

The Main Memory

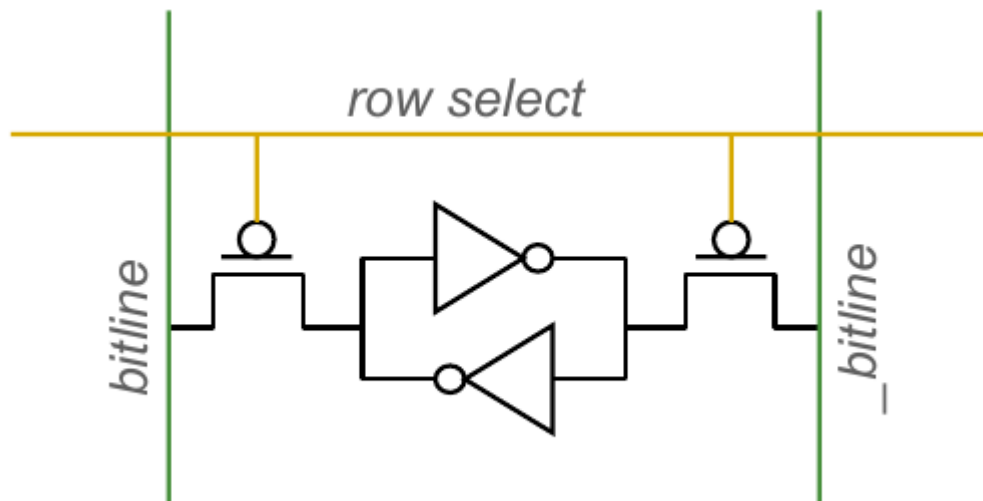
Technology, Organization and Architecture

Computer System Architecture (CS5202)
IIT Tirupati

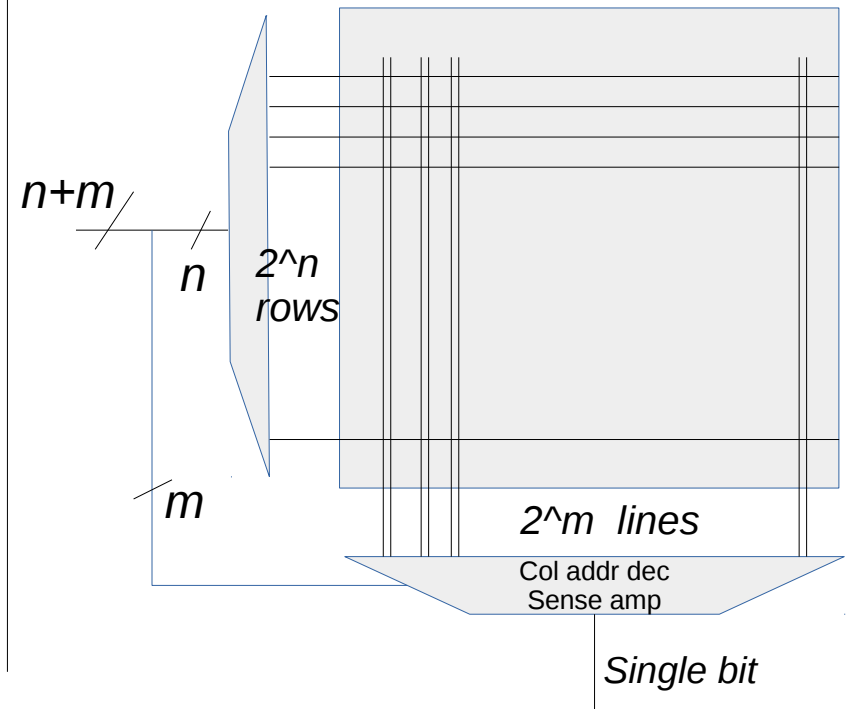
Jaynarayan t tudu
jtt@iittp.ac.in
17th March 2020

Memory Cell: storing a bit

SRAM Cell (6 Transistors cell)



SRAM Memory Array



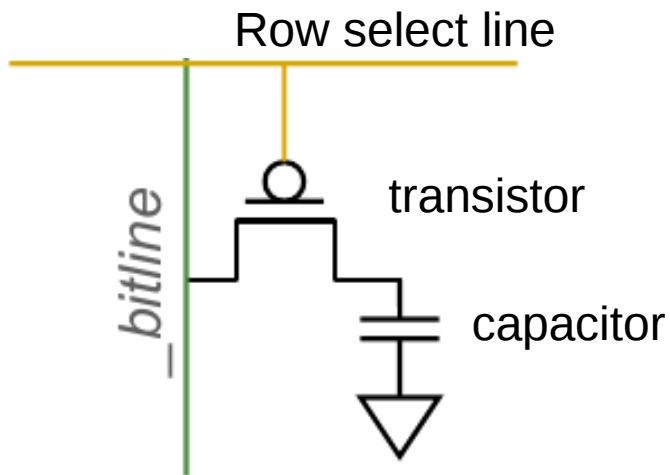
- 1) Address decode
- 2) Drive row select
- 3) Selected bit-cells drive bitlines
- 4) Sense amplifier senses the bit difference
- 5) Column address decode and select
- 6) Pre-charge all the bitlines for next read/write

The same structure can be replicated for simultaneous read or write of multiple data

Step 2 and 3 dominates the access time, Step 2, 3 and 5 dominates the cycle time

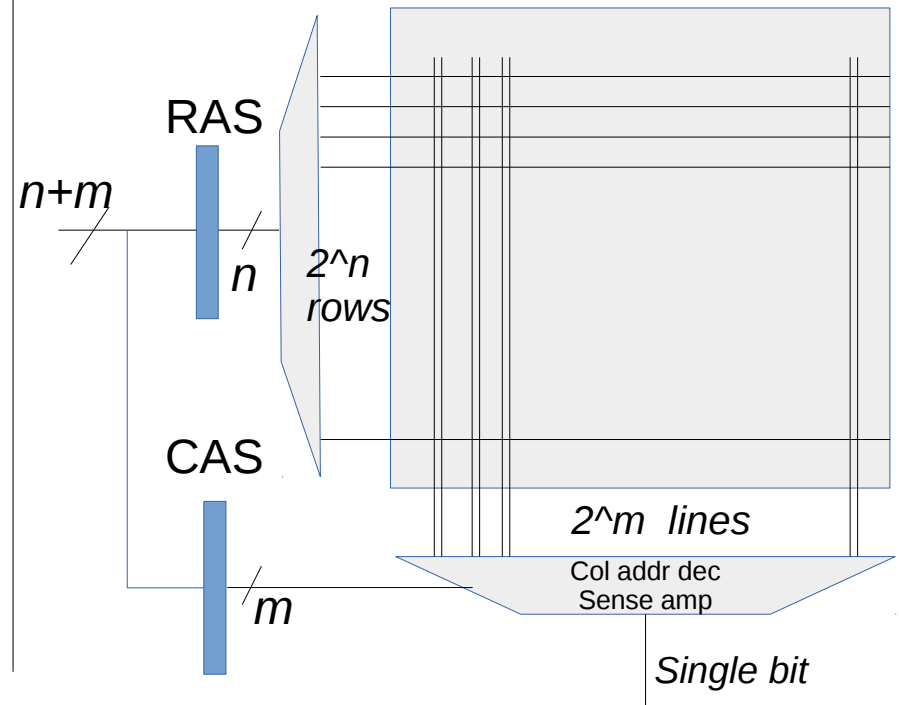
Memory Cell: storing a bit

DRAM Cell (1 transistor + 1 capacitor)
[Dynamic RAM]



- 1) Row address decode from RAS
- 2) Drive row select
- 3) Selected bit-cells drive bitlines
- 4) Sense amplifier senses the bit difference
- 5) Column address decode from CAS and select
- 6) Pre-charge all the bitlines for next read/write

DRAM Memory Array



RAS: Row Address Strobe
CAS: Column Address Strobe

The same structure can be replicated for simultaneous read or write of multiple data

Refresh: Needed to restore the charge stored in capacitor.
This has to be done periodically (typically 10s of ms)

SRAM compared to DRAM

- SRAM

- Used for L1, L2 caches and reg file
- Fast access
- No refresh
- Lower density
 - To store a single bit needs 6 transistor
- Higher cost

- DRAM


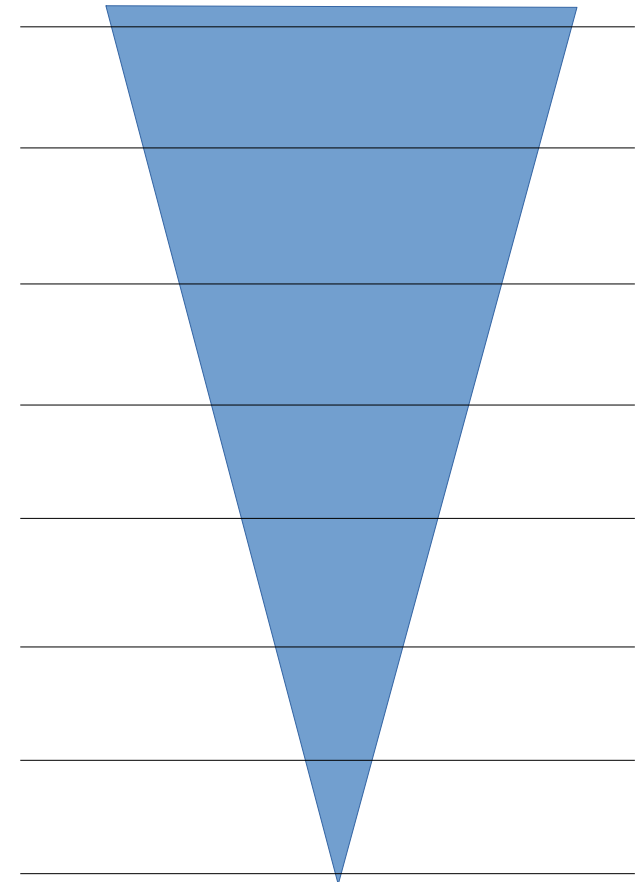
- Main memory
- Higher access time
- Refreshing required
- Higher capacity
- Higher density
 - To store a single bit just 1 tran and 1 cap
- Cost per bit less

SRAM: **Static** Random Access Memory

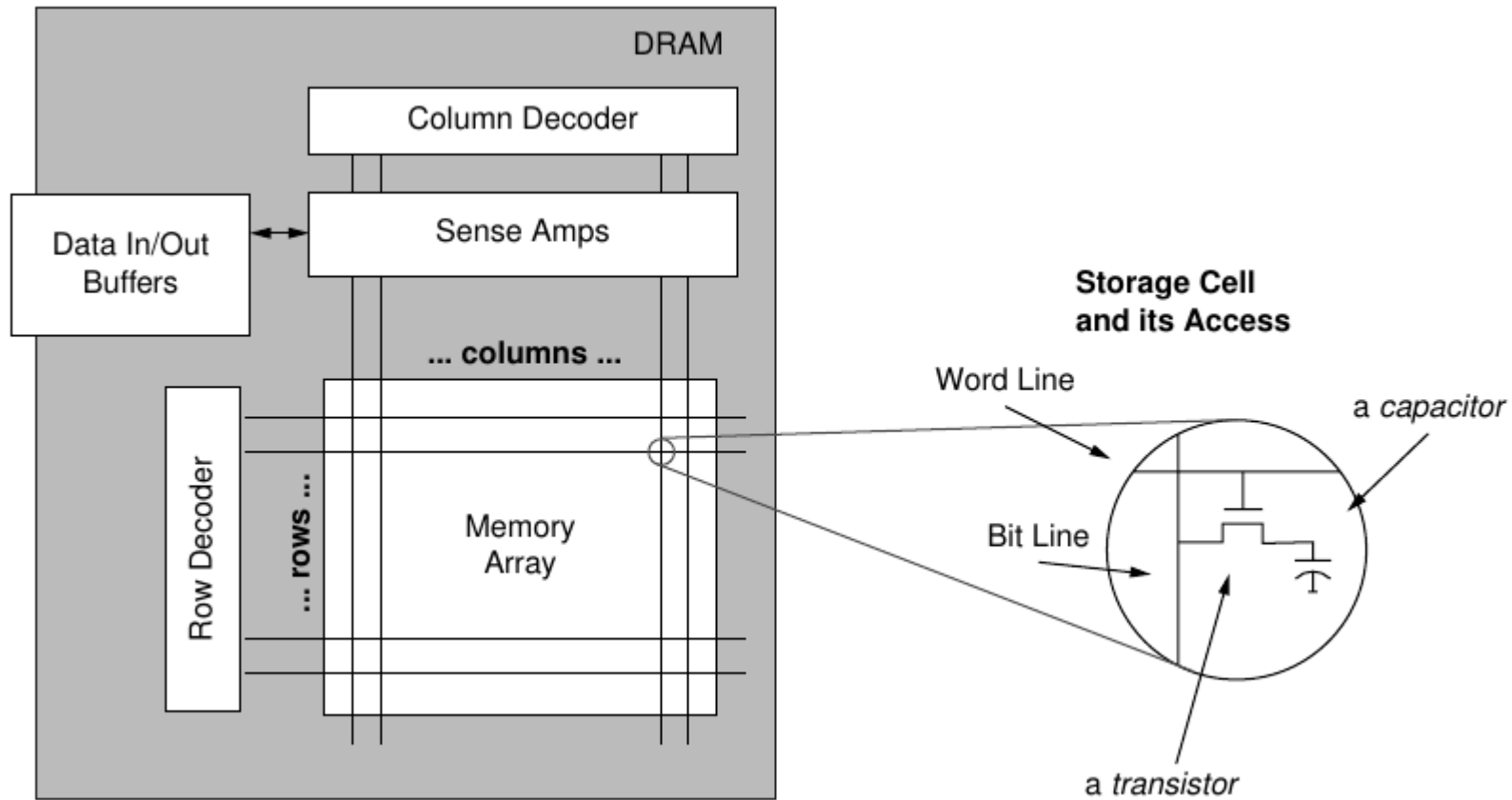
DRAM: **Dynamic** Random Access Memory

DRAM/Main Memory Organization

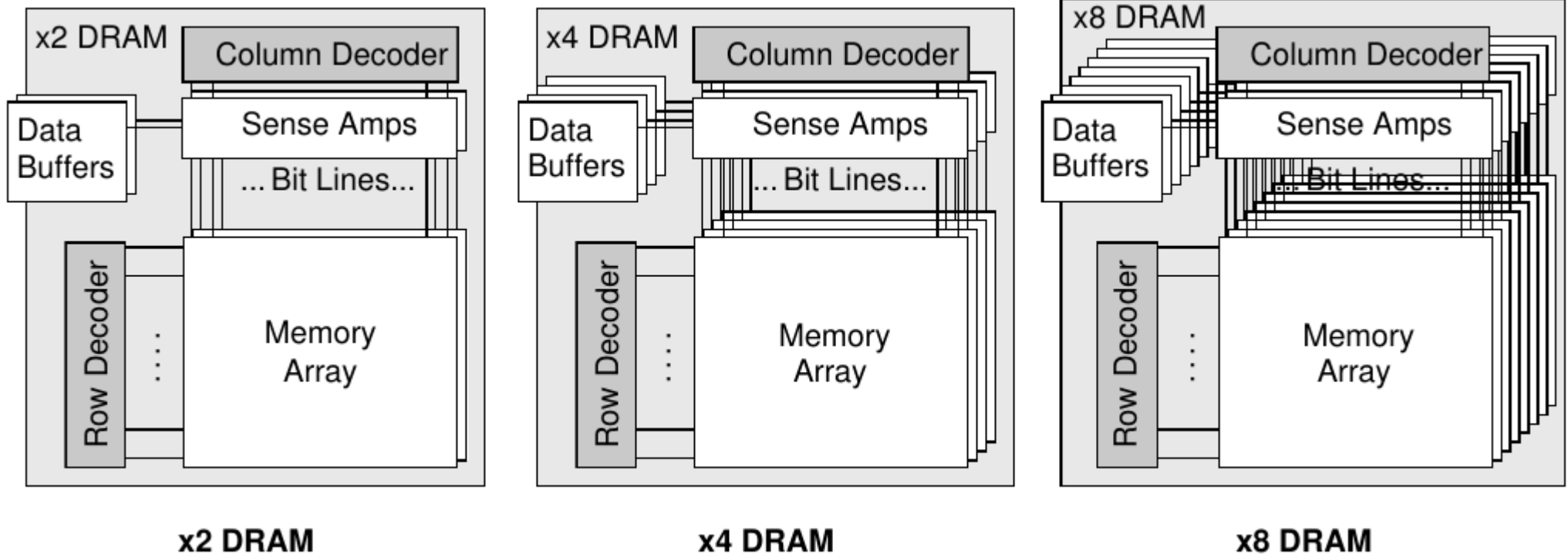
Channel
DIMM
Rank
Bank (Chip)
Memory Array
Row/Column
Memory Cell
Transistor/Capacitor

An upward-pointing arrow is positioned to the right of the text list, indicating the direction of increasing abstraction or organization level from the physical components at the bottom to the system-level components at the top.

Memory Array: Storing a set of bits

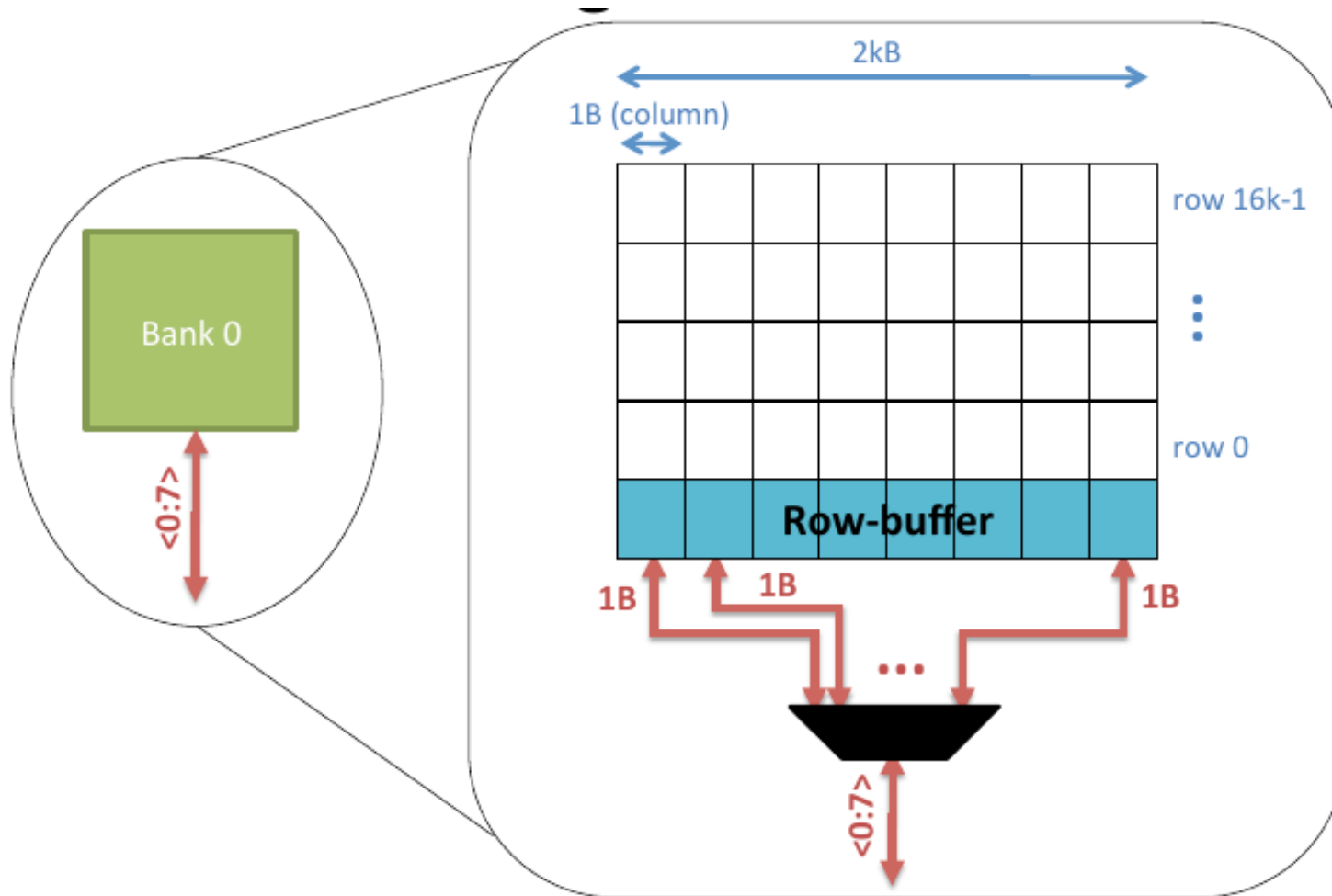


Bank: Multiple Memory Array



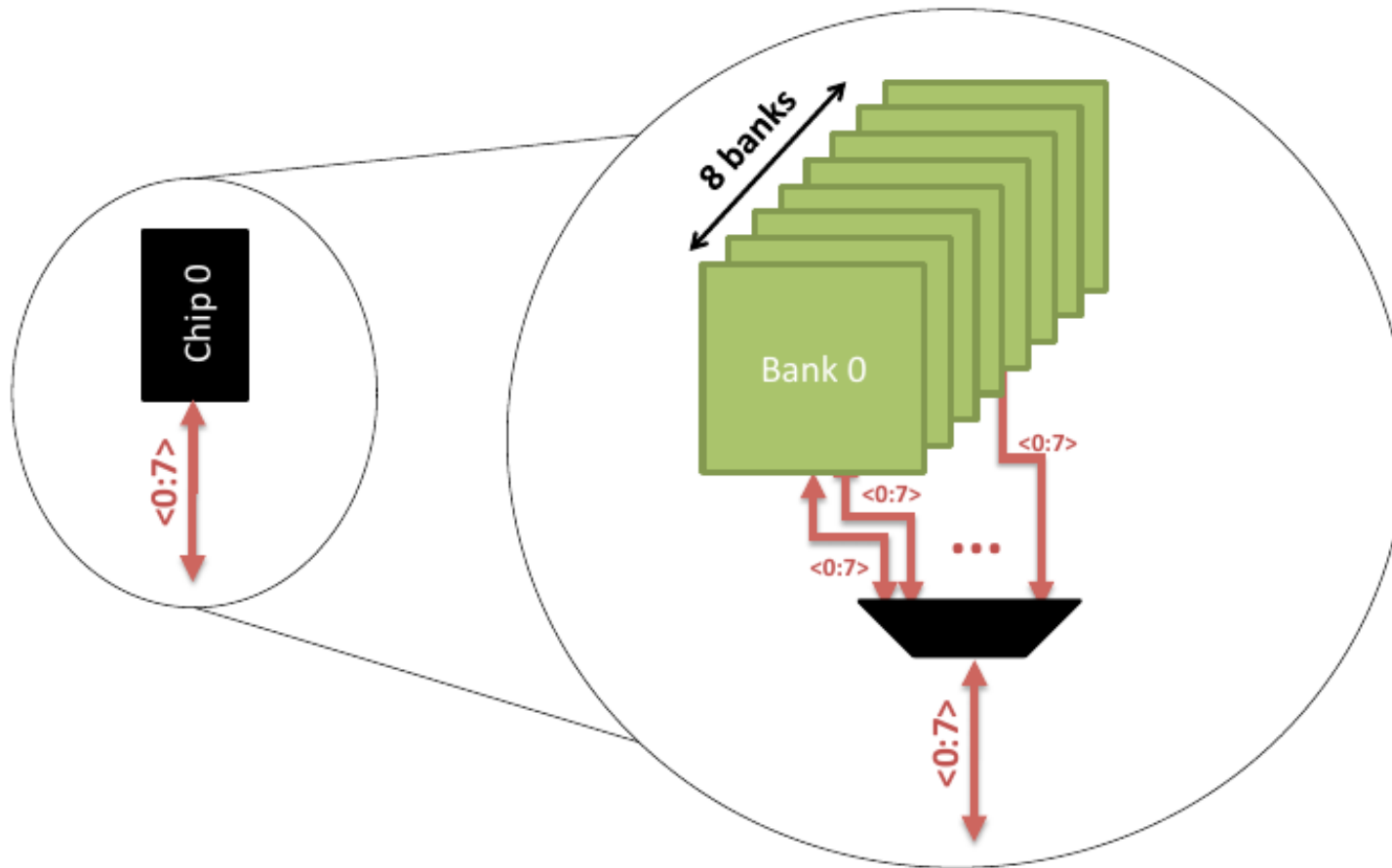
x2 DRAM: by two DRAM array. It can outputs 2 bits at a time.
x4 DRAM can output 4 bits and
X8 DRAM can output 8 bits at a time

Bank: Multiple Memory Array



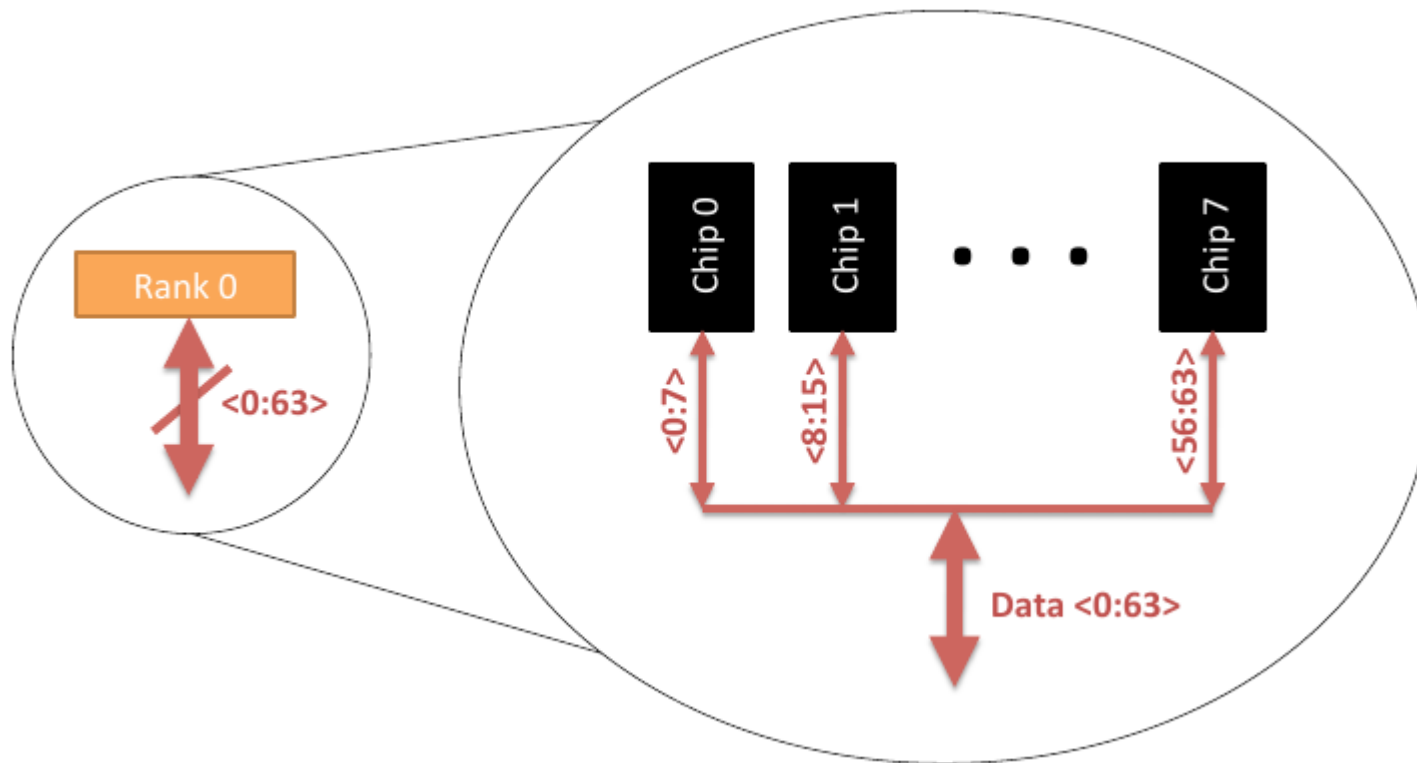
- x8 DRAM organization: at a time a byte of data can be read out or written into. Such organization is called Bank.
- Multiple bank can be organized for larger size of DRAM

Chip: Multiple Banks



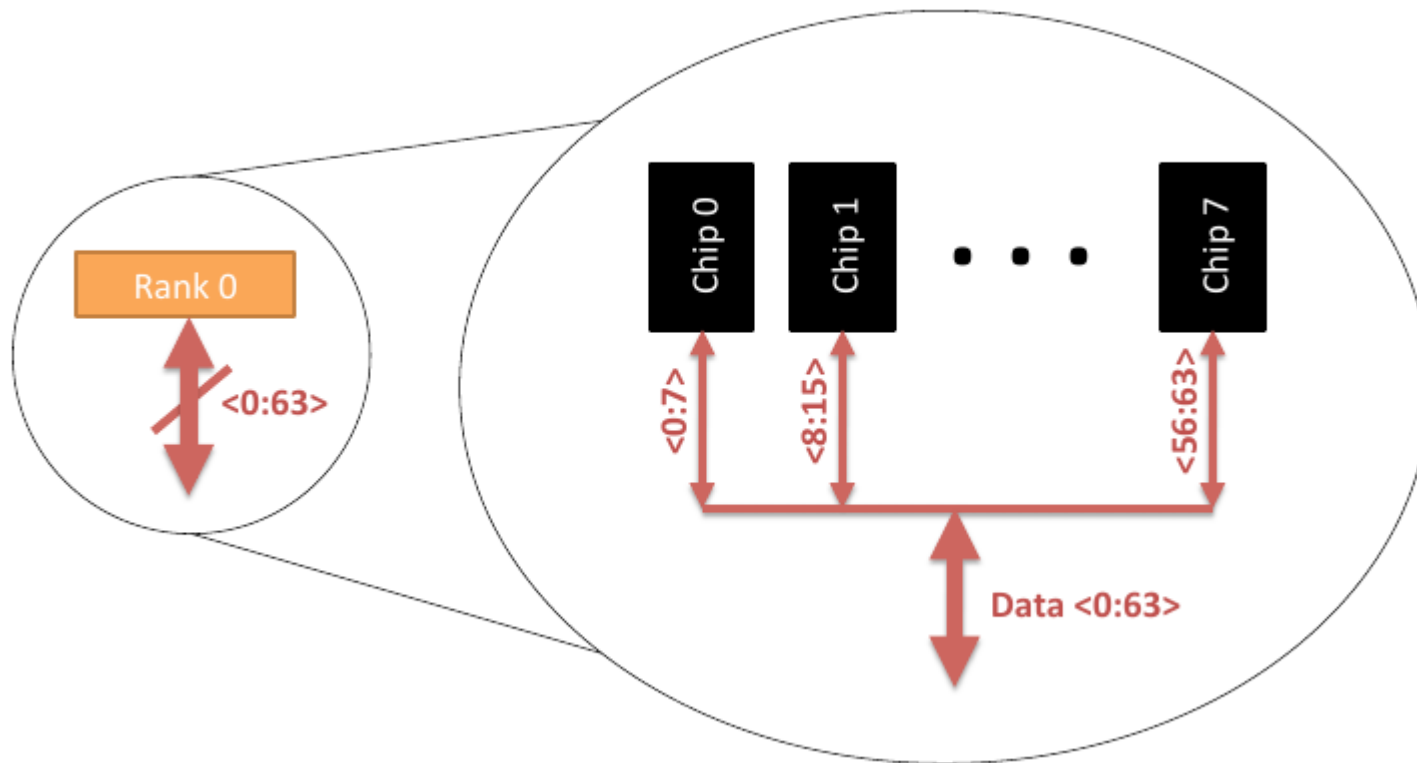
- 8 banks are organized into a single chip
- At a time 8 bits of data can read out and write in

Rank: Organization of Multiple Chips



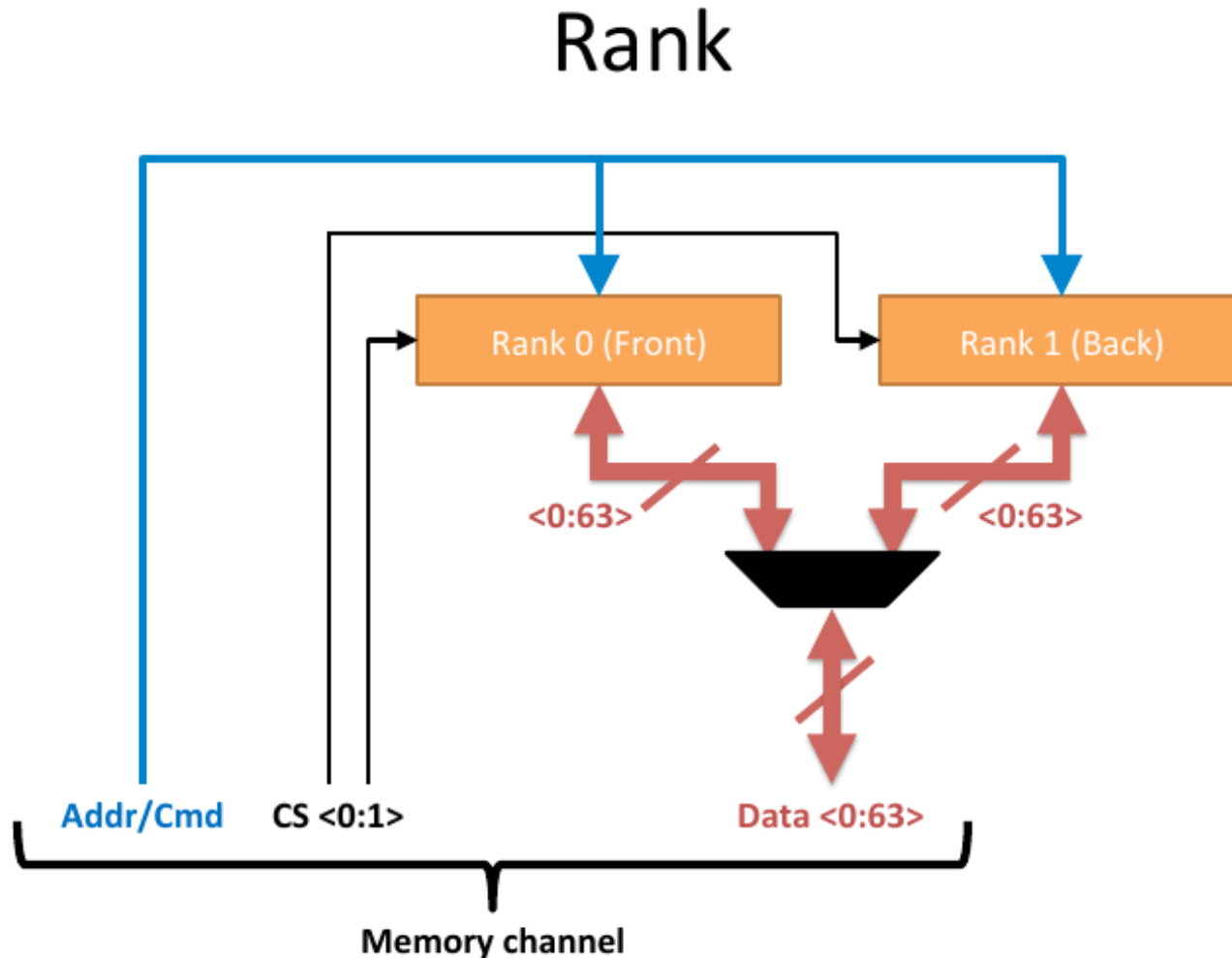
- Increasing the bandwidth to 64 bits (8 bytes)
- A single Rank can be formed by organizing multiple DRAM chips

Rank: Organization of Multiple Chips



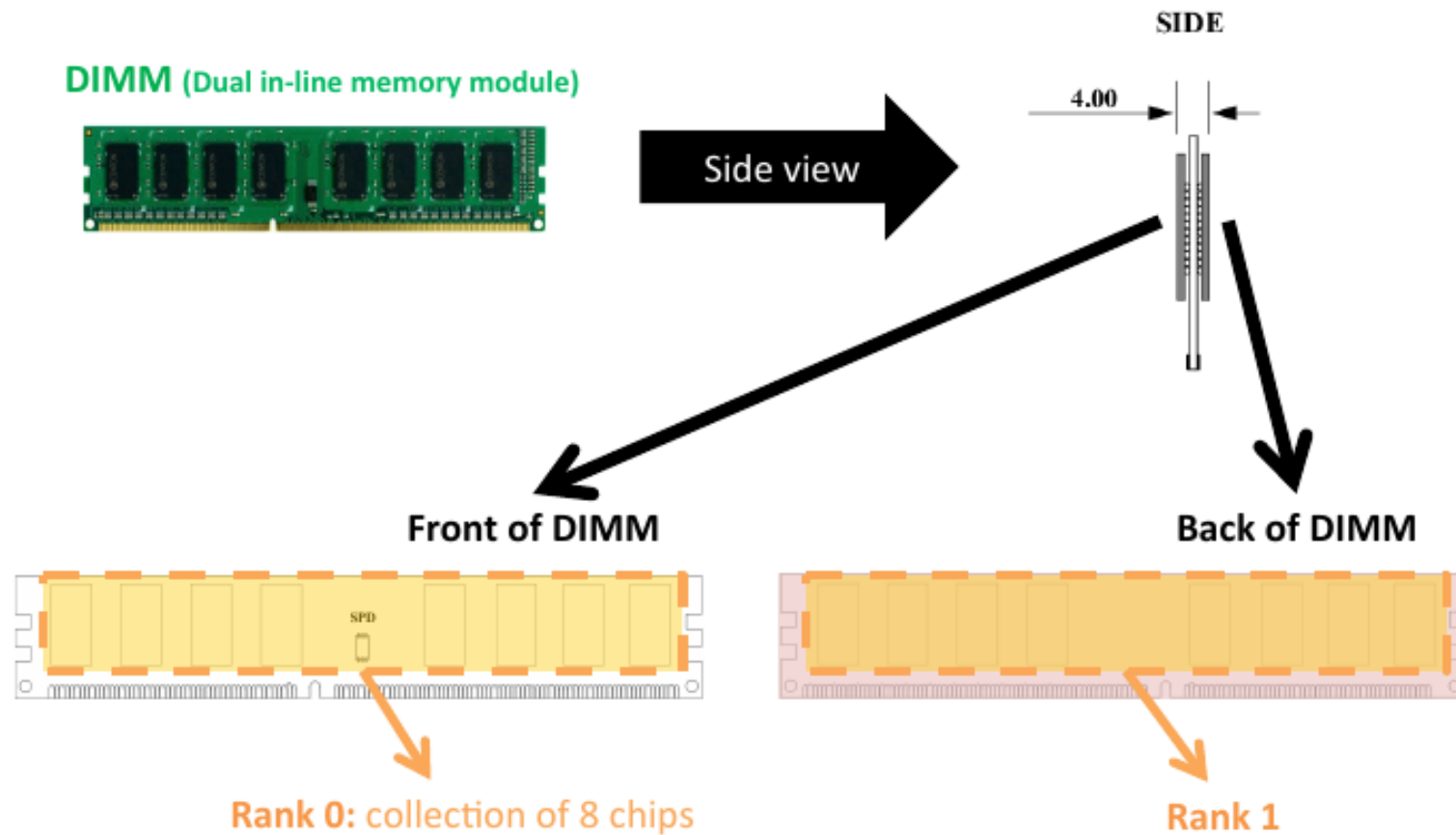
- Increasing the bandwidth to 64 bits (8 bytes)
- A single Rank can be formed by organizing multiple DRAM chips

Organization of Multiple Ranks



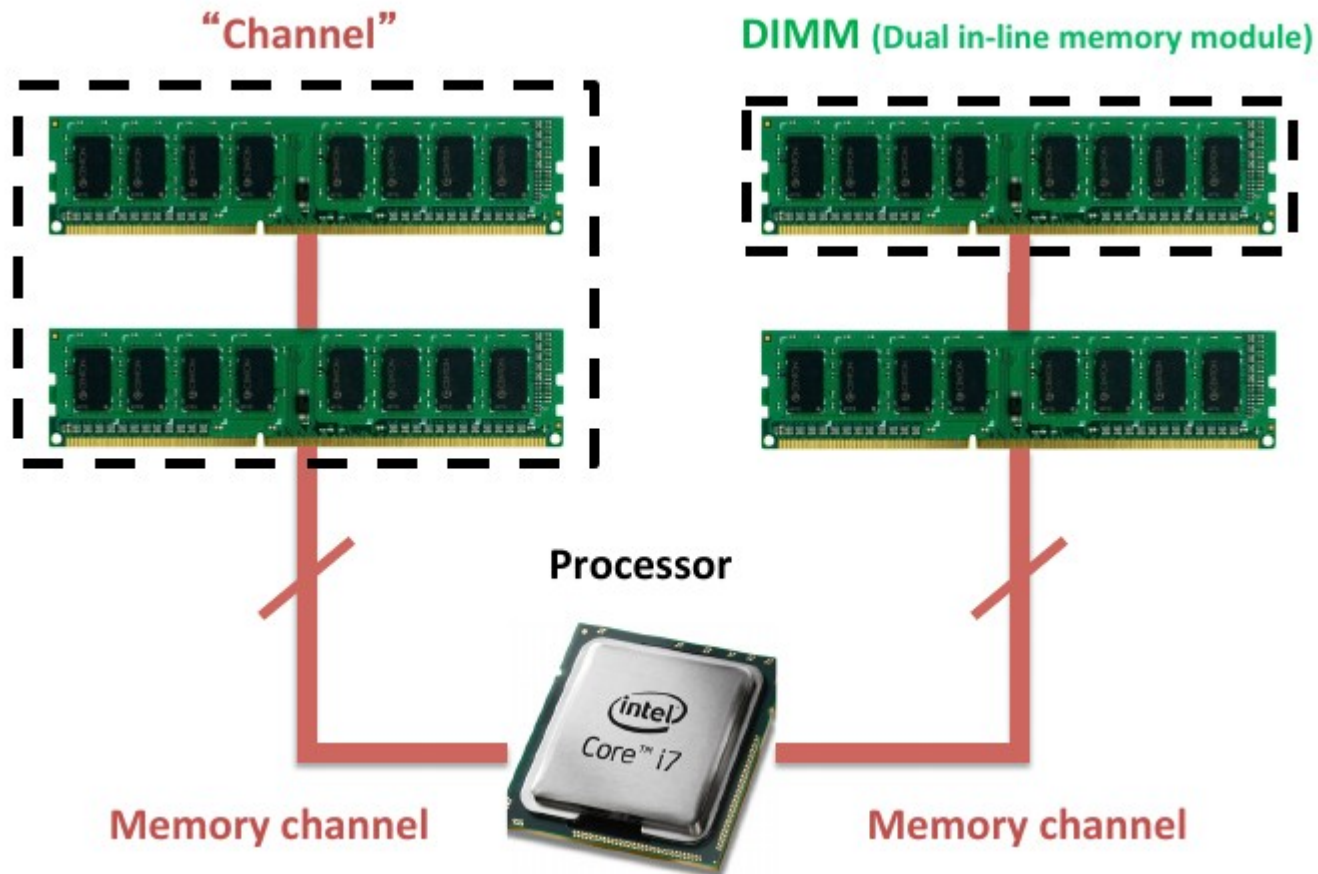
- Address and command (Addr/Cmd) to locate the 8 bits data in memory bank
- CS <0:1> is chip select line to select all the chips in a Rank
- Data <0:63>: Data bus getting readout data from either of the Ranks

Organization of Multiple Ranks into DIMM



- Each Rank is having 8 chips
- Mounted back-to-back for form a DIMM

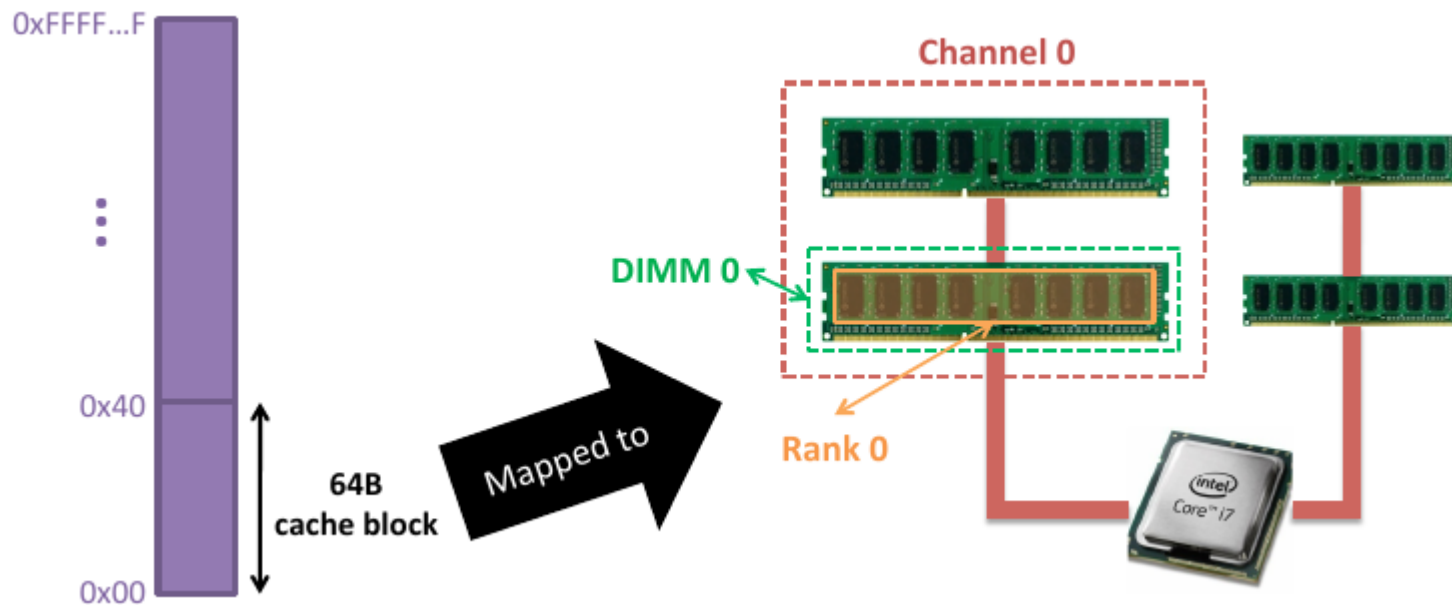
Channel: Organization of Multiple DIMMs



- Two DIMMs in a single channel
- Two channels connected to i7 core processor

Block Transfer: Address to Device Mapping

Physical Address Space

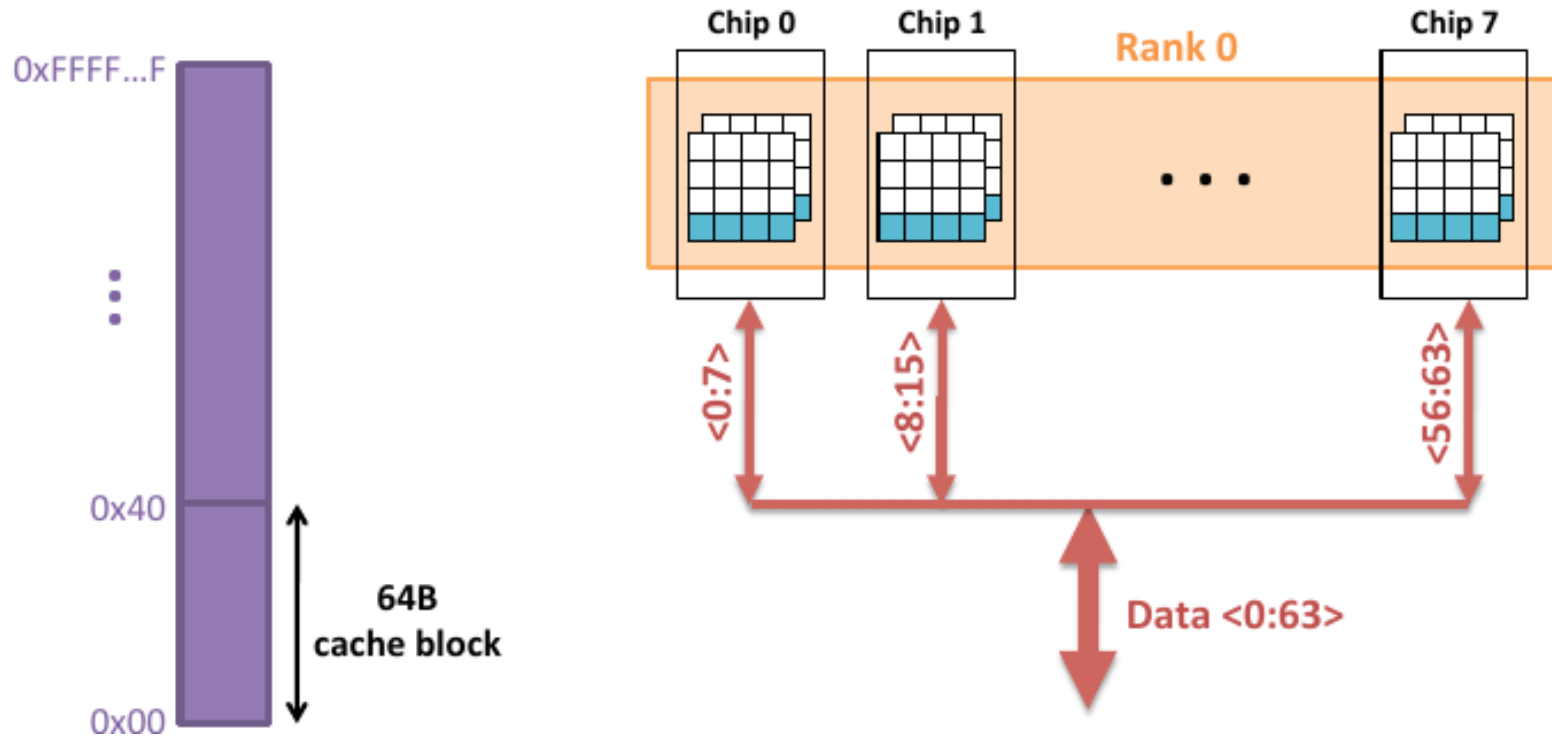


- Cache block size = 64 Byte
- Channel Bandwidth: 64 bits

Cache block is the same as the Cache block

Block Transfer: Address to Device Mapping

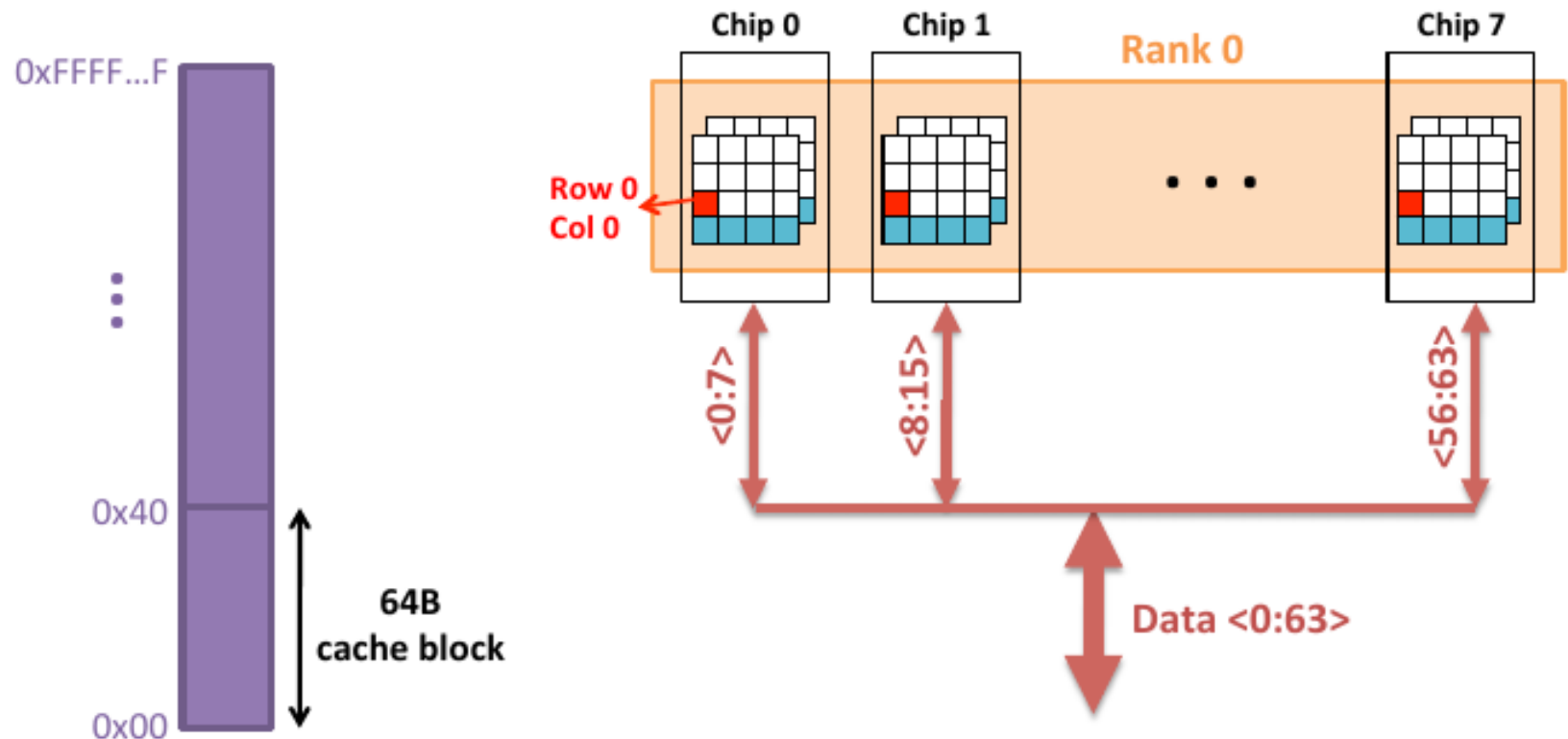
Physical Address Space



- Cache block size = 64 Byte
- Channel Bandwidth: 64 bits
- A Rank is having 8 DRAM Chips (Chip 0 to 7)

Block Transfer: Address to Device Mapping

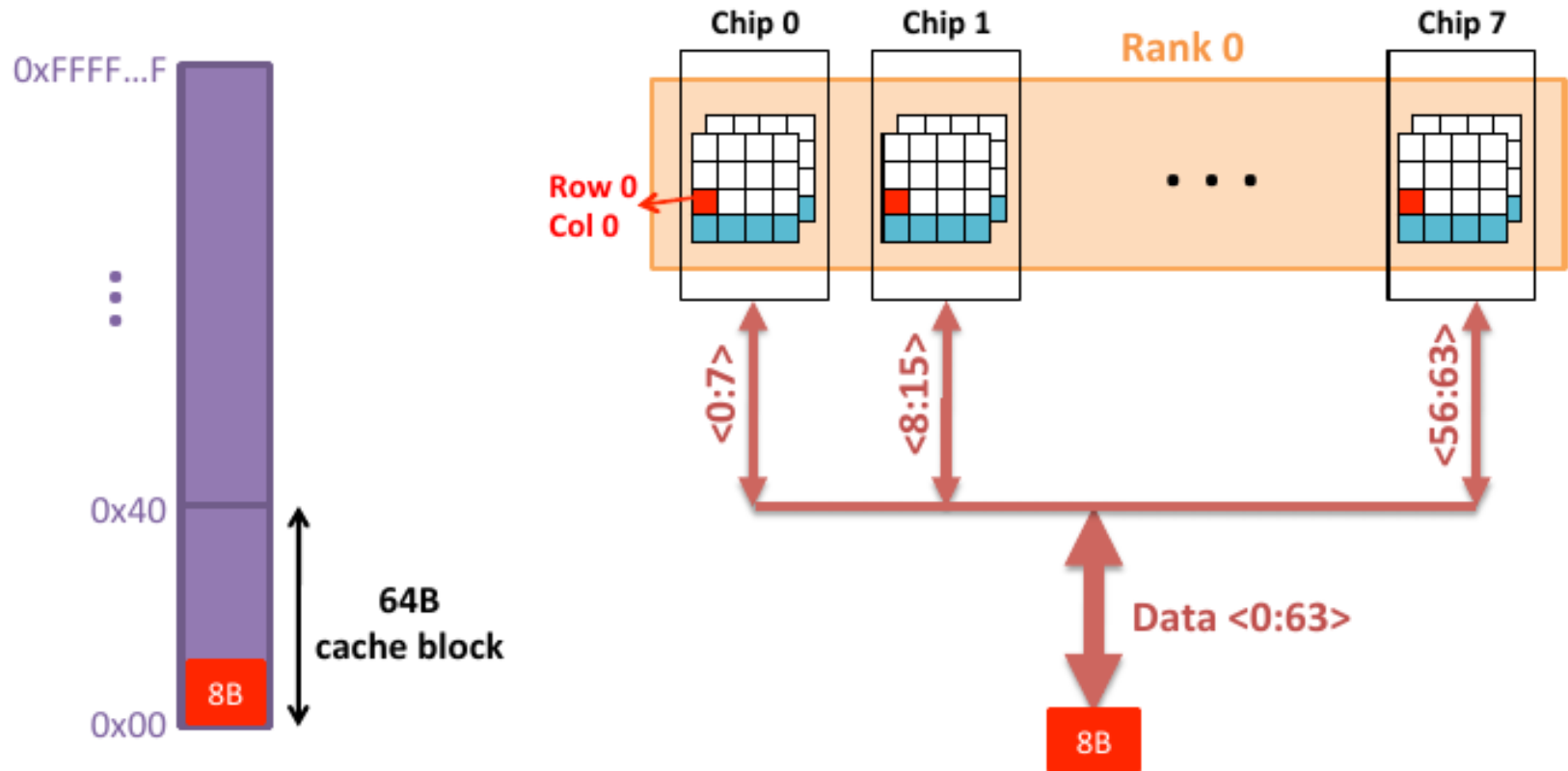
Physical Address Space



- Cache block size = 64 Byte
- Channel Bandwidth: 64 bits
- A Rank is having 8 DRAM Chips (Chip 0 to 7)
- **64 bits to be transferred in a memory cycle**

Block Transfer: Address to Device Mapping

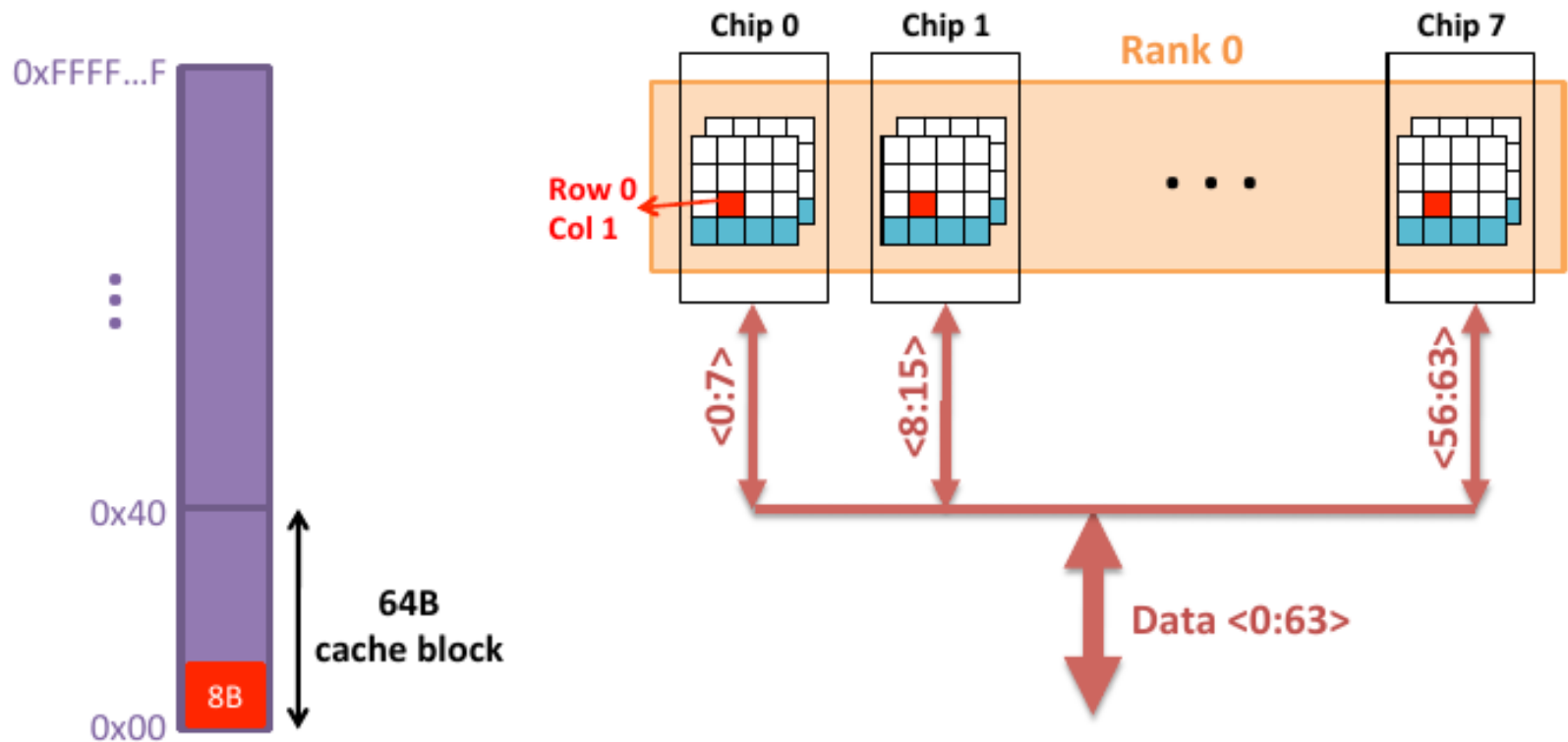
Physical Address Space



- Cache block size = 64 Byte
- Channel Bandwidth: 64 bits
- A Rank is having 8 DRAM Chips (Chip 0 to 7)
- **64 bits to be transferred in a memory cycle**

Block Transfer: Address to Device Mapping

Physical Address Space

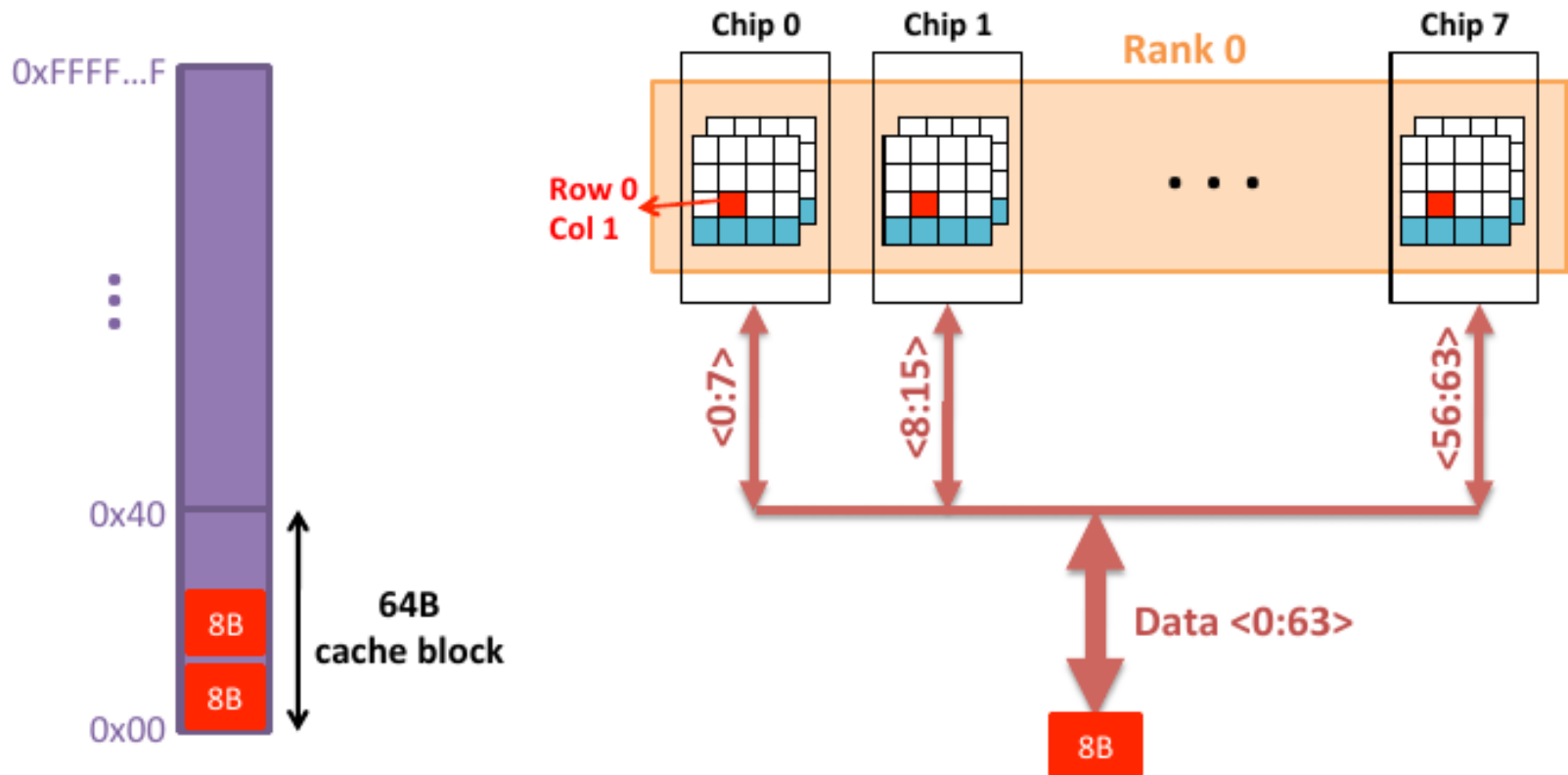


- Cache block size = 64 Byte
- Channel Bandwidth: 64 bits
- A Rank is having 8 DRAM Chips (Chip 0 to 7)
- **64 bits to be transferred in a memory cycle**

The next set of 8B will
Be transferred from **Col1**

Block Transfer: Address to Device Mapping

Physical Address Space

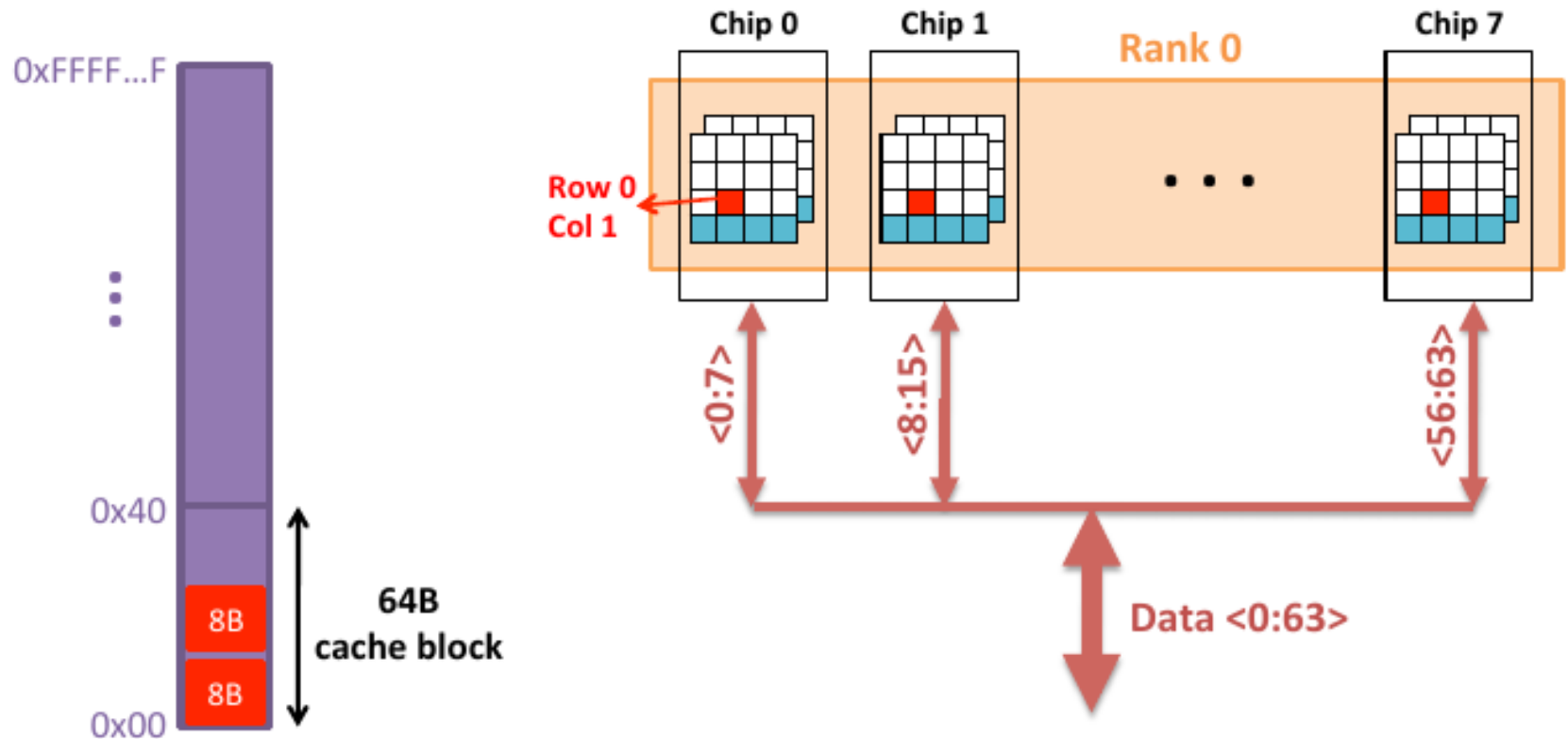


- Cache block size = 64 Byte
- Channel Bandwidth: 64 bits
- A Rank is having 8 DRAM Chips (Chip 0 to 7)
- **64 bits to be transferred in a memory cycle**

The next 8B is transferred

Block Transfer: Address to Device Mapping

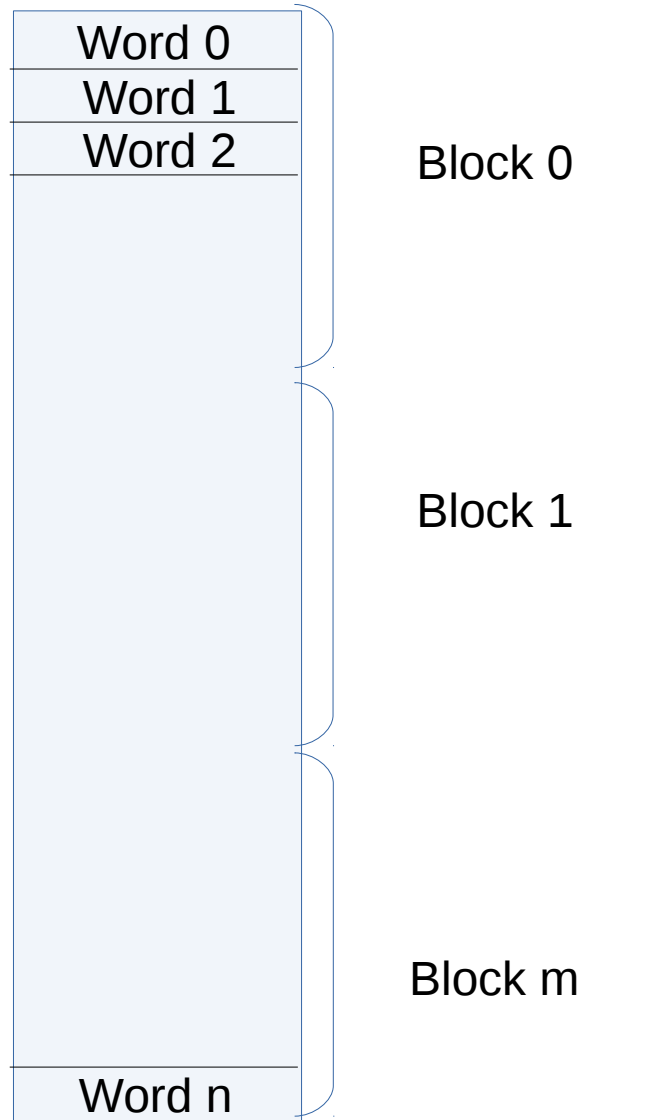
Physical Address Space



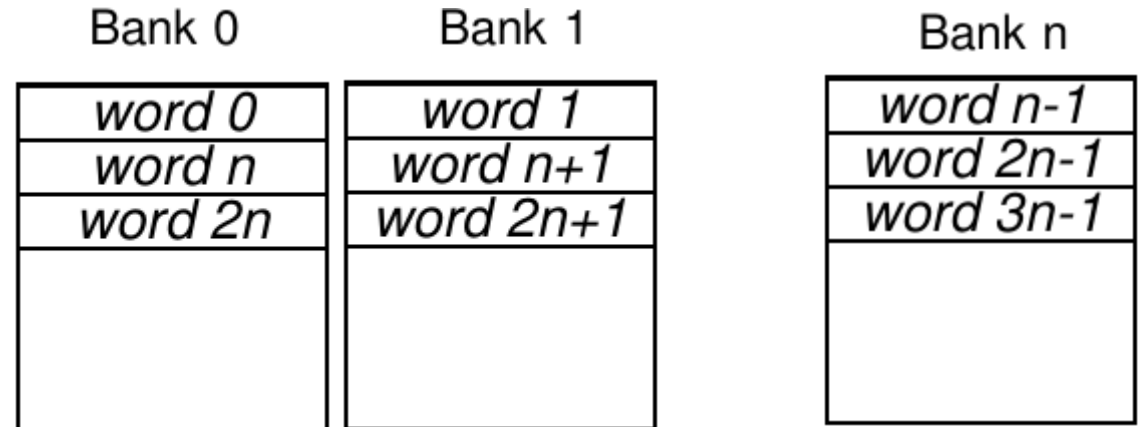
- To transfer 64B it takes 8 I/O cycles
- Each 8B at a single cycle

Concept of Interleaving

Without Bank



Interleaving with Bank



Addressing:

Word 0 is located in Bank $(0 \bmod n)$

Word n is located in Bank $(n \bmod n)$

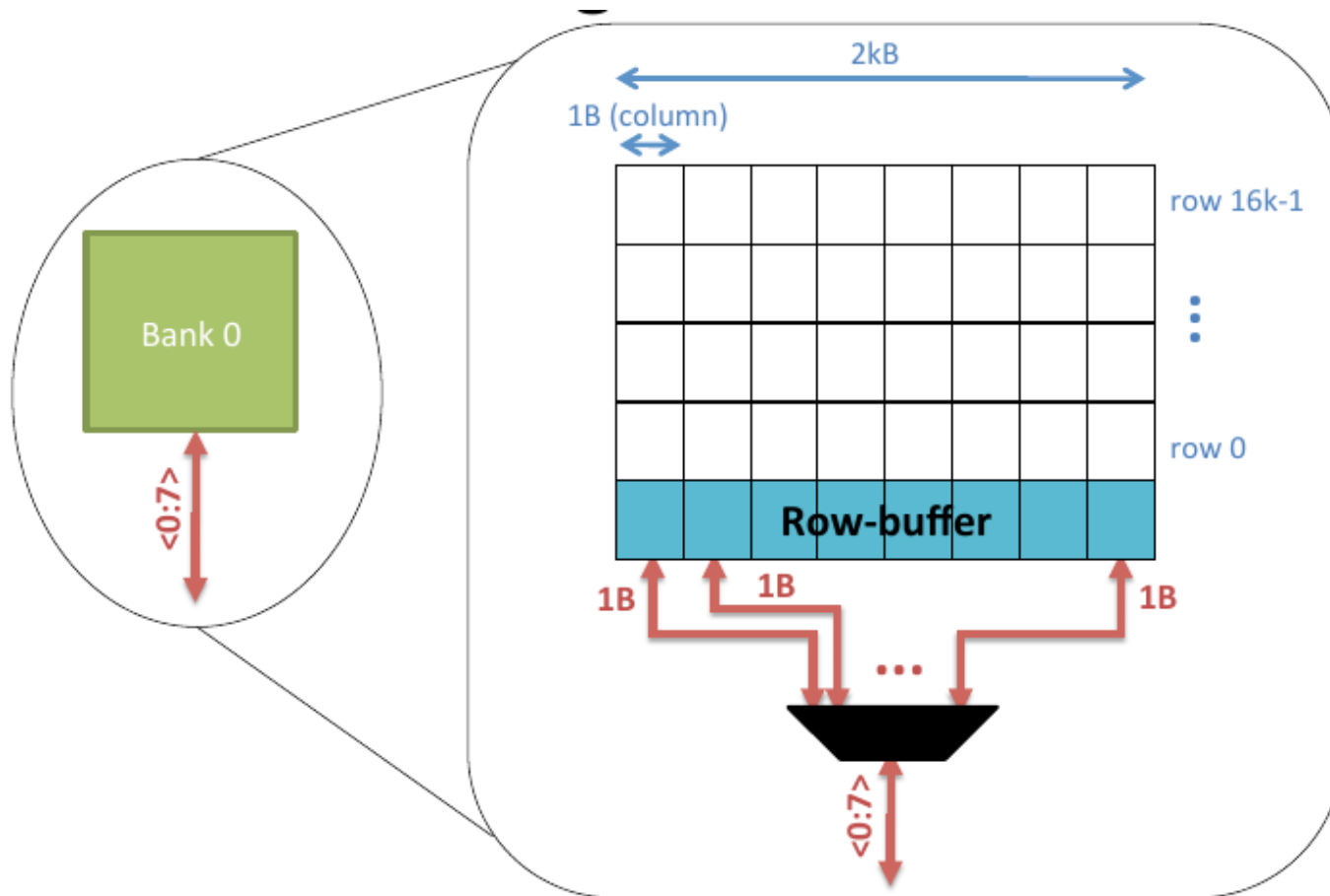
Word 0 in Bank 0 is located at $(0 \div n)$ place

Word n in Bank 0 is located at $(n \div n)$ place

Increase in Bandwidth:

- All the banks can be accessed in interleaved fashion
- Bank 1 need not wait for Bank 0 to finish

Addressing in Bank



Addressing in Bank

- A bank is addressed with **Row and Column number**
- A column reads out a 1B of data for byte addressable
- Whether 1B or 2B depends on design

Example: Row-major addressing

<Row 0, Column 0>

<Row 0, Column 1>

<Row 0, Column 2>

<Row 0, Column 3>

<Row 1, Column 0>

<Row 1, Column 1>

<Row 1, Column 2>

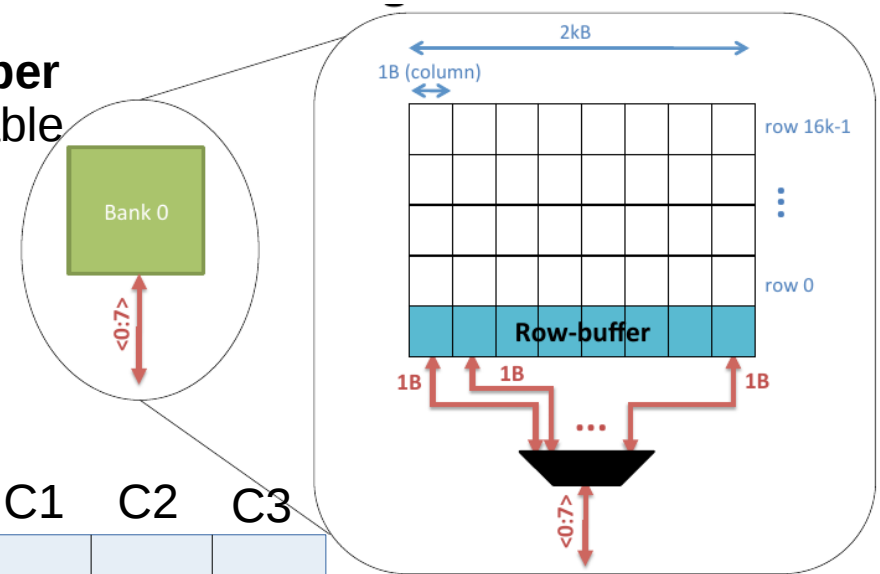
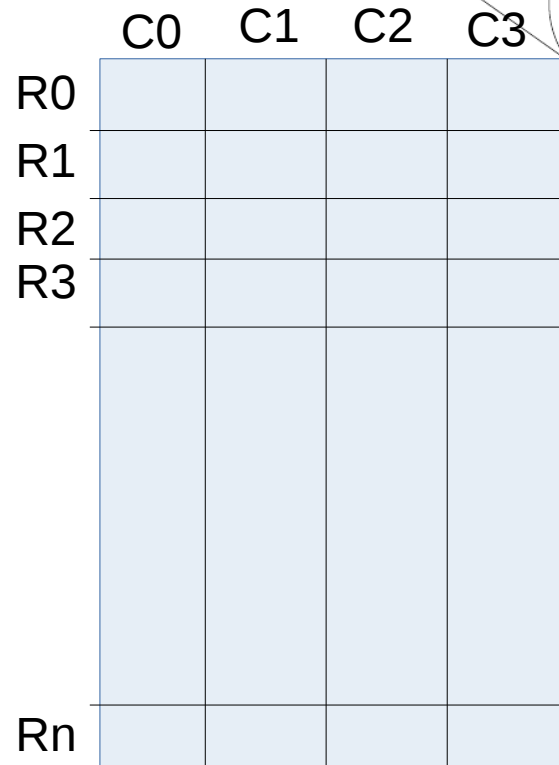
<Row 1, Column 3>

<Row n, Column 0>

<Row n, Column 1>

<Row n, Column 2>

<Row n, Column 3>



Addressing in Bank

- A bank is addressed with **Row and Column number**
- A column reads out a 1B of data for byte addressable
- Whether 1B or 2B depends on design

Example: Column-major addressing

<Row 0, Column 0>

<Row 1, Column 0>

<Row 2, Column 0>

<Row 3, Column 0>

<Row 4, Column 0>

<Row 5, Column 0>

<Row 6, Column 0>

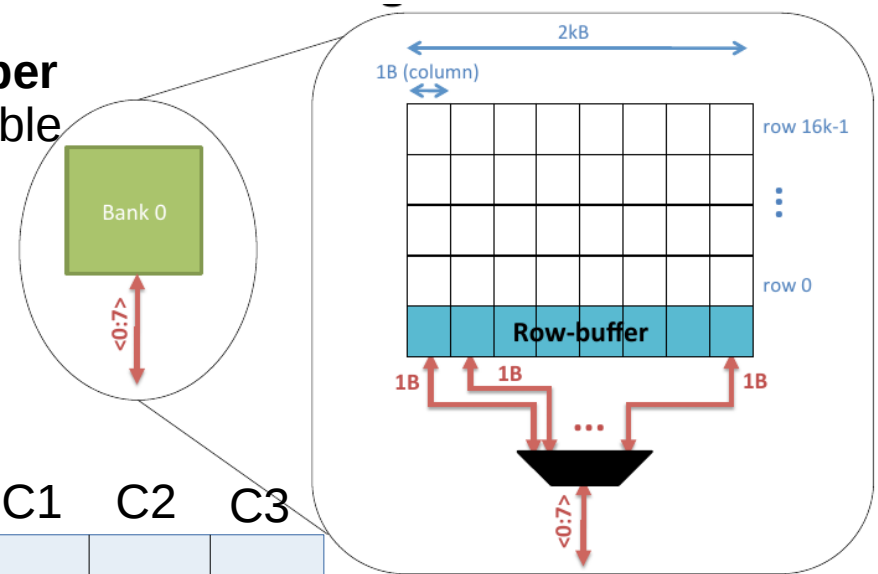
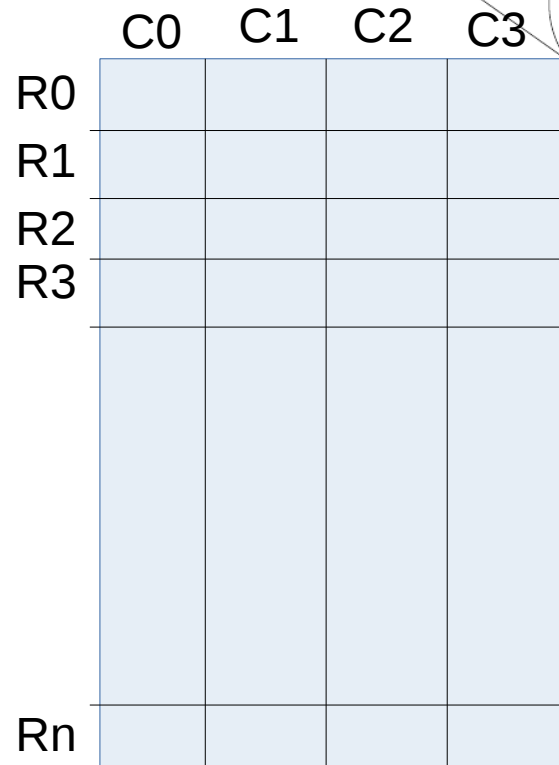
<Row 7, Column 0>

<Row n-3, Column 0>

<Row n-2, Column 0>

<Row n-1, Column 0>

<Row n, Column 0>



Address Mapping in Single Channel

Example: A Single channel system with 8 byte memory bus, 2GB memory, 8 banks, 16K row and 2K columns per bank

Row Interleaving:

Consecutive rows of memory in consecutive banks

16K = 2^{14} rows

8 banks = 2^3 banks

2K Col = 2^{11} columns

Byte addressable = 2^3 bits offset

Row: 14 bits

Bank: 3 bits

Column: 11 bits

Offset: 3 bits

Row 0 ---- Bank 0 ---- Col 0

Row 0 ---- Bank 0 ---- Col 1

.....

Row 0 ---- Bank 0 ---- Col 2047

Row 0 ---- Bank 1 ---- Col 0

Row 0 ---- Bank 1 ---- Col 1

.....

Row 0 ---- Bank 1 ---- Col 2047

Row 1 ---- Bank 0 ---- Col 0

Row 1 ---- Bank 0 ---- Col 1

.....

Row 1 ---- Bank 0 ---- Col 2047

Row 1 ---- Bank 1 ---- Col 0

Row 1 ---- Bank 1 ---- Col 1

.....

Row 1 ---- Bank 1 ---- Col 2047

Address Mapping in Single Channel

Example: A Single channel system with 8 byte memory bus, 2GB memory, 8 banks, 16K row and 2K columns per bank

Block Interleaving:

- Consecutive cache block addresses in consecutive banks.
- 64 byte block

16K = 2^{14} rows

8 banks = 2^3 banks

2K Col = 2^{11} columns

Byte addressable = 2^3 bits offset

Row: 14 bits	High Col: 8 bits	Bank: 3 bits	Low Col: 3 bits	Offset: 3 bits
Row 0	---- Bank 0	---- Col 0	Row 0	---- Bank 0 ---- Col 8
Row 0	---- Bank 0	---- Col 1	Row 0	---- Bank 0 ---- Col 9
.....				
Row 0	---- Bank 0	---- Col 7	Row 0	---- Bank 0 ---- Col 15
Row 0	---- Bank 1	---- Col 0	Row 0	---- Bank 1 ---- Col 8
Row 0	---- Bank 1	---- Col 1	Row 0	---- Bank 1 ---- Col 9
.....			
Row 0	---- Bank 1	---- Col 7	Row 0	---- Bank 1 ---- Col 15

Address Mapping in Multi Channel

Example: A Multi-channel system with 8 byte memory bus, 2GB memory, 8 banks, 16K row and 2K columns per bank

Row Interleaving:

Consecutive rows of memory in consecutive banks

16K = 2^{14} rows

8 banks = 2^3 banks

2K Col = 2^{11} columns

Byte addressable = 2^3 bits offset

Chanel: 1 bit	Row: 14 bits	Bank: 3 bits	Column: 11 bits	Offset: 3 bits
Row: 14 bits	Chanel: 1 bit	Bank: 3 bits	Column: 11 bits	Offset: 3 bits
Row: 14 bits	Bank: 3 bits	Chanel: 1 bit	Column: 11 bits	Offset: 3 bits
Row: 14 bits	Bank: 3 bits	Column: 11 bits	Chanel: 1 bit	Offset: 3 bits

Address Mapping in Multi Channel

Example: A Multi-channel system with 8 byte memory bus, 2GB memory, 8 banks, 16K row and 2K columns per bank

Block Interleaving:

- Consecutive cache block addresses in consecutive banks.
- 64 byte block

16K = 2^{14} rows

8 banks = 2^3 banks

2K Col = 2^{11} columns

Byte addressable = 2^3 bits offset

Ch: 1bit	Row: 14 bits	High Col: 8 bits	Bank: 3 bits	Low Col: 3 bits	Offset: 3 bits
----------	--------------	------------------	--------------	-----------------	----------------

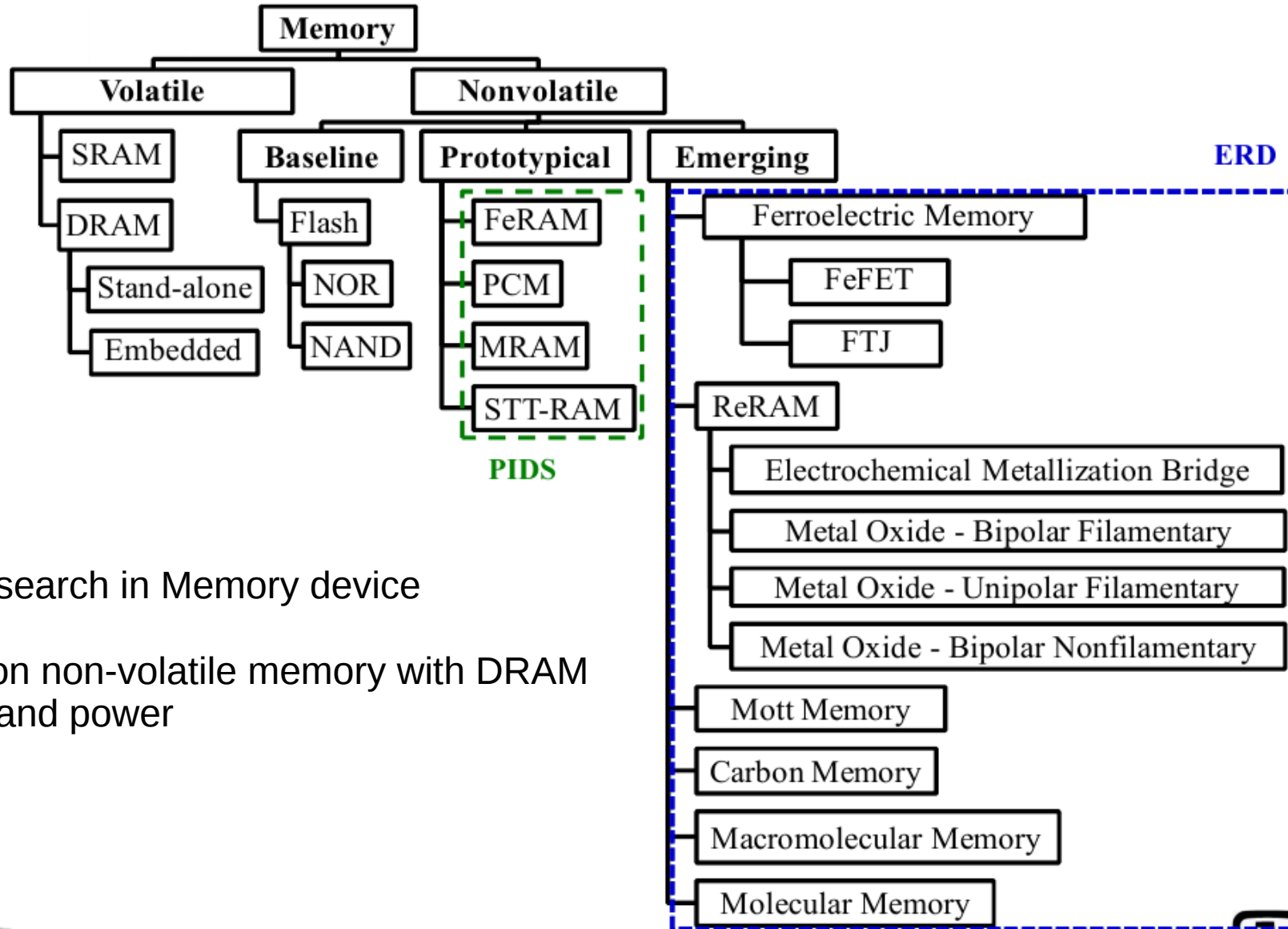
Row: 14 bits	Ch: 1bit	High Col: 8 bits	Bank: 3 bits	Low Col: 3 bits	Offset: 3 bits
--------------	----------	------------------	--------------	-----------------	----------------

Reference:

Chapter 13, Memory Systems: Cache, DRAM, Dsik; Bruce Jacob et al

What Next in Main Memory

Technology: [ITRS 2015 Report]



- Emerging Research in Memory device
- The focus is on non-volatile memory with DRAM performance and power

What Next in Main Memory

Architectural Ideas:

- In-memory computation:
 - Bring computation to memory instead of taking data to computation
- Think of bringing an additional hierarchy with non-volatile memory
 - Adaptive or application specific memory hierarchy
- Increase on-chip memory

A nice trend road-map for next research in memory:

<https://ece.umd.edu/~blj/talks/Sun-Workshop.pdf>

Reading Materials

- Bruce Jacob, Spencer Ng, and David Wang; ***Memory Systems: Cache, DRAM, Disk***; 2008, Elsevier. (Refer: Chapter 10 and Chapter 8)
- Onur Mutlu, Scalable Memory System Lectures, Lecture 1, Lecture 2 and Lecture 3;
<http://users.ece.cmu.edu/~omutlu/acaces2013-memory.html>
- Text Book (H&P): Memory Technology and Optimization

Thank You

Additional Pages
