3D-Stacked Integrated Circuits: from technology to application

Dragomir Milojevic

9.6.2015

Lille

“No thing under the sun is new”

history of image representation

2000BC, Egypt

“fressco a secco”

1450AD, Italy

perspective

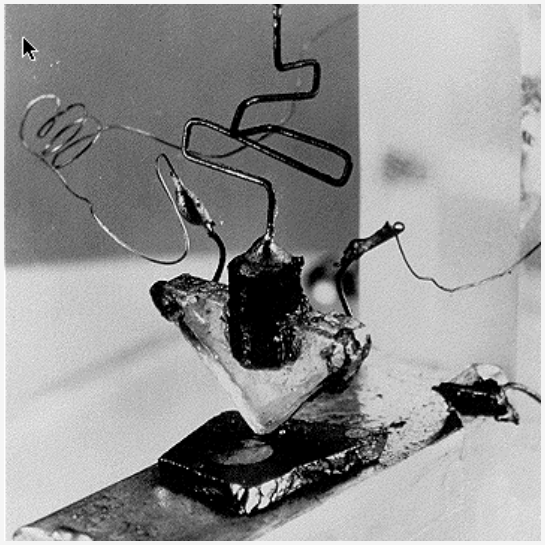
1995AD, California (US)

computer animated 3D

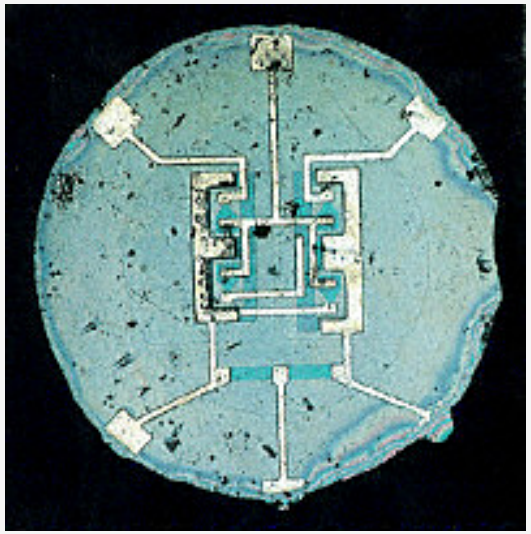
Same happened to ICs

In the beginning everything was 2D …

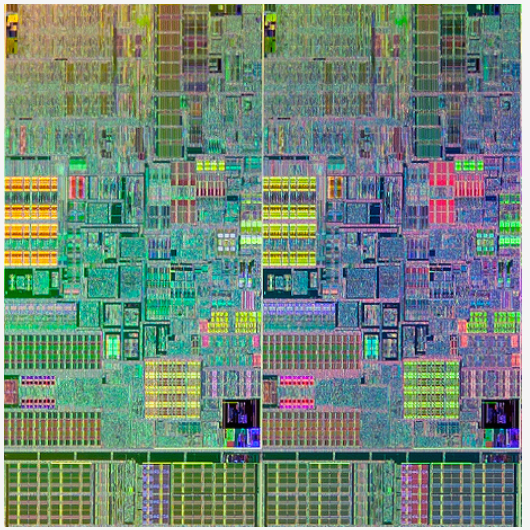
Transistor



IC



VLSI



!!! Enter the 3D world !!!

Outline

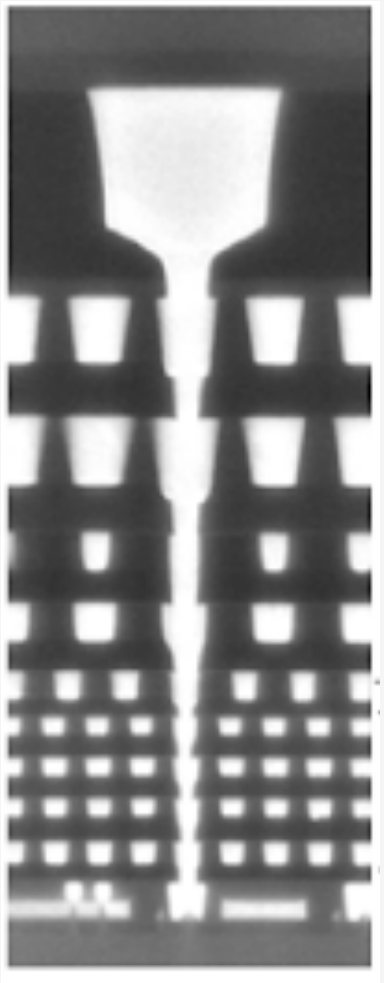
1. 2D ASICs
2. CMOS scaling (and problems)
3. 3D integration
4. Applications and benefits
5. Conclusion

Outline

1. **2D ASICs**
2. CMOS scaling (and problems)
3. 3D integration
4. Applications and benefits
5. Conclusion

ASIC structure

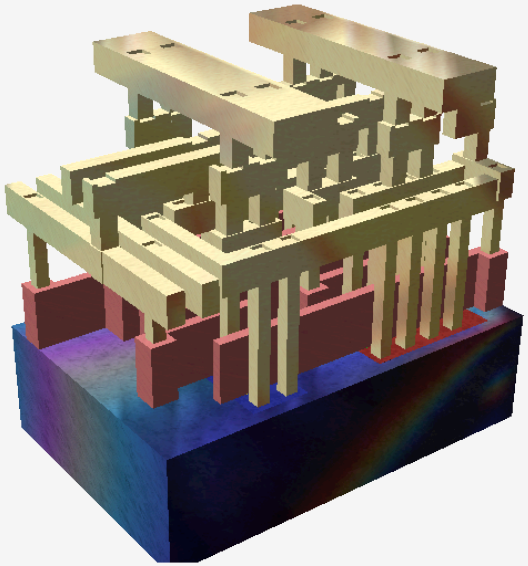
- Layered structure, each layer is processed in sequential manner (meaning one after another) in a manufacturing line
- **Substrate** – support for mechanical handling of the IC (initial wafer thickness, can be thinned)
- **Front End Of Line – FEOL**
  (so called because processed first in the line)
  Active layer, contains transistors used to build gates, this is VERY thin
- **Back End Of Line (BEOL)**
  (processed last in the manufacturing line)
  - ✦ **Metal layers** – conductors in a plane with possibly alternated preferred direction
  - ✦ **Via layers** – connect different metal layers



ASIC manufacturing

- Variety of physical and chemical processes performed on a semiconductor substrate (e.g. silicon)
- Various processes (film deposition, patterning, semiconductor doping, ...)
  - **conductors** – such as polysilicon, aluminum, and more recently copper
  - **insulators** – various forms of silicon dioxide, silicon nitride, ...

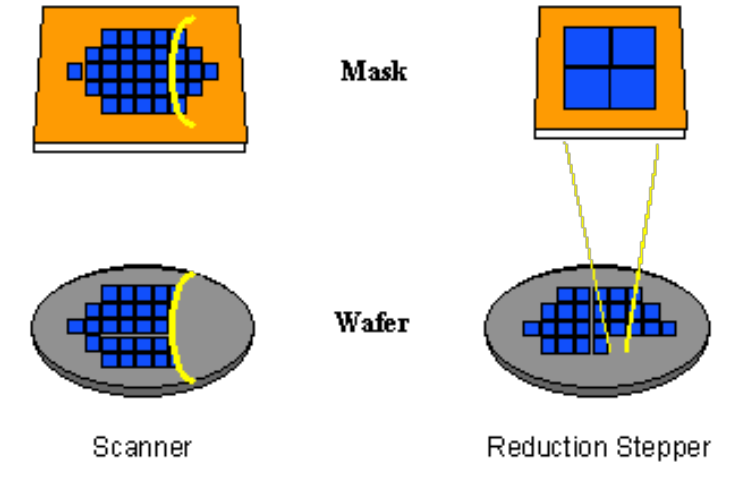
  are used to create, connect and isolate transistors and their components



- Fundamental to all of these processes is **lithography**: formation of three-dimensional relief images on the substrate for subsequent transfer of the pattern to the substrate on wafers

Printing system: resolution

- **Projection lithography tools**
  - **contact** – simultaneous patterning of the complete wafer
  - **scanning** – scans throughout the mask
  - **step-and-repeat systems** – expose a part (reticle), step out of the process



- **Tools → resolution**, i.e. the smallest printable feature is limited by:
  - the smallest image that can be projected onto the wafer,
  - and the resolving capability of the photoresist to make use of that image
- **Projected image resolution *R*** is determined by:
  - k1 – process related factor (increases by definition)
  - the wavelength of the imaging light (λ)
  - and the numerical aperture (NA)

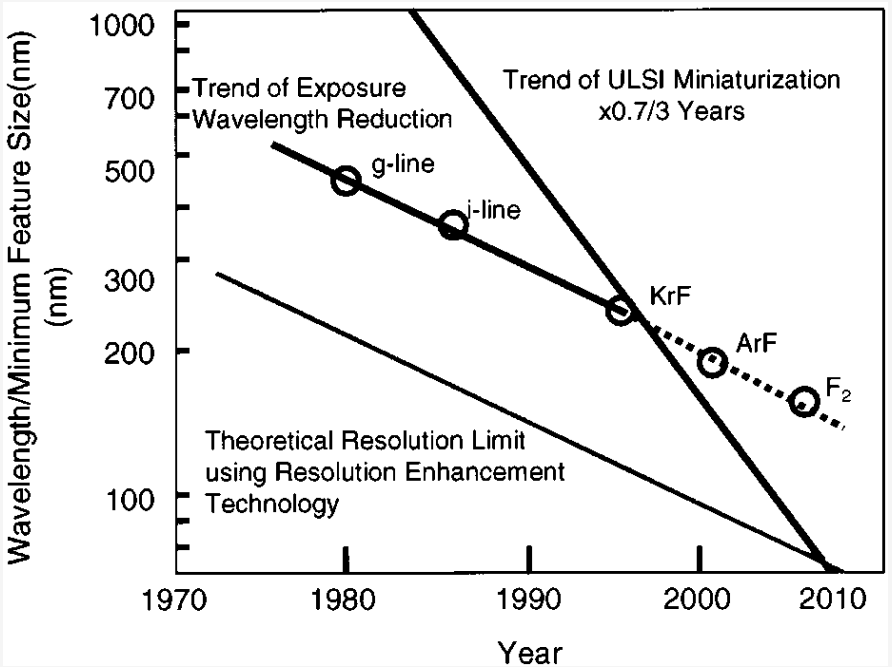
$$R = k_1 \frac{\lambda}{NA}$$

Lithography vs. Moore

- From: $$R = k_1 \frac{\lambda}{NA}$$

  to increase the printing resolution:

  - **decrease wavelength**
  - **increase NA**



- Light sources:
  435→248→193 and 157nm (~2010), they are already huge !
- Plotted against Moore → there is a breakdown ...
- Wavelength reduction only is not enough to follow the scaling requirements !!!
- **Supplementary tricks to print features smaller then “light”**

Practical considerations

- Photolithography needs high resolution, high sensitivity, precise alignment and low defect density
- Advanced ICs **more than 30 patterning steps**; each one must align with the previous one precisely to successfully transfer the pattern of the chip design → **lengthy process**
- Photolithography takes 40–50% of total wafer-processing time
- **In practice this can last from six to eight weeks!**
  (from bare wafers to finished IC as of 2001)
- Solution: optimize & parallelize processing to increase the wafer throughput, but this increases the cost significantly !
  - 100.000 wafers/moth facility → 10 billion US$ (2012)

Outline

1. 2D ASICs
2. CMOS scaling (and problems)
3. 3D integration
4. Applications and benefits
5. Conclusion

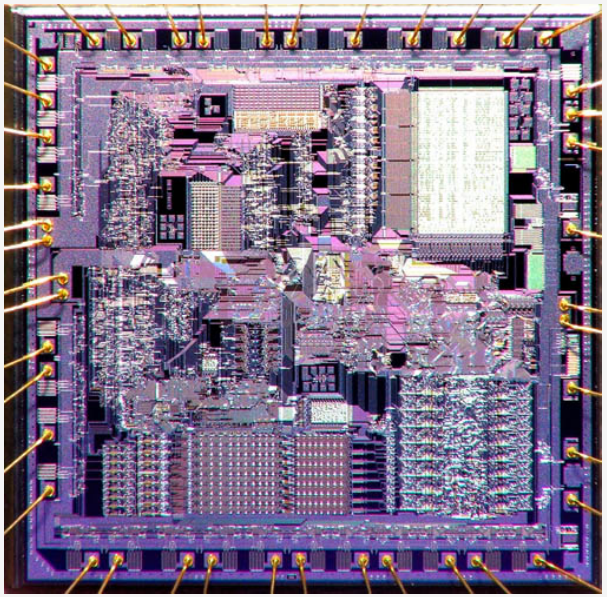
Scaling principals

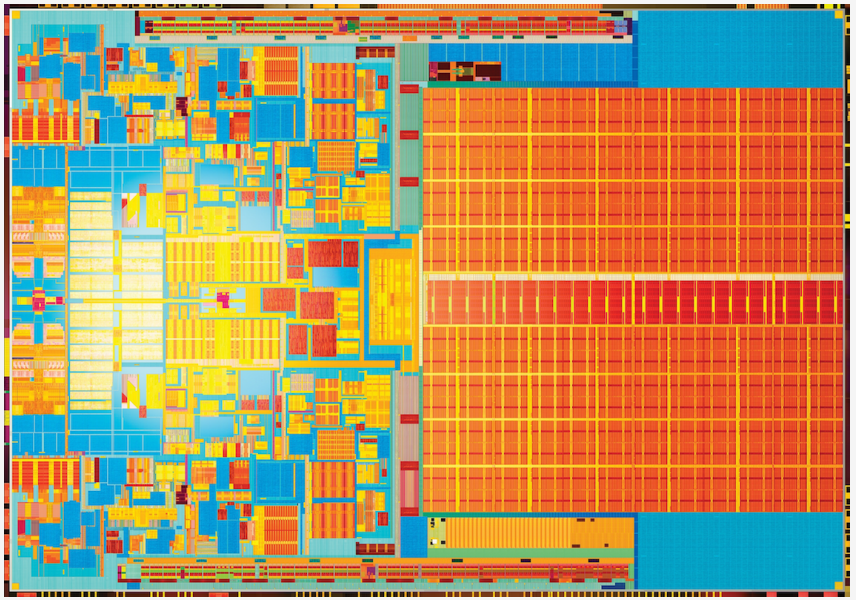
- From node to node, the size of the minimum (linear) feature (dimension) is reduced by a constant factor λ → **scaling**
- Reduced feature size means smaller logic gate and better performance (smaller delay, lower power)
- Reduction of the feature size (λ) can be predicted, and time showed that it is constant (Moore's law)
  - λ = 0.7
  - Area goes down with factor 0.5 (0.7x0.7)
  - All other performance parameters are f(λ)

feature size	area	capacitance (C)	frequency (f)	V_{dd}	power ($CV_{dd}^2 f$)	power density
0.7X	0.5X	0.616X	1.0X	0.925X	0.527X	1.054X

Good old days …

- You make your design in technology node n
- Once you benchmark your design for this node you could of predict what would happen in the node n+1
  … and this was true for about 50 years !
- Key design parameters: area, power, performance etc. → from n to n+1 constant evolution





If only the car industry did the same ...

Speed	180.000.000 km/h
Fuel	0,04 l/100km
Price	0,0003$

Scaling values for current techs

tech node	feature size	area	capacitance (C)	freq (f)	V_{dd}	power ($CV_{dd}^2 f$)	power density
45-¿32nm	0.755X	0.57X	0.665X	1.10X	0.925X	0.626X	1.096X
32-¿22nm	0.755X	0.57X	0.665X	1.08X	0.95X	0.648X	1.135X
22-¿14nm	0.755X	0.57X	0.665X	1.05X	0.975X	0.664X	1.162X
14-¿10nm	0.755X	0.57X	0.665X	1.04X	0.985X	0.671X	1.175X

- Area continues to scale as well as capacitance (same λ)
- Frequency gains are lower
- Power supply voltage slows down
- Less power gains
- Increased power density (impact on cooling)

Scaling is hitting the wall !

Many reasons cause scaling wall

In this talk we focus on:

- Interconnect related problems
  - Delay
  - Impact of long interconnects
  - Interconnect power
- Cost and manufacturing efficiency
- Lithography issues

Interconnect delay – RC model

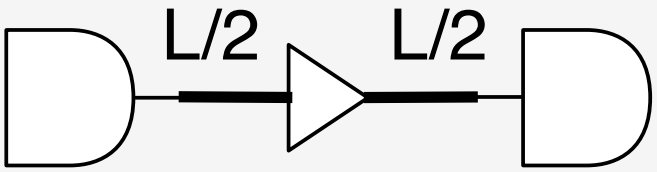
$$d_w = rc\frac{L^2}{2}$$

Technology

r – wire resistance / unit length

c – wire capacitance / unit length

- Wire delay: function of the material properties and length
- Not much can be done on material properties to reduce the delay
- Wire length is the dominating factor! (true for power too) P&R can do something but not much

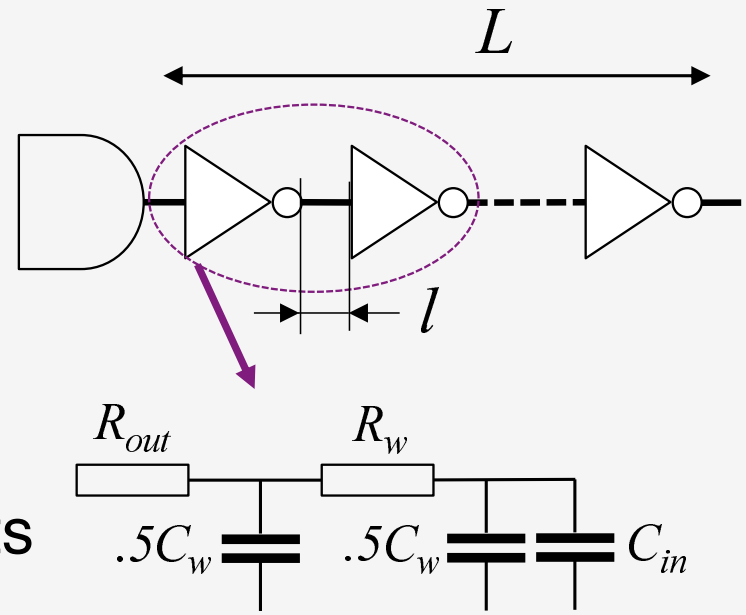


$$d_w = d_{INV} + 2 \times rc\left(\frac{L}{2}\right)^2$$

- Solution: insert a repeater ! (as long as the introduced delay is low)
- You insert inverter delay but you reduce by 2 the impact of the wire length
- For long wires you will insert as many as you need (can be a lot!, done automatically at P&R)

Long wires delay

- Unit delay: from repeater to repeater
- Two inverters connected with a wire of given length
- Repeated segment delay model, as on the previous slide
- Total delay the sum of all delay segments
- Will depend on the connected gate properties *Rout* and *Cin*



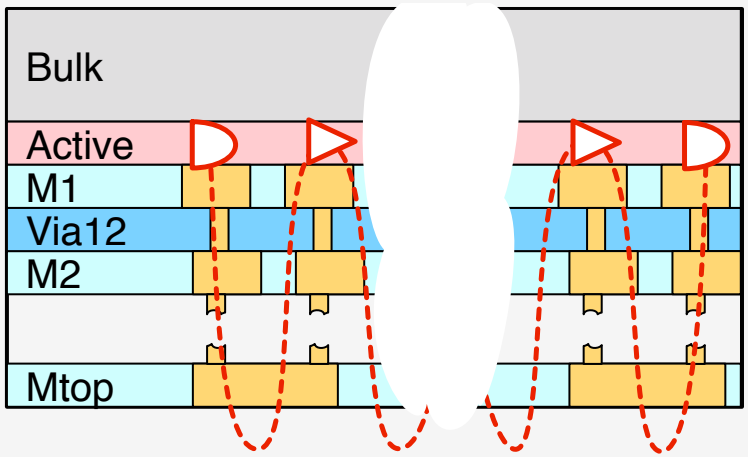
$$D_l = R_{out}\frac{C_w}{2} + (R_{out} + R_w)\left(\frac{C_w}{2} + C_{in}\right)$$

$$D_L = \frac{L}{l} \cdot (D_l + D_{rep})$$

Typically every 100nm…
this means a lot of repeaters !!!

Repeater insertion: consequences

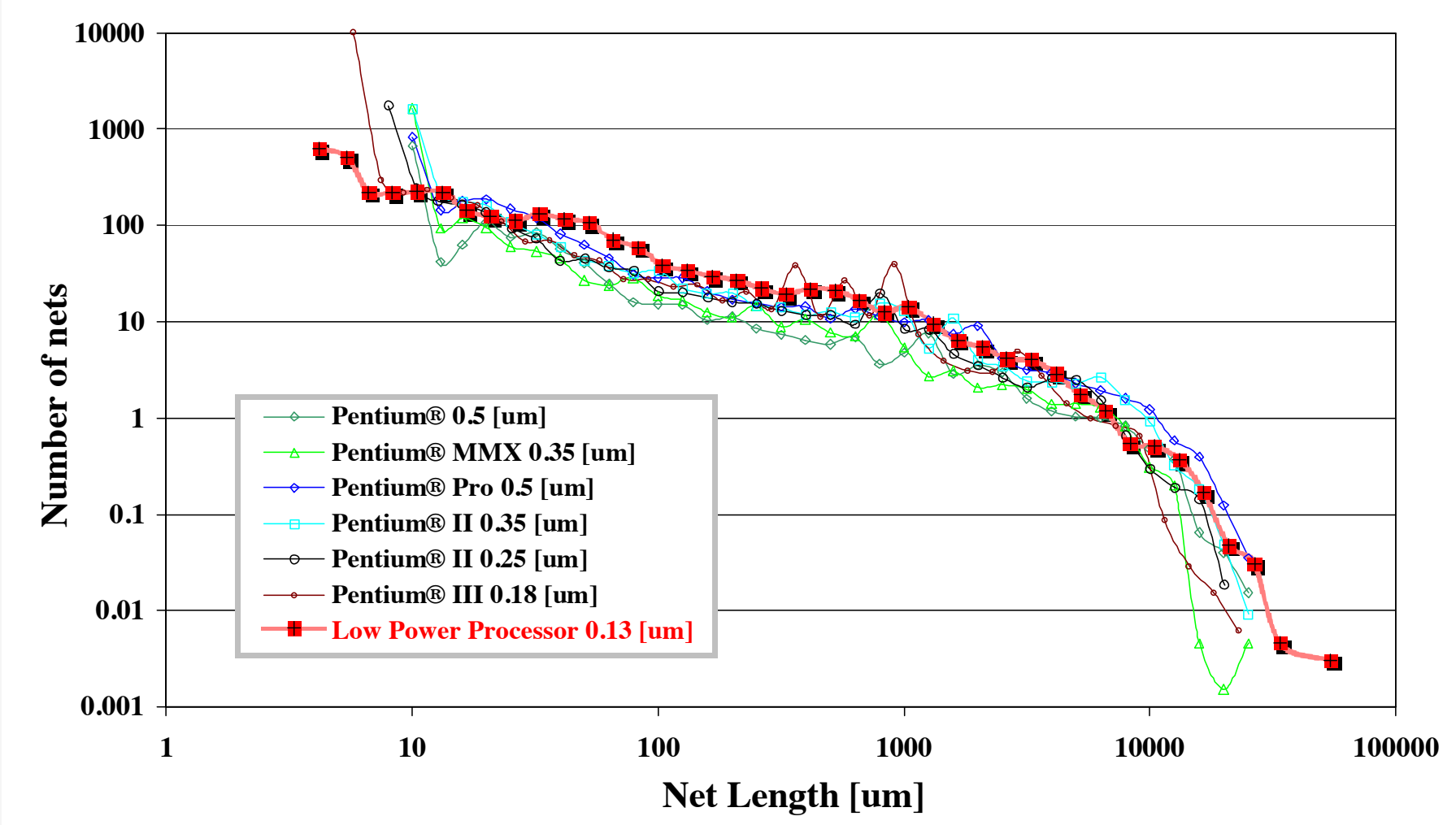
- Extra area (read cost)
- More power
- But also (and even worse) :
  - Many via cuts from upper metal layers down to substrate
  - Use of many routing resources, scarce for advanced technologies → **increased routing congestion**
- To avoid congestion further area increase (& cost)



How many repeaters will be inserted is design/target perf. dependent, and is directly linked to **wirelength distribution**

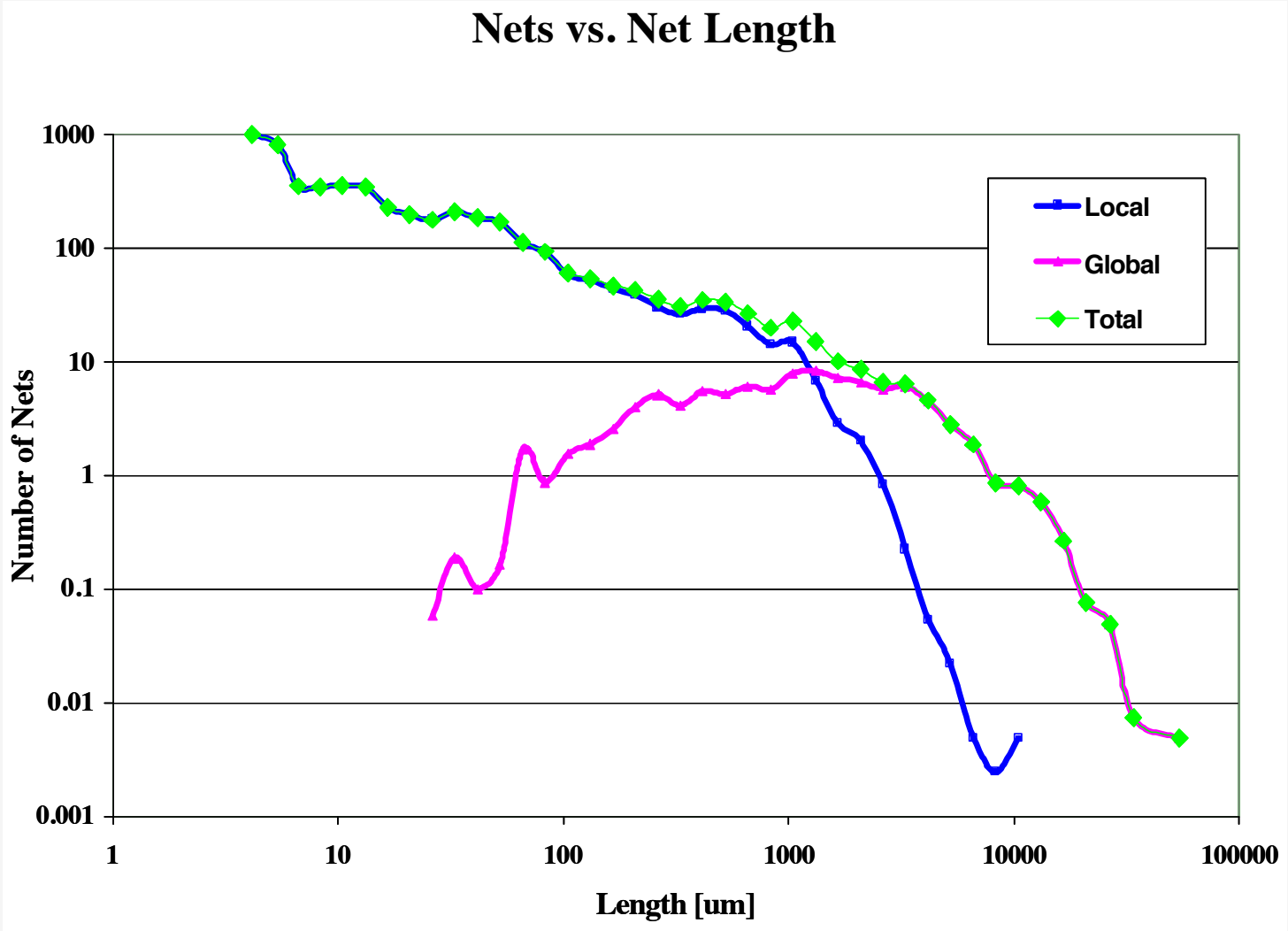
Wirelength distribution of a CPU

Lot's of wires are long …



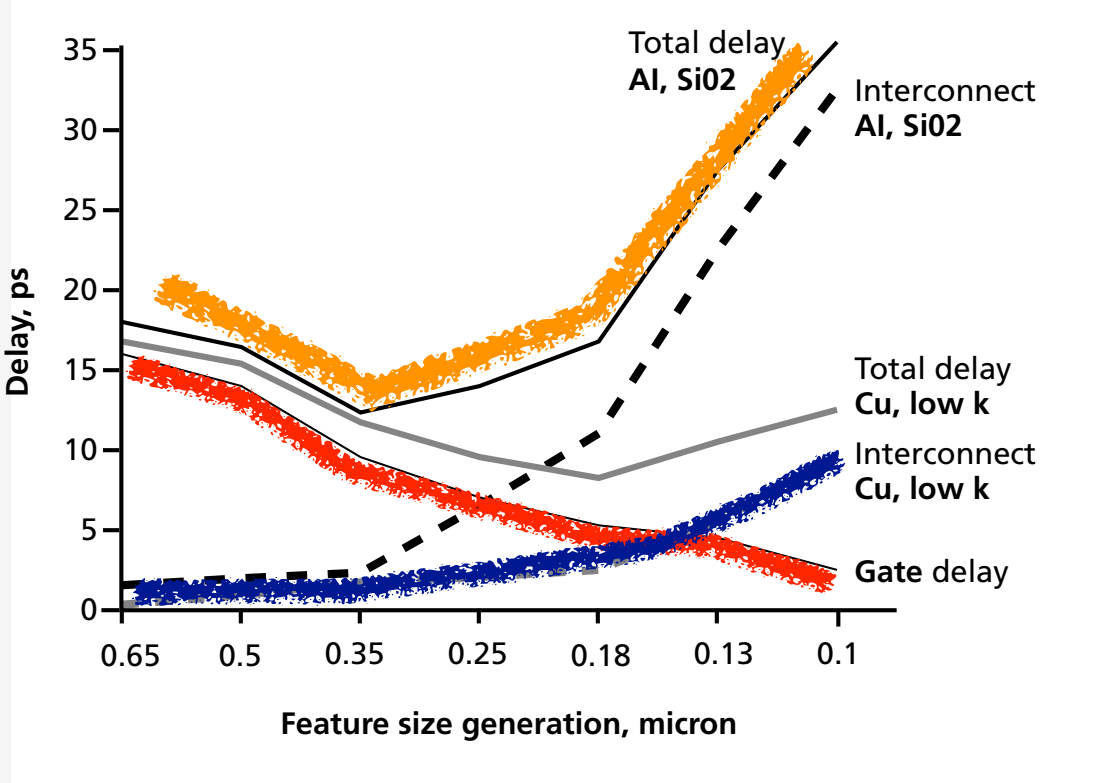
Wirelength and BEOL

Local vs. global wires: significant number of global wires



Solving wire delay: there is a tech limit!

Gate vs. wire delay across technologies



Gate delay

Wire delay

Total delay

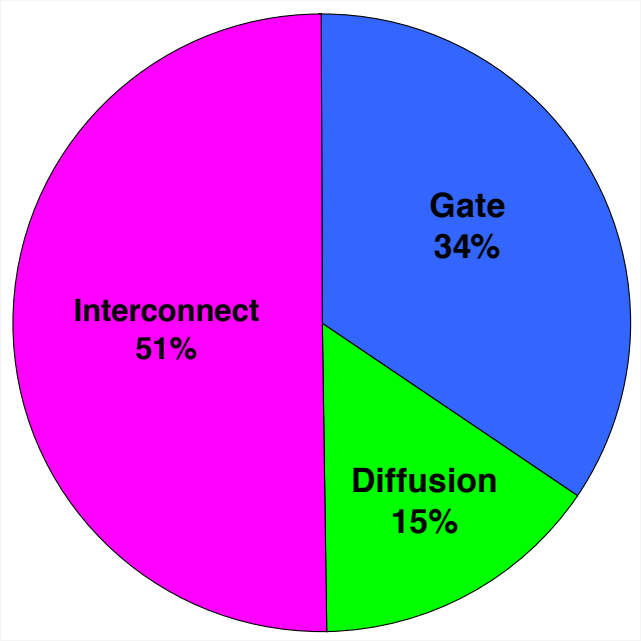
After 180nm (1999) inversion of tendency:
interconnect delay dominates gate delay

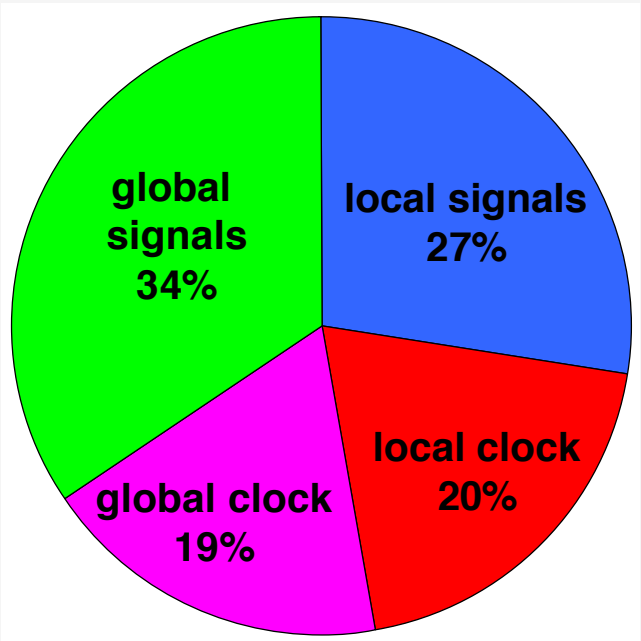
Globally, delay is increasing !

Interconnect power is important !

Typical CPU

- Interconnect consumes 50% of total dynamic power of the IC
- This power dissipation is due to
  - parasitic capacitance of wires
  - and repeaters (they can not be gated)!
- 90% of power consumed by 10% of nets
- Clock power: 40% of interconnect power
- Interconnect design is NOT power-aware (at this level it is difficult to do anything)



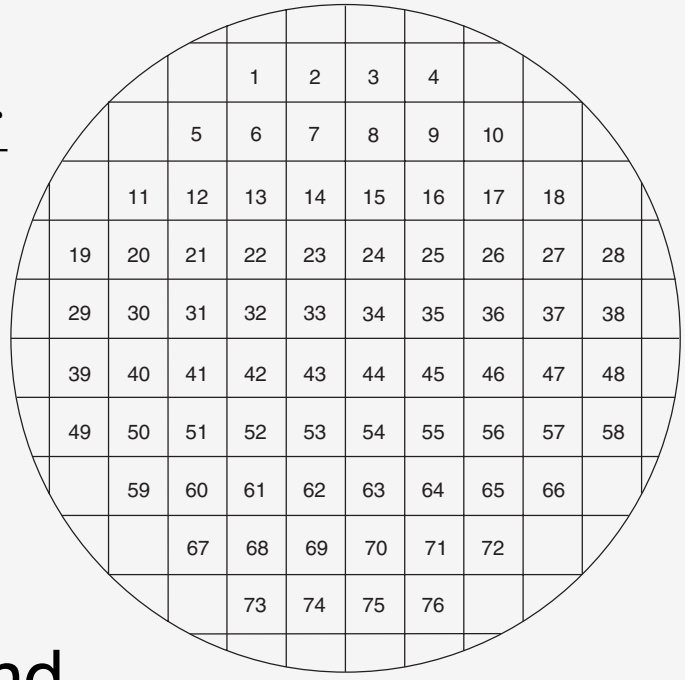


Manufacturing issues & cost

- Wafers are cylindrical, and IC's rectangular
- There is an area loss that is function of the wafer & die size :

$$\text{Die per wafer} = \frac{\pi(\text{wafer diameter}/2)^2}{\text{die area}} - \frac{\pi \times \text{wafer diameter}}{\sqrt{2 \times \text{die area}}}$$

  - 1st term – wafer/single die area
  - 2nd term – **area loss of rectangular die that do not entirely fit the wafer**



- Smaller dies mean less edge effect and hence lower per die cost

Yield

- Density of defects and complexity of the manufacturing process determine **the die yield** – the percentage of functional dies
- Assuming defects are uniformly distributed across the wafer, the die yield is estimated as:

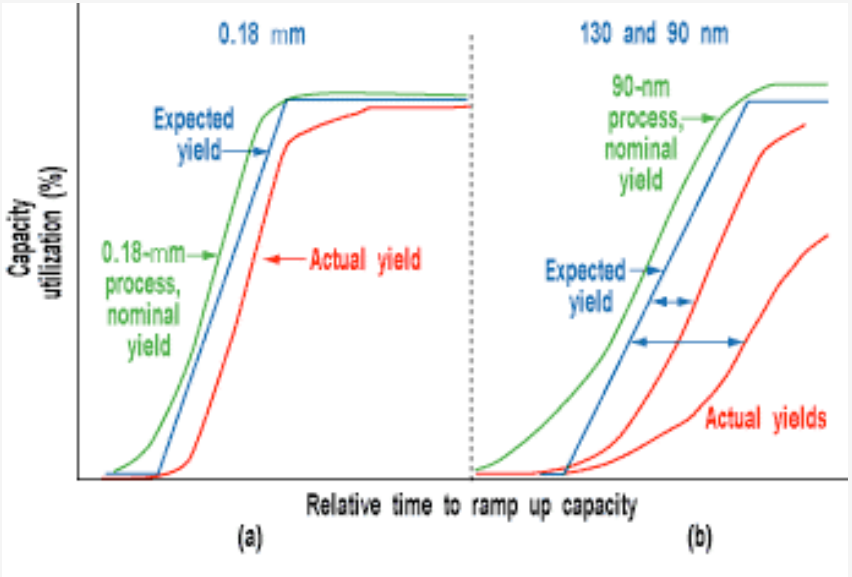
$$\text{Die yield} = \text{wafer yield} \times \left(1 + \frac{\text{defects per area} \times \text{die area}}{\alpha} \right)^{-\alpha}$$

  where α is a measure of the complexity of the fabrication process (for modern CMOS processes is $\alpha = 4$)

- Wafer yield = percentage of successfully processed wafers (often close to 100 percent, but after certain time, see next slide)
- Yield = function of the frequency of defects and the size of the die
- In 2001: defects per area >0.4 and <0.8 defects/cm2 (more processing steps lead to a higher value)

Yield as function of time

- Ramp-up shape
- In the beginning the yield is low, the process is fine-tuned so that very quickly reaches the nominal value of almost 100% (typically 95%)
- But this is function of technology!
- As we move towards more and more aggressive nodes actual yield curves don't reach the expected yield and in more time
- This is related to cost …
  → economics is the worst CMOS enemy
  (even bigger then physics)



Test & packaging

- Dies are tested before wafer slicing
- Only good dies will be packaged, because packaging adds extra and non-negligible cost
- **Die assembly and package cost**: base cost plus cost depending on the number of pins (~1000k pins would be typical high pin count)
  - Base cost – depends on thermal : few dollars +
  - Per pin cost – typically = 0.5 cents/pin
    - But limit the total power to less than 3 W
- High-cost, high-performance packages might allow power densities up to 100 W/cm2, but have base costs of $10 to $20 plus 1 to 2 cents per pin !
- **Since for high performance processors power density is increasing, packaging could represent significant part of the total cost**

Lithography issues

- Lithography scaling is the key enabler of the Moore's Law
  - resolution, tech
  - **critical dimension** (smallest feature R) control, tech
  - overlay accuracy, tech
  - **throughput**, $$$
- We have solutions for high-resolution printing methods that can go well beyond 30nm, but the ultimate limit to lithography scaling will be set by:
  - critical dimension control requirements &
  - economics rather than purely resolution performance
- **Scaling will stop not because we can't do it, but because we will not be able to afford it !**

Planar CMOS is hitting the wall

- **Lithography:** advanced nodes are becoming more and more tricky, causing yield decrease → **increased cost**
- **Scaling:** doesn't work that good; from node to node **gains are less** (frequency doesn't scale up)
- **Power density:** is constantly increasing, leading to **longer design times** (cost) and **expensive cooling** solutions
- **Interconnect wall:** we compute fast, but "communicate" slow, **critical paths are worse**
- **Cost**: all the above contribute to **exponential cost rise** from node to node

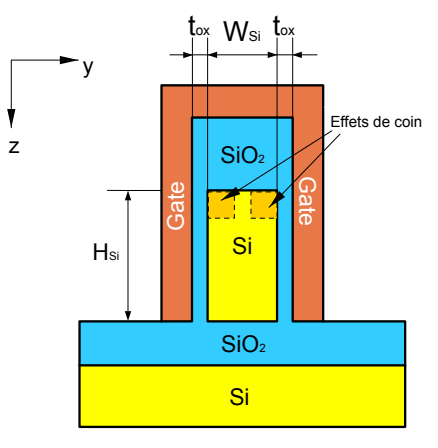
Outline

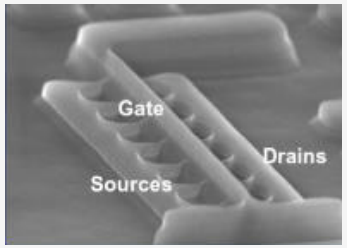
1. 2D ASICs
2. CMOS scaling (and problems)
3. 3D integration
4. Applications and benefits
5. Conclusion

Two approaches (not mutually exclusive)

Microscopic scale

Multi-gate transistors (fin-FETs)





Intel (2011): 1/2 power dissipation gain with ~35% more speed

Monolithic integration

Stacked FEOLs

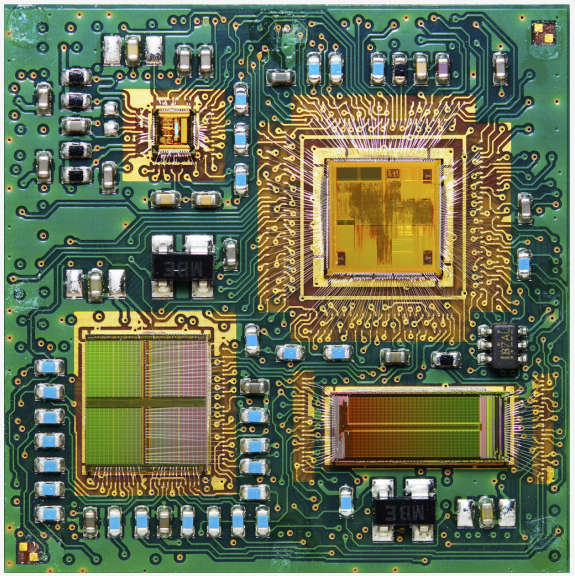
Macroscopic scale

Multi-die integration in one package

System-in-Package (SiP),

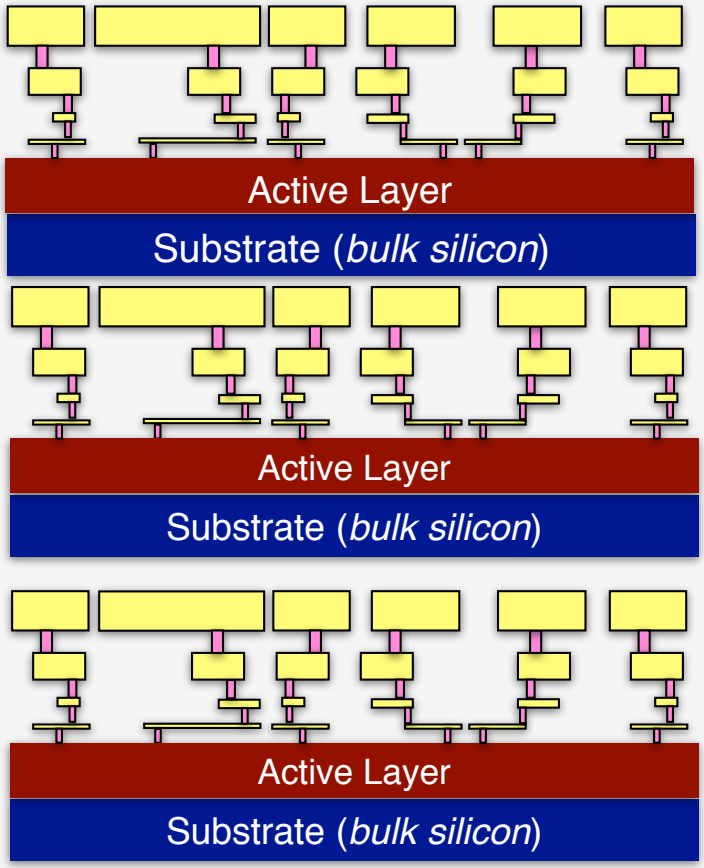
Multi-Chip Modules (MCM)

→ multiple ICs implemented as one package horizontally or vertically



Vertical 3D integration

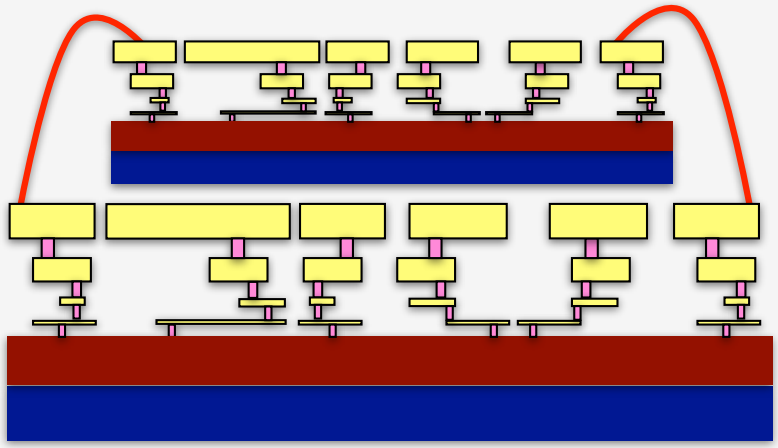
- 3 Dimensional Integrated Circuit (3D-IC): two or more layers of active electronic components integrated into a single circuit (package)
- Many ways on how to do 3D integration: **different 3D structures** allow die-to-die connection

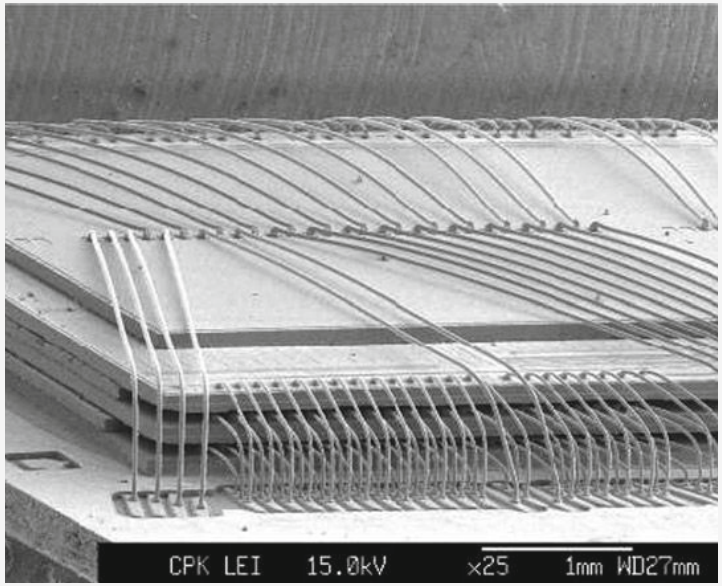


- CMOS is planar technology (planar sounds like 2D)
- In 2D 3rd dimension is used for Metal layers and Vias only, not for active devices
- How to exploit the 3rd dimension?

1. Wire bonding

- Peripheral routing
- Huge pitch
- Limited N° of connexions with bad performance
- **THIS IS NOT VIABLE !**
  (although one of the iPhones used this ...)

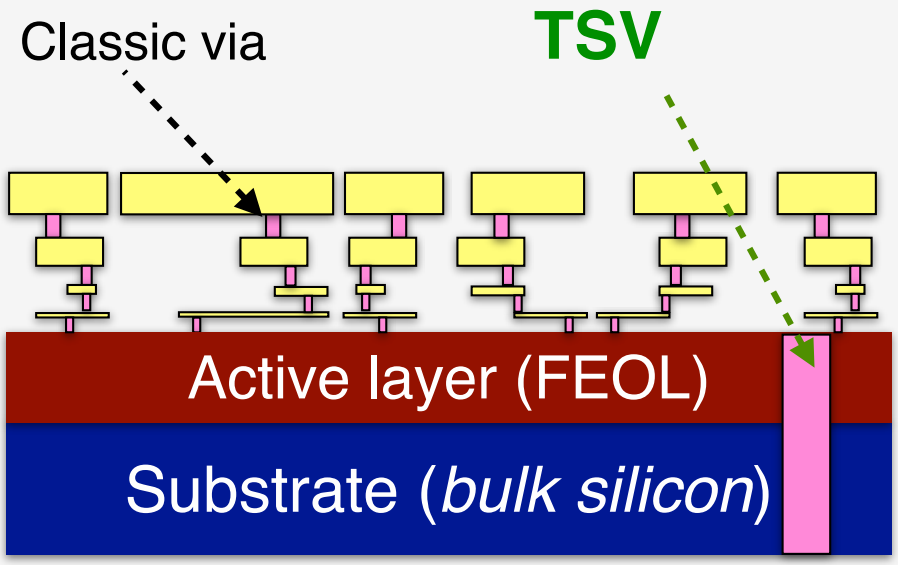


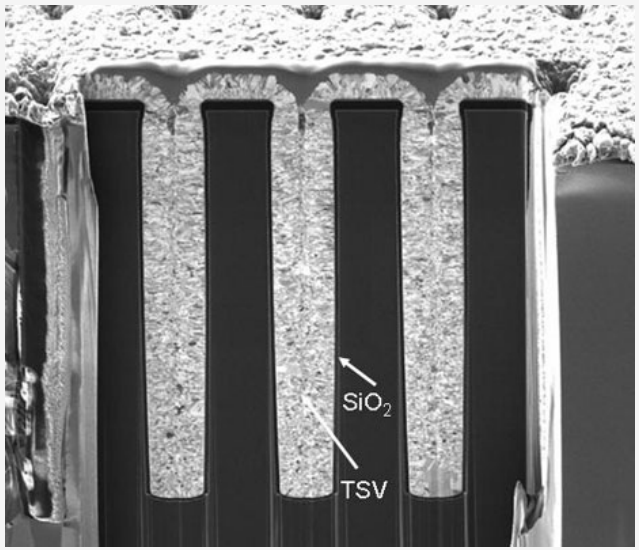


2. Through Silicon Vias (TSV)

- TSV = connection(s) from the front (active layer) to the back-side of the die
- Direct die-to-die routing
- Small pitch (<10μm)
  - ➡ Huge number
  - ➡ Fast connexions

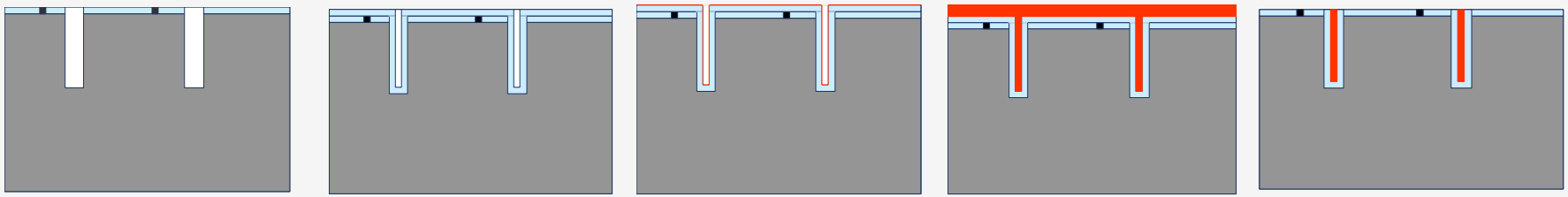
Viable technology !!!





TSV processing: via first

1. TSV Manufacturing



Deep silicon etching

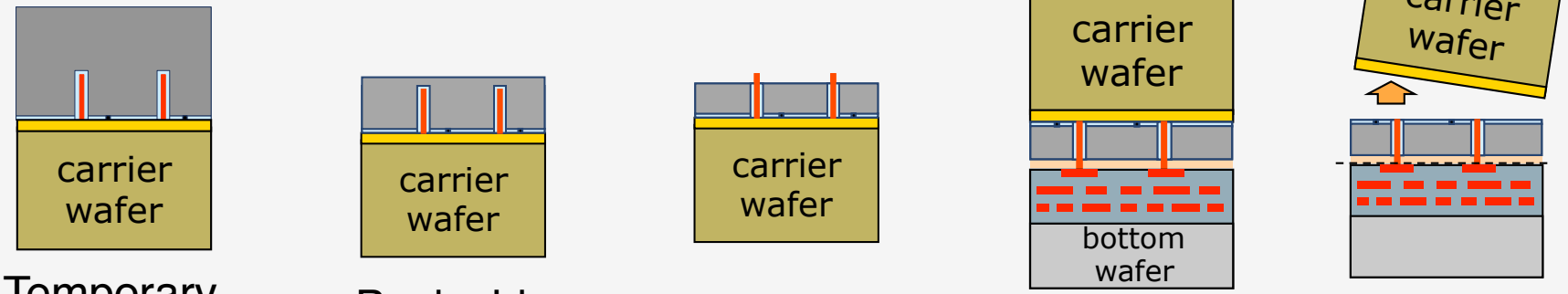
Via oxide deposition

Cu seed deposition

Cu Plating

Chemical Mechanical Polishing

2. Wafer Thinning and Bonding



Temporary carrier bonding

Back side thinning

Exposed Cu nails

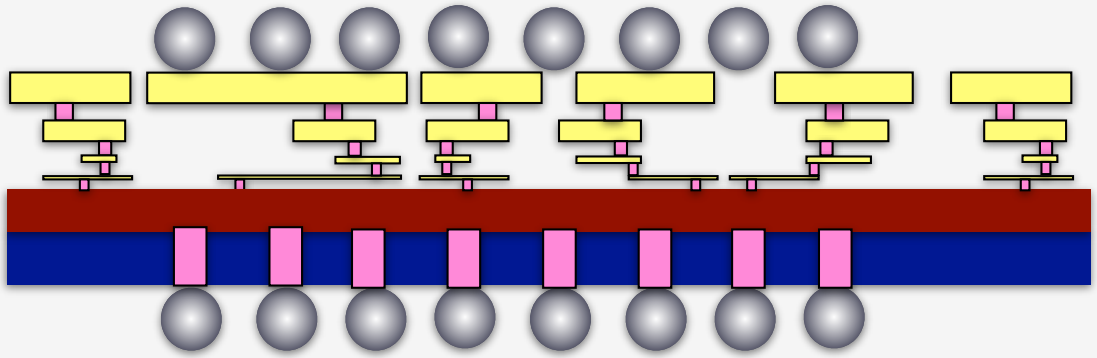
Permanent bonding

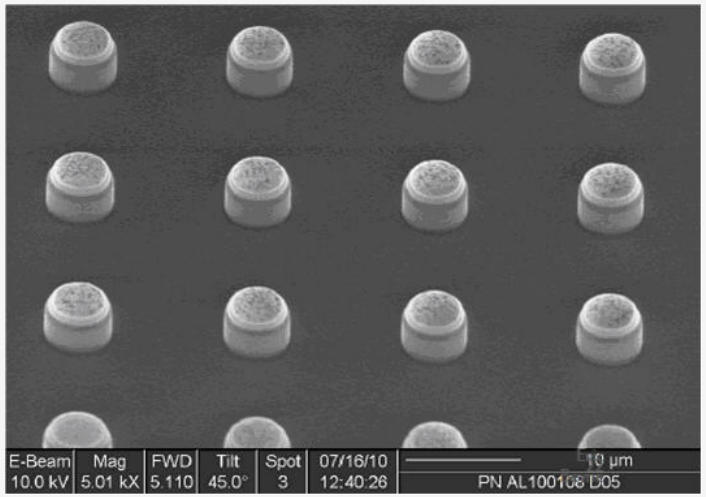
Temporary carrier de-bonding

3. Micro (μ) bumps

- Processed on the top of the front side (top of the BEOL)
- Pitch ~ 30μm (aggressive 10μm) (high pitch but low cost)
- VIABLE TOO !!!
- Direct die-to-die routing
- Small pitch (<10μm)
  - ➡ Huge number
  - ➡ Fast connexions

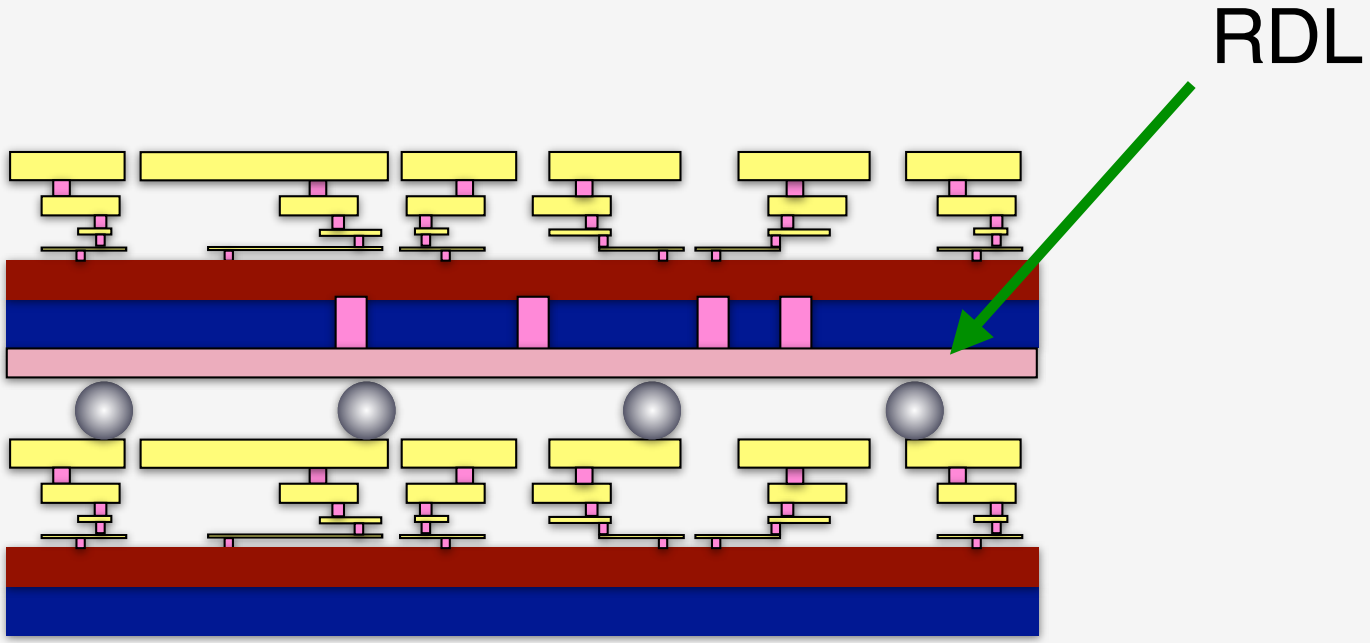
Viable technology !!!





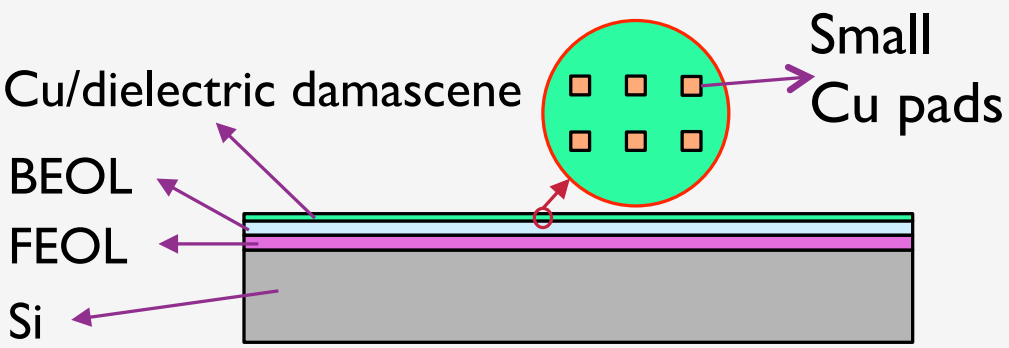
4. ReDistribution Layer — RDL

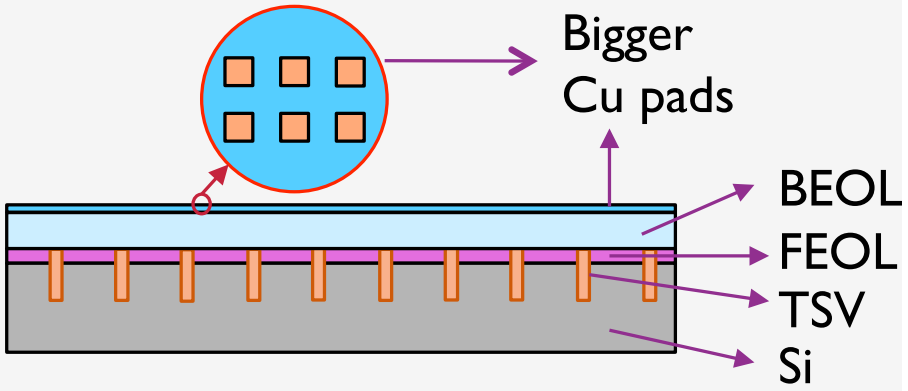
Metal layer on the backside of the existing die, can be used route TSVs and *µ*bumps



TSVs and *µ*bumps do not have necessarily to be aligned → more freedom for the placement & route of the TSVs on the top die.

5. CuCu bonding 1/2



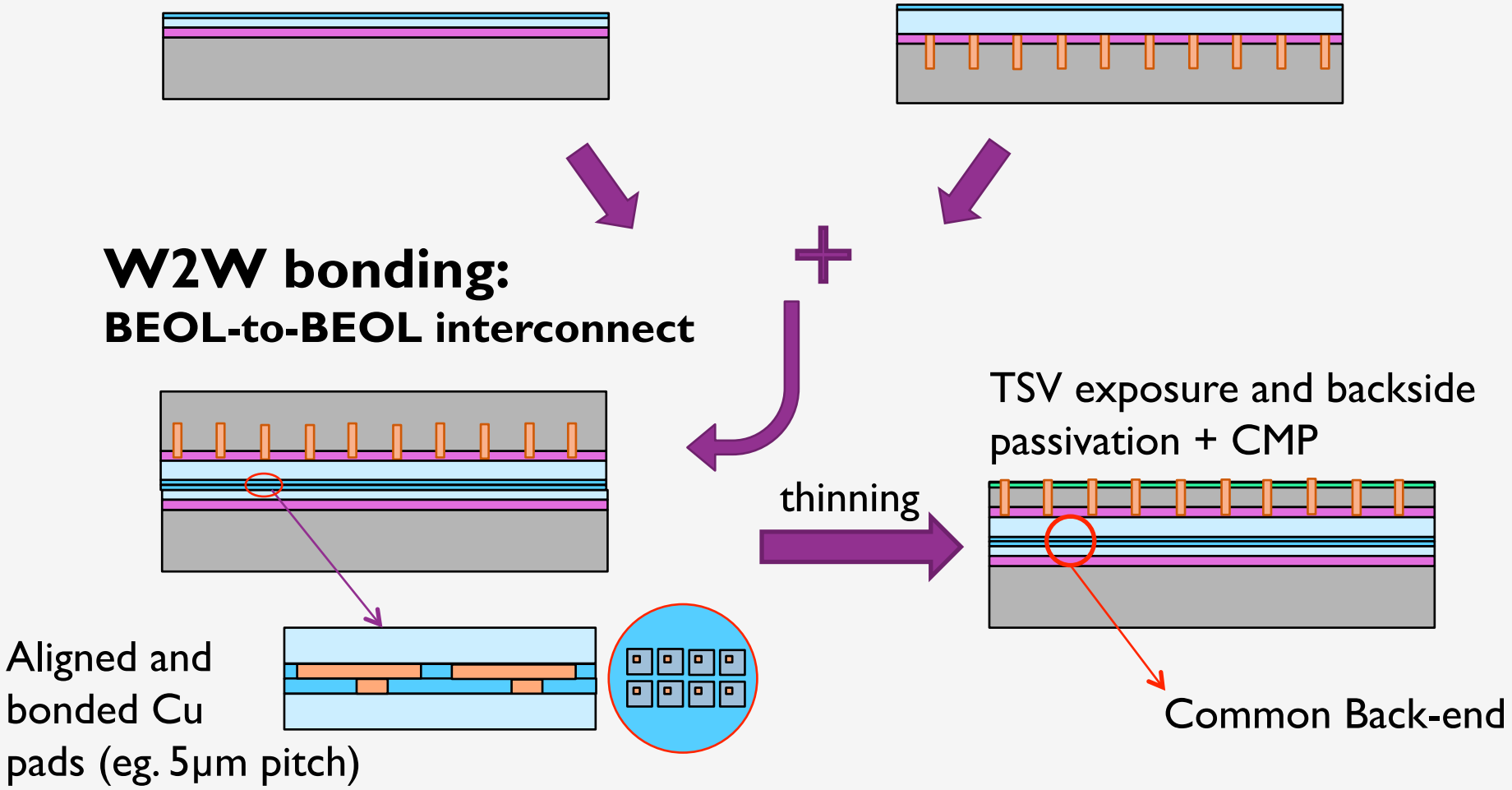


- No TSV & regular FEOL
- Contact pads below 1x1μm2
- Full or limited back-end interconnect stack, depending on application

- Via-middle TSV (after FEOL)
- Contact pads below 4x4μm2
- Full or limited back-end interconnect stack, depending on application

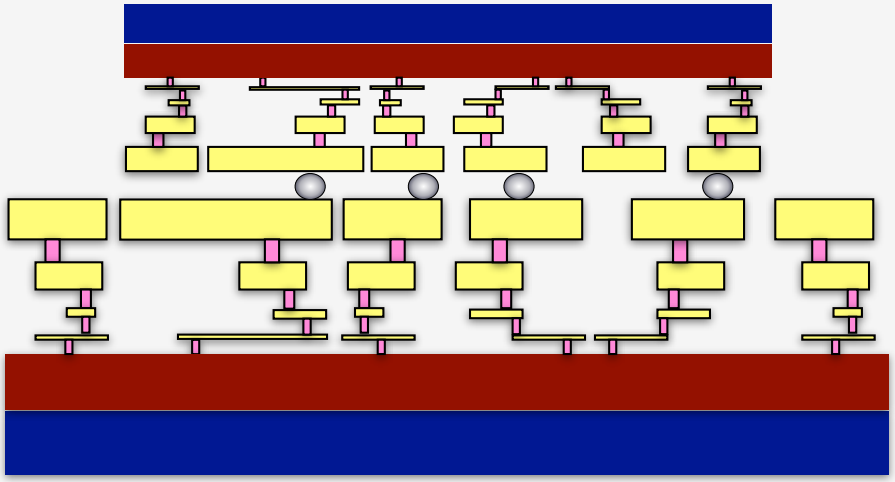
5. CuCu bonding 2/2

N+1 — Advanced process



Flavors of 3D Integration 1/3

a) Face-to-Face (F2F)

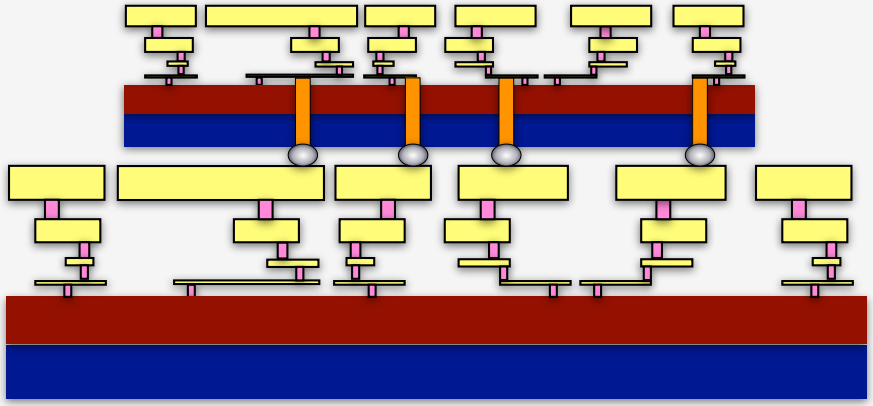


- Based on ubumps or CuPads
- Still need TSVs for the IOs, but typically the number of IOs is not that high (few hundreds to thousand)
- For small 3D structure pitch, allows integration of many die-to-die connections

Flavors of 3D Integration 2/3

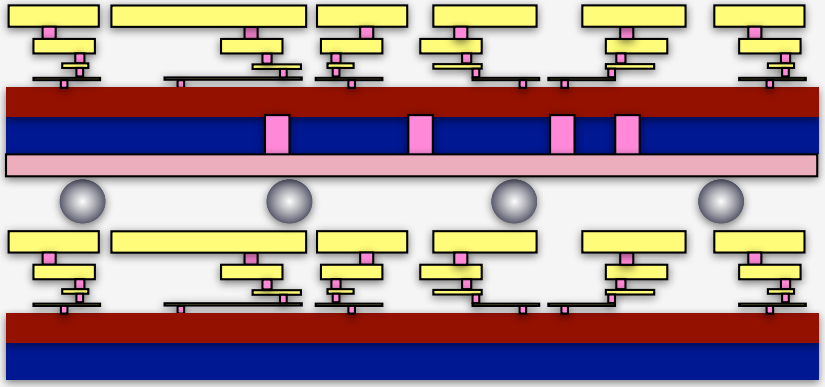
b) Face-to-Back (F2B) no RDL

- ubumps and TSV need to be aligned
- Appears like a constraint for physical design
- Arbitrary number of dies (DRAMs 8 or more)



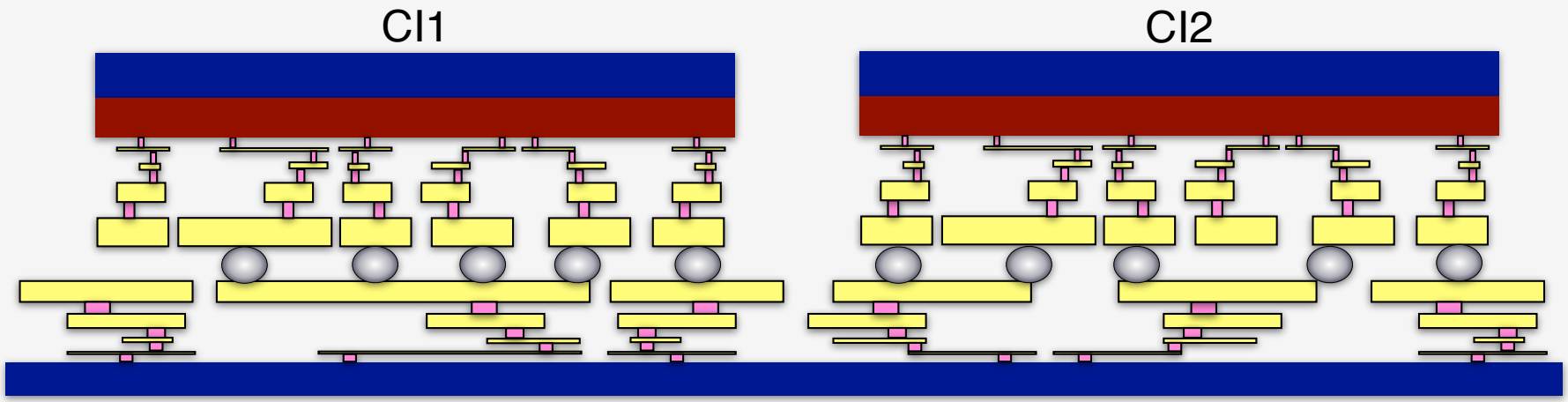
b') F2B with RDL

- ubumps/TSVs do not need to be aligned
- Adds extra cost for RDL processing
- RDLs can not be that long (no active area for repeaters)



Flavors of 3D Integration 3/3

c) *Silicon Interposer*



Only bulk silicon with BEOL:

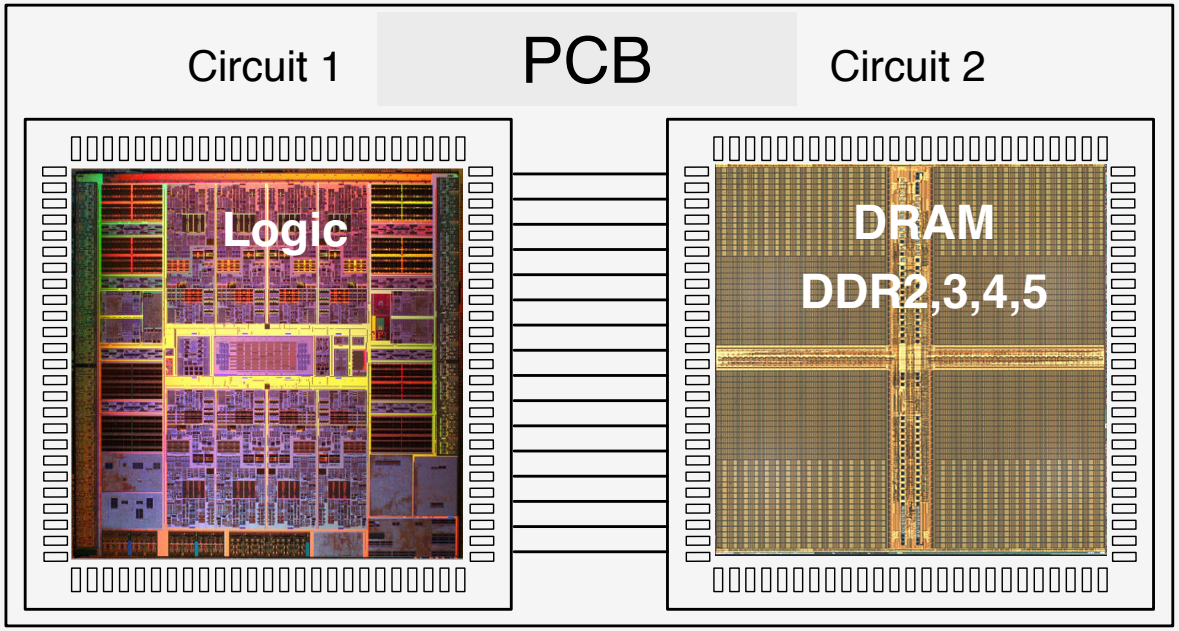
→ No active devices; semi-active interposers are in vogue

→ Basically a reticle size limited routing resource

→ Looks like PCB, but at much smaller scale

What to do with 3D?

- **Inter die connection density increases !!!**
- Allow functional block with high IO count (number of pins) to be moved in another die
- Blocks can be either from the existing design or from the outside of the package (PCB) → Example off-chip DRAM



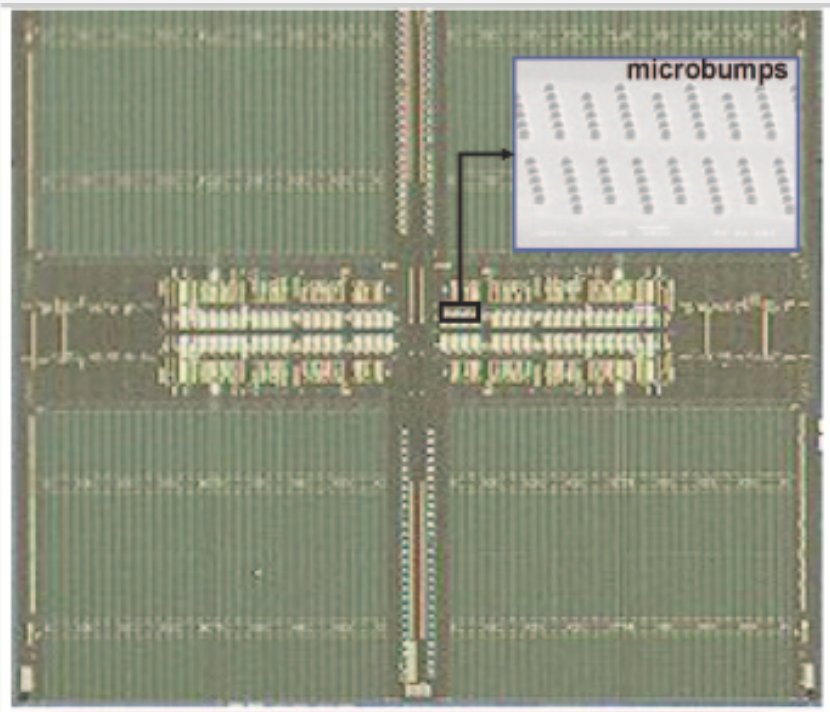
PCB wires

huge capacitive load to logic circuitry → big drivers that are **area and power hungry!**

Example

- If IO is that cheap, why do not increase the datapath width?
- Until today the pin cost is main blocker for this approach
- With 3D integration, this is not true any more
- Birth of new possibilities
  → **Wide IO DRAMs**
  instead of 64
  → **1200 bit wide data bus !!!!!**

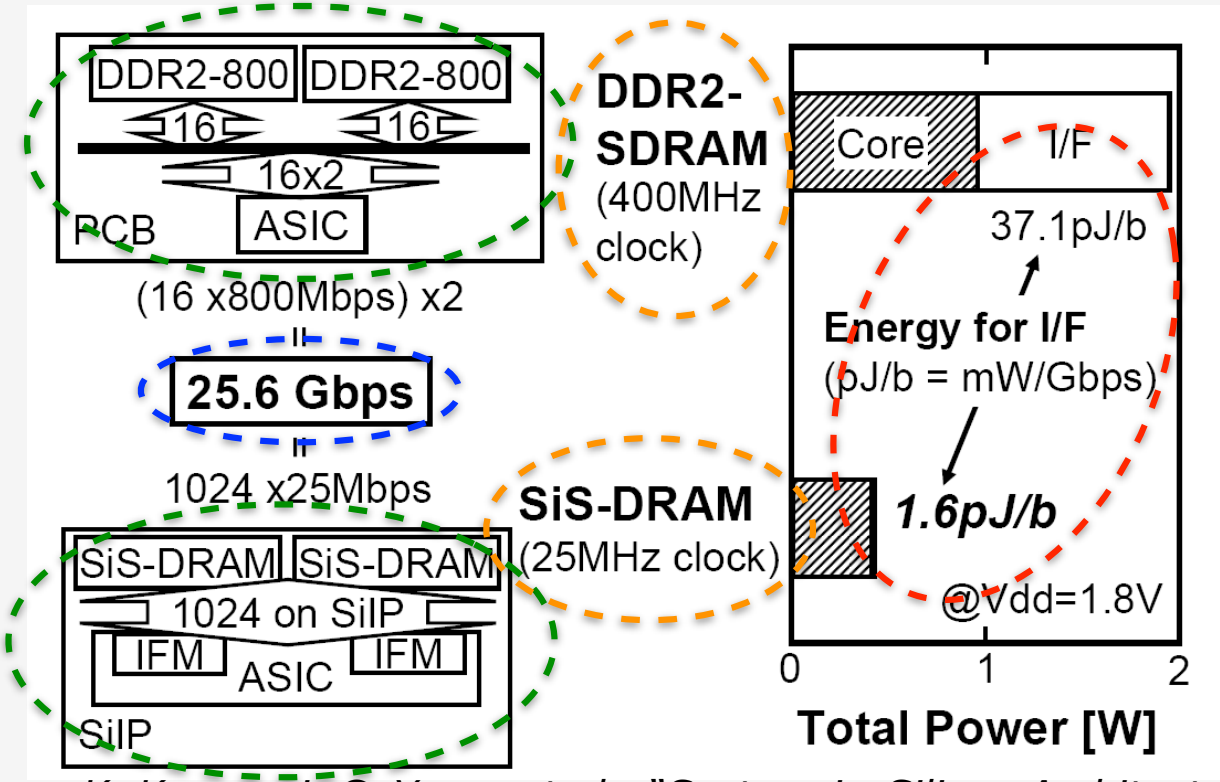
SAMSUNG
WideIODRAM



Less load capacitance mean smaller drivers, less area, power so better access to DRAM (that is the bottleneck from the system perspective anyhow)

Wide IO DRAM power savings

The N° of TSVs does not influence the cost of manufacturing; impact is on area overhead, but @5µm diam. and 10µm pitch this is not an issue any more; huge impact on design:



Increasing the data path width:

16x < F

= BW

23x < P

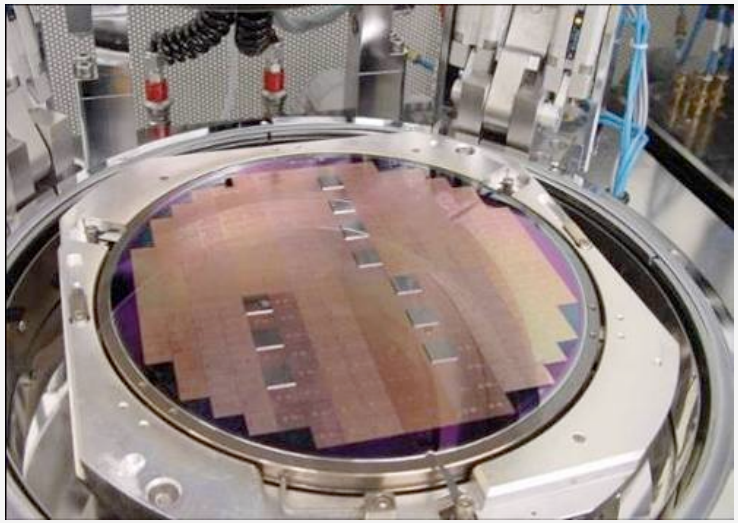
K. Kumagai, C. Yang, et al., "System-in-Silicon Architecture and its Application to H.264/AVC Motion Estimation for 1080HDTV", ISSCC 2006.

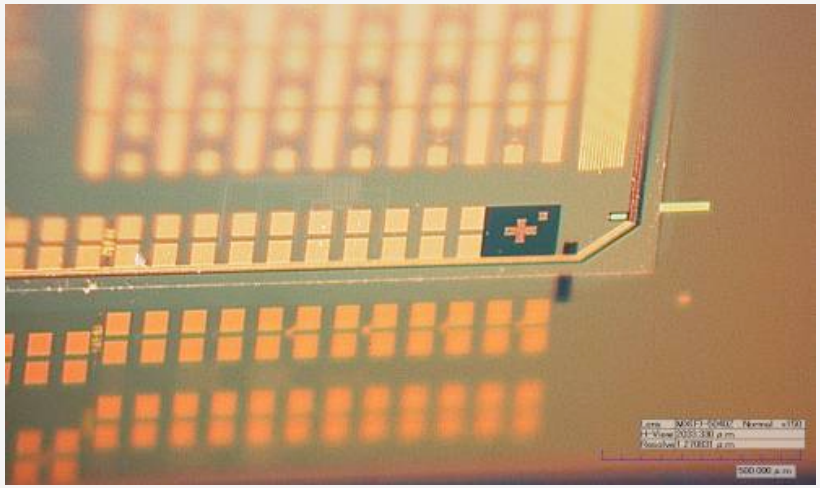
3D integration: advantages

- **More functionality –** Increased density for the same footprint and a little bit bigger volume (important for mobile)
- **Smaller delays –** Closer, tightly coupled blocks, shorter wires
- **Lower power –** Shorter wires, mean less interconnect power, but also less repeater insertion (area savings too)
- **Heterogeneous integration –** combine circuits manufactured in different technologies: memory-on-logic, logic-on-logic, devices that don't scale with those that can scale etc.
- **Higher bandwidth –** Huge inter-die interconnect density → thousands, rather then dozens, of die-to-die connections
- **New product opportunities –** design of new systems (e.g WideIO DRAM)
- **Lower cost –** Smaller dies allow better yield and wafer usage

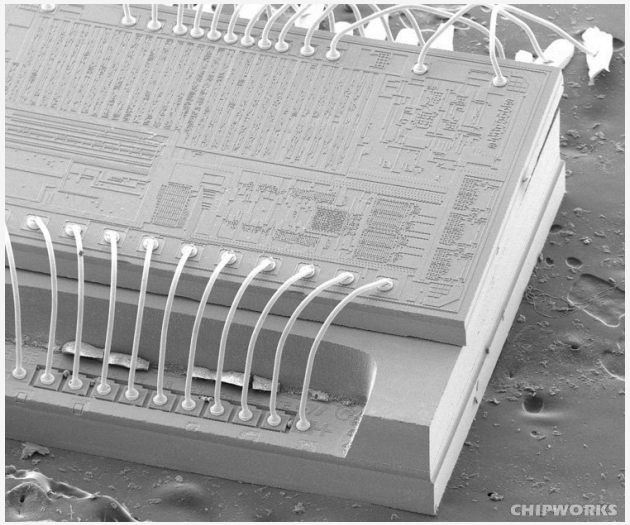
3D is real...

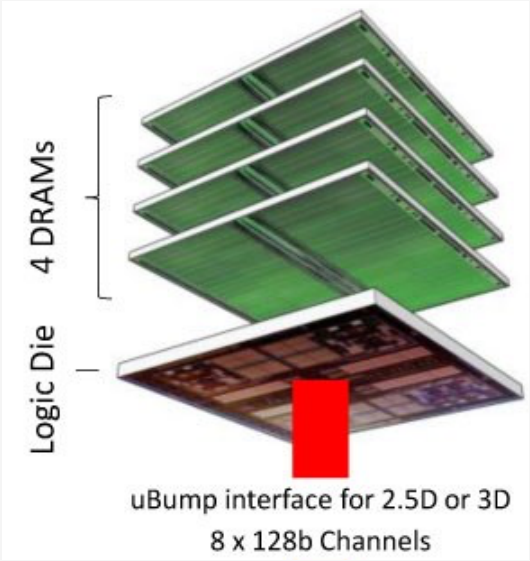
In research



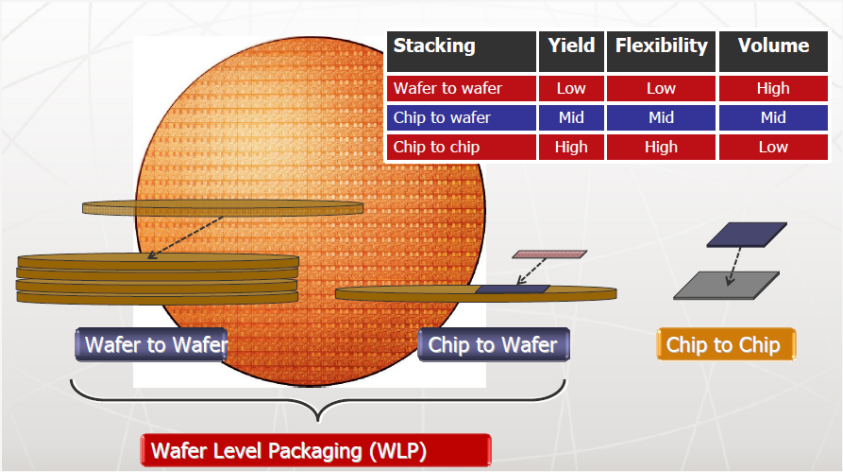


and in practice

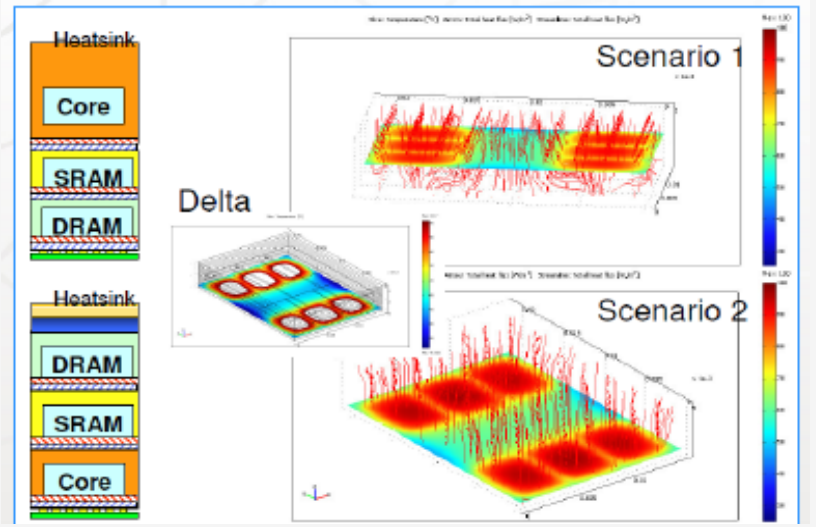




… but is not mainstream tech yet !

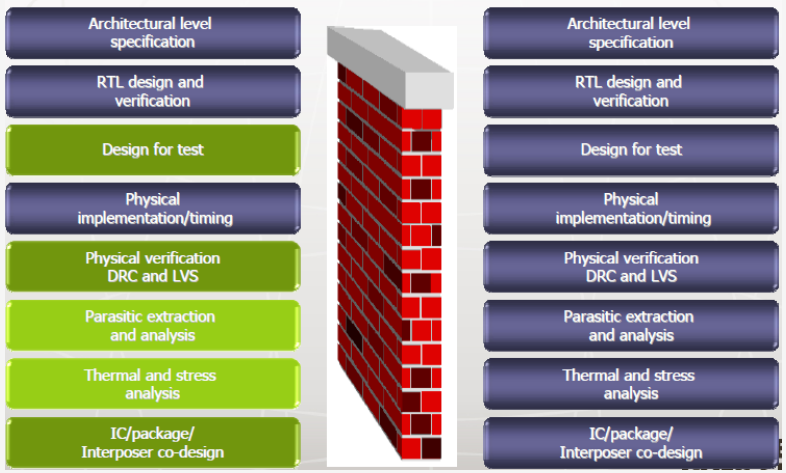


Manufacturing, Test & Yield



Power density, peak temp.

Standardization, supply chain



Design flow

Outline

1. 2D ASICs
2. CMOS scaling (and problems)
3. 3D integration
4. **Applications and benefits**
5. Conclusion

Design flow for 3D

Design Planning

- Perform design exploration early in the flow
- Based on Atrenta's SpyGlass Physical3D®
- Fast, enables many iterations

Compact models

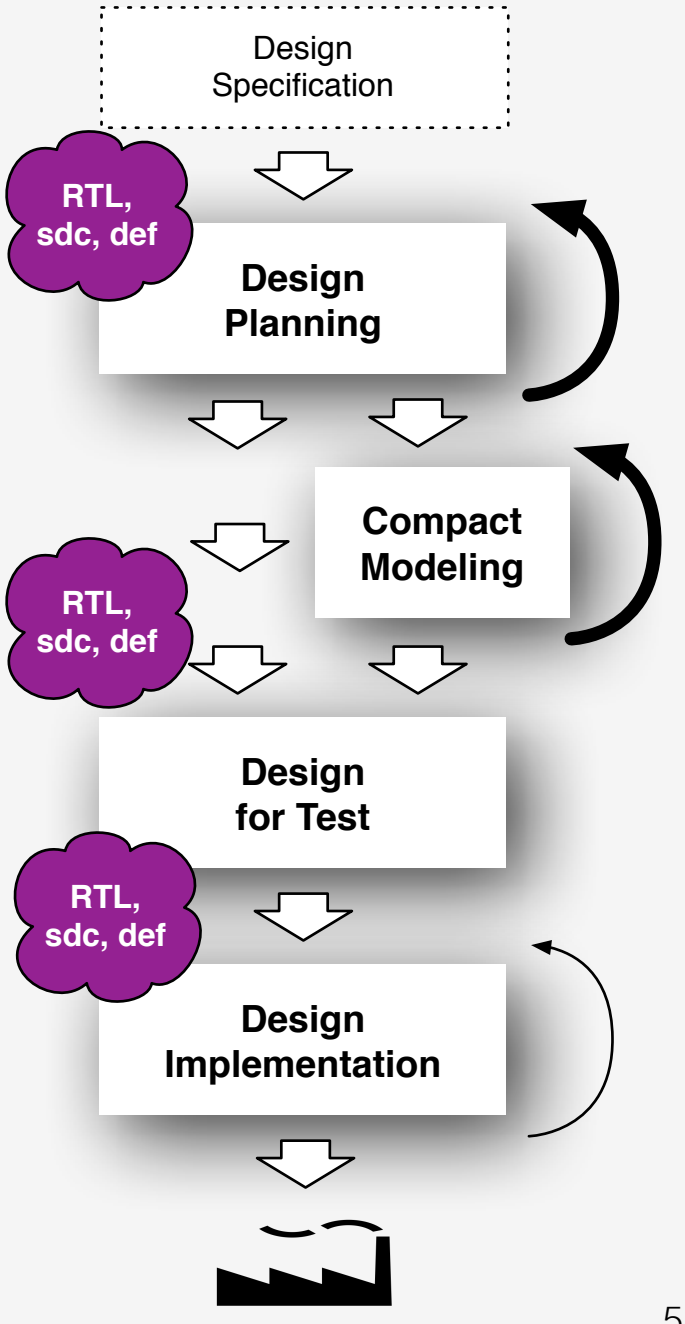
- Check for other design properties
  - thermo-mechanical, delay, cost, ...
- Fast & accurate since validated using silicon

Design for test

- Automated addition of DfT structures

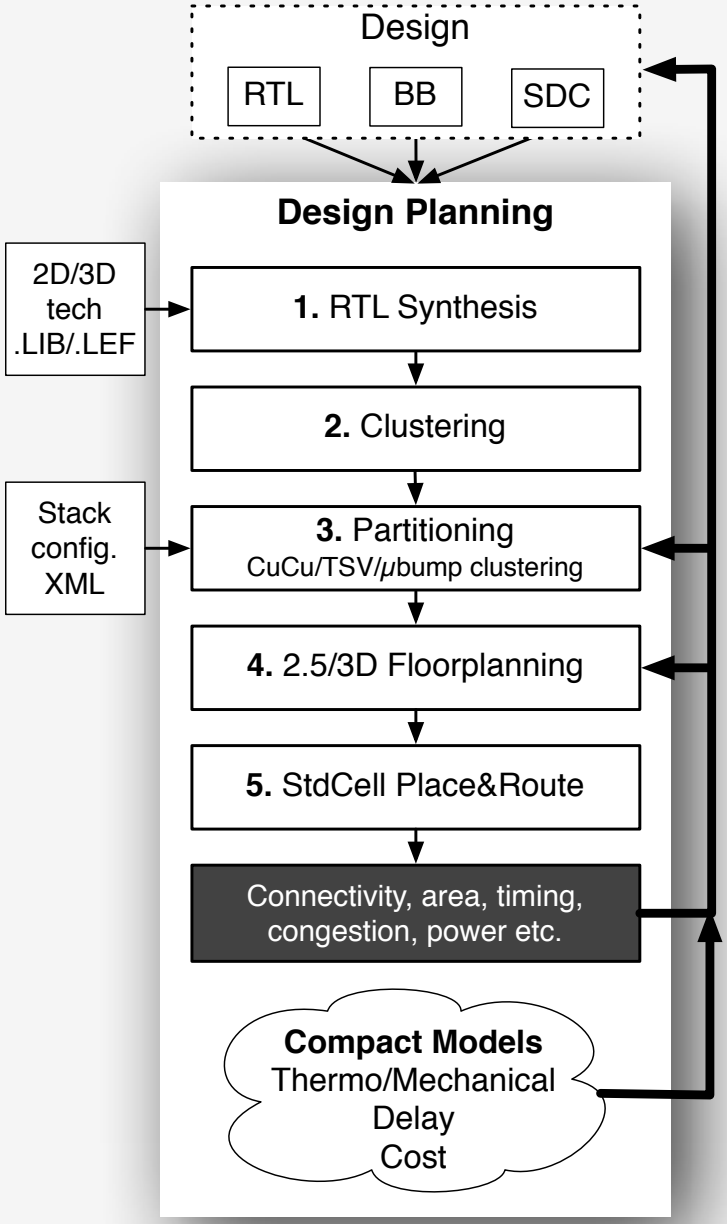
Design implementation

- Generate the actual GDSII
- Minimize the number of iterations in the bottom parts of the flow
- Any industry standard back-end flow



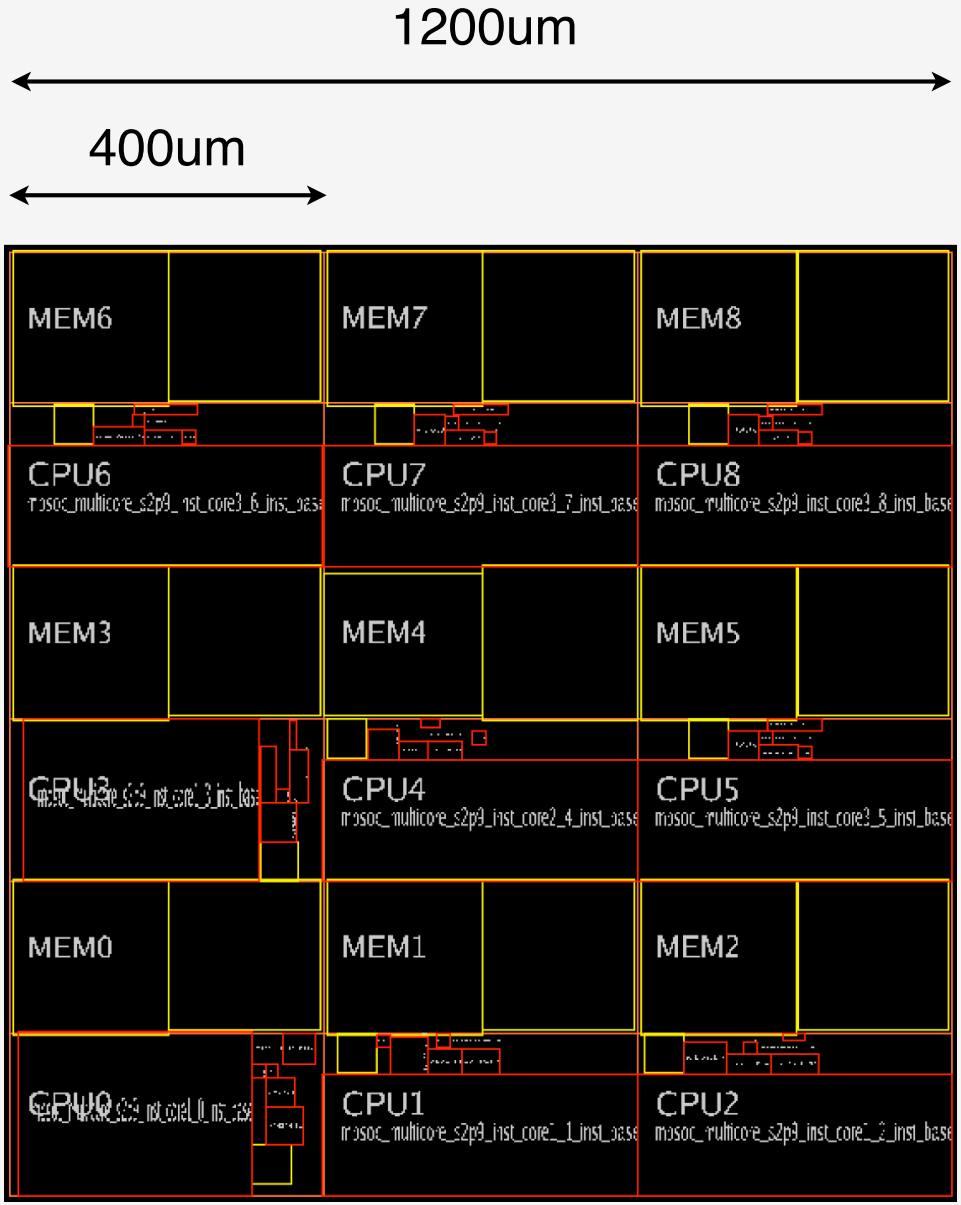
Design planning: our 3D flow

- Inputs
  - Incomplete design specification support (RTL + BlackBox)
  - Industry std. constraints (.sdc)
  - Fully technology aware flow (.lib/.lef)
- Fast synthesis and physical clustering
- Flexible 3D configuration (.XML)
- Automated 3D gate-level netlist partitioning
- Automatic inter-die net extraction
- Support for TSV/μbump clustering, P&R, technology features exploration
- Std cell placement & front/RDL routing
- Links to Thermo-Mechanical/Delay/Cost Compact Models



Mobile MPSoC architecture

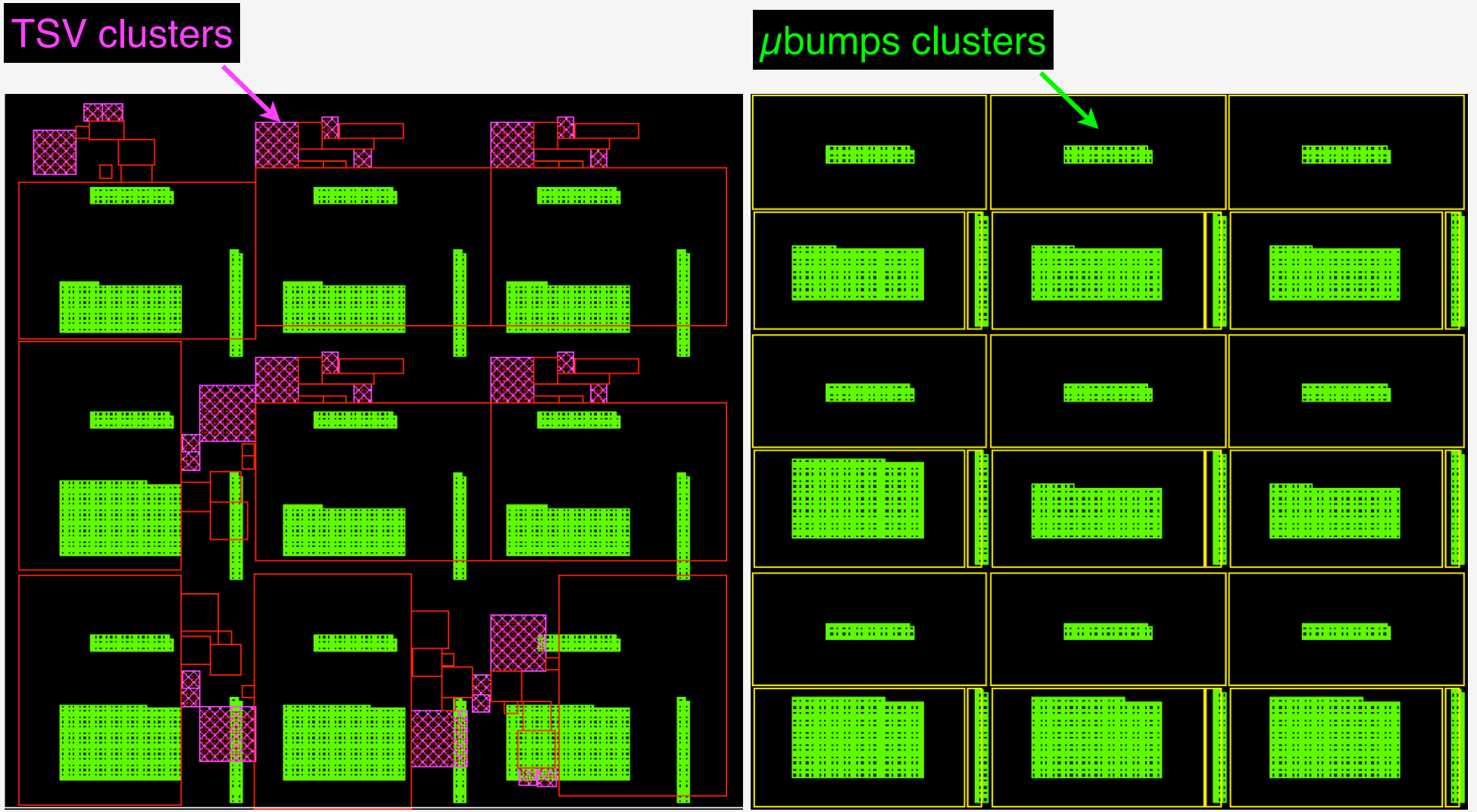
- MPSoC with 9 cores, bus interconnect
- 3 different architectures for cores
- Small cores (~.16mm^2/core in 28nm)
- L1/L2: 3 memory instances per core (total 64kB/core) ≈ size of core
- Different memory interface sizes: 641 and 449 pins
- 2D and 3D implementations (memory-on-logic)
  - Equal partitions (W2W)
  - ~5k 3D wires (signal wires only)
- 3D flavors
  - 3D Face-to-Back (3D F2F)
  - 3D Face-to-Face (3D F2B)



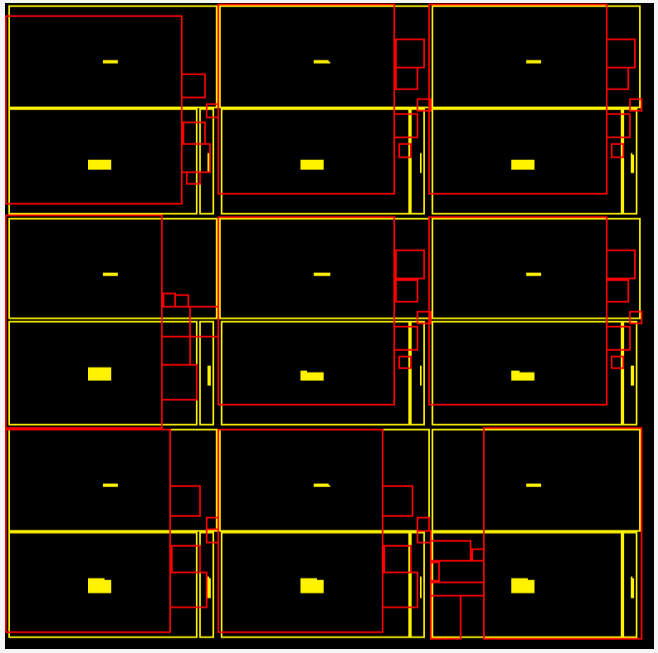
3D F2B: 6/10μm TSV/μbump pitch

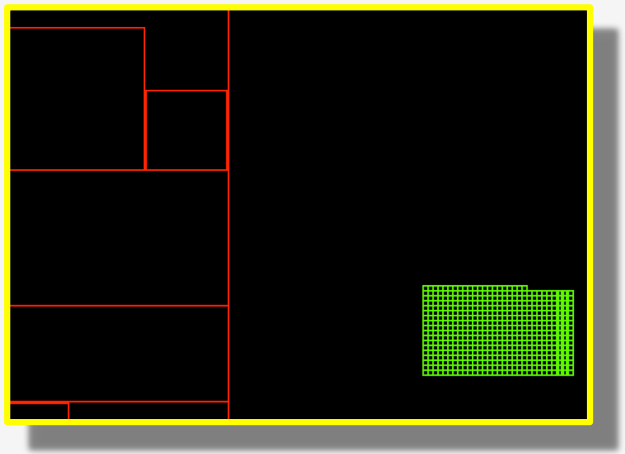
Bottom die

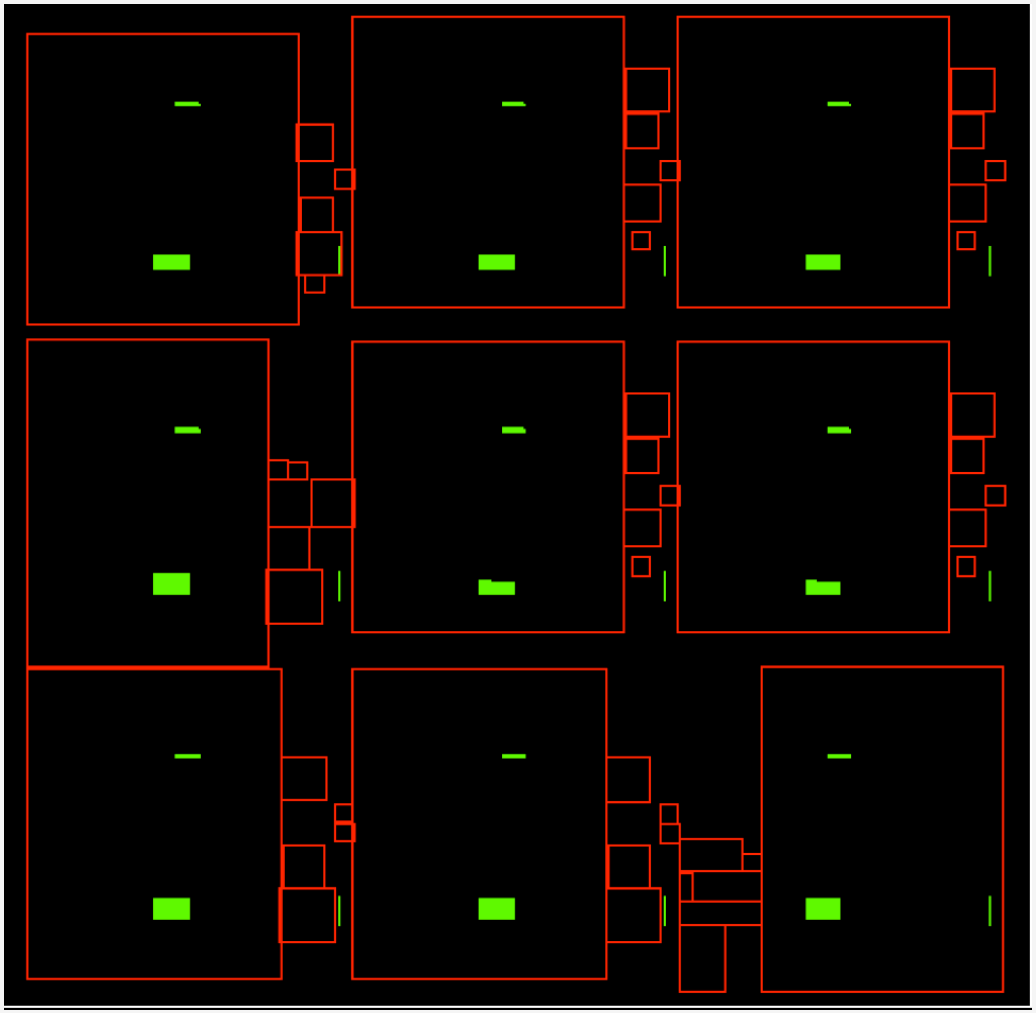
Front die



3D F2F: 1μm Cu-Pad pitch



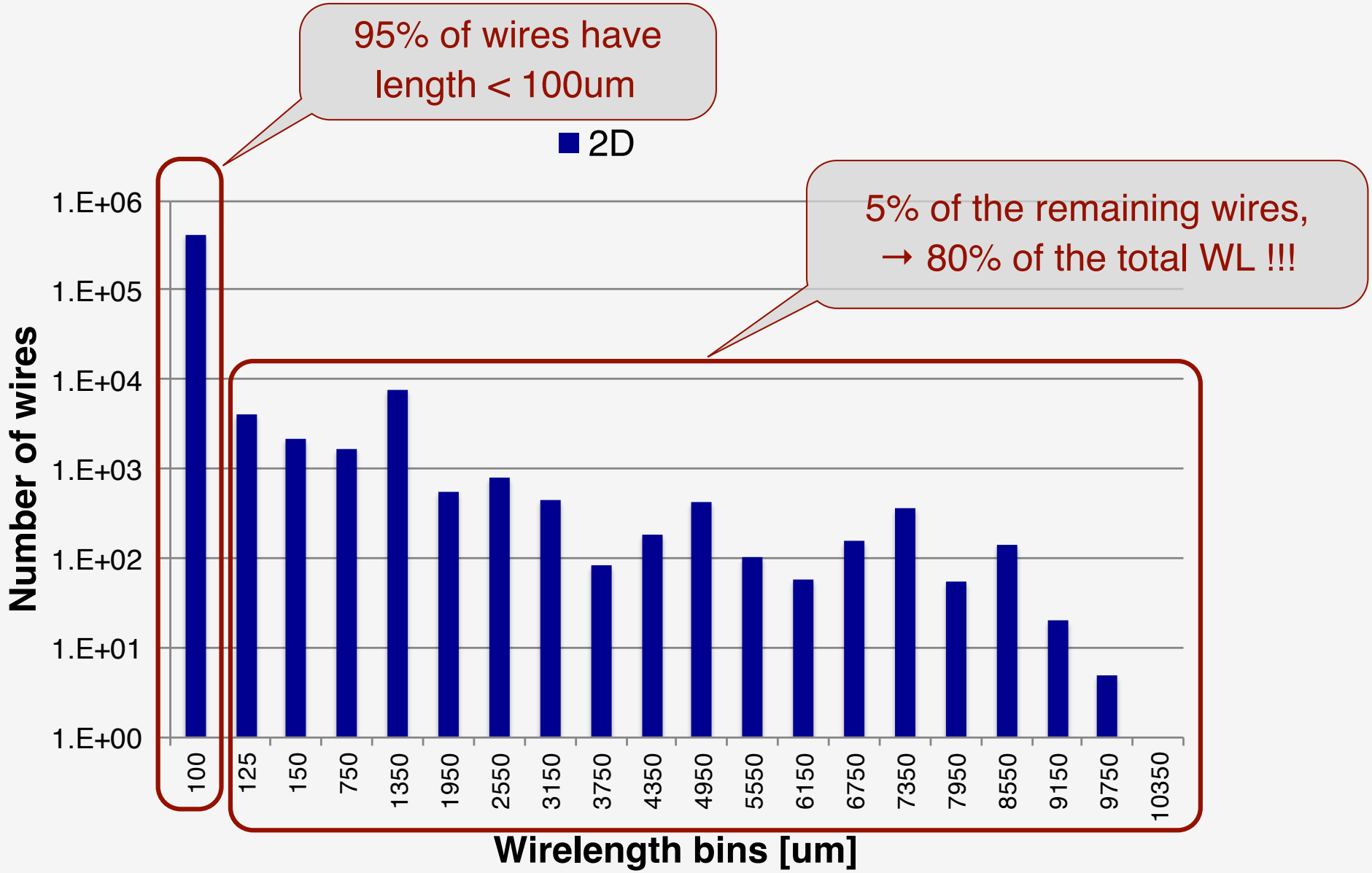




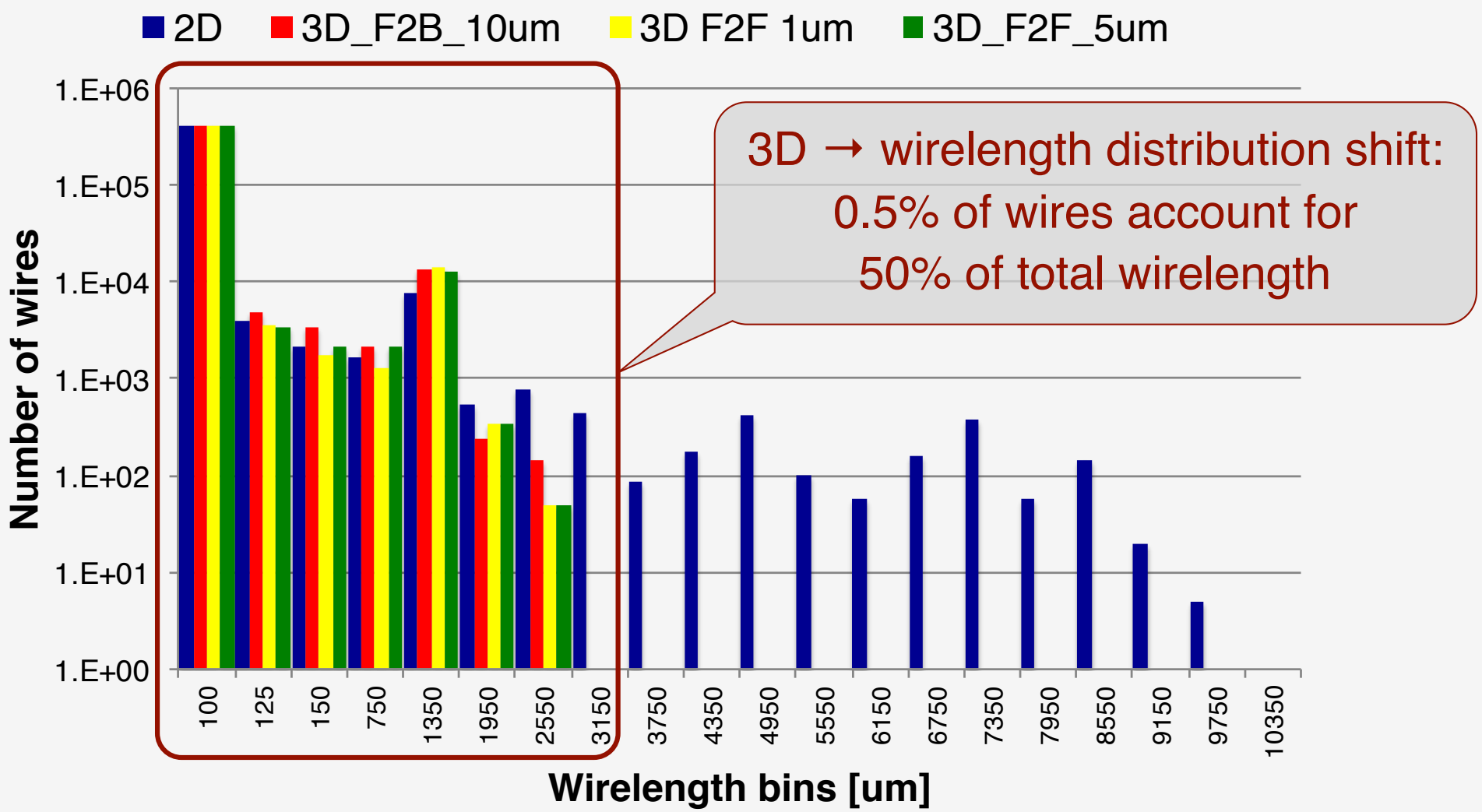
Design characterization

- 2D vs. 3D total wirelength distribution analysis (total and critical path length analysis)
  - Impact on BEOL congestion and cost (Nº of metal layers)
- Critical path delay (due to wires)
  - Impact on performance
- Area savings
  - Impact on cost
- Block-to-block interconnect power
  - Cost, cooling and autonomy

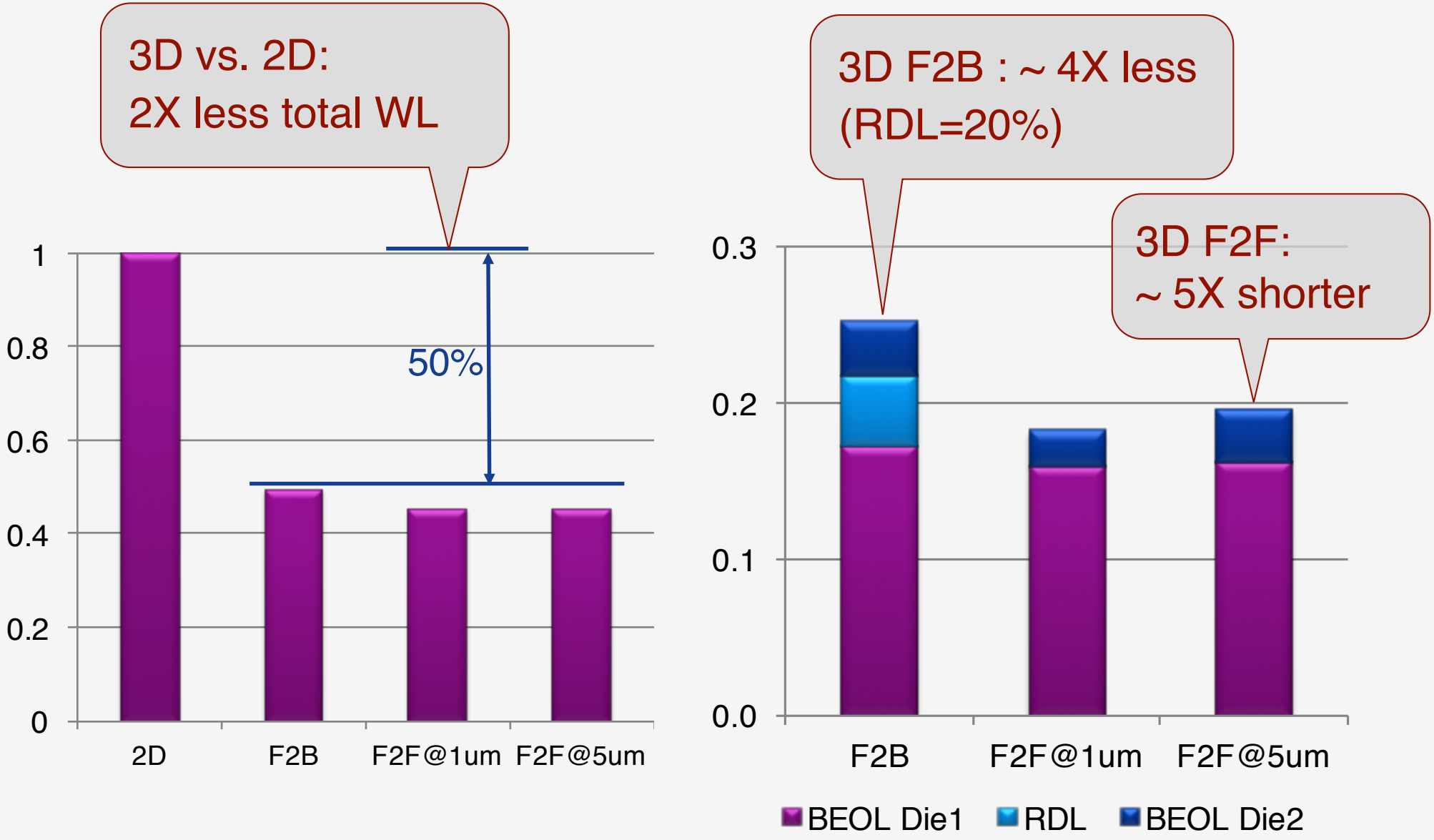
Wirelength distribution 2D



Wirelength distribution 2D + 3D



Total and critical path wirelength



Area savings

[1] before timing optimization (post logic synthesis & optimization)
[2] very aggressive, needed to compensate repeater insertion
[3] more realistic TSV size
[4] assuming 120 IOs, and C4 bumps to connect 2D, F2B to package
[5] assuming identical repeaters

Area [%]	2D	3D F2B	3D F2@5um	3D F2@1um
Total std. cell area[1]	100	100	100	100
Signal TSVs (~5k wires) ø3µm [2] ø5µm [3]	–	4.2 10.0		–
IO TSVs (ø10um)	–	–	1[4]	1[4]
Repeater area[5]	8.7	2.5	2.4	2.2
Total overhead	**8.7**	**6.7**	**3.4**	**3.2**

For realistic TSV sizes, no area gain

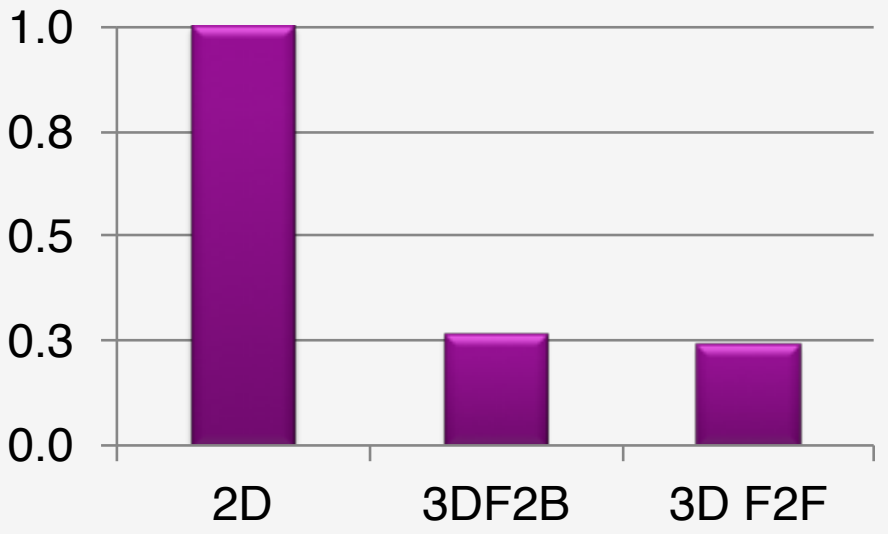
Critical path delay

- 2D: 8900um wire
- F2B:
  - 1500um Bottom die
  - 400um RDL
  - 400um Top die

  → 70% reduction vs. 2D
- F2F:
  - 1500um Bottom die
  - 400um Top die

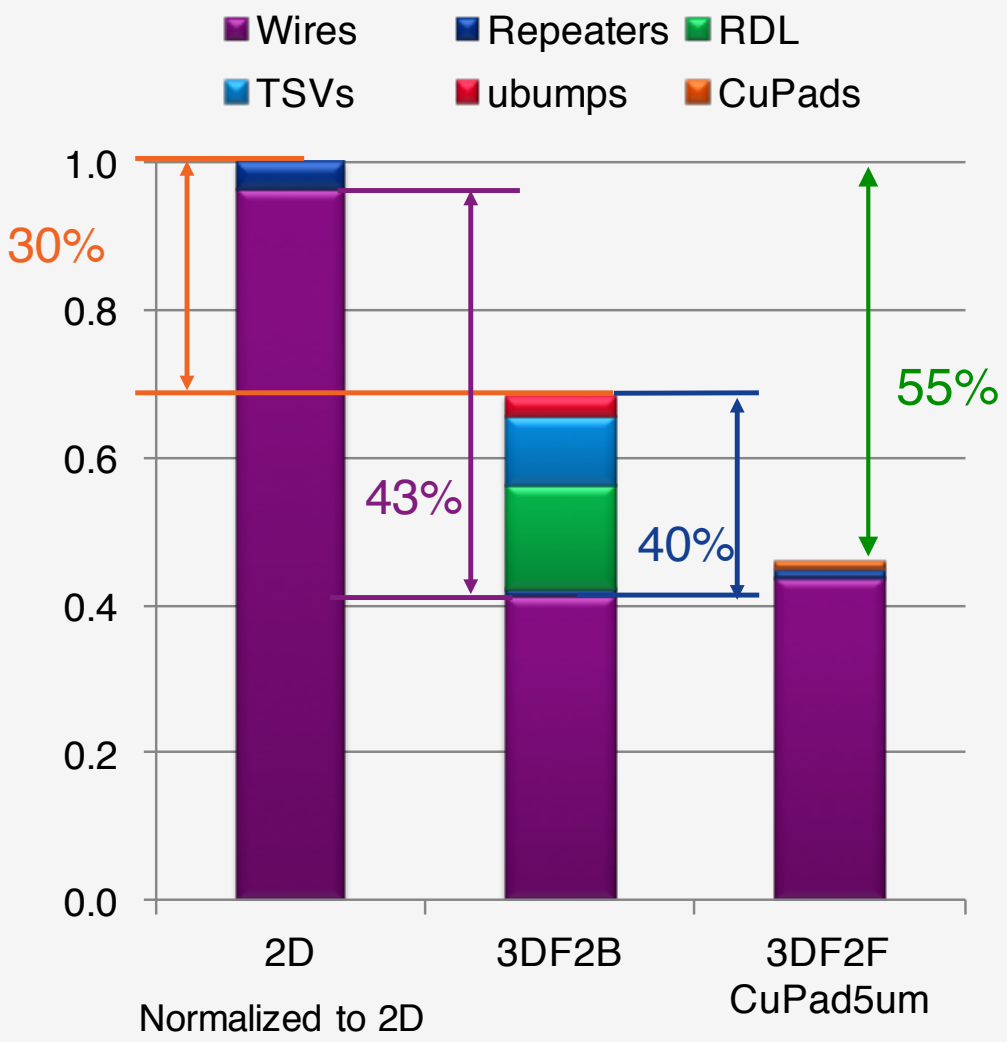
  → No significant gain compared to F2B

RC parameters		
	R [mOhm]	C [fF]
[1]TSV 5X50um	22	55
[1]ubump 25um	50	16
[1]RDL[/um]	4	0.4
[2]Cu Pad	100	3



Interconnect power

- 3D-F2B vs. 2D:
  30% less total interconnect power compared
  (**43%** less net wire power)
- F2B: **40%** of the total power in 3D nets
  (triplet TSV+RDL+ubump)
- 3D-F2F vs. 2D: **55%** less interconnect power
  (net wire power gain similar to F2B, but less for 3D nets



Outline

1. 2D ASICs
2. CMOS scaling (and problems)
3. 3D integration
4. Applications and benefits
5. **Conclusion**

Technology aspects

- 2D integration is becoming more and more complex because lithography process is reaching the limit
  - Technology is becoming more and more costly, with less and less gains
- Despite, pure CMOS scaling is still on the agenda and it will be there for the next couple of years
- Alternative technologies are required and 3D integration looks like a very attractive option
- Most likely the 3D will enable heterogeneous integration
- But still some things that need to be solved

Practical (design) gains

- Even for small designs (less demanding in global interconnect), **repeater insertion** uses area, resulting in **increased die cost** and **higher power dissipation**
- 3D F2B and F2F improve wirelength distribution: **reduced total wirelength and critical path** (less BEOL cost and stress)
- For fine grain partitioning and small dies, **F2B needs aggressive TSV diameter** (pitch) to enable **area gains** (due to lesser repeater insertion)
- Both F2B and F2F provide better performance compared to 2D due to shorter wires; however power wise F2F outperforms F2B
- For considered die size & inter-die nets count, there is no significant gain in reducing the Cu pad pitch