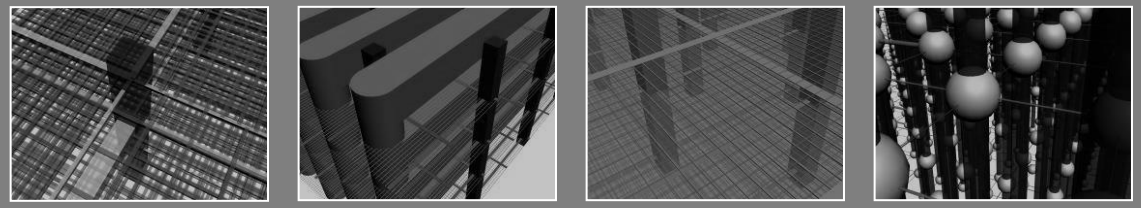
Thermal Impact Study of Block Folding and Face-to-Face Bonding in 3D IC



Yarui Peng[1], Moongon Jung[2], Taigon Song[1], Yang Wan[3], and Sung Kyu Lim[1]

[1]School of ECE, Georgia Institute of Technology, Atlanta, GA, U.S.A.

[2]Intel Corp., Santa Clara, CA, USA. [3]Google Inc., Mountainview, CA, USA.

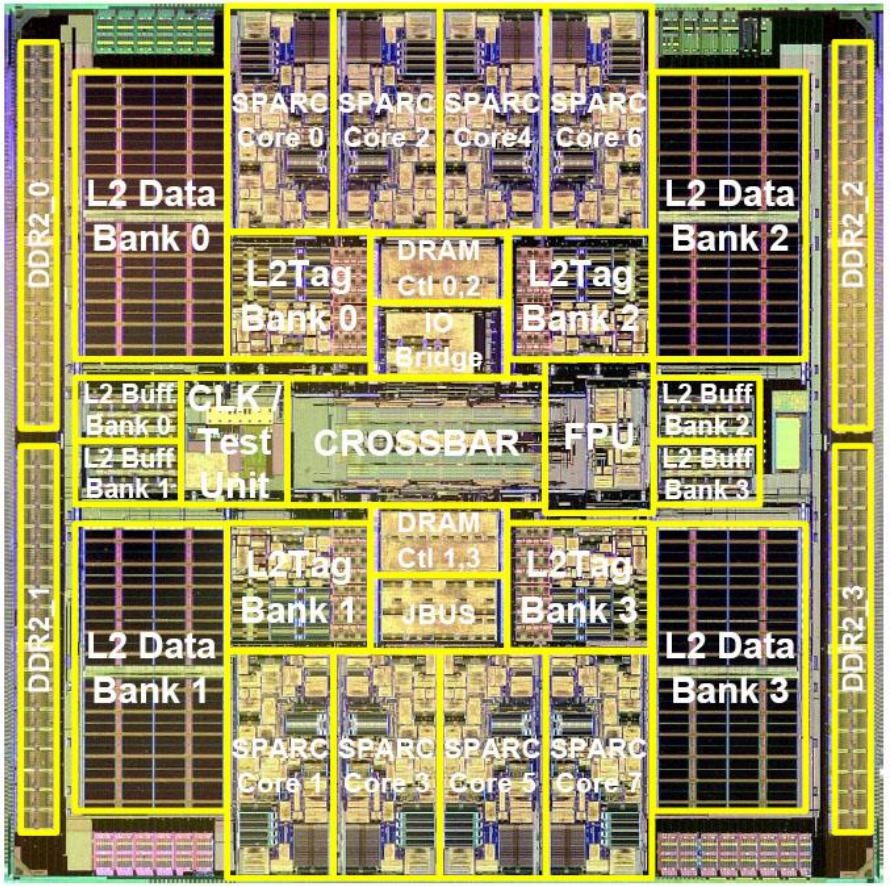
Contents

Motivation

- **Existing 3D IC design studies**
  - **Block folding *[M. Jung et al., CICC13']***
  - **F2F-bonding-based design *[D. Kim et al., TC'14] [M. Jung et al., DAC14']***
- **Little in-depth study on thermal impact**
  - **Most previous works focused on wire length reduction, power reduction, and performance improvement**
  - **Full-chip thermal impacts were not studied in detail**
- ***What are the thermal impacts in 3D IC design?***
  - ***Block folding and F2F bonding***

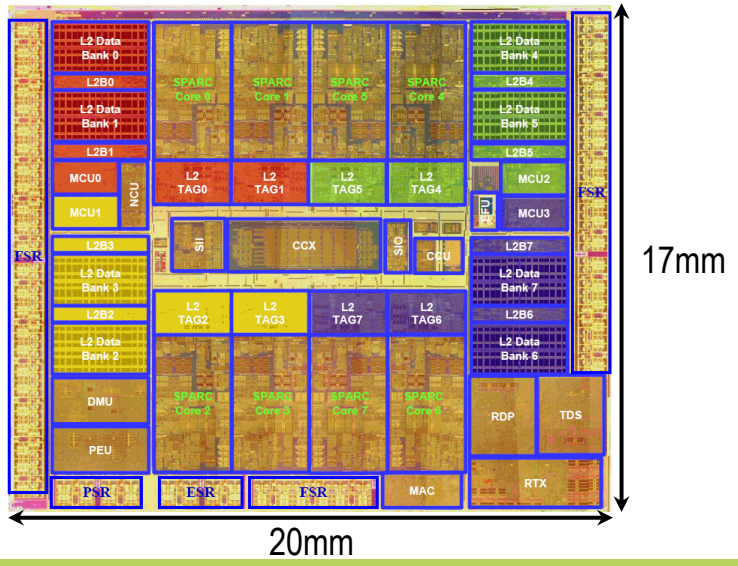
Benchmark Design

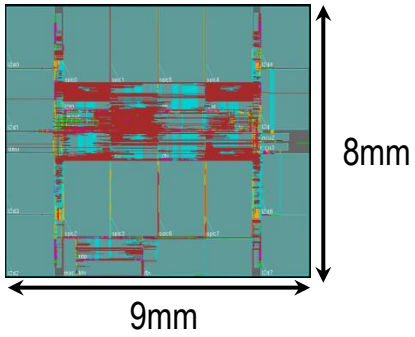
- OpenSPARC T2
  - Commercialized in 2007
    - UltraSPARC T2
  - Manufactured in 65nm @ 1.4GHz
  - RTL in public: 8 cores, 64-threads
  - Big (& meaningful) core: 500M TR
  - Digital & memory designed (~90%)
  - I/O and analog removed



Comparison with Commercial T2

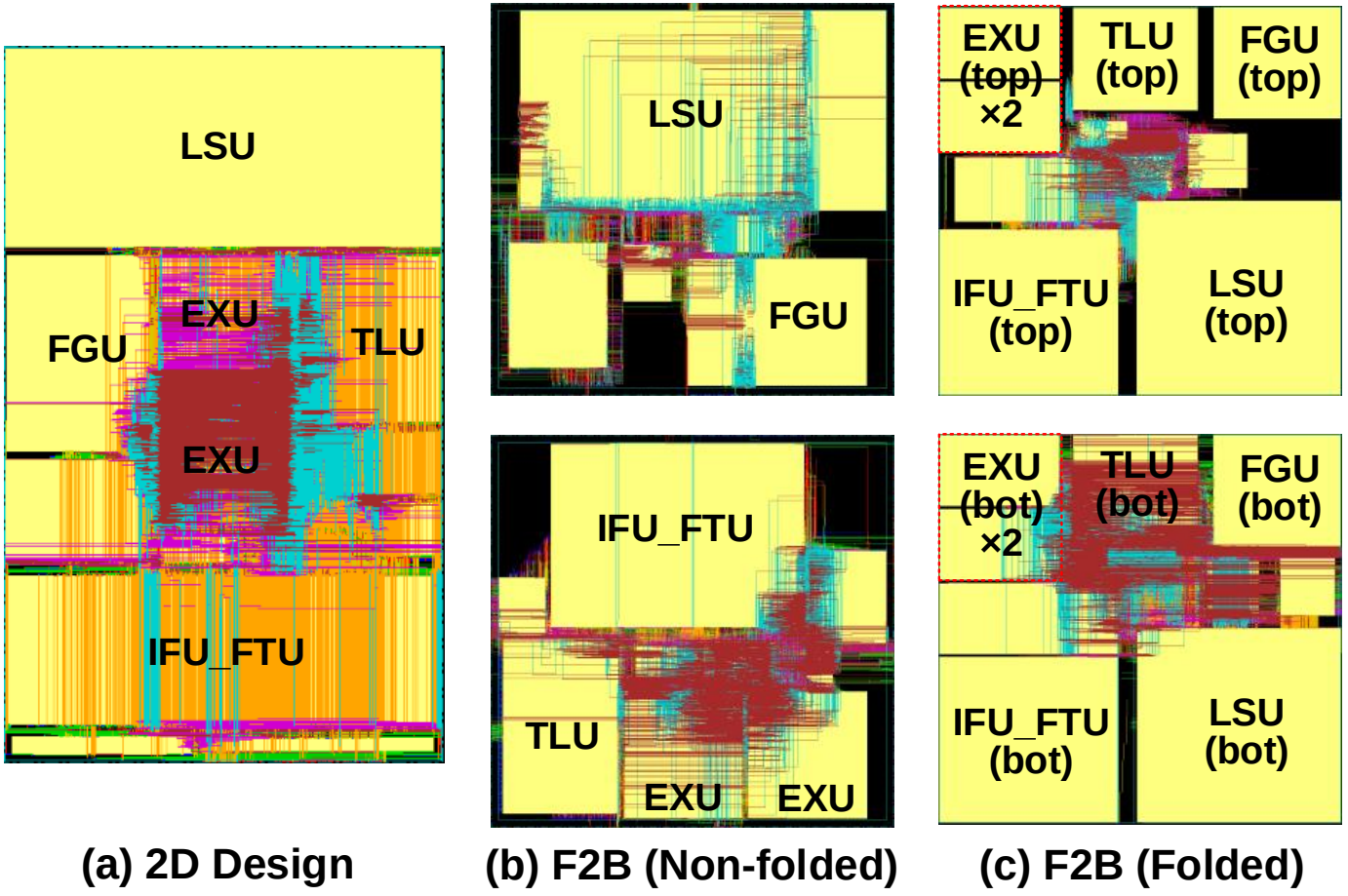
	UltraSPARC T2	GT OpenSPARC T2
Technology	65nm	28nm
Area (mm^2)	342 (20 x 17mm)	72 (9 x 8mm)
Complexity	503M Transistors	7.5M Logic Std. Cells + 5548 Mem. Macros
Power	95 W (123W max)	8.96 W
Max. clock freq.	1.4GHz	0.5MHz





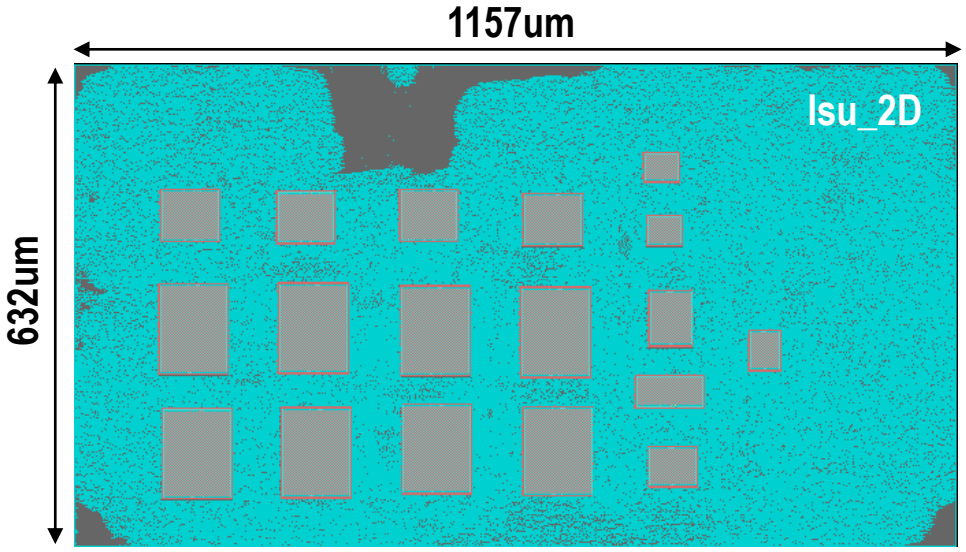
T2 Core Layouts: 2D vs. 3D

- **We designed our T2 core in two ways:**
  - **Non-folding and block-folding**

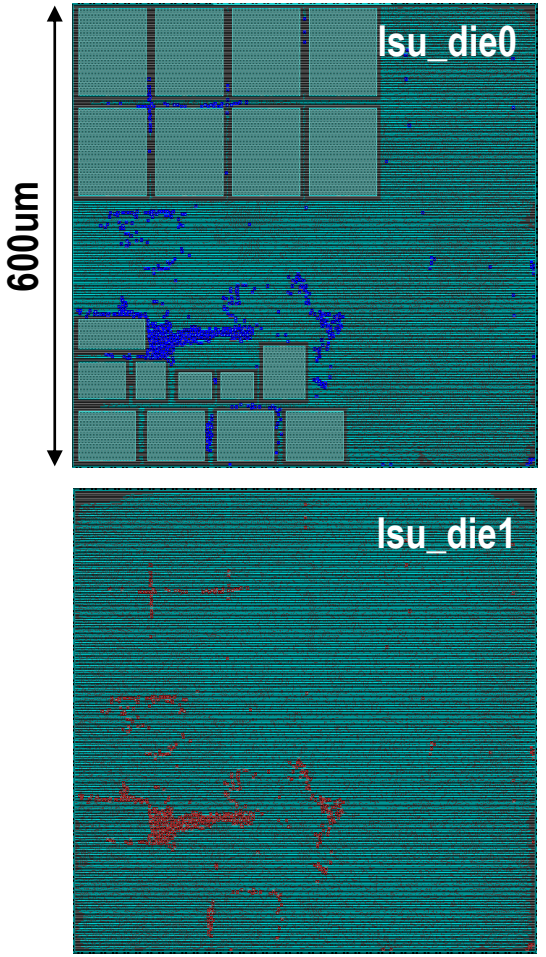


(a) 2D Design (b) F2B (Non-folded) (c) F2B (Folded)

What is Block-Folding?



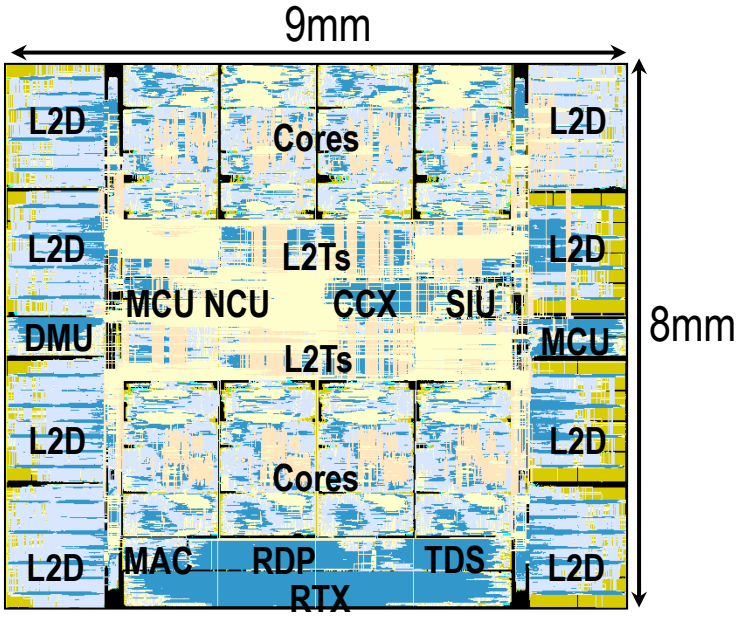
(a) 2D LSU



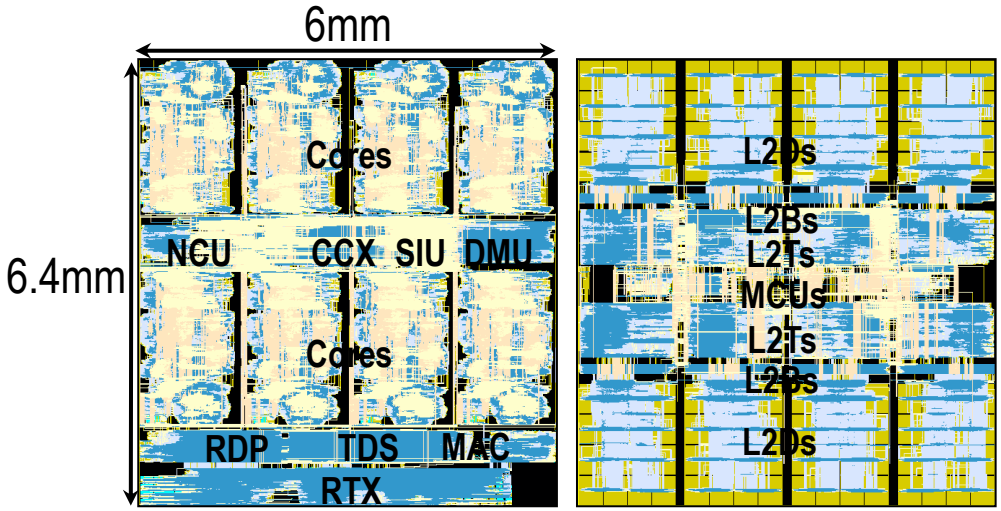
(b) 3D LSU

- **Block-folding: Designing a single module (block) in two dies**
  - **Power reduction INSIDE modules**

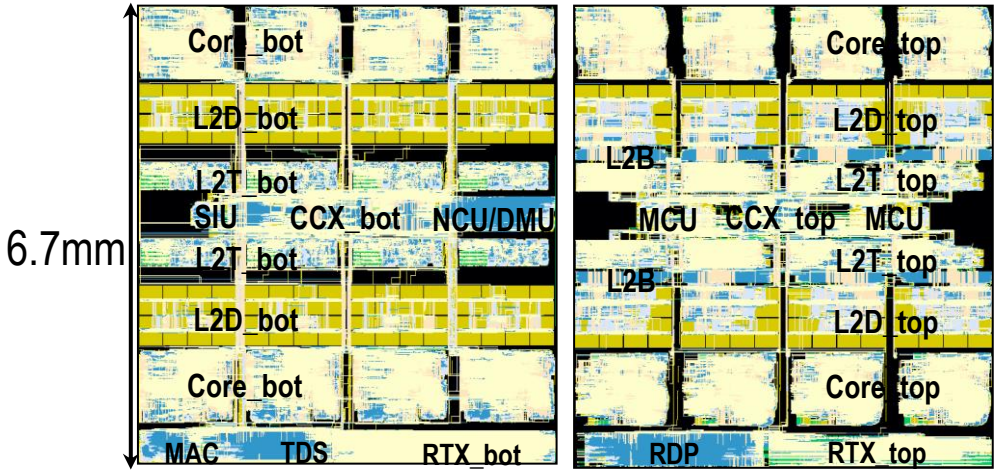
T2 Full-chip Layouts: 2D vs. 3D



(a) 2D



(b) 3D w/o folded blocks (#TSV: 3,263)



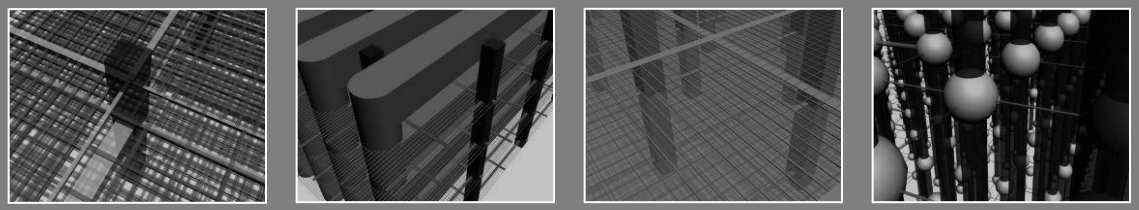
(c) 3D w/ folded blocks (#F2F: 101,555)

Block-Folding: Impact in Full-chip Design

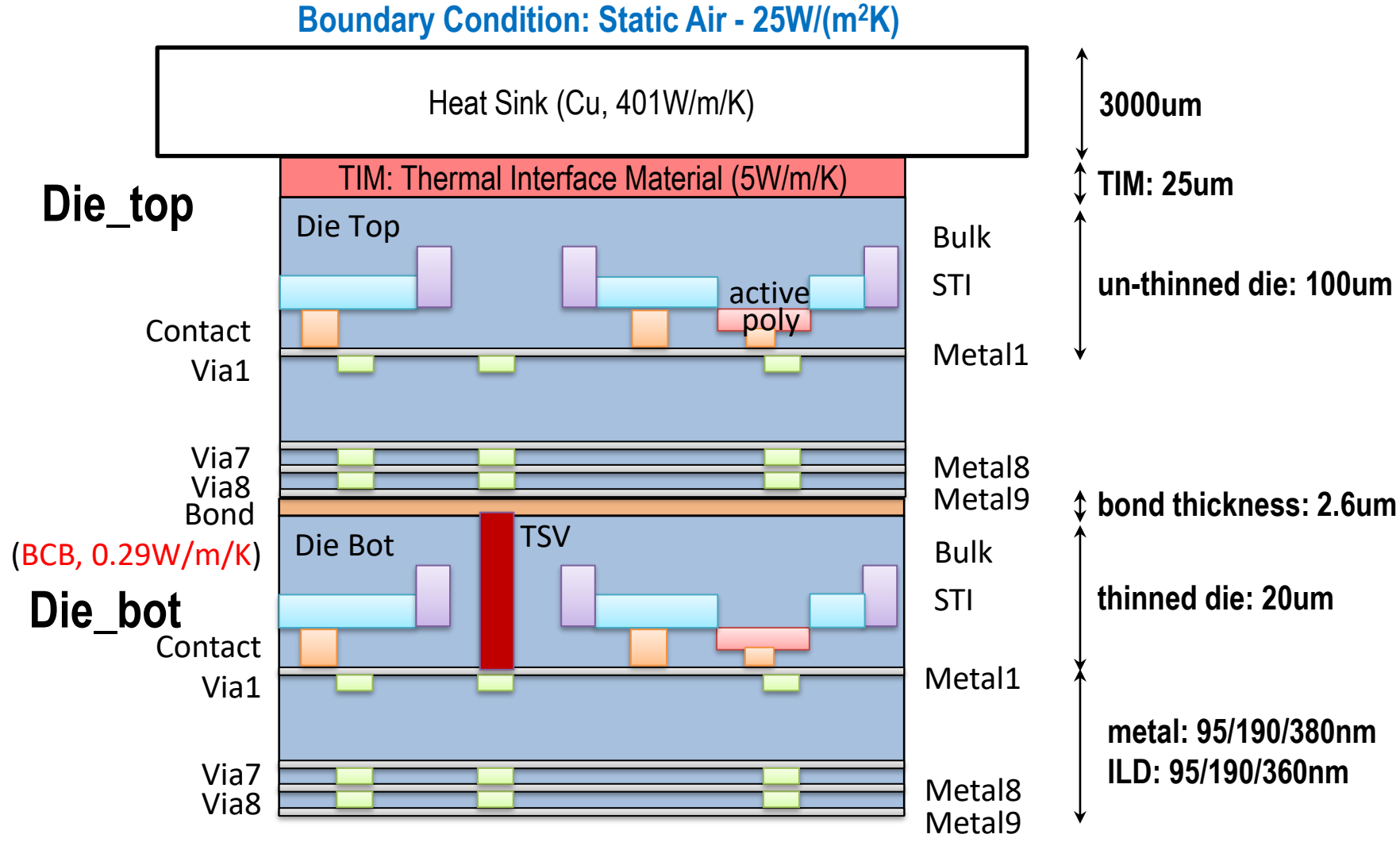
T2	2D	3D w/o folding	diff	3D w/ folding	diff
Footprint (mm^2)	71.1	38.4	-46.0%	40.8	-42.6%
WL (m)	339.7	320.3	-5.7%	306.9	-9.7%
# Cells	7.41M	7.02M	-5.3%	6.72M	-9.3%
# Buffers	2.89M	2.37M	-18.0%	1.97M	-31.8%
# HVT cells	6.40M (86.3%)	6.38M (90.0%)		6.58M (95.1%)	
# TSV	-	3,263		69,091	
WNS (ns) @2ns clk	-0.050	-0.040		+0.022	
gcc power (W)	20.5	17.4	-15.1%	16.2	-21.0%
spice power (W)	21.3	18.1	-15.0%	16.8	-21.1%

- Significant power saving in 3D with block-folding
- 3D uses more HVT cells than 2D, thanks to better timing
  - Dual-Vt design: RVT (regular Vt) & HVT (high Vt)

Thermal Impacts of Block-Folding

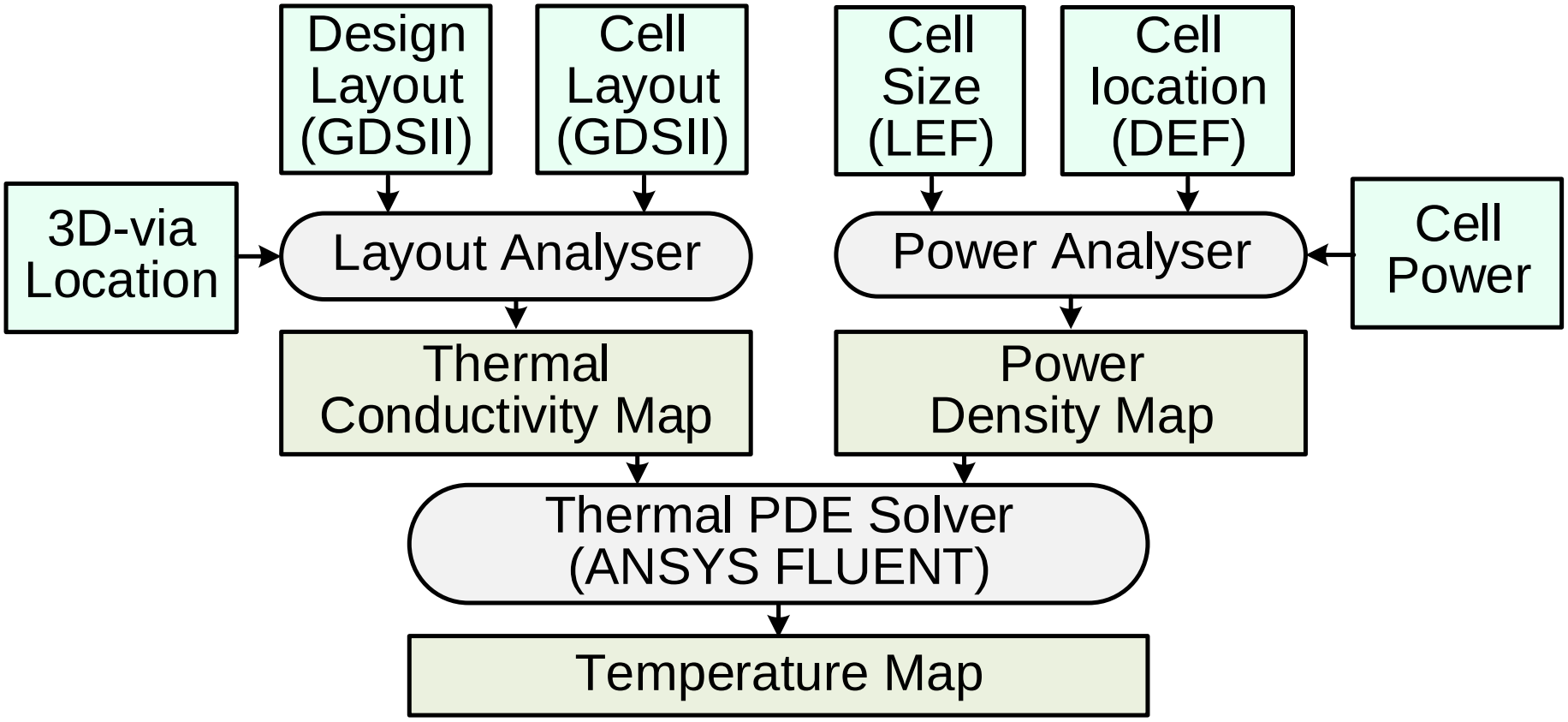


F2B Thermal Analysis Structure



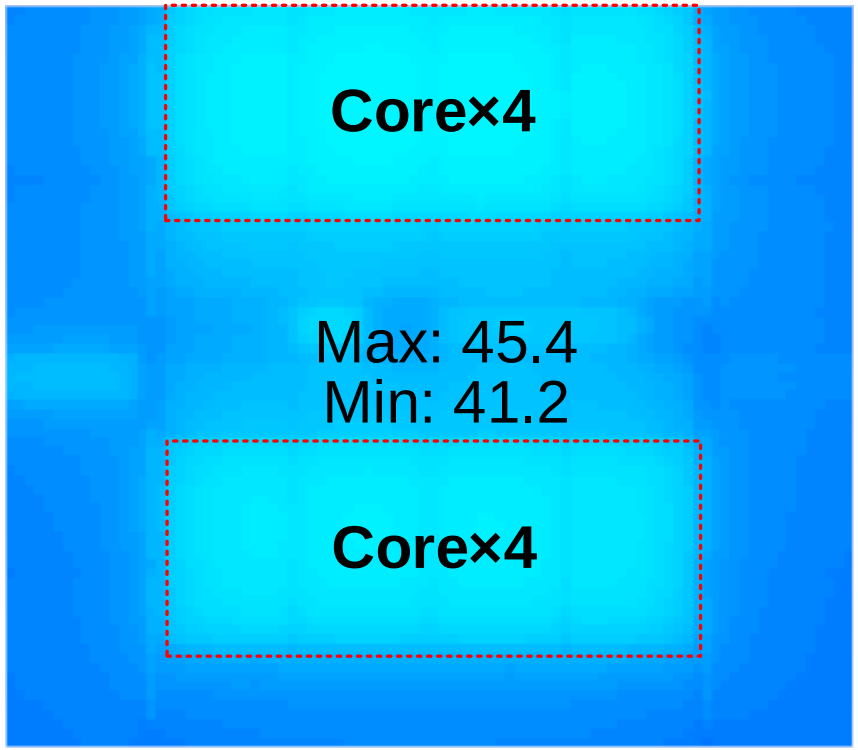
Our GDSII-Level Simulation Flow

- **Our simulation flow combines both commercial tools and in-house design tools to provide maximum accuracy.**

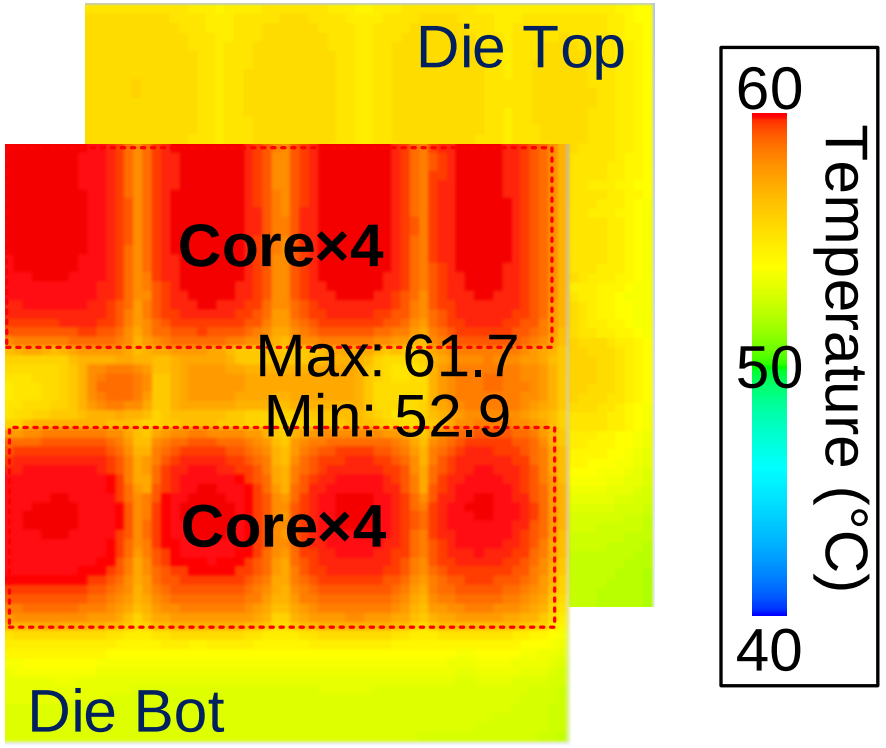


3D Thermal Challenges

- **3D design introduces significant thermal challenges due to increased power density on a smaller footprint**



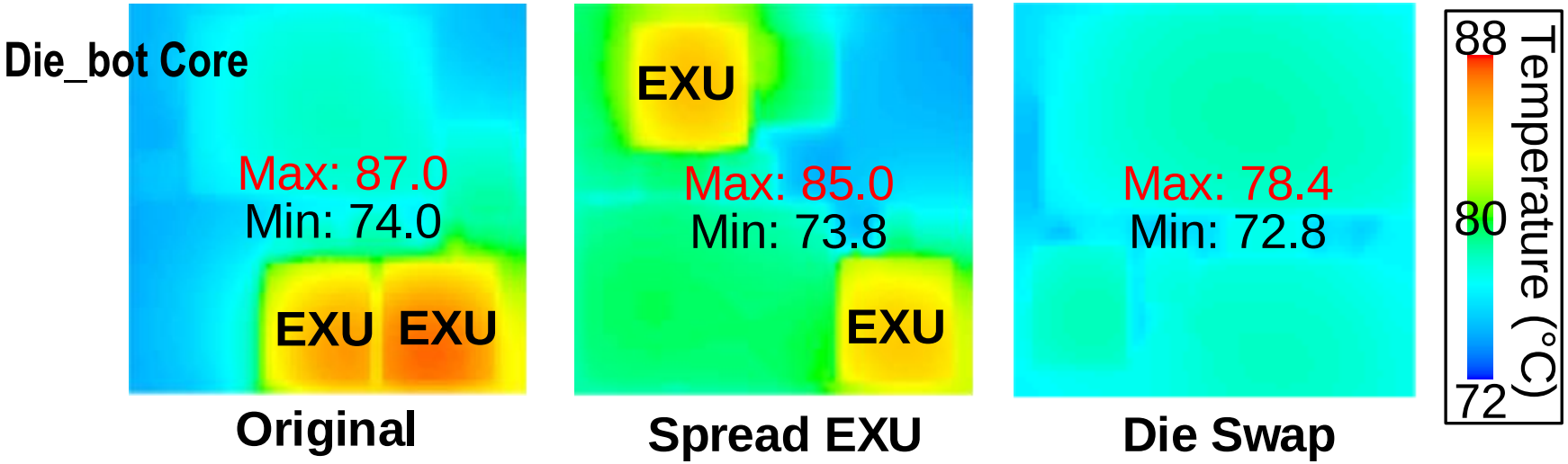
2D full-chip



3D full-chip (F2B, non-folded)

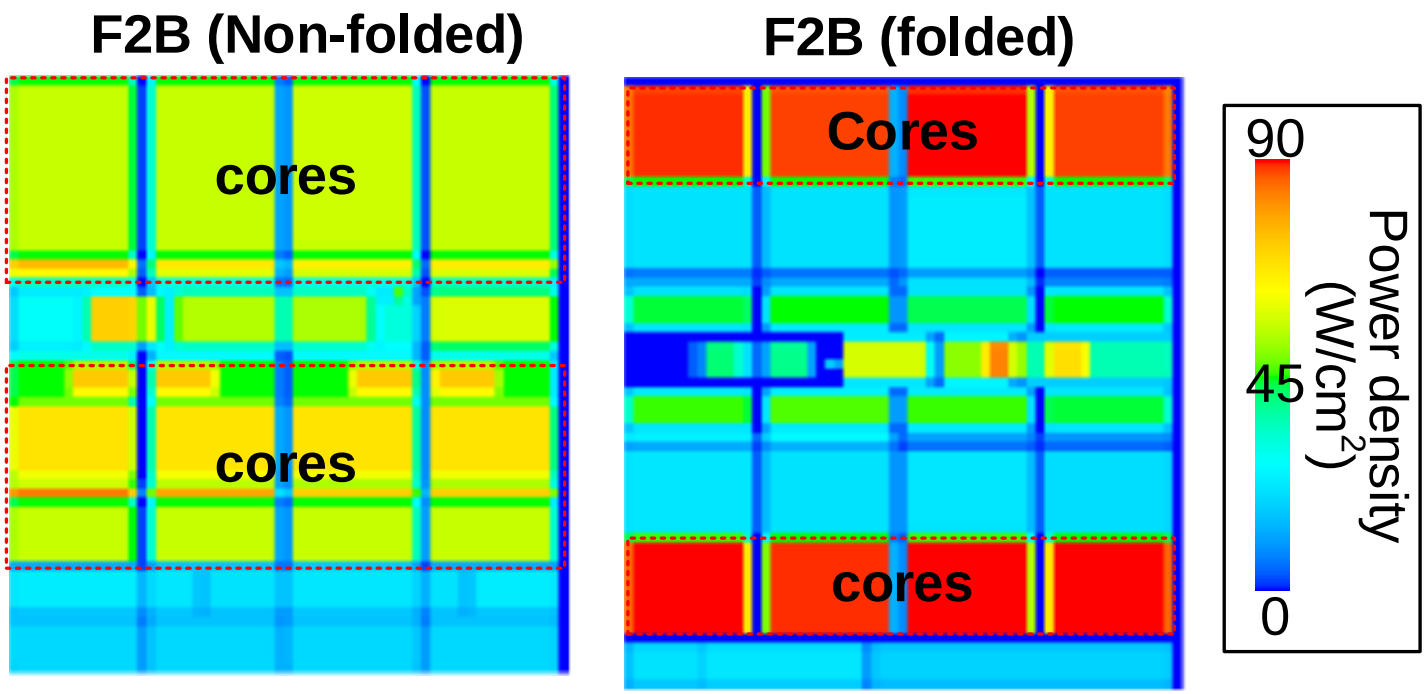
Hot Spot Redistribution Helps Thermal

- **Distributing hot spots help reducing maximum temperature**
  - **Spreading hot spots reduces thermal accumulation**
  - **High power density modules on the top die reduces vertical heat flow**

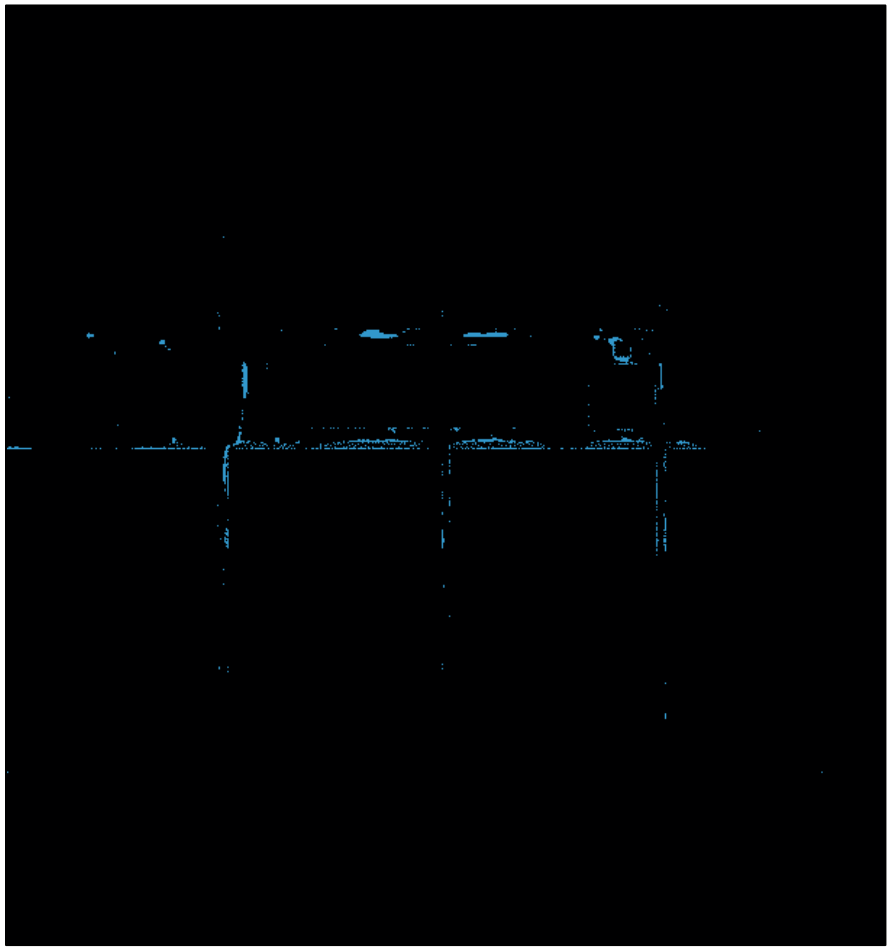


Thermal Perspectives in Block-Folding

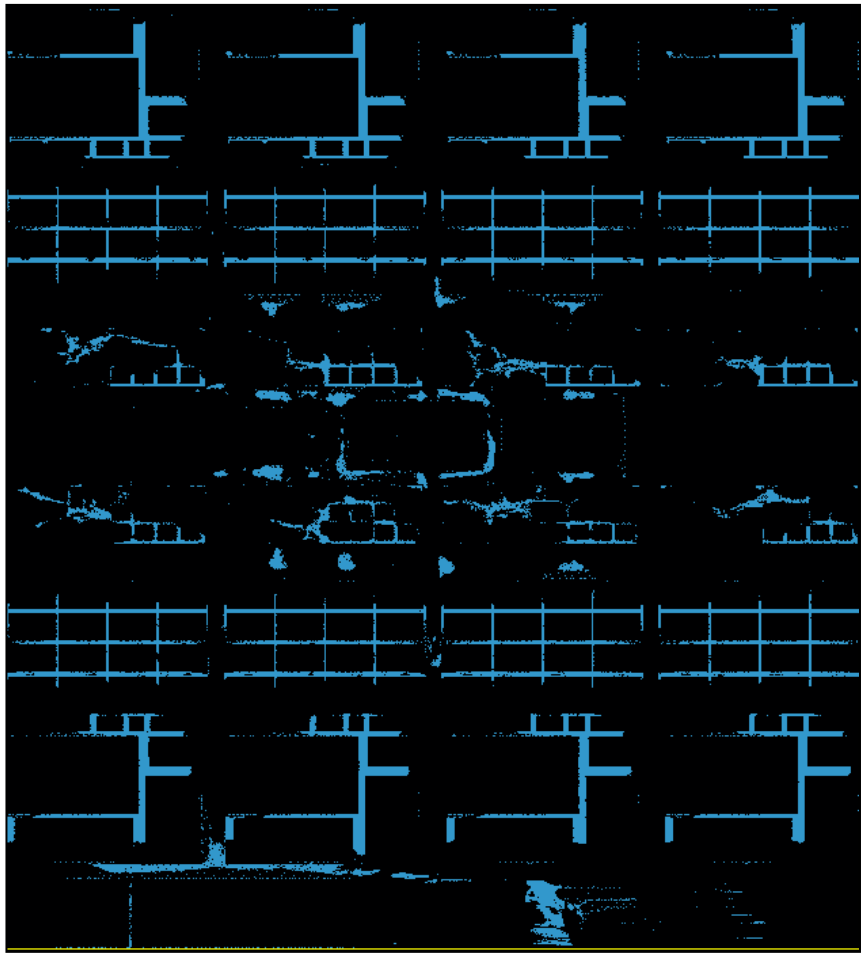
- **Two aspects of block-folding:**
  - **Folded-blocks increase power density**
  - **Folded-blocks move half of power onto the top die**



TSV Locations in 3D Designs



(a) TSV locations (3D w/o Folding, die bot)
#TSV: 3,263



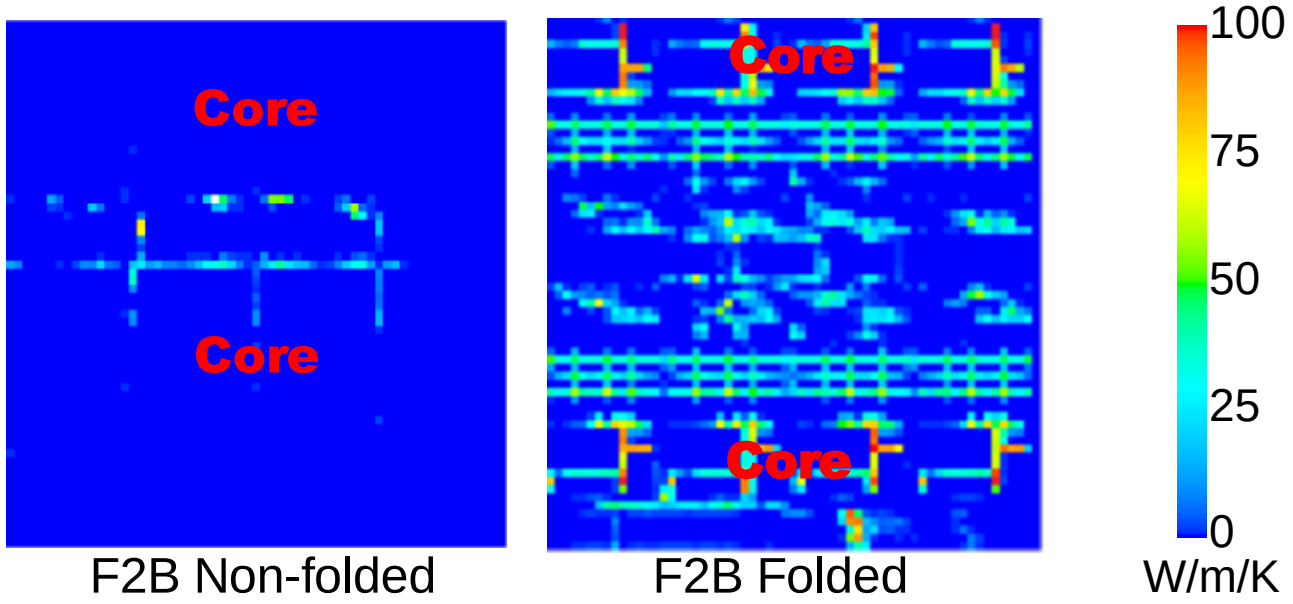
(b) TSV locations (3D w/ Folding, die bot)
#TSV: 69,091

Block Folding: Bonding Layer

- **With block folding, TSV count increases significantly. This helps the vertical heat flow and reduces the maximum temperature on the bottom die**

TSV count	Non folded (TSV)	Folded (TSV)
Core	2,979	9,551
Full chip	3,265	**69,151**

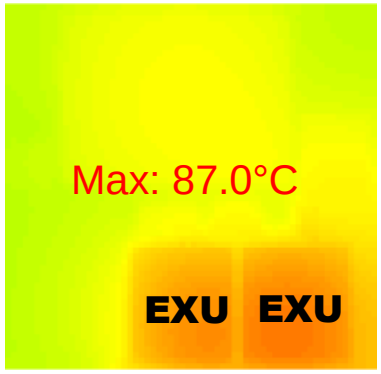
Die_bot full-chip



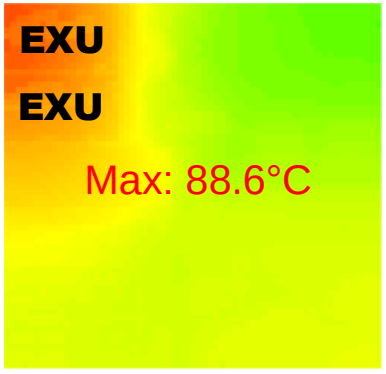
F2B Non-folded F2B Folded

Block-Folding: Overall Impact

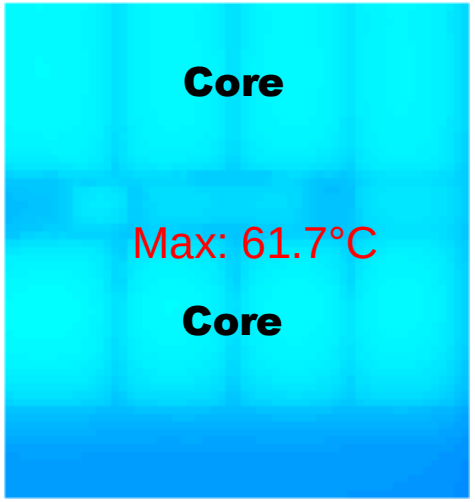
- **Block-folding does not worsen thermal results**
  - Power decreases with FUB folding
  - More TSVs help vertical heat dissipation
  - More blocks are moved onto the top die
  - But, unavoidable **hotspot overlapping** occurs
    - (folded EXU, highest power density)



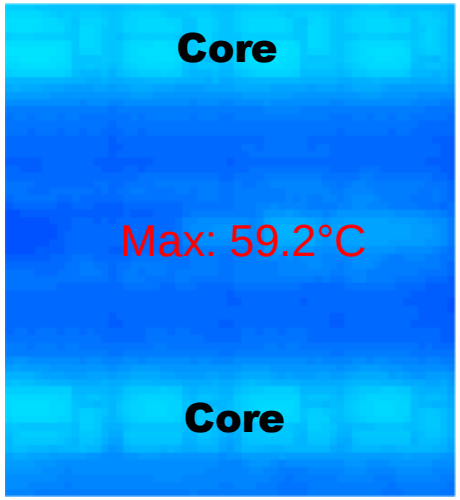
F2B non-folded core



F2B folded core



F2B non-folded full-chip



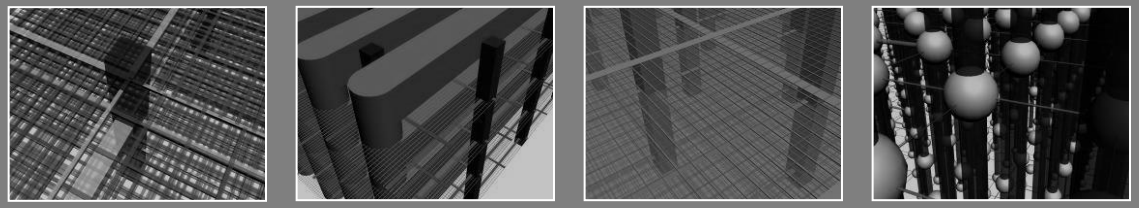
F2B folded full-chip

Block-Folding: Summary

- **Block-folding does not worsen thermal results**

Design	Benchmark	Folded?	Temperature (°C)	
			Bottom die	Top die
Core	gcc	No	74.0~87.0	75.9~78.5
		Yes	70.6~88.6	70.1~82.9
	spice	No	78.7~92.6	76.6~85.2
		Yes	79.1~91.7	77.8~86.6
Full-chip	gcc	No	53.6~61.7	52.9~57.6
		Yes	51.1~59.2	50.6~57.3
	spice	No	54.6~63.2	53.9~58.5
		Yes	52.1~60.7	51.6~58.6

Thermal Impact of F2F (Face-to-Face) Bonding

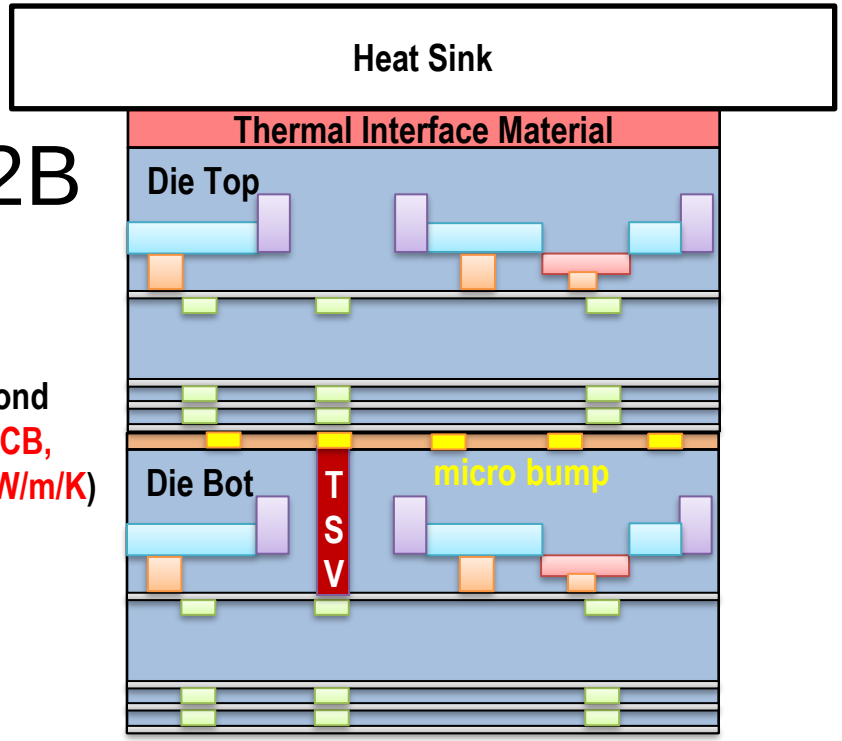


Thermal Analysis Structure: F2B vs. F2F

- **Our simulation covers both F2B and F2F bonding structures.**

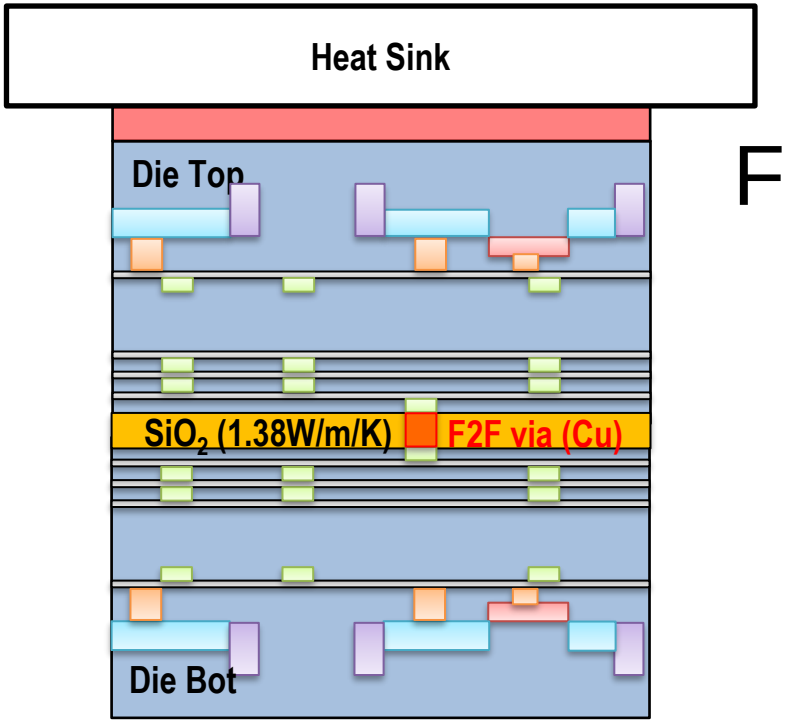
Boundary Condition: Static Air - 25W/(m^{2}K)

F2B



The bonding material (BCB, 2.6um thick)
is limiting the vertical heat flow
TSV is large (3um) → helps vertical heat flow

F2F



The bonding material (SiO_2, 1um thick)
has better thermal conductivity than BCB
F2F via is small (0.5um) →
smaller per-via impact than TSV

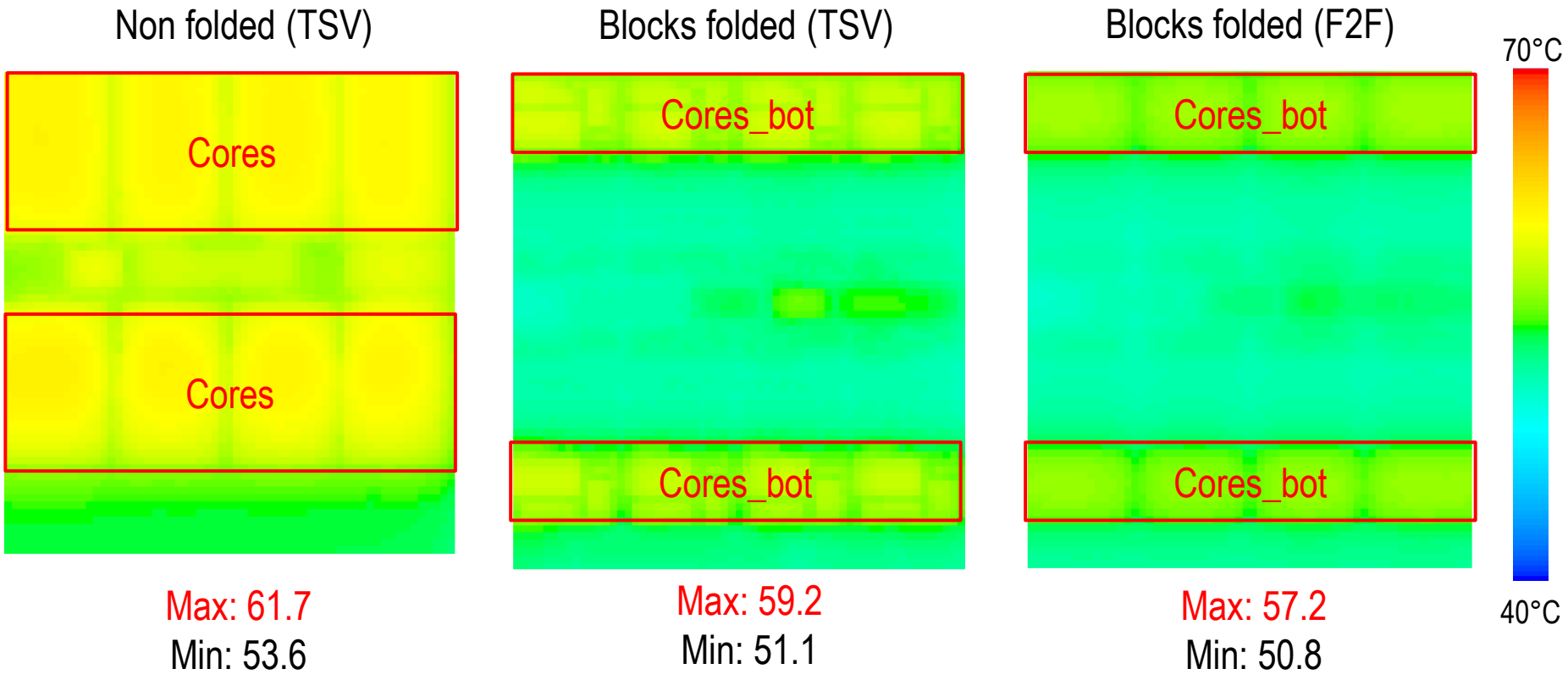
F2F Bonding: Design Comparisons

T2	2D	3D (F2B)	diff	3D (F2F)	diff
Footprint (mm^2)	71.1	40.8	-42.6%	40.8	-42.6%
WL (m)	339.7	306.9	-9.7%	302.7	-10.9%
# Cells	7.41M	6.72M	-9.3%	6.59M	-11.1%
# Buffers	2.89M	1.97M	-31.8%	1.85M	-36.0%
# TSV/F2F	-	69,091		101,555	
gcc power (W)	20.5	16.2	-21.0%	15.9	-22.4%
spice power (W)	21.3	16.8	-21.1%	16.6	-22.1%

- **F2F saves more power than F2B**
  - F2F vias are much smaller and do not occupy silicon area
  - Block quality is better in spite of inter-block routing issue

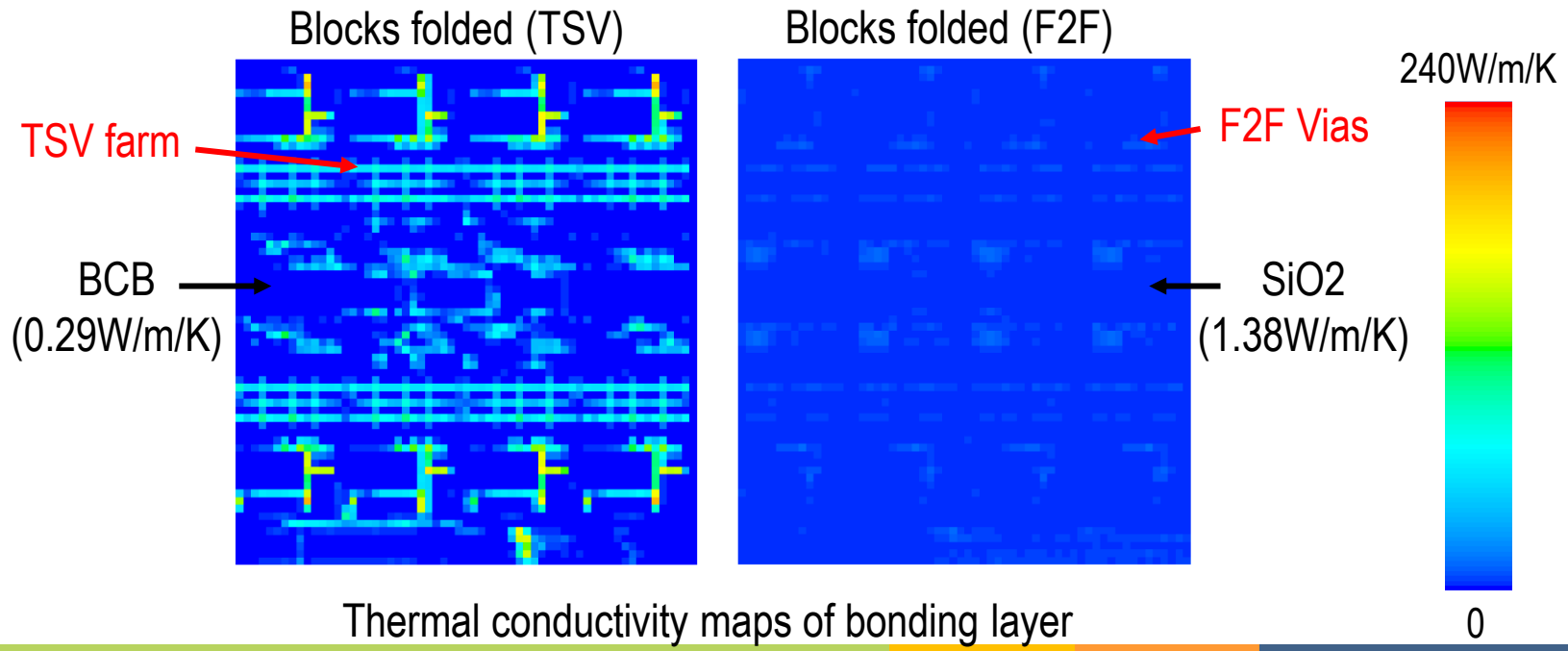
F2F Helps Thermal in Full Chip

- **F2F bonding shows the lowest max temperature**
  - **Reduced power consumption**
  - **Increased background thermal conductivity**



Bonding Layer Comparison (6 FUBs Folded)

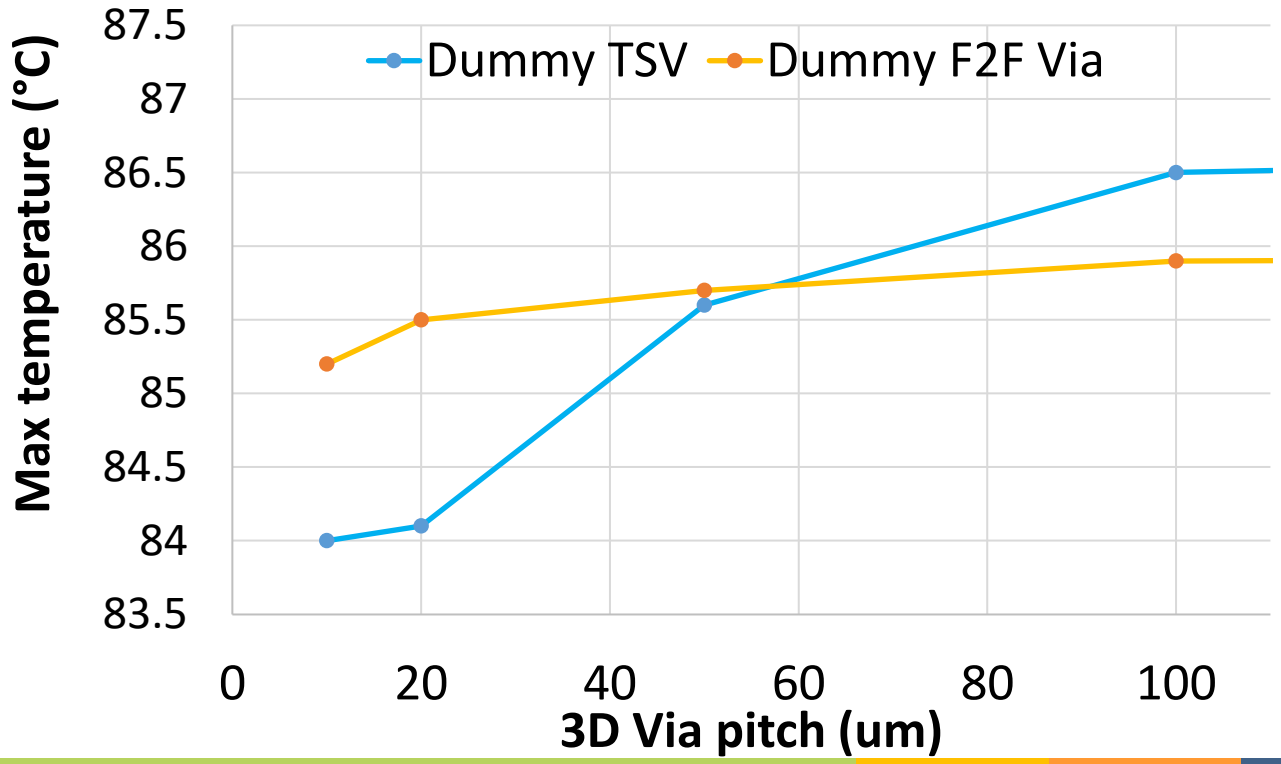
- **TSVs (3um) help thermal more compared with F2F vias (0.5um) in similar via count**
- **But, background thermal conductivity is different**
  - **TSV structure (0.29W/m/K) vs. F2F structure (1.38W/m/k)**



Thermal conductivity maps of bonding layer

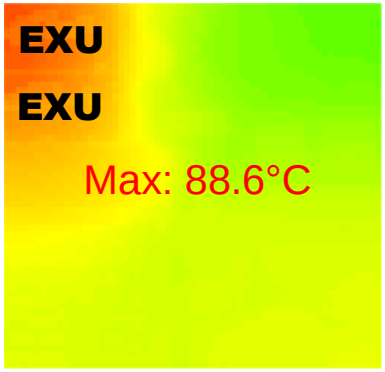
F2F Via Impact

- **Unlike TSVs, additional F2F vias show much smaller impact than TSVs and µbumps**
  - **F2F vias are much smaller than TSVs**
  - **F2F bonding layer is not a limiting factor for vertical heat flow**

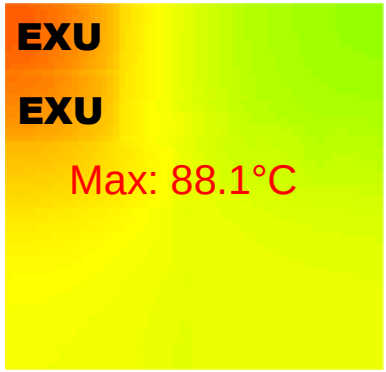


F2F Bonding: Thermal Maps

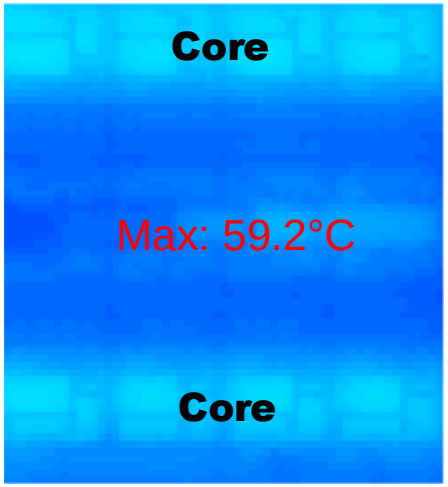
- **Though F2F bonding helps reducing overall power consumption, its impact on thermal is small**
  - **The F2F bonding layer has a much better background thermal conductivity**
  - **But small-sized F2F via reduces the thermal benefits (Compared to TSVs)**



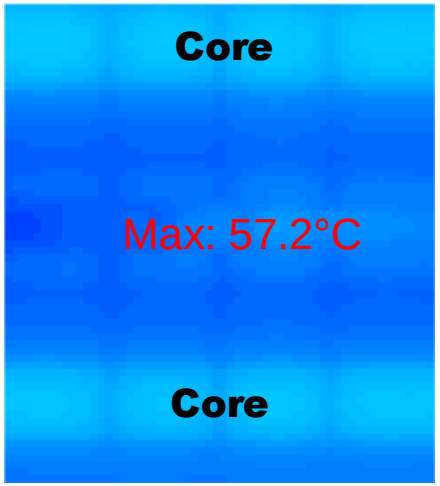
F2B folded core



F2F folded core



F2B folded full-chip



F2F folded full-chip

F2F Bonding: Summary

- **Despite its power benefits, F2F does not help thermal significantly**

Design	Benchmark	Bonding	Temperature (°C)	
			Bottom die	Top die
Core	gcc	F2B	70.6~88.6	70.1~82.9
		F2F	72.5~88.1	71.7~83.1
	spice	F2B	79.1~91.7	77.8~86.6
		F2F	80.9~91.5	79.3~86.9
Full-chip	gcc	F2B	51.1~59.2	50.6~57.3
		F2F	50.8~57.2	50.6~56.5
	spice	F2B	52.1~60.7	51.6~58.6
		F2F	51.8~58.6	51.6~57.8

Summary

- Separating hot spots and moving high power blocks closer to the heat sink helps
- BCB bonding layer blocks vertical heat flow in F2B
  - Additional TSVs and µbumps help
- Block folding does not worsen thermal
  - Low power design + TSVs as heat conductors
- F2F bonding does not improve temperature too much
  - Small F2F via sizes

Thank You

- **yarui.peng@gatech.edu**