3D VLSI: Next Generation 3D Integration Technology

Karim Arabi
Vice President, R&D

Continued smartphone momentum
World's largest technology platform

~8B

Cumulative smartphone unit shipments forecast between 2014-2018

Source: Gartner, Sept '13

Mobile scale

Cumulative global unit shipments, 2013–2017

Smartphones & Tablets

>

- PCs
- Vehicles
- Digital media adapters
- Digital video recorders
- Portable game consoles
- Portable media players
- Audio systems
- Blu-ray disc players
- Digital cameras
- Game consoles
- Flat-panel TVs
- Set-top-boxes

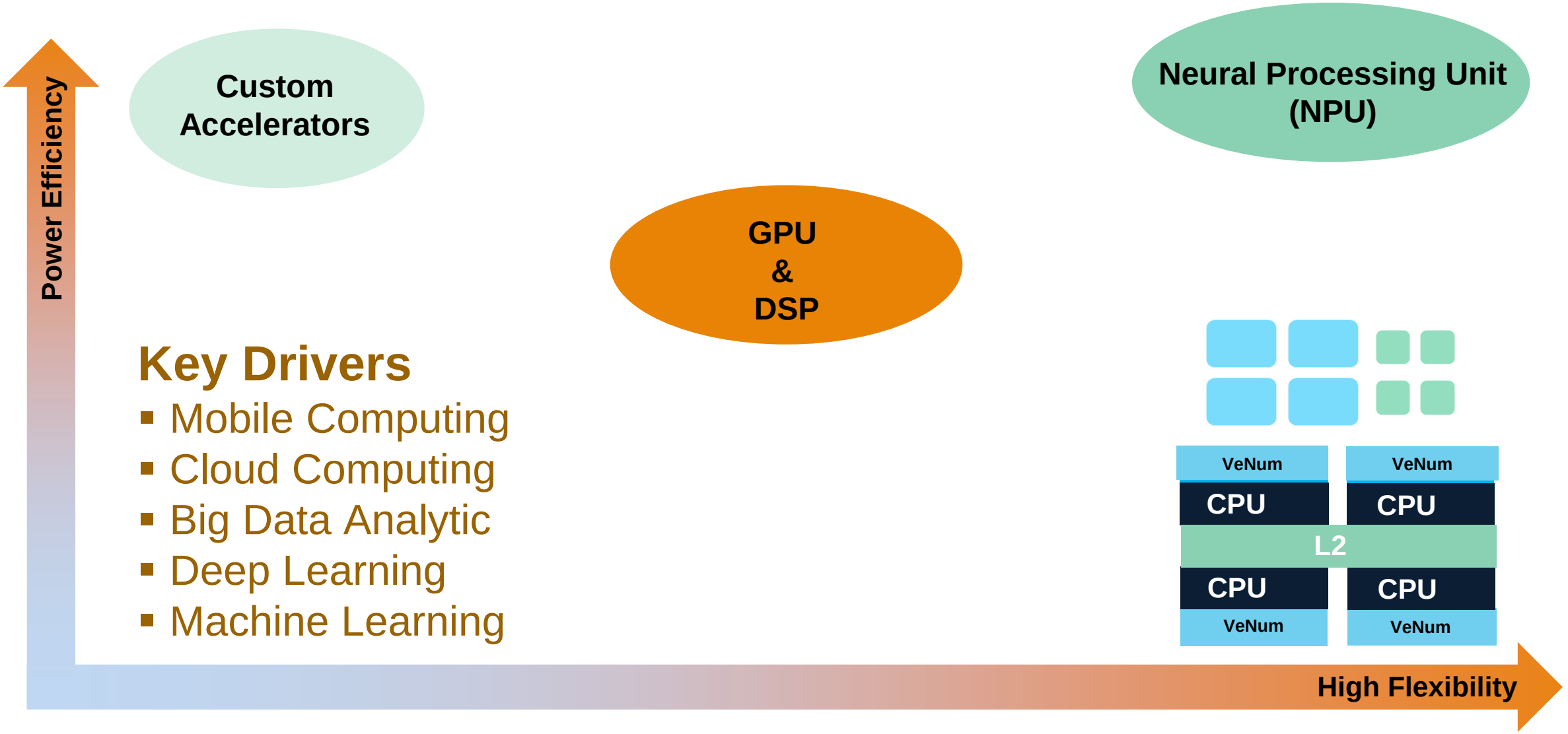
Sources: Smartphones, tablets and PCs: Gartner, Sep. '13; Vehicles: ABI, Apr. 2013; All others: Strategy Analytics, Mar. '13

Cloud and Mobile Computing

Big Data
and abundant computing power are pushing computing to the **Cloud**

Instant Data
generated by sensors and users are pushing computing to the **Edge**

Mobile Heterogeneous Compute Units to Lower Power

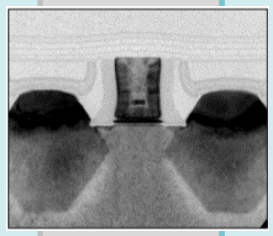


Key Drivers

- Mobile Computing
- Cloud Computing
- Big Data Analytic
- Deep Learning
- Machine Learning

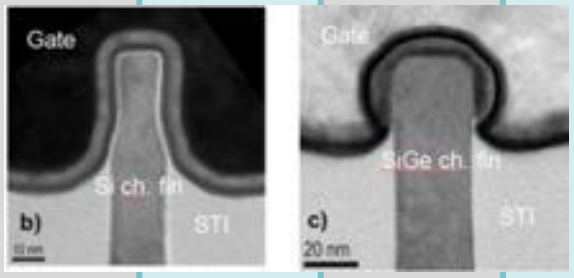
CMOS Scaling Outlook – The Roadmap Ahead

	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
Technology	20nm		14nm		10nm		7nm		5nm	
Vdd	1.0/0.9V 0.5V		0.9/0.8/0.7V			0.7/0.6V		0.6V		
Transistor	Planar/Tri-gate, Si channel		FinFET, Si/SiGe/Ge		FinFET, Si/SiGe/Ge		FinFET/Nanowire SiGe/Ge/ III-V?		Tunnel FET SiGe/Ge/III-V?	

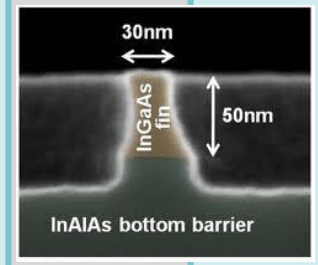


Planar bulk RMG

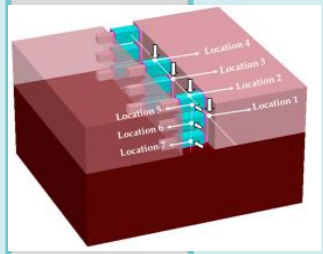
Tri-gate, FinFET Si



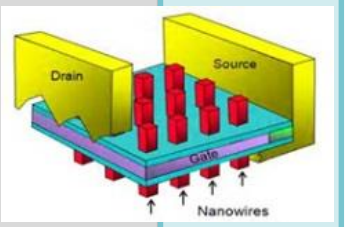
FinFET Si/SiGe/Ge channel



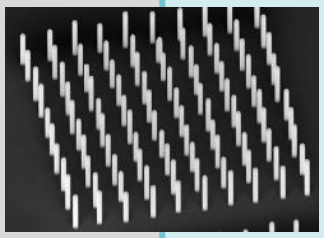
FinFET III-V channel



Nanowire, horizontal



Nanowire, vertical?



Tunnel FET, vertical?

CMOS Scaling Outlook – The Roadmap Ahead

Potential Issues to Address for 7nm and Beyond:

- FinFET/NWFET self-heating (phonon confinement)
- Feasibility of TFET and vertical channel devices
- System performance degradation (contact/BEOL bottleneck)

More Moore

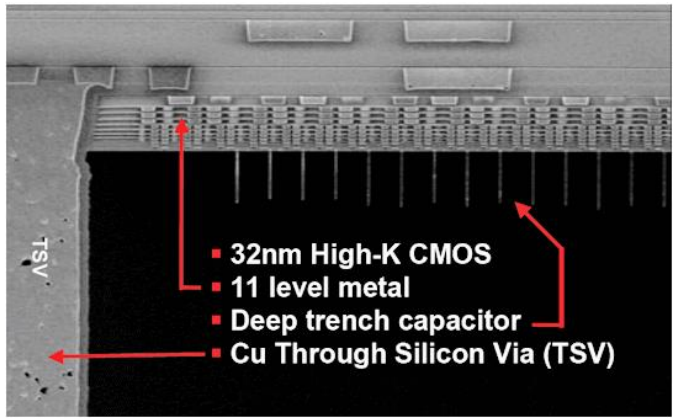
- Nano-wire
- TFET
- Ge/III-V

The 3rd Dimension

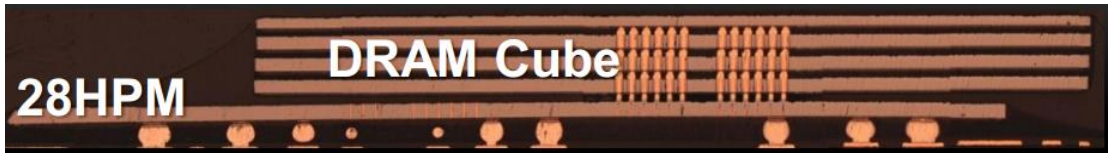
- 3DVLSI (3DV)
- MRAM &RRAM

Introduction: TSV-based 3D ICs

- **TSV-based 3D ICs are close to market**



IBM (VLSI 2011)

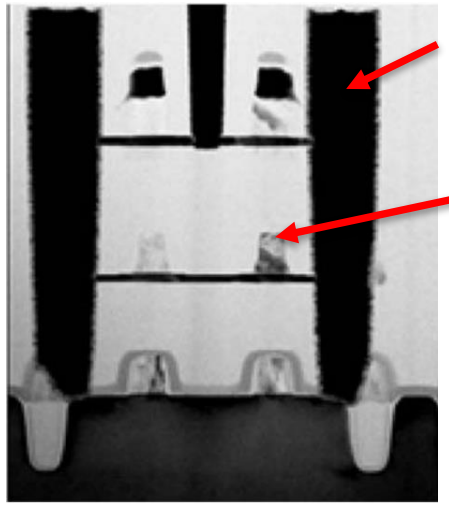


TSMC (ISPD 2014)

- **TSV-based 3D ICs shortens the interconnects, still quite large (5-10um, C=10-30fF)**

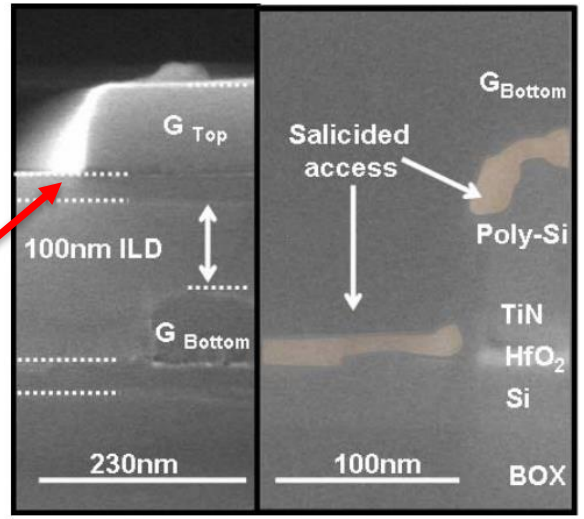
Courtesy Panth et al., ISLPED'14

3D VLSI - An Emerging 3D Technology

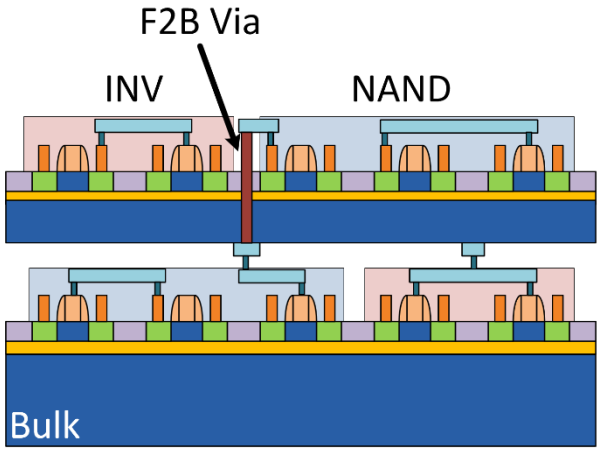


3D VLSI SRAM
Samsung (2010)

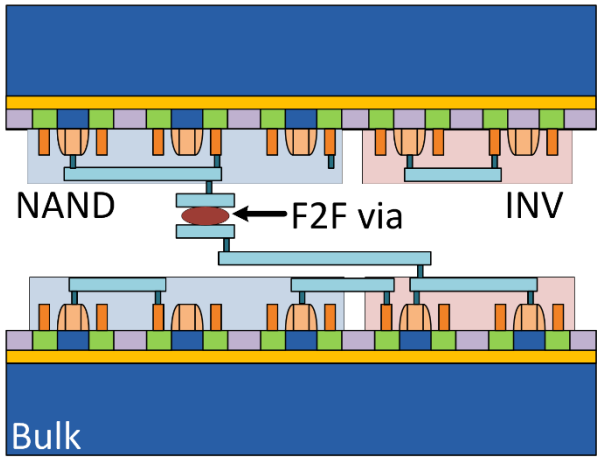
High quality thin silicon
(single crystal)



3D VLSI for general
logic LETI (2011)



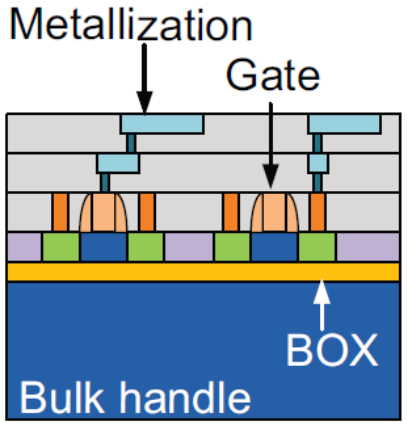
(a) Gate-Level F2B



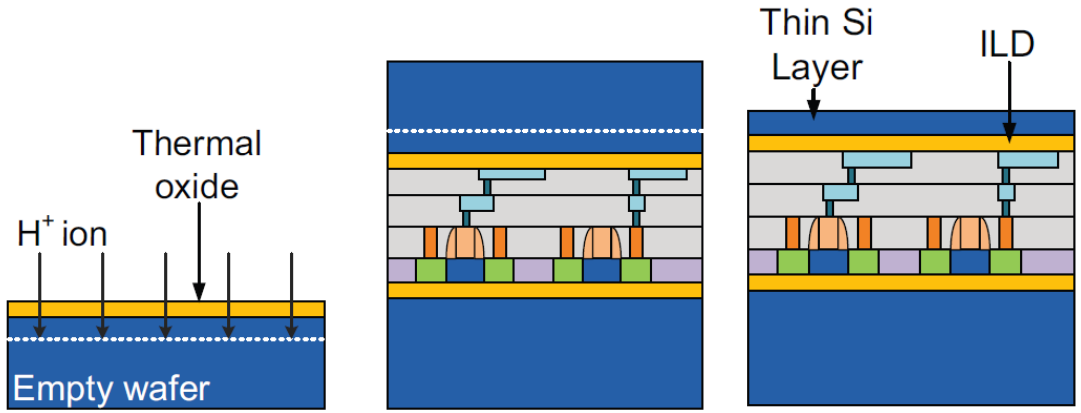
(b) Gate-Level F2F

Courtesy Panth et al., ISLPED'14

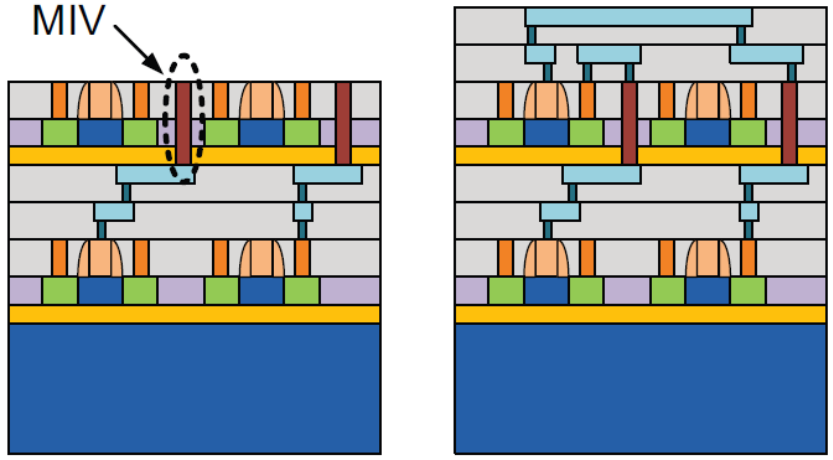
3D VLSI: Face-to-Back Fabrication Process



Bottom tier is created as usual



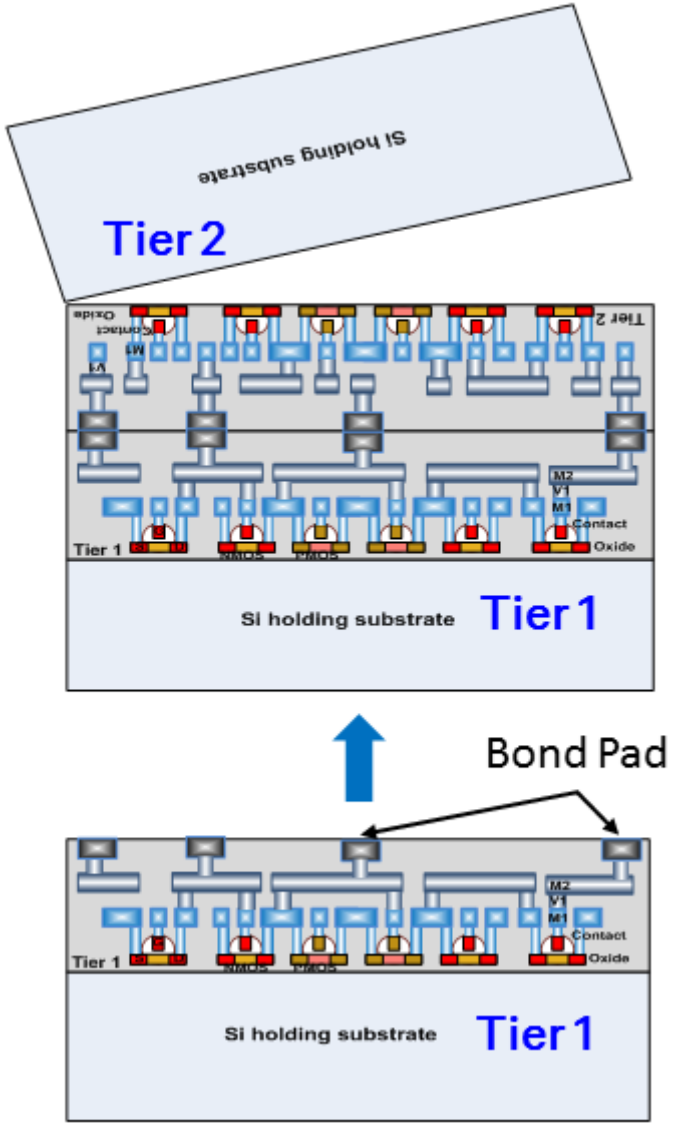
Thin Si layer is attached



Fabricate top-tier devices + interconnects

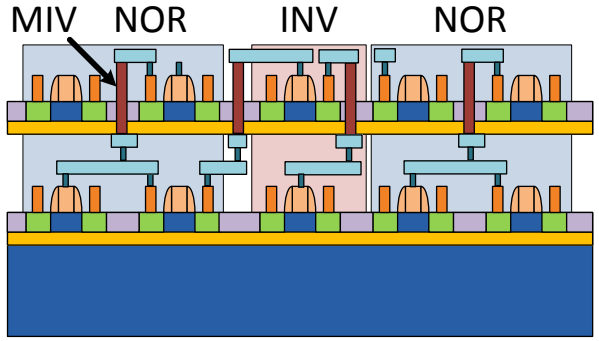
Courtesy Panth et al., ISLPED'14

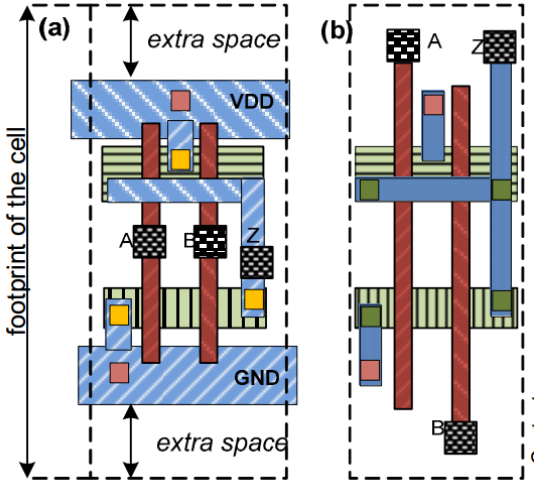
3D VLSI: Face-to-Face Fabrication Process



Design Styles Available in 3D VLSI (1/2)

- **Transistor-level [1]**
  - Each standard cell is folded
  - Pin density increases significantly
  - Footprint reduction is ~40%, not 50%
  - Standard cell re-design required
- **CELONCEL [2]**
  - Gate-level, but cell redesign required
  - Simplified design flow
  - Same disadvantages as transistor-level

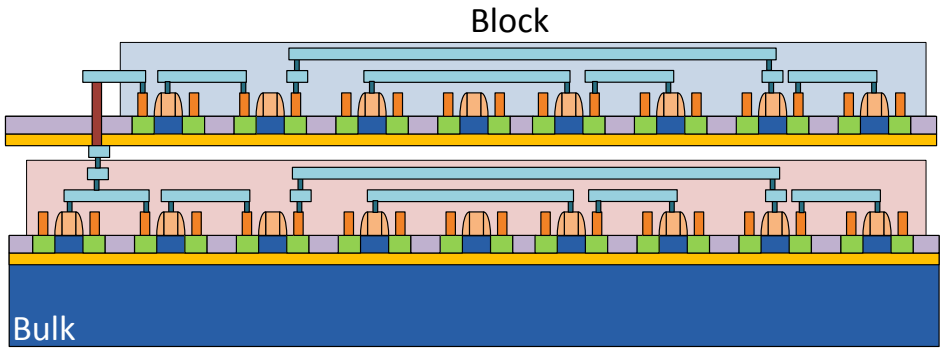




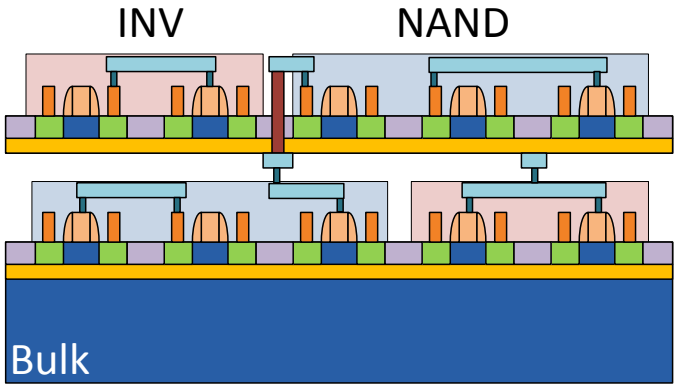
[1] Y.-J. Lee, D. Limbrick and S. K. Lim, “Power Benefit Study for Ultra-High Density Transistor-Level Monolithic 3D ICs,” DAC, 2013.
[2] S Bobba et al., “CELONCEL: Effective Design Technique for 3-D Monolithic Integration targeting High Performance Integrated Circuits,” ASPDAC, 2011.
Courtesy Panth et al., ISLPED’14

Design Styles Available in 3D VLSI (2/2)

- **Block-level** [1]
  - Functional blocks are 2D & they are floorplanned on to a 3D space
  - Reuse of IP
  - Does not fully take advantage of the high density offered by M3D



- **Gate-level**
  - Use existing standard cells & place them in 3D
  - Reuse of cells

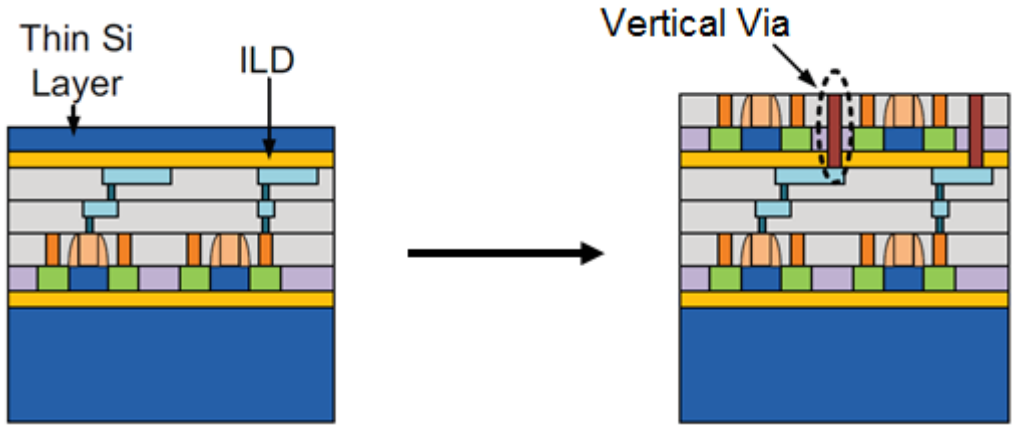


[1] S. Panth et al. "Power-Performance Study of Block-Level Monolithic 3D-ICs Considering Inter-Tier Performance Variations", DAC, 2014.

Courtesy Panth et al., ISLPED'14

Block-Level 3D VLSI Physical Implementation

Sequential 3D: Source of Inter-Tier Performance Variation



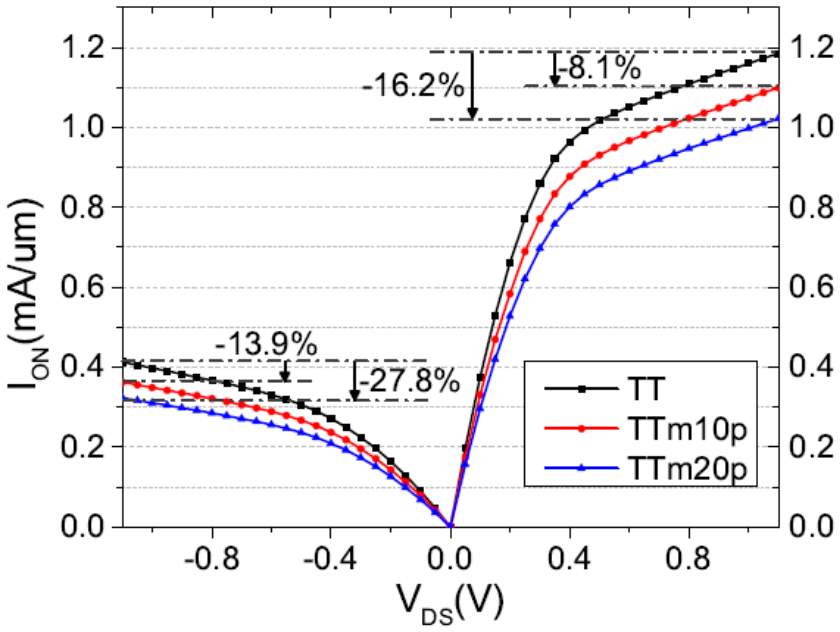
- **FEOL processing of top tier**
  - **RTA at 1200C will damage both devices and interconnects**

- **Process improvement: < 625C without performance loss → still too high for Cu interconnect**
- **Preventing damage to interconnects – Two options:**
  - **Use Tungsten (W) on the bottom tier → Worse interconnects on bottom tier**
    - **Identical devices on both tiers**
  - **~ 450C processing on the top tier → Worse transistors on the top tier**
    - **Identical interconnects on both tiers**

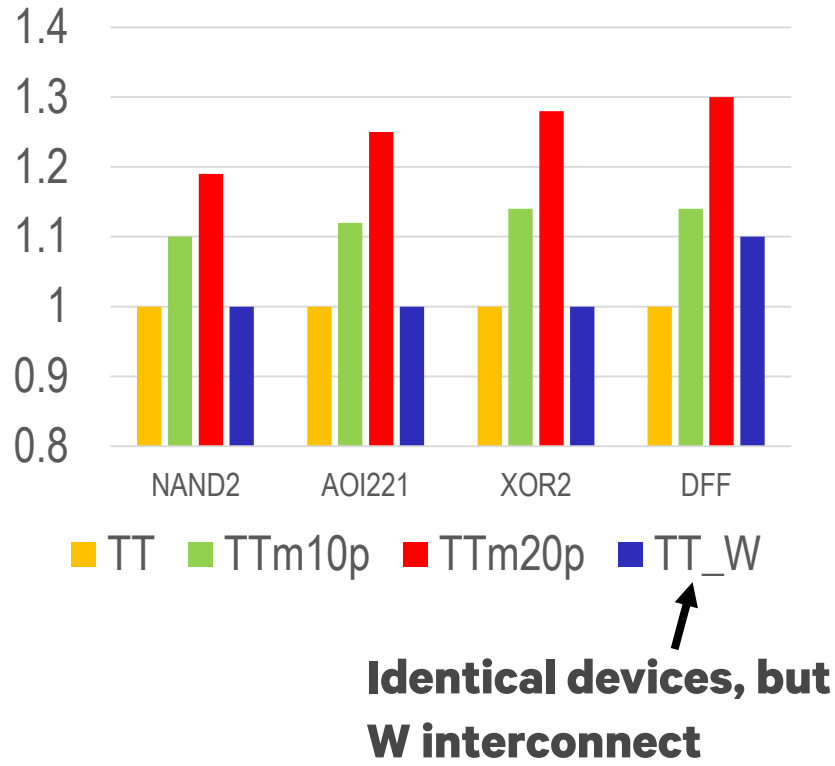
Courtesy Panth et al., DAC'14

Degraded Transistors

- **PMOS worsens by 27.8% and NMOS worsen by 16.2% (TTm20p corner)**



Change in delays of select standard cells

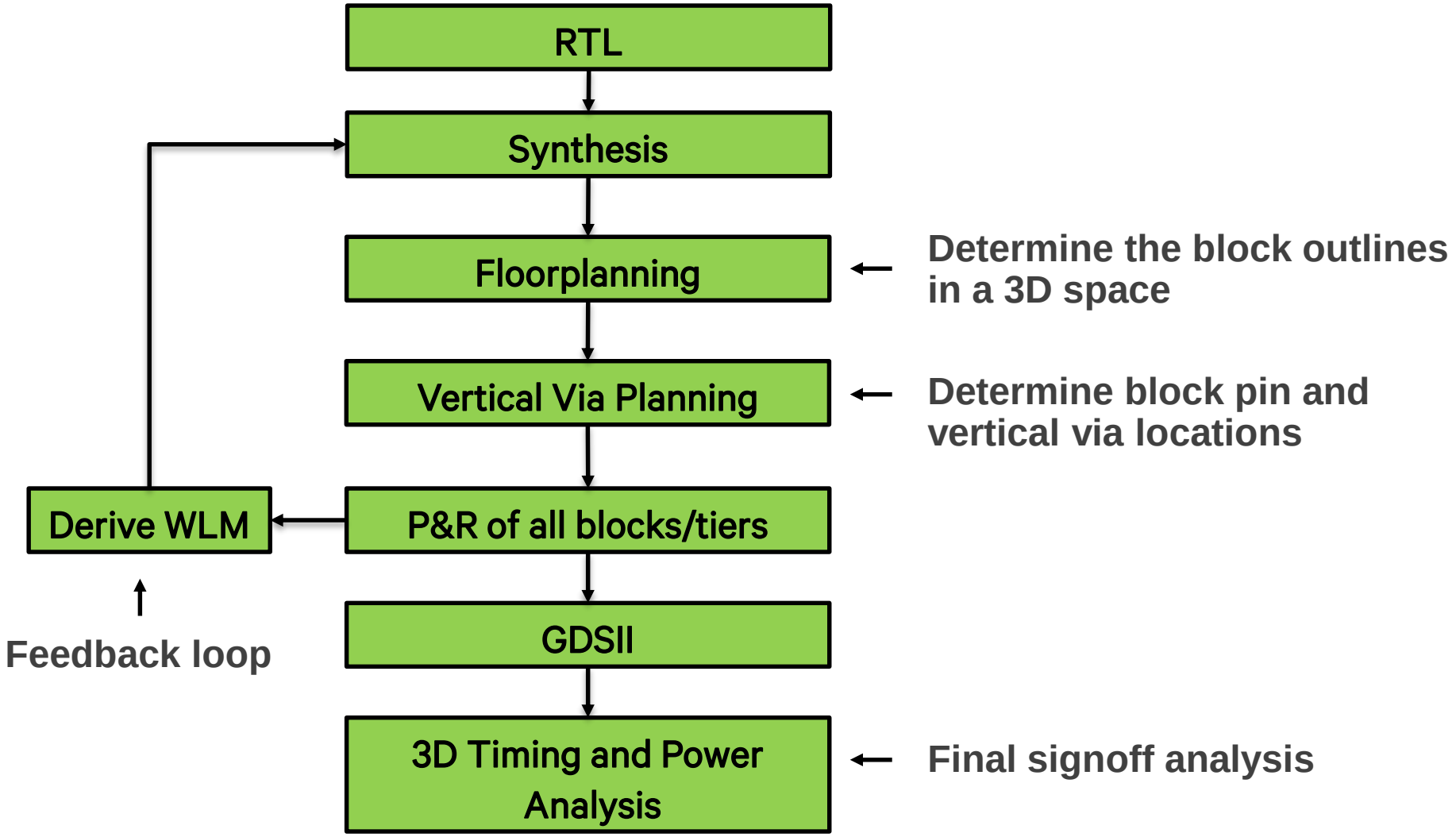


Identical devices, but W interconnect

[1] Low Thermal Budget Processing for Sequential 3D IC Fabrication, Rajendran et al., TED, 2007

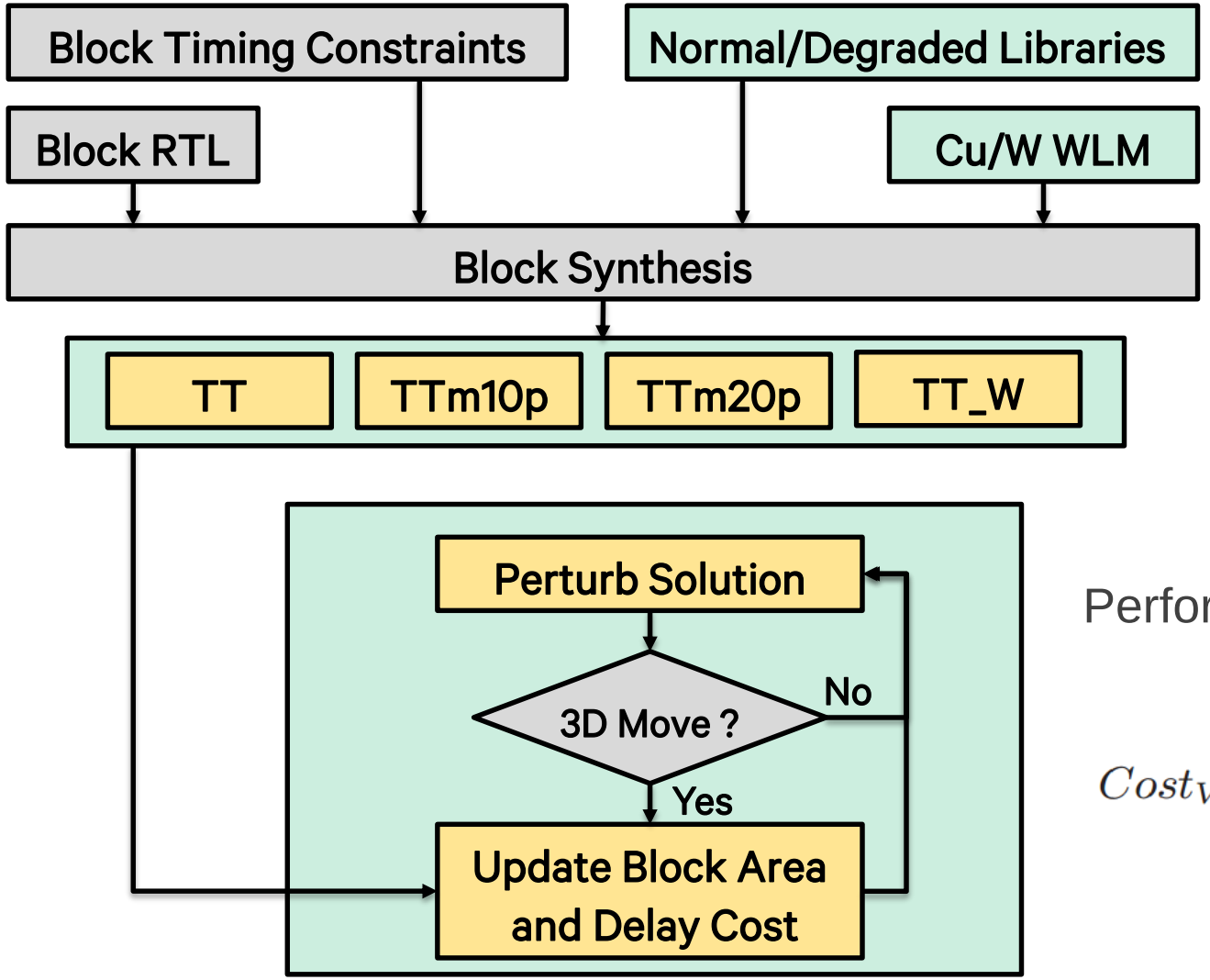
Courtesy Panth et al., DAC'14

Design Flow



Courtesy Panth et al., DAC'14

Variation-Aware 3D Block-Level Floorplanning



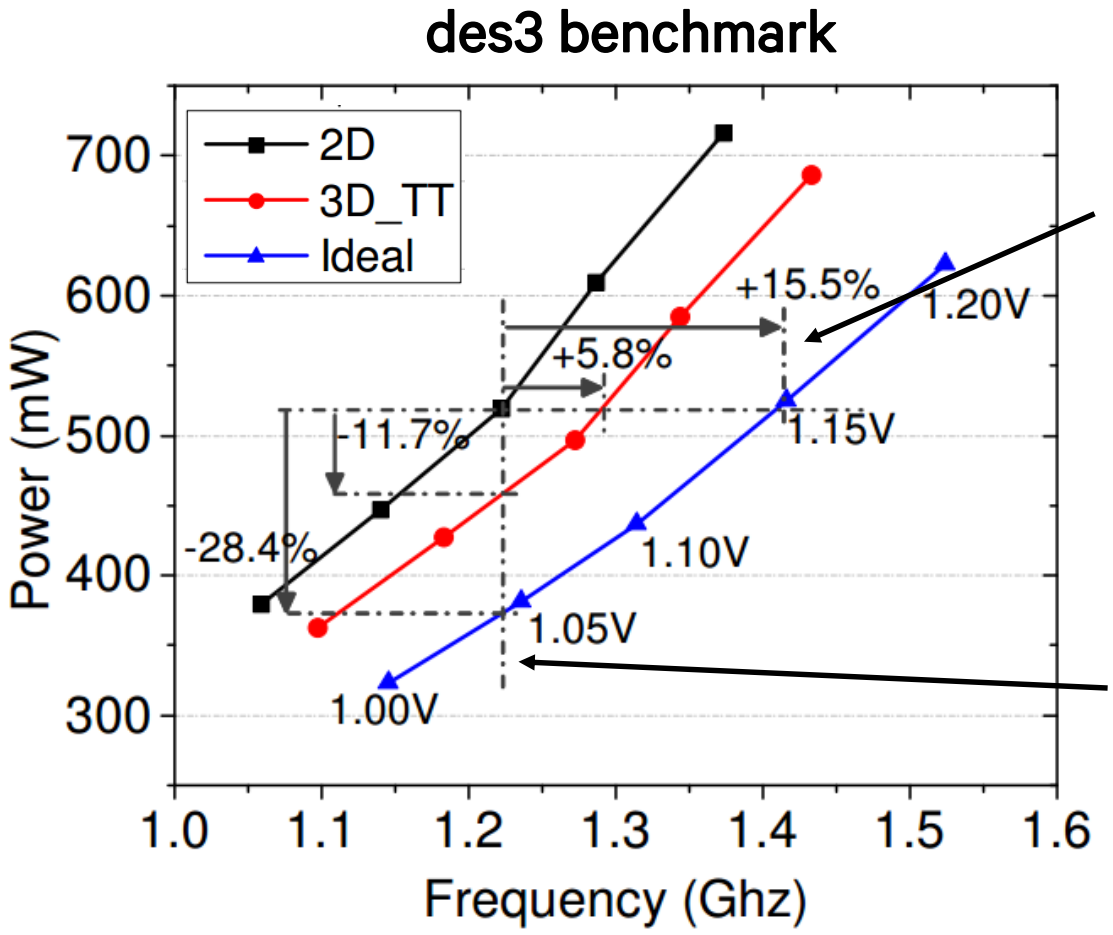
Block flavors

Performance-aware floorplanner

$$Cost_{VA} = \alpha.WL + \beta.Area + \gamma \sum_{i=1}^{N_{Block}} LPD(b_i)$$

Courtesy Panth et al., DAC'14

Power-Performance Study: Identical Tier Performance



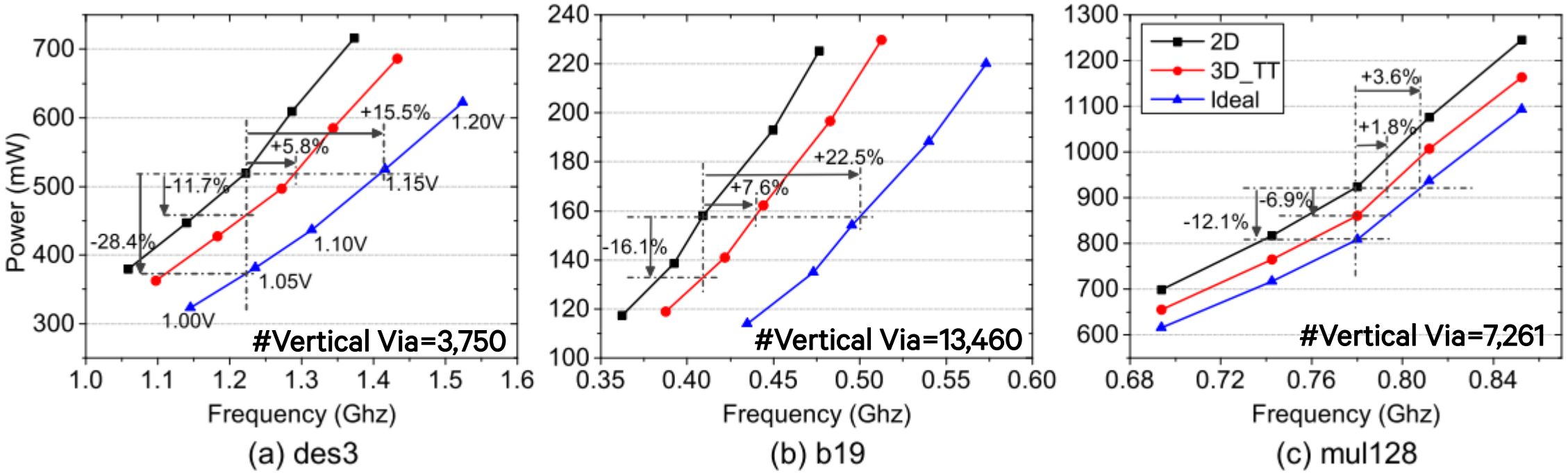
M3D closes 37% of the performance gap to ideal

M3D closes 41% of the power gap to ideal

- **Ideal: Zero RC for inter-block nets: Best possible block-level implementation**

Courtesy Panth et al., DAC'14

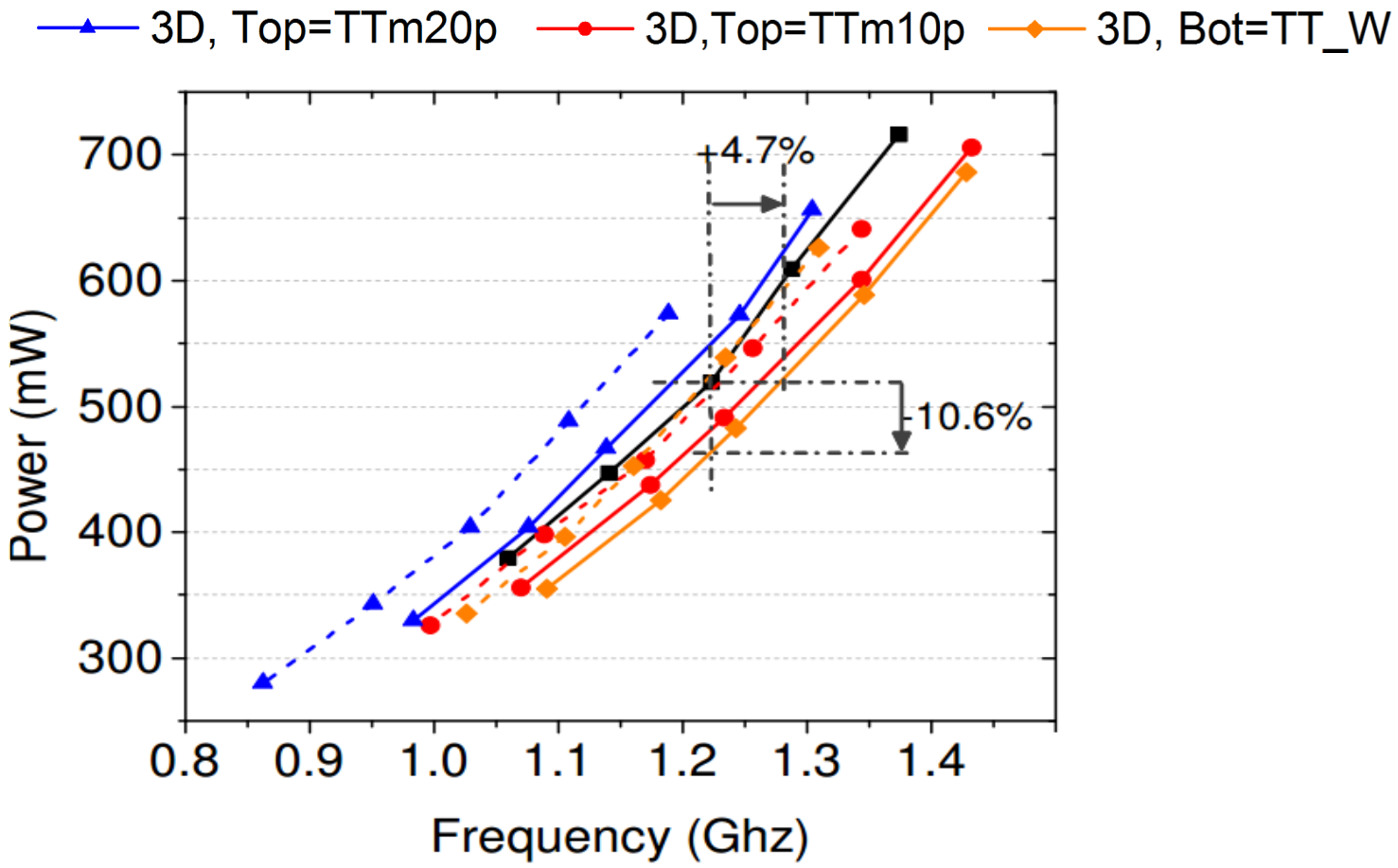
Power-Performance Study: Identical Tier Performance



- **Similar results for all benchmarks**
- **3D closes the power gap to ideal by at least 40%**
- **3D closes the performance gap to ideal by up to 50% and 40% on average**

Courtesy Panth et al., DAC'14

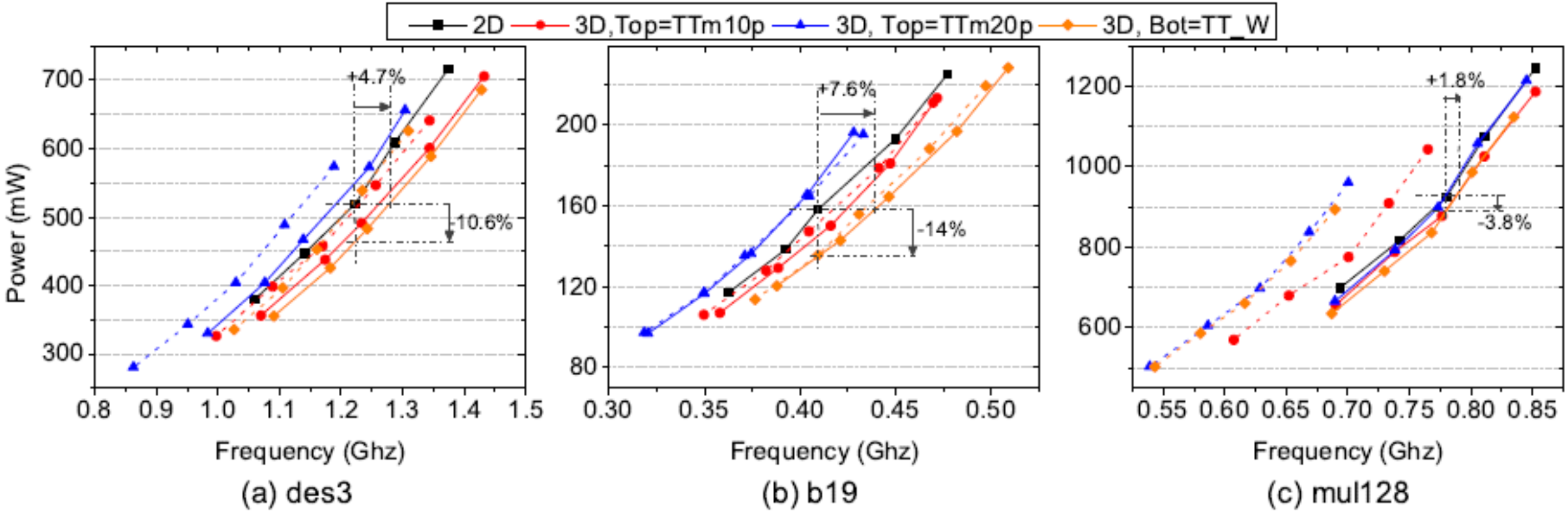
Variation-Aware Power-Performance Results



- **Dashed lines = no variation-aware floorplanning**
- **Solid lines = variation-aware floorplanning**

Courtesy Panth et al., DAC'14

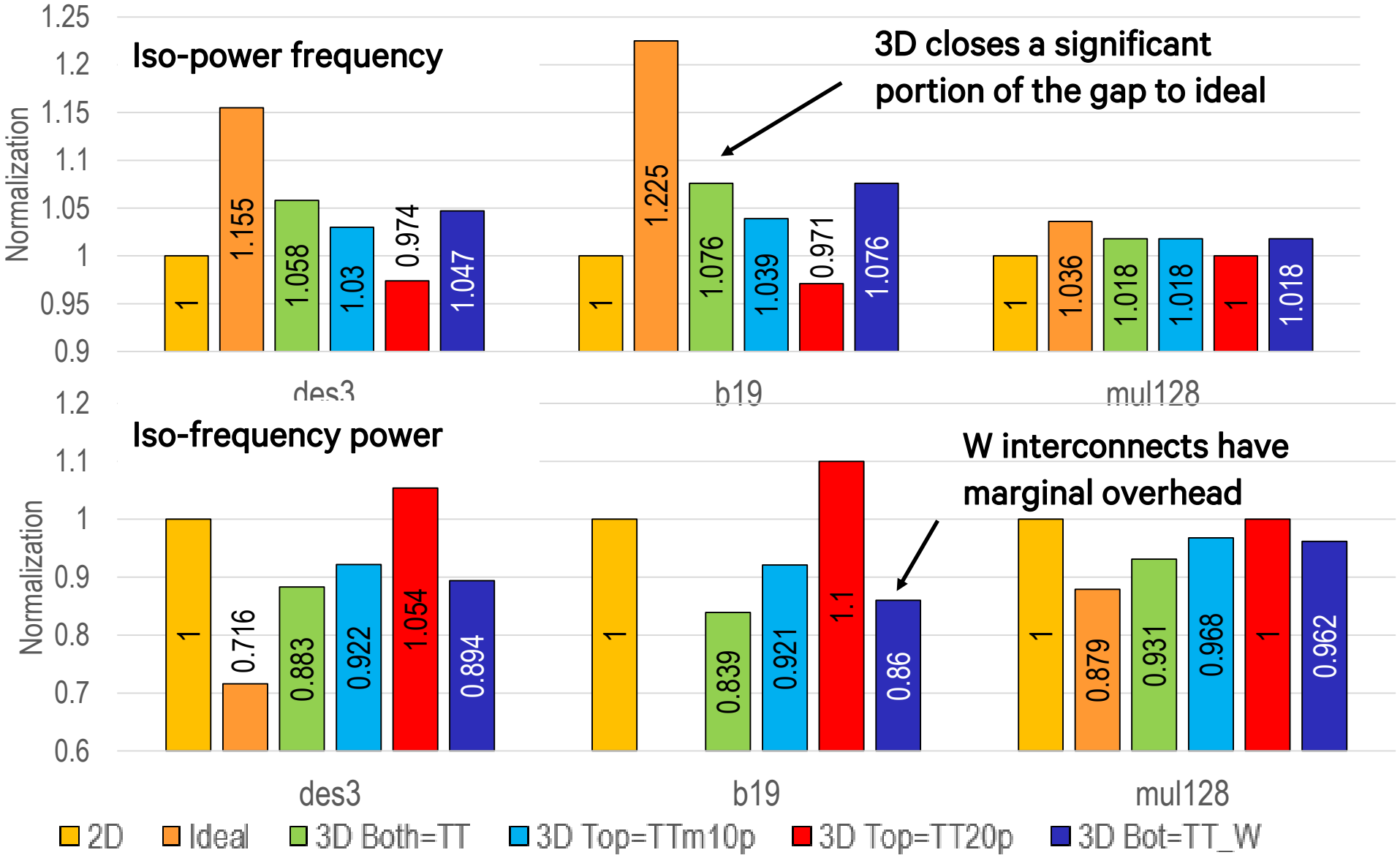
Variation-Aware Power-Performance Results



- **Variation-aware floorplanning always gives better results**
- **W on the bottom tier seems to be the best option**

Courtesy Panth et al., DAC'14

Summary of Results



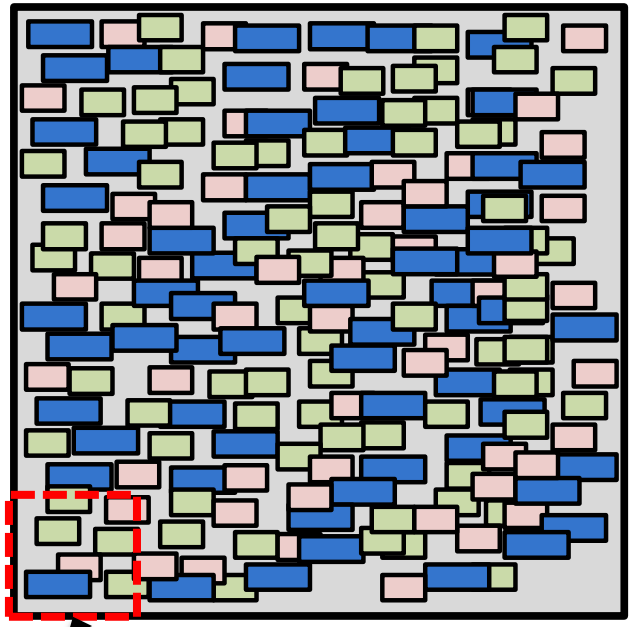
Courtesy Panth et al., DAC'14

Gate-Level 3D VLSI Physical Implementation

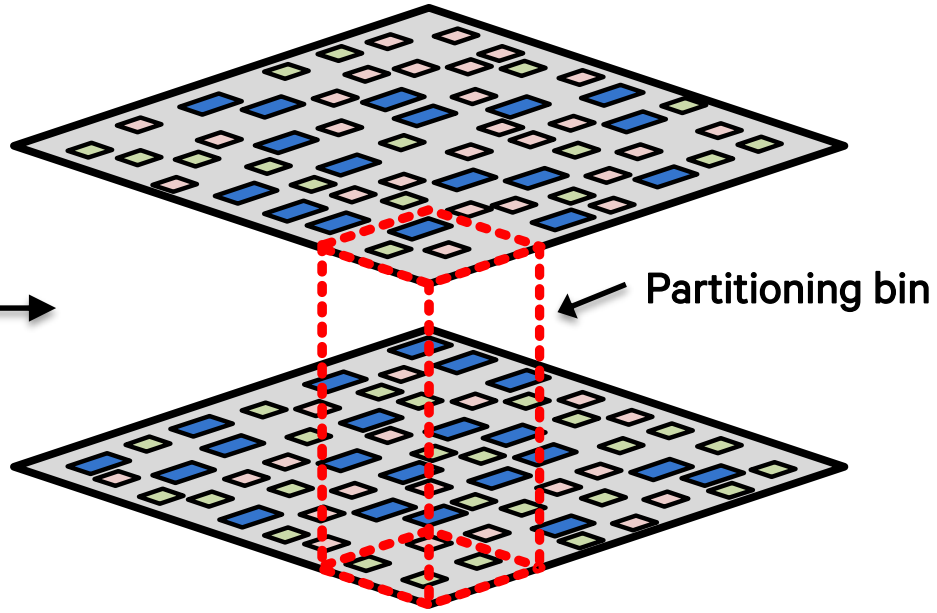
Initial Work in Gate-level 3D VLSI

- **Placement-driven partitioning using academic placers [1]**

First, make the 3D footprint 50% of 2D



In a 2D placer, double the placement capacity of each global bin (for two-tier)



Partition the design, maintaining local area balance within each partitioning bin
"Placement-driven Partitioning"

[1] S. Panth et. al., "Placement-Driven Partitioning for Congestion Mitigation in Monolithic 3D IC Designs", ISPD, 2014.

Courtesy Panth et al., ISLPED'14

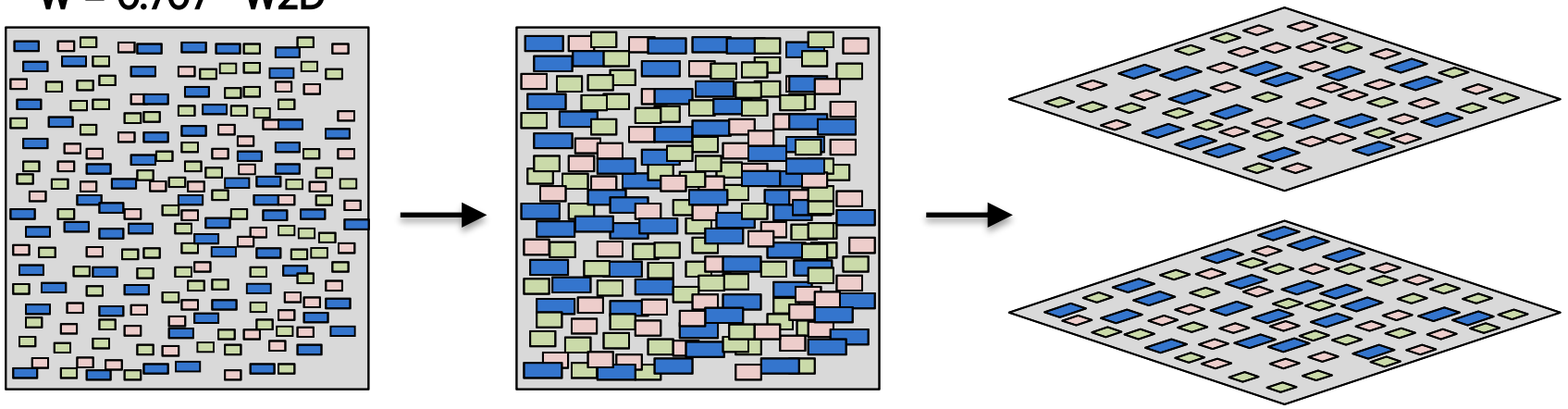
"Shrunk 2D" Placement using a Commercial Tool

- In a commercial tool, we cannot "double" the supply.
- Instead, we first halve the std. cell areas (multiply W/H by 0.707)

Original 2D Std. Cells → Shrunk 2D Std. Cells

Note: We do not touch the .lib file → Timing information is maintained

W = 0.707 * W2D

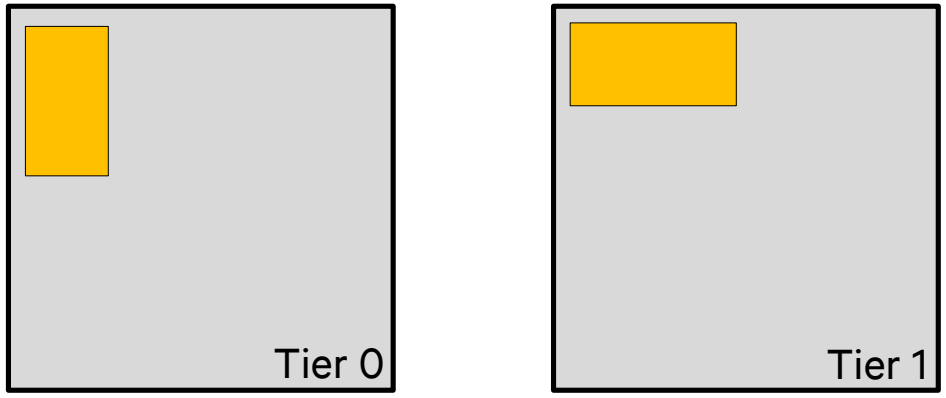


Shrunk 2D Placement → Cell Expansion → Placement-driven Partitioning

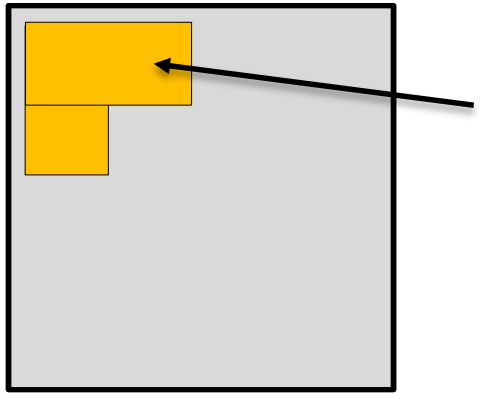
Courtesy Panth et al., ISLPED'14

Handling Memory Macros: Issues

- **Memory is usually pre-placed before placement starts**



- **We cannot simply superimpose them before feeding it to the commercial tool for shrunk 2D P&R**



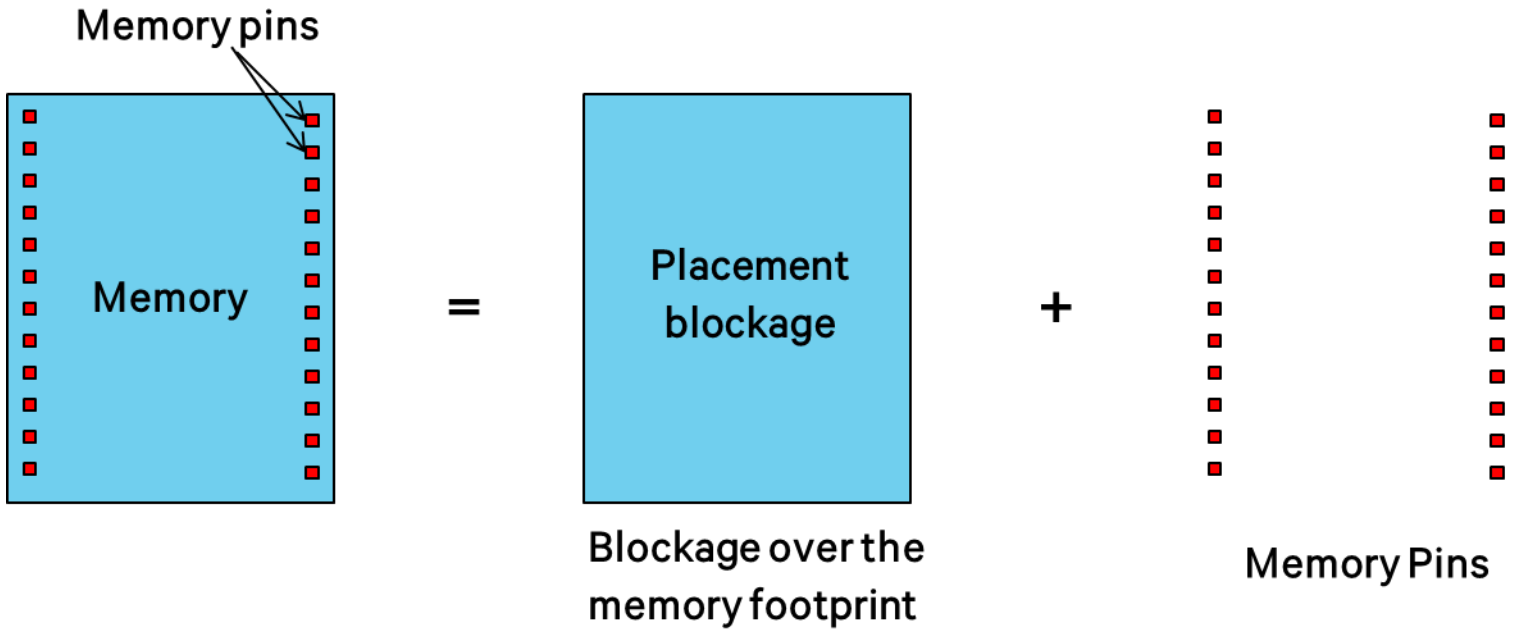
Shrunk 2D Footprint

This will cause a placement blockage in these regions, which is wrong

Courtesy Panth et al., ISLPED'14

Handling Memory Macros: Decomposition

- **Memory macros can be thought of as a combination of a placement blockage and memory pins**



- **If we can isolate each component, then they can be handled separately during shrunk 2D P&R**

Courtesy Panth et al., ISLPED'14

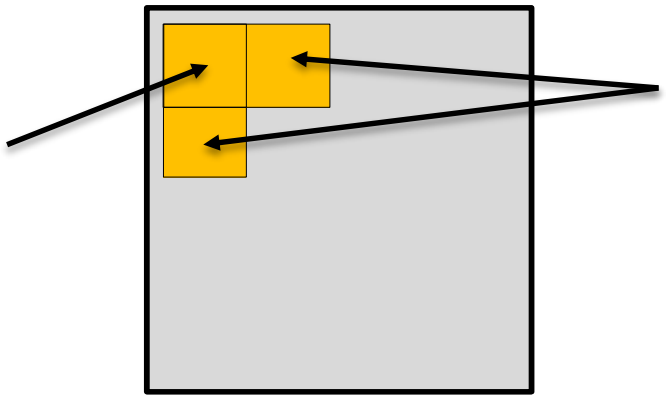
Handling Memory Placement Blockages (1/2)

- **Consider the two memory regions overlapping as shown earlier**

This region has memory in both tiers

After partitioning, neither tier will contain cells

Therefore, it will be a full placement blockage in the shrunk 2D footprint



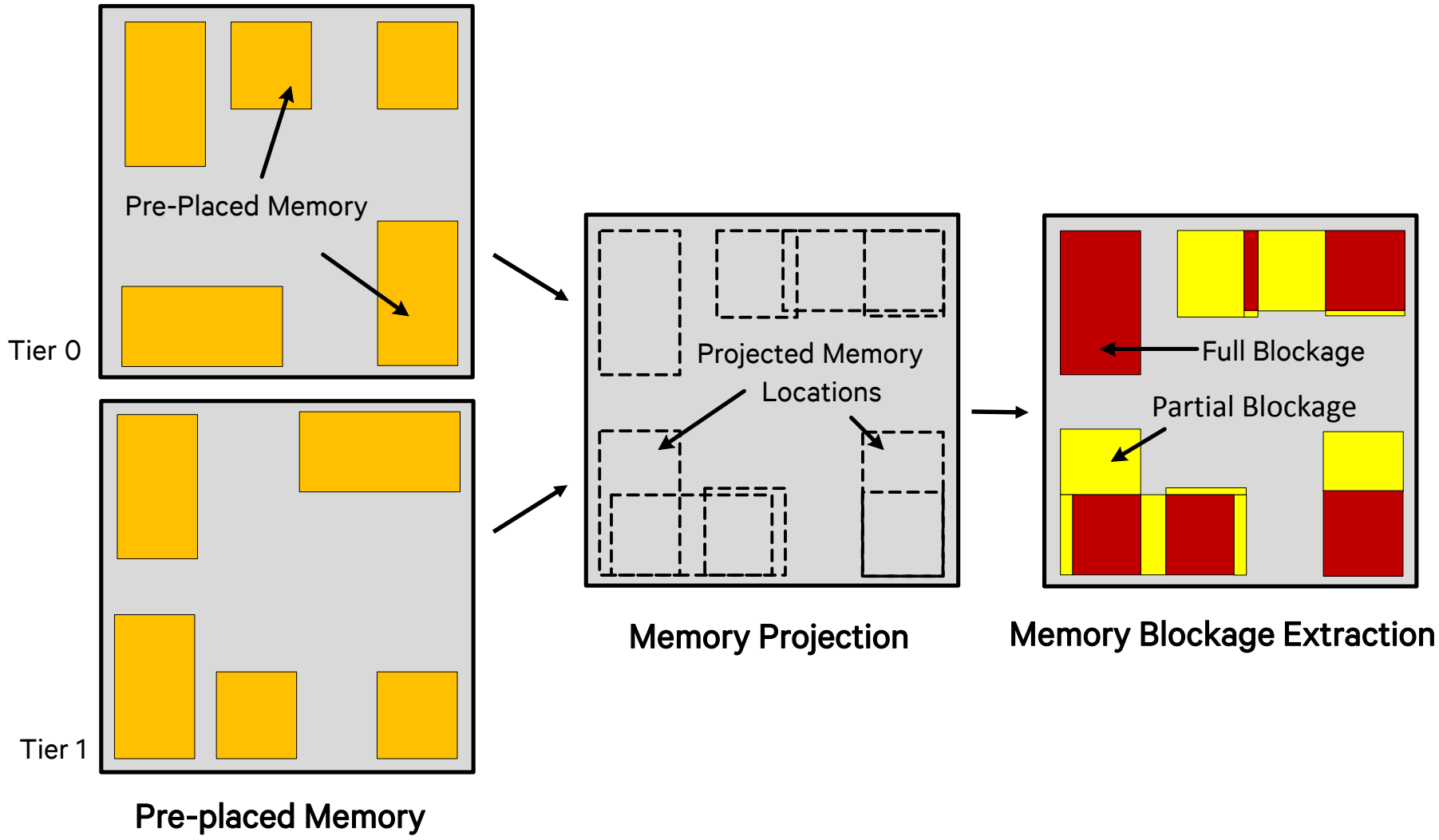
These regions have memory in one tier only → The other tier can contain cells

If the target density = 70% in the final 3D design, we set the max density of these regions = 35% (=70% once cells are expanded back to original area)

This can be achieved by creating a partial placement blockage in these areas

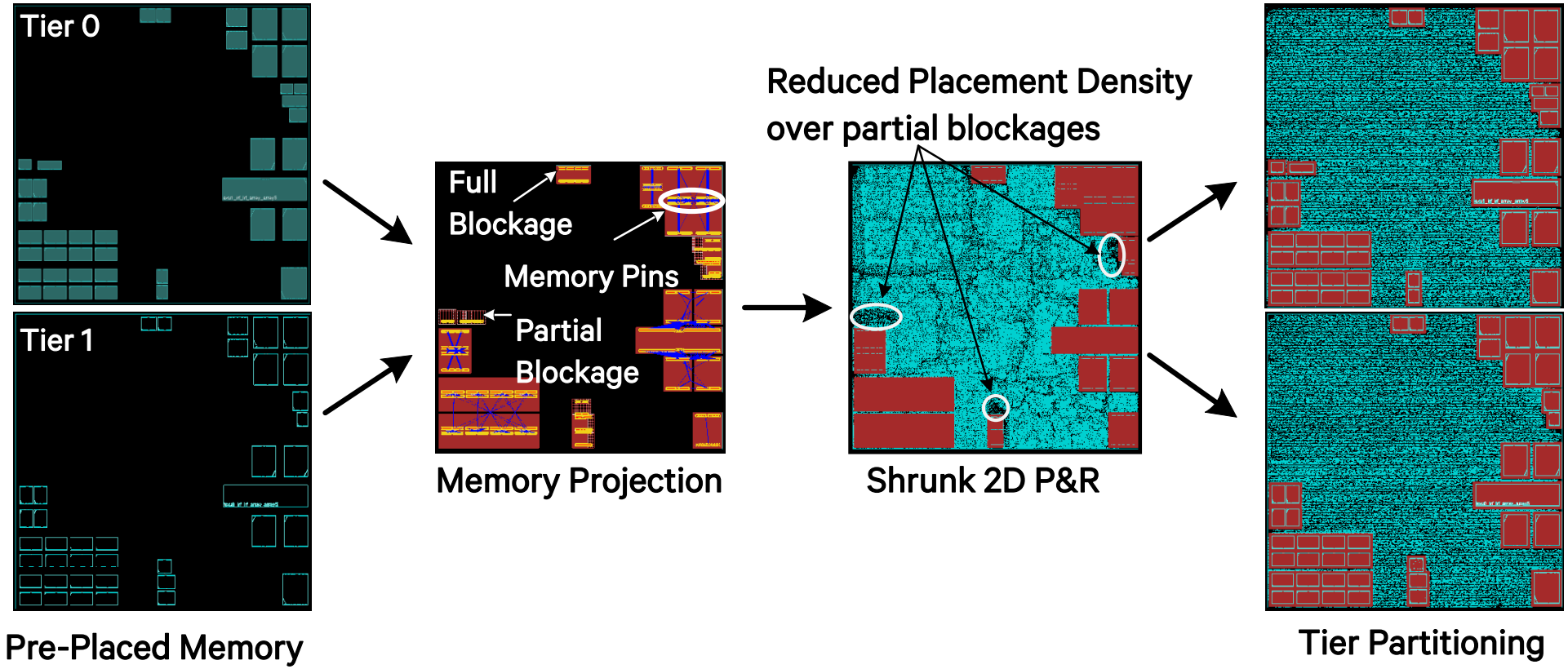
Courtesy Panth et al., ISLPED'14

Handling Memory Placement Blockages (2/2)



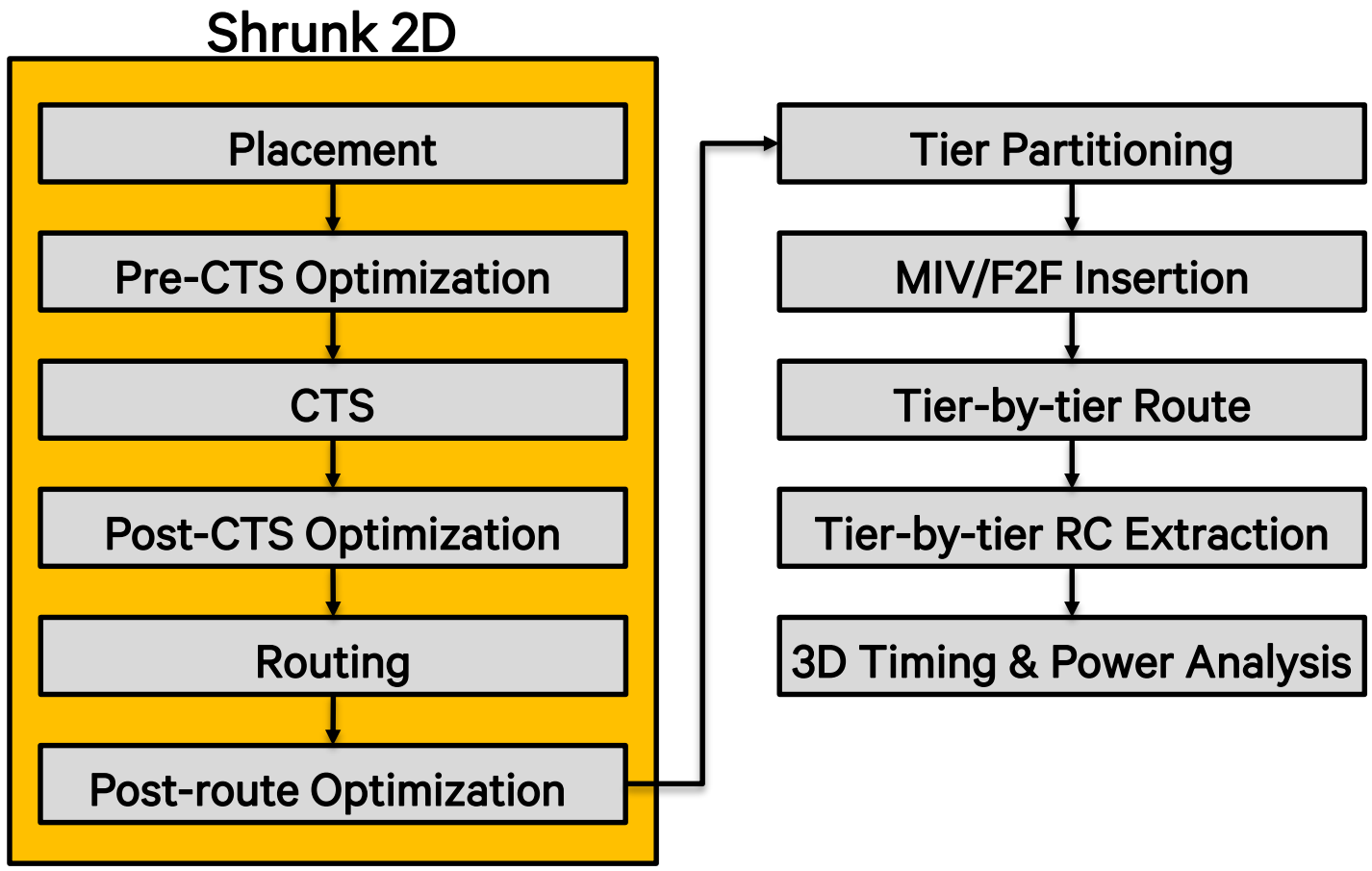
Courtesy Panth et al., ISLPED'14

Design Flow Screenshots



Courtesy Panth et al., ISLPED'14

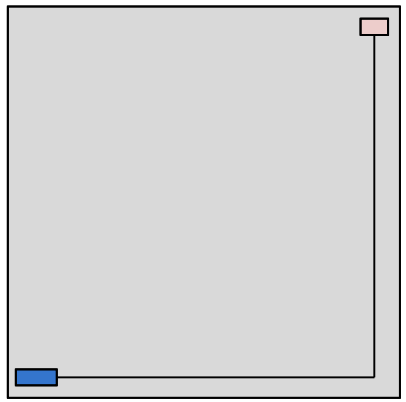
Design Flow



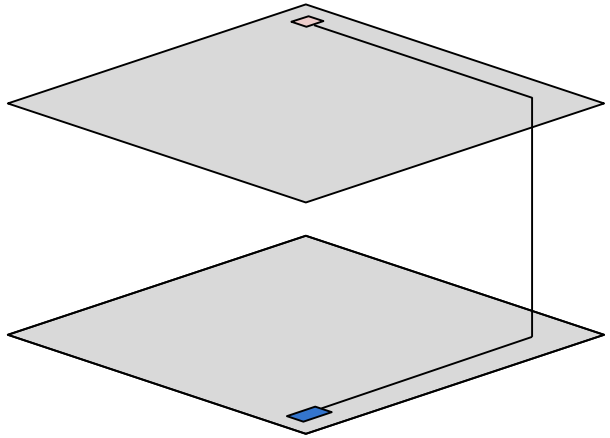
Courtesy Panth et al., ISLPED'14

Matching Wire Parasitics between Shrunk 2D and 3D

- Consider two cells connected to each other in Shrunk 2D & then in 3D



$L_{S2D} = L_x + L_y$



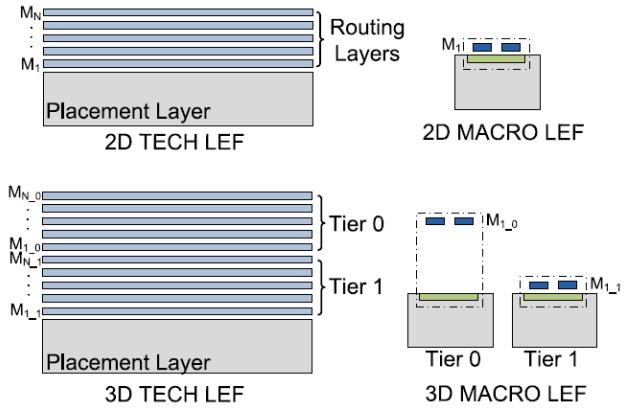
$L_{3D} = L_x + L_y + L_{MIV}$

- But L_{MIV} < 1um. Therefore $L_{S2D} \approx L_{3D}$
- However, the wire widths are different; $W_{S2D} = 0.707\ W_{3D}$
- Since we want $R_{S2D} \approx R_{3D}$ and $C_{S2D} \approx C_{3D}$, we **do not scale** the per-unit-length RC values in the cap table file for shrunk 2D design.

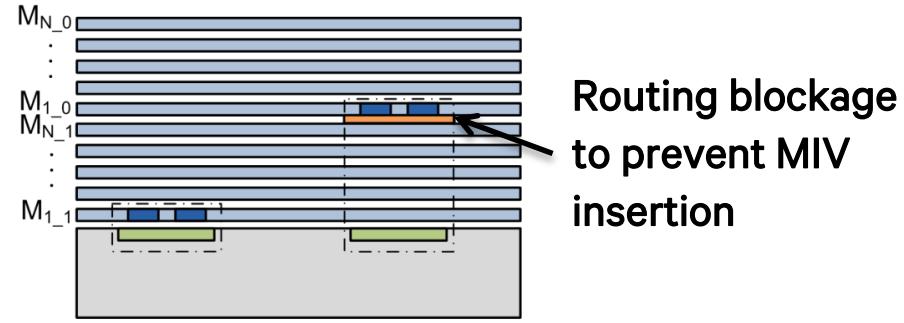
Courtesy Panth et al., ISLPED'14

Vertical Via Insertion (Face-to-Back)

- **Trick the commercial router into inserting MIVs for us [1]**

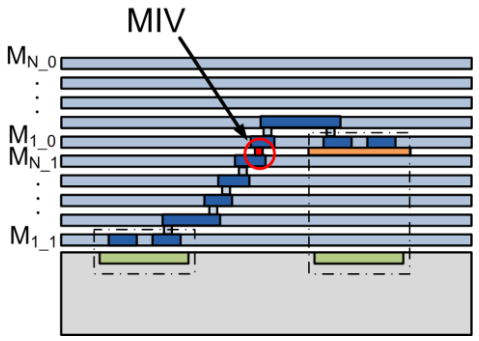


LEF files are modified for 3D

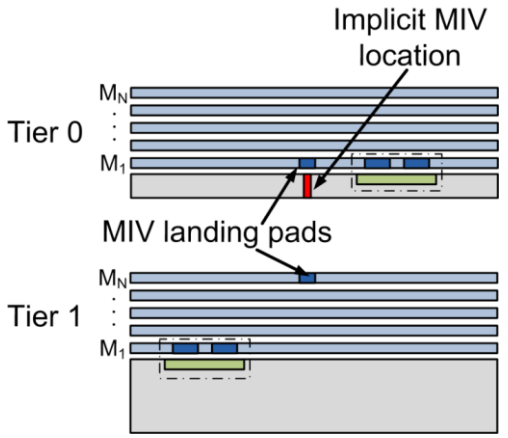


Routing blockage to prevent MIV insertion

All gates are then placed in the same placement layer



Route with Encounter

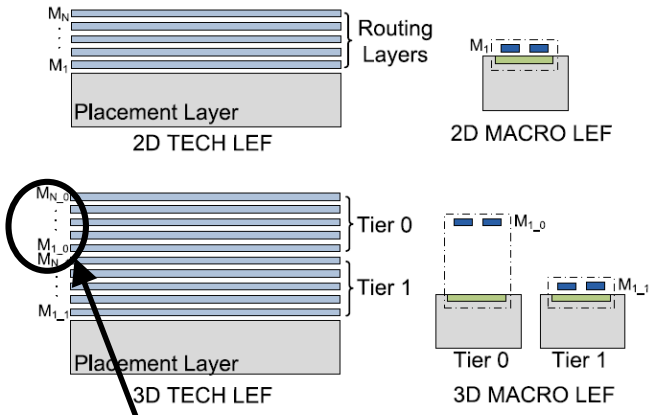


Create separate verilog/DEF for each tier

[1] S. Panth et. Al., "Placement-Driven Partitioning for Congestion Mitigation in Monolithic 3D IC Designs", ISPD 2014.

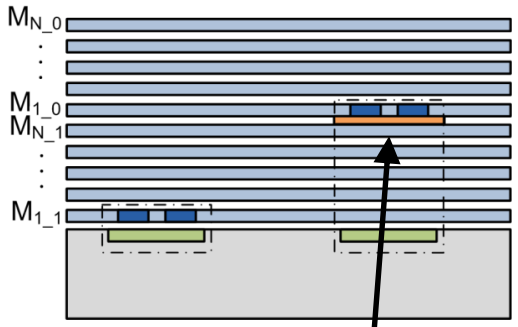
Courtesy Panth et al., ISLPED'14

Vertical Via Insertion (Face-to-Face)



LEF files are modified for 3D

Reverse the order of metals in Tier 0

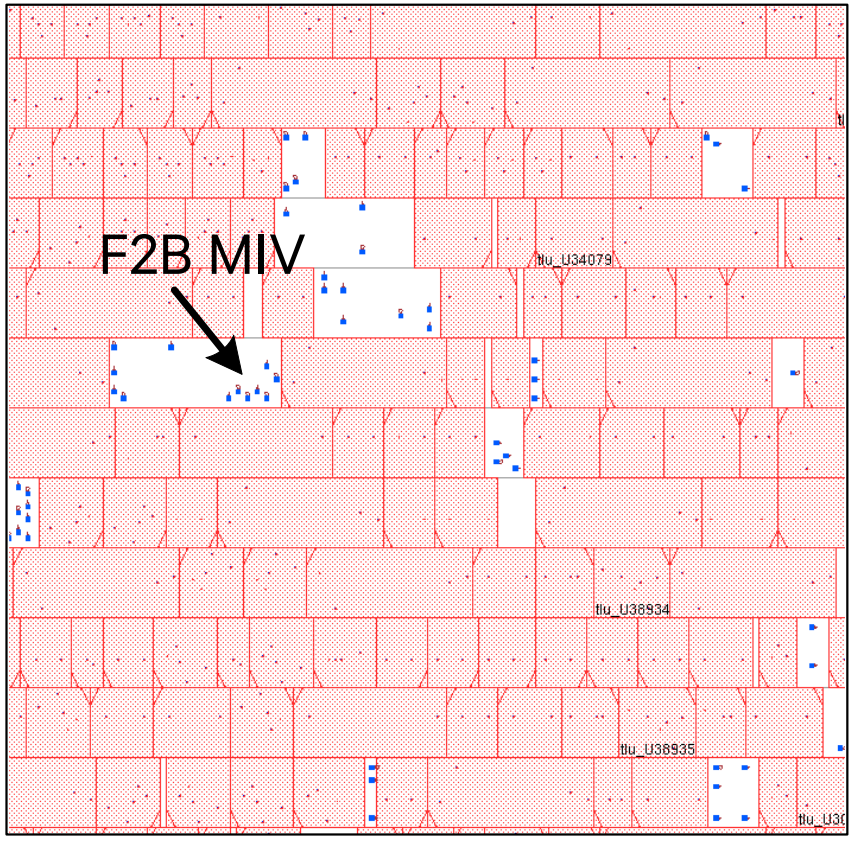


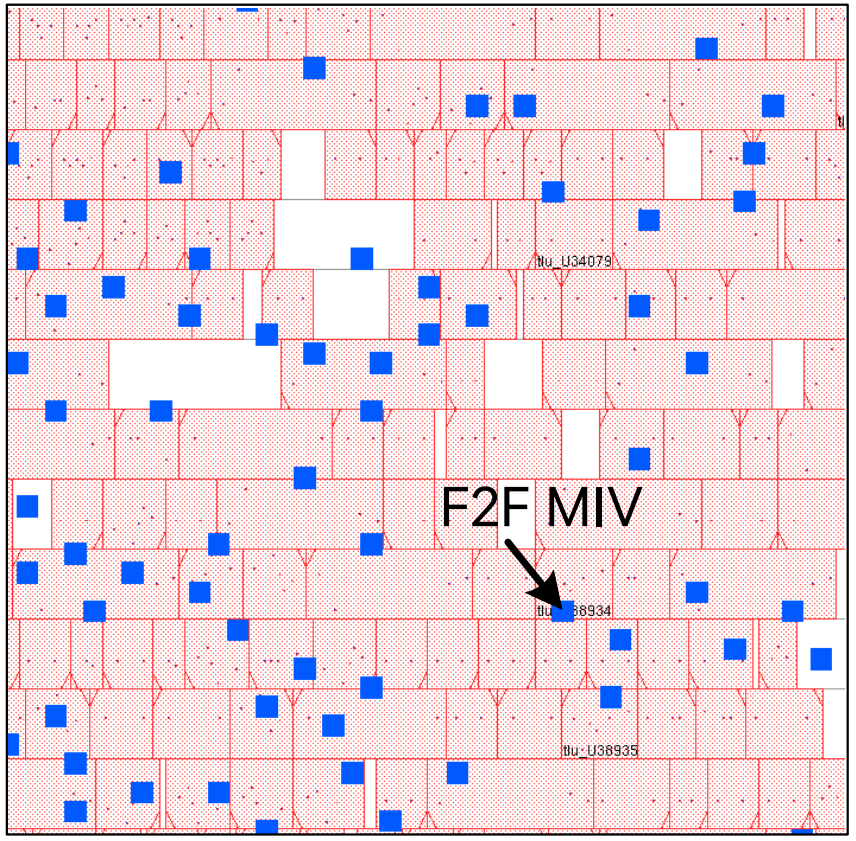
All gates are then placed in the same placement layer

Do not add any routing blockage to prevent F2F insertion over cells

Courtesy Panth et al., ISLPED'14

F2B / F2F MIV Insertion Screenshots

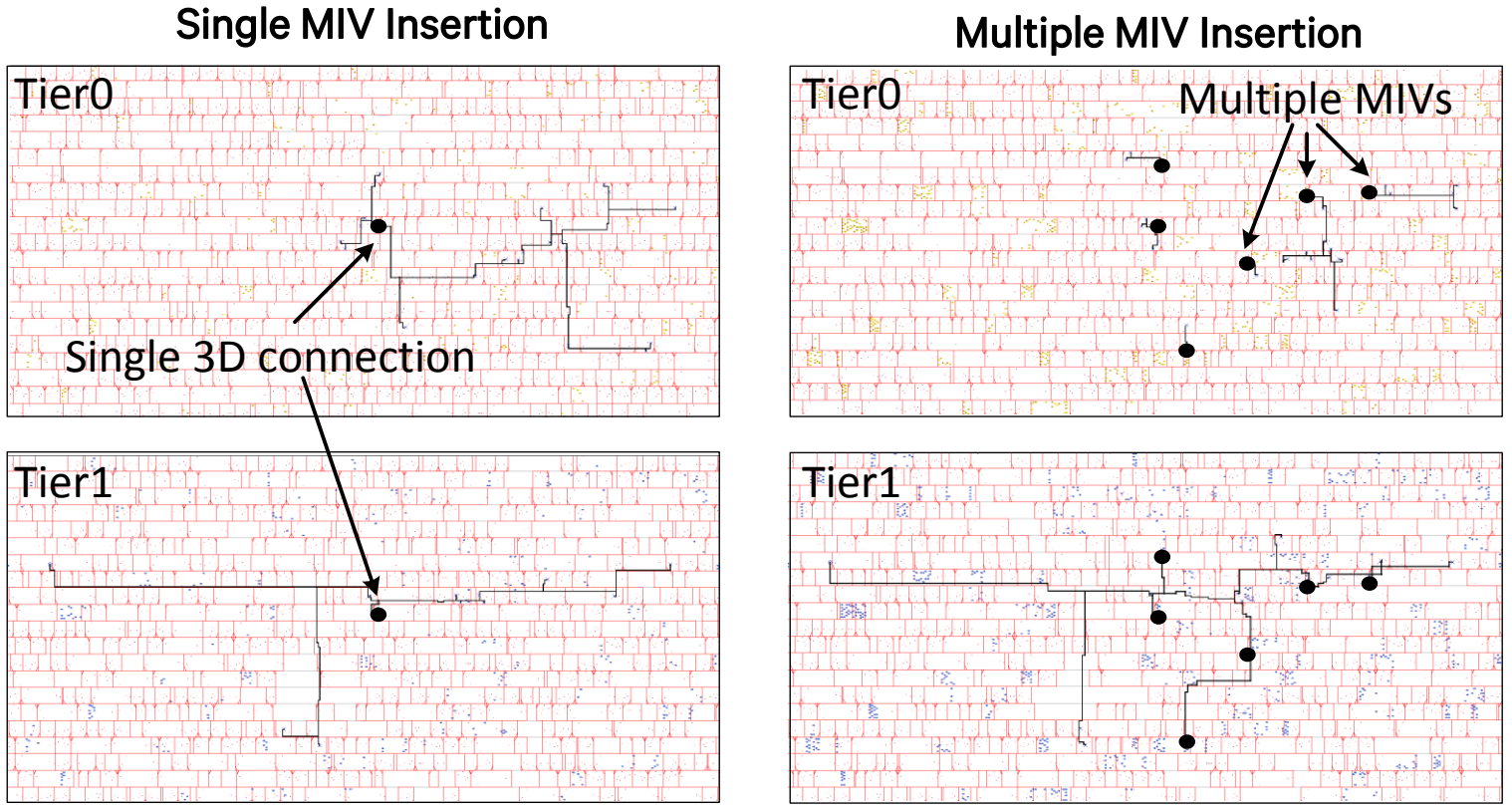




Courtesy Panth et al., ISLPED'14

Single vs. Multiple MIV Insertion: Screenshots

- Conventional 3D flows have a tier-by-tier optimization step
- It is very difficult to derive timing budgets for multiple MIVs per net
- Shrunk 2D flow enables multiple MIV insertion → Lower WL and power



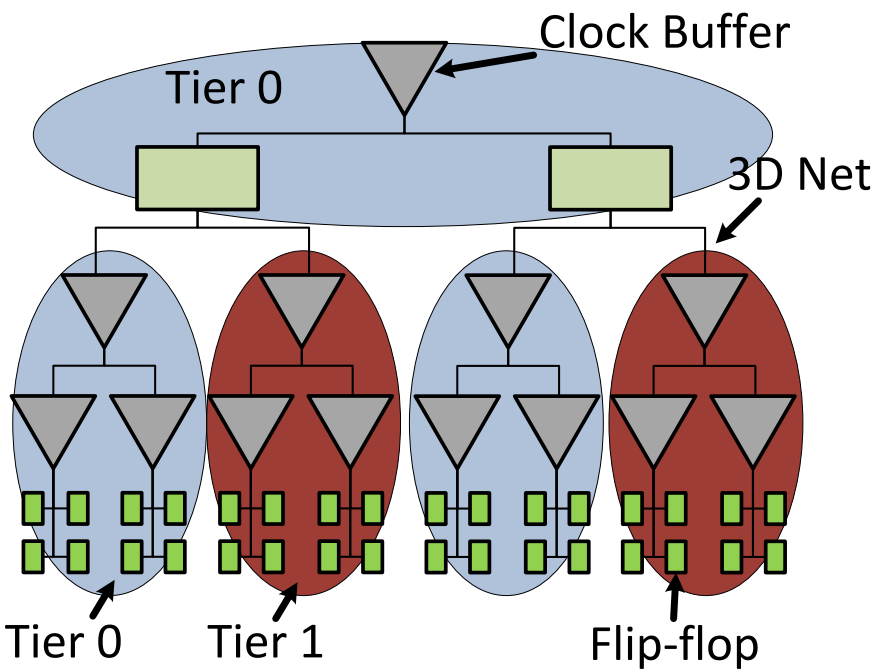
Courtesy Panth et al., ISLPED'14

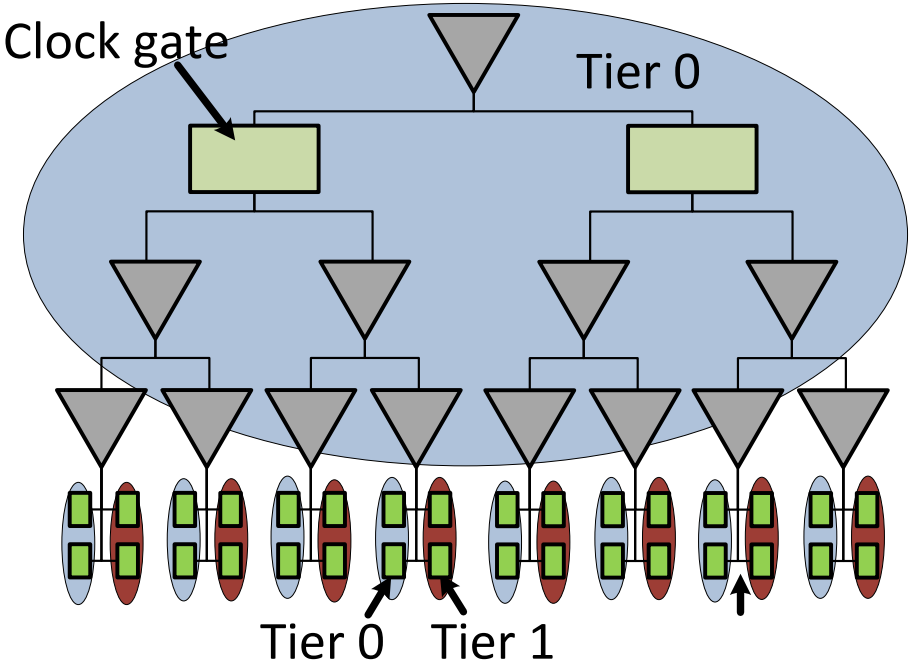
Single vs. Multiple Vertical Via Insertion: Results

	F2B 3D		F2F 3D	
	Single	Multiple	Single	Multiple
Total WL (m)	15.61	14.29 (-8.43%)	15.44	13.89 (-10.05%)
#MIV/F2F	106k	235k (+120.44%)	106k	202k (+89.72%)
Total Power (mW)	534.10	522.10 (-2.25%)	538.30	524.00 (-2.66%)
Cell Power (mW)	126.90	126.10 (-0.63%)	127.30	126.40 (-0.71%)
Net Power (mW)	293.90	282.70 (-3.81%)	297.80	284.30 (-4.53%)
Leak. Power (mW)	113.30	113.30 (+0.0%)	113.30	113.30 (0.00)

Courtesy Panth et al., ISLPED'14

3D Clock-Tree Synthesis





Traditional 3D CTS: Source-level

One clock-tree per clock-gating group in each tier, tied together at the root level

Proposed 3D CTS: Leaf-level

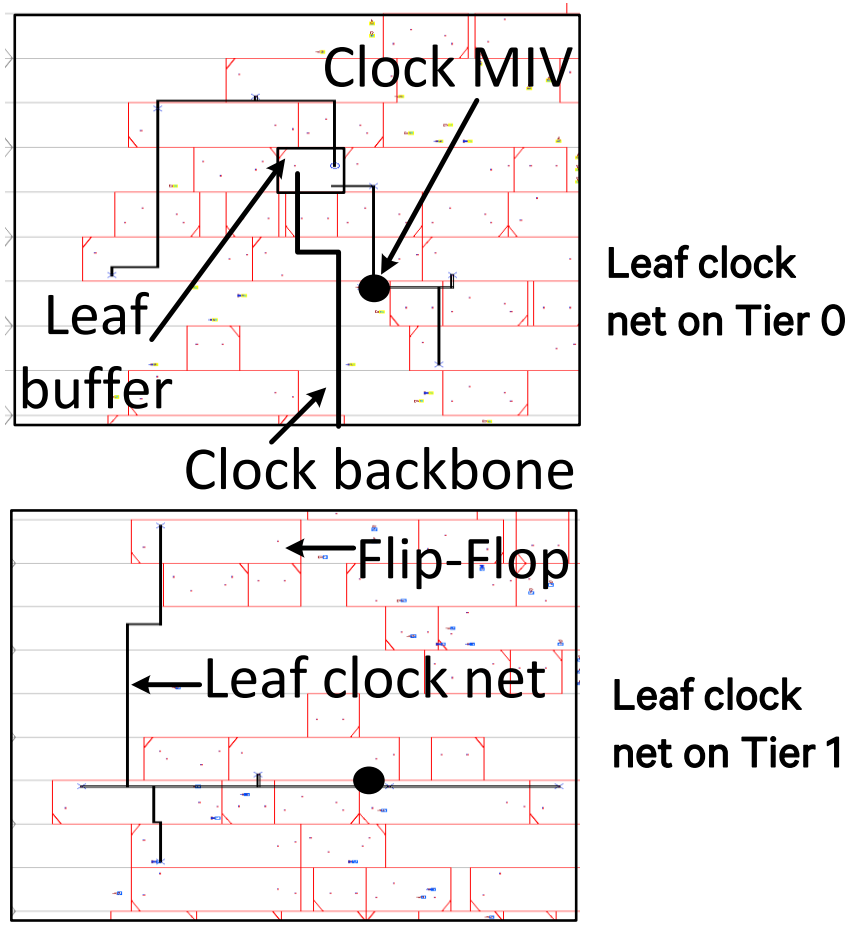
Keep the entire backbone on one tier. Only insert clock MIVs to connect the FF on different tiers at the leaf level

Courtesy Panth et al., ISLPED'14

Leaf-Level CTS: Screenshots

Clock back-bone on Tier 0

Zoom in of red rectangle



Leaf clock net on Tier 0

Leaf clock net on Tier 1

Courtesy Panth et al., ISLPED'14

CTS Results

	F2B 3D		F2F 3D	
	Source-level	Leaf-level	Source-level	Leaf-level
#Vertical Via	871	11,376 (+1.2k%)	871	11,376 (+1.2k%)
Clock Skew (ps)	197.42	103.00 (-47.83%)	172.90	117.07 (-32.29%)
Clock Power (mW)	68.40	48 (-29.82%)	69.00	48.50 (-29.71%)
Clk WL – Tier 0 (m)	0.55	0.62 (+11.89%)	0.53	0.62 (+16.61%)
Clk WL – Tier 1 (m)	0.48	0.19 (-60.50)	0.48	0.17 (-64.85%)
Total Clk WL (m)	1.03	0.80 (-21.67%)	1.01	0.79 (-21.91%)
# Clk Buf – Tier 0	14,610	21,687 (+48.44%)	14,958	21,687 (+44.99%)
# Clk Buf – Tier 1	12,444	0 (-100%)	12,691	0 (-100%)
Total # Clk Buf	27,054	21,687 (-19.84%)	27,649	21,687 (-21.56%)

Courtesy Panth et al., ISLPED'14

Single Vt Power Comparisons (mW)

	Encounter 2D	Shrunk 2D	F2B 3D	F2F 3D
Total	618.40	514.40 (-16.82%)	522.10 (-15.57%)	524.00 (-15.27%)
Cell	135.60	126.80 (-6.49%)	126.10 (-7.01%)	126.40 (-6.78%)
Net	356.30	274.30 (-23.01%)	282.70 (-20.66%)	284.30 (-20.21%)
Leakage	126.50	113.30 (-10.43%)	113.30 (-10.43%)	113.30 (-10.43%)
Memory	49.00	45.10 (-7.96%)	45.10 (-7.96%)	45.00 (-8.16%)
Combinational	385.10	300.00 (-22.10%)	305.30 (-20.72%)	306.80 (-20.33%)
Clock Tree	62.50	46.90 (-24.96%)	48.00 (-23.20%)	48.50 (-22.40%)

Courtesy Panth et al., ISLPED'14

Dual Vt Power Comparisons (mW)

	Encounter 2D	Shrunk 2D	F2B 3D	F2F 3D
Total	572.10	471.4 (-17.60%)	480.10 (-16.08%)	482.20 (-15.71%)
Cell	131.80	122.5 (-7.06%)	123.00 (-6.68%)	123.30 (-6.45%)
Net	356.60	274.2 (-23.11%)	282.70 (-20.72%)	284.30 (-20.27%)
Leakage	83.60	74.7 (-10.65%)	74.70 (-11.00%)	74.60 (-10.77%)
Memory	48.80	45.1 (-7.58%)	45.10 (-7.58%)	45.00 (-7.79%)
Combinational	361.60	278.6 (-22.95%)	283.00 (-21.74%)	284.30 (-21.38%)
Clock Tree	62.50	47.3 (-24.32%)	48.00 (-23.20%)	48.50 (-22.40%)

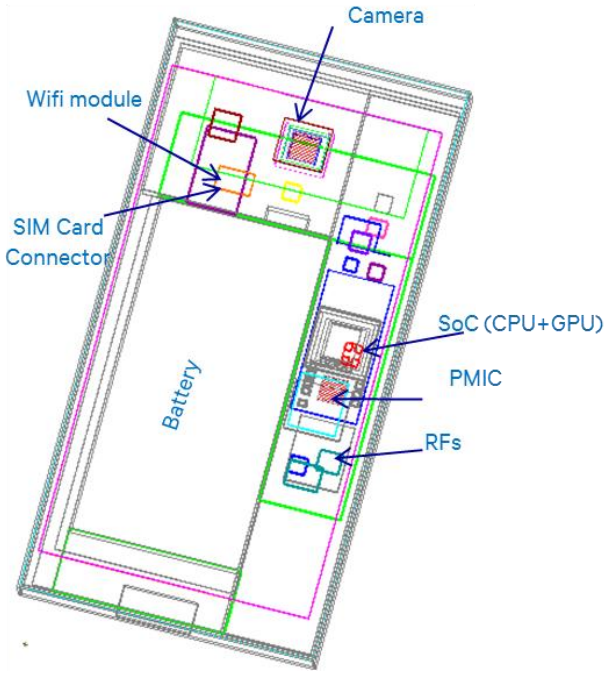
Courtesy Panth et al., ISLPED'14

Thermal Implications of 3D ICs

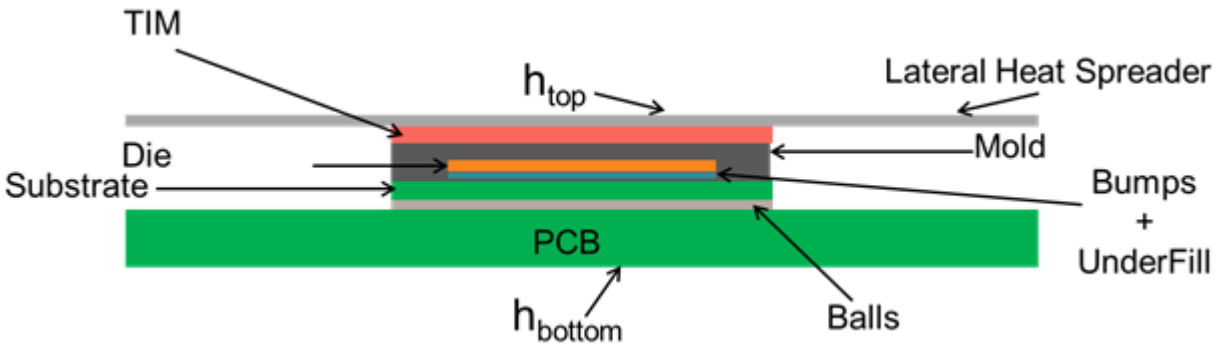
Thermal Model Setup – Die, Package and Cooling Mechanism

- Compact model derived from full system cellphone model
- SoC temperature matched with full system model

Typical Mobile Full System Thermal Model

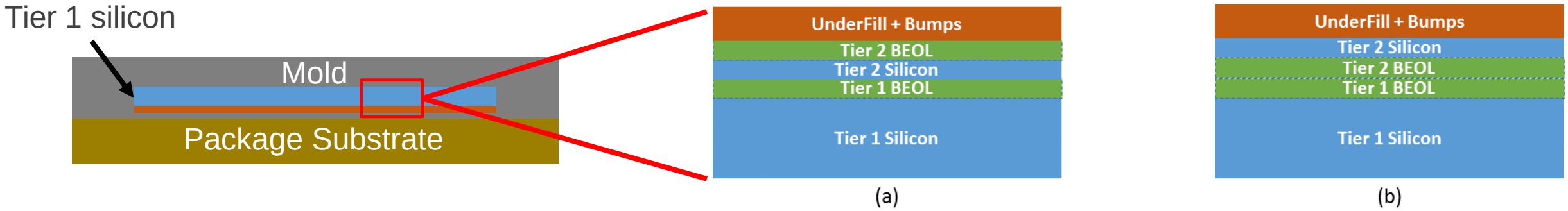


Compact Model
(SoC Package, board and top side heat spreader only)



Thermal Model Setup – F2B and F2F Stack up

- 3D stack-up for Face-to-Face (F2F) and Face-to-Back (F2B)
  - Back-End-Of-Line (BEOL) modeled as one layer with effective thermal properties
  - Tier 2 Silicon with vertical connections modeled as one layer with effective thermal properties

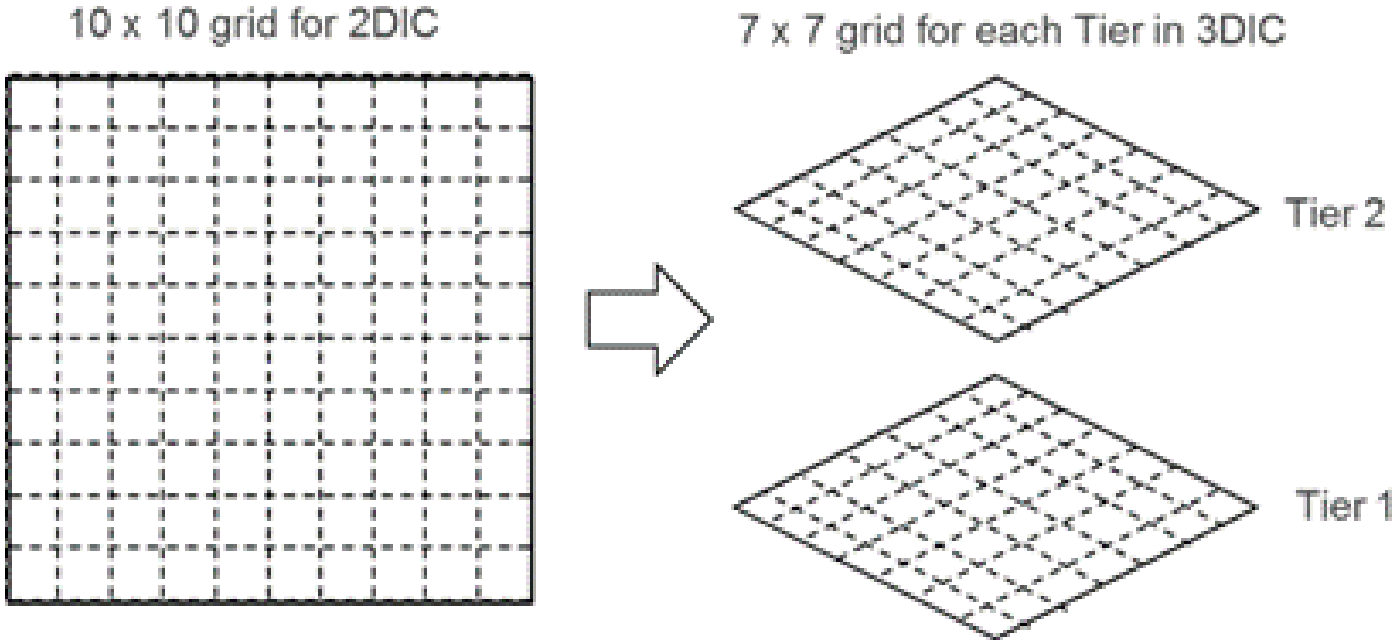


(a) F2B stack-up

(b) F2F stack-up

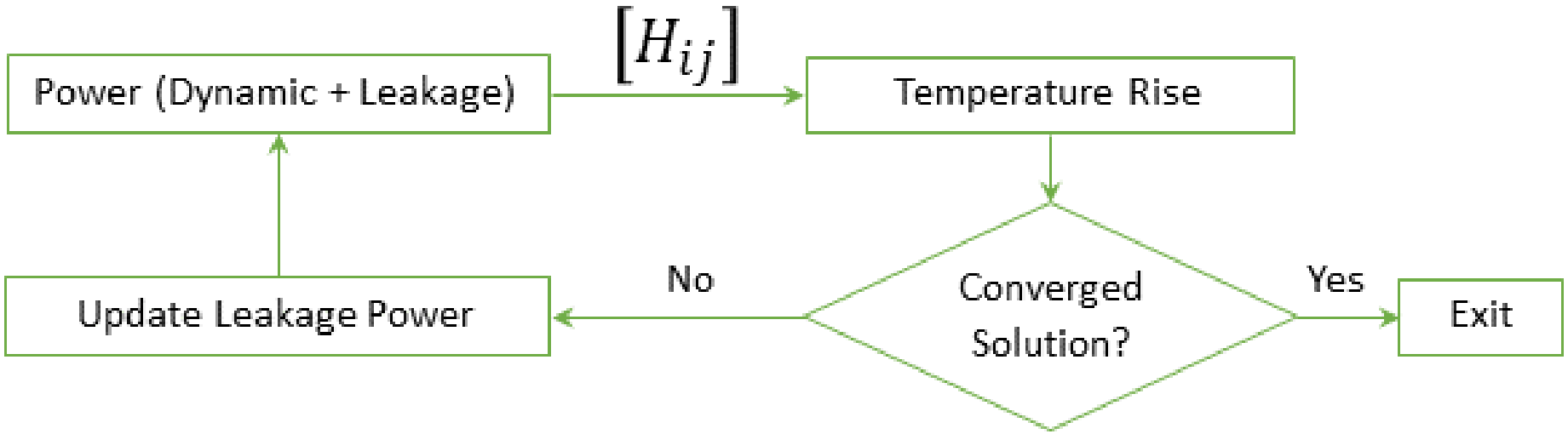
Model Generation – Power Mapping

- 2D and 3D silicon areas are similar ~ 50% area footprint shrink
- 1x1 mm^2 squares → 100 mm^2 2D design



Model Generation – Simulation Methodology

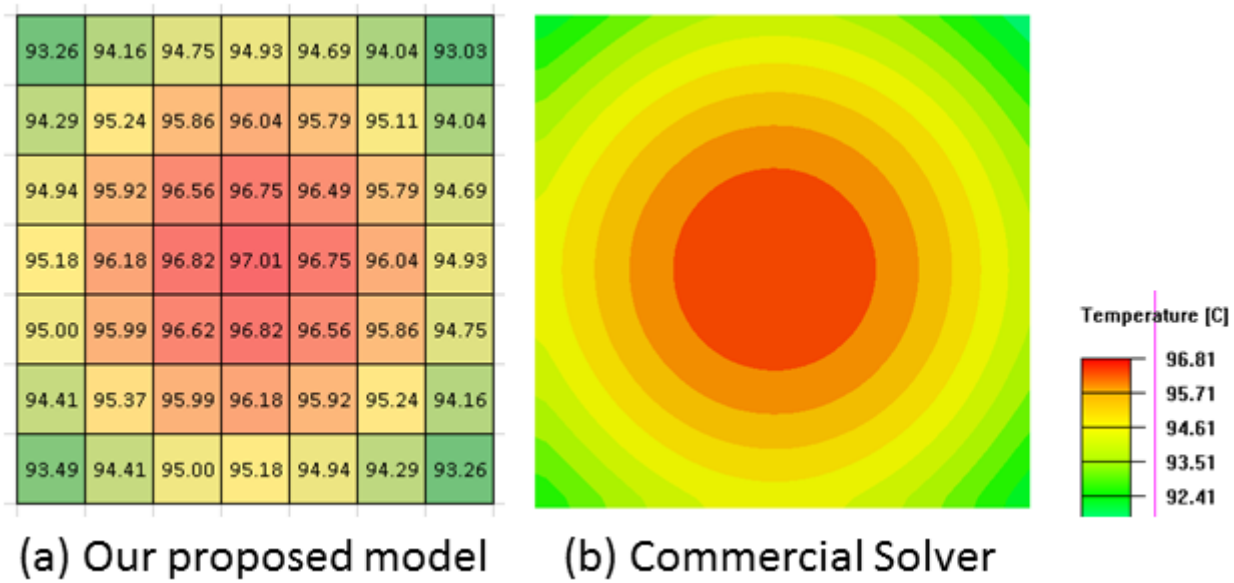
- Influence coefficient methodology
- $\Delta T = \sum H_{ij} \times P_i$
- Temperature-dependent leakage power loop



ICM Model Accuracy

- ICM based solver data compared with simulation using commercial thermal analysis tool, ICEPAK
  - Temperature delta (ICM- ICEPAK) = 0.2˚C
- Excellent match achieved between finite-volume analysis vs. ICM method

SoC Temperature distribution



(a) Our proposed model (b) Commercial Solver

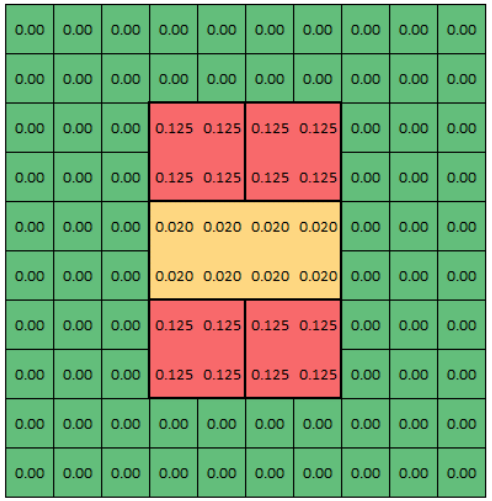
SoC Temperature Difference

ICM Max Temp (°C)	Commercial Solver Max Temp (°C)	Temp Rise Diff (°C)	% Temp Rise Diff
97.0	96.8	0.2	0.3

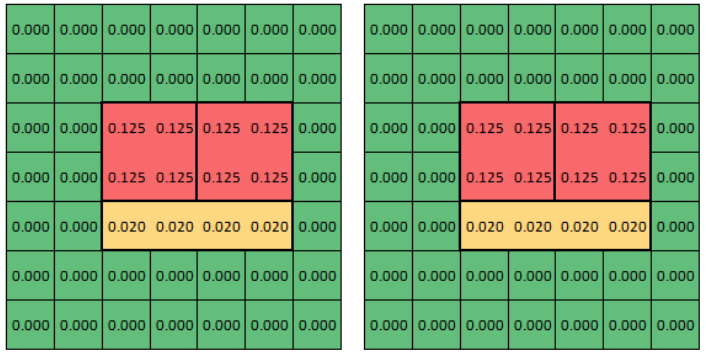
3D Thermal Characteristics

- Power input: power distribution for a typical quad-core CPU case
- Results: Temperature difference between tiers is small < 1˚C
  - Distance between tiers is very small → small thermal resistance between tiers → good mutual heating
  - Power is low on each tier: ΔT = R x P

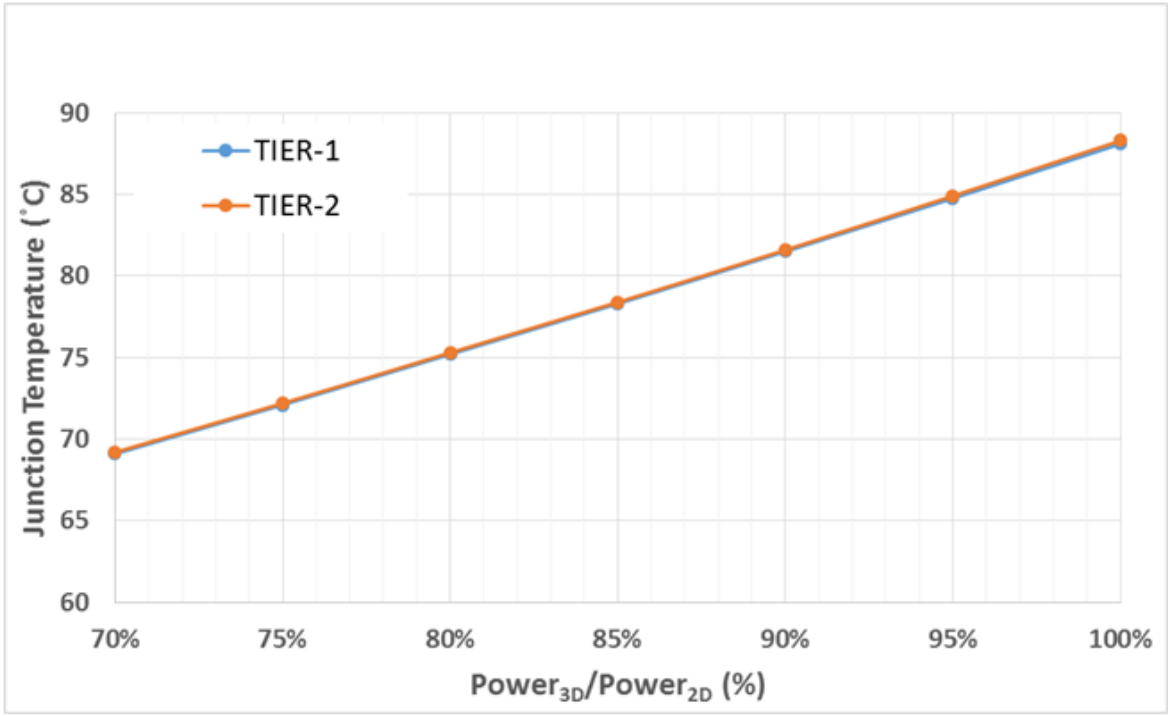
Dynamic Power Distribution



(a) 2D design

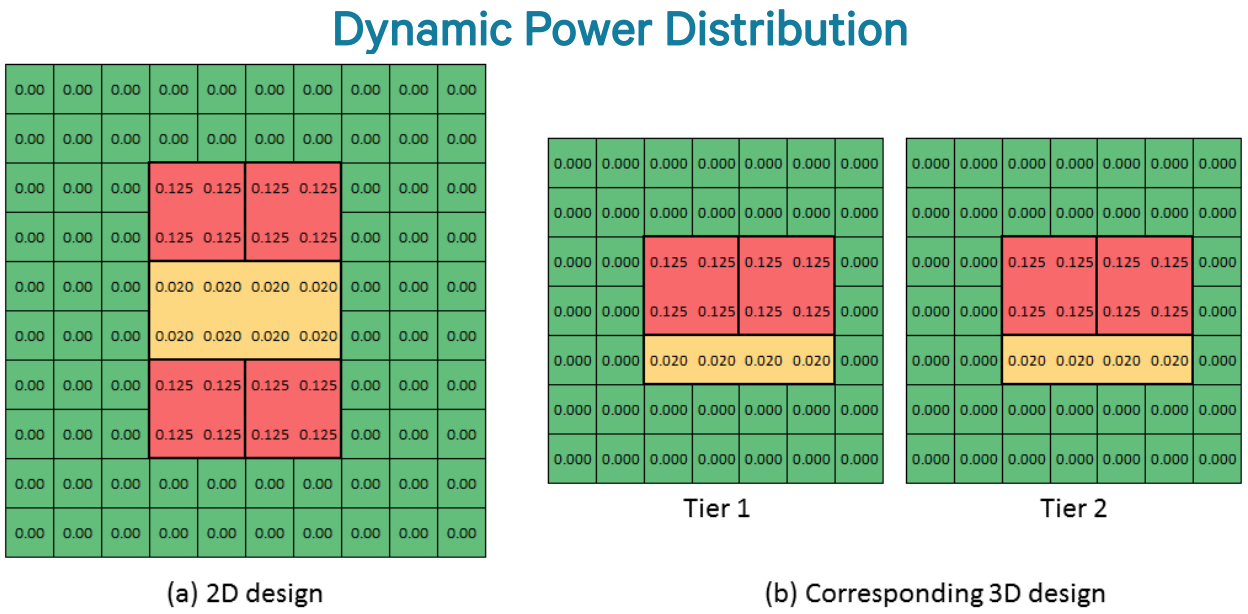


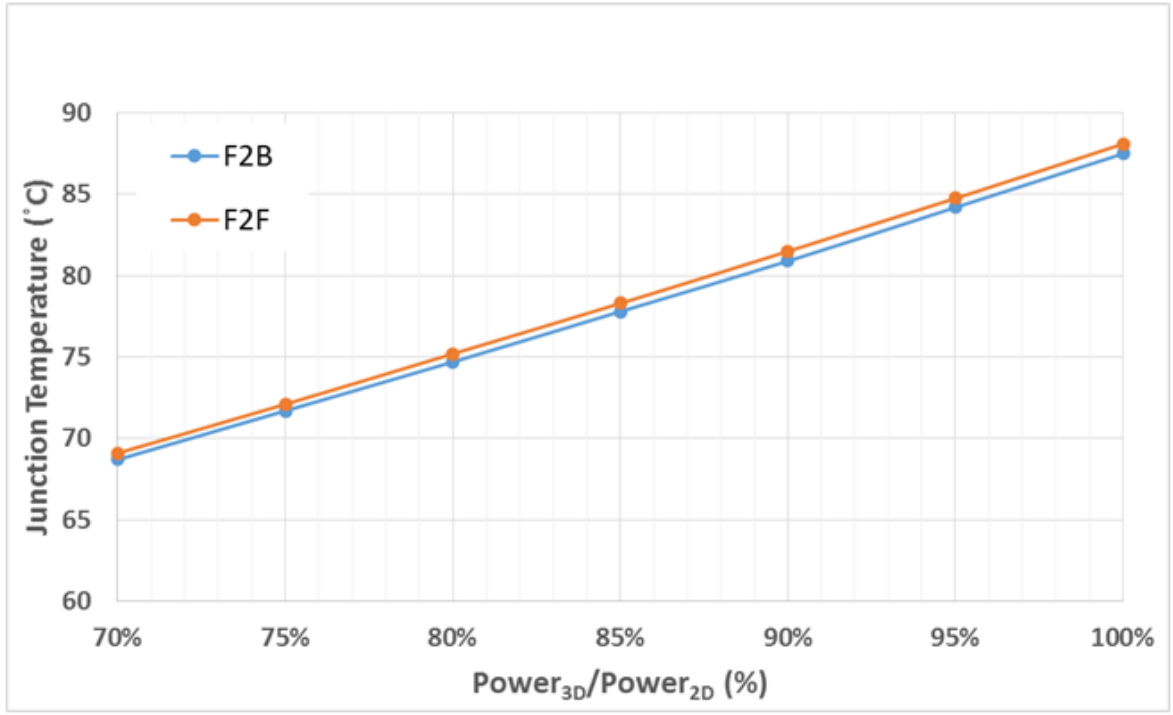
(b) Corresponding 3D design



3D Thermal Characteristics – F2B vs. F2F Integration

- Power input: power distribution for a typical quad-core CPU case
- F2F integration is slightly hotter than F2B (~ 1˚C)
  - In F2F, the active layers are closer to each other
  - In F2B there is a layer of thin silicon between the two tiers which slightly helps with temperature reduction

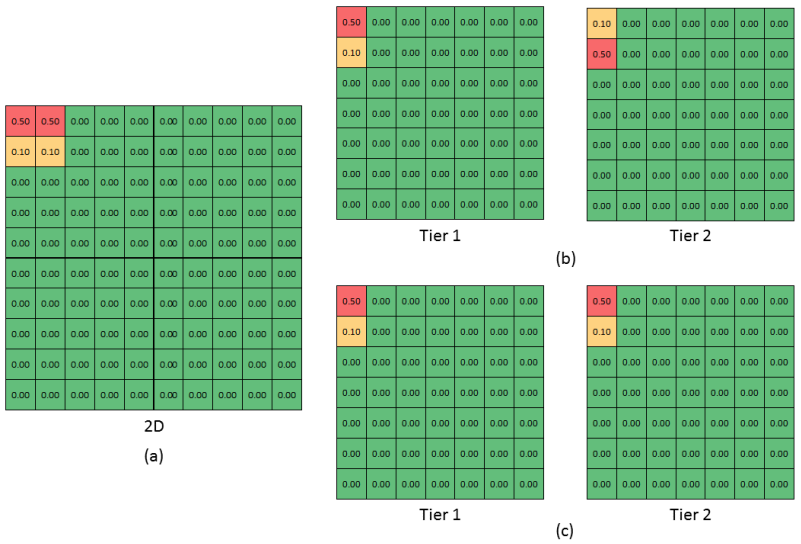




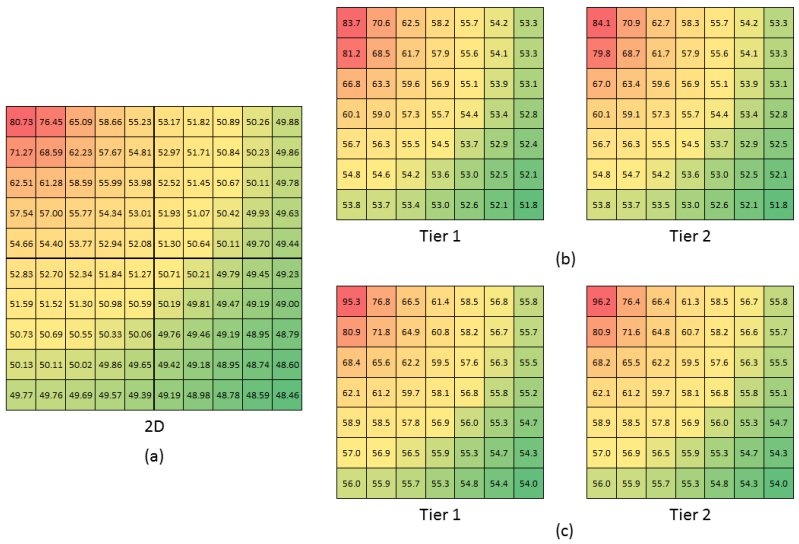
Temperature Rise in 3D vs. 2D: IP Block Partitioning Impact

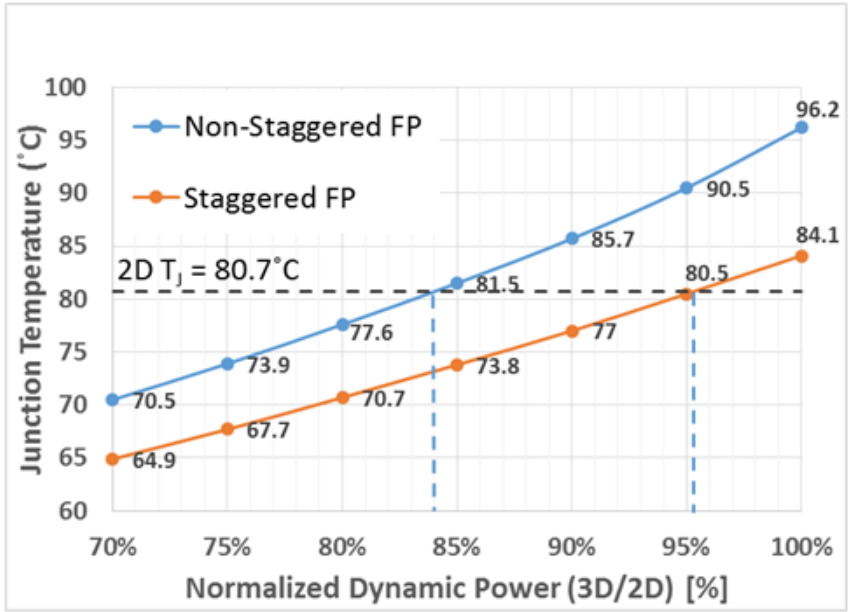
- With the same power inputs, 3D temperature is higher than 2D
- 3D temperature is very sensitive to IP block partitioning
  - Non-staggered partitioning results is higher junction temperatures, requiring 16% power reduction in 3D to match 2D
  - With only 5% power reduction in staggered partitioning, 3D temperature matches with 2D

Dynamic Power Distribution (W)



Temperature Distribution (˚C)

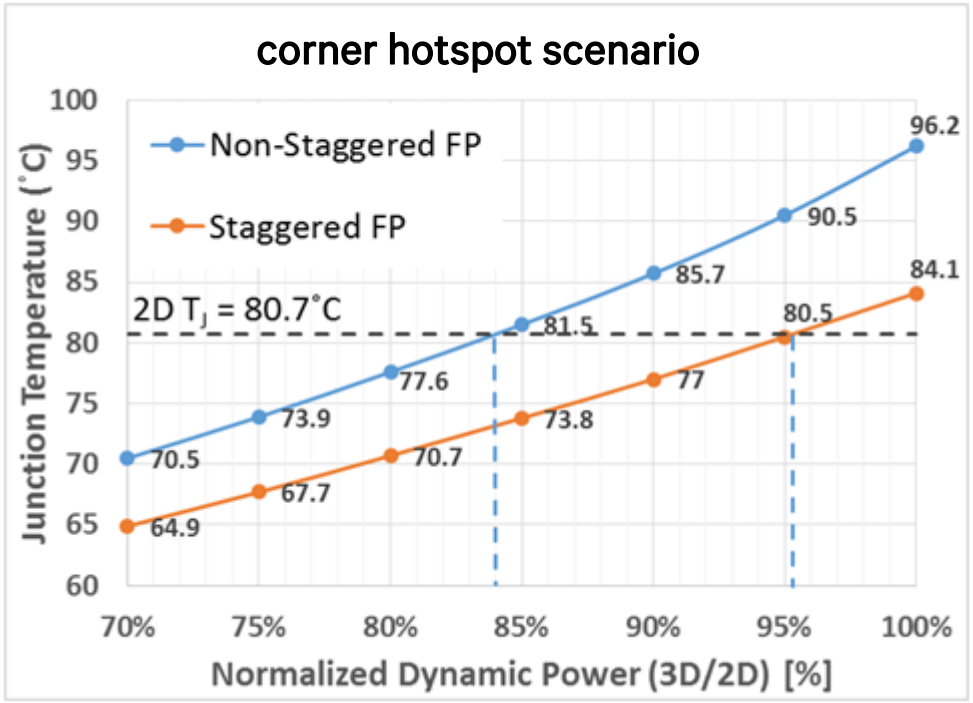


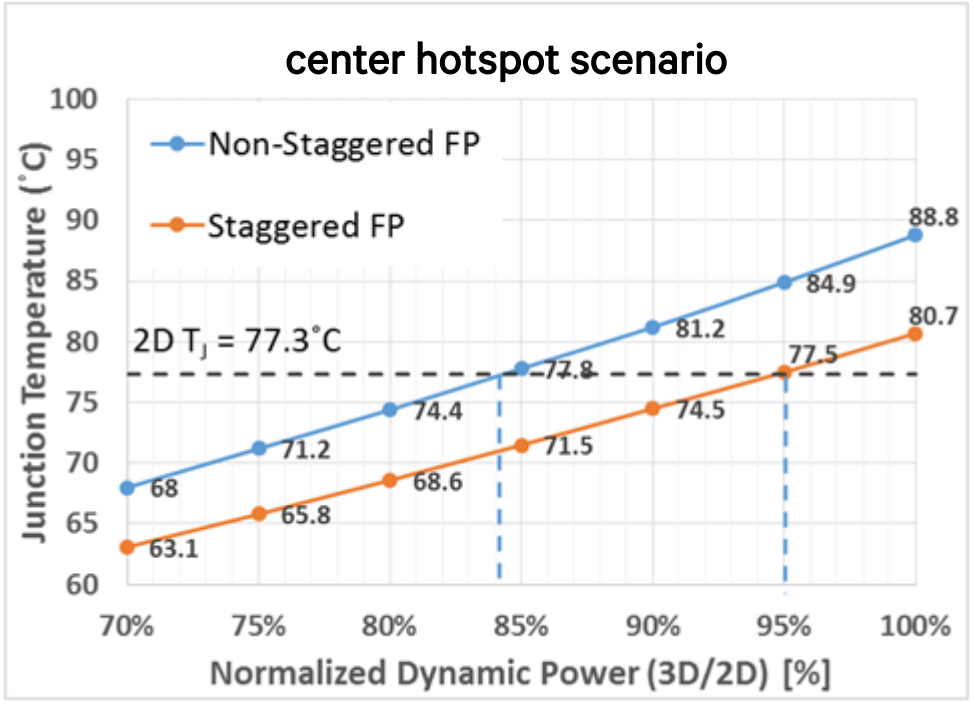


(a) Baseline 2D, (b) Staggered and (c) Non-staggered designs

Temperature Rise in 3D vs. 2D: Floorplanning Impact

- 3D thermal risk is lower if the high power density is placed in the center of the die
  - Temperature rise is significantly lower for center IP block (96.2C vs. 88.8C)
- Power saving requirement is the same for both floorplans
  - ~5% for staggered partitioning
  - ~16% for non-staggered partitioning
- Power saving is more sensitive to partitioning than floorplanning





Power Saving Opportunities in 3D

- Power savings are primarily coming from wirelength and buffer reductions
- $P_{total} = p_{internal} + p_{switching} + p_{leakage}$
- $p_{switching}^{3D} = \left(\alpha \cdot c_{pin} + \beta \cdot c_{wire}\right) \cdot p_{switching}^{2D}$
- Internal and leakage components are proportional to total cell area

Floorplanning and Partitioning Options in 3D

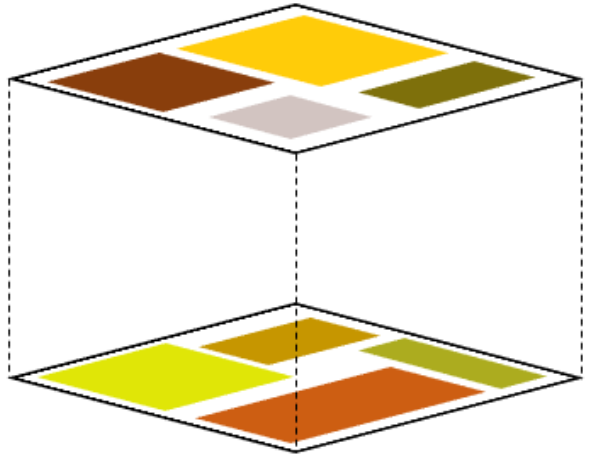
(a) 2D block partitioning with 2D floorplanning

(b) 2D block partitioning with 3D floorplanning

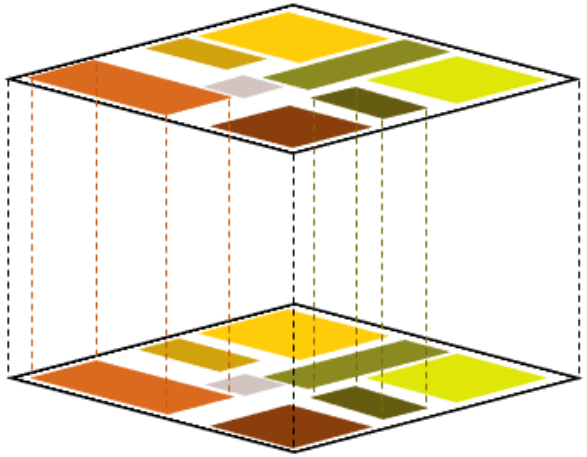
(c) 3D block partitioning with 2D floorplanning

(d) Combination of 2D and 3D block partitioning with 3D floorplanning

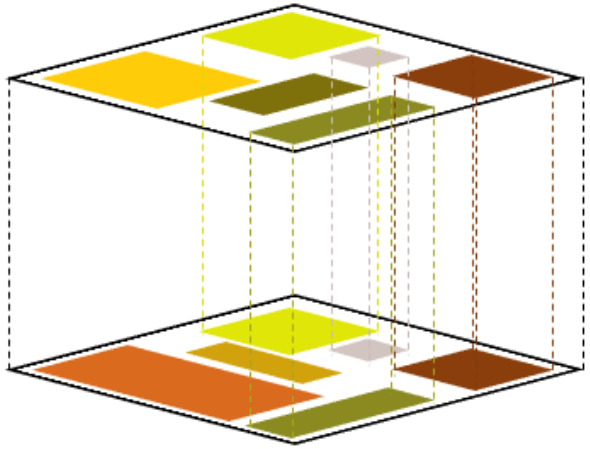
(a)



(b)

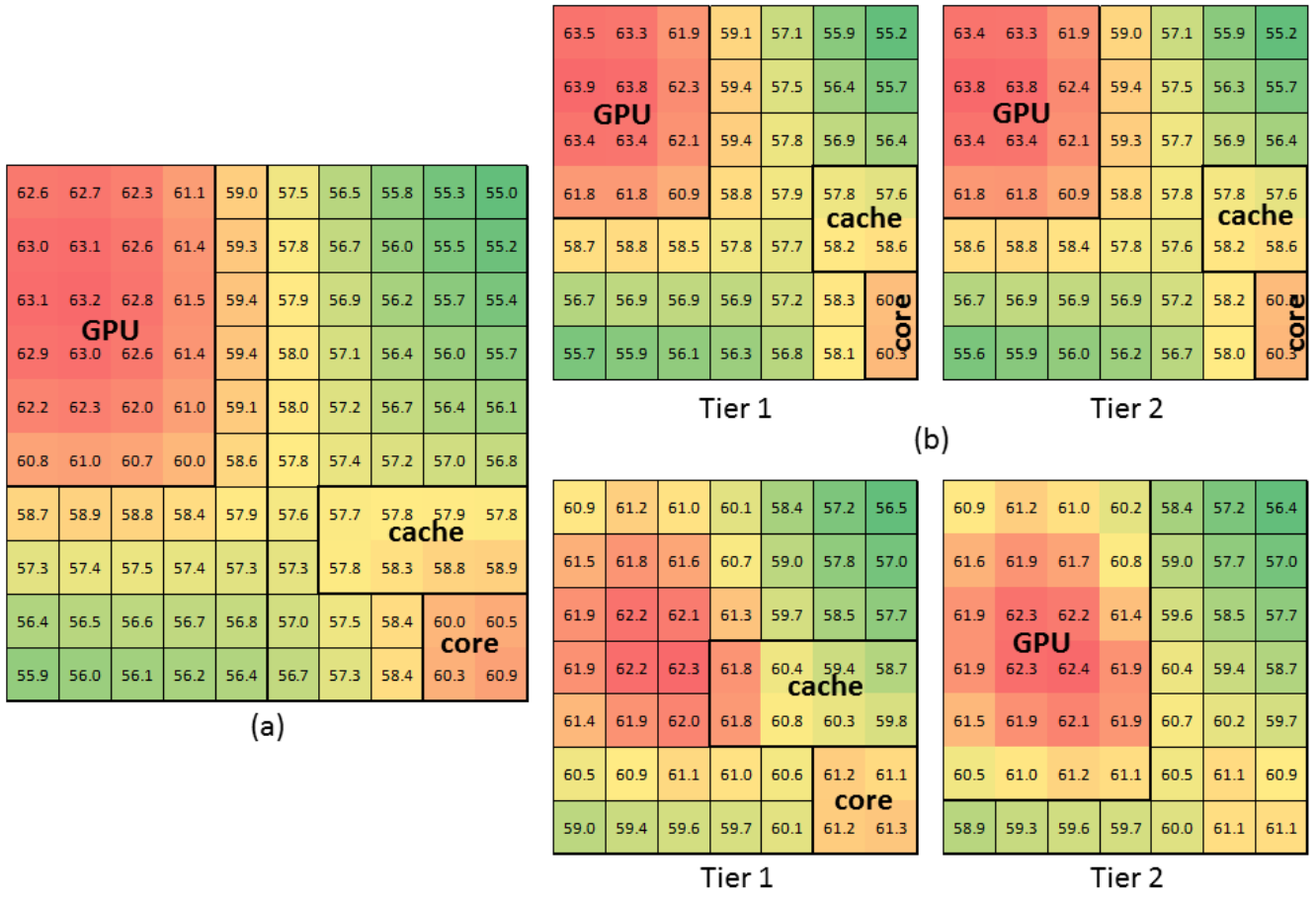


(c)



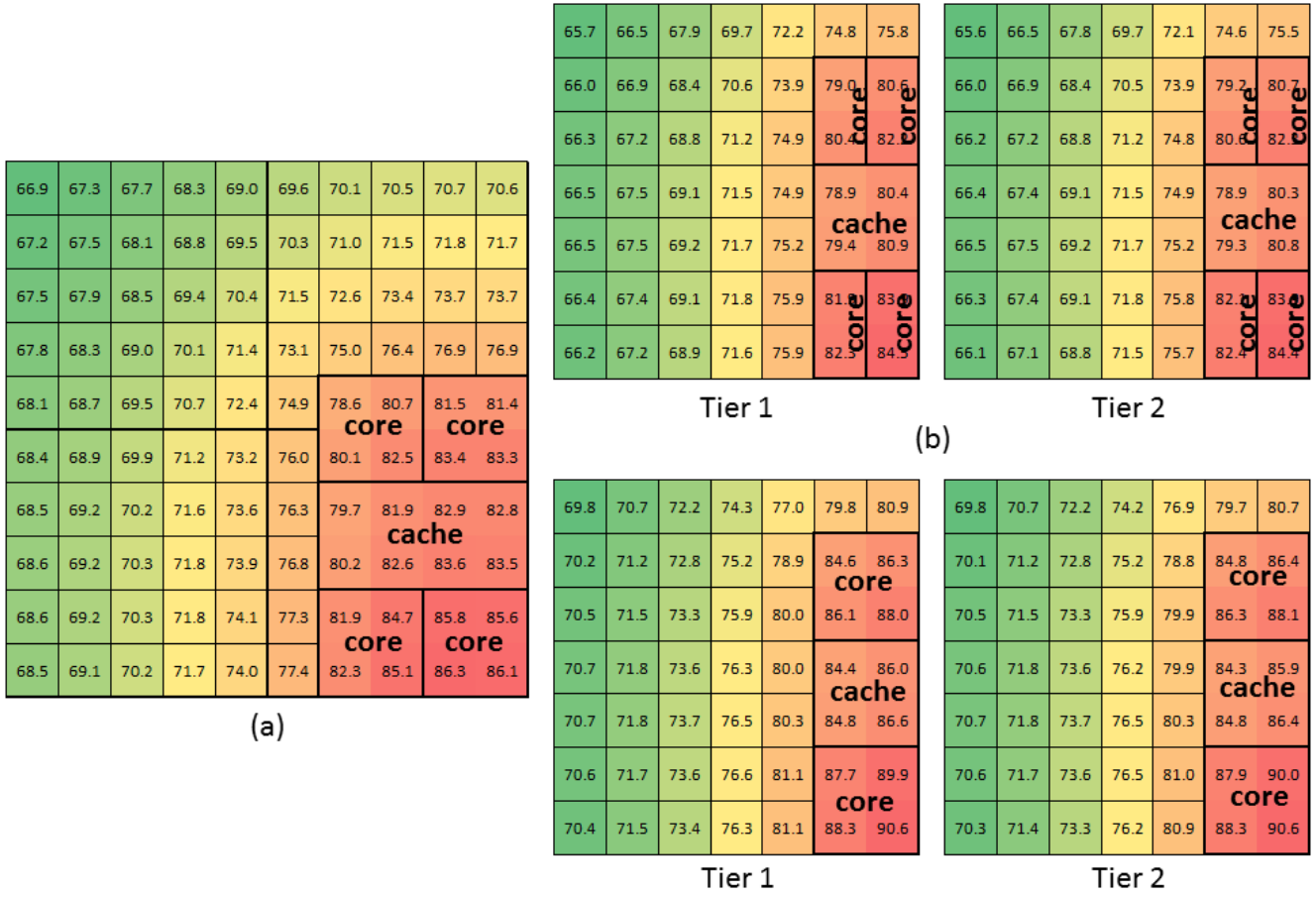
(d)

3D Floorplanning and Partitioning Case Study: GPU Power Intensive Case



Thermal maps of mobile MPSoC GPU intensive use-case for (a) 2D, (b) All 3D and (c) All 2D configurations

3D Floorplanning and Partitioning Case Study: CPU Power Intensive Case



Thermal maps of mobile MPSoC CPU intensive use-case for (a) 2D, (b) All 3D, and (c) All 2D configurations

Summary of 3D Floorplanning and Partitioning Case Study

- 3D temperatures with appropriate partitioning / floorplanning are comparable (or even) better than 2D

Scenario	Configuration	Partitioning	TIER-1 Tj [C]	TIER-2 Tj [C]	Leakage TIER-1 [W]	Leakage TIER-2 [W]	Total power (W)
GPU Intensive	2D		63.2		0.49		2.07
	3D (Baseline)	All 2D	64.9	65	0.21	0.32	2.11
	3D	All 2D	**62.3**	**62.4**	0.19	0.28	1.971
	3D (Baseline)	All 3D	68.9	68.9	0.28	0.28	2.14
	3D	All 3D	63.9	63.8	0.24	0.24	1.901
	3D (Baseline)	Hybrid	68.9	68.9	0.38	0.17	2.13
	3D	Hybrid	64.3	64.3	0.33	0.15	1.921
CPU Intensive	2D	N/A	86.3		0.85		3.01
	3D (Baseline)	All 2D	98.3	98.3	0.57	0.57	3.3
	3D	All 2D	90.6	90.6	0.45	0.45	2.952
	3D (Baseline)	All 3D	98.3	98.3	0.57	0.57	3.3
	3D	All 3D	**84.5**	**84.4**	0.39	0.39	2.68

Conclusions

- Practical and cost efficient 3DVLSI technologies are emerging
- A new generation of implementation tools are required to take full advantage of 3DVLSI technology
  - Floorplanner
  - Place & Route
  - Extraction
  - Timing
  - CTS
- New design methodologies are required
  - New Architectures
  - New foundation IP structures

Courtesy Saeidi et al., 3DIC'14

Qualcomm Research

Thank you

QUALCOMM®

www.qualcomm.com/research