VLSI Cost Model Cont’d

Effects of scaling, distribution of area in a processor

Cost at Different Scales

	45nm	32nm	20nm	7nm
Wafer cost	7000	7500	8000	9000
Diameter (cm)	30	30	30	30
Chip area (cm^2)	2.65	1.33	0.51	0.064
Defects/unit (cm^2)	0.5	0.35	0.29	0.25
test sites	0	0	0	0
P factor	4.3	4.5	4.8	5.1
Wafer yield	0.999	0.999	0.999	0.999
Test time	10	10	10	10
Tester cost	1000	1200	1500	1800
Package cost	12	12	12	5
Percent system cost	7	7	7	7
Final yield	0.97	0.97	0.97	0.97
Chips per wafer	225	473	1292	10781
Die yeld	0.315	0.641	0.864	0.983
Die cost	98.77	24.74	7.17	0.85
Test cost	8.82	5.2	4.82	5.09
Packaged cost	123.29	43.24	24.73	11.28
System cost	1761.29	617.71	353.29	161.14

System Cost

- Can be approximated from processor chip cost
- For high-end systems, about 7%
- For consumer systems, about 22%

Architecture Fits Cost

- Marketing & corporate goals dictate cost
- Architecture/design has to fit
- Given cost goal, work backward to get chip size for a given technology
- Within a cost constraint, plenty of options for design performance

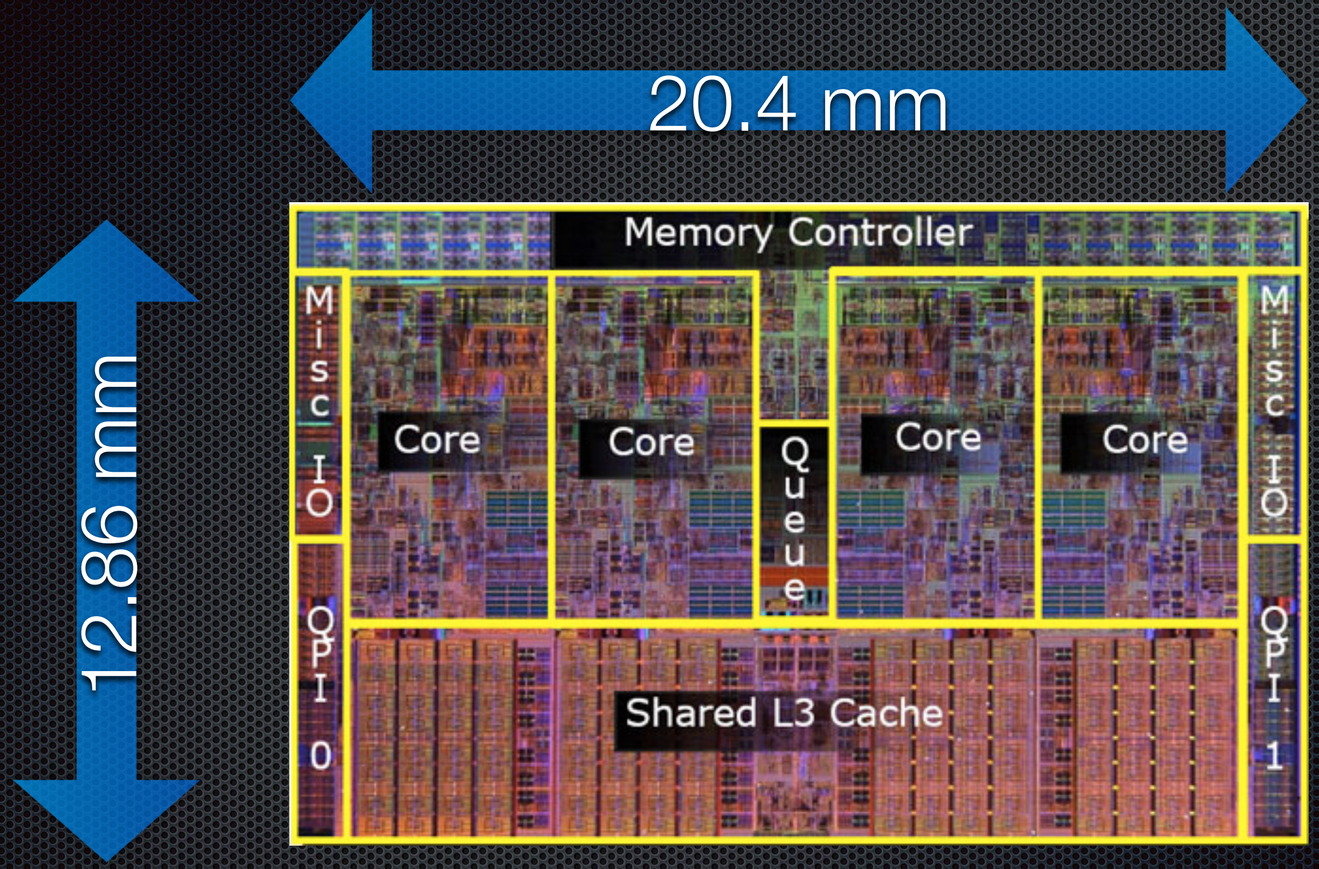
Chip Space Distribution

Intel Core i7

Core i7 Specs

- 4 cores, 2.66 to 3.2 GHz
- 731 million transistors, 45nm process
- Die size 263mm^2 = 20.40mm x 12.86mm = 0.8 x 0.5 in
- 32K L1 I-cache, 32K L1 D-cache, 1M L2, 8M L3
- On-chip memory controller w/3 channels, DDR3
- 1366 pins

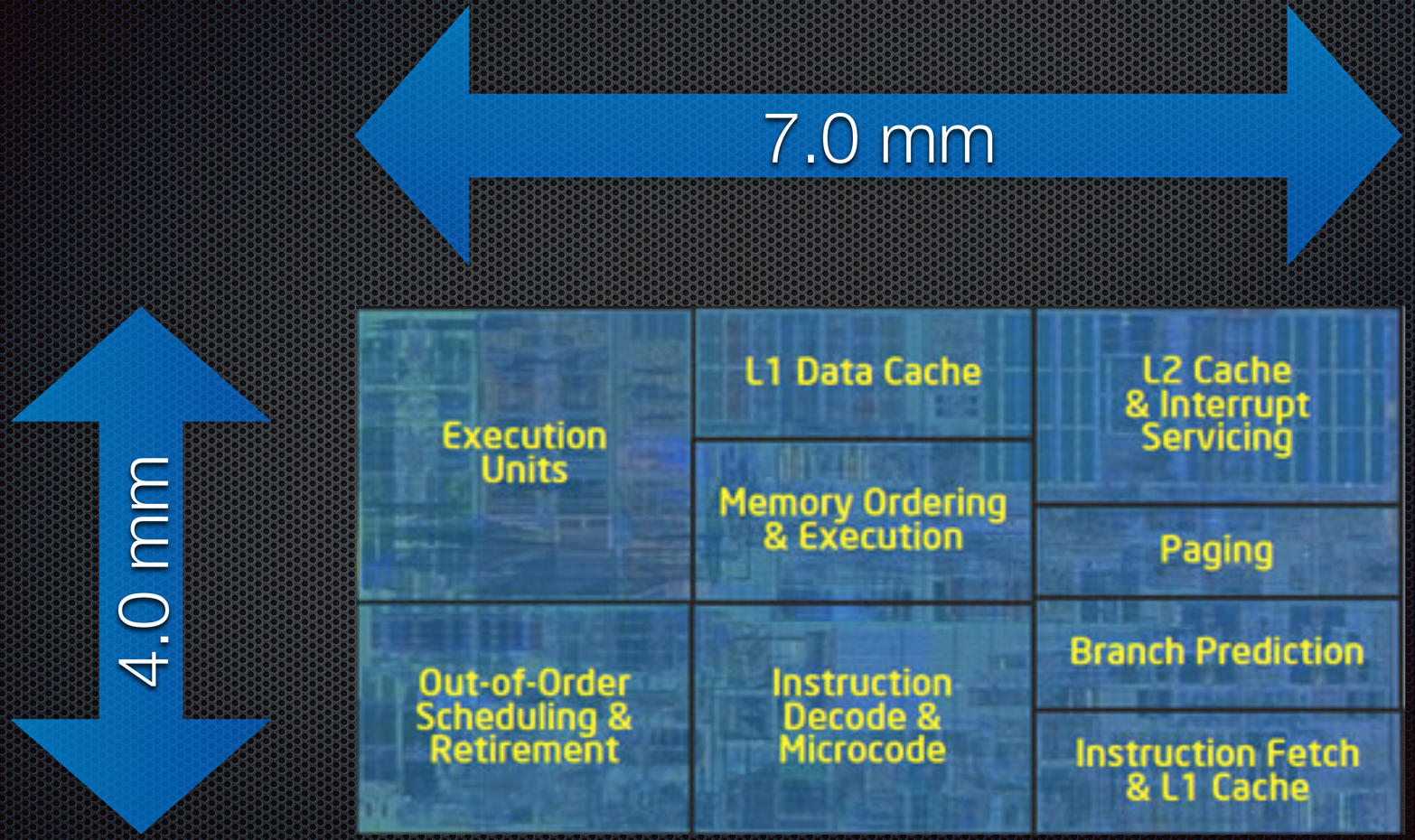
Die Dimensions



Major Block Distribution

- L3 Cache: 30%
- Cores: 45% (11.2% each)
- Memory Controller: 13% (more than one core)
- Misc. I/O: 5%
- Quickpath: 5%
- Queue: 2%

Core Dimensions



Core Area Distribution

- Execution Units: 19%
- L1 Data Cache: 8%
- L1 Instruction Cache: 9%
- L2 Cache: 10%
- Memory Ordering and Execution: 10%
- Out of order scheduling: 15%
- Paging (TLB): 6%
- Branch Prediction: 7%
- Decode: 15%

Overall Distribution

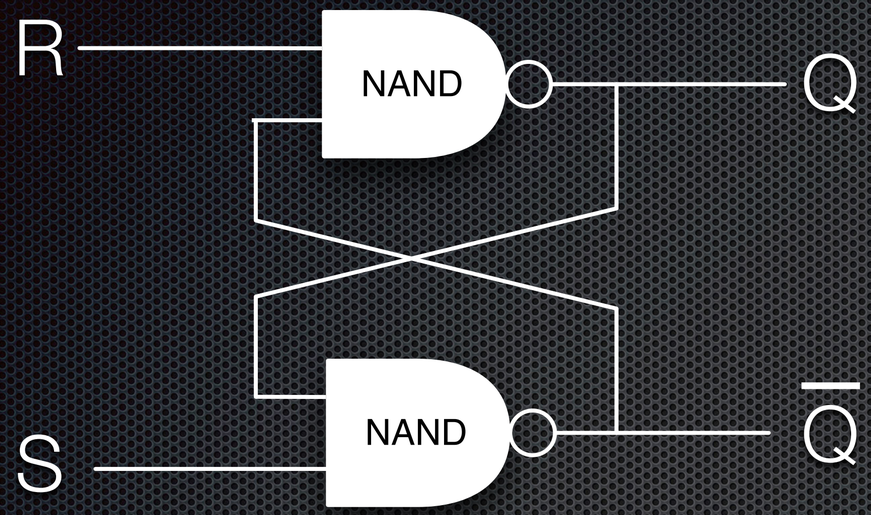
- Execution units: 8.6%
- Memory: 42% (60% including controllers)
- Instruction flow management: 17%

Summary

- Processors are becoming “smart memory”
- Heat is concentrated in a small fraction of the chip
  - but spread out across 45% of area
- More area dedicated to external access

Memory Technology

Basic Register Bit

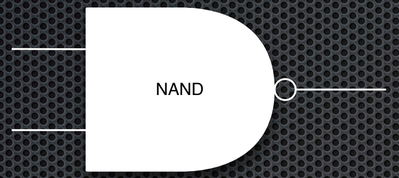


A	B	NAND
0	0	1
0	1	1
1	0	1
1	1	0

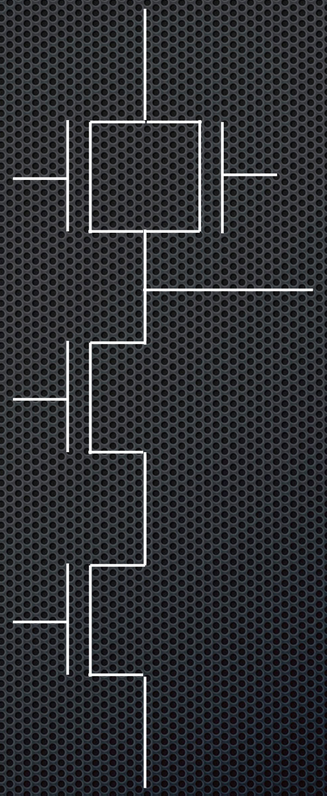
S	R	Q	Q
0	0	1	1
0	1	0	1
1	0	1	0
1	1	M	M

Cost of Register Bit

Each

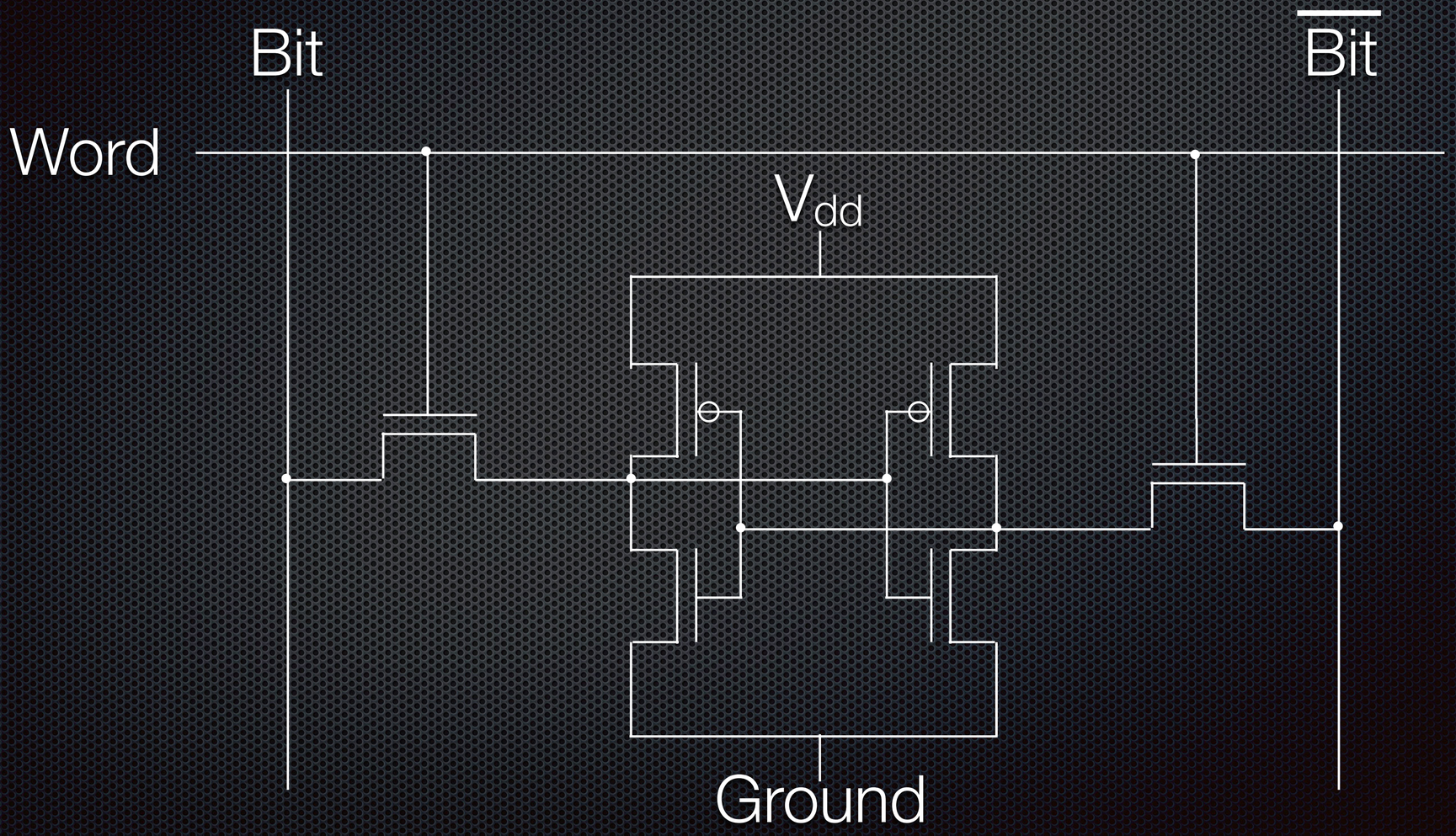


=

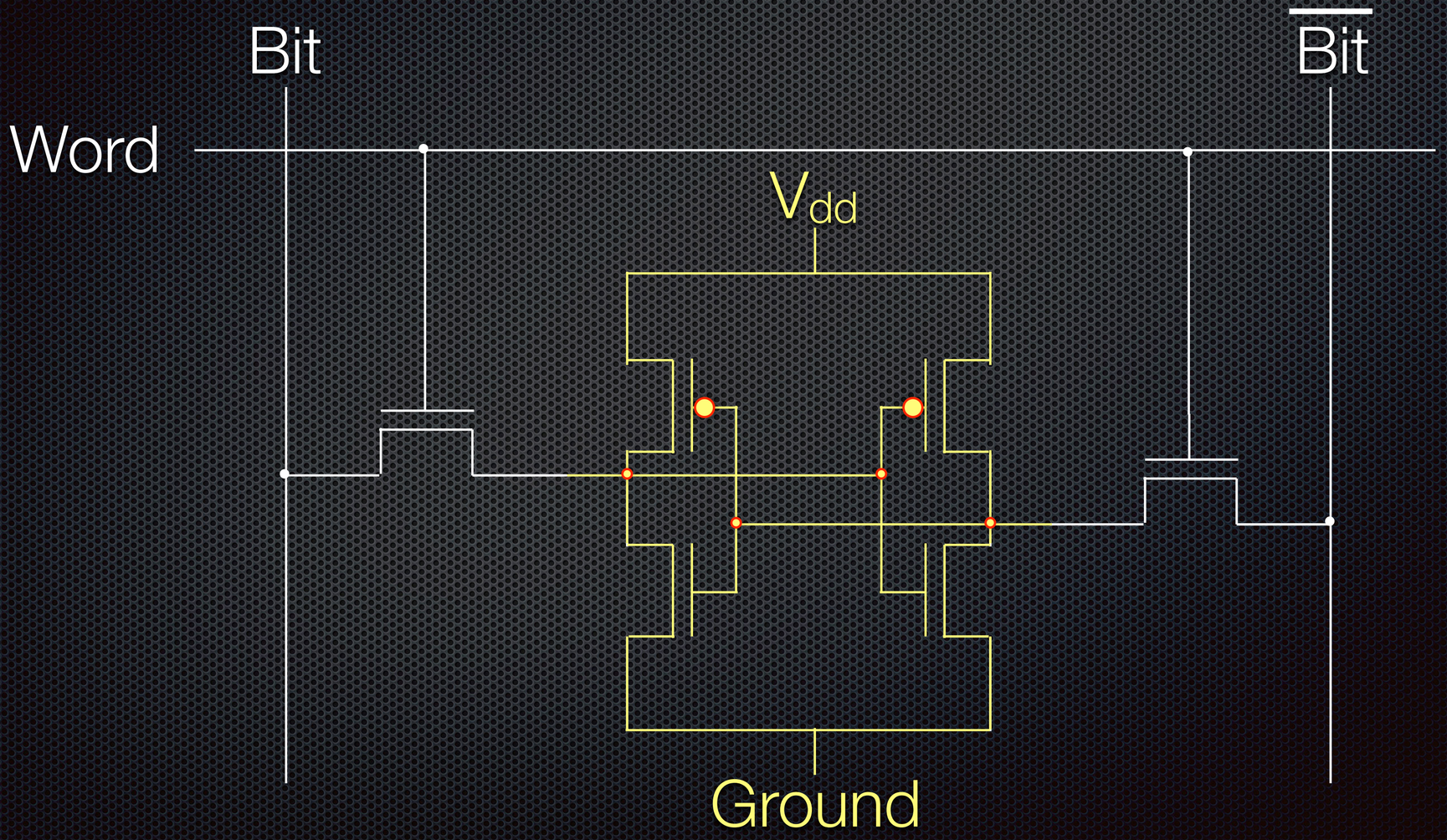


So each bit has 8 (full size) transistors

SRAM Cell



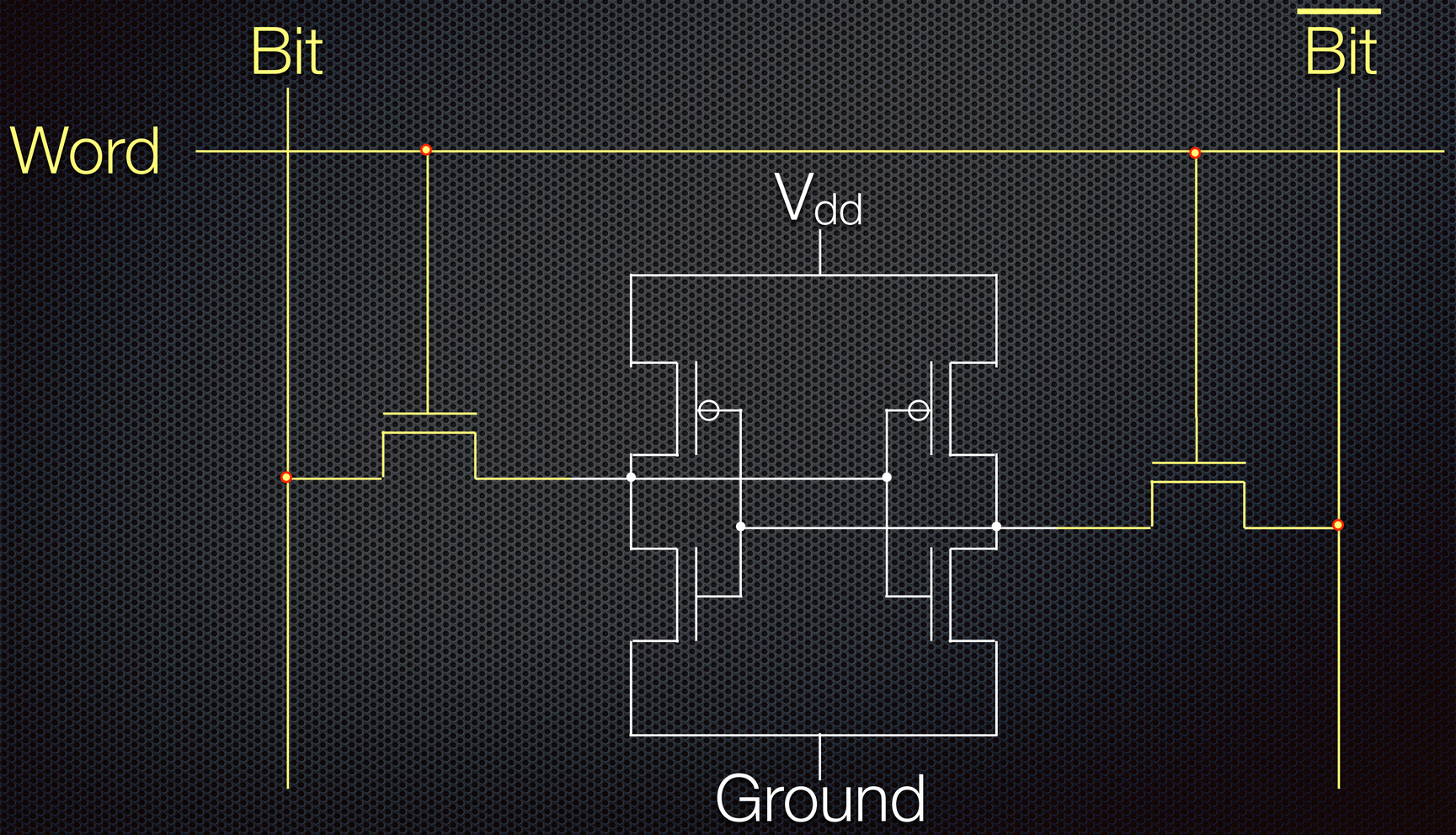
Cross-Coupled Inverters



Cross-Coupled Inverters

- Hold value in a feedback loop
- Built with undersized transistors
- Can be externally set/reset by more powerful signal

Word and Bit Lines



Word and Bit Lines

- When word line is active, bit lines are connected to the cross-coupled inverters
- If the input is active on the bit lines, it drives the inverters into a new state
- If the input is disconnected, the values in the inverters appear on the bit lines

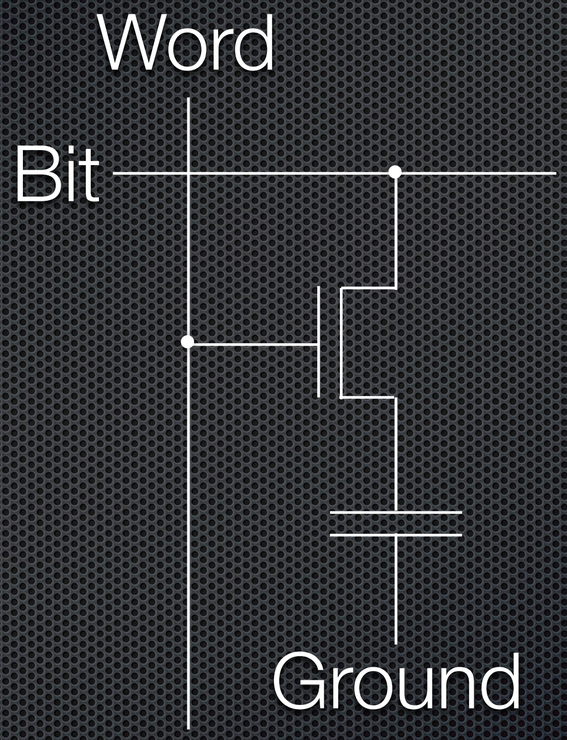
And then there’s reality...

- The inverter transistors are too small to drive the capacitance of the bit lines
- The bit lines are precharged to 1 before the word line is activated
- The inverter holding 0 partly discharges its bit line before being overwhelmed
- A differential amplifier at the end of the bit lines catches the momentary difference and enlarges it
- A latch (register-style bit) captures the value
- The value is written back into the cell

Cost of SRAM

- 6 transistors per bit cell (versus 8)
- 4 of the transistors are smaller
- Precharge, setup, latch, rewrite all combine to reduce speed

DRAM Cell



DRAM Write Operation

- Word line turns transistor on
- Bit line charges (1) or discharges (0) capacitor
- Word line turns transistor off
- Value is stored on capacitor

DRAM Read Operation

- Precharge bit line
- Word line turns transistor on
- If capacitor is charged, no change
- If capacitor is discharged, momentary drain on capacitance of bit line = voltage drop
- Sense amplifier chain enlarges the drop
- Latch captures the data
- Value is written back into the cell

More Reality

- VLSI capacitors are leaky
- Values shift toward indeterminate
- Periodically, all values in RAM must be read out and restored (refresh)
- Actual geometry uses pairs of odd/even bit lines, where one line acts as a reference for differential amplification, and both are precharged to an intermediate value

Cost of DRAM

- Just one transistor, plus a capacitor
- Special fabrication process puts capacitor in a well to minimize area
- Precharge, setup, amplifier chain, latch, restore make DRAM slower than SRAM
- Refresh adds small overhead
- Different process prevents mixing with logic, so signals must go off-chip

Memory Technology Hierarchy

Technology	Cost Per Bit	Speed
Register	At least 8 large transistors (34 typ)	Fast (1 cycle)
SRAM	4 small, 2 large transistors	Medium (2 - 20 cycles)
DRAM	1 transistor, 1 capacitor	Slow (100 cycles)

DRAM

Fuel tank for the processor

Growth in Size

- Most applications are not high performance
  - Even so, growth in features requires more RAM
- High performance applications depend on memory
  - Old mantra: A megabyte per mega-FLOP
  - Often limited by memory capacity
- Manufacturers emphasize size over speed
- 64K to 8G = factor 128K times bigger in 32 years

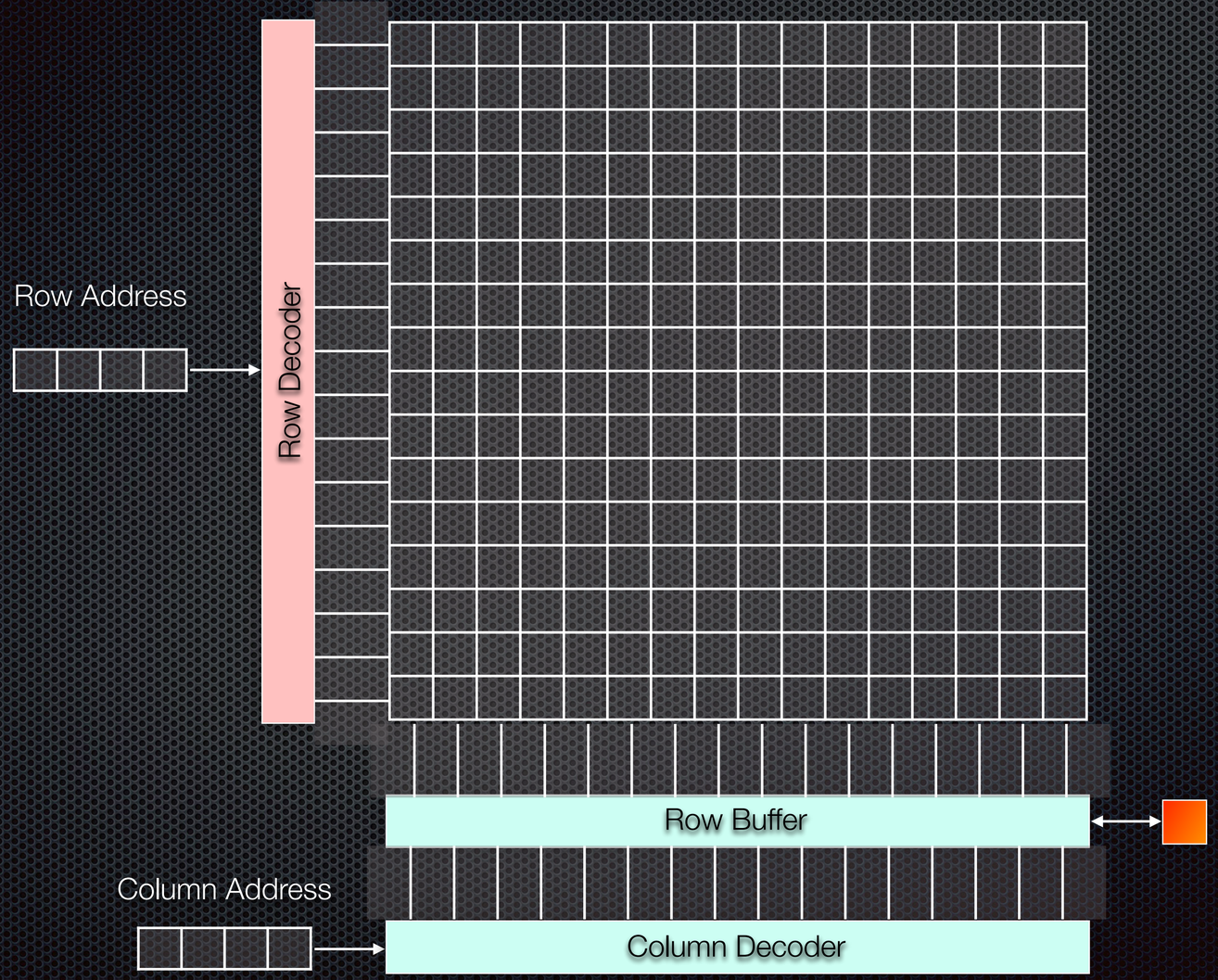
Effects of Size Scaling

- Fewer memory chips for a small system
- More compact large systems
- Reduced cost
- Reduced opportunity for innovation in memory architecture
  - Architecture is almost entirely within the chip
  - Only manufacturers can implement new ideas

Organization

- Bits arranged in a rectangular array: rows and columns
- Address is divided into row and column portions

Organization



Organization

- Bits arranged in a rectangular array: rows and columns
- Address is divided into row and column portions
- Row address selects a row to read or write
- Row is destructively read out to buffer registers
- Column address selects a register within the buffer
- Write overwrites the register, read copies it out
- Buffer is rewritten to row

Multiple Banks

- A large RAM will be divided into multiple banks
  - Partly for row access speed, partly for yield
  - Each bank has its own buffers, read/write logic
  - Address is further divided into bank portion
- Access is really to buffers
- Opportunity for increased performance

Buffer Access

- Once a row is in a buffer, sequential accesses can be to the buffer without extra row access cycles
- Adding a layer of buffering allows a new row access to start while a prior row is still being read
- Multiple bank buffers enable different pages to be open for fast access at the same time
- Internally, much higher bandwidth is available
  - But manufacturers won’t allow us to use it

DRAM Performance

- Access time decreased about 6X in 32 years (180ns to 30ns cycle) vs 36X increase in clock rate
- Bandwidth has improved more
  - Double data rate
  - Open pages, fast page mode, synchronous transfer
- Still slower than rate of CPU performance improvement

DDR

- Transfer data on rising + falling clock (double data rate)
- DDR 2.5 volts, 200MHz
- DDR2 1.8 volts, 400MHz
- DDR3 1.5 volts, 800MHz
- DDR4 1 to 1.2 volts, 1600MHz
- GDDR specialized for GPU, 32-bit bus, faster clock

Early Architectures

- FPMDRAM: Fast Page Mode (holds page open)
- EDODRAM: Extended Data Out (adds latch)
- SDRAM: Synchronous (multi-byte transfer)
- ESDRAM: Enhanced Synchronous (adds latch)
- SLDRAM: Synch Link (packet based, split transaction)
- RDRAM: Rambus (packet based, split, open pages)
- DRDRAM: Direct Rambus (lower cost, more transactions)

Current Architectures

- DDR-SDRAM: Uses double-width internal accesses and multiplexes interface with rising and falling clock edge transfers of half-access units
- Internal access rate is about 200 million per second
- Subsequent versions (DDR2, DDR3) increase internal access width, and interface bandwidth for up to 6400 MB/s transfer rate

DDR4 SDRAM

- Approx 3X to 5X increase in clock rate
- More internal banks
- Direct connect from each DIMM to memory controller
- Lower voltage (1.2V), 20nm process
- More pins (288 vs 240) per DIMM
- Parity on command/address bus, CRC on data bus

DDR4 SDRAM

- Released 2014
- Emphasis on ECC applications first
- Supported by Intel Haswell-EX (Mid-2015)
- Expected to achieve broader use in 2016, but…
- Much slower adoption than DDR3
- No current plans underway for DDR5
- Successor may be 3D memory technology

Flash Memory

- Will be covered in a later class in more detail
- Slower and cheaper than DRAM, faster and more expensive than disk
- Nonvolatile (though not permanent), thus lower power
- Limited life (100K cycles)
- Erase blocks to 1s, write pages with 0s
  - Slower to erase than read or write
- MLC stores 2 bits per site, but is slower, has shorter life