3D-FPGA

Abbas El Gamal

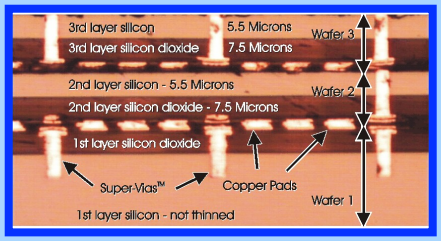
Joint work with: Mingjie Lin, Yi-Chang Lu, Simon Wong

Work partially supported by DARPA 3D-IC program

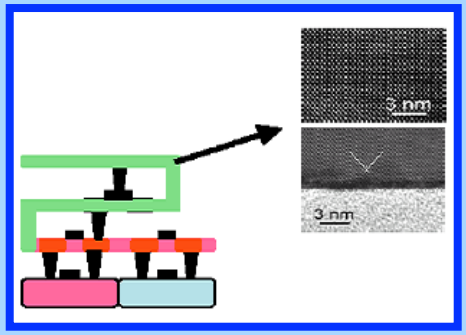
Stanford University

3D-IC Technologies

- Chip stacking
  - Vertical interconnect density < 20/mm

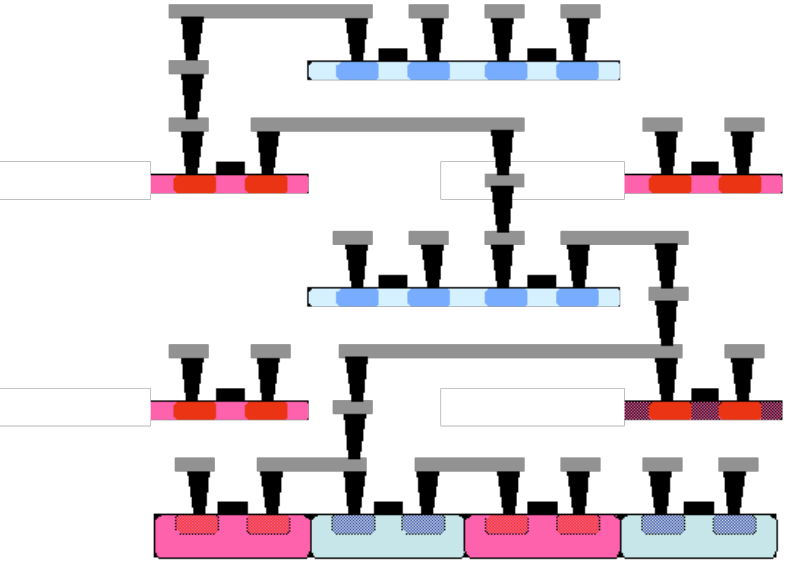


- Wafer Stacking
  - Through Silicon Via (TSV) pitch: 3-5X Via 3
    → 0.54µm-0.9µm in 45nm CMOS
  - TSV pitch today 2-5µm



- Monolithic Stacking
  - Roughly same size via as CMOS
  - Can only add few monolithic layers to CMOS

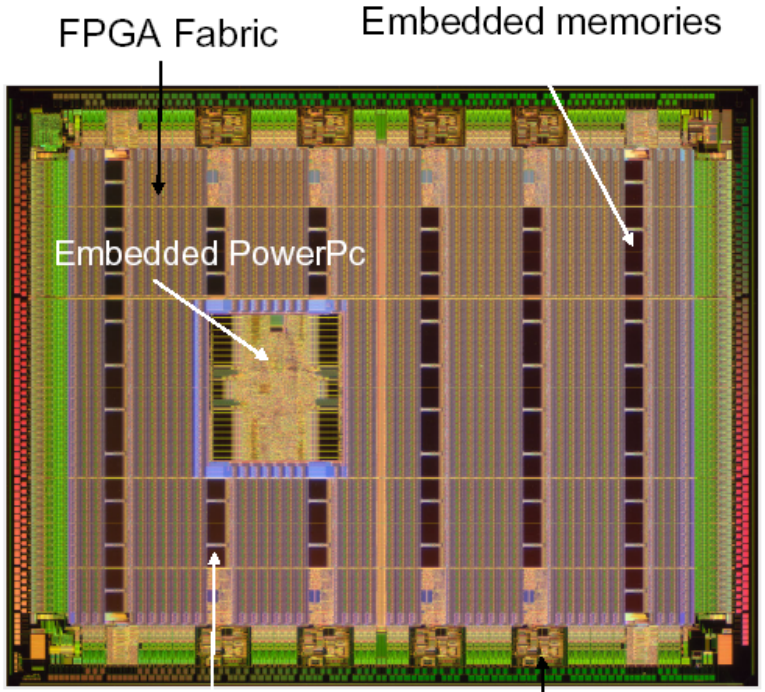
Stanford 3D-IC Program



- Funded through DARPA 3D-IC program
- Interdisciplinary team of nanotechnologists, device physicists, IC technologists, circuit designers, architects
- Goals:
  1. Develop technology to monolithically stack active layers on top of CMOS
     - Challenge: Low temperature processing
  2. Investigate 3D-IC architectures
     - Demonstrate the performance gain of 3D-IC in logic density, delay, and power consumption
     - Demonstration vehicles: FPGA, SRAM

A Modern FPGA

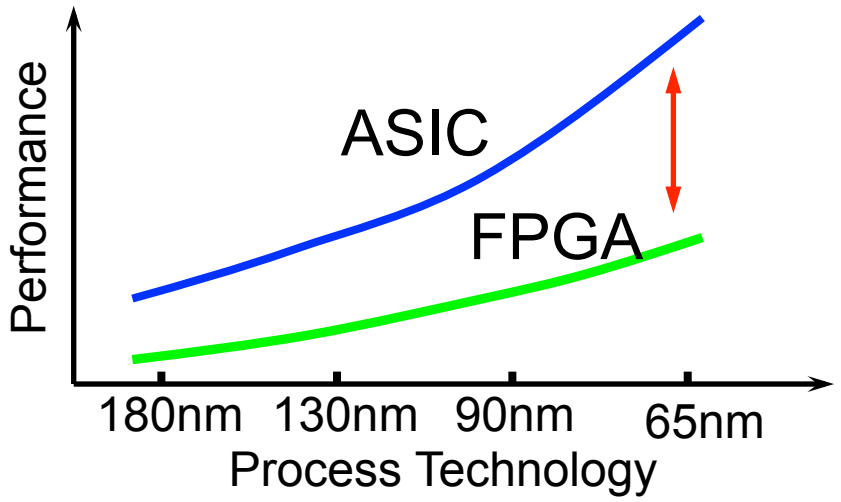
- SRAM programmable
- Fabricated in 65nm CMOS
- Integrates up to 0.33M logic blocks (each roughly 20 gates)
- Up to 10Mb of SRAM block memory
- Embedded microprocessor
- DSP slices
- 550MHz internal clock speed
- 1200 user I/Os including many types of high speed I/Os



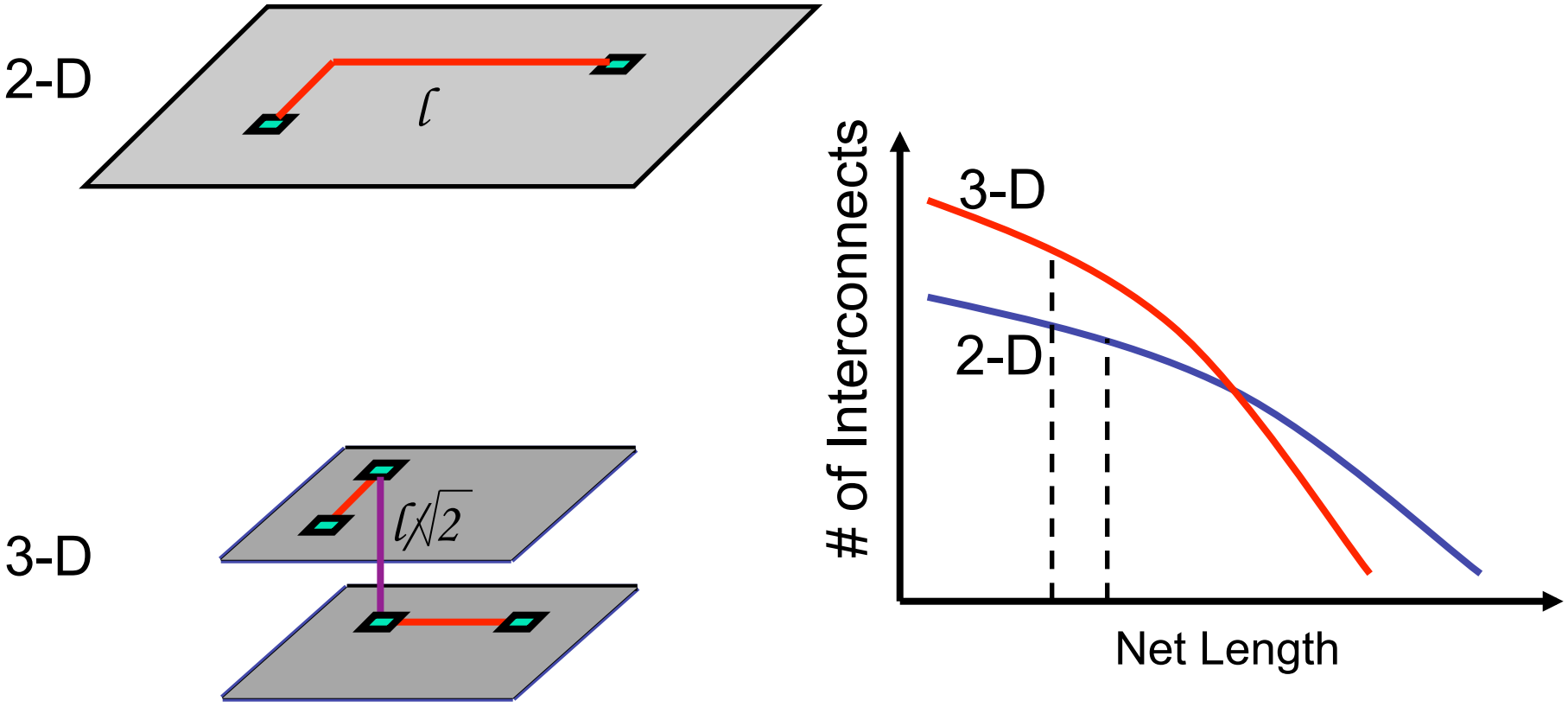
Courtesy Xilinx

FPGA: The Good and the Bad

- FPGAs are becoming increasingly attractive for digital system design:
  - Escalating cell-based ASIC design and prototyping costs in deep submicron
  - Prototyping and field re-programmability
- But, FPGA performance is much lower than cell-based ASIC [Kuon *et al.* 07]:
  - 10-40X lower logic density
  - 3-4X higher delay
  - 5-12X higher dynamic power



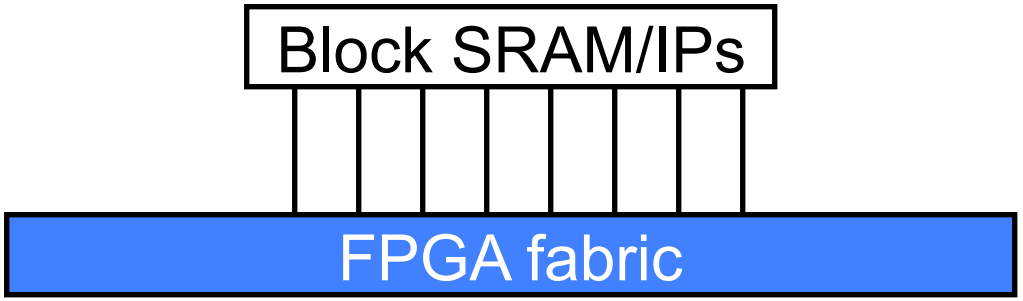
How 3D Can Help Digital



Shorter interconnects → Lower delay and power consumption

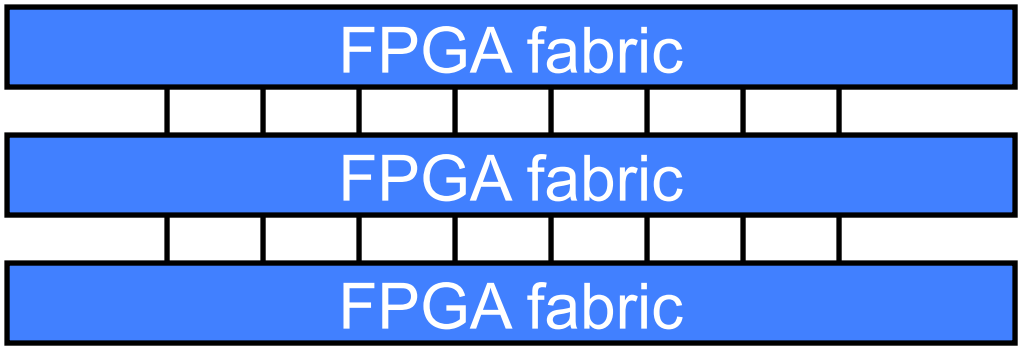
Wafer Stacking: Scenario 1

- Stack block memory and hard IPs on top of FPGA logic fabric
  - Relaxed TSV pitch requirement
  - Delay-power benefits depend on size of IPs included



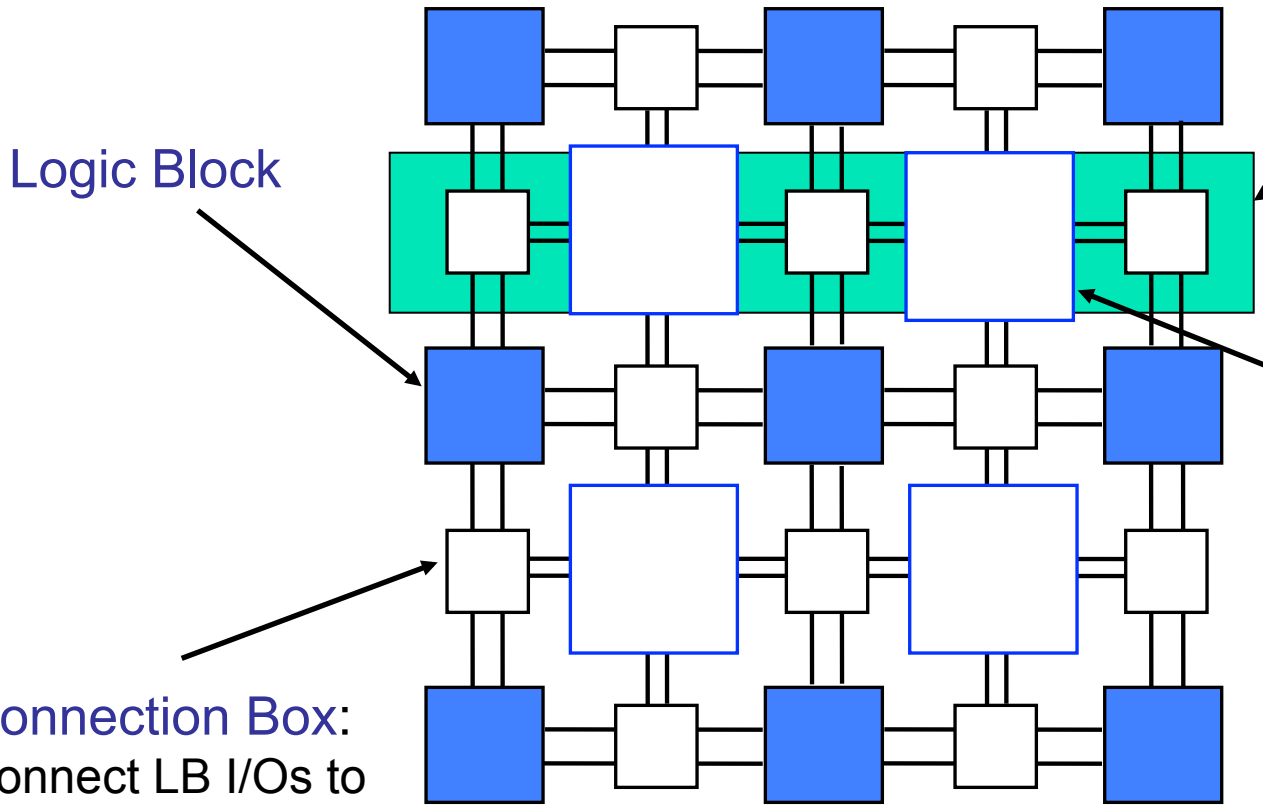
Wafer Stacking: Scenario 2

- Homogeneous stacking (true 3D):



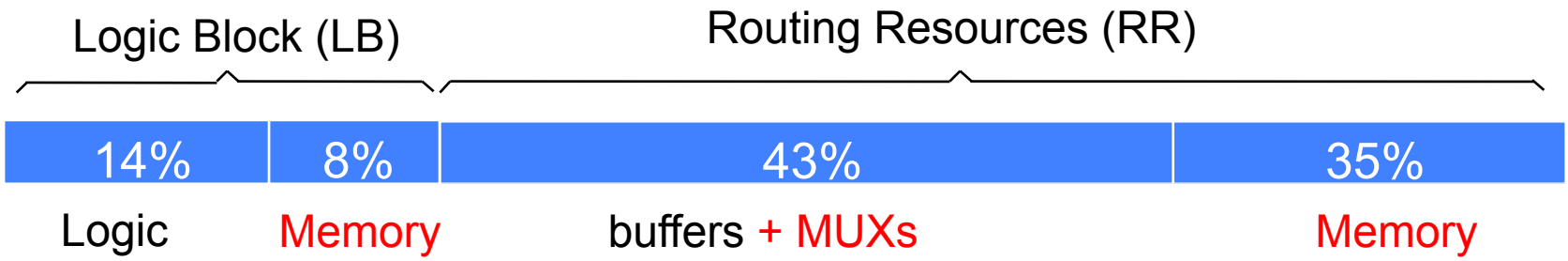
- Good news: TSV pitch of 3-5x Via 3 allows for stacking several layers with small footprint overhead
- Issues:
  - Performance benefits may diminish with increased number of layers
    - Programming overhead for vertical dimension
    - TSV parasitics
    - Heat dissipation in intermediate layers
  - Significant modifications to routing architecture and CAD tools needed

Island-Style FPGA Architecture



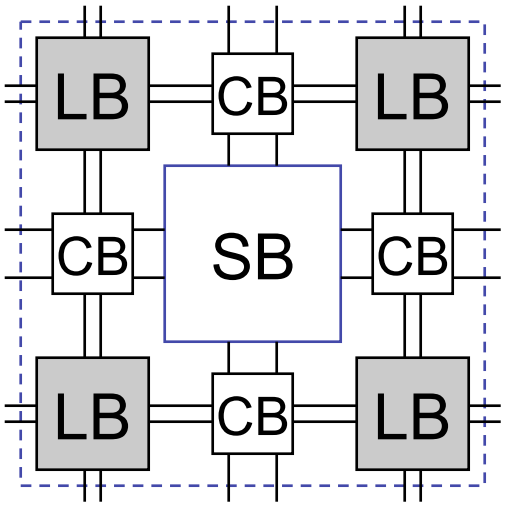
Logic block, connection box, switch box are SRAM programmable

Why is FPGA Inefficient?

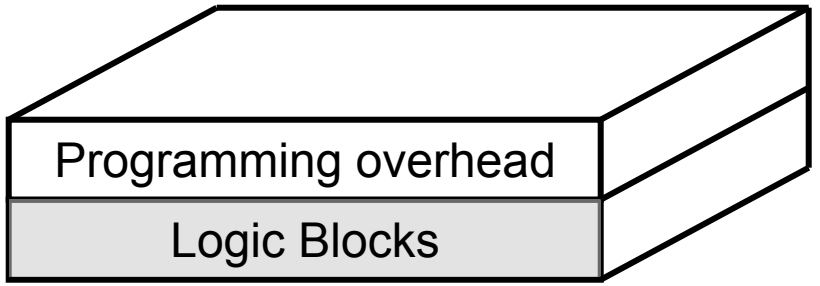


Around 80% of the area is overhead

Idea: Stack FPGA Overhead



Move programming overhead on top of logic

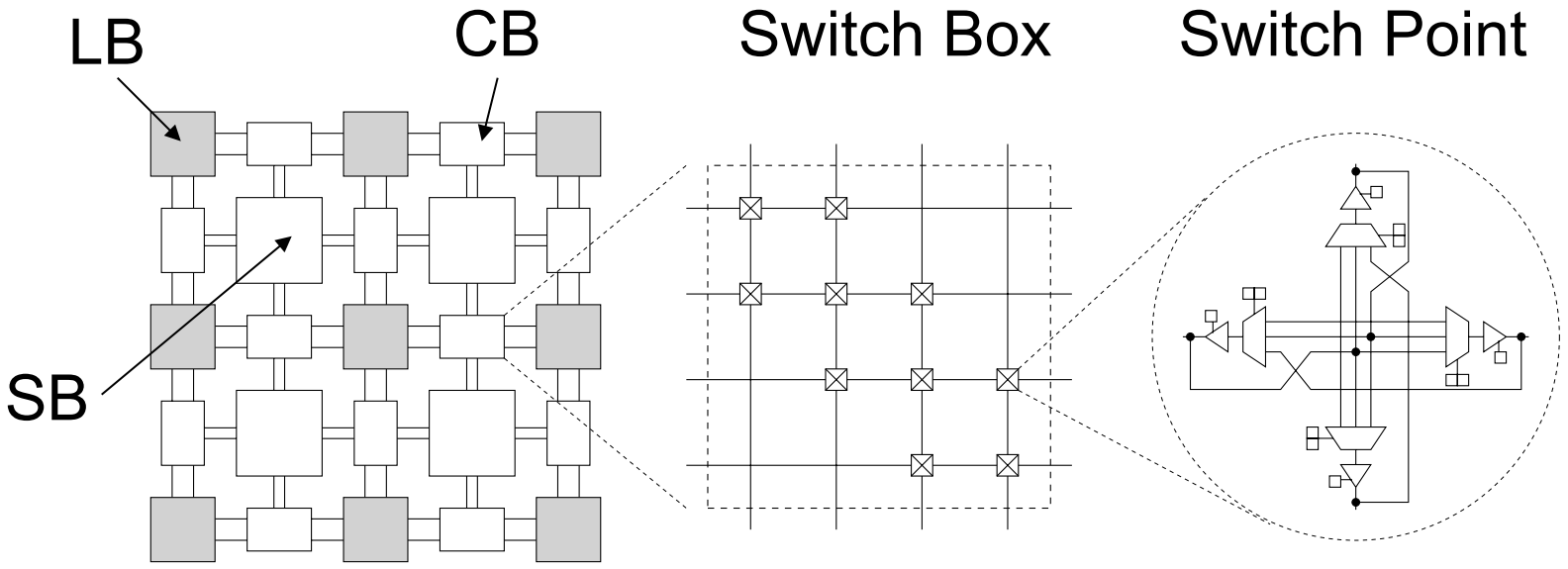


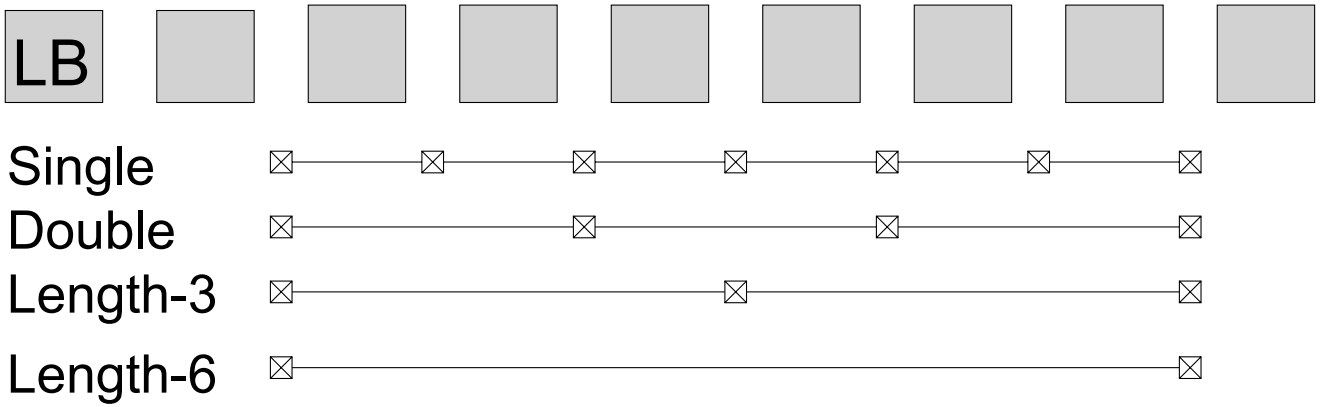
- Potential benefit: reduce footprint by up to 6 times

Digression: Delay and Power Estimation

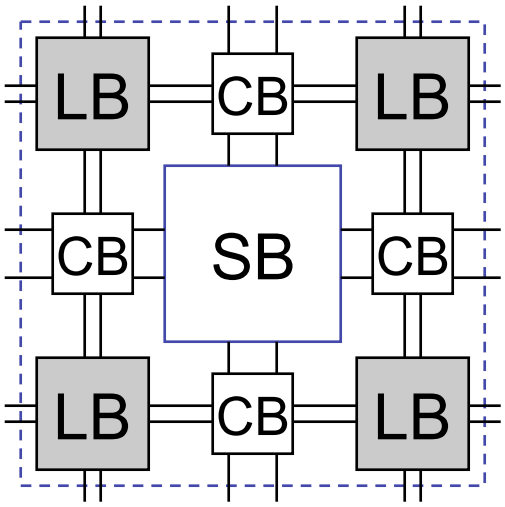
- Delay:
  - HSPICE simulations with BPT model used to extract device and metal wire parameters
  - Elmore delay used to determine buffer and switch point sizes
  - VPR from University of Toronto used to perform P&R on 20 largest MCNC benchmark designs, extract delay
  - Delay improvement: geometric average of all pin-to-pin net delays/ critical path delays relative to a baseline 2D-FPGA in 65nm CMOS
- Dynamic power:
  - Assume typical breakdown of FPGA power consumption between logic block, routing fabric, clock network
  - Logic block power doesn't change with 3D
  - Sum up capacitances of routed nets in benchmark designs
  - Power improvement: total capacitance relative to baseline 2D-FPGA, factor in fixed logic power consumption

Digression: Baseline 2D-FPGA

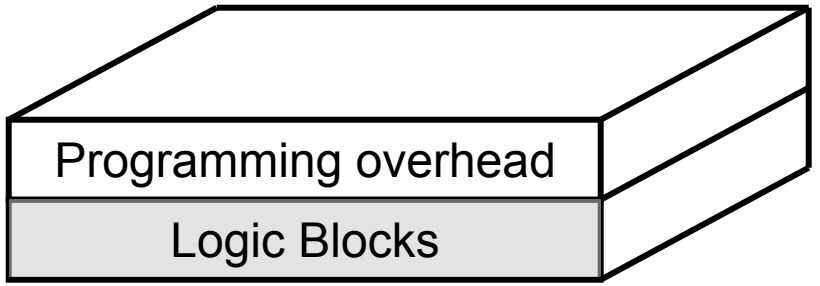




Idea: Stack FPGA Overhead



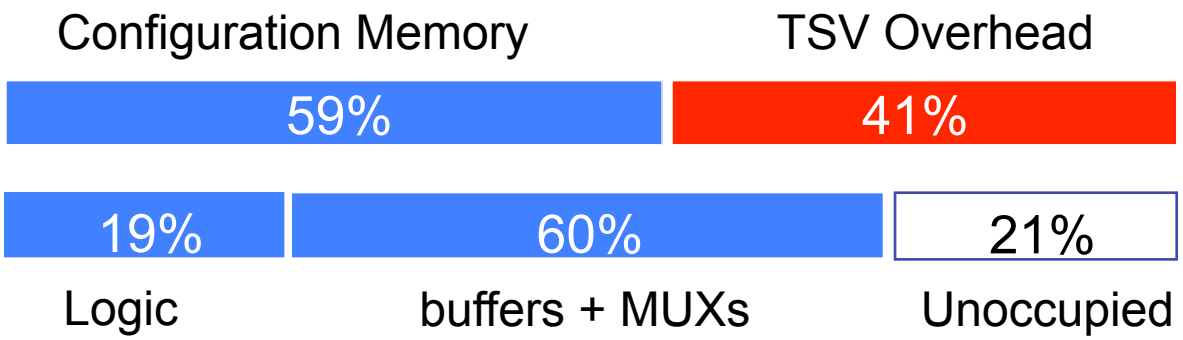
Move programming overhead on top of logic



- Potential benefit: reduce footprint by up to 6 times
- Can we do this using wafer stacking?

Wafer Stacking: Scenario 3

- Stack configuration memory



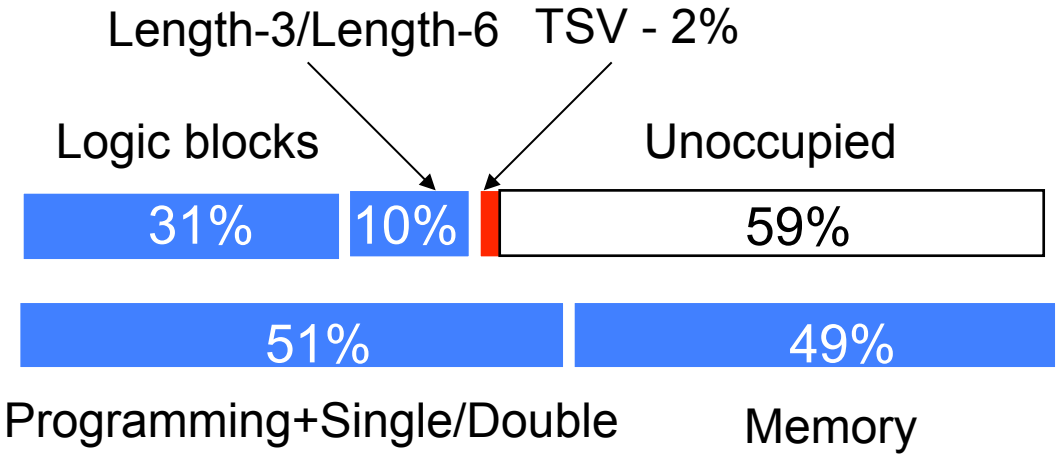
72% of baseline footprint

- Assume TSV pitch = 4x Via 3
  - TSV area ≈ 1k λ^2 vs 1.5k λ^2 for SRAM cell
- Delay improvement = 1.21X
- Dynamic power improvement = 1.14X

  Relative to baseline 2D-FPGA in 65nm CMOS

Wafer Stacking: Scenario 3

- Stack logic blocks and some interconnect



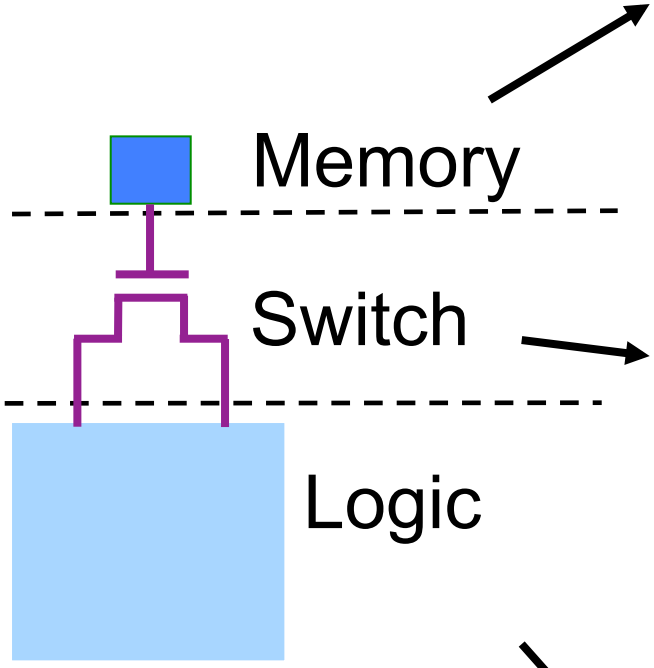
71% of baseline footprint

- Many fewer TSVs needed
- Delay improvement = 1.21X
- Dynamic power improvement = 1.14X
  Relative to baseline 2D-FPGA in 65nm CMOS
- Unoccupied area may be used for block memory, IPs

Wafer Stacking Conclusions

- Stacking IPs on top of FPGA fabric feasible but marginally beneficial to delay and power
- Homogeneous stacking promising---needs more investigation
- Stacking FPGA overhead on top of logic provides marginal improvements in delay and power
  - Need significantly finer vertical via pitch to realize potential --> Use monolithic stacking

Monolithically Stacked 3D-FPGA



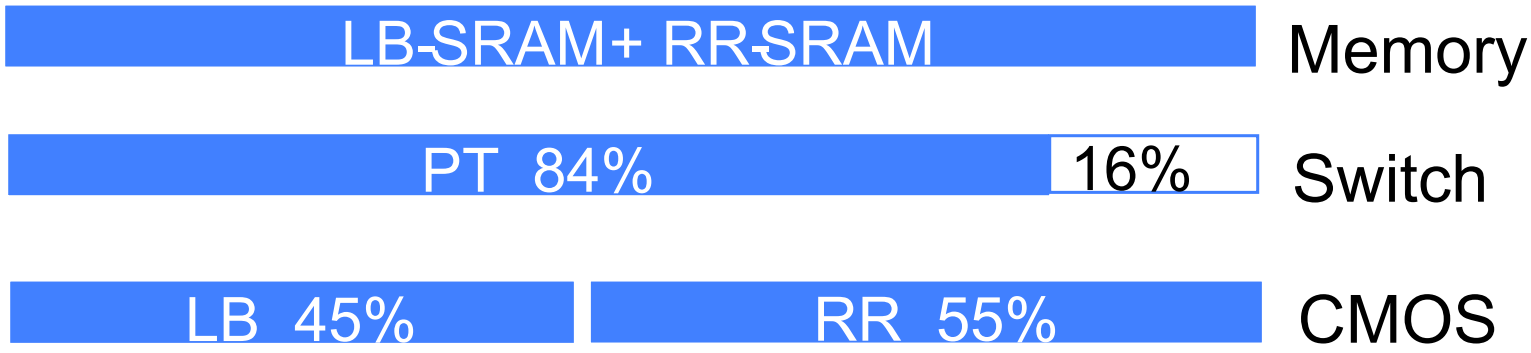
- 2-T flash [Cao'94]
- 3D SRAM cells [Hitachi'04]
- Reprogrammable via

Ge PMOS:

- Ge rapid growth
- Metal induced Si Epitaxy
- Template based Epitaxy
- Ge nanowire

- Standard CMOS

Monolithically Stacked 3D-FPGA [Lin et al 06]

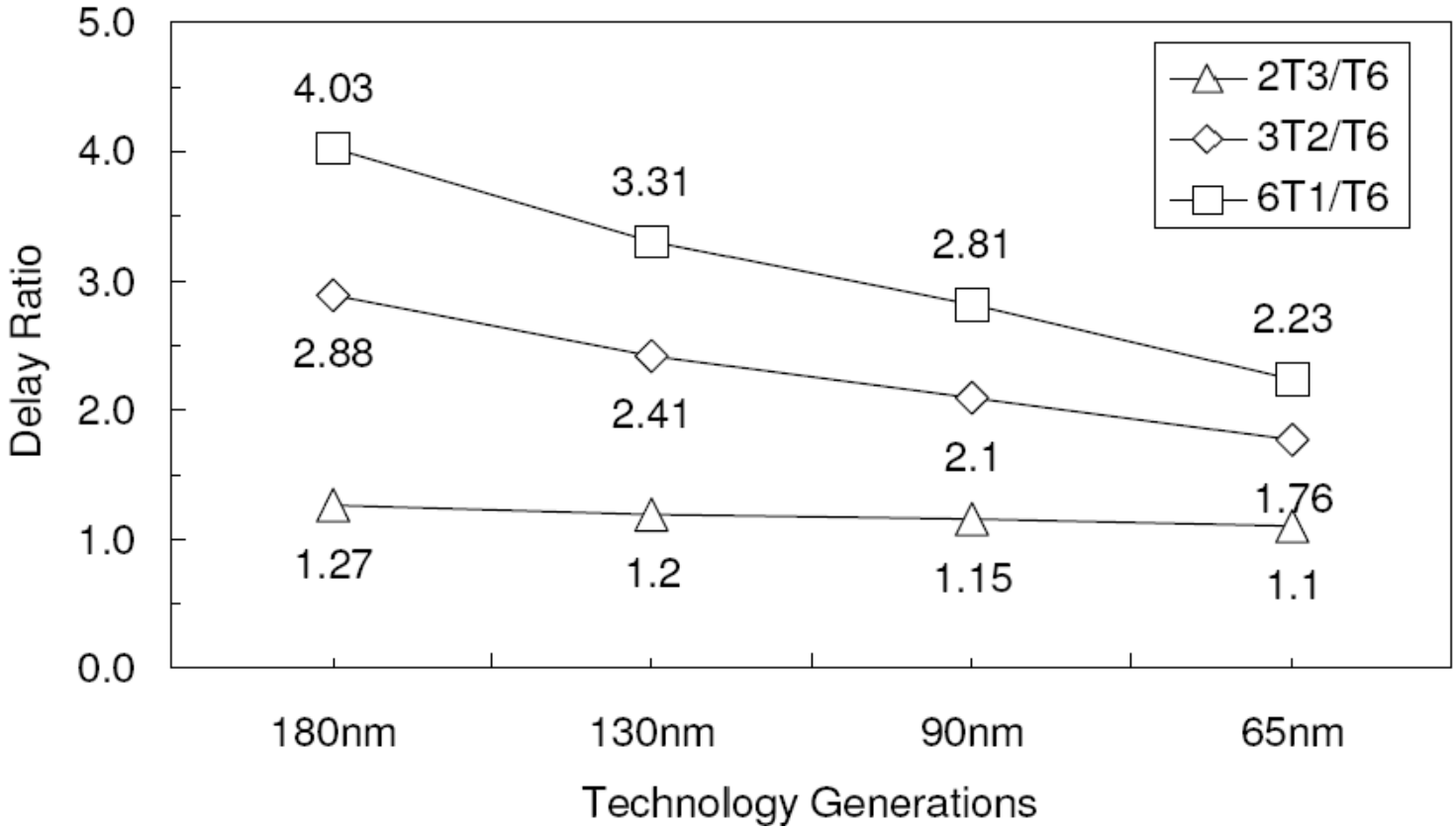


- Can achieve:
  - 3.2X higher logic density
  - 1.7X lower geometric mean delay
  - 1.7X less dynamic power consumption

  Relative to baseline 2D-FPGA in 65nm CMOS
- Improvements achieved using very few layers on top of CMOS
- How much better can we do by optimizing the FPGA architecture?

Observation 1

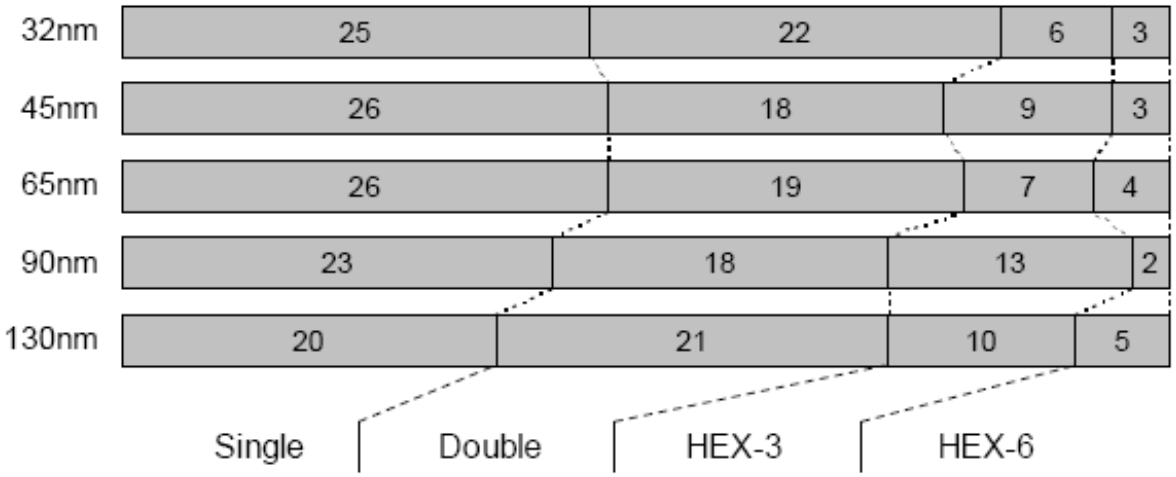
Utility of long segments (length 3 and 6) decreases with technology scaling and 3D



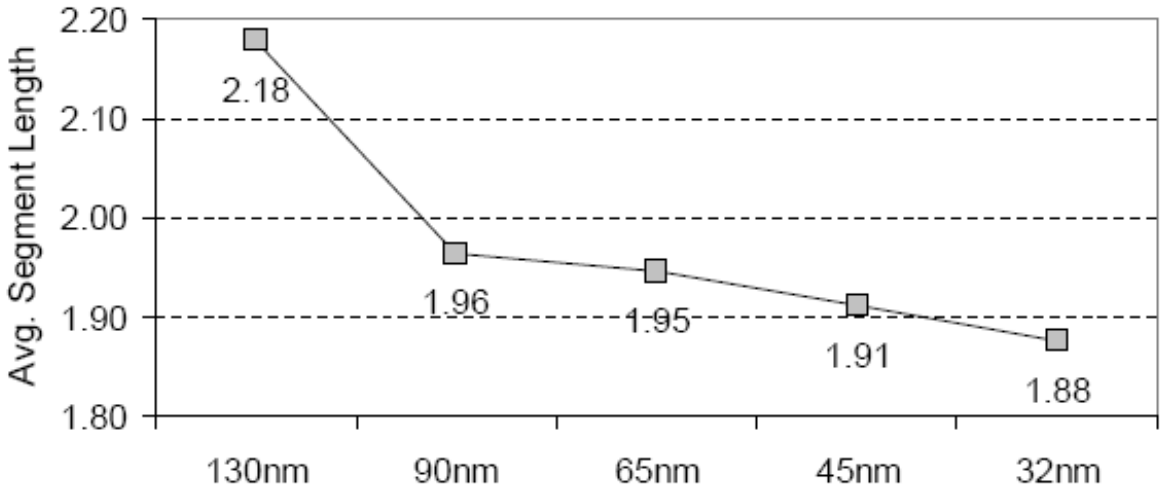
TORCH [Lin et al 08]

- Design tool for segmented routing channel
- Input:
  - FPGA architecture with initial (random) segmentation
  - CMOS technology parameters
  - A set of benchmark circuits
- Cost function:
  - Product of average net delay and interconnect power consumption averaged over benchmarks
- Use simulated annealing and incremental routing to incrementally optimize cost function--computation easily parallelized
- Output:
  - An optimized segmentation

Results: Technology Scaling



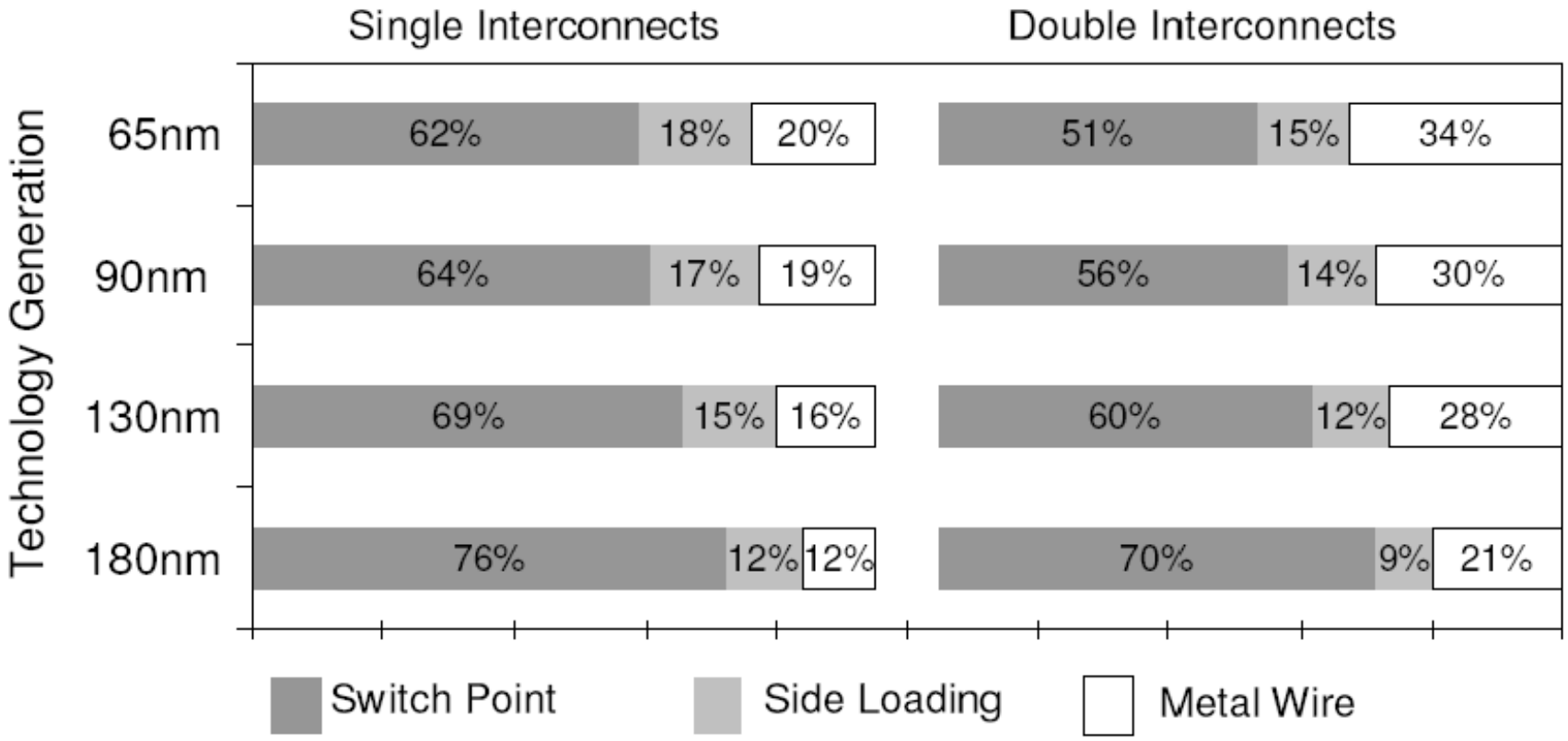
(a)



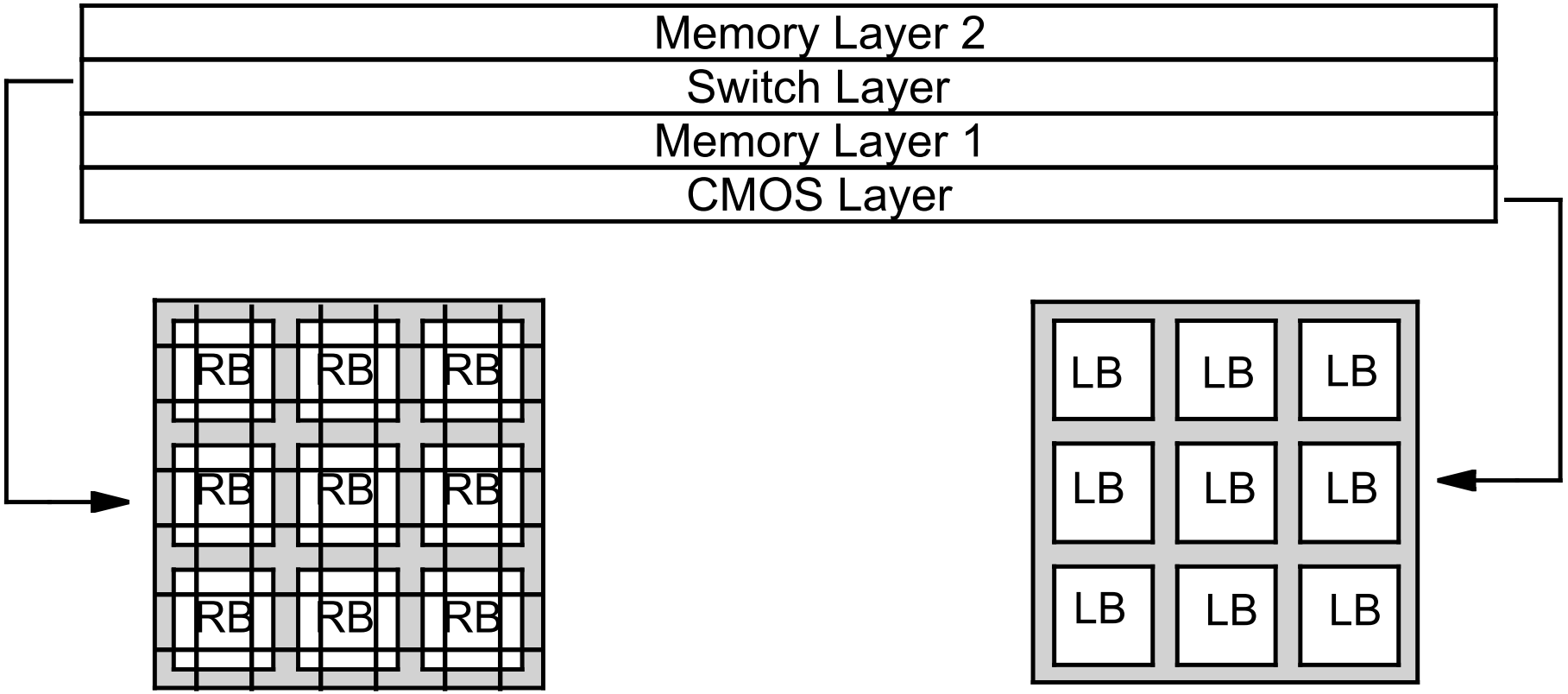
(b)

Observation 2

Switch point is the bottleneck of interconnect delay/power



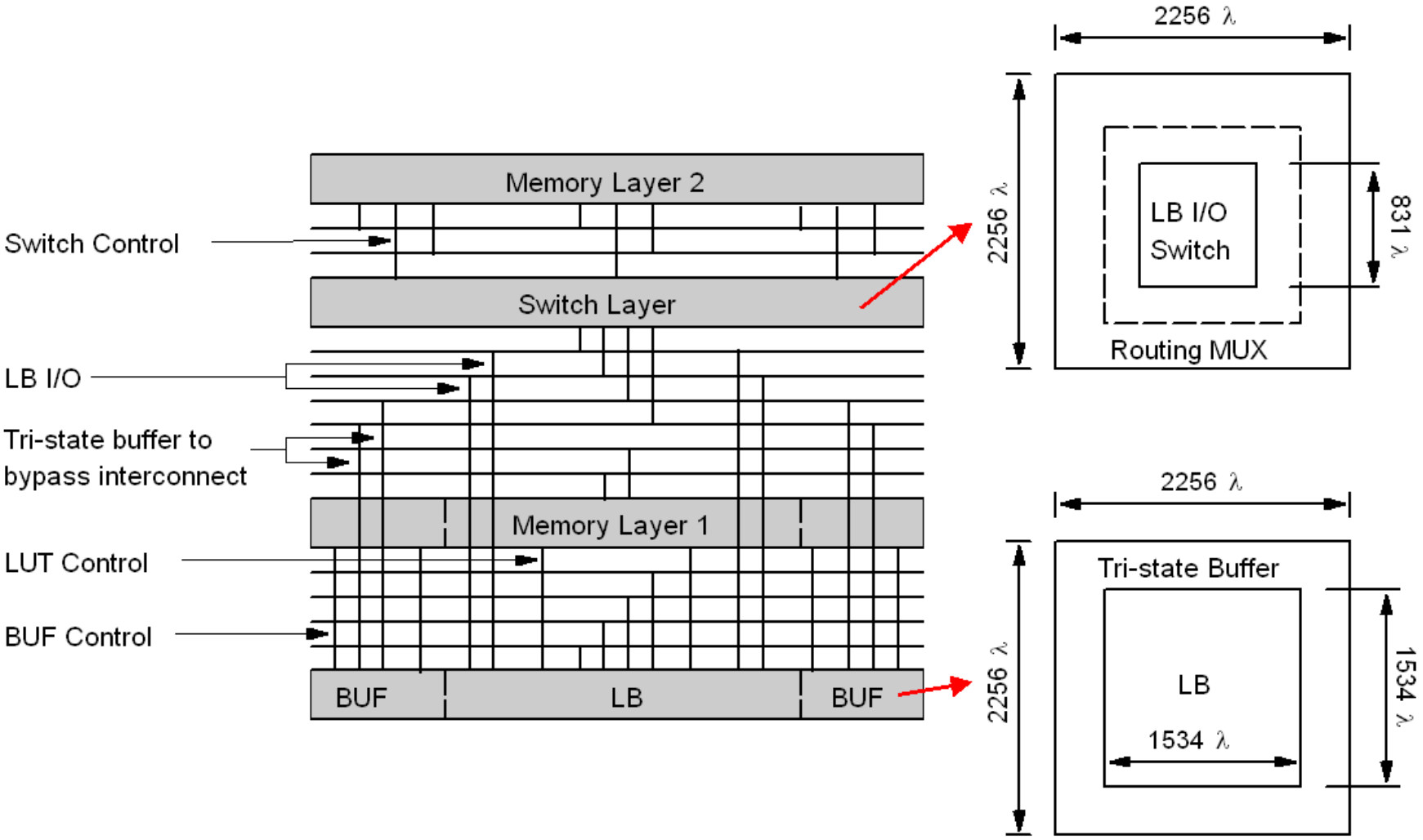
New Routing Fabric [Lin et al. 07]



Highlights

- Horizontal and vertical routing channels
  - Each comprised only of single, double interconnects
  - Longer segments formed by directly connecting several single and double segments without going through routing blocks
- Routing block (RB)
  - Integrates connection and switch box functionalities
  - Provides direct connects between neighboring LBs
  - Provides extended switching capability

3D Implementation Feasibility



Performance Benefits [Lin et al. 07]

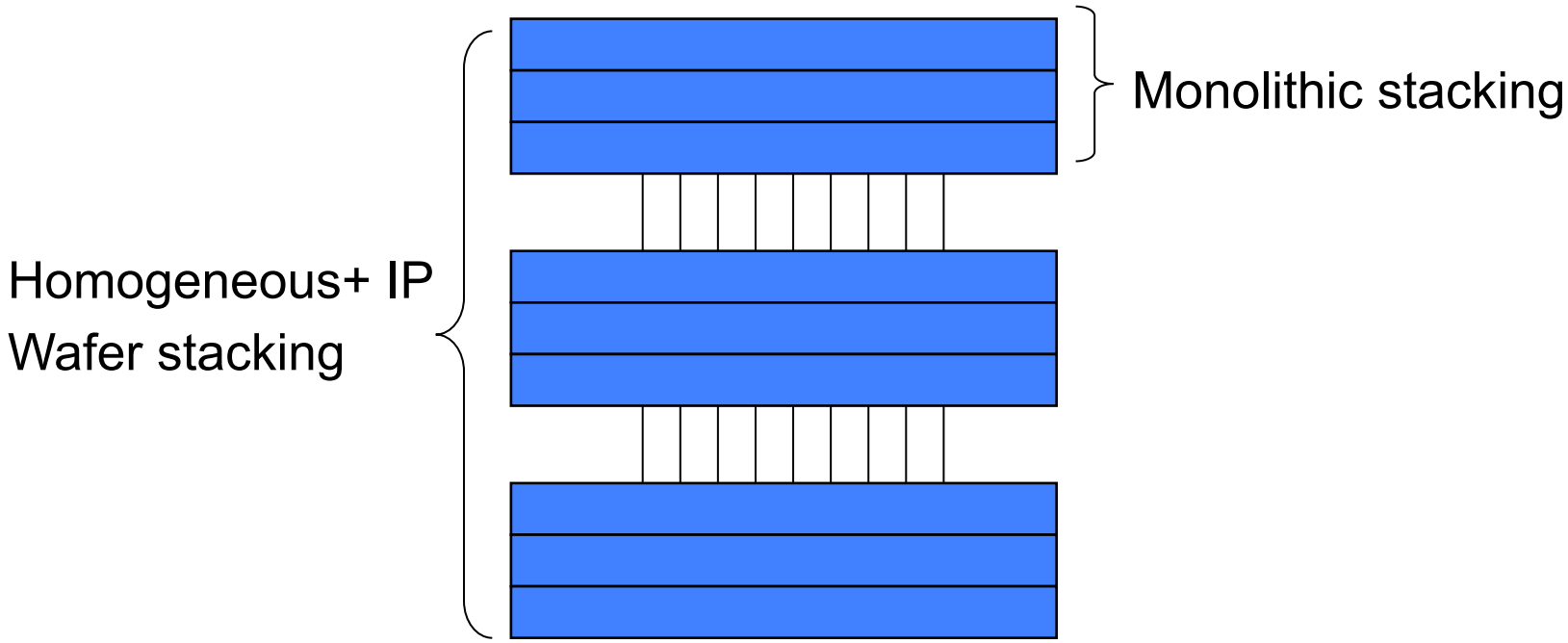
- 3.30X higher logic density
- 2.35X lower delay (vs. 1.7x)
- 2.82X less dynamic power (vs. 1.7x)

  Relative to baseline 2D-FPGA
- Further logic density improvement can be obtained by optimizing the logic block [Lin 08]

Conclusion

- 3D can help close the performance gap between FPGA and ASIC
- Wafer stacking:
  - Stack IPs on top of FPGA logic fabric
  - Homogeneous stacking--true 3D
- Monolithic stacking:
  - Stack programming overhead on top of logic
  - Significant performance advantages relative to baseline 2D
  - Need few monolithic layers on top of CMOS

Ultimate 3D-FPGA



Has potential to approach 2D-ASIC performance

Thank You!


LELAND STANFORD JUNIOR UNIVERSITY
DIE LUFT DER FREIHEIT WEHT
1891