



## GPU Teaching Kit

Accelerated Computing



# Module 6.1 – Memory Access Performance

DRAM Bandwidth

# Objective

- To learn that memory bandwidth is a first-order performance factor in a massively parallel processor
  - DRAM bursts, banks, and channels
  - All concepts are also applicable to modern multicore processors

# Global Memory (DRAM) Bandwidth

— Ideal

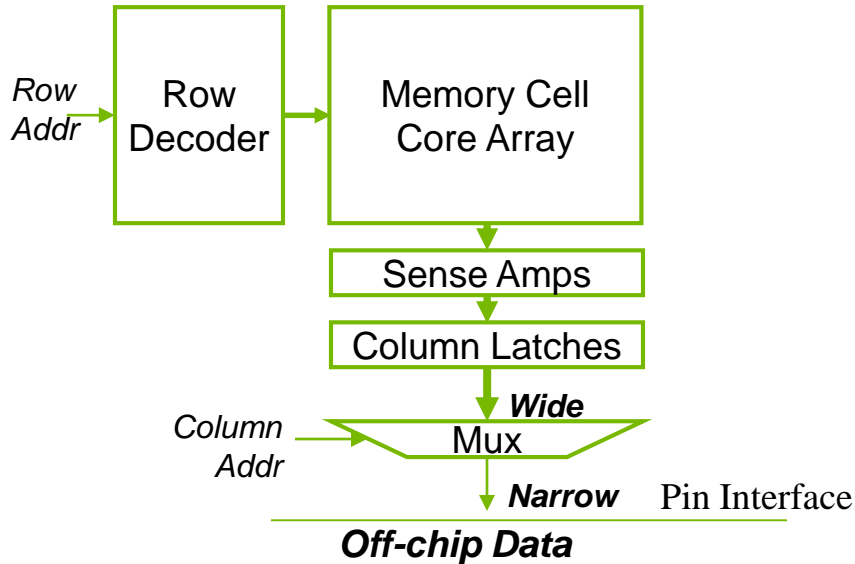


— Reality

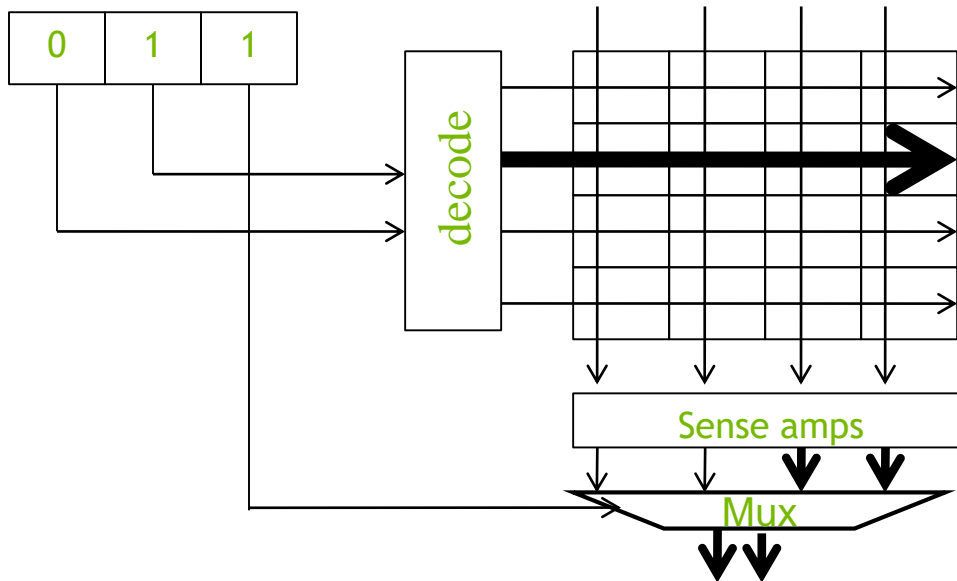


# DRAM Core Array Organization

- Each DRAM core array has about 16M bits
- Each bit is stored in a tiny capacitor made of one transistor

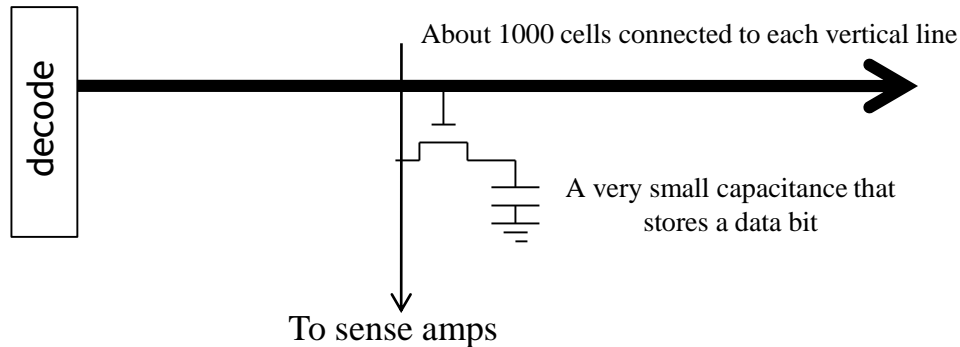


# A very small (8x2-bit) DRAM Core Array



# DRAM Core Arrays are Slow

- Reading from a cell in the core array is a very slow process
  - DDR: Core speed =  $\frac{1}{2}$  interface speed
  - DDR2/GDDR3: Core speed =  $\frac{1}{4}$  interface speed
  - DDR3/GDDR4: Core speed =  $\frac{1}{8}$  interface speed
  - ... likely to be worse in the future

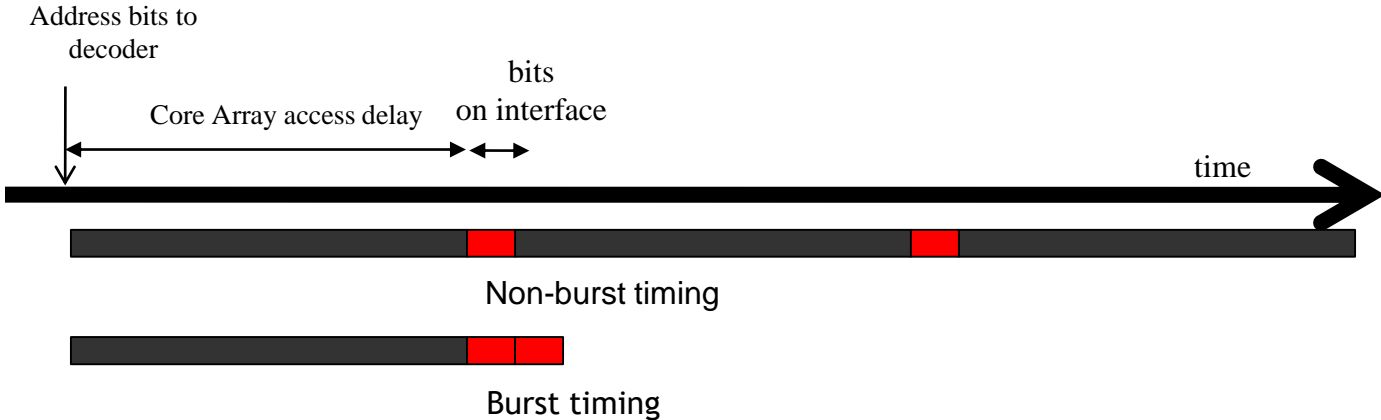


# DRAM Bursting

- For DDR{2,3} SDRAM cores clocked at  $1/N$  speed of the interface:
  - Load ( $N \times$  interface width) of DRAM bits from the same row at once to an internal buffer, then transfer in  $N$  steps at interface speed
  - DDR3/GDDR4: buffer width =  $8X$  interface width



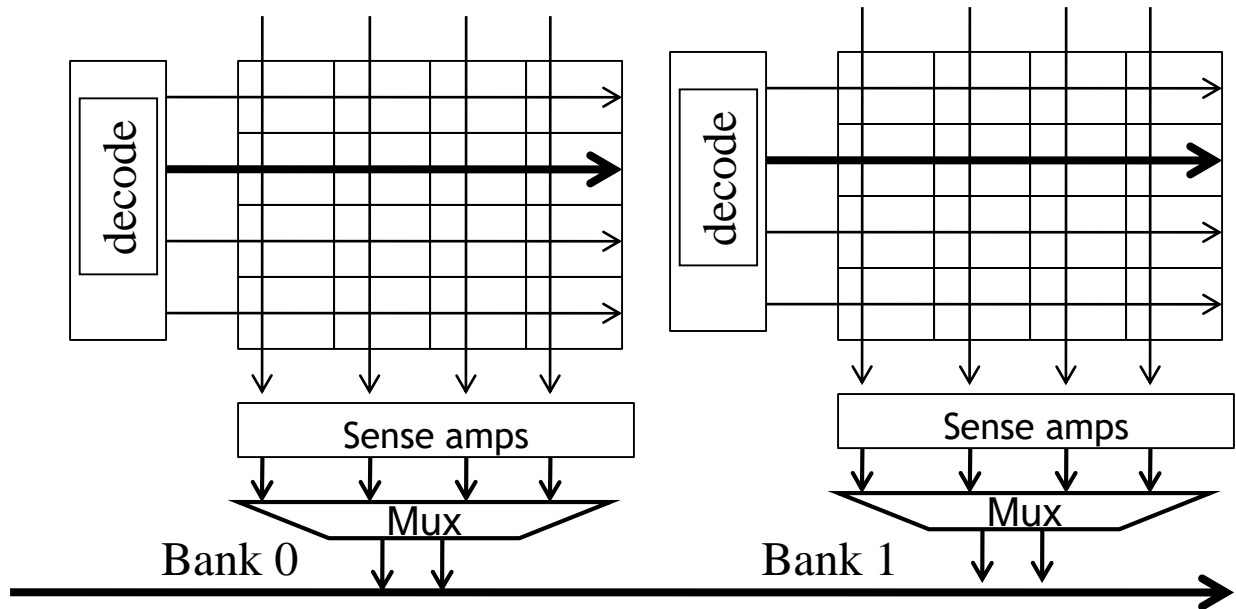
# DRAM Bursting Timing Example



Modern DRAM systems are designed to always be accessed in burst mode. Burst bytes are transferred to the processor but discarded when accesses are not to sequential locations.



# Multiple DRAM Banks



# DRAM Bursting with Banking



Single-Bank burst timing, dead time on interface



Multi-Bank burst timing, reduced dead time

# GPU off-chip memory subsystem

- NVIDIA RTX6000 GPU:
  - Peak global memory bandwidth = 672GB/s
- Global memory (GDDR6) interface @ 7GHz
  - 14 Gbps pin speed
  - For GDDR6 32-bit interface, we can sustain only about 56 GB/s
  - We need a lot more bandwidth (672 GB/s) – thus 12 memory channels



## GPU Teaching Kit



The GPU Teaching Kit is licensed by NVIDIA and the University of Illinois under the [Creative Commons Attribution-NonCommercial 4.0 International License](https://creativecommons.org/licenses/by-nc/4.0/).