

ESE532: System-on-a-Chip Architecture

Day 22: November 16, 2022
Estimating Chip Area and Costs



Today

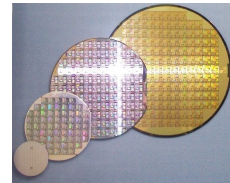
- Part 1: Chip Costs from Area
- Part 2: Chip Area
 - IO
 - Interconnect – Rent's Rule
 - Infrastructure
- Part 3: Some Areas

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

Penn ESE5320 Fall 2022 -- DeHon

7

7

...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/marketing

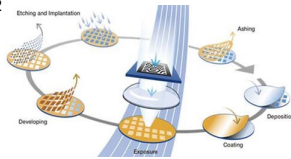
Penn ESE5320 Fall 2022 -- DeHon

8

8

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-that-is-involved-in-the-lithography-business/ja/2814575d-44709>

Penn ESE5320 Fall 2022 -- DeHon

9

9

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

Penn ESE5320 Fall 2022 -- DeHon

10

10

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

Penn ESE5320 Fall 2022 -- DeHon

11

11

Chip Yield

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

Penn ESE5320 Fall 2022 -- DeHon

12

12

Preclass 2

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

Area (mm ²)	Yield Rate
10	
50	
100	
500	

Penn ESE5320 Fall 2022 -- DeHon

13

13

Preclass 2

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

Area (mm ²)	Yield Rate
10	0.90
50	0.60
100	0.37
500	0.0066

Penn ESE5320 Fall 2022 -- DeHon

14

14

Yielded Die

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

Penn ESE5320 Fall 2022 -- DeHon

15

15

Preclass 3

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

Area	Raw Cost	Yield Rate	Cost/Yielded Chip
10		0.90	
50		0.60	
100		0.37	
500		0.0066	

Penn ESE5320 Fall 2022 -- DeHon

16

16

Preclass 3

- P=0.99 0.06 / mm²
- Die cost for:
 - 10 mm², 50 mm², 100 mm², 500 mm²

Area	Raw Cost	Yield Rate	Cost/Yielded Chip
10	0.60	0.90	
50	3.00	0.60	
100	6.00	0.37	
500	30.00	0.0066	

Penn ESE5320 Fall 2022 -- DeHon

17

17

Preclass 3

- P=0.99 0.06 / mm²
- Die cost for:
 - 10 mm², 50 mm², 100 mm², 500 mm²

Area	Raw Cost	Yield Rate	Cost/Yielded Chip
10	0.60	0.90	0.67
50	3.00	0.60	5.00
100	6.00	0.37	16.39
500	30.00	0.0066	4500

Penn ESE5320 Fall 2022 -- DeHon

18

18

Yielded Die Cost

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

Penn ESE5320 Fall 2022 -- DeHon

19

19

Yielded Die Cost

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

- 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

Penn ESE5320 Fall 2022 -- DeHon

20

20

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

Penn ESE5320 Fall 2022 -- DeHon

21

21

Slightly Fuller Story

- Chip cost = die + test + package

Penn ESE5320 Fall 2022 -- DeHon

22

22

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

Penn ESE5320 Fall 2022 -- DeHon

23

23

Packaging

- Pay for density and performance

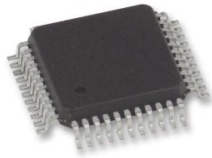
Penn ESE5320 Fall 2022 -- DeHon

24

24

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.electroons.com/doku.php?id=ic_packages

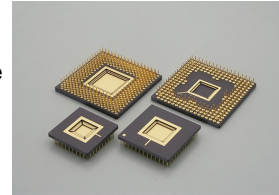
Penn ESE5320 Fall 2022 -- DeHon

25

25

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

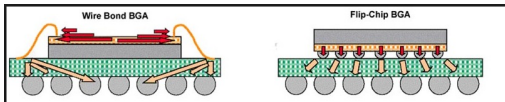
Penn ESE5320 Fall 2022 -- DeHon

26

26

Flip Chip Packages

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



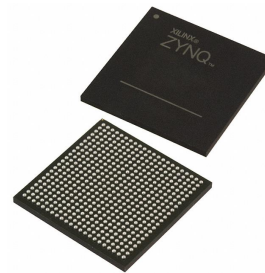
Source: <http://mantravlsi.blogspot.com/2014/10/flip-chip-and-wire-bonding.html>

Penn ESE5320 Fall 2022 -- DeHon

27

27

Zynq Land Grid Package



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

Penn ESE5320 Fall 2022 -- DeHon

28

28

Don't Forget NRE

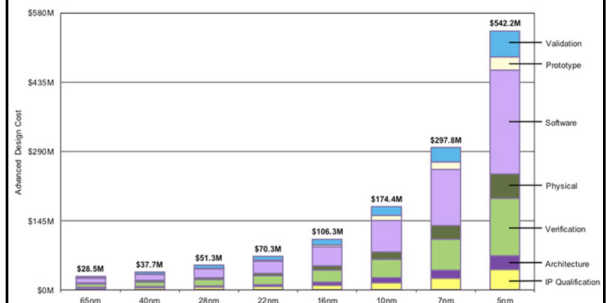
- This is all about recurring costs
- Cost = RecurringCost + (NRE/NumParts)
- NRE
 - Mask costs in millions
 - Design costs in 10s to 100s of millions

Penn ESE5320 Fall 2022 -- DeHon

29

29

Bonus: Chip Design Costs



<https://wccftech.com/apple-5nm-3nm-cost-transistors/>

Penn ESE5320 Fall 2022 -- DeHon

30

30

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 ESE5320 Fall 2022 -- DeHon

31

31

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 ESE5320 Fall 2022 -- DeHon

32

32

Area

Part 2

Penn ESE5320 Fall 2022 -- DeHon

33

33

Area

- Simple story
 - Sum up component areas

$$A = \sum_i A_i$$

Penn ESE5320 Fall 2022 -- DeHon

34

34

Too Simplistic

- Area may be driven by
 - I/O
 - Interconnect
- Will need to pay for infrastructure
 - Clocking, Power
 - ...maybe linear overhead factor

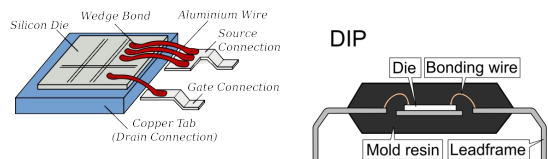
Penn ESE5320 Fall 2022 -- DeHon

35

35

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

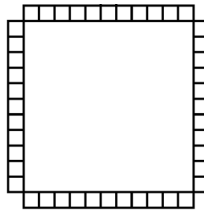
Penn ESE5320 Fall 2022 -- DeHon

36

36

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



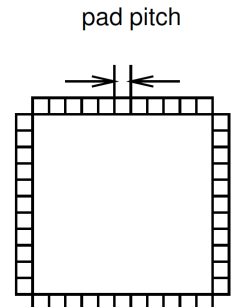
Penn ESE5320 Fall 2022 -- DeHon

37

37

Preclass 4

- 400 pads
- 25µm pad spacing
- Square chip dimensions?



Penn ESE5320 Fall 2022 -- DeHon

38

38

I/O Limits

- Perimeter grows as 4s
- Area grows as s²
- Area grows (NumIO/4)²
- IO may drive chip area

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

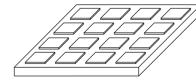
Penn ESE5320 Fall 2022 -- DeHon

39

39

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

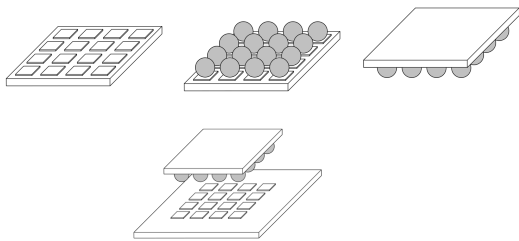


Penn ESE5320 Fall 2022 -- DeHon

40

40

Flip Chip I/O



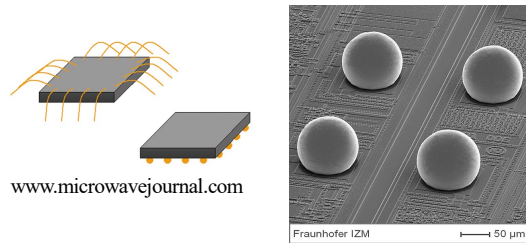
Source: https://en.wikipedia.org/wiki/Flip_chip

Penn ESE5320 Fall 2022 -- DeHon

41

41

Flip Chip, Area IO



www.microwavejournal.com

http://www.izm.fraunhofer.de/cv/abteilungen/high_density_interconnectwaferlevelpackaging/arbeitsgebiete/arbeitsgebiet1.html

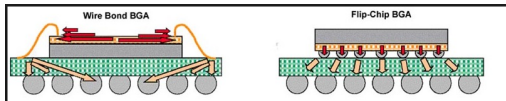
Penn ESE5320 Fall 2022 -- DeHon

42

42

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>

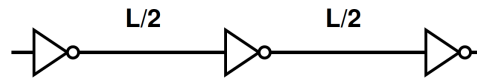
Penn ESE5320 Fall 2022 -- DeHon

43

43

Interconnect

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



Penn ESE5320 Fall 2022 -- DeHon

44

44

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

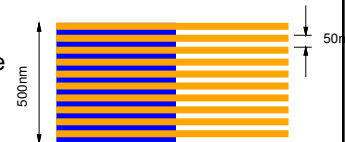
Penn ESE5320 Fall 2022 -- DeHon

45

45

Wiring Requirements

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



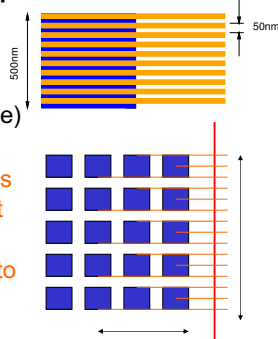
Penn ESE5320 Fall 2022 -- DeHon

46

46

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?



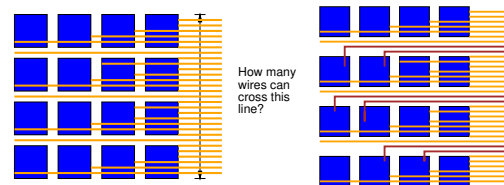
Penn ESE5320 Fall 2022 -- DeHon

47

47

Need More Wires

- What if need more wires to cross to right?



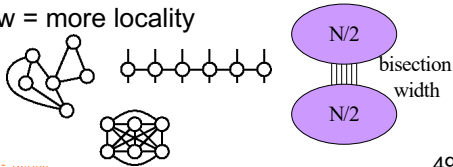
Penn ESE5320 Fall 2022 -- DeHon

48

48

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 ESE5320 Fall 2022 -- DeHon

49

49

Rent's Rule

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

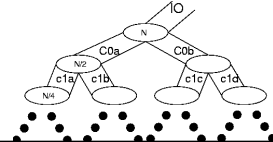
$$BW=IO = c N^p$$

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

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

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

Penn ESE5320 Fall 2022 -- DeHon

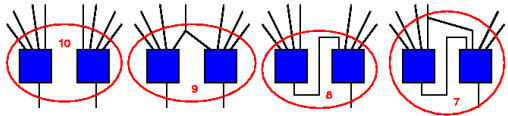


50

Rent and Locality

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

$$IO = c N^p$$



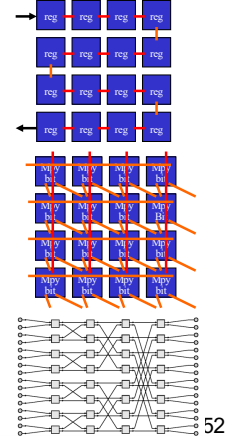
Penn ESE5320 Fall 2022 -- DeHon

51

51

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

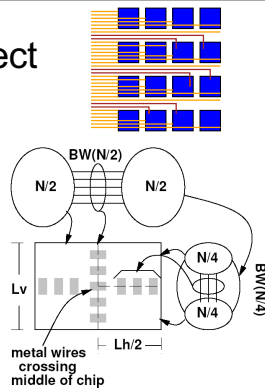


Penn ESE5320 Fall 2022 -- DeHon

52

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 ESE5320 Fall 2022 -- DeHon

53

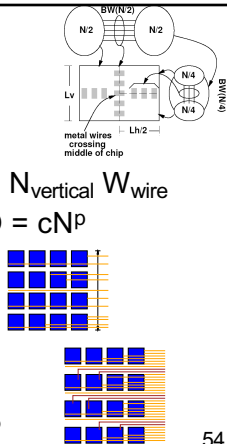
53

As a function of Bisection

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

$$A_{chip} \propto N$$

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



Penn ESE5320 Fall 2022 -- DeHon

54

In terms of Rent's Rule

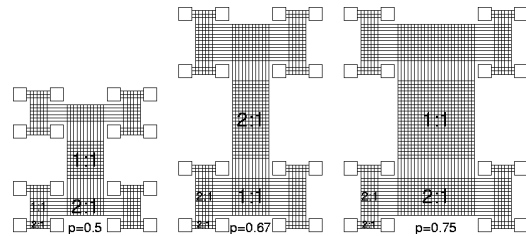
- If $p < 0.5$, $A_{\text{chip}} \propto N$
- If $p > 0.5$, $A_{\text{chip}} \propto N^{2p}$
- **Typical** designs have $p > 0.5$
 - **interconnect dominates**
 - $A_{\text{chip}} > \sum A_{\text{elements}}$

Penn ESE5320 Fall 2022 -- DeHon

55

55

Rent Network Richness



Penn ESE5320 Fall 2022 -- DeHon

56

56

Infrastructure: Clocking

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

$$A = F_{\text{infra}} \times \sum_i A_i$$

Penn ESE5320 Fall 2022 -- DeHon

57

57

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 ESE5320 Fall 2022 -- DeHon

58

58

Area

$$A = F_{\text{infra}} \times \sum_i A_i$$

- 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
 - Include as components
 - Then linear factor for distribution

Penn ESE5320 Fall 2022 -- DeHon

59

59

Some Areas

Part 3

Penn ESE5320 Fall 2022 -- DeHon

60

60

Processor Areas

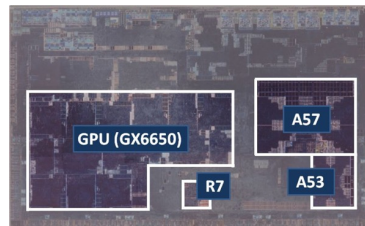
- ARM Cortex A53 about 2mm² in 28nm
 - Zynq UltraScale+ processor
 - SuperScalar core
- A5 (scalar) about 0.25mm²
- A9 (superscalar) about 1mm²
- A15 (higher performance) about 3mm²
- A57 (big core to A53 little) about 3mm²
- A72 (Amazon a1) about 4.6mm²
 - (ARM claim 1.15mm² in 16nm FinFet)

Penn ESE5320 Fall 2022 -- DeHon

61

61

R-Car H3 from Renesas Quad A57, Quad A53



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

Penn ESE5320 Fall 2022 -- DeHon

62

62

Zynq Compute Blocks

Crude estimate, including interconnect

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

Penn ESE5320 Fall 2022 -- DeHon

63

63

CACTI

- Standard program for modeling memories and caches
 - More sophisticated version of the simple modeling we've been doing

Penn ESE5320 Fall 2022 -- DeHon

64

64

CACTI – Memories on Zynq

- 32nm (closest technology it models to 28nm in Zynq)
- 36Kb BRAMs 0.025mm²
 - 2 port, 72b output
- ARM L1 cache 0.08mm²
 - 32KB 4-way associative
- ARM L2 cache 1.5 mm²
 - 512KB 8-way associative
 - (older, not yours)

Penn ESE5320 Fall 2022 -- DeHon

65

65

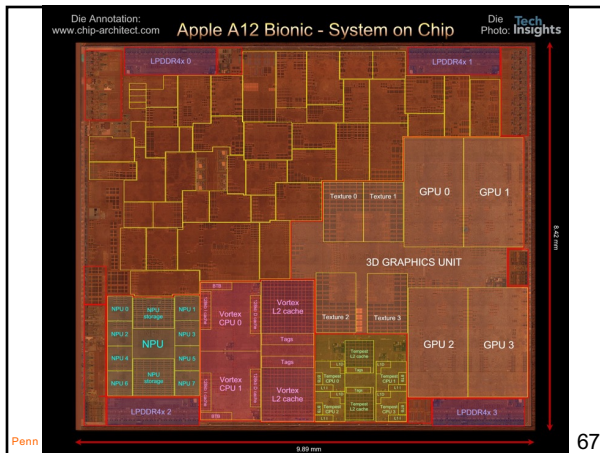
Zynq Component Estimates (32nm)

- 6-LUT 0.0005 mm²
- DSP Block 0.1 mm²
- 36Kb BRAMs 0.025mm²
- ARM L1 cache 0.08mm²
- ARM L2 cache (512KB, 8-way) 1.5 mm²
- ARM Cortex A53 2.0 mm²

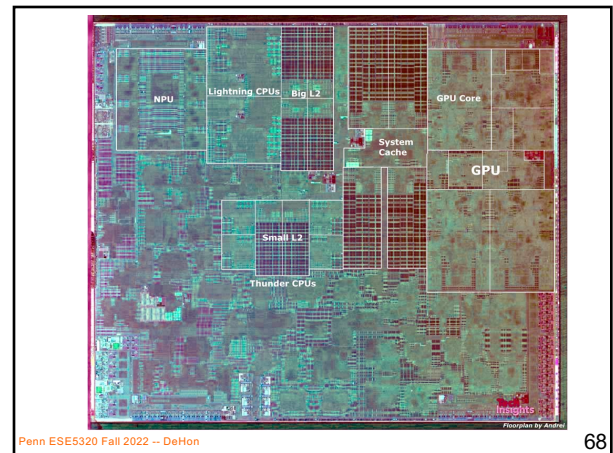
Penn ESE5320 Fall 2022 -- DeHon

66

66



67



68

Apple A13 Bionic

- 98mm², 7nm
- 8.5 Billion Tr.
- iPhone 11 +
- 6 ARM cores
 - 2 fast (2.6GHz)
 - 4 low energy
- 4 custom GPUs
- Neural Engine
 - 5 Trillion ops/s?

69

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

70

Zynq Component Estimates

• 6-LUT	0.0005 mm ²
• DSP Block	0.1 mm ²
• 36Kb BRAMs	0.025mm ²
• ARM L1 cache	0.08mm ²
• ARM L2 cache (512KB, 8-way)	1.5 mm ²
• ARM Cortex A53	2.0 mm ²

71

Rough Zynq 7020 Mockup (32nm)

Component	Area Each	Number	Total Area
A53 Core	2.0mm ²	2	
L2 cache	1.5mm ²	1	
(total processor)			
LUTs	0.0005 mm ²	53,200	
BRAMs	0.025 mm ²	140	
DSPs	0.1 mm ²	220	
(total PL)			
Total			

Crude model – omit many parts – for rough illustration.

72

Rough Zynq 7020 Mockup (32nm)

Component	Area Each	Number	Total Area
A53 Core	2.0mm ²	2	4.0
L2 cache	1.5mm ²	1	1.5
(total processor)			5.5
LUTs	0.0005 mm ²	53,200	26.6
BRAMs	0.025 mm ²	140	3.5
DSPs	0.01 mm ²	220	22.0
(total PL)			52.1
Total			57.6

Crude model – omit many parts – for rough illustration.

Penn ESE5320 Fall 2022 -- DeHon

73

73

Big Ideas $A = F_{infra} \times \sum_i A_i$

- 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))

Penn ESE5320 Fall 2022 -- DeHon

74

74

Admin

- Reading for Monday on Web
- P3 due Friday
- P4 posted
 - due 2 weeks, next Friday Thanksgiving break

Penn ESE5320 Fall 2022 -- DeHon

75

75