ESE532: System-on-a-Chip Architecture

Day 17: March 22, 2017
Estimating Chip Area and Costs

Today

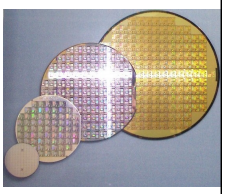
- Chip Costs from Area
- Chip Area
- Some Areas
  - CACTI – for modeling memories

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. $3,500
- 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

Preclass 1

- Rough cost per mm of silicon?
  - $3500 for 300mm wafer

Implication

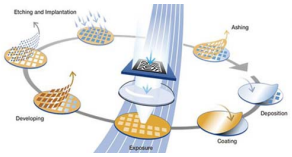
- Raw silicon die cost is roughly proportional to area
  - Larger the die, the fewer we get on the wafer

…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/marking

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/s28145?rid=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

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 mm^2, 50 mm^2, 100 mm^2, 500 mm^2

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 mm^2, 50 mm^2, 100 mm^2, 500 mm^2

Yielded Die Cost

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

Yielded Die Cost

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

- 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

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

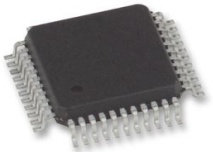
- Testing costs proportional to test time
  - Time on expensive test unit
  - Depends on complexity of tests need to run
    - Can motivate spending silicon area on on-chip test structures to reduce
- Can dominate on small chips or complex testing

Packaging

- Pay for density and performance

Plastic Packages

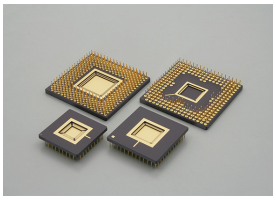
- Simple plastic packages cheap
  - Limited number of pins
    - Limited to perimeter
  - Limited heat removal (few Watts)
  - Can be large (due to pins)
  - Higher inductance on pins



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

Ceramic Packages

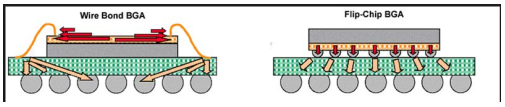
- Better thermal characteristics
  - Add heat-sink, tolerate hotter chips
    - To 100 W
  - More pins



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

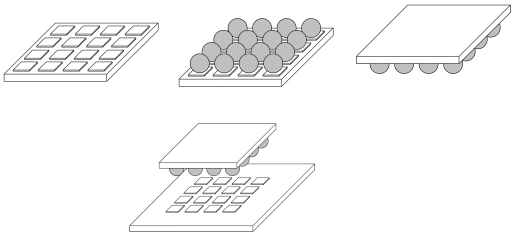
Flip Chip Packages

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



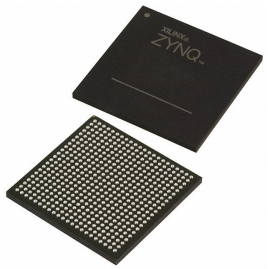
Source: http://mantravlsi.blogspot.com/2014/10/flip-chip-and-wire-bonding.html

Flip Chip I/O



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

Zynq Land Grid Package



Don't Forget NRE

- This is all about recurring costs

$$Cost = \mathrm{Re}curreningCost + \frac{NRE}{NumParts}$$

- NRE
  - Mask costs in millions
  - Design costs in 10s to 100s of millions

Price vs. Cost

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

Area

Area

- Simple story
  - Sum up component areas

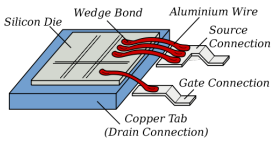
$$A = \sum_i A_i$$

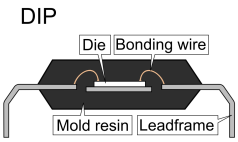
Too Simplistic

- 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



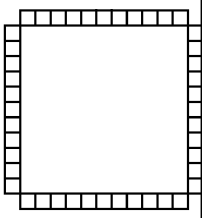


Src: http://en.wikipedia.org/wiki/File:Wirebonding2.svg

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

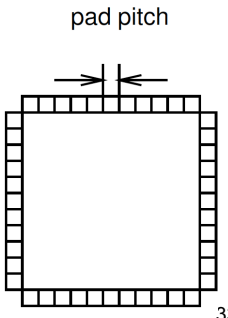
Pad Ring

- 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
- 25μm pad spacing
- Minimum chip dimensions?



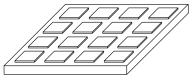
I/O Limits

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

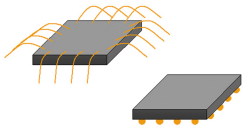
$$A = Max\left(\left(\sum_i A_i\right), \left(\frac{NumIO}{4}\right)^2\right)$$

Area I/O

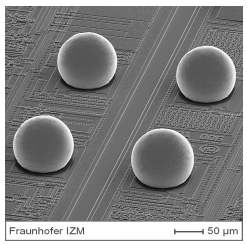
- Put I/O in grid over chip
- I/O pads still large and take up space
- Avoid perimeter scaling
- Requires more expensive flip-chip package



Flip Chip, Area IO



www.microwavejournal.com



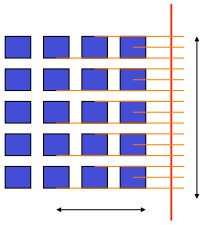
http://www.izm.fraunhofer.de/en/abteilungen/high_density_interconnectwaferlevelpackaging/arbeitsgebiete/arbeitsgebiet1.html

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

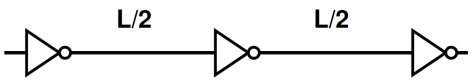
Wiring Requirements

- Wires 50nm pitch
- Gates 500nm tall
- How many gates per row can provide outputs before saturate edge?
- What if want more outputs to exit across edge?



Interconnect

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



Infrastructure: Clocking

- PLL (Phased-Lock-Loop) to generated lock
- Clock drivers are big (drive big load)
- Need buffering all over chip

Infrastructure: Power

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

Area $A = \sum_i A_i$

- Mostly sum of components, but…
- Area may be driven by
  - I/O
  - Interconnect
- Will need to pay for infrastructure
  - Clocking, Power

Some Areas

Processor Areas

- ARM Cortex A9 about 1mm^2 in 28nm
  - Zynq processor
  - SuperScalar core
- A5 (scalar) about 0.25mm^2
- A15 (higher performance) about 3mm^2

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

CACTI Parameters

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

Example Output

```
- Total cache size (bytes): 32768
  Number of banks: 1
  Associativity: 4
  Block size (bytes): 64
  Read/write Ports: 1
  Read ports: 0
  Write ports: 0
  Technology size (nm): 32

  Access time (ns): 1.09421
  Cycle time (ns):  1.25458
  Total dynamic read energy per access (nJ): 0.0234295
  Total dynamic write energy per access (nJ): 0.018806
  Total leakage power of a bank without power gating, including its netwo
  Cache height x width (mm): 0.152304 x 0.523289
```

CACTI – Memories on Zynq

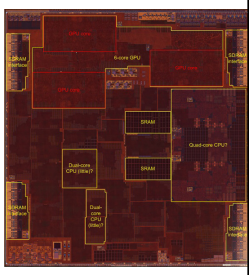
- 32nm (closest technology it models to 28nm in Zynq)
- 36Kb BRAMs 0.025mm
  - 2 port, 72b output
- ARM L1 cache 0.08mm
  - 32KB 4-way associative
- ARM L2 cache 1.5 mm
  - 512KB 8-way associative

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 1.5 mm^2
- ARM Cortex A9 1.0 mm^2

Today SoC: Apple A10

- Quad=Dual Dual Core
  - Dual 64-bit ARM 2.3GHz
  - Dual 64-bit ARM low energy
- 3MB L2 cache
- 6 GPU cores
- Custom accelerators
  - Image Processor?
- 125mm^2 16nm FinFET
- 3.3B transistors



Chipworks Die Photo

Big Ideas

- First order:
  - Chip cost proportional to Area $A = \sum_i A_i$
  - 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))

Admin

- Project Analysis Milestone due Friday
  - Including teaming