

ESE532: System-on-a-Chip Architecture

Day 23: November 18, 2019
Estimating Chip Area and Costs



Today

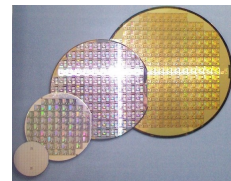
- Chip Costs from Area
- Chip Area
 - IO
 - Interconnect – Rent's Rule
 - Infrastructure
- Some Areas
 - CACTI – for modeling memories

Message

- First order:
 - Chip cost proportional to Area
 - Area = Sum(Area(Components))
- But appreciate the simplification:
 - Yield makes cost superlinear in area
 - I/O, Interconnect, infrastructure
 - Can make Area > Sum(Area(Components))

Wafer Cost

- Incremental cost of producing a silicon wafer is fixed for a given technology
 - Independent of the specific design
 - E.g. \$4,000
- Can fill wafer with copies of chip



By German Wikipediabiatch, original upload 7. Okt 2004 by Stahlkocher de:Bild:Wafer 2 Zoll bis 8 Zoll.jpg, CC BY-SA 3.0, <https://commons.wikimedia.org/w/index.php?curid=928106>

16nm Wafer Costs

	16/14nm FinFET		14nm FD SOI	
	\$	%	\$	%
Depreciation	2,303.94	58.79	1,972.26	53.24
Equipment maintenance	581.42	14.84	445.32	12.02
Direct labor	64.78	1.65	45.52	1.23
Indirect labor	238.76	6.09	208.36	5.62
Facilities	232.47	5.93	189.37	5.11
Wafer cost	99.93	2.55	475.00	12.82
Consumables	359.28	9.17	331.47	8.95
Monitor wafers	38.62	0.99	37.24	1.01
TOTAL Unyielded wafer cost	3,919.20	100.00	3,704.54	100.00
Line yield (%)	96.03	--	97.96	--
TOTAL Yielded wafer cost	4,081.22	--	3,781.69	--

Source: https://www.eetimes.com/author.asp?section_id=36&doc_id=1329887

Preclass 1

- Rough cost per mm of silicon?
 - \$4000 for 300mm wafer

Implication

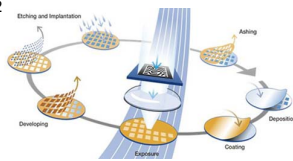
- Raw silicon die cost is roughly proportional to area
 - Larger the die, the fewer we get on the wafer

...but

- Limits to how big we can make chips
 - Manufactures are prepared to create
 - Can be reliably manufactured
- ...and how small we can make chips
 - I/O pads
 - Cutting/handling/markings

Imaging

- Limit to how large optical imaging supports
- Reticle – imagable region for photo lithography
 - Around 600mm^2



Source: <https://www.asml.com/the-asml-exposure-apparatus-is-the-most-expensive-and-complex-step-in-the-chip-fabrication-process-what-is-involved-in-the-lithography-business/ja/2314576d-44709>

Yield

- Chips won't be manufactured perfectly
 - Dust particles can impact imaging
 - Manufacturing processes are statistical
- If chips must be defect-free,
 - larger chips are more likely to have defects than smaller chips

Simple Yield Model

- Probability of a region being perfect
 - E.g. probability of one sq. mm being defect-free
- Chip yields if its entire area is defect free

(look at how to tolerate defects in a couple of weeks)

Chip Yield

- P = defect-free probability per sq. mm
- What is probability a chip of A sq. mm yields (symbolic) ?

Preclass 2

- P=0.99
- Probability of yield for
 - 10 mm², 50 mm², 100 mm², 500 mm²

Yielded Die

- For a yield rate, Y, how many raw die need to manufacture per yielded die?

Preclass 3

- P=0.99
- Die cost for:
 - 10 mm², 50 mm², 100 mm², 500 mm²

Yielded Die Cost

$$\text{Cost} = \frac{\text{Raw}}{\text{Yield}} = \frac{A^* \text{Cost/mm}^2}{P^A}$$

Yielded Die Cost

$$\text{Cost} = \frac{\text{Raw}}{\text{Yield}} = \frac{A^* \text{Cost/mm}^2}{P^A}$$

- Ultimately exponential in Area
- Means
 - Expensive above knee in exponential curve
 - Close to linear below knee in curve
- E.g.
 - Below P^A=0.5
 - effect of Yield term is less than 2

Design Dependent Cost

- P can be design dependent
 - More aggressive designs have higher defect rates
 - Can tune design to ease manufacturing
- Contrast with point that wafer manufacture cost independent of design

Slightly Fuller Story

- Chip cost = die + test + package

Test

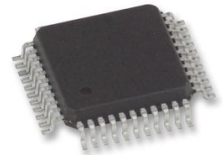
- Testing costs proportional to test time
 - Time on expensive test unit
 - Depends on complexity of tests need to run
 - Can motivate spending silicon area on on-chip test structures to reduce
- Can dominate on small chips or complex testing

Packaging

- Pay for density and performance

Plastic Packages

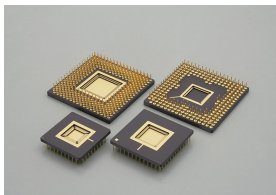
- Simple plastic packages cheap
 - Limited number of pins
 - Limited to perimeter
 - Limited heat removal (few Watts)
 - Can be large (due to pins)
 - Higher inductance on pins



http://wiki.electrooons.com/doku.php?id=ic_packages

Ceramic Packages

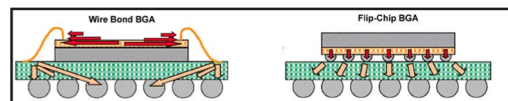
- Better thermal characteristics
 - Add heat-sink, tolerate hotter chips
 - To 100 W
 - More pins
 - More expensive



Source: https://www.ngkntk.co.jp/english/product/semiconductor_packages/htcc.html

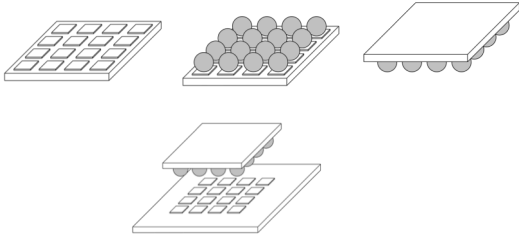
Flip Chip Packages

- Support Area-IO
 - More, denser pins
 - Smaller die if IO limited
 - Lower inductances
 - Smaller packaged chip



Source: <http://mantravlsi.blogspot.com/2014/10/flip-chip-and-wire-bonding.html>

Flip Chip I/O

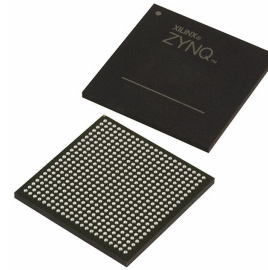


Source: https://en.wikipedia.org/wiki/Flip_chip

Penn ESE532 Fall 2019 -- DeHon

25

Zynq Land Grid Package



SBVA 484 – flip chip, Ball-Grid Array (UG 1075)

Penn ESE532 Fall 2019 -- DeHon

26

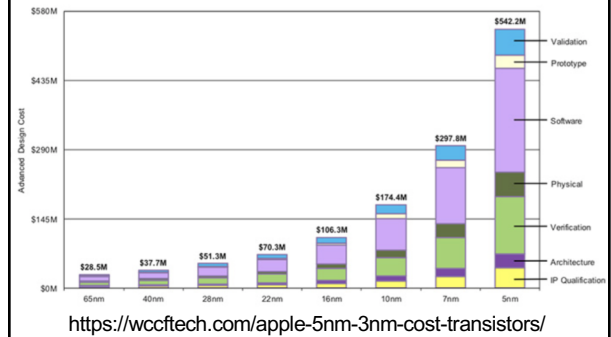
Don't Forget NRE

- This is all about recurring costs
- Cost = RecurringCost + (NRE/NumParts)
- NRE
 - Mask costs in millions
 - Design costs in 10s to 100s of millions

Penn ESE532 Fall 2019 -- DeHon

27

Bonus: Chip Design Costs



Penn ESE532 Fall 2019 -- DeHon

28

Putting Together

- 100mm² die -- \$5.6 raw
 - Maybe \$6--16 yielded -- call it \$7
- NRE \$100 M -- \$1
 - Sell 100 M units
- Put in \$1 package -- \$1
- Test -- \$1
- Total: \$10

Penn ESE532 Fall 2019 -- DeHon

29

Price vs. Cost

- ...and this is all about **cost**
 - What it takes to manufacture
- Price
 - What people will pay for it
- Profit = Price - Cost

Penn ESE532 Fall 2019 -- DeHon

30

Area

Area

- Simple story
 - Sum up component areas

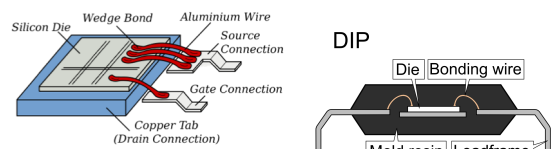
$$A = \sum_i A_i$$

Too Simplistic

- Area may be driven by
 - I/O
 - Interconnect
- Will need to pay for infrastructure
 - Clocking, Power

I/O Pads

- Must go on edge for wire bonding
 - Esp. for cheap packages

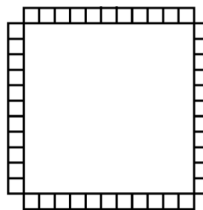


Src: <http://en.wikipedia.org/wiki/File:Wirebonding2.svg>

Source: https://commons.wikimedia.org/wiki/File:DIP_package_sideview.PNG

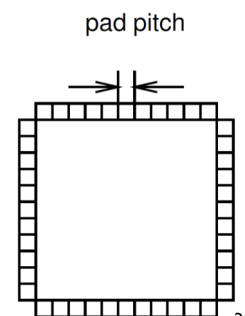
Pad Ring

- Pads must go on side of chip
- Pad spacing large to permit bonding
- I/O pads may set lower bound on chip size



Preclass 4

- 400 pads
- 25 μ m pad spacing
- Square chip dimensions?



I/O Limits

- Perimeter grows as $4s$
- Area grows as s^2
- Area grows $(\text{NumIO}/4)^2$
- IO may drive chip area

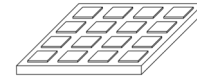
$$A = \text{Max} \left(\left(\sum_i A_i \right), \left(\frac{\text{NumIO}}{4 \times \text{PadPitch}} \right)^2 \right)$$

Penn ESE532 Fall 2019 -- DeHon

37

Area I/O

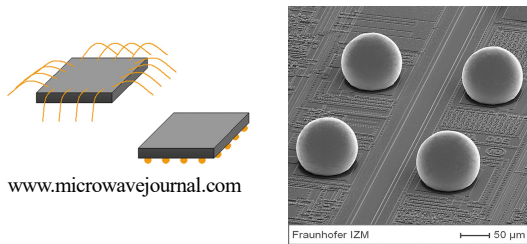
- Put I/O in grid over chip
- I/O pads still large and take up space
- Avoid perimeter scaling
- Requires more expensive flip-chip package



Penn ESE532 Fall 2019 -- DeHon

38

Flip Chip, Area IO



www.microwavejournal.com

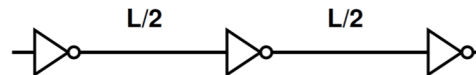
http://www.izm.fraunhofer.de/en/abteilungen/high_density_interconnect/waferlevelpackaging/arbeitsgebiete/arbeitsgebiet1.html

Penn ESE532 Fall 2019 -- DeHon

39

Interconnect

- Long wires need buffering
- Buffers take up space
 - Weren't in simple accounting of logic and memory blocks



Penn ESE532 Fall 2019 -- DeHon

40

Interconnect

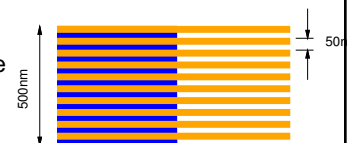
- Wires take up space
- Similar issue to pad I/O
 - Wires crossing into region grow as perimeter
 - Logic inside grows as area
- Region size may be dictated by wires entering/leaving

Penn ESE532 Fall 2019 -- DeHon

41

Wiring Requirements

- Wires 50nm pitch
- Gates 500nm on side
 - (500nm x 500nm)
- How many wires fit across the side of one gate?

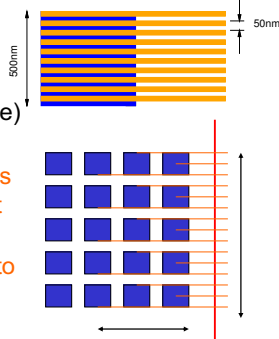


Penn ESE532 Fall 2019 -- DeHon

42

Wiring Requirements

- Wires 50nm pitch
- Gates 500nm on side
 - (500nm x 500nm)
- Wires/gate side (prev slide)
- If have SxS gate on left, how many wires can cross over the line of S gates at the right?
- What if need more wires to cross to right?

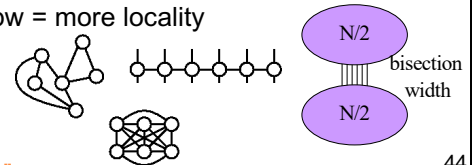


Penn ESE532 Fall 2019 -- DeHon

43

Bisection Width

- Partition design into two equal size halves
 - Minimize wires (nets) with ends in both halves
- Number of wires crossing is **bisection width**
 - Information crossing
- lower bw = more locality



Penn ESE532 Fall 2019 -- DeHon

44

Rent's Rule

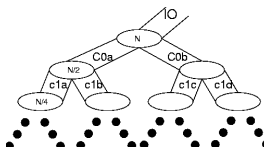
- If we recursively bisect a graph, attempting to minimize the cut size, we typically get:

$$BW=IO = c N^p$$

$$-0 \leq p \leq 1$$

- $p \leq 1$ means many inputs come from within a partition

[Landman and Russo, IEEE TR Computers p1469, 1971]

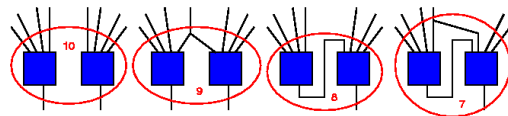


Penn ESE532 Fall 2019 -- DeHon

Rent and Locality

- Rent and IO quantifying locality
 - local consumption
 - local fanout

$$IO = c N^p$$

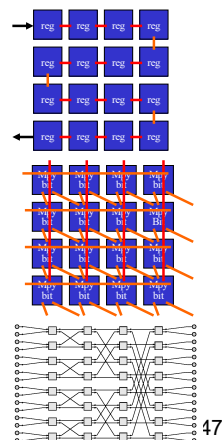


Penn ESE532 Fall 2019 -- DeHon

46

Common Applications

- Rent $p=0$
 - Shift-register, 1D filter
- Rent $p=0.5$
 - Array multiplier
 - 2D Window Filter
 - nearest-neighbor
- Rent $p=1.0$
 - FFT, Sort

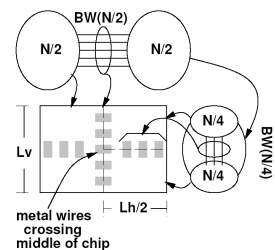


Penn ESE532 Fall 2019 -- DeHon

47

VLSI Interconnect Area

- Bisection width is lower-bound on IC width
 - When wire dominated, may be tight bound
- (recursively)
- Rent's Rule tells us how big our chip must be



Penn ESE532 Fall 2019 -- DeHon

48

As a function of Bisection

- $A_{\text{chip}} \geq N \times A_{\text{gate}}$
- $A_{\text{chip}} \geq N_{\text{horizontal}} W_{\text{wire}} \times N_{\text{vertical}} W_{\text{wire}}$
- $N_{\text{horizontal}} = N_{\text{vertical}} = \text{IO} = cN^p$
- $A_{\text{chip}} \geq (cN)^{2p}$
- If $p < 0.5$

$$A_{\text{chip}} \propto N$$

- If $p > 0.5$

$$A_{\text{chip}} \propto N^{2p}$$

Penn ESE532 Fall 2019 -- DeHon

49

In terms of Rent's Rule

- If $p < 0.5$, $A_{\text{chip}} \propto N$
- If $p > 0.5$, $A_{\text{chip}} \propto N^{2p}$

- **Typical** designs have $p > 0.5$

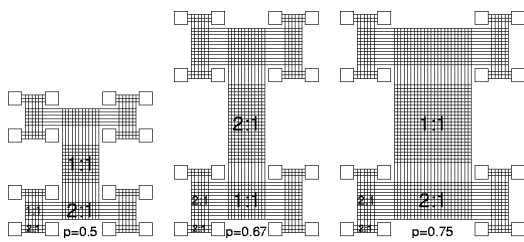
→ **interconnect dominates**

$$\rightarrow A_{\text{chip}} > \sum A_{\text{elements}}$$

Penn ESE532 Fall 2019 -- DeHon

50

Rent Network Richness



Penn ESE532 Fall 2019 -- DeHon

51

Infrastructure: Clocking

- PLL (Phased-Lock-Loop) to generate and synchronize clock
- Clock drivers are big (drive big load)
- Need buffering all over chip

Penn ESE532 Fall 2019 -- DeHon

52

Infrastructure: Power

- Need many I/O Pads
 - Carry current
 - Keep inductance low
- Wires to distribute over chip
- Maybe
 - Capacitance to stabilize power
 - Voltage converters

Penn ESE532 Fall 2019 -- DeHon

53

Area

$$A = \sum_i A_i$$

- Mostly sum of components, but...
- Area may be driven by
 - I/O
 - Interconnect $A \geq N^{2p}$
- Will need to pay for infrastructure
 - Clocking, Power

Penn ESE532 Fall 2019 -- DeHon

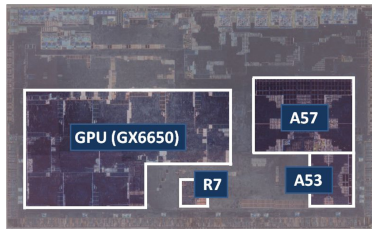
54

Some Areas

Processor Areas

- ARM Cortex A53 about 2mm^2 in 28nm
 - Zynq UltraScale+ processor
 - SuperScalar core
- A5 (scalar) about 0.25mm^2
- A9 (superscalar) about 1mm^2
- A15 (higher performance) about 3mm^2
- A57 (big core to A53 little) about 3mm^2

R-Car H3 from Rensas Quad A57, Quad A53



https://en.wikichip.org/wiki/arm_holdings/microarchitectures/cortex-a53

Zynq Compute Blocks

Crude estimate, including interconnect

- 2000 6-LUTs per sq. mm
- DSP Block ~ 0.1 sq. mm

CACTI

- Standard program for modeling memories and caches
 - More sophisticated version of the simple modeling we've been doing
- Will ask you to use for custom area estimates (P4 milestone, final report)

CACTI Parameters

- Technology
- Capacity
- Output Width
- Ports
- Cache ways

A13 Die Areas

- https://en.wikipedia.org/wiki/Apple_A13

SoC	A13 (7 nm)	A12 (7 nm)
Process Node	TSMC N7P	TSMC N7
Total Die	98.48	83.27
Big Core	2.61	2.07
Small Core	0.58	0.43
CPU Complex (incl. cores)	13.47	11.16
GPU Core	3.25	3.23
GPU Total	15.28	14.88
NPU	2.09	1.23

Penn ESE532 Fall 2019 -- DeHon

67

Big Ideas

- First order:
 - Chip cost proportional to Area
 - Area = Sum(Area(Components))
- But appreciate the simplification:
 - Yield makes cost superlinear in area
 - Limited range over which “linear” accurate
 - I/O, Interconnect, infrastructure
 - Can make Area > Sum(Area(Components))

$$A = \sum_i A_i$$

Penn ESE532 Fall 2019 -- DeHon

68

Admin

- P4 due Friday

Penn ESE532 Fall 2019 -- DeHon

69