Technology Challenges

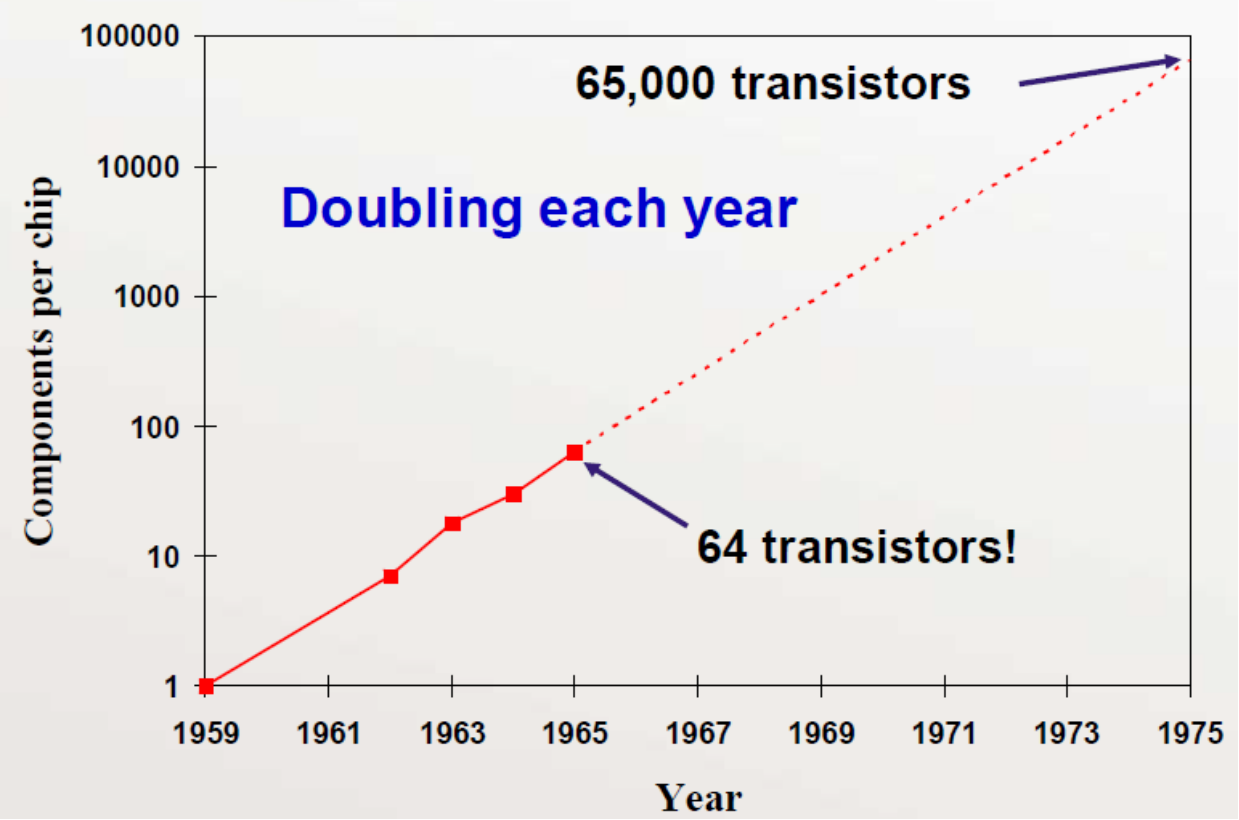
ECE/CS 752 Fall 2017

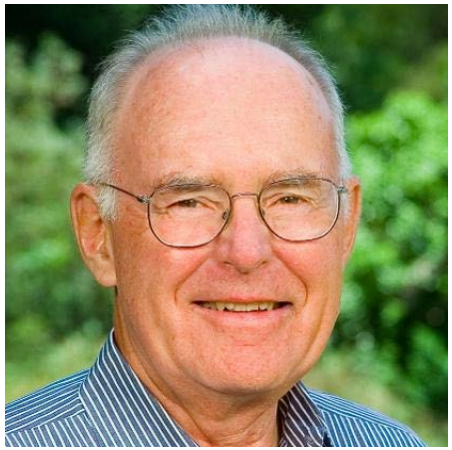
Prof. Mikko H. Lipasti
University of Wisconsin-Madison

Readings

- Read on your own:
  - Shekhar Borkar, Designing Reliable Systems from Unreliable Components: The Challenges of Transistor Variability and Degradation, IEEE Micro 2005, November/December 2005 (Vol. 25, No. 6) pp. 10-16.
  - 2015 ITRS Roadmap -- Executive Summary. Read sections 1, 5, 6, 8, 9, and skim the rest.
- Review by Wed 9/13/2017:
  - Jacobson, H, et al., “Stretching the limits of clock-gating efficiency in server-class processors,” in Proceedings of HPCA-11, 2005.

Moore's Law (1965)





G. E. Moore, "Cramming More Components onto Integrated Circuits," *Electronics* Vol. 38, No. 8 (Apr. 19, 1965) pp. 114-117.

Dennard Scaling (1974)

Device/Circuit Parameter	Scaling Factor
Device dimension/thickness	$1/\lambda$
Doping Concentration	λ
Voltage	$1/\lambda$
Current	$1/\lambda$
Capacitance	$1/\lambda$
Delay time	$1/\lambda$
Transistor power	$1/\lambda^2$
Power Density	1

R. Dennard et. al, "Design of ion-implanted MOSFET's with very small physical dimensions" *IEEE Journal of Solid State Circuits. SC–9 (5)*

End of Dennard Scaling

- Everything was great! No tradeoffs at all…
  - Density doubling every two years (Moore's law)
    - Feature size
    - Device density
  - Device switching speed improves 30-40%/generation
  - Supply & threshold voltages decrease ($V_{dd,}$ V_{th})
- This ended around 2000
  - Now, feature size, device density scaling continues
    - Roadmap well below 10nm generation
  - Switching speed improves ~20%/generation or less
  - Voltage scaling has tapered off
    - SRAM cell stability becomes an issue at ~0.7V V_{dd}
- Still cheaper (or denser) but power-limited

Summary of Challenges

Late 20th Century	The New Reality
Moore's Law – 2x transistors/chip every 18-24 months	Transistor count still 2x every 18-24 months, but see below
Dennard Scaling – near-constant power/chip	Gone. Not viable for power/chip to double (with 2x transistors/chip growth)
The modest levels of transistor unreliability easily hidden (e.g., via ECC)	Transistor reliability worsening, no longer easy to hide
Focus on computation over communication	Restricted inter-chip, inter-device, inter-machine communication; communication more expensive than communication
One-time (non-recurring engineering) costs growing, but amortizable for mass-market parts	Expensive to design, verify, fabricate, and test, especially for specialized-market platforms

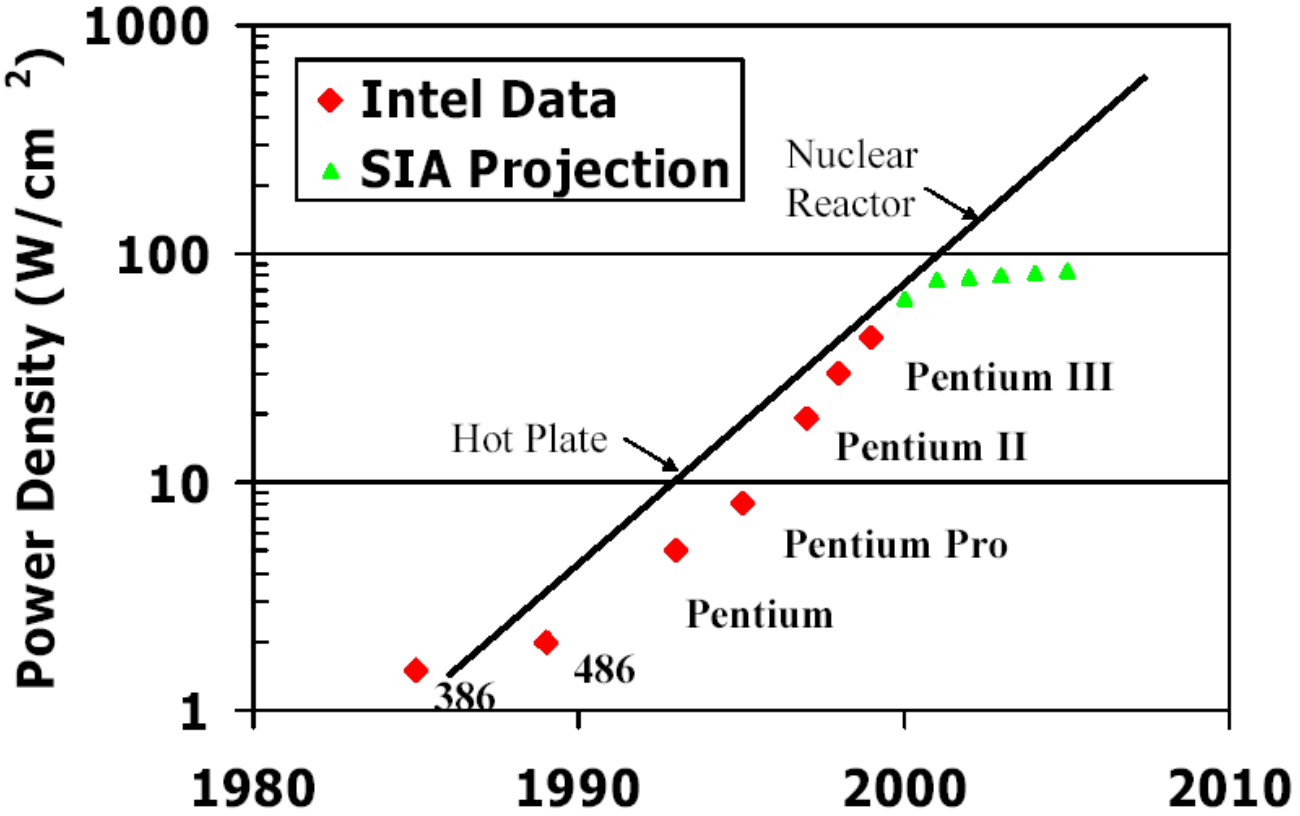
From "21st Century Computer Architecture: A community white paper," Computing Research Assocation (CRA), 2012

Solutions?

THE UNIVERSITY of WISCONSIN MADISON

- New markets will drive demand and need for innovation
  - Datacenters
  - IoT
  - Mobile
- Incremental (one-time) improvements
  - Copper, high-K dielectric, FinFETs, …
  - Packaging: 2.1D, 2.5D, 3D
- Beyond CMOS
  - Carbon nanotubes, spin FET,
- “More than Moore”
  - Analog, RF, accelerators, non-Von Neumann, …

Power Density [Hu et al, MICRO '03 tutorial]



- Power density increasing exponentially
  - Power delivery, packaging, thermal implications
  - Thermal effects on leakage, delay, reliability, etc.

Dynamic Power

$$P_{dyn} \approx kCV^2Af$$

- Aka AC power, switching power
- Static CMOS: current flows when transistors turn on/off
  - Combinational logic evaluates
  - Sequential logic (flip-flop, latch) captures new value (clock edge)
- Terms
  - C: capacitance of circuit (wire length, no. & size of transistors)
  - V: supply voltage
  - A: activity factor
  - f: frequency
- Moore's Law: which terms increase, which decrease?
  - Voltage scaling has been saving our bacon!

Reducing Dynamic Power

- Reduce capacitance
  - Simpler, smaller design
  - Reduced IPC
- Reduce activity
  - Smarter design
  - Reduced IPC
- Reduce frequency
  - Often in conjunction with reduced voltage
- Reduce voltage
  - Biggest hammer due to quadratic effect, widely employed
  - Can be static (binning/sorting of parts), and/or
  - Dynamic (power modes)
    - E.g. Transmeta Long Run, AMD PowerNow, Intel Speedstep

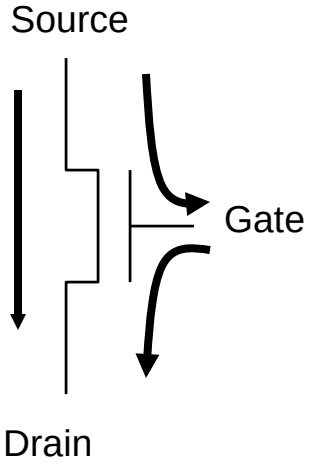
Frequency/Voltage relationship

- Lower voltage implies lower frequency
  - Lower V_{th} increases delay to sense/latch 0/1
- Conversely, higher voltage enables higher frequency
  - Overclocking
- Sorting/binning and setting various V_{dd} & V_{th}
  - Characterize device, circuit, chip under varying stress conditions
  - Black art – very empirical & closely guarded trade secret
  - Implications on reliability
    - Safety margins, product lifetime
    - This is why *overclocking* is possible

Frequency/Voltage Scaling

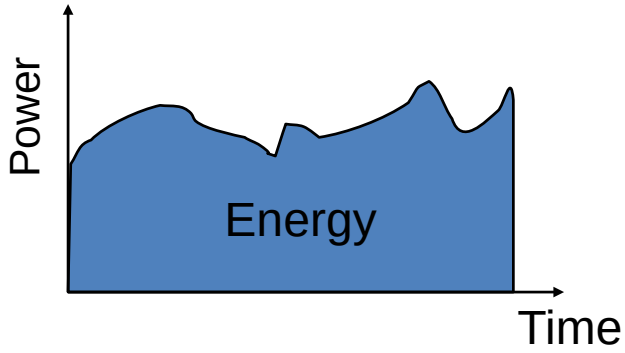
- Voltage/frequency scaling rule of thumb:
  - +/- 1% performance buys -/+ 3% power (3:1 rule)
- Hence, any power-saving technique that saves less than 3x power over performance loss is uninteresting
- Example 1:
  - New technique saves 12% power
  - However, performance degrades 5%
  - Useless, since 12 < 3 x 5
  - Instead, reduce f by 5% (also V), and get 15% power savings
- Example 2:
  - New technique saves 5% power
  - Performance degrades 1%
  - Useful, since 5 > 3 x 1
- Does this rule always hold?

Leakage Power (Static/DC)



- Transistors aren't perfect on/off switches
- Even in static CMOS, transistors leak
  - Channel (source/drain) leakage
  - Gate leakage through insulator
    - High-K dielectric replacing SiO_2 helps
- Leakage compounded by
  - Low threshold voltage
    - Low V_{th} => fast switching, more leakage
    - High V_{th} => slow switching, less leakage
  - Higher temperature
    - Temperature increases with power
    - Power increases with C, V^2, A, f
- Rough approximation: leakage proportional to area
  - Transistors aren't free, unless they're turned off
- Could be a huge problem in future technologies
  - Estimates are 40%-50% of total power

Power vs. Energy



- Energy: integral of power (area under the curve)
  - Energy & power driven by different design constraints
- Power issues:
  - Power delivery (supply current @ right voltage)
  - Thermal (don't fry the chip or make user uncomfortable)
  - Reliability effects (chip lifetime)
- Energy issues:
  - Limited energy capacity (battery)
  - Efficiency (work per unit energy)
- Different usage models drive tradeoffs

Power vs. Energy

- With constant time base, two are “equivalent”
  - 10% reduction in power => 10% reduction in energy
- Once time changes, must treat as separate metrics
  - E.g. reduce frequency to save power => reduce performance => increase time to completion => consume more energy (perhaps)
- Metric: energy-delay product per unit of work
  - Tries to capture both effects, accounts for quadratic savings from DVS
  - Others advocate energy-delay^2 (accounts for cubic effect)
  - Best to consider all
    - Plot performance (time), energy, ed, ed^2

Usage Models

- Thermally limited => dynamic power dominates
  - Max power ("power virus" contest at Intel)
  - Must deliver adequate power (or live within budget)
  - Must remove heat
    - From chip, from case, room, building, **pocket**
  - Chip *hot spots* cause problems
- Efficiency => dynamic & static power matter
  - E.g. energy per DVD frame
  - Analogy: cell-phone "talk time"
- Longevity => static power dominates
  - Minimum power while still "awake"
  - Cellphone "standby" time
  - Laptop still responds quickly
    - Not suspend/hibernate
  - "Power state" management very important
    - Speedstep, PowerNow, LongRun

Worst Case

Average Case

Best Case

Circuit-Level Techniques

- Multiple voltages
  - Realize non-critical circuits with slower transistors
  - Voltage islands: Vdd and Vth are lower
    - Problem: supplying multiple Vdd
  - MTCMOS: only Vth is lower
- Multiple frequencies
  - Globally Asynchronous Locally Synchronous (GALS)
- Exploiting safety margins
  - Average case vs. worst case design
  - Razor latch [UMichigan]:
    - Sample latch input twice, then compare, recover
- Body biasing
  - Reduce leakage by adapting V_{th}

Architectural Techniques

- Multicore chips (later)
- Clock gating (dynamic power)
  - 70% of dynamic power in IBM Power5 [Jacobson et al., HPCA 04]
  - Inhibit clock for
    - Functional block
    - Pipeline stage
    - Pipeline register (sub-stage)
  - Widely used in real designs today
  - Control overhead, timing complexity (violates fully synchronous design rules)
- Power gating (leakage power)
  - Sleep transistors cut off V_{dd} or ground path
  - Apply to FU, cache subarray, even entire core in CMP

Architectural Techniques

- Cache reconfiguration (leakage power)
  - Not all applications or phases require full L1 cache capacity
  - Power gate portions of cache memory
  - State-preservation
    - Flush/refill (non-state preserving) [Powell et al., ISLPED 00]
    - Drowsy cache (state preserving) [Flautner, Kim et al., ISCA 02]
  - Complicates a critical path (L1 cache access)
  - Does not apply to lower level caches
    - High V_{th} (slower) transistors already prevent leakage

Architectural Techniques

- Heterogeneous cores [Kumar et al., MICRO-36]
  - Prior-generation simple core consumes small fraction of die area
  - Use simple core to run low-ILP workloads
  - E.g. ARM's big.LITTLE
- Configurable cores: dynamically vary ILP vs. power
- Filter caches (dynamic power)
  - Many references are required for correctness but result in misses
    - External snoops [Jetty, HPCA '01]
    - Load/store alias checks [Sethumadhavan et al., MICRO '03]
  - Filter caches summarize cache contents (e.g. Bloom filter)
  - Much smaller filter cache lookup avoids lookup in large/power-hungry structure
- And many others…check proceedings of
  - ISLPED, MICRO, ISCA, HPCA, ASPLOS, PACT

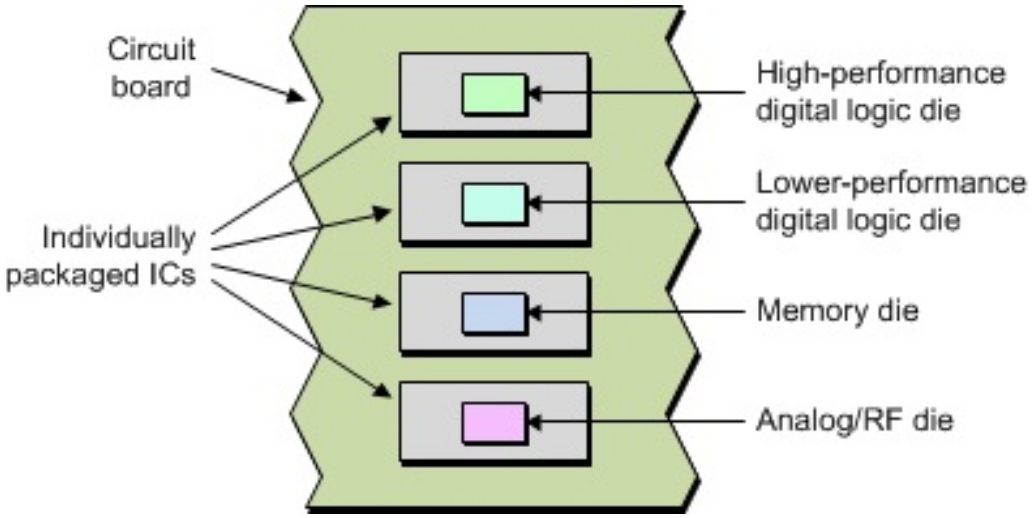
Variability

- Shrinking device dimensions lead to sensitivity to minor processing variations

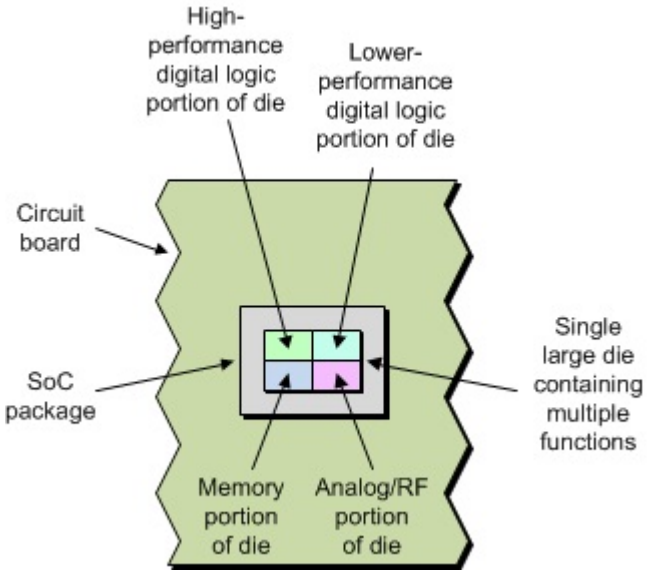
  "*No two transistors are the same*"

  – Die-to-die variations
    - Across multiple die on same wafer, across wafers
  – Within-die variations
    - Systematic and random
    - E.g. line edge roughness due to sub-wavelength lithography or dopant variations (~10 molecules)
  – Dynamic variations
    - E.g. temperature-induced variability (hot spots)

2D Packaging



Board level

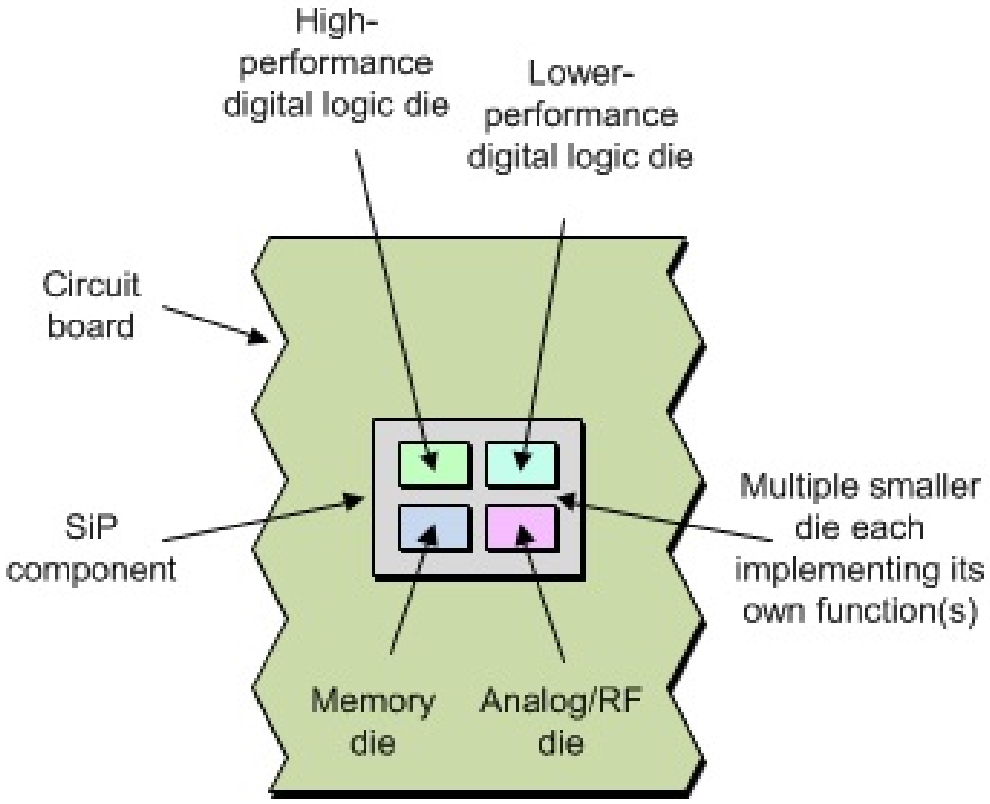


System on Chip

[M. Maxfield, “2D vs. 2.5D vs. 3D ICs 101,” EE Times, April 2012]

Conventional packaging approaches

2D Packaging

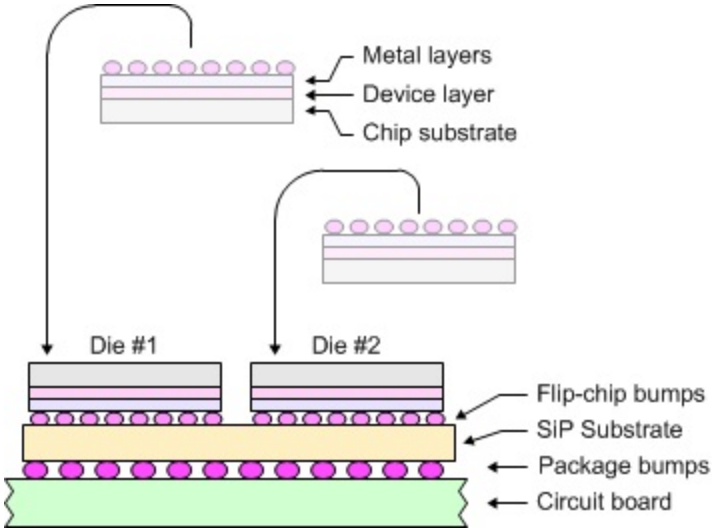


[M. Maxfield, "2D vs. 2.5D vs. 3D ICs 101," EE Times, April 2012]

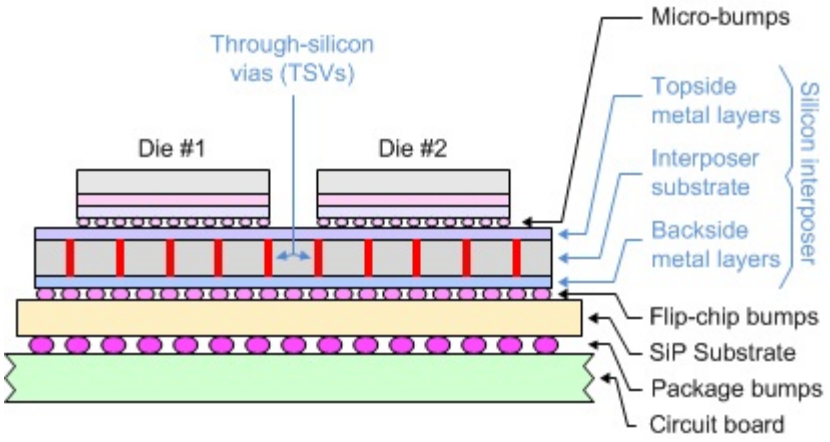
Move toward System in Package (SIP)

- PCB, ceramic, semiconductor substrates

2.5D Packaging



2D Packaging

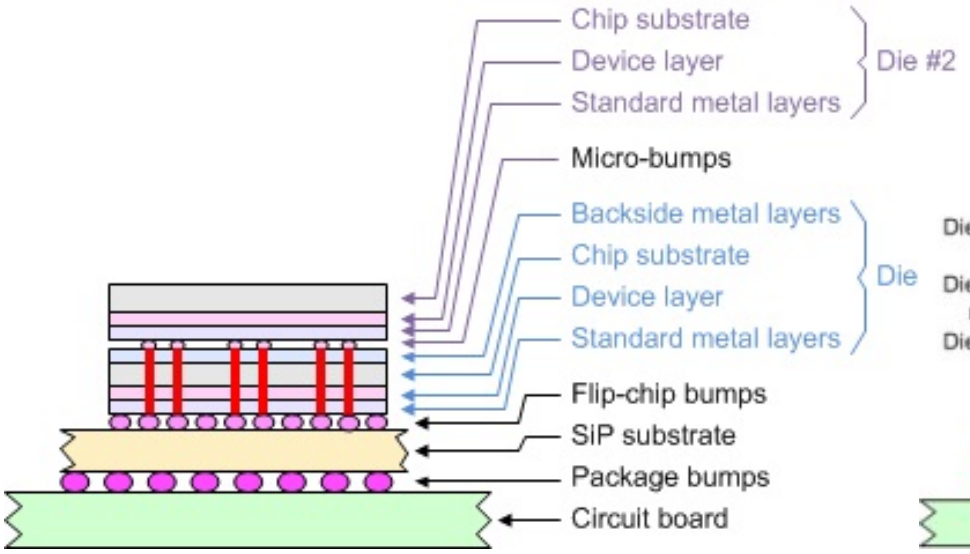


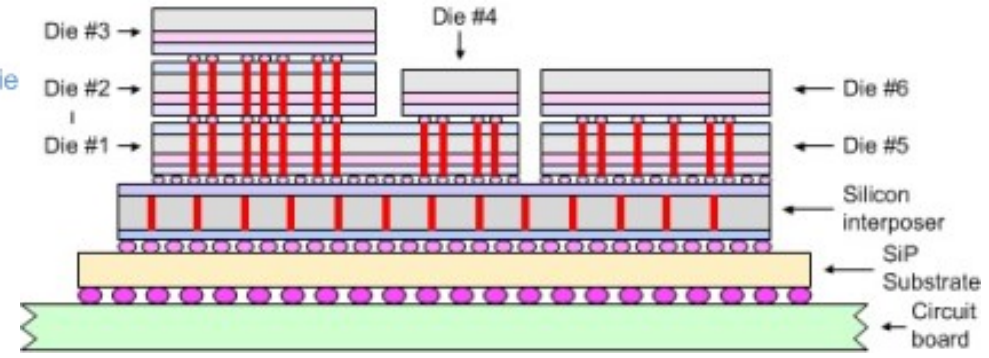
2.5D Packaging

[M. Maxfield, "2D vs. 2.5D vs. 3D ICs 101," EE Times, April 2012]

2.5D uses silicon interposer, through-silicon vias (TSV)

3D Packaging





3D Homogeneous

3D Heterogeneous

[M. Maxfield, "2D vs. 2.5D vs. 3D ICs 101," EE Times, April 2012]

3D uses through-silicon vias (TSV) and/or interposer

Packaging Discussion

- Heterogeneous integration
  - RF, analog (PHY), FG/PCM/ReRAM, photonics
- Cost
- Silicon yield
- Bandwidth, esp. interposer
- Thermals
- It's real!
  - DRAM: HMC, HBM
  - FPGAs
  - GPUs: AMD, NVIDIA
  - CPUs: AMD Zen, EPYC

Summary

- Technology scaling, Moore vs. Dennard
- Power: dynamic, static
  - CMOS scaling trends
  - Power vs. Energy
  - Dynamic power vs. leakage power
- Usage Models: thermal, efficiency, longevity
- Circuit Techniques
- Architectural Techniques
- Variability
- Packaging