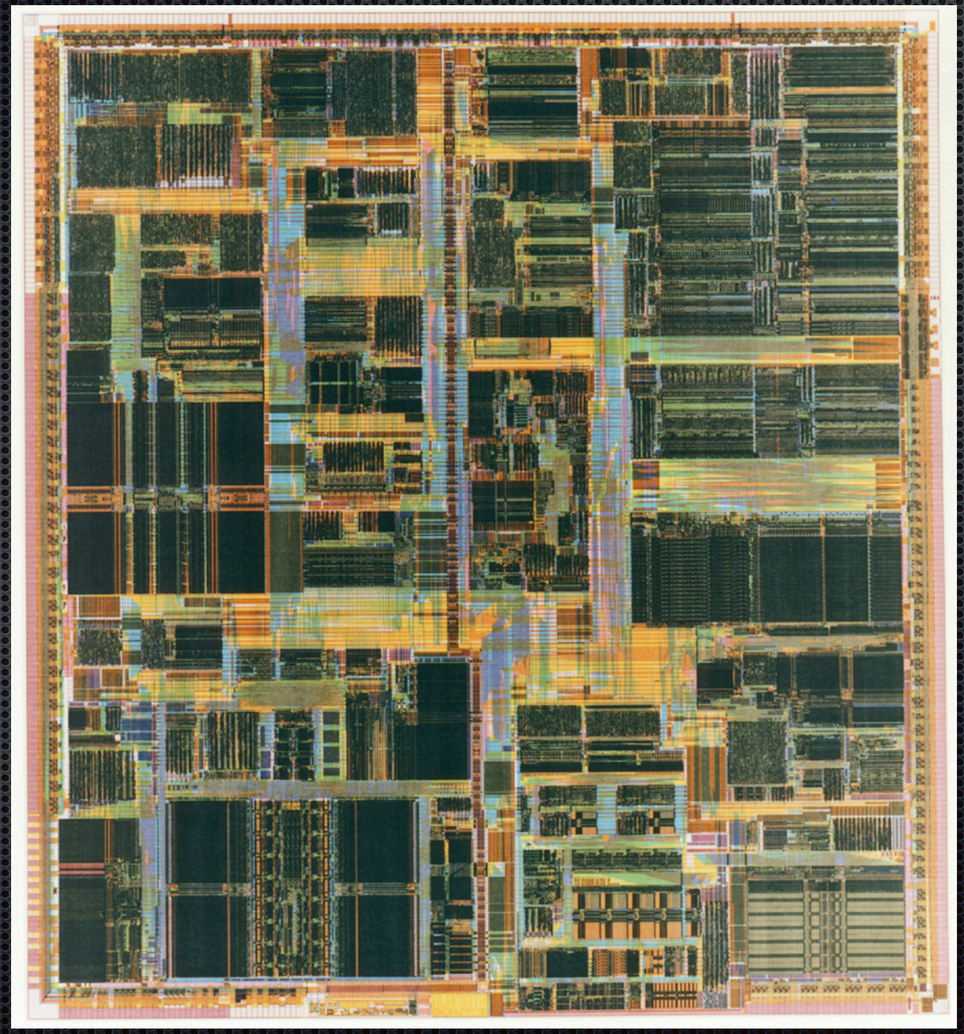
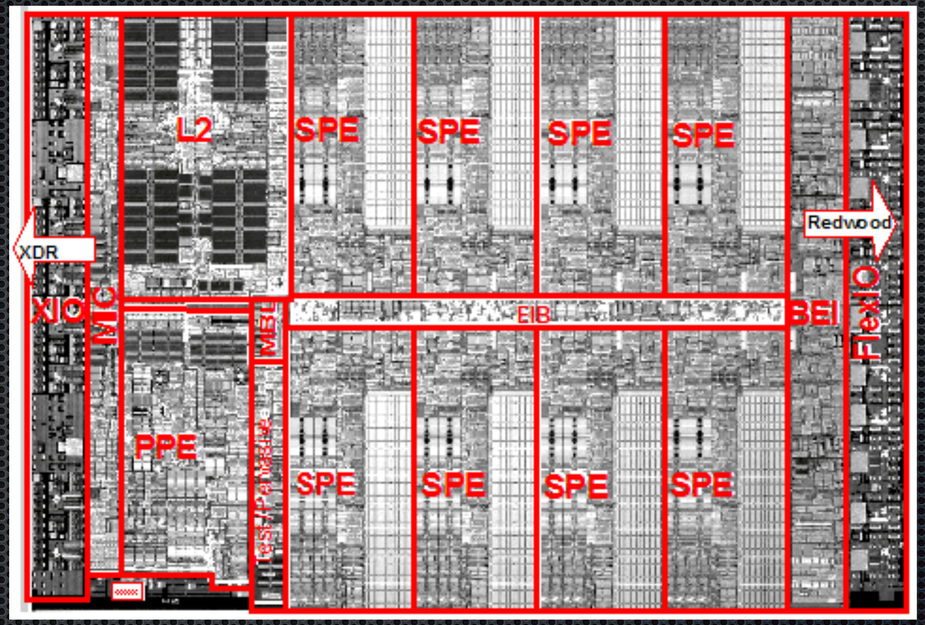
VLSI Technology

Fabrication and Structures

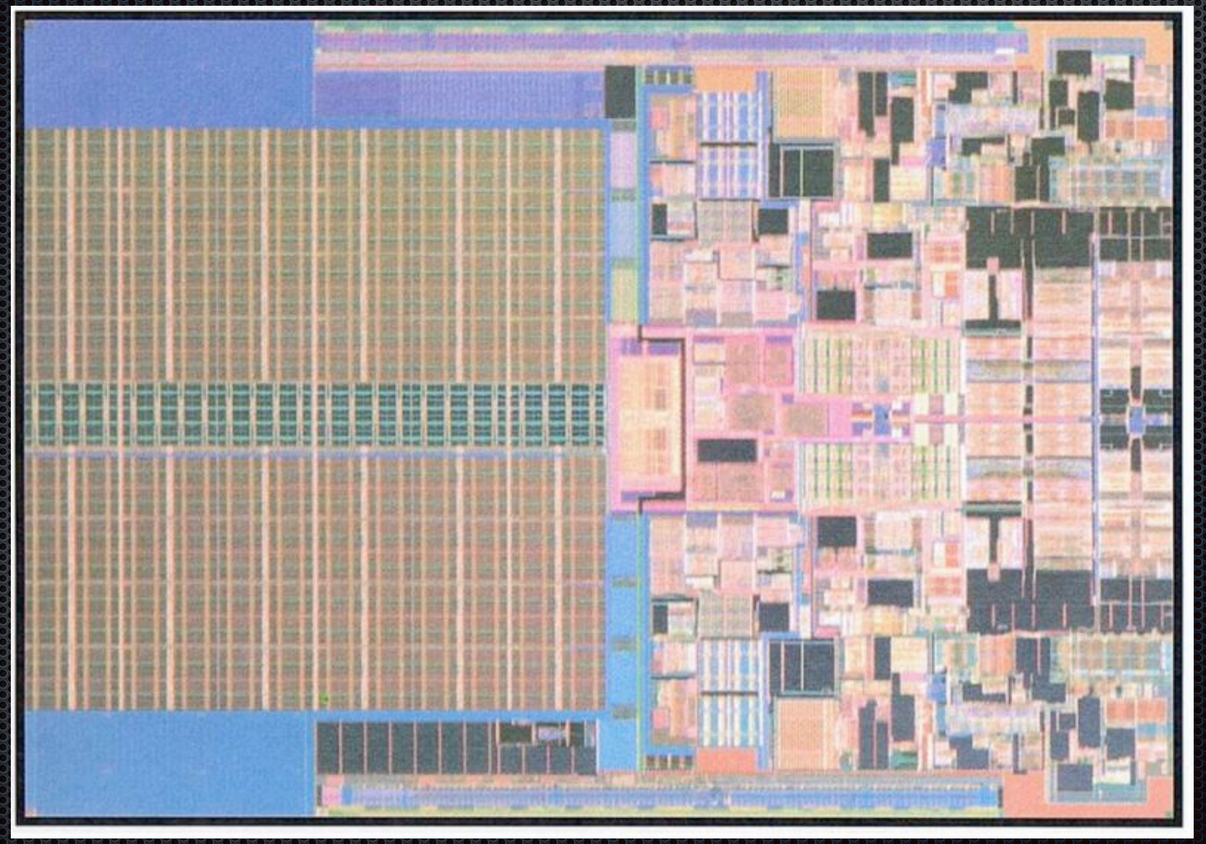
Chip Die Photo (P6)



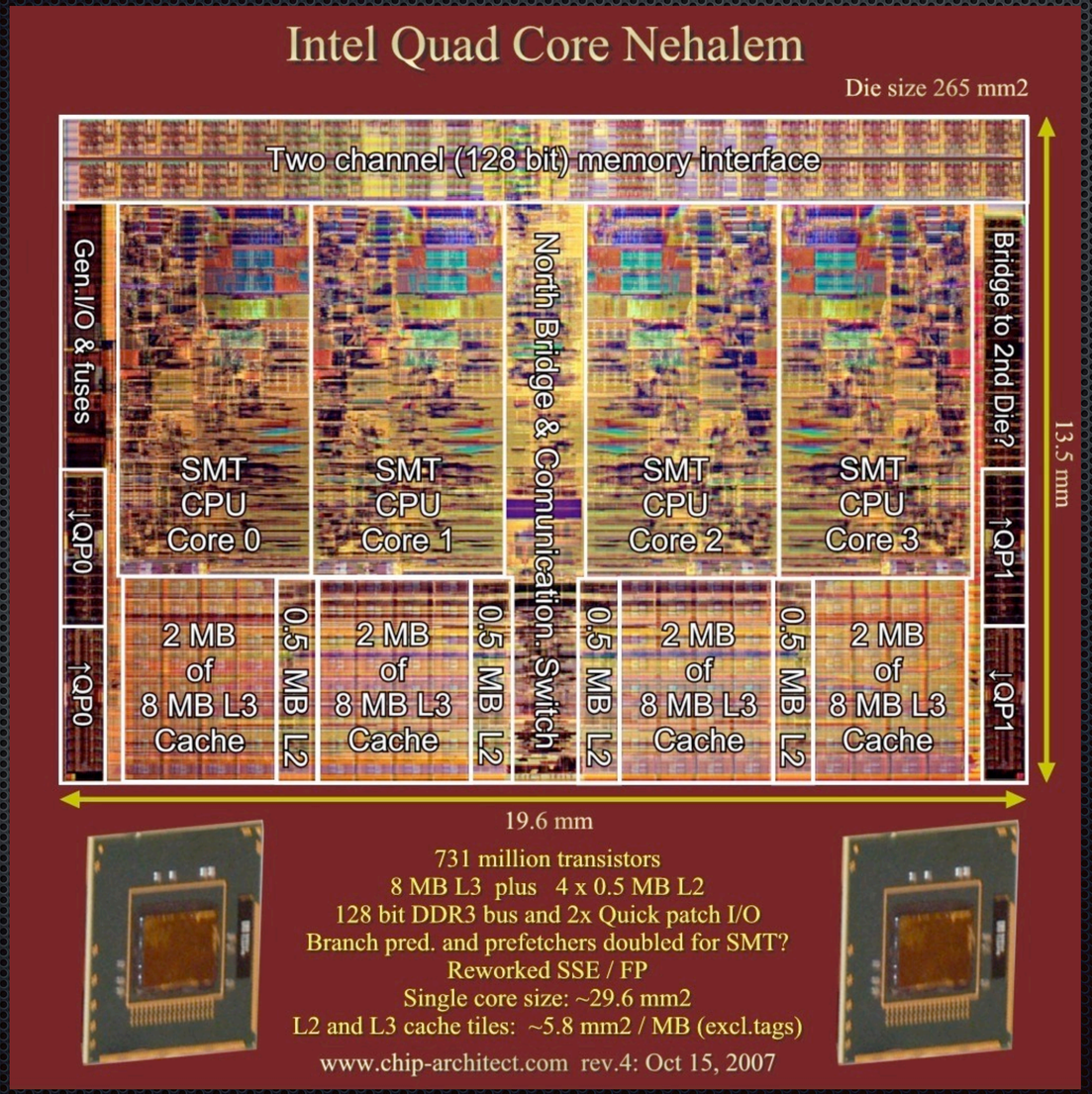
Another Chip (Cell)



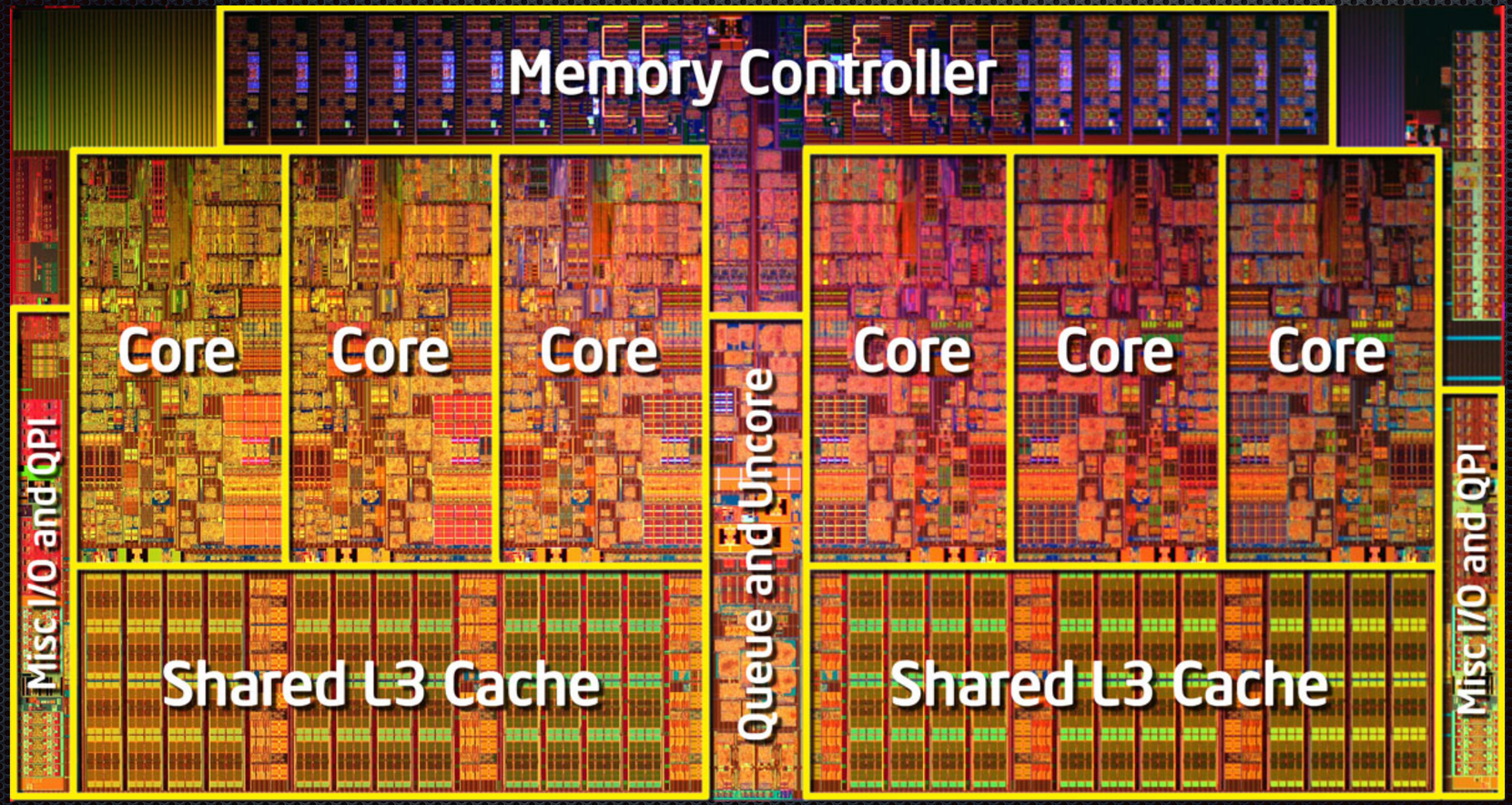
Dual-Core (Penryn)



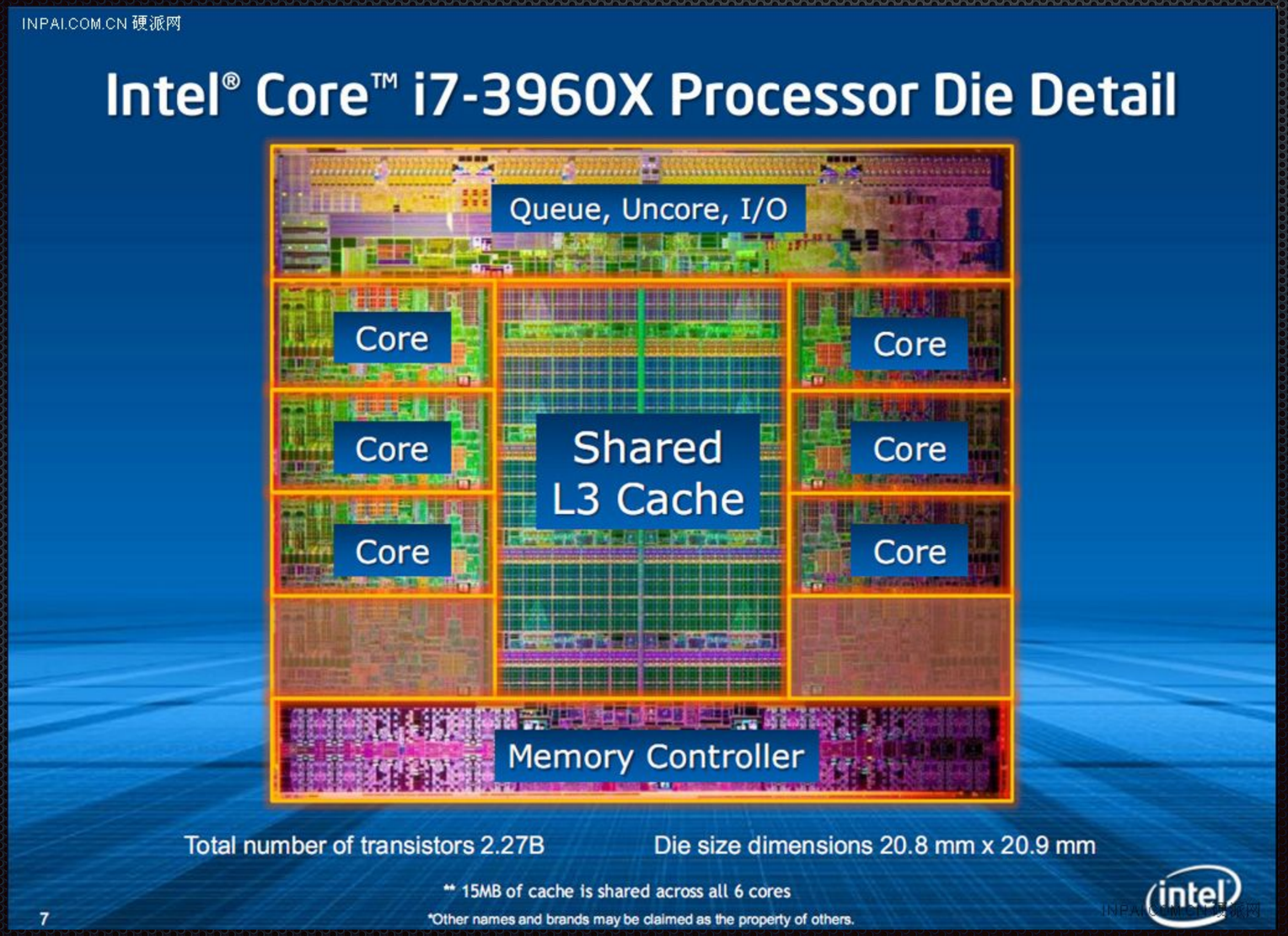
Quad-Core (Nehalem)



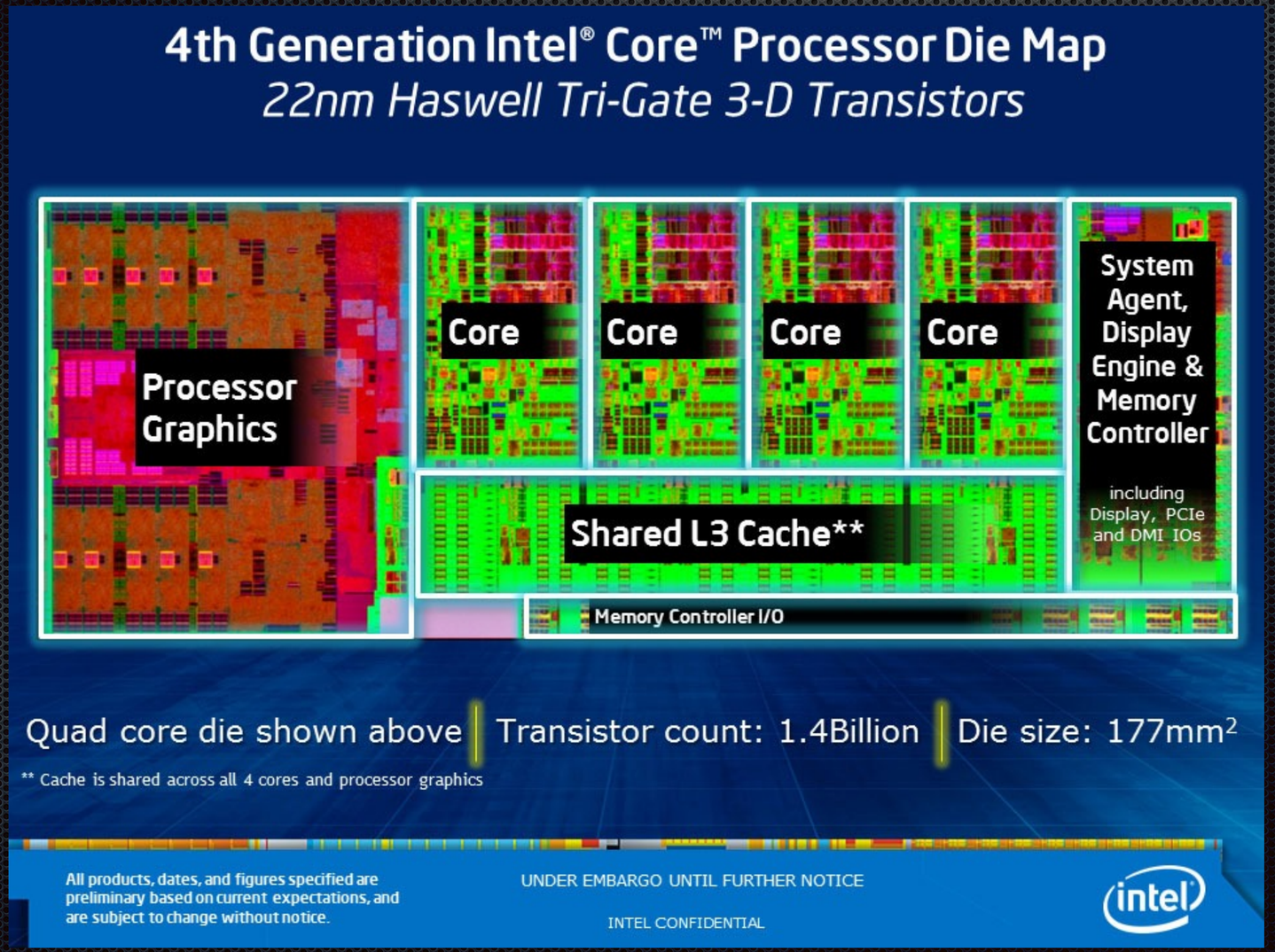
Six-core Gulftown



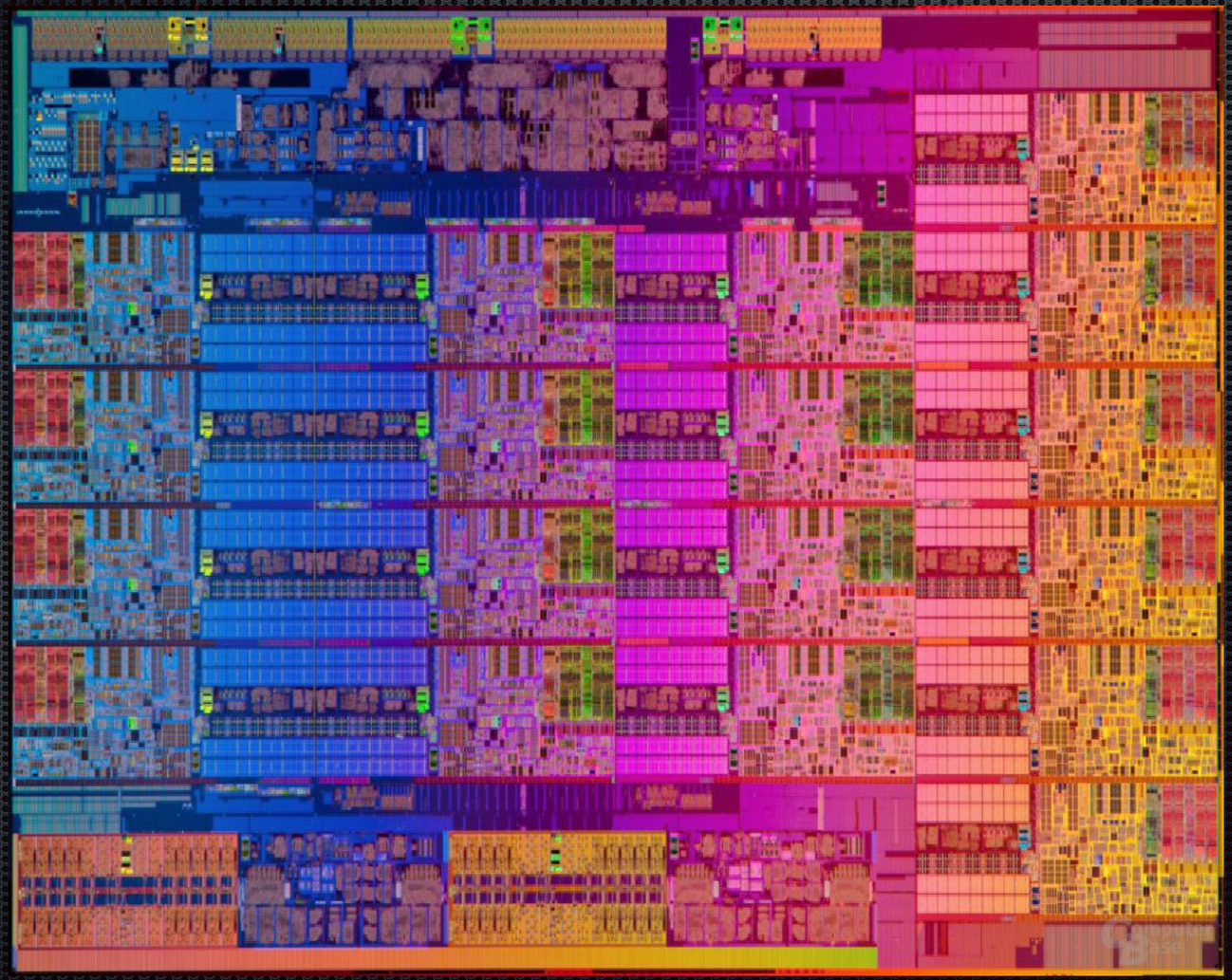
Six-core i7(Sandy Bridge)



Quad-Core Haswell w/GPU



18-core Haswell

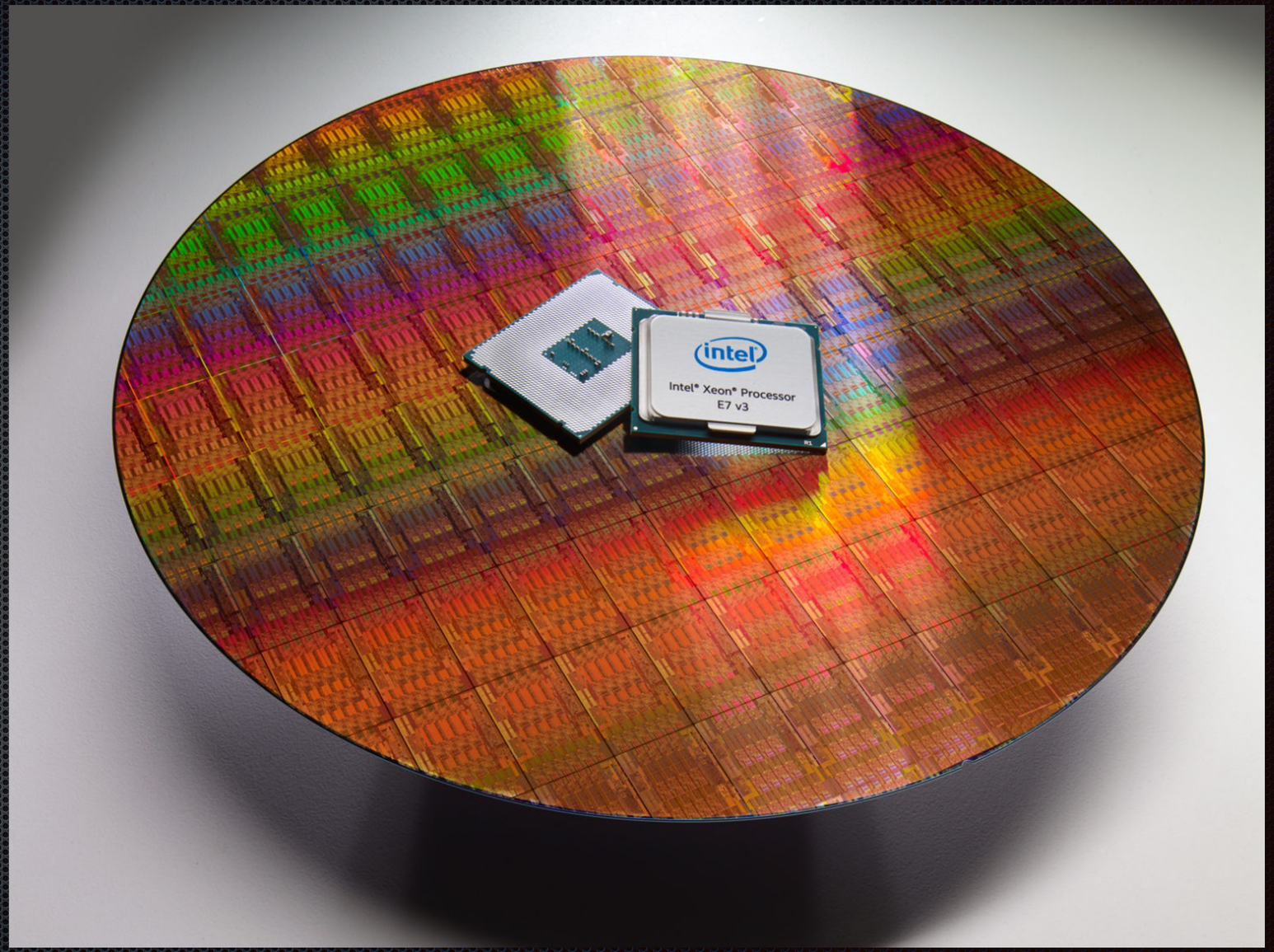


300mm Wafer

9 chips by 14 chips

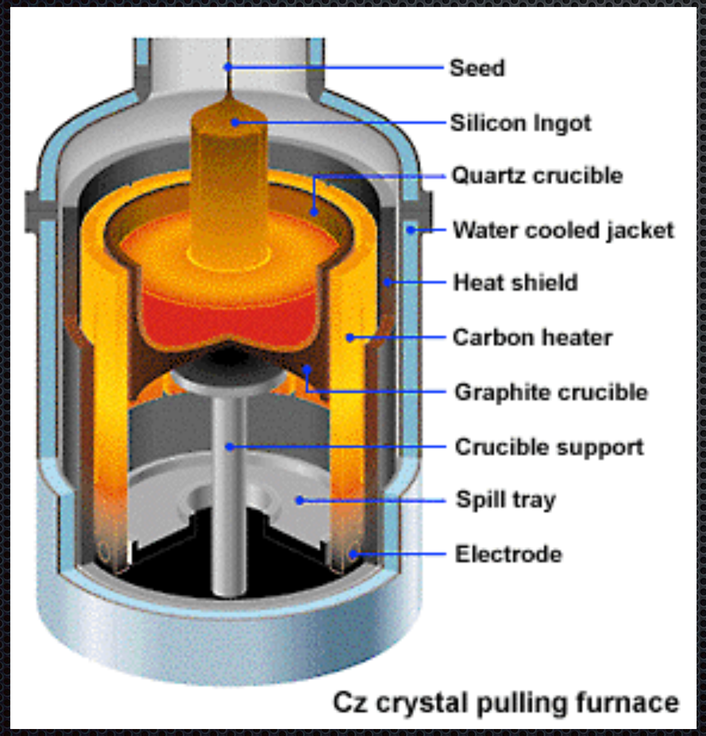
=> 33mm x 21mm

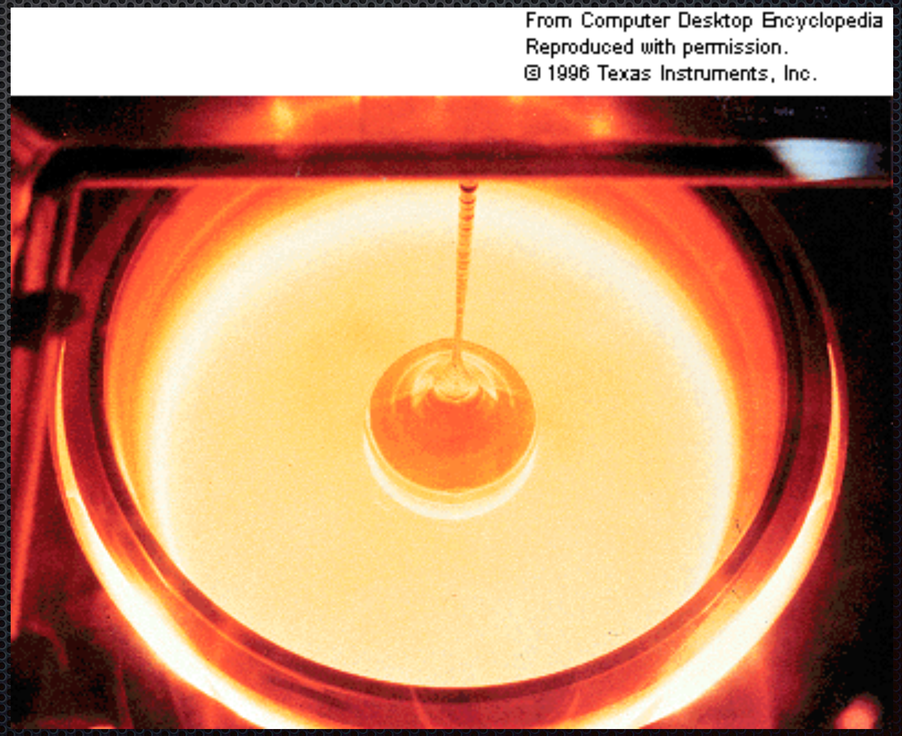
Only 79 whole chips



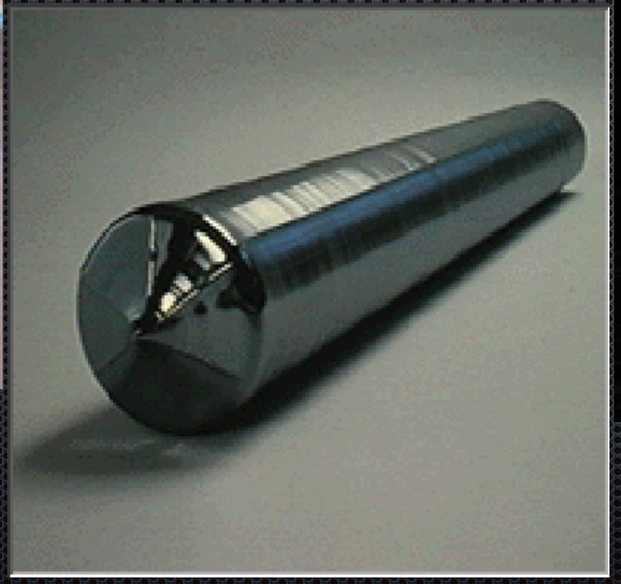
Fabrication Begins

- 99.9999999% pure Silicon, purified from old sand
- Slowly draw a single crystal boule (ingot) out of a melt, starting from a seed



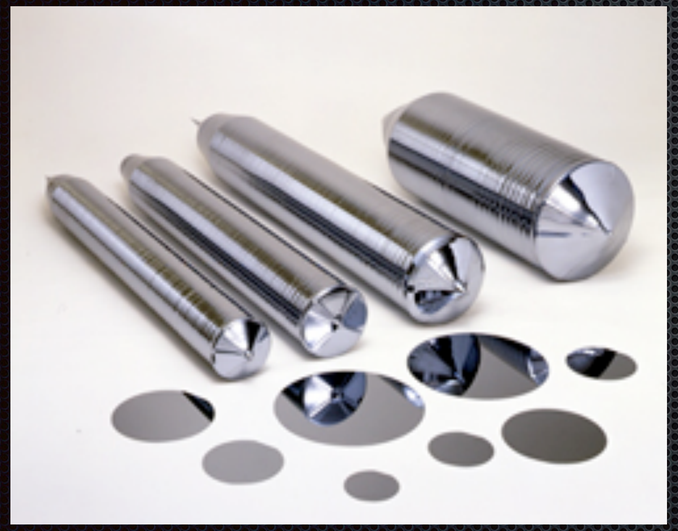


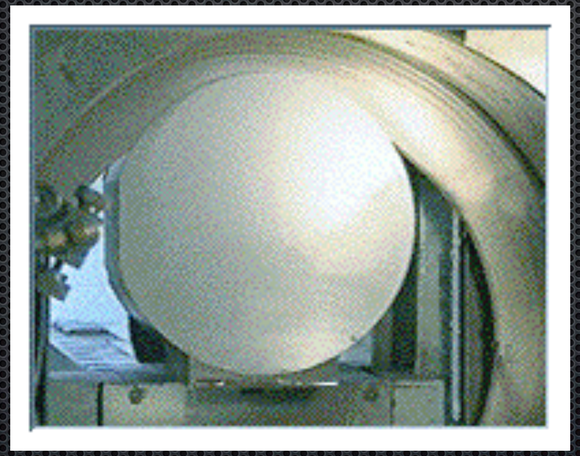
Silicon Boules

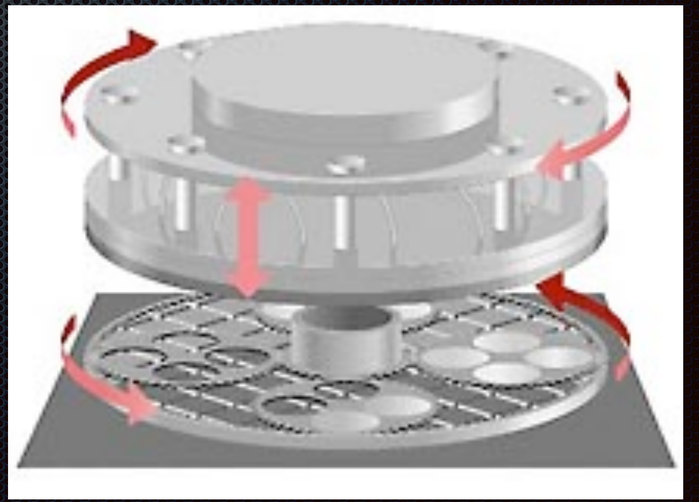


Turn into Wafers

- Lathe to a cylinder, machine flat or notch on one side
- Slice with inside diameter diamond saw, or parallel wire saw
- Polish, clean, and overcoat



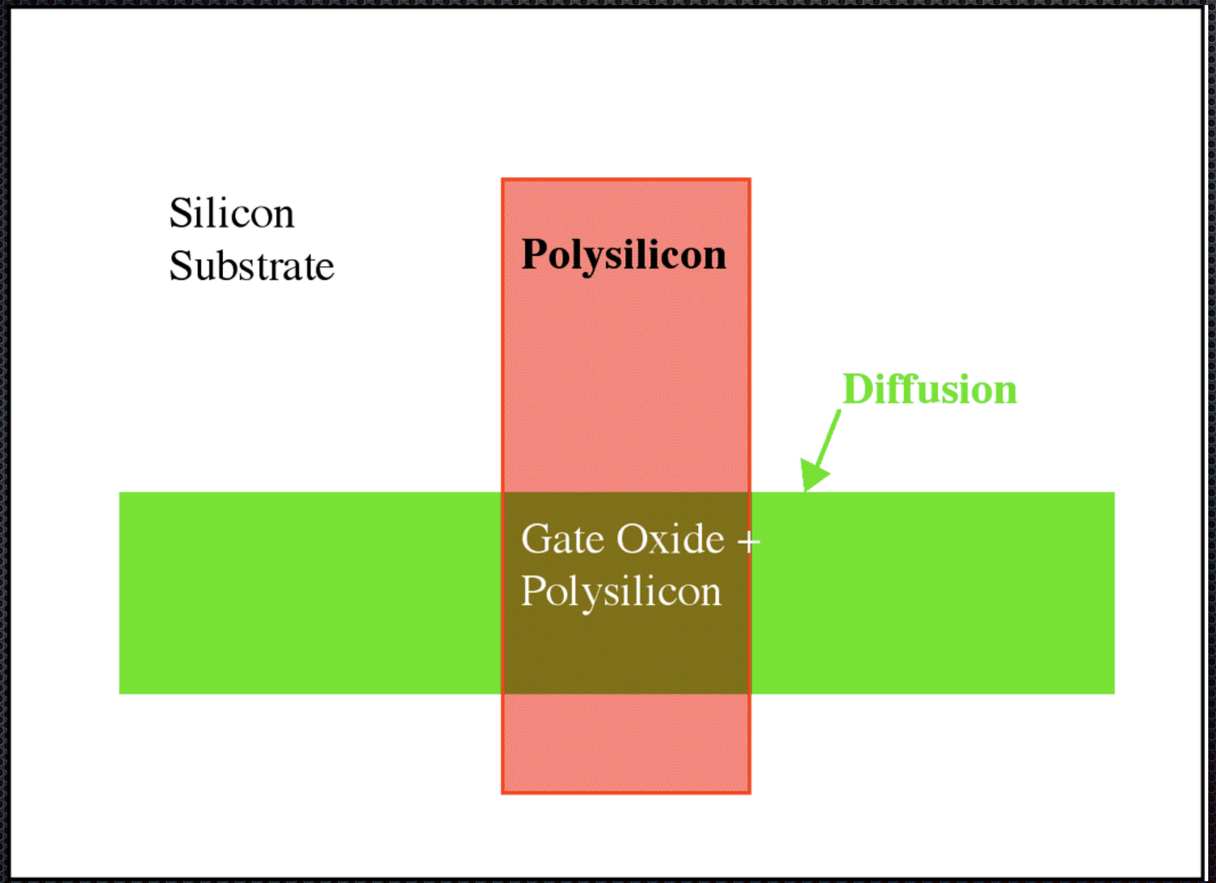




Processing Cycle

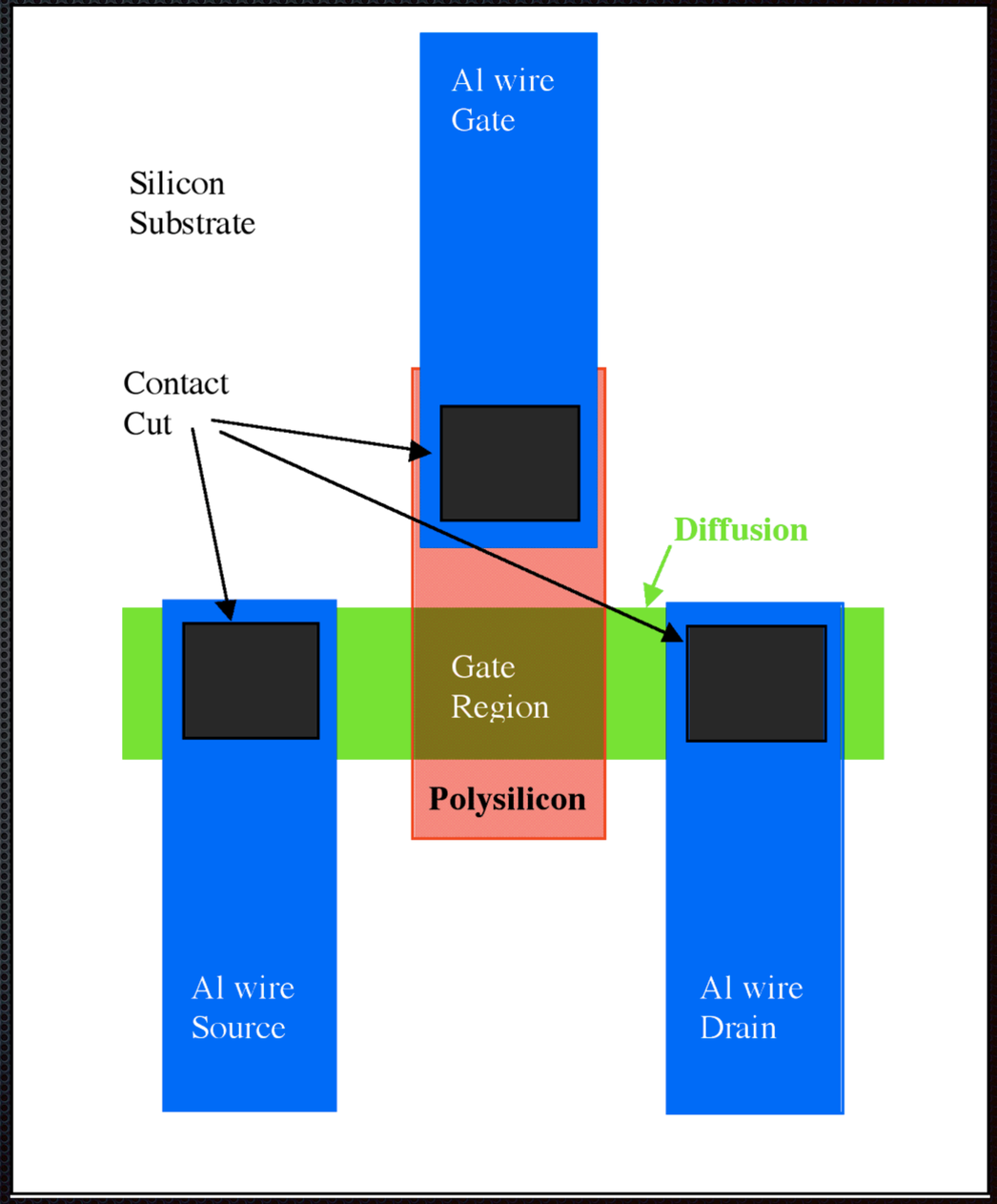
- Deposit a material
- Deposit a UV-sensitive photoresist
- Expose through a quartz/chrome mask
- Wash away unwanted resist
- Acid etch material into pattern
- Repeat many times (35 masks, 700 steps)

Typical Pattern



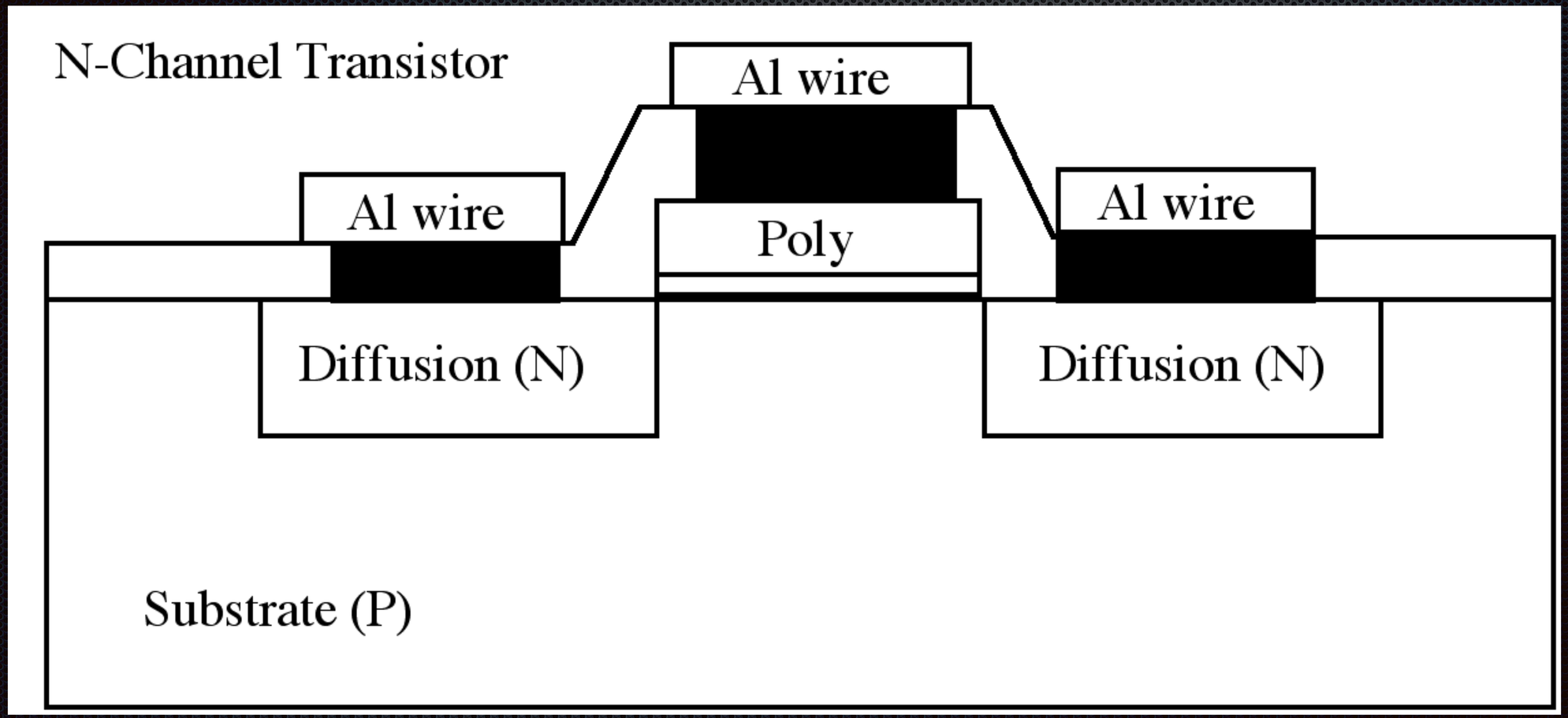
After More Steps

- Highly simplified layout
- Doesn’t show wells, guard rings, multiple metal layers, etc.

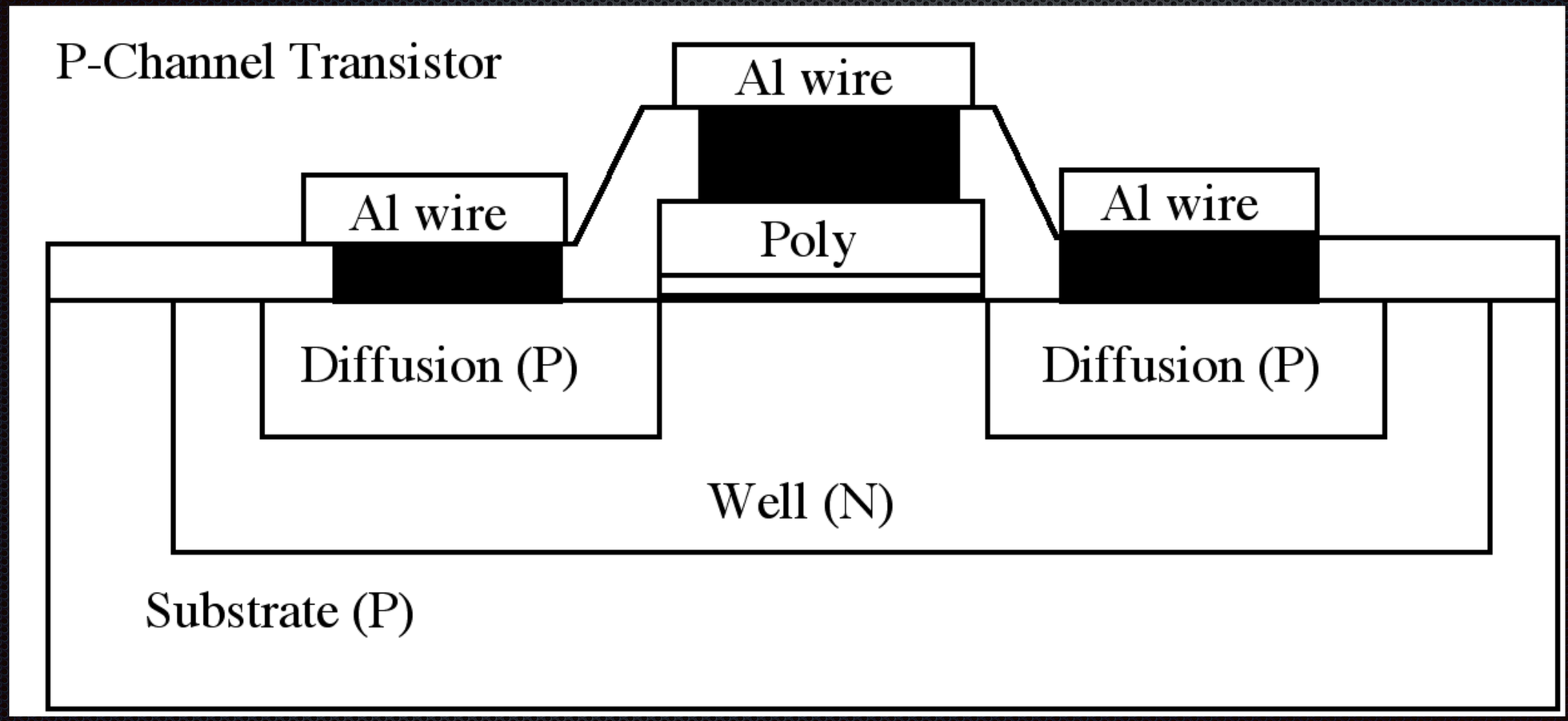


Cross Section

(NMOS transistor -- 1 = on)

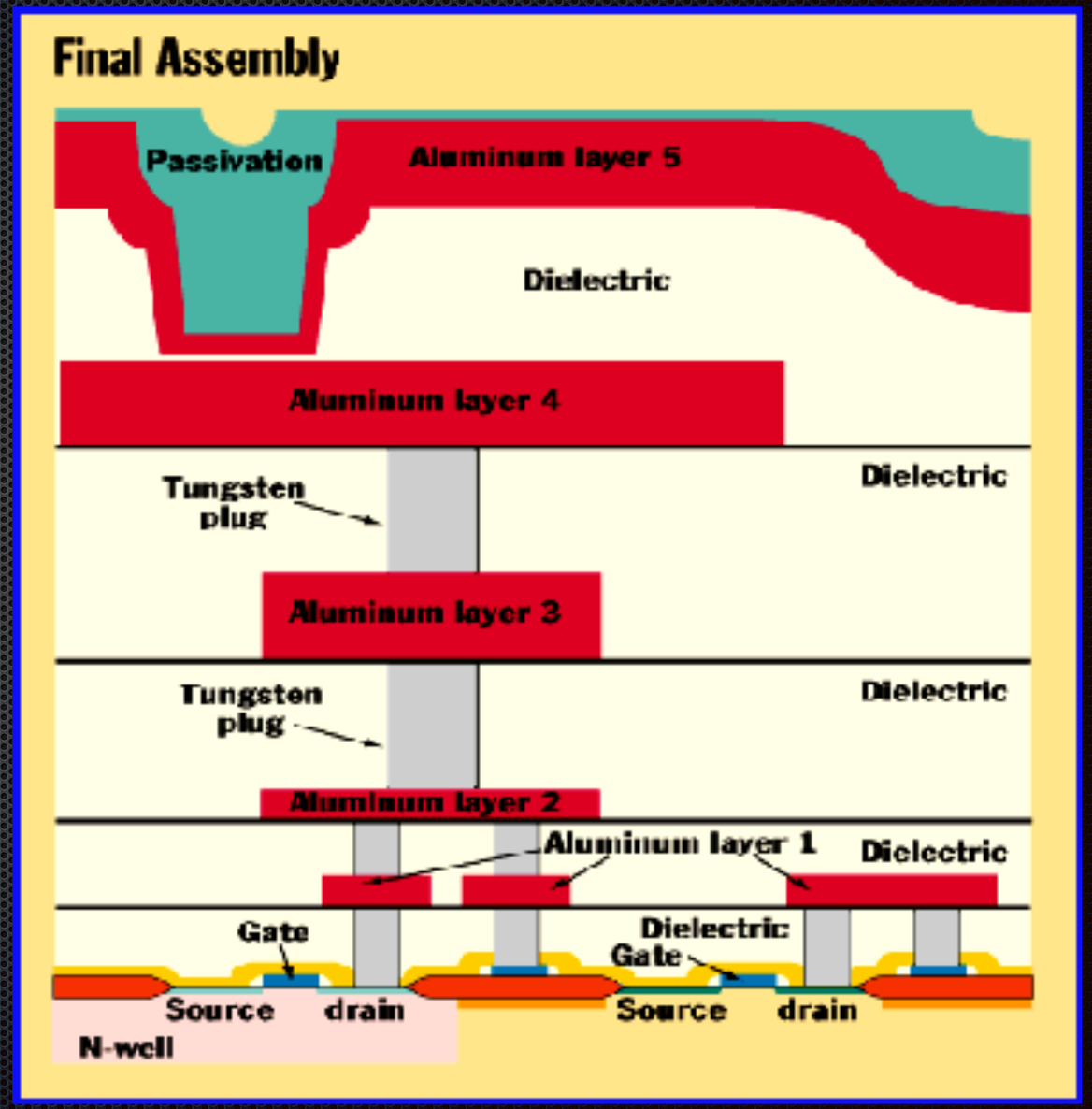


N-well PMOS Transistor
(0 = on)



Multiple Wire Layers

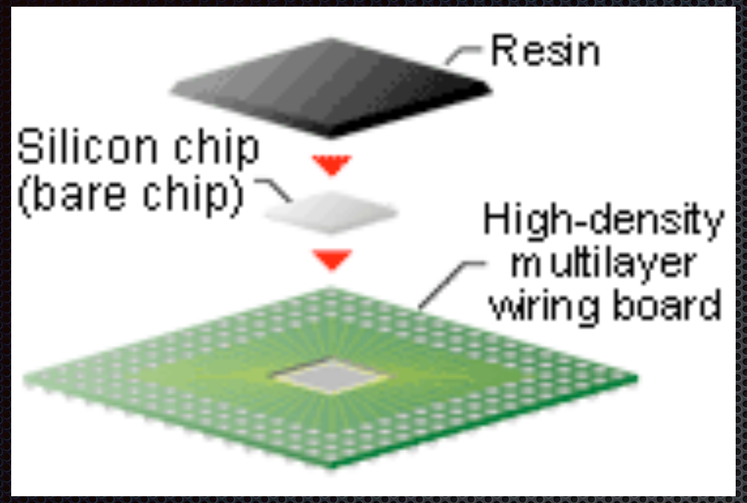
- Even with polishing, surface is uneven
- Higher layers must be thicker, and wires wider to be reliable. Reduces wiring density.
- Insulates heat-generating regions
- Up to 7 layers (4 in memory)

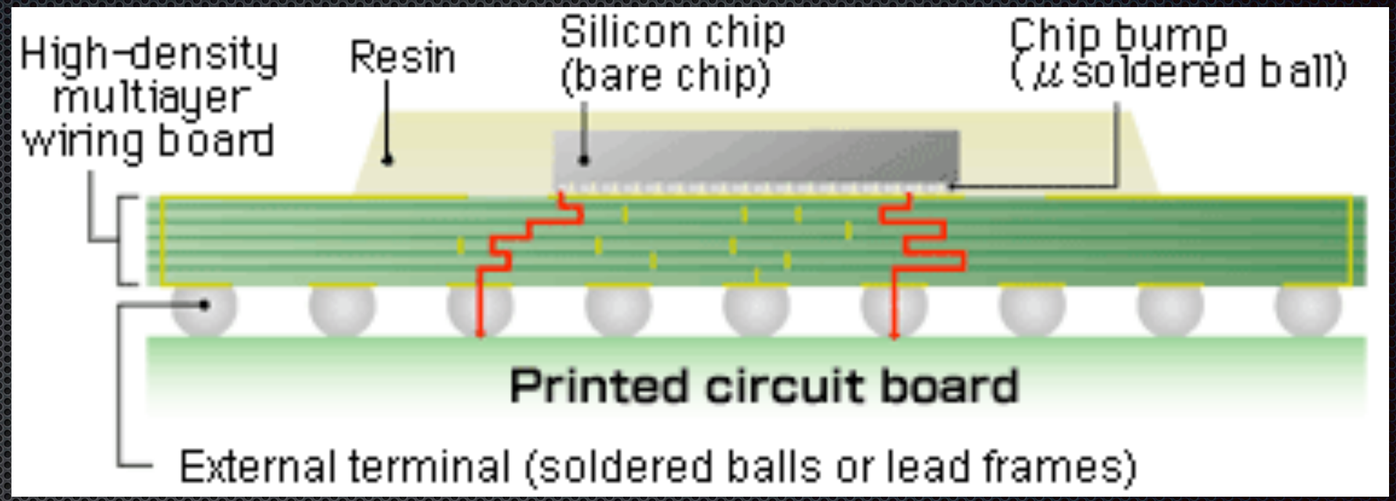


Bare Die Testing

- Optical inspection for defects
- Micro-probe tester contacts test pads, powers chip and runs test sequence
- Probes wear out, must be replaced
- Bad die are marked as rejects

Packaging





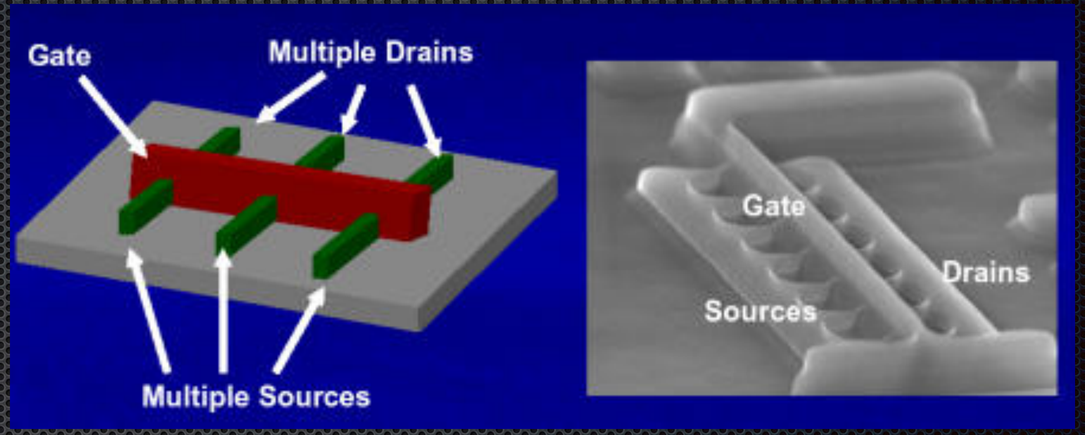
- Solder-ball or wire bonding
- Multi-level wiring in package
- Bare chip for heat sink contact in some cases

Packaged Chip Testing

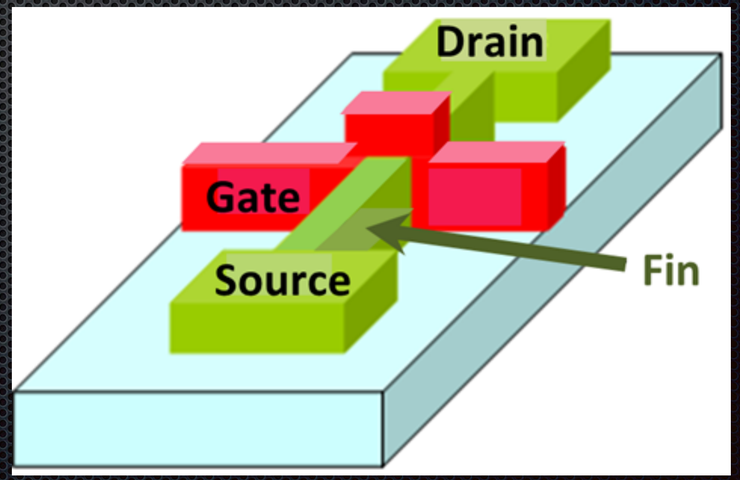
- Initial test (some fail in bonding)
- Performance grading test
- Power and thermal cycle tests
- Vibration tests (for some)
- Packages with bad chips recycled

Newer technologies

Intel Tri-gate
(in production)

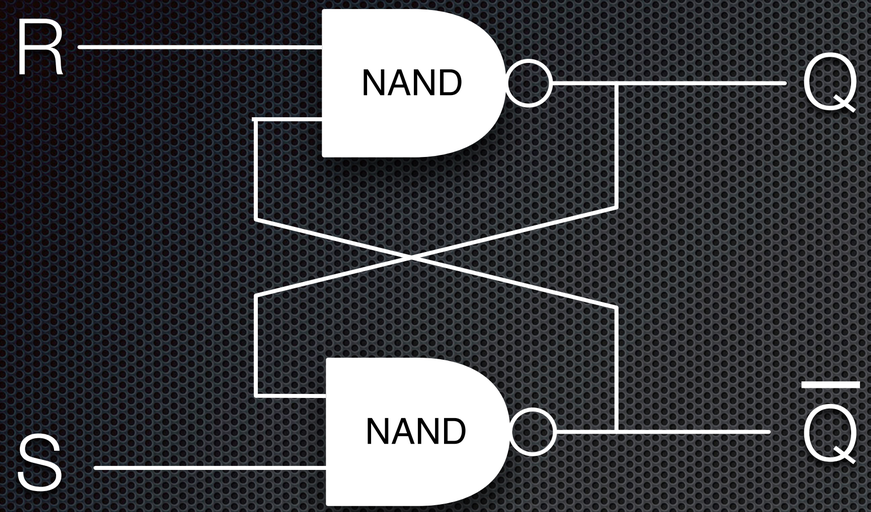


Fin-FET



Memory Technology

Basic Register Bit

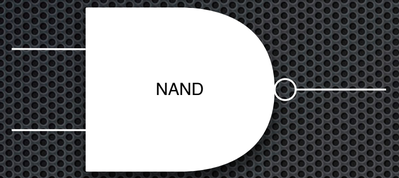


A	B	NAND
0	0	1
0	1	1
1	0	1
1	1	0

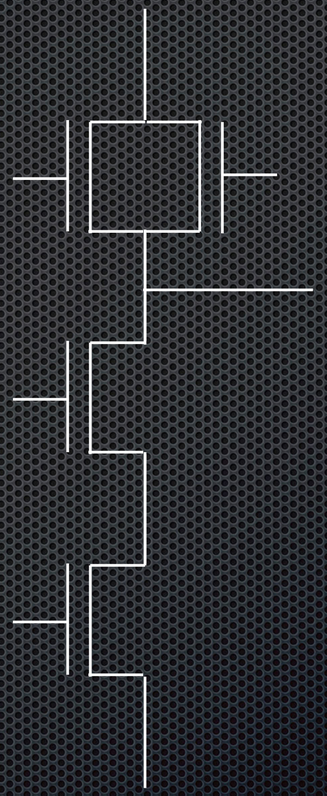
S	R	Q	Q
0	0	1	1
0	1	0	1
1	0	1	0
1	1	M	M

Cost of Register Bit

Each

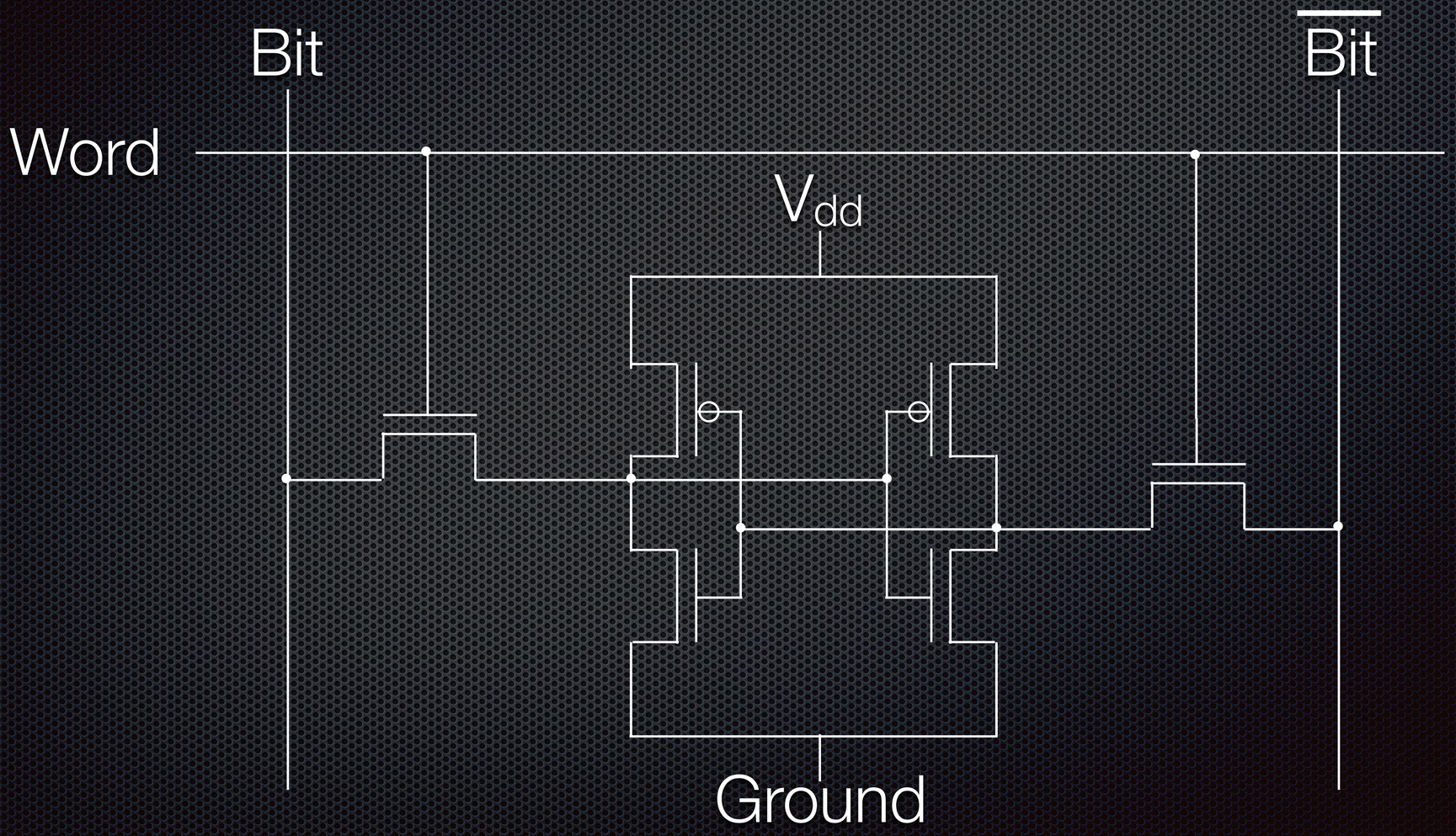


=

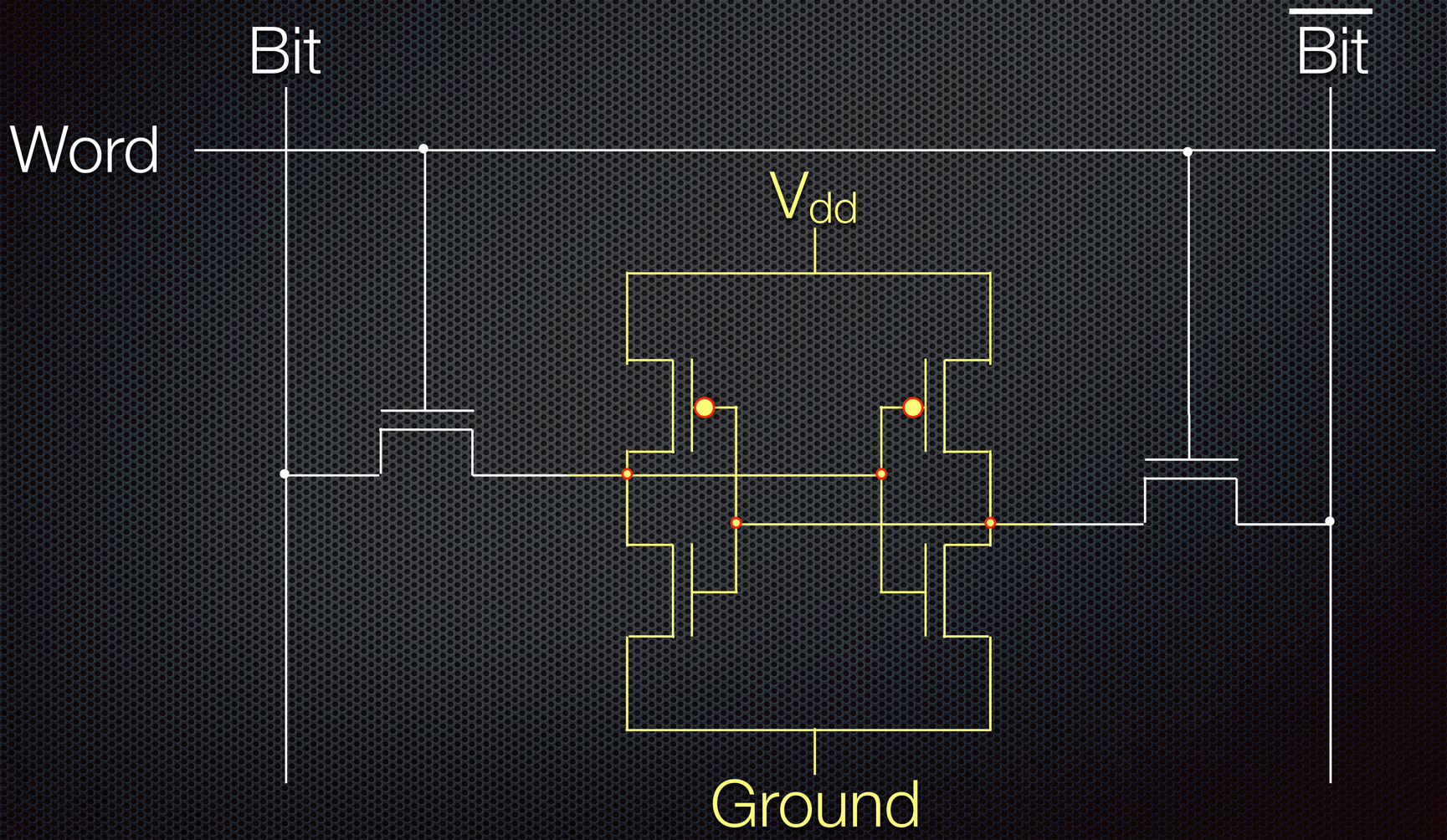


So each bit has 8 (full size) transistors

SRAM Cell



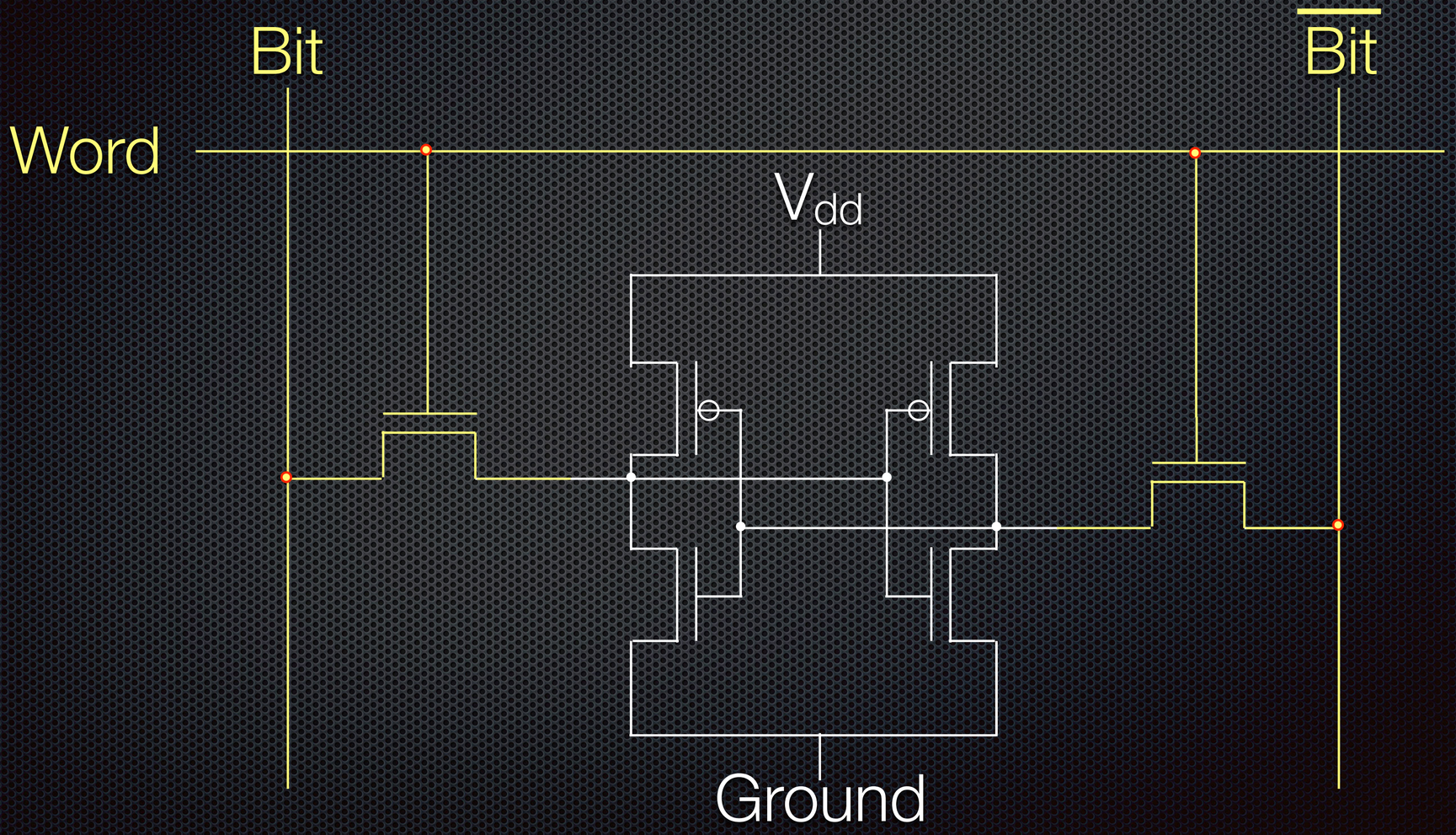
Cross-Coupled Inverters



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

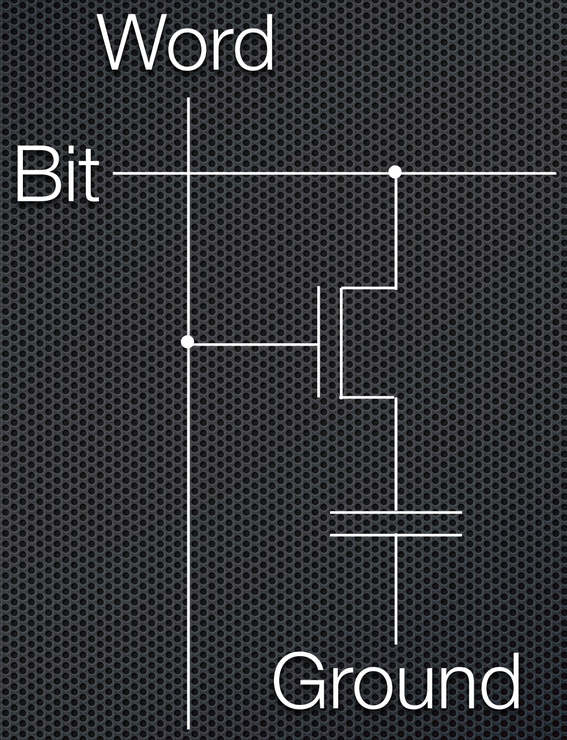
And then there’s reality...

- 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	8 large transistors	Fast (1 cycle)
SRAM	4 small, 2 large transistors	Medium (2 - 20 cycles)
DRAM	1 transistor, 1 capacitor	Slow (100 cycles)

VLSI Cost Model

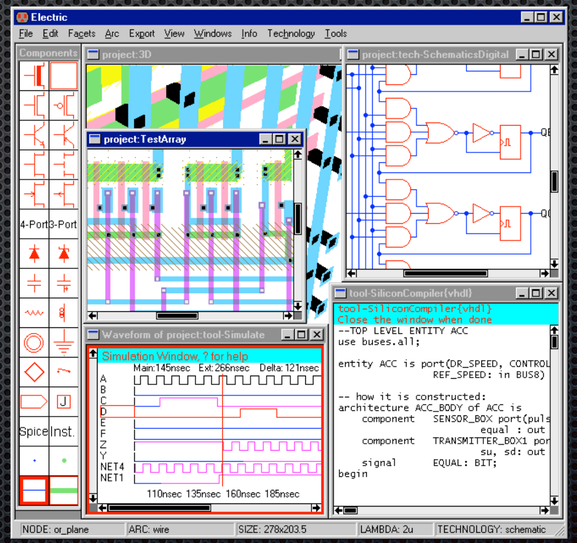
Estimating Cost of Chips and Systems for a Given Technology

Two Kinds of Cost

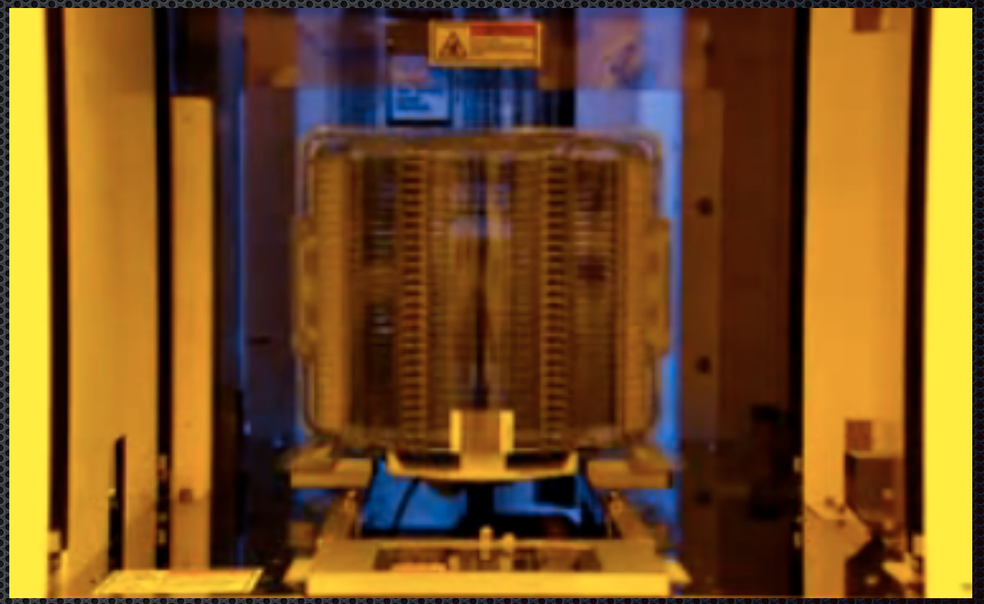
- Non-recurring (NRE)
- Recurring

Nonrecurring Cost

- Chip design ($400M),
- Plant capitalization ($4B),
- Mask set ($1.5M)



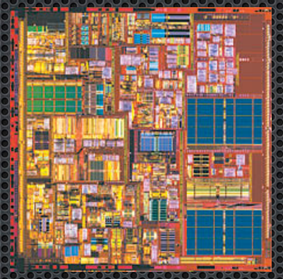
Recurring Cost

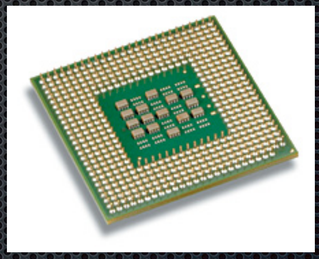


- Recurring: Manufacturing, packaging and testing, marketing, distribution, warranty, research, overhead, legal
- We will focus on manufacture, package, & test

Packaged Chip Cost

- $\text{Cost}_{\text{package}}$ and $\text{Yield}_{\text{final}}$ are given
- Other terms are computed





$$\text{Cost}_{\text{IC}} = \frac{\text{Cost}_{\text{chip}} + \text{Cost}_{\text{test}} + \text{Cost}_{\text{package}}}{\text{Yield}_{\text{final}}}$$

Cost of Chip

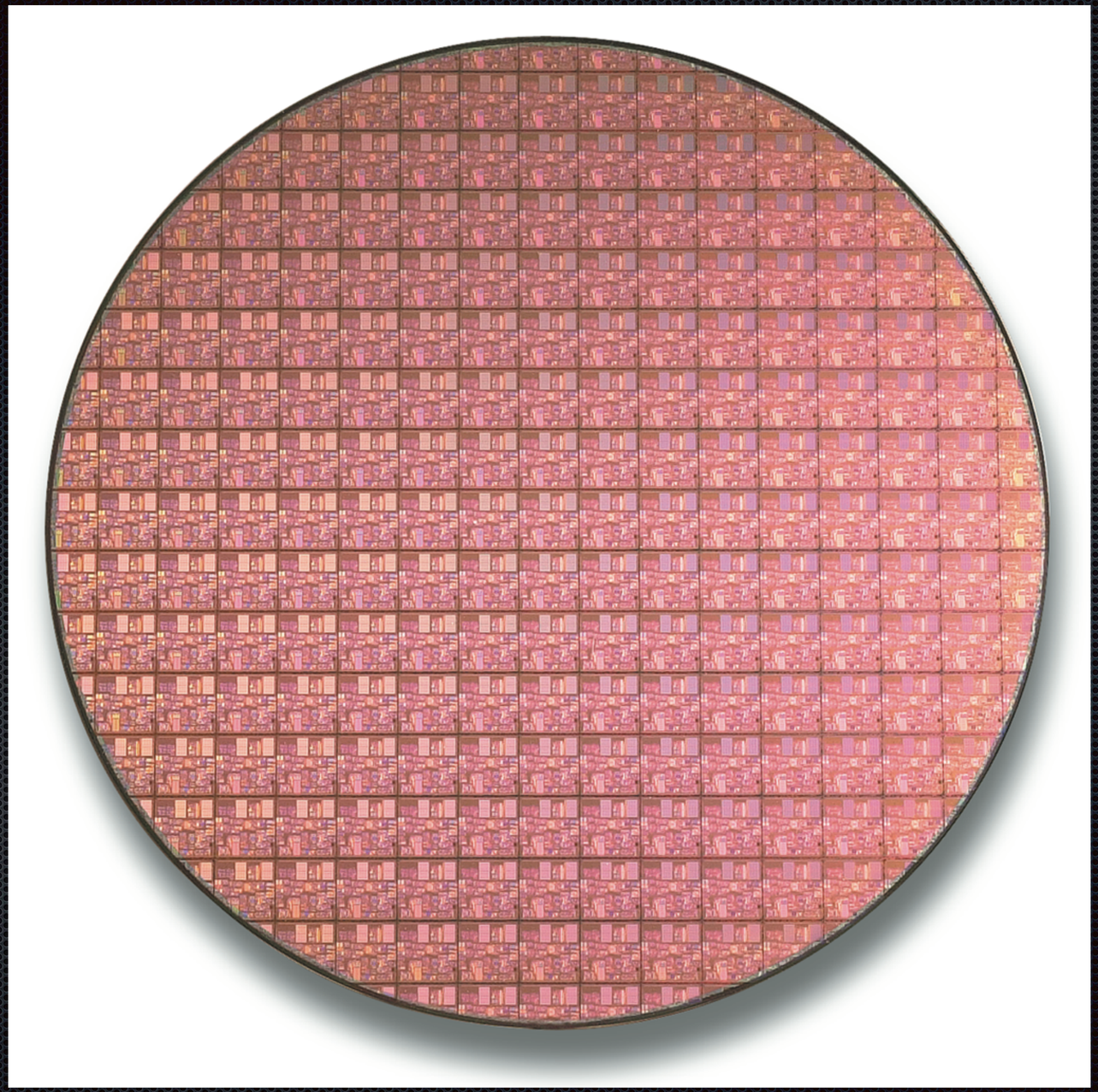
$$\text{Cost}_{\text{chip}} = \frac{\text{Cost}_{\text{wafer}}}{\text{Chips}_{\text{wafer}} \times \text{Yield}_{\text{die}}}$$

- $\text{Cost}_{\text{wafer}}$ is the cost of the finished wafer. Typically around $6000 to $8000
- Other terms are calculated

Chips Per Wafer

$$\text{Chips}_{\text{wafer}} = \left\lfloor \frac{\pi \times \left(\frac{\text{Diameter}_{\text{wafer}}}{2} \right)^2}{\text{Area}_{\text{chip}}} - \frac{\pi \times \text{Diameter}_{\text{wafer}}}{\sqrt{2 \times \text{Area}_{\text{chip}}}} - \text{TestSites}_{\text{wafer}} \right\rfloor$$

- Square peg in a round hole formula
- Chips per wafer less ones at edge



Chips Per Wafer

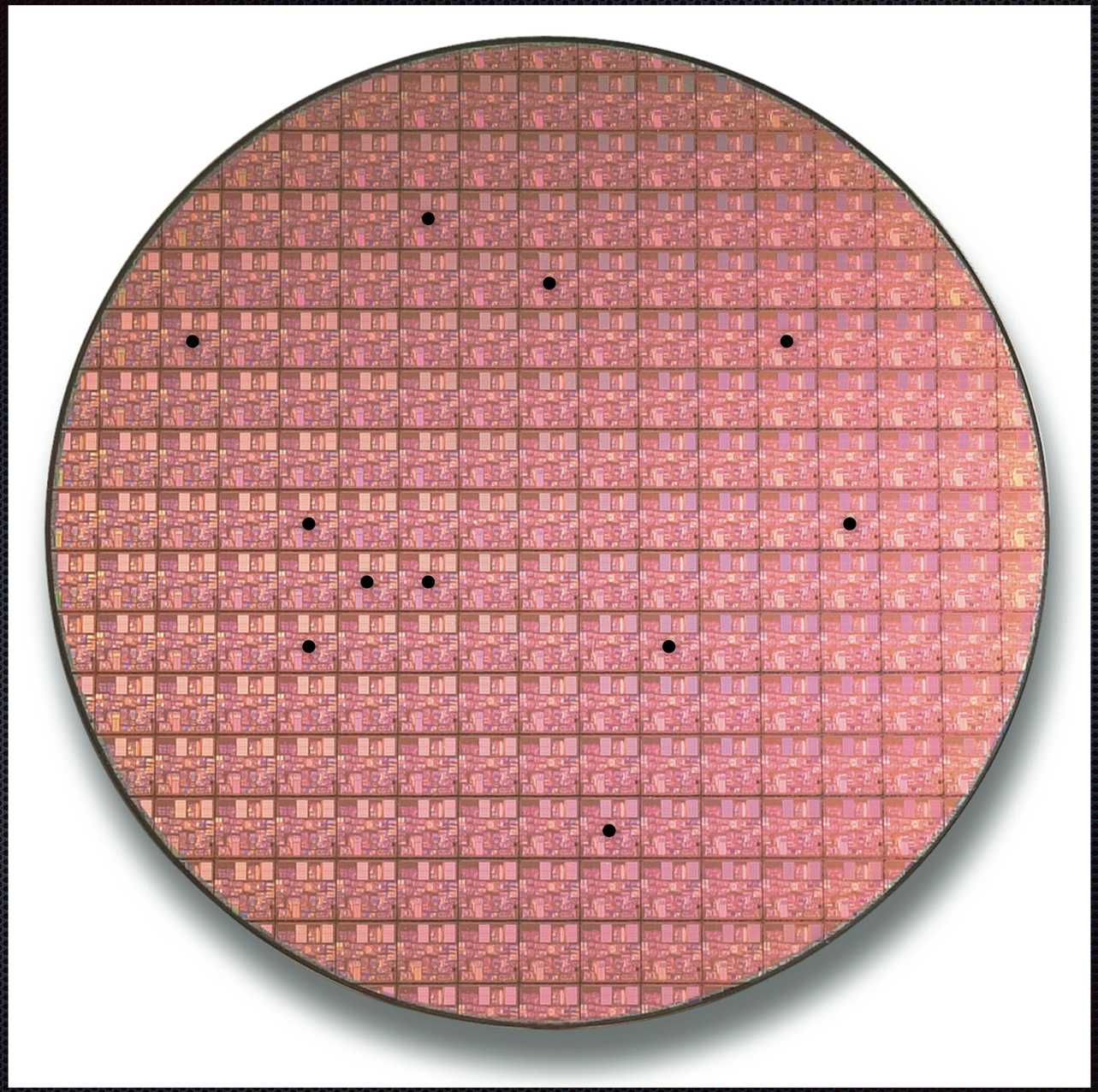
$$\text{Chips}_{\text{wafer}} = \left\lfloor \frac{\pi \times \left(\frac{\text{Diameter}_{\text{wafer}}}{2} \right)^2}{\text{Area}_{\text{chip}}} - \frac{\pi \times \text{Diameter}_{\text{wafer}}}{\sqrt{2 \times \text{Area}_{\text{chip}}}} - \text{TestSites}_{\text{wafer}} \right\rfloor$$

- Square peg in a round hole formula
- Chips per wafer less ones at edge

Die Yield

$$\text{Yield}_{\text{die}} = \text{Yield}_{\text{wafer}} \times \left(1 + \frac{\text{Defects}_{\text{unit-area}} \times \text{Area}_{\text{chip}}}{\text{P}}\right)^{-\text{P}}$$

- Fraction of good die. Depends on process.
- $\text{Defects}_{\text{unit-area}}$ and $\text{Yield}_{\text{wafer}}$ are given
- P is a process complexity factor



Die Yield

$$\text{Yield}_{\text{die}} = \text{Yield}_{\text{wafer}} \times \left(1 + \frac{\text{Defects}_{\text{unit-area}} \times \text{Area}_{\text{chip}}}{\text{P}}\right)^{-\text{P}}$$

- Fraction of good die. Depends on process.
- $\text{Defects}_{\text{unit-area}}$ and $\text{Yield}_{\text{wafer}}$ are given
- P is a process complexity factor

Test Cost

$$\text{Cost}_{\text{test}} = \frac{\text{Cost}_{\text{hour}} \times \text{Time}_{\text{test}}}{\text{Yield}_{\text{die}}}$$

- Testers are expensive to operate
- Cost must be amortized over good chips

Cost Practice

- $Cost_{wafer}$ = $7000
- $Diameter_{wafer}$ = 300mm
- $Area_{chip}$ = 13.5 X 19.6 mm
- $Defects_{unit\text{-}area}$= 0.5 per cm^2
- $Yield_{wafer}$ = 0.999
- $Yield_{final}$ = 0.97
- P = 4.3
- $Cost_{hour}$ = $1000
- $Time_{test}$ = 10 seconds
- $TestSites_{wafer}$ = 0
- Package Cost = $12

Cost at Different Scales

18-core

	45nm	32nm	10nm
Wafer cost	7000	7500	9000
Diameter (cm)	30	30	30
Chip area (cm^2)	2.65	1.33	6.8
Defects/unit (cm^2)	0.5	0.35	0.2
test sites	0	0	0
P factor	4.3	4.5	5.0
Wafer yield	0.999	0.999	0.999
Test time	10	10	6
Tester cost	1000	1200	5000
Package cost	12	12	12
Percent system cost	7	7	7
Final yield	0.97	0.97	0.98
Chips per wafer	225	473	78
Die yeld	0.315	0.641	0.3
Die cost	98.77	24.74	384.62
Test cost	8.82	5.2	27.78
Packaged cost	123.29	43.24	433.06
System cost	1761.29	617.71	6186.57

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

Reliability

Bad things happen...

Not So Solid State

- Early solid state circuits were considered to have nearly unlimited lifetime -- large devices
- Shrinking feature size makes wires and transistors so small that formerly minor effects become important
- Carriers migrate in and out of gate region under normal operation, and metal atoms can flow in wires
- Reliability becomes a factor in architecture

MTTF, etc.

- Mean time to failure -- average operating time
- Mean time to repair -- average downtime
- Mean time between failures -- MTTF + MTTR
- 1/MTTF = failure rate
- Availability = MTTF/MTBF

Failure Rate

- Multiplicative -- failure of any component causes system to fail
- Rate for any one component is usually small (1 per 100,000 hours = 11 years)
- More complex systems have many more components
- Failure rate can grow to unusable levels

Redundancy

- Components that operate in parallel or switch in on failure to avoid downtime
- MTTF depends on probability that one fails while the other is being repaired.
- Assuming independent failures, can estimate as:
- $MTTF^2/(2*MTTR)$
- Resulting MTTF can be much greater

Independence is Key

- Independence of failures isn’t guaranteed
- If there is a fire in the data center, redundant components are only independent if they are far enough apart
- Some components have distinct age profiles
- Power surges, overheating, malicious acts, etc., can all violate independence

Abishek Tiwari and Josep Torellas

Micro 2008

Facelift:

Hiding and Slowing Down Aging in Multicores

Submicron Technology

- Circuits age with use
- Aging takes the form of increased transistor switch time (slower active circuit elements)
- Engineers design circuits with a guard band of timing to allow for slowdown
  - Enables chips to have longer effective lifetimes
- Chips run slower because of the guard band

Aging Factors

- Negative Bias Temperature Instability (NBTI)
- Hot Carrier Injection (HCI)
- Metal migration

NBTI

- Apply a 0 to a PMOS transistor (negative bias)
- Hole insertion, together with heating, allows H atoms to diffuse away from the gate
- Si is thus ionized, which means higher threshold voltage V_t is needed to activate the gate
- With 1 input, H returns but not as quickly and not completely, so V_t gradually increases

HCI

- Energetic (hot) electrons enter the gate oxide
- Mainly affects NMOS
- Increases V_t because more holes are needed to induce a field through the carrier saturated oxide
- Higher temperature increases HCI rate, as does higher frequency switching

Metal Migration

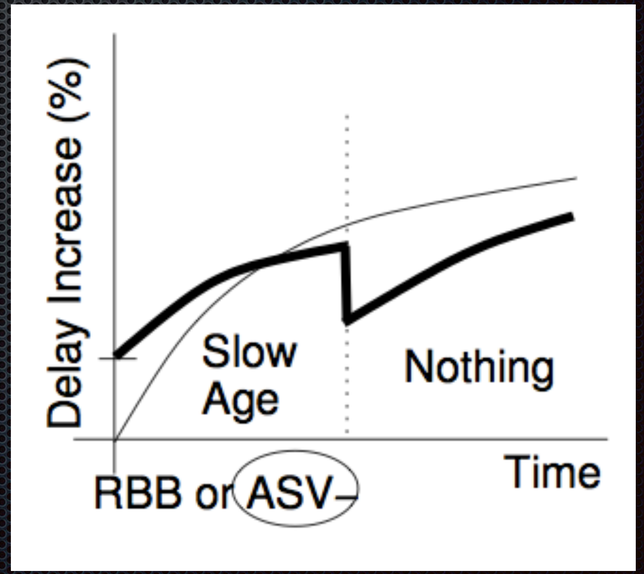
- Not addressed in this paper
- Tendency in narrow regions of wires for the metal atoms to be pulled along by the moving electrons
- Narrow regions have higher resistance, thus higher voltage, and fewer atomic bonds in their cross sections
- Positive feedback cycle, since migration further narrows the wire; eventually breaks

Multicore Technology

- Because of process variations in manufacturing, some cores start out faster than others
- With a single timing domain, all cores must run at the rate of the slowest
- Slower cores are slow because they have weaker transistors, thus they also age faster

Facelift Scheduling

- Run intense loads on fast cores, preserve slow cores with light loads
- Balancing rates of aging so they all hit slowdown limit at the same time
- Chips can run faster or be designed for shorter lifetimes
- Longer lifetime also possible, but not desirable from maker's perspective, and needed by few customers

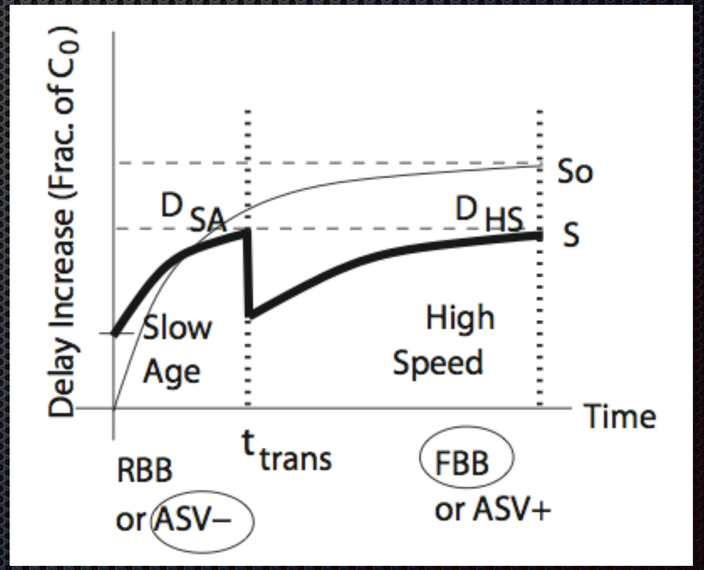


Facelift Aging Reduction

- Adaptive Supply Voltage (ASV)
- Adaptive Body Bias (ABB)
- Reducing voltage and reverse biasing slow aging
- Increased voltage and forward biasing increase speed
  - Can compensate for aging slowdown
- Rates of aging vary over time

When to Slow or Speed?

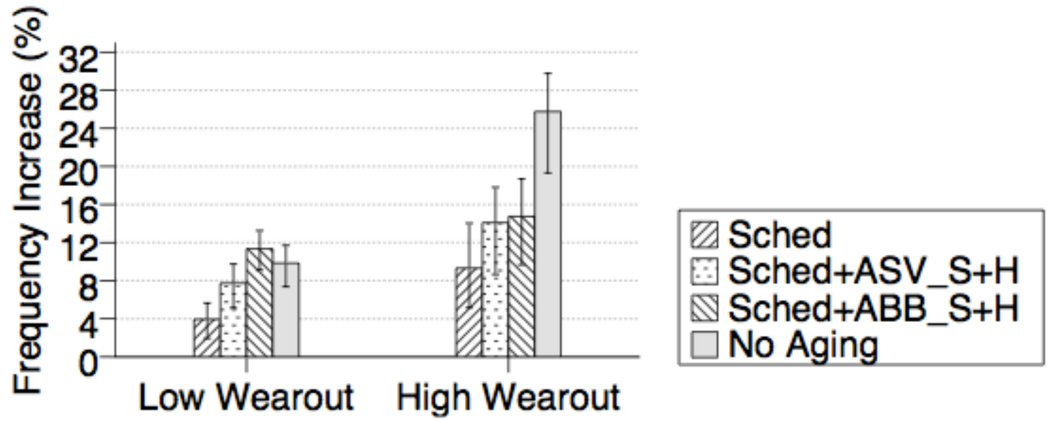
- Slowing aging early has most effect, and less impact late
- Increasing speed more effective near end of life
  - Early in life it significantly accelerates aging
  - Later it compensates more and affects aging rate less



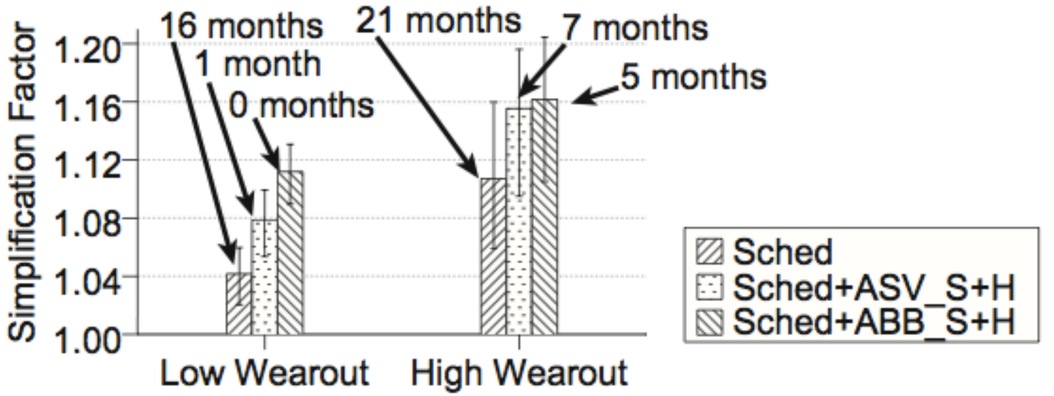
Benefits

- Can run at higher average clock rate
- Can spend less effort tuning the design to get optimal timing guard bands, which may also increase fabrication yield (more chips pass less stringent requirements)
- Can design for reduced service life and still get full life

Results



(a)



(b)

Figure 13: Frequency increases (a) and Simplification factors (b) enabled by the combination of aging-hiding and aging-slowing techniques.

Discussion

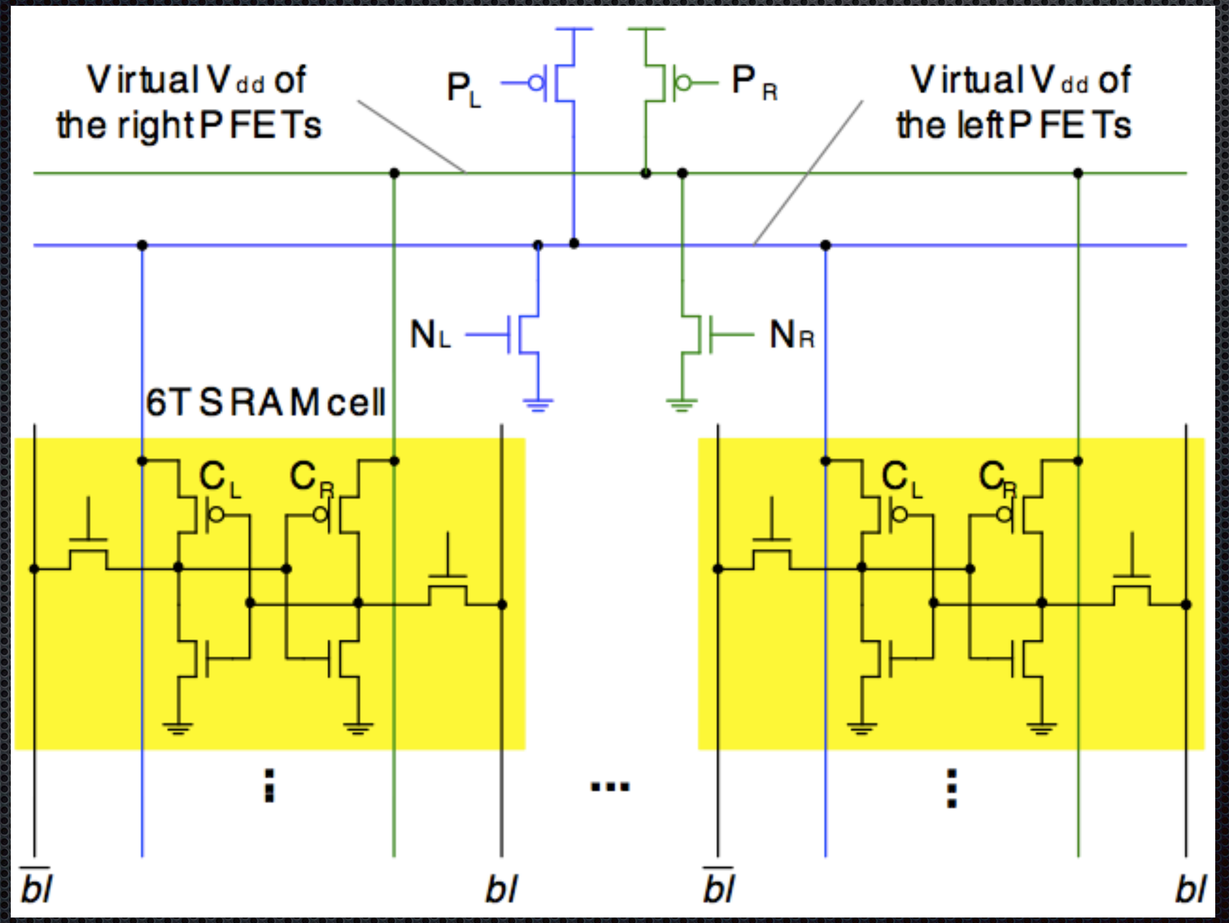
Jeonghee Shin ISCA 2008

A Proactive Wearout Recovery Approach for Exploiting Microarchitectural Redundancy to Extend Cache SRAM Lifetime

NBTI Again

- Negative Bias Temperature Instability
- Affects PMOS transistors in SRAM, causing wearout
- H atoms migrate out with high voltage and temp
- Tend to return as voltage and temp are lowered, but not fully
- Add ability to reverse the field and attract them back

Revised SRAM Cells



Circuit Description

- The extra transistors enable reverse biasing
- They can be shared among multiple cells
- The power lines are already present
- At the cell level, the extra area is minimal (< 1%)

The Catch

- When SRAM cells are in recovery mode, they cannot be used to store data for some period
- The SRAM of interest is the on-chip cache
- We don’t want to stall the processor for recovery
- Need an architectural solution to enable use of recovery

Invalidation Approach

- Invalidate the line before it enters recovery
- Cache flushes to main level below, and invalidates level above (to preserve inclusion)
- After recovery, data is restored via normal miss process

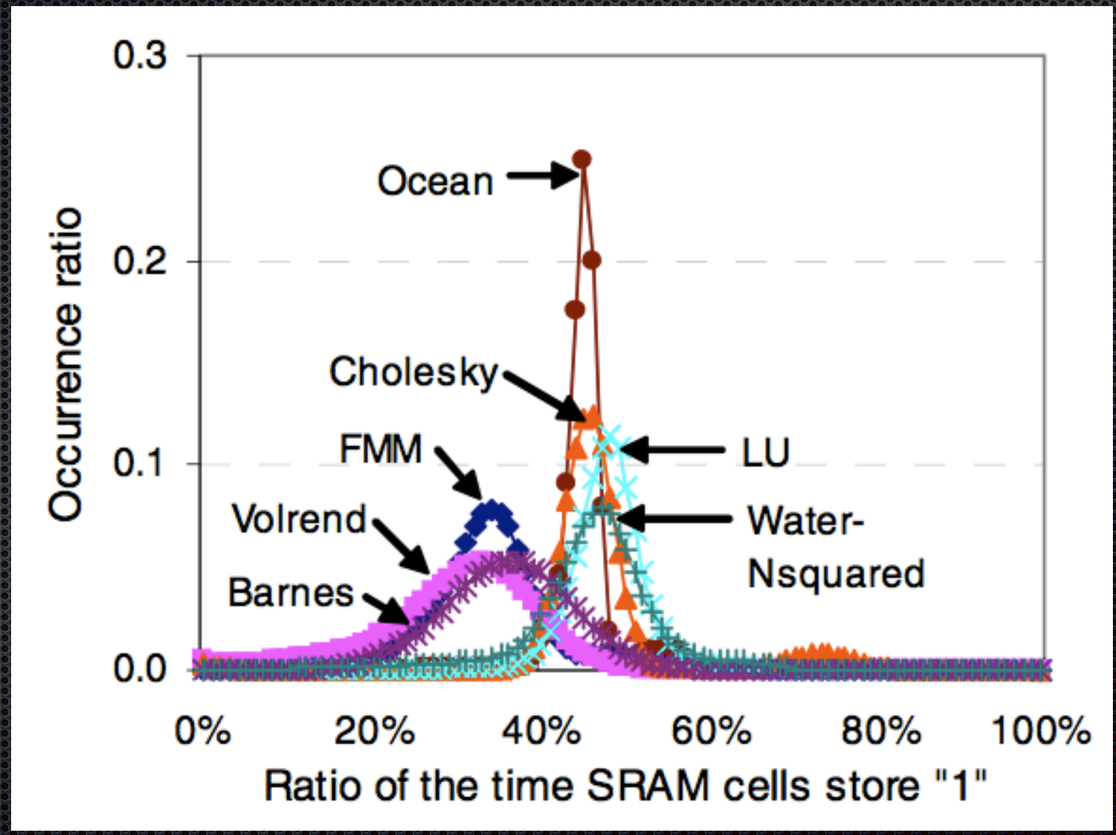
Extra Bank

- Caches are already arranged in banks
- Add one more (sometimes done to improve chip yield)
- Lines that are dirty, shared, or exclusive have to be copied into the spare bank
- Accesses have to be redirected to the spare while recovery is being done
- Data gets moved back after recovery

Workload Balancing

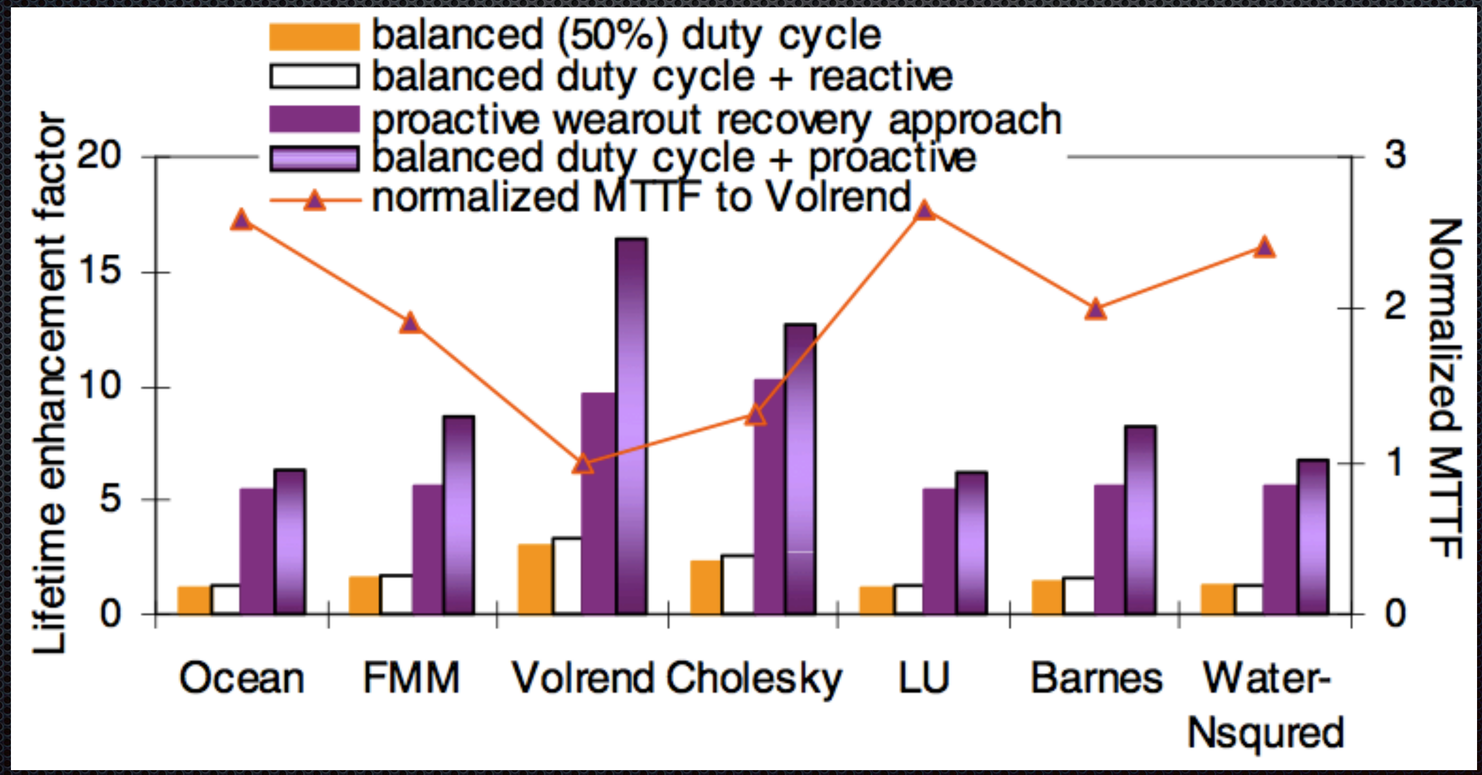
- Some applications are more wearing than others
- volrend 2.6 X more than LU
- Due to balance of 1/0 writes
- Artificially balancing gives 3X longer life
- Increase lifetime for swapping in an extra bank after one fails is negligible

Balance of 1/0 load



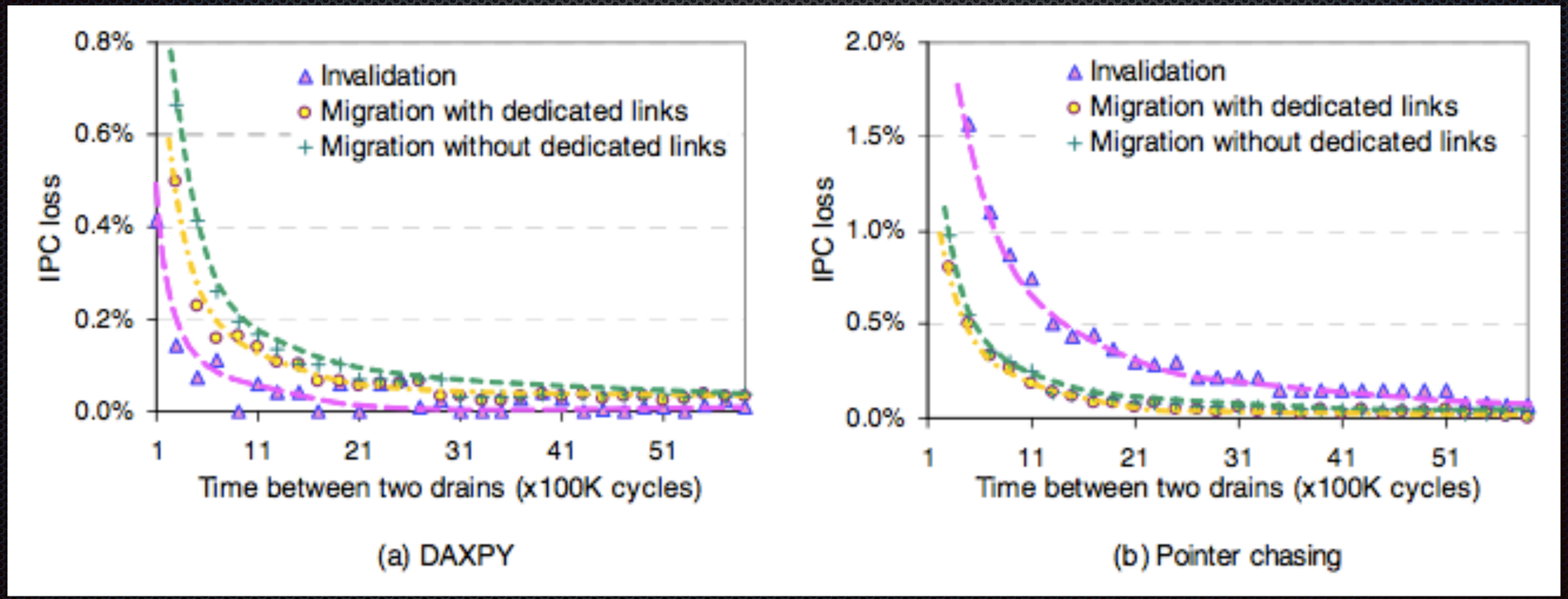
Recovery Impact

- 5.5 to 10.2 X for recovery cycles alone
- With balancing, 6.0 to 16.5X



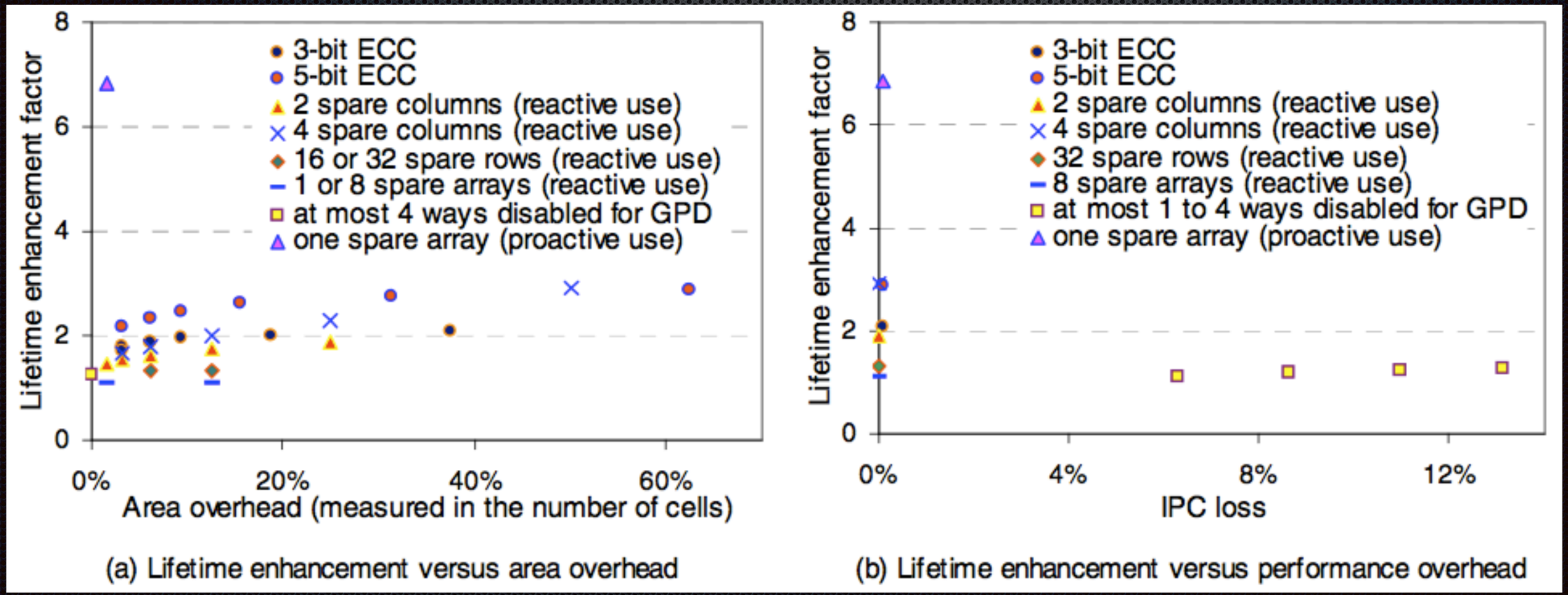
Performance Impact

- Less than 1% for 100K cycle interval



Versus area and time

- Compared with other approaches



Discussion

Benjamin Lee ISCA 2009

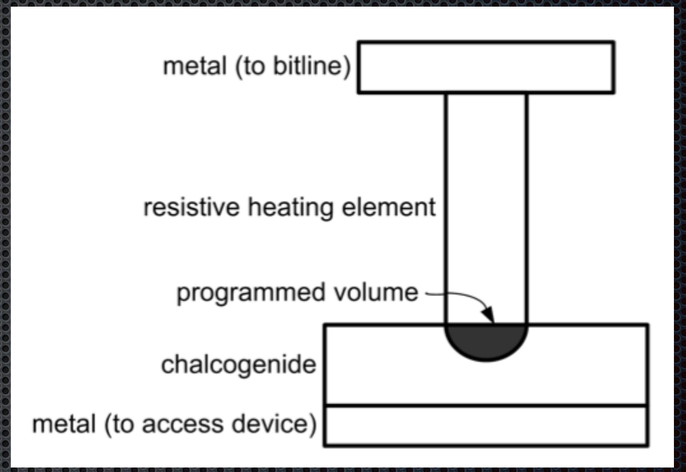
Architecting Phase Change Memory as a Scalable DRAM Alternative

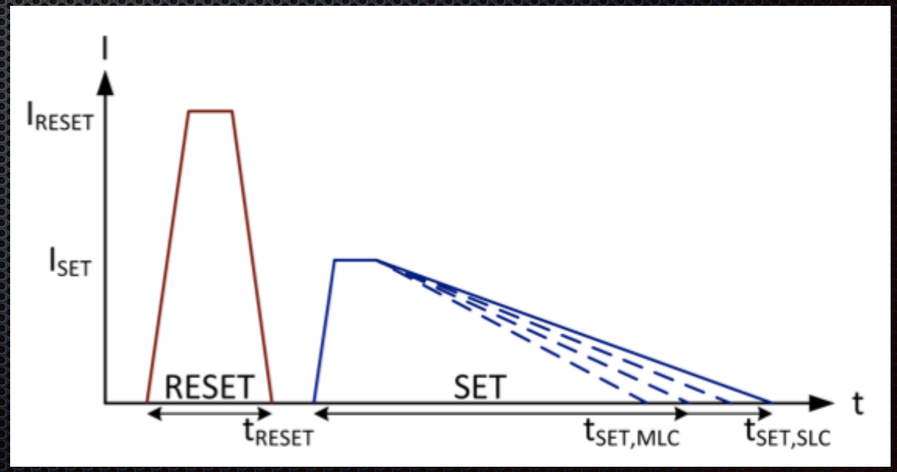
DRAM Scaling Limit (?)

- No known way to scale below 40nm (at time of article)
- Assuming this is true, time to look for alternative
- Phase Change Memory (PCM)
- Theoretically scales to 9nm

PCM operation

- Heat a spot
- Cool quickly -- amorphous or polycrystaline
- Cool slowly -- crystalizes
- Different levels of resistance
- Nonvolatile, no power when idle, no refresh, nondestructive read

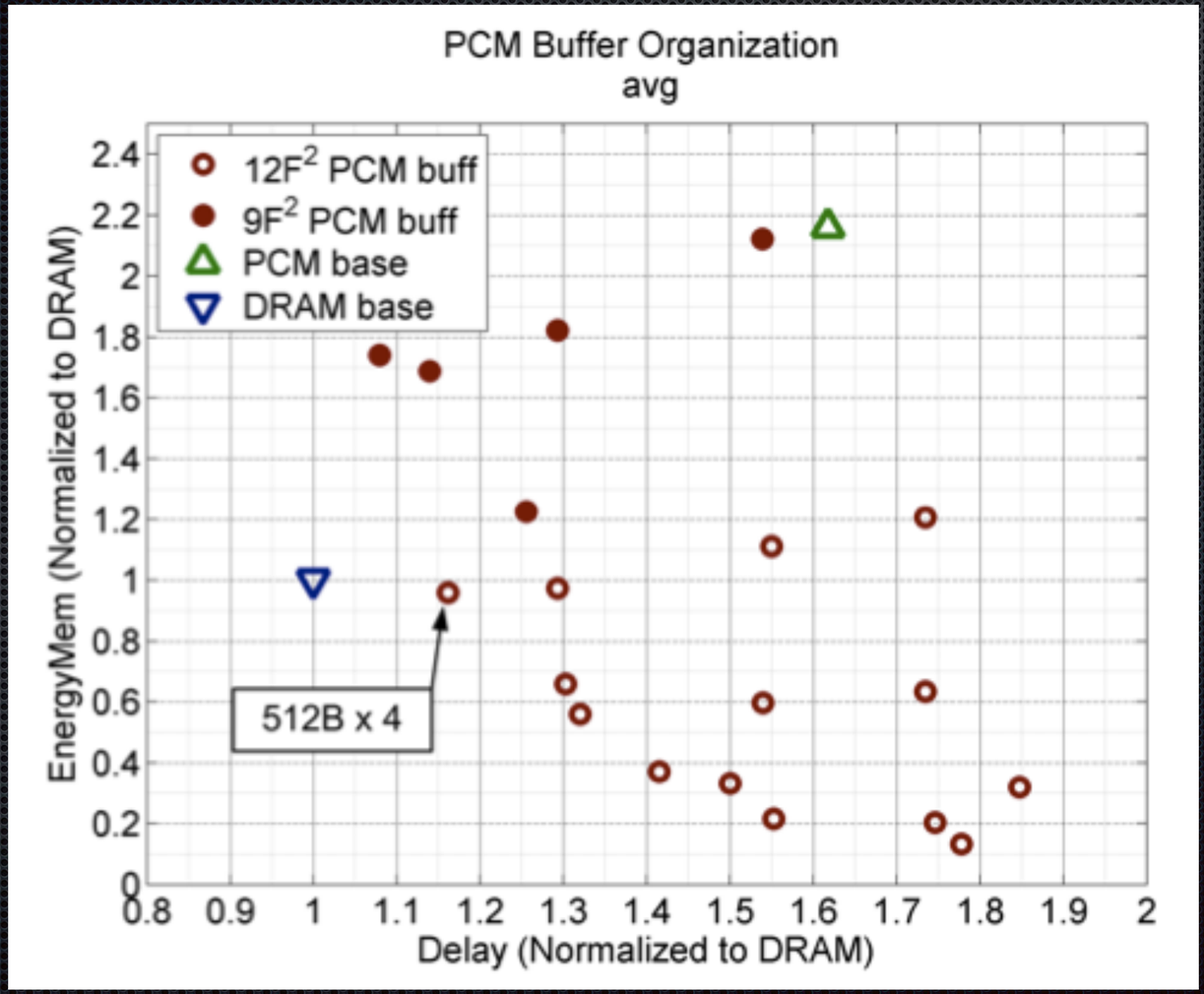




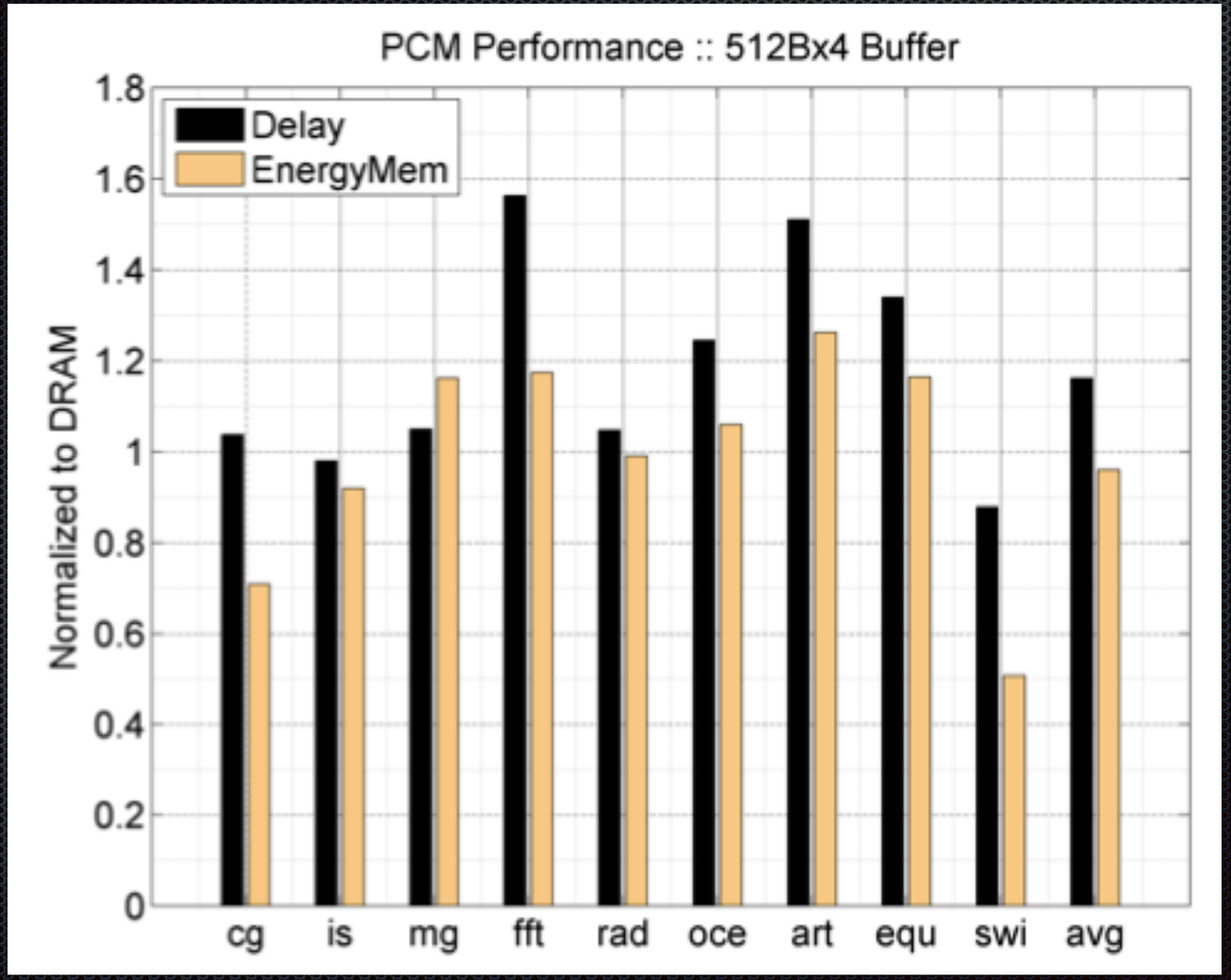
Characteristics

- Slow 1 writes (150ns), but 0 writes only 40ns
- Wearout, estimated as 100,000,000 cycles
- Read latency estimated as 48ns (4.4 x DRAM)
- Write latency 12x greater than DRAM
- Allows multiple levels (2 bits/cell) but reduces life

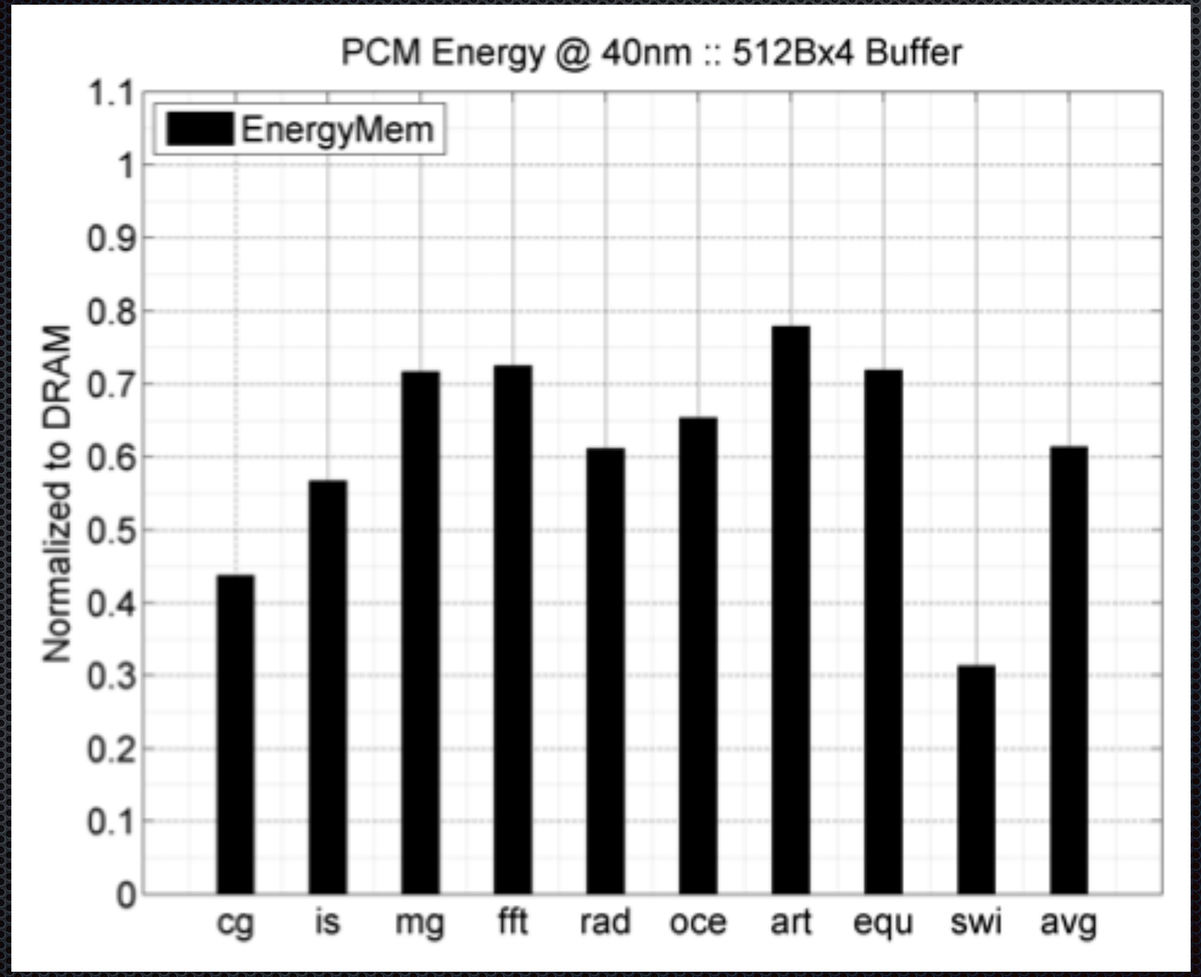
Buffer Size Analysis



Performance with Buffer

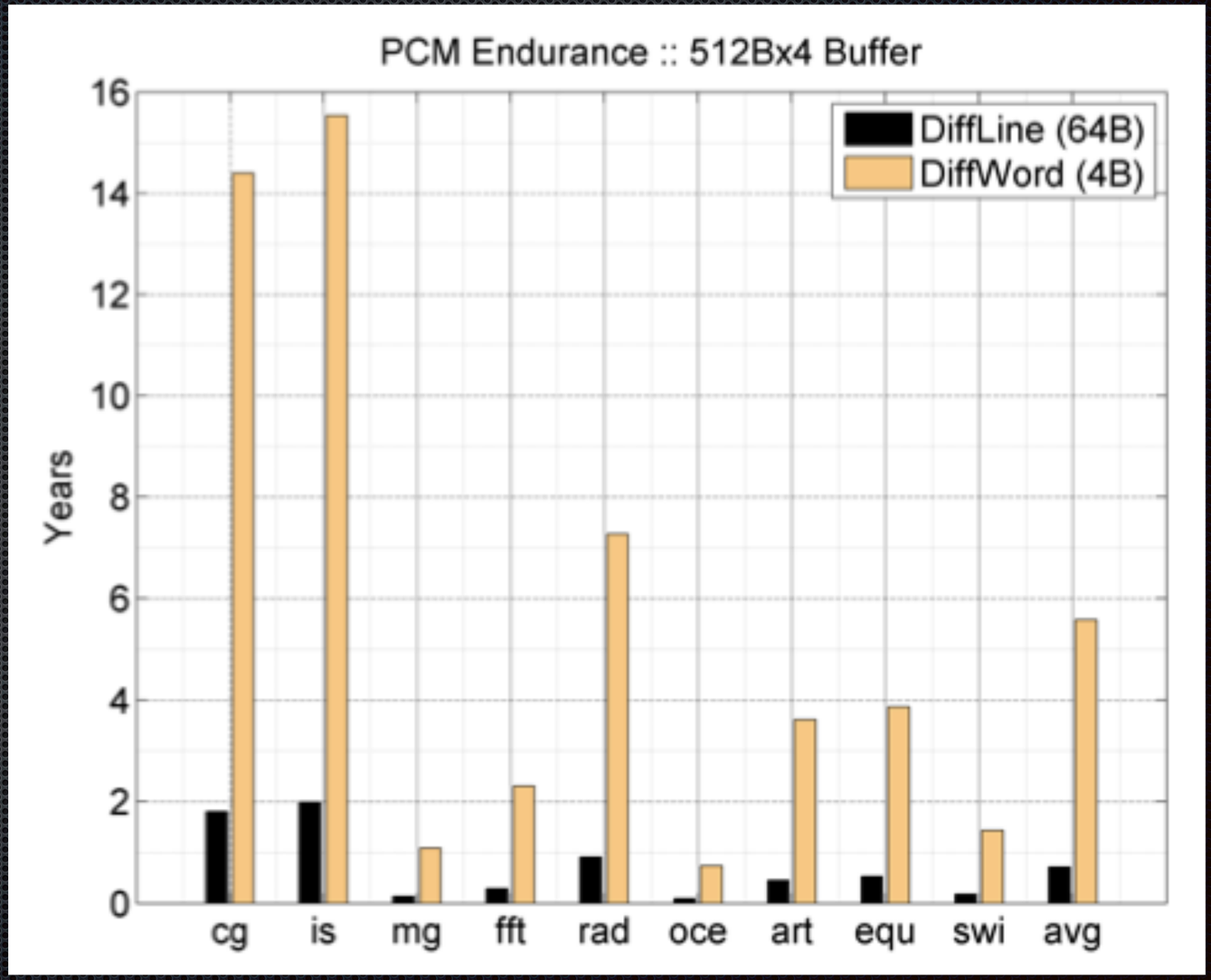


Energy with Buffer



Endurance

- Changing by line versus by word



Discussion

Future?

- DRAM shifts to speed emphasis to outpace PCM
- L3/L4 Fast DRAM layer ~1GB, with active sharing
- ~16 GB Fast PCM main memory, for idle pages
- Slow PCM or Flash replaces disk for active data
- Disk for warehousing less active data
- Integrated nonvolatile RAM/file system avoids paging