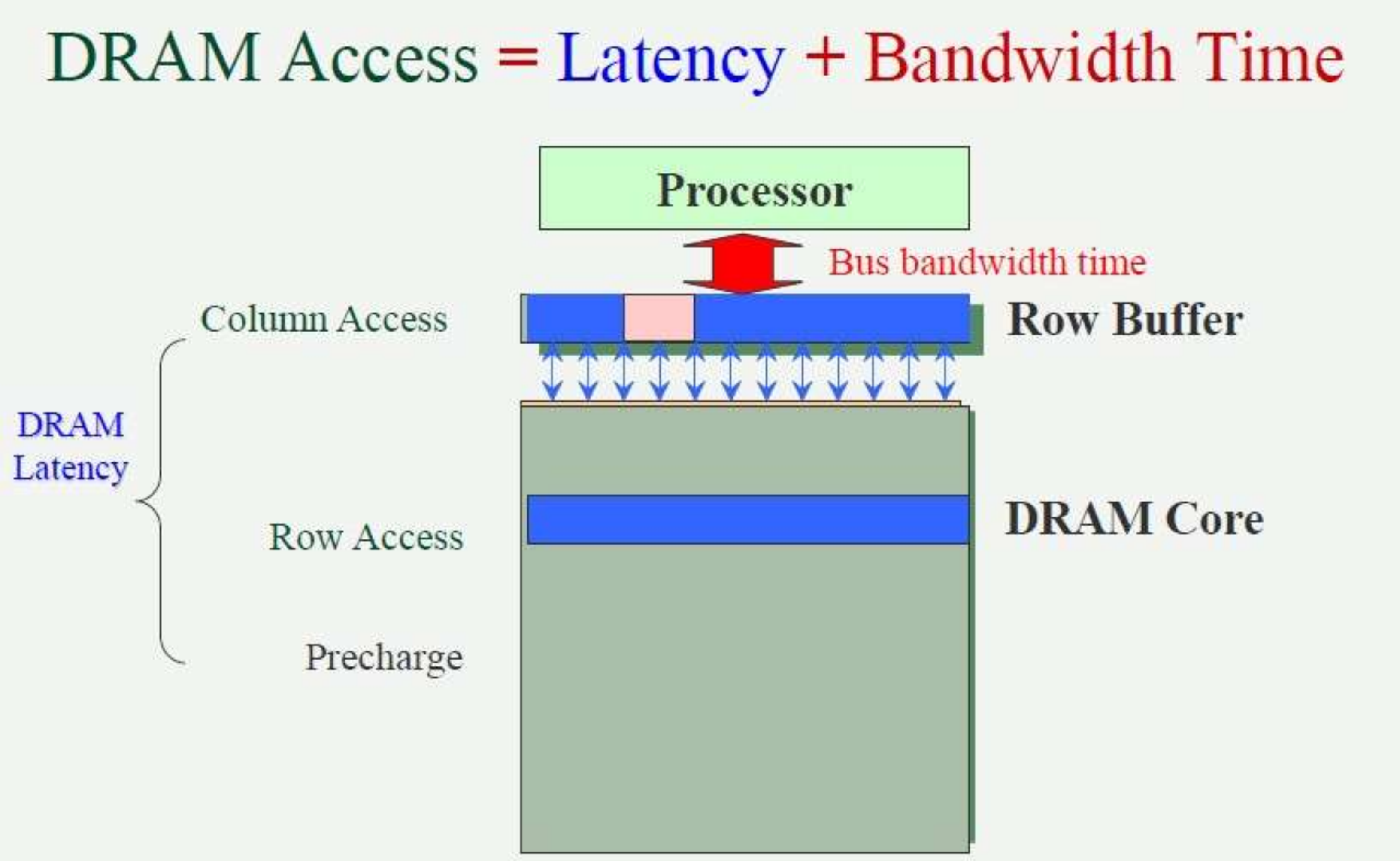
图解访问步骤6


CPU
Bus
Memory Controller
Interconnect
Data In/Out Buffers
Column Decoder
Sense Amps
Row Decoder
Memory Array
DRAM

DRAM bank选择

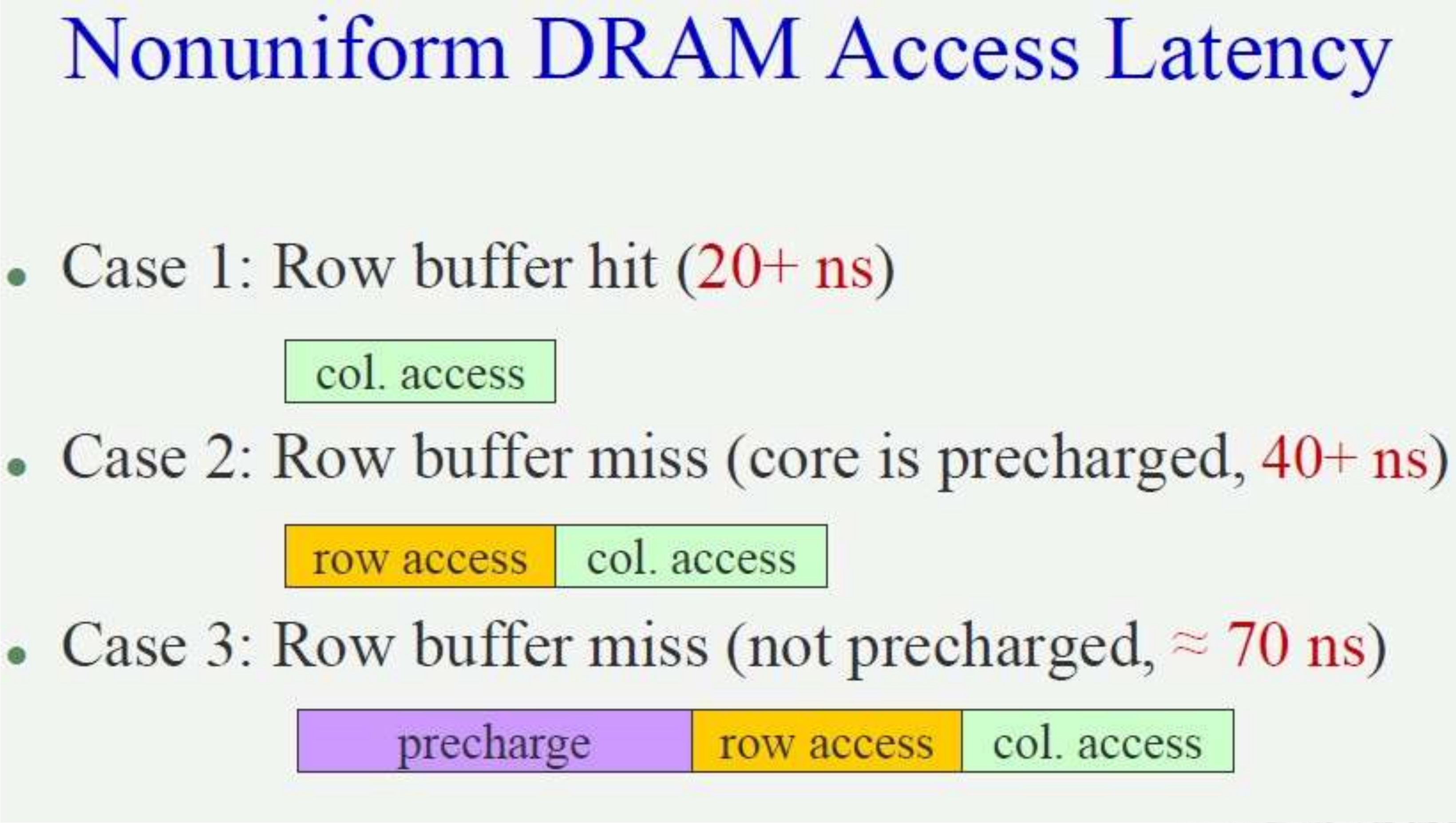

Page 0
Page 1
Page 2
Page 3
Page 4
Page 5
Page 6
Page 7
…
…
…
…
Bank 0
Bank 1
Bank 2
Bank 3
Address format
r
k
p
page index
bank
page offset

DRAM访问延时分布



DRAM Access = Latency + Bandwidth Time

DRAM不同情况下访问延迟



Nonuniform DRAM Access Latency

- Case 1: Row buffer hit (20+ ns)

col. access

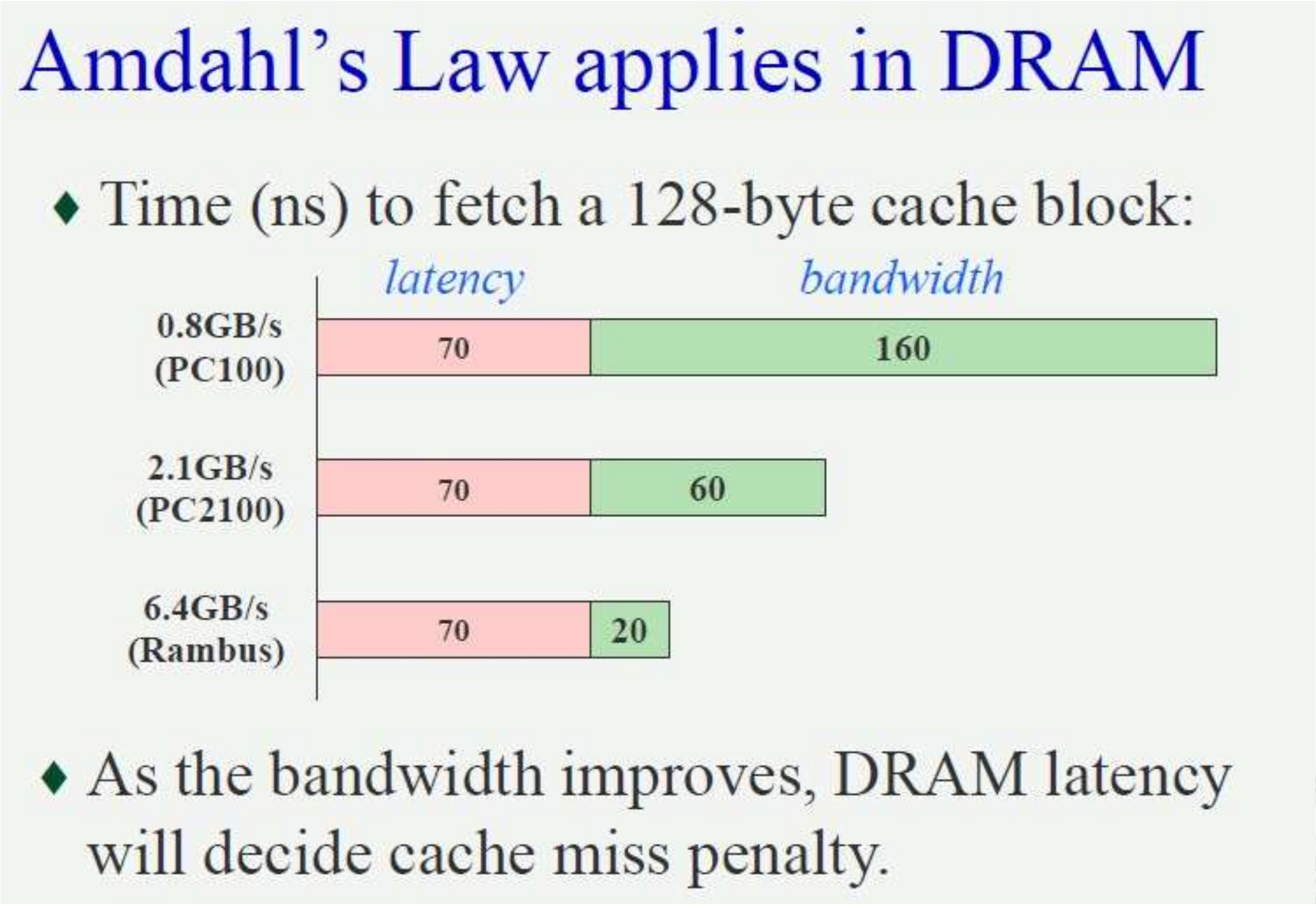
- Case 2: Row buffer miss (core is precharged, 40+ ns)

row access | col. access

- Case 3: Row buffer miss (not precharged, ≈ 70 ns)

precharge | row access | col. access

DRAM延时演化



Amdahl’s Law applies in DRAM

- Time (ns) to fetch a 128-byte cache block:

- As the bandwidth improves, DRAM latency will decide cache miss penalty.

参考

- http://en.wikipedia.org/wiki/Prefetch_buffer
- 详解服务器内存带宽计算和使用情况测量: http://blog.yufeng.info/archives/1511
- DDR3 内存带宽如何计算: http://zhidao.baidu.com/question/107154668
- hwconfig查看硬件信息: http://blog.yufeng.info/archives/2086
- Exploiting Locality in DRAM, Xiaodong Zhang

提问时间

谢谢大家！