CMPE 110: Computer Architecture

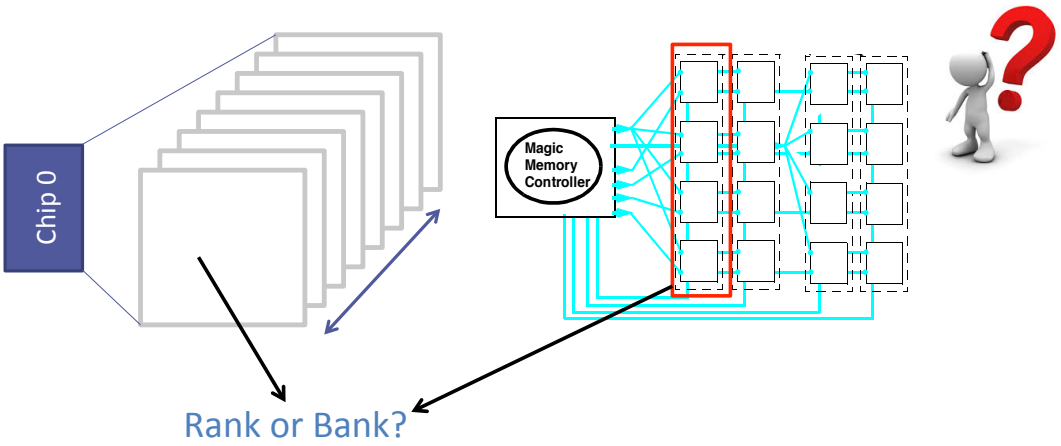
Week 9

Main Memory

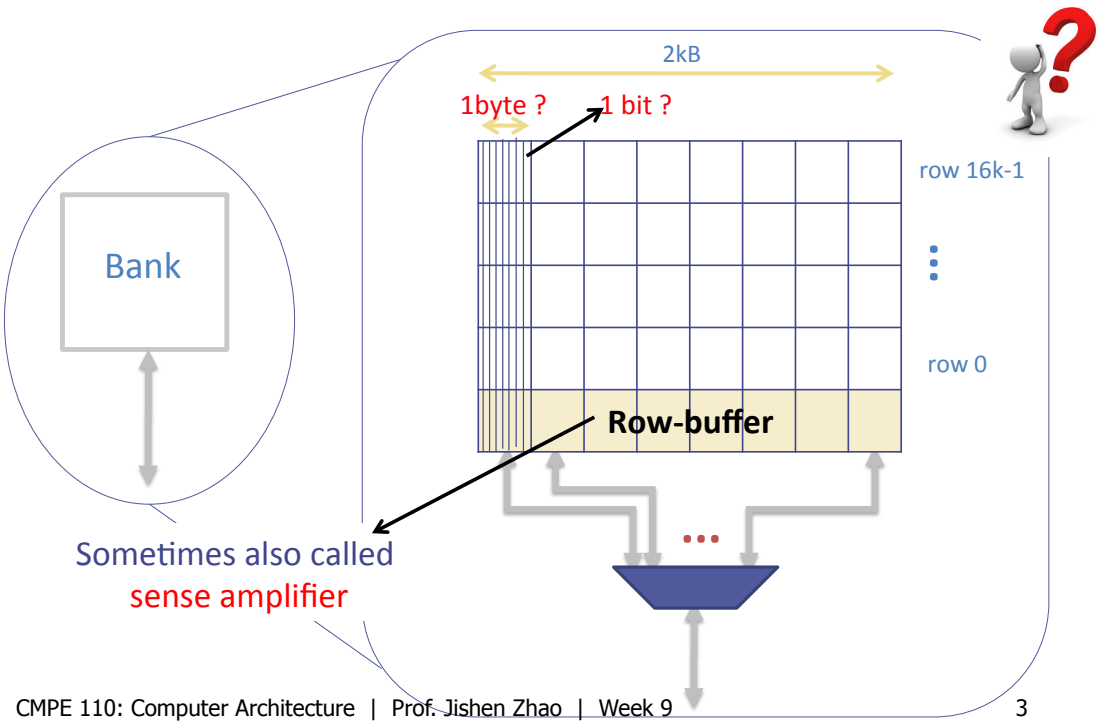
Jishen Zhao (http://users.soe.ucsc.edu/~jzhao/)

[Adapted in part from Jose Renau, Mary Jane Irwin, Joe Devietti, Onur Mutlu, and others]

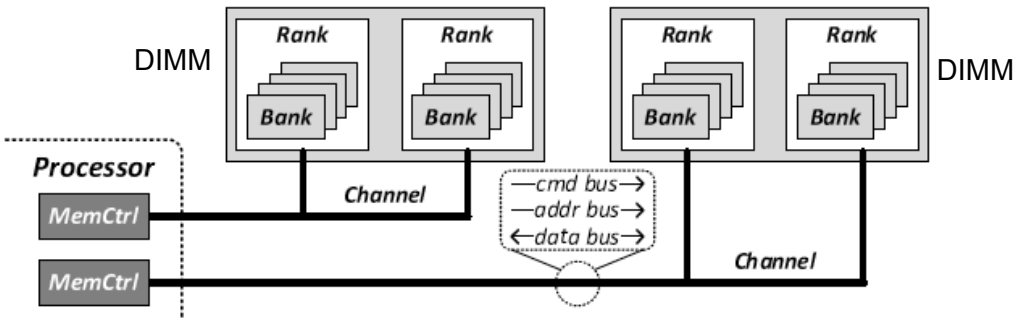
Review: Test yourself -- Rank and bank



Review: Test yourself -- row and column



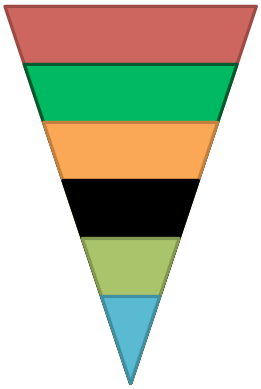
Review: Generalized Memory Structure



Source: Kim+, "A Case for Exploiting Subarray-Level Parallelism in DRAM," ISCA 2012.

Review: main memory organization

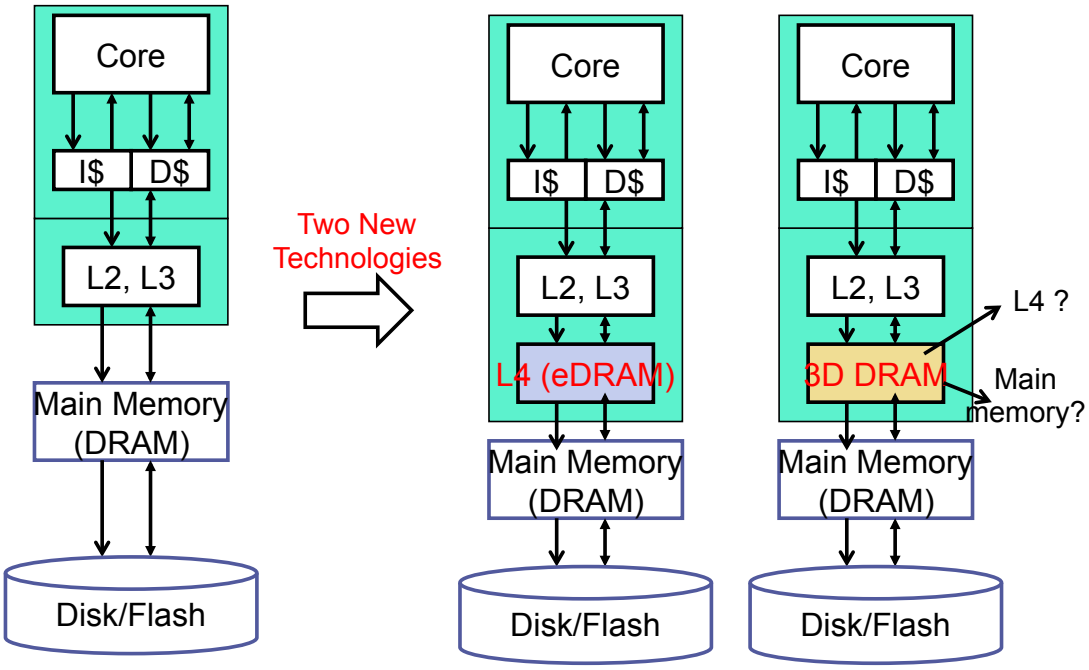
- Channel
- DIMM
- Rank
- Chip
- Bank
- Row/Column



Today

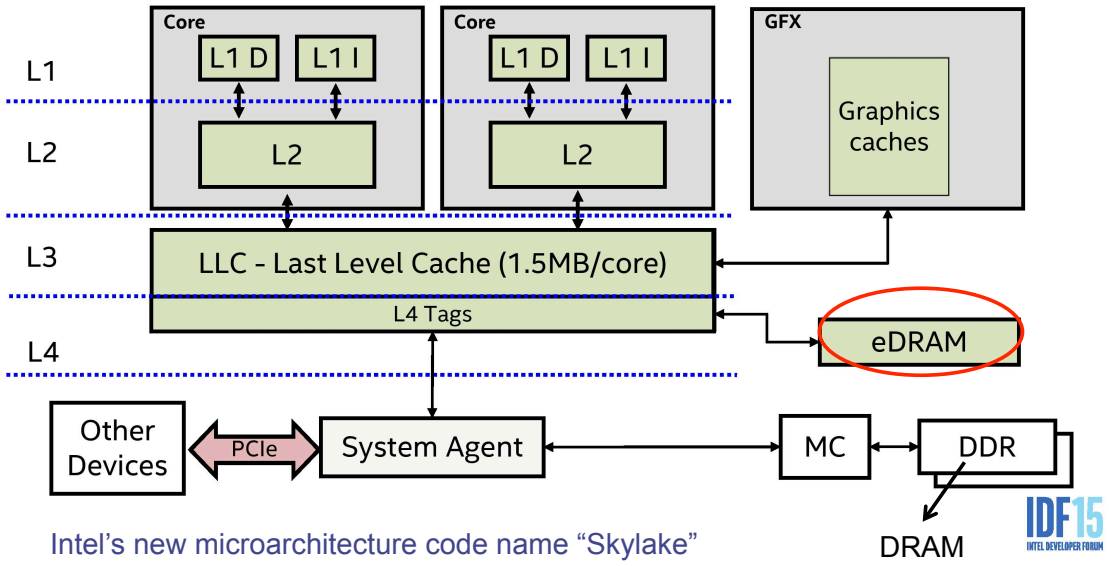
- On-chip memory
  - Memory can be integrated in processor chip
- Where is data?
  - How to access main memory with address

Integrate DRAM on the processor chip



eDRAM (embedded DRAM)

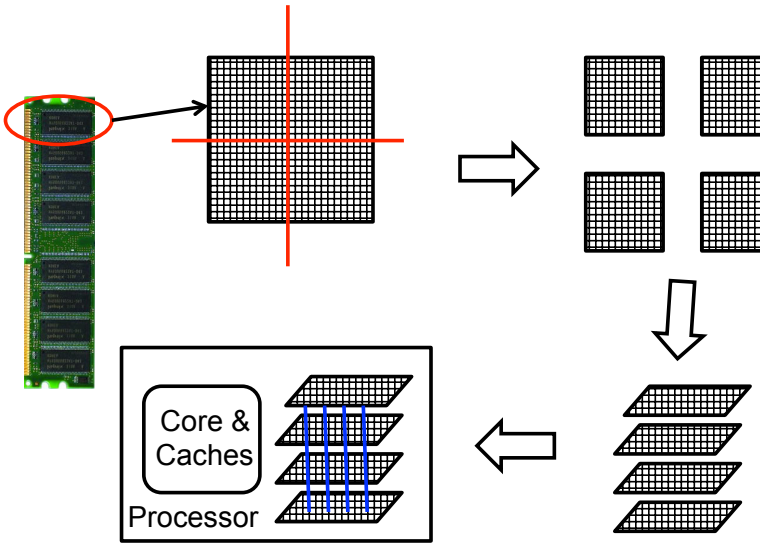
eDRAM Based Cache



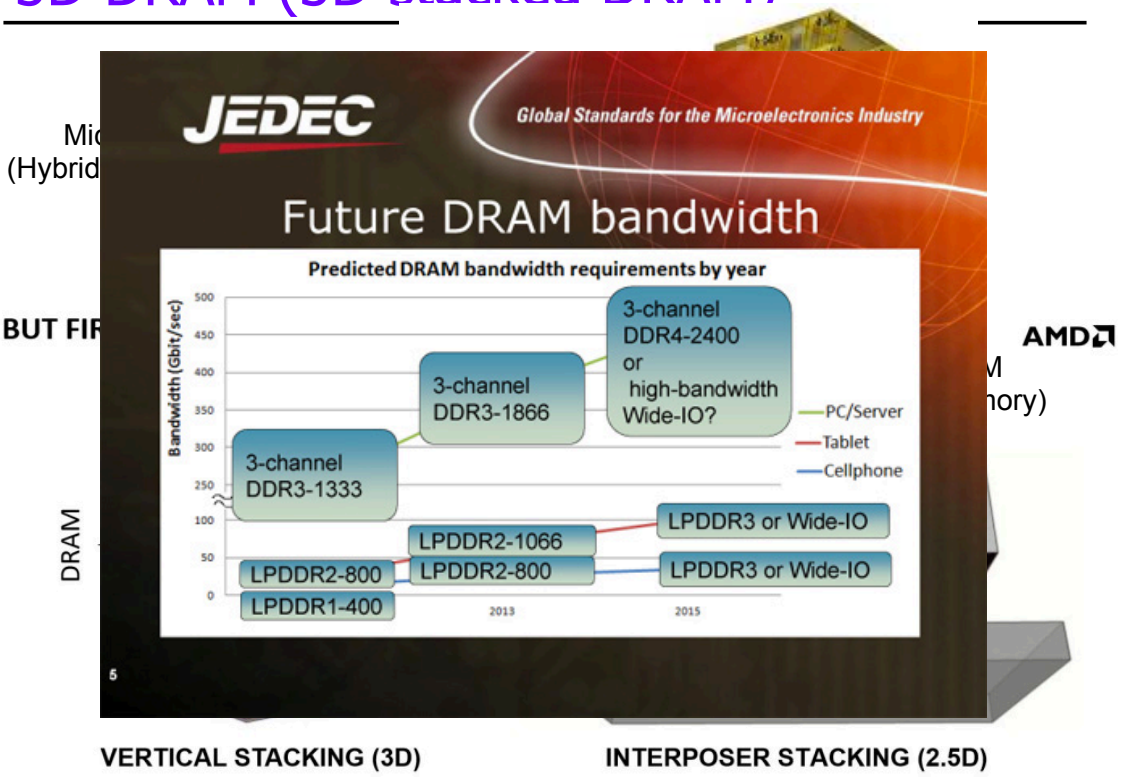
26 Intel's new microarchitecture code name "Skylake"

From 2D to 3D architectures

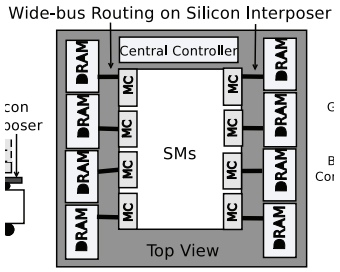
3D DRAM (3D stacked DRAM)



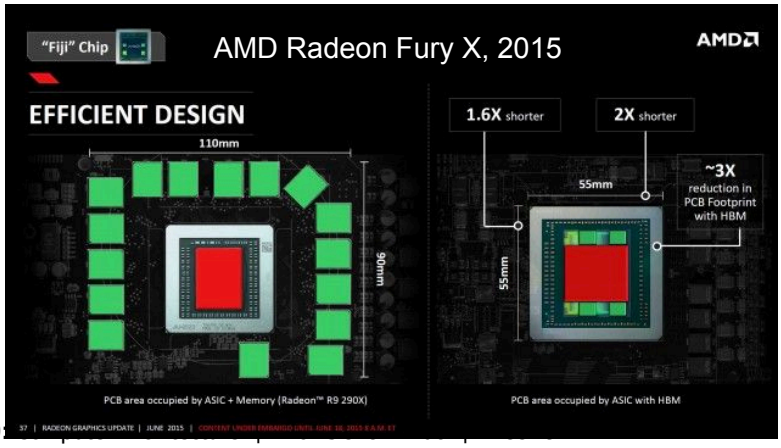
3D DRAM (3D stacked DRAM)



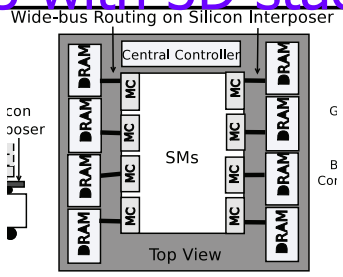
From research to real product: GPU with 3D stacked DRAM

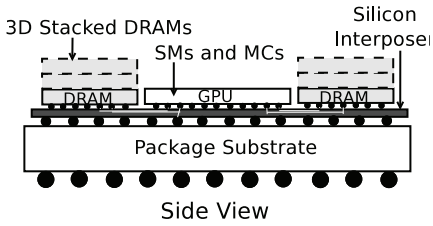


Zhao et al., "Energy-efficient GPU Design with Reconfigurable In-package Graphics Memory," ISLPED'12

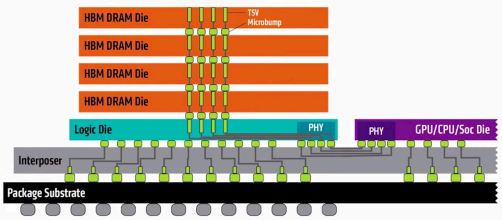


From research to real product: GPU with 3D stacked DRAM



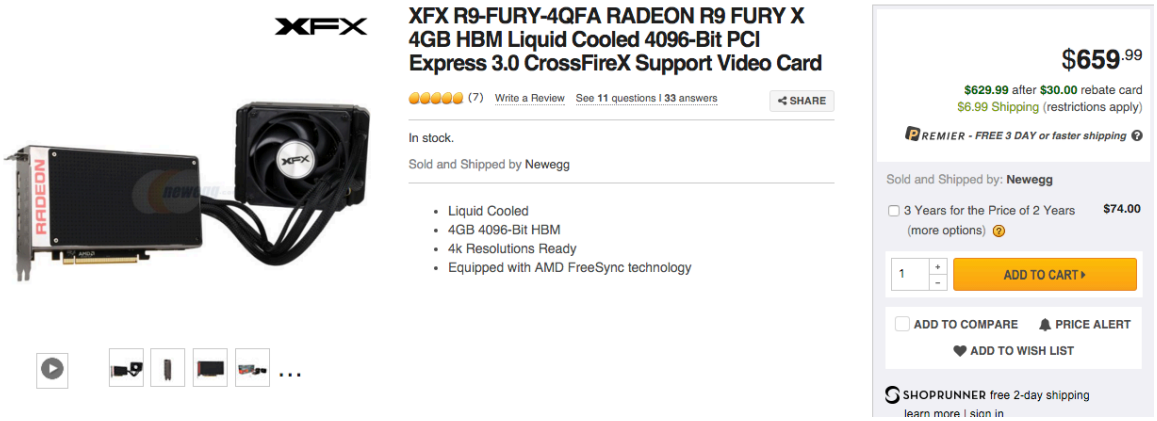


Zhao et al., "Energy-efficient GPU Design with Reconfigurable In-package Graphics Memory," ISLPED'12

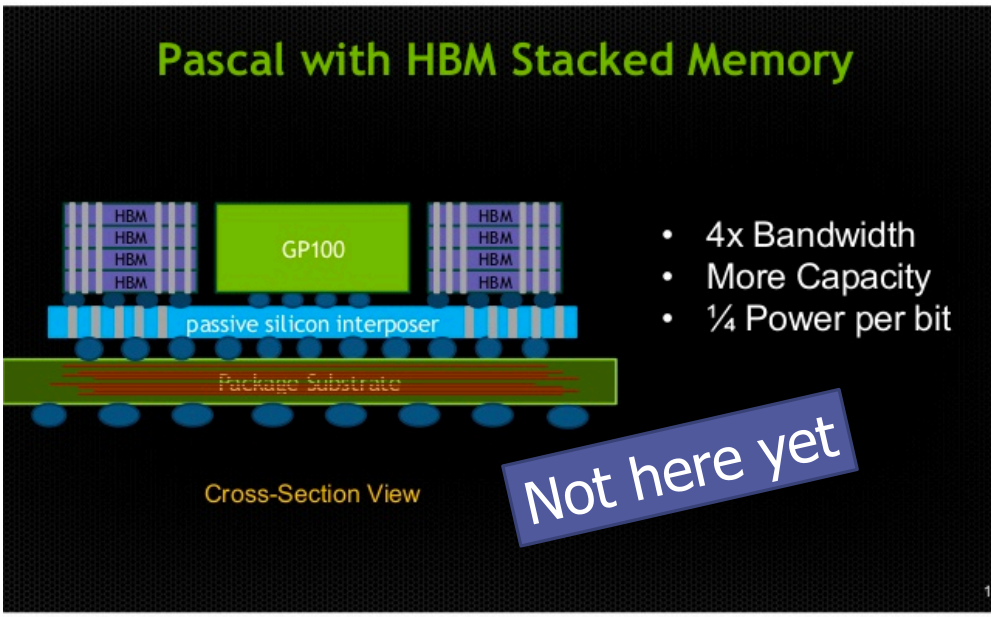


AMD Radeon Fury X, 2015

You can buy it



GPU with 3D stacked DRAM

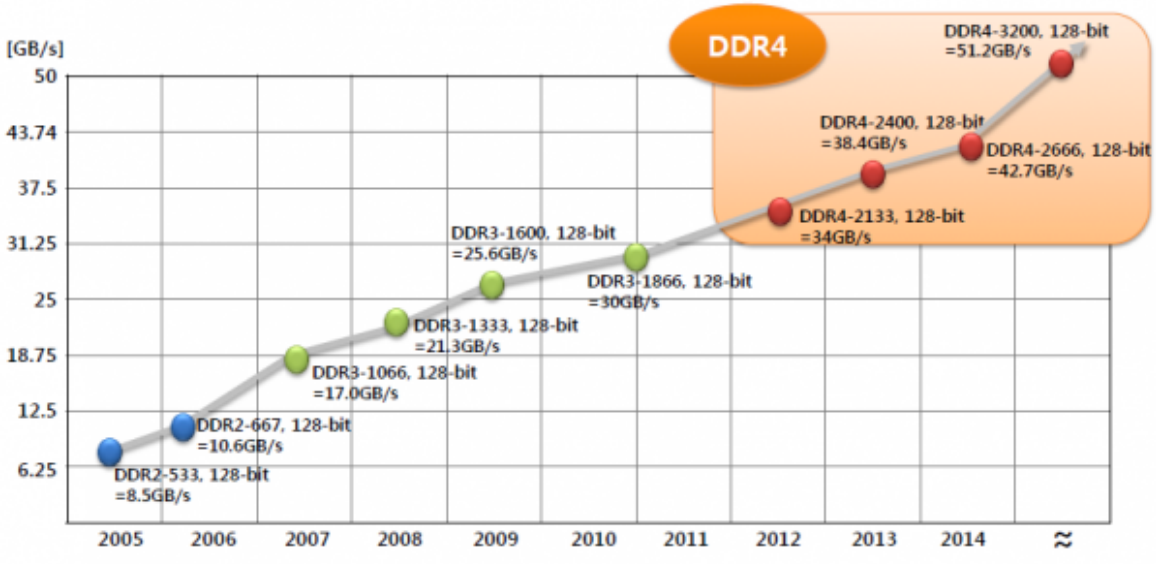


CPU system

Why 3D stacked DRAM?

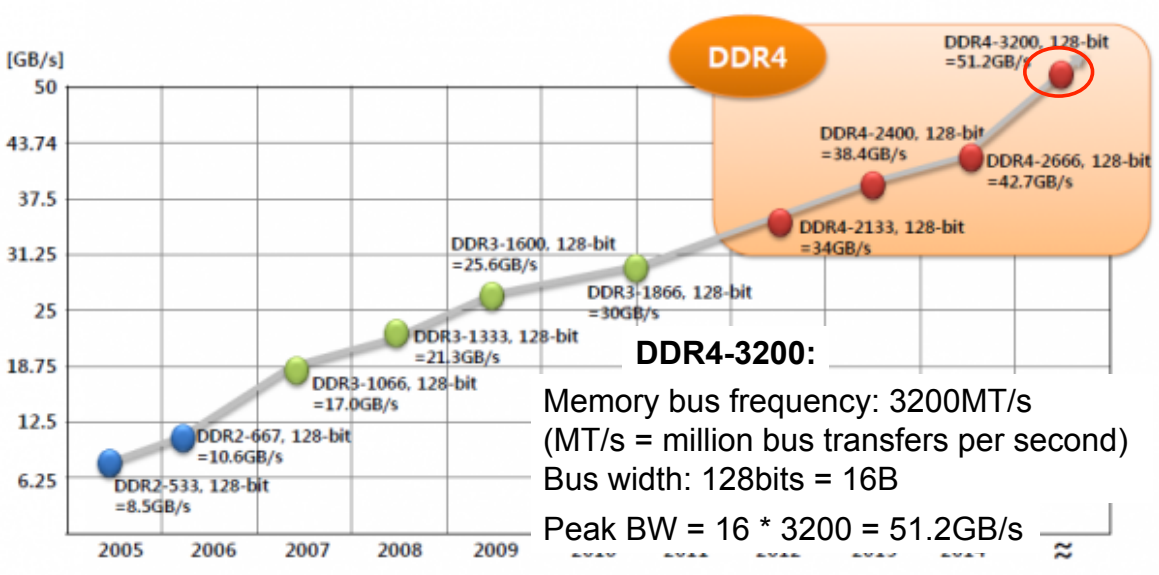
- One important reason:
  - Much higher "memory bandwidth" than traditional DRAM
- Memory performance: Latency -- we already know
- What about memory throughput? → called "memory bandwidth"
  - How many bytes main memory can send per second?
  - Unit: GB/s or MB/s
  - Read bandwidth & write bandwidth
- "Peak memory bandwidth"
  - The maximum bandwidth a main memory device can achieve
  - This is what vendors like to advertise about

Peak memory bandwidth example



Peak memory bandwidth calculation

Peak memory bandwidth = memory bus width * memory bus frequency



DDR4-3200:

Memory bus frequency: 3200MT/s
(MT/s = million bus transfers per second)
Bus width: 128bits = 16B

Peak BW = 16 * 3200 = 51.2GB/s

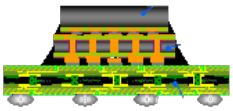
How about 3D stacked DRAM peak BW?

- May have the same or even lower memory clock frequency
- But much wider bus

	Radeon R9 390X	GeForce GTX 780 Ti	GeForce GTX 980	**Radeon Fury X**	GeForce GTX 980 Ti	GeForce GTX Titan X	Radeon R9 295X2
Shader Units	2816	2880	2048	**4096**	2816	3072	2x 2816
ROPs	64	48	64	**64**	96	96	2x 64
Graphics Processor	Hawaii	GK110	GM204	**Fiji**	GM200	GM200	2x Hawaii
Transistors	6200M	7100M	5200M	**8900M**	8000M	8000M	2x 6200M
Memory Size	8192 MB	3072 MB	4096 MB	**4096 MB**	6144 MB	12288 MB	2x 4096 MB
Memory Bus Width	512 bit	384 bit	256 bit	**4096 bit**	384 bit	384 bit	2x 512 bit
Core Clock	1050 MHz	876 MHz+	1126 MHz+	**1050 MHz**	1000 MHz+	1000 MHz+	1018 MHz
Memory Clock	1500 MHz	1750 MHz	1750 MHz	**500 MHz**	1750 MHz	1750 MHz	1250 MHz
Price	$430	$390	$480	**$650**	$650	$1050	$620

Radeon Fury X Market Segment Analysis

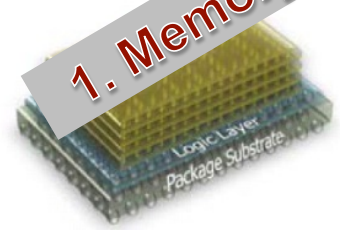
3D stacking – other than memory

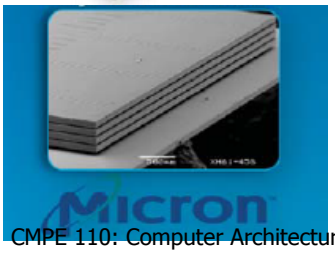


3D Die Stacking
(Integrated Ver…)

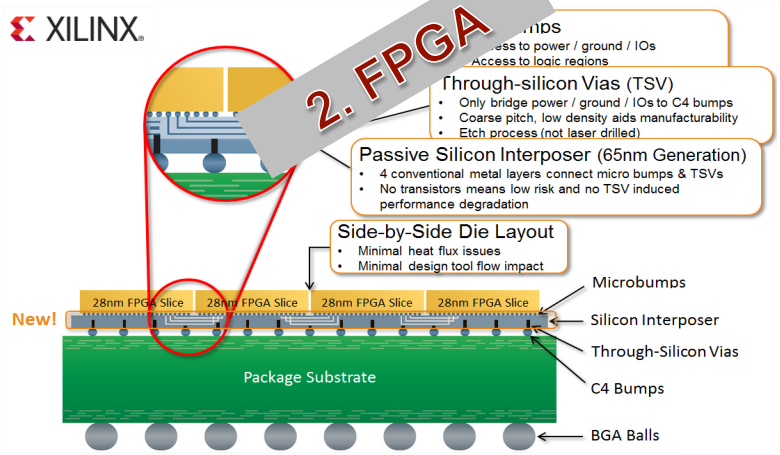
2.5D Integration (Si Interposer)
(Integrated side-by-side)

1. Memory





2. FPGA



MORGAN & CLAYPOOL PUBLISHERS

Hello. Sign in to personalize your visit. New user? Register now.

Home | Synthesis | Colloquium | Search | Profile | Author | Help | About

Quick search: Enter keyword within: This issue go

Home > Series home > Abstract

Die-stacking Architecture

Synthesis Lectures on Computer Architecture

June 2015, 127 pages, (doi:10.2200/S00644ED1V01Y201505CAC031)

Yuan Xie
University of California, Santa Barbara

Jishen Zhao
University of California, Santa Cruz

Abstract

The emerging three-dimensional (3D) chip architectures, with their intrinsic capability of reducing the wire length, promise attractive solutions to reduce the delay of interconnects in future microprocessors. 3D memory stacking enables much higher memory bandwidth for future chip-multiprocessor design, mitigating the "memory wall" problem. In addition, heterogenous integration enabled by 3D technology can also result in innovative designs for future microprocessors. This book first provides a brief introduction to this emerging technology, and then presents a variety of approaches to designing future 3D microprocessor systems, by leveraging the benefits of low latency, high bandwidth, and heterogeneous integration capability which are offered by 3D technology.

Table of Contents: Preface / Acknowledgments / 3D Integration Technology / Benefits of 3D Integration / Fine-granularity 3D Processor Design / Coarse-granularity 3D Processor Design / 3D GPU Architecture / 3D Network-on-Chip / Thermal Analysis and Thermal-aware Design / Cost Analysis for 3D ICs / Conclusion / Bibliography

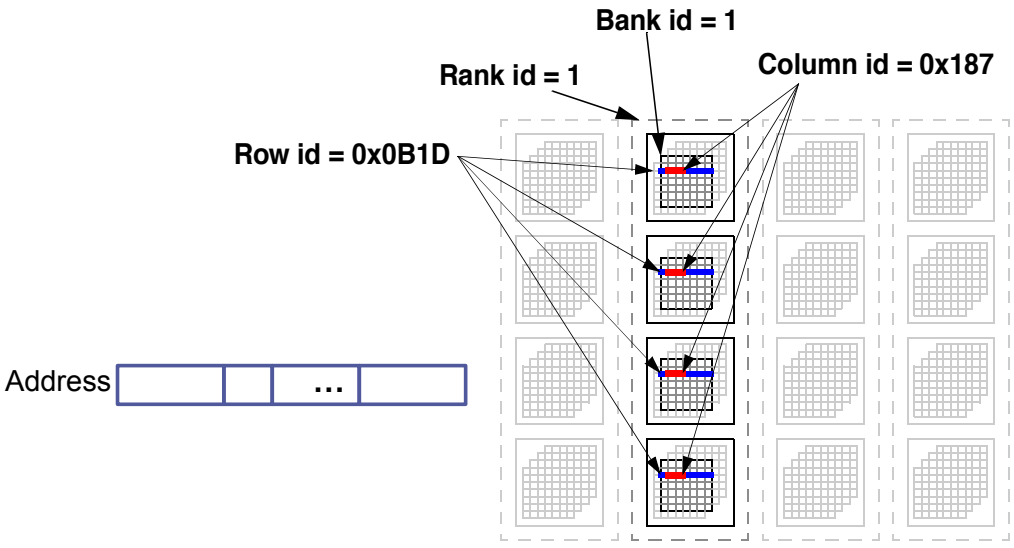
Computer Architecture

Prev. lecture | Next lecture
View/Print PDF (3463 KB)
View PDF Plus (3377 KB)
Add to favorites
Email to a friend
XML TOC Alert | Citation Alert What is RSS?

Quick Links
- Alert me when: New articles cite this article
- Download to citation manager

Today

- On-chip memory
  - Memory can be integrated in processor chip
- Where is data?
  - How to access main memory with address



Where is the data

Page Mode DRAM

- [Known] A DRAM bank is a 2D array of cells:
  - rows x columns
- A "DRAM row" is also called a "DRAM page"
- [Known] "Sense amplifiers" also called "row buffer"

- **A closed row vs. an open row**
  (analogy: write-through vs. write-back cache)
  - A closed row: data are not buffered in the row buffer after each access (similar to "write-through")
  - An open row: data are buffered in the row buffer after access (similar to "write-back")
  - Can only have one open row at a time ← only one row buffer in a bank

Open row vs. closed row

- Access to an "open row" – data in the row buffer
  - Read/write command reads/writes column in the row buffer
- Access to a "closed row" – data not in the row buffer
  - Activate command opens row (placed into row buffer)
  - Read/write command reads/writes column in the row buffer
  - Precharge command closes the row and prepares the bank for next access