Power, Performance and Area Improvements with Monolithic 3D-IC for Equivalent Moore's Law Scaling

ARM

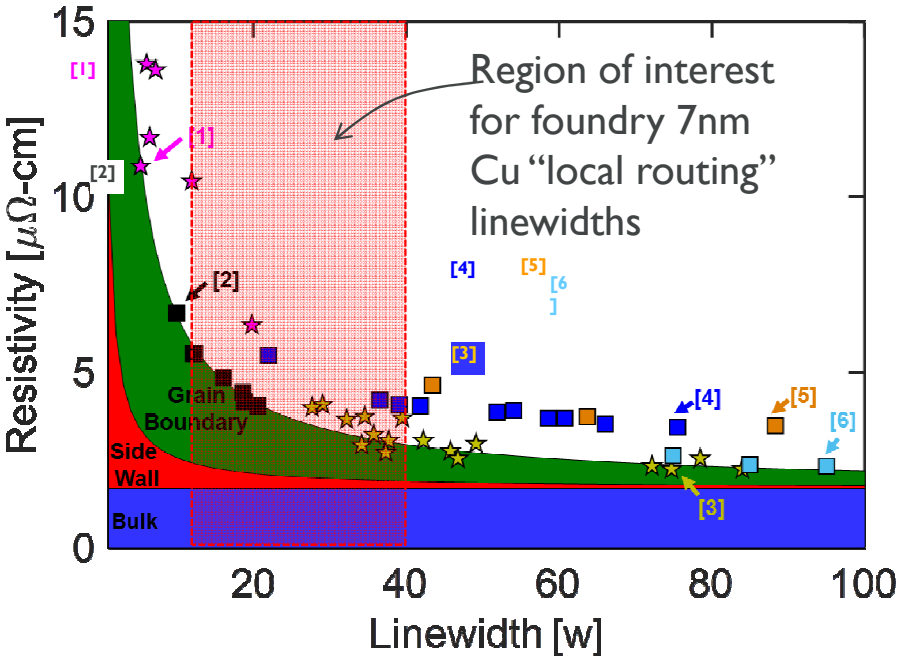
Saurabh Sinha, Brian Cline and Shidhartha Das

ARM Research

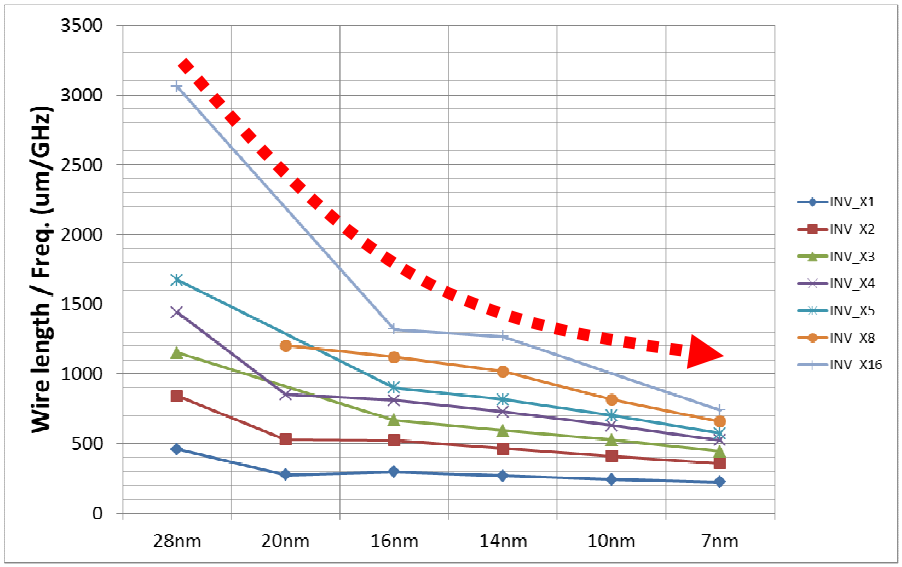
CEA-LETI
June 27th, 2017

The Wire-Parasitics Problem

Copper Resistivity



Buffer-insertion with Technology scaling



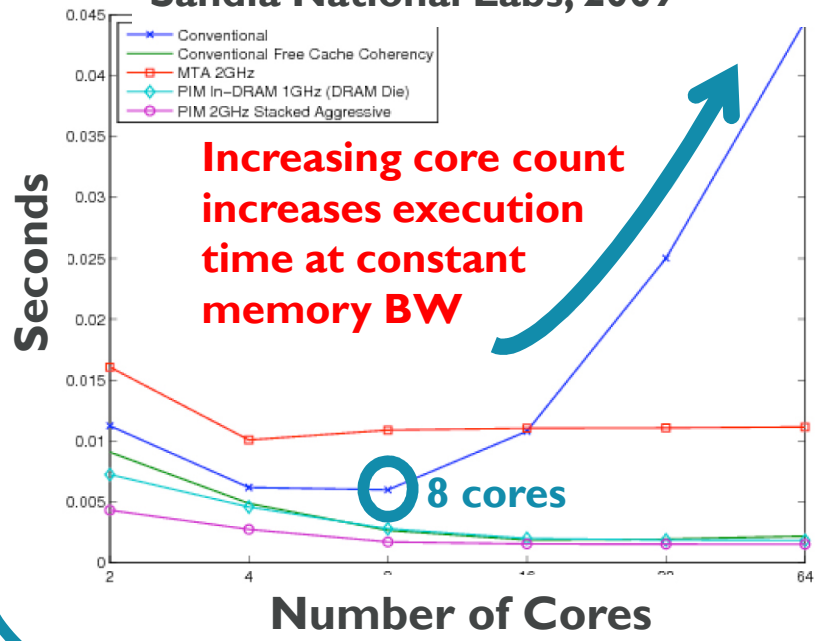
Copper resistivity increases exponentially with reduced line-widths

- No technology scaling relief @ 5nm or 3nm → ~10 years (w. r. t. new devices or materials)

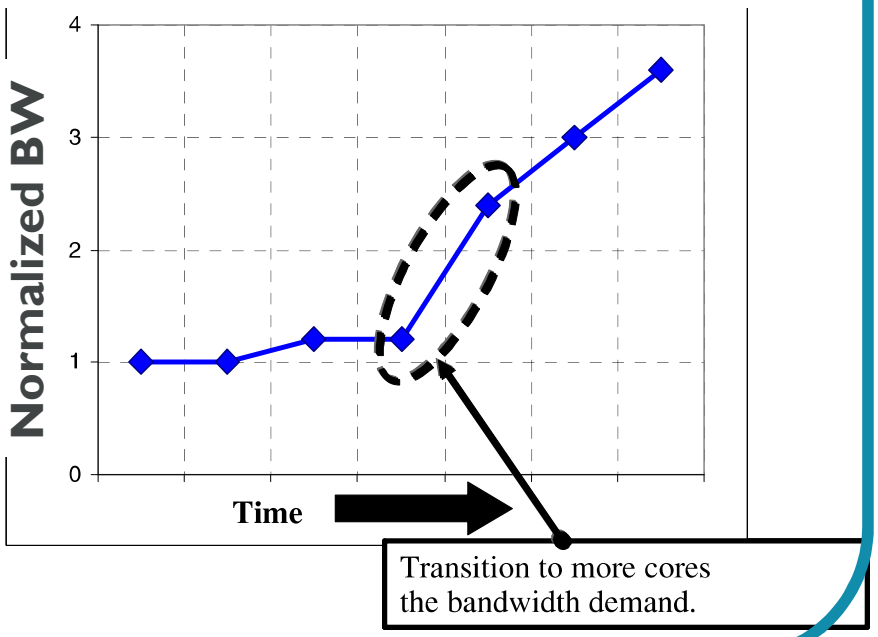
ARM

The Memory "Wall" Problem

Multicore meets Memory Wall - Sandia National Labs, 2009



Multicore clusters increase memory BW demand – Intel Technology Journal, 2007

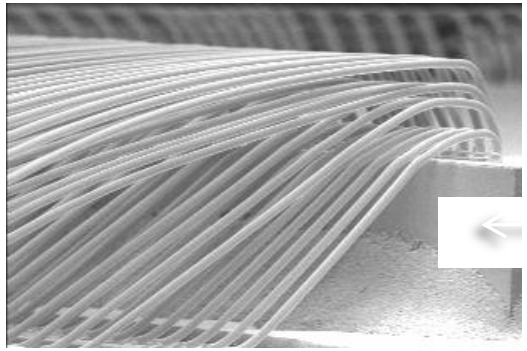


Multicore clusters exacerbate the "memory wall" problem

ARM

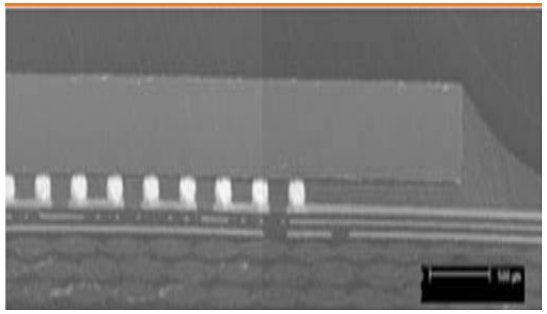
Package Interconnect Evolution

Wire Bond Interconnect
Typically 250 – 500 per chip



IC Packaging
Circa 2000

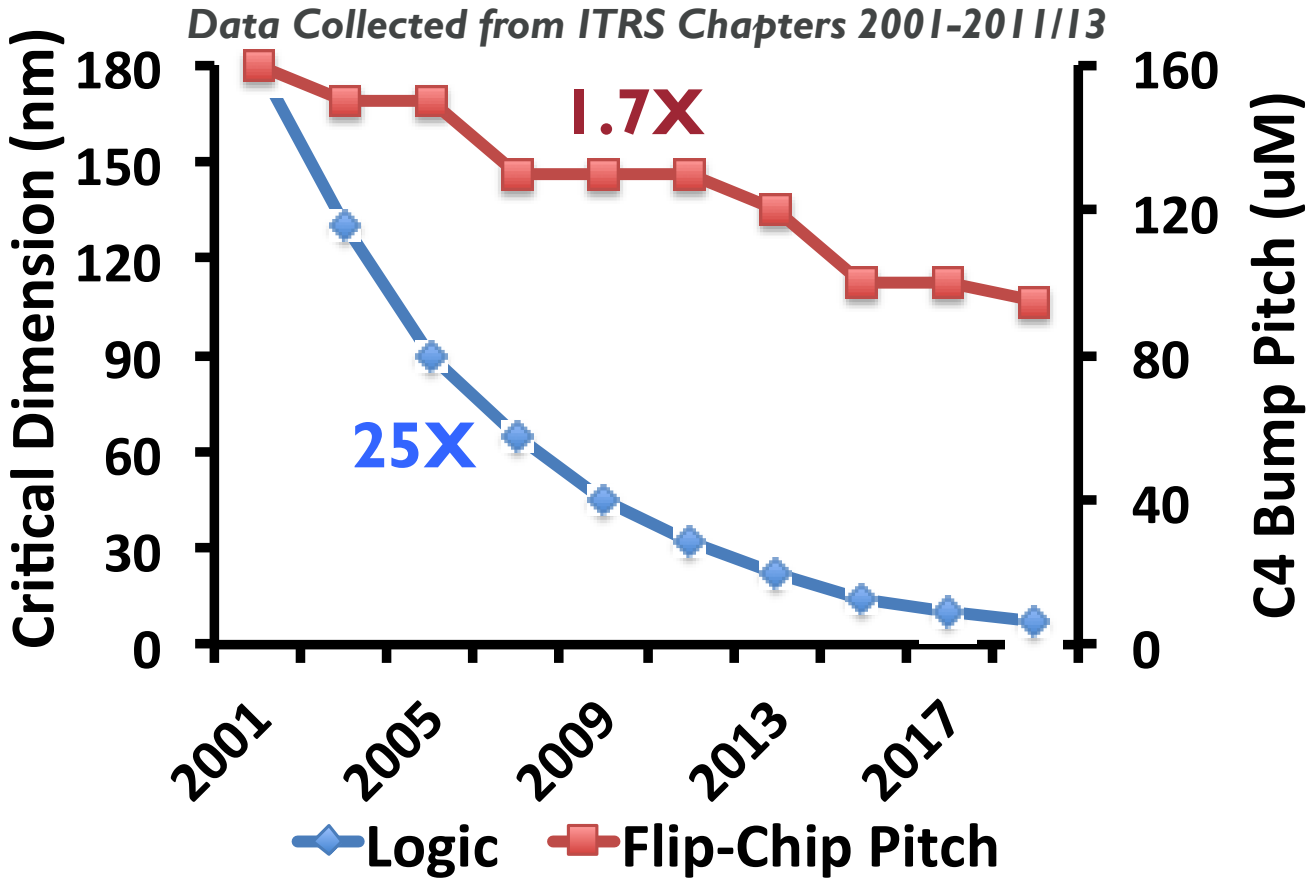
Flip-Chip Interconnect
Typically 10X higher per chip



IC Packaging
Circa 2010

ARM

Flip-Chip Pitch Trends

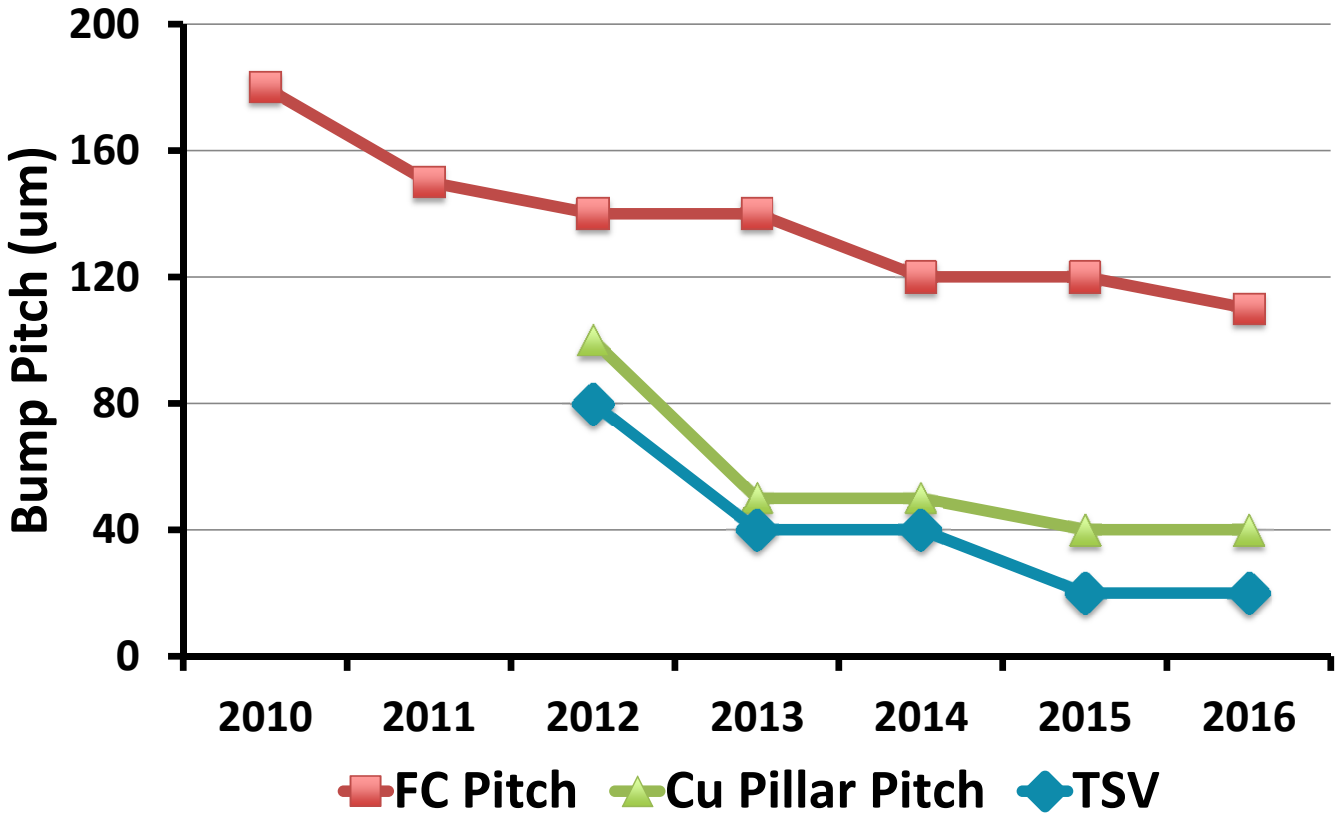


Bump pitch fails to scale at proportional rate compared to logic scaling

ARM

IO Density Trends Drives Transition to TSV

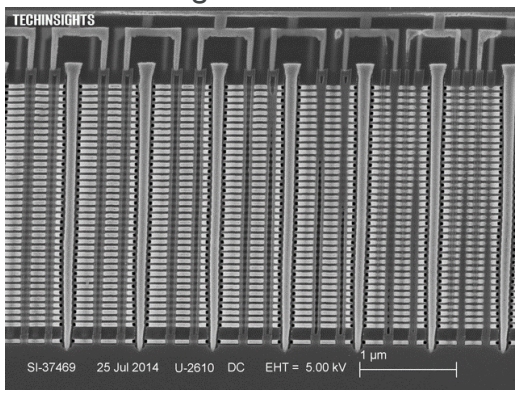
- Silicon Interconnect using Through-Silicon Via (TSV) necessary for IO Density
- TSV and Cu Pillars are the fundamental technologies driving stacked packaging



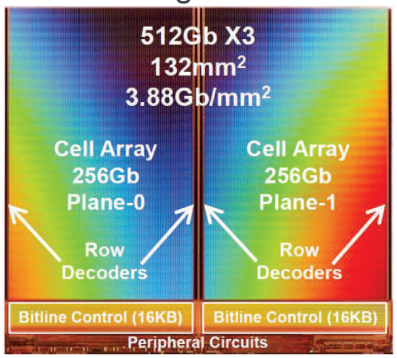
ARM

3D-ICs in Products Today

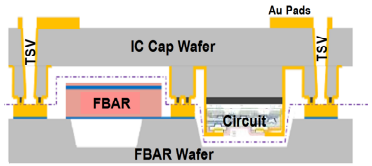
Samsung 3D V-NAND



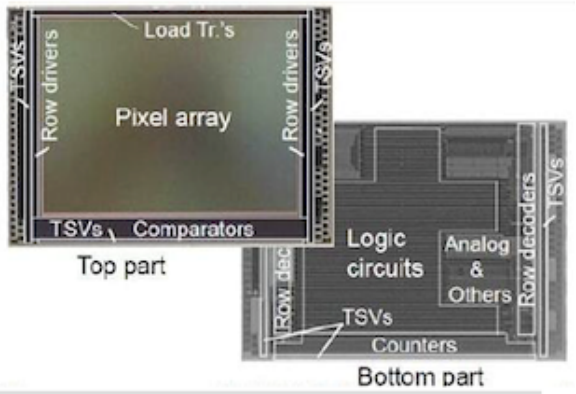
Western Digital – V-NAND



3D- NAND Flash

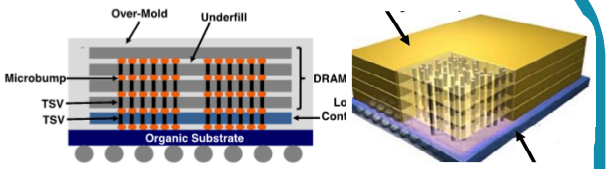


MEMS+RF

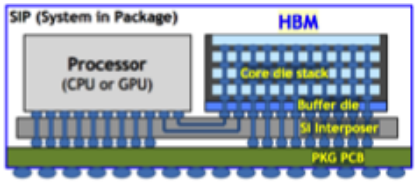


Sony's CMOS Image Sensors

TSV-Enabled CIS



HMC



HBM (DRAM+GPU)

High Bandwidth Memory (HBM)
AMD Radeon GPUs

High-Performance DRAM

ARM

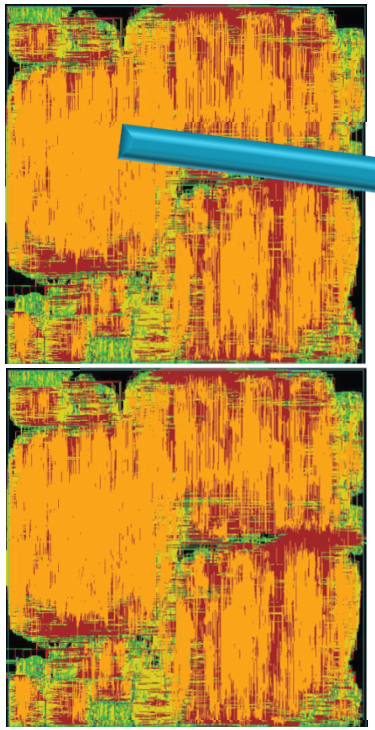
Monolithic 3DIC – TSV to MIV

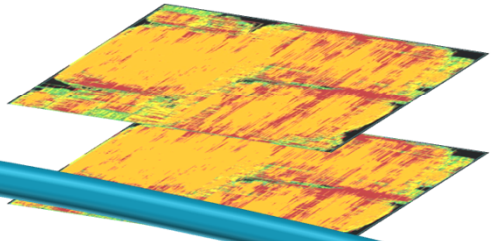
2D

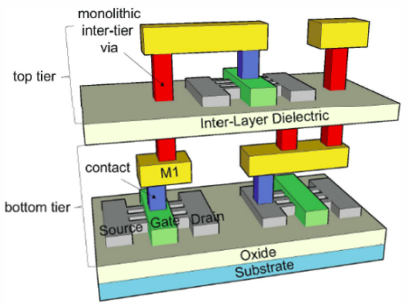
Core / block folding

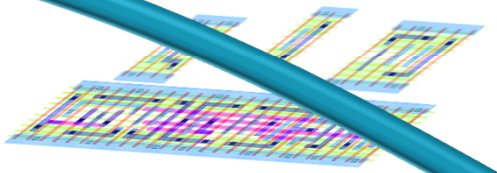
Gate-level folding

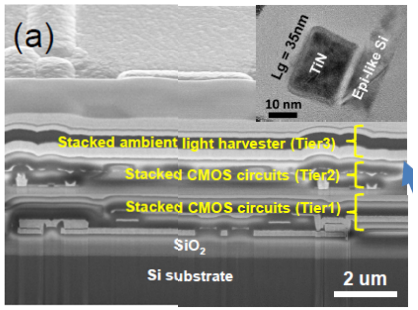
Monolithic 3D











Power/Performance:
~One Moore's Law Node

Chang, Acharya, Sinha, Cline, Yeric, Lim; ISLPED 2015

ARM

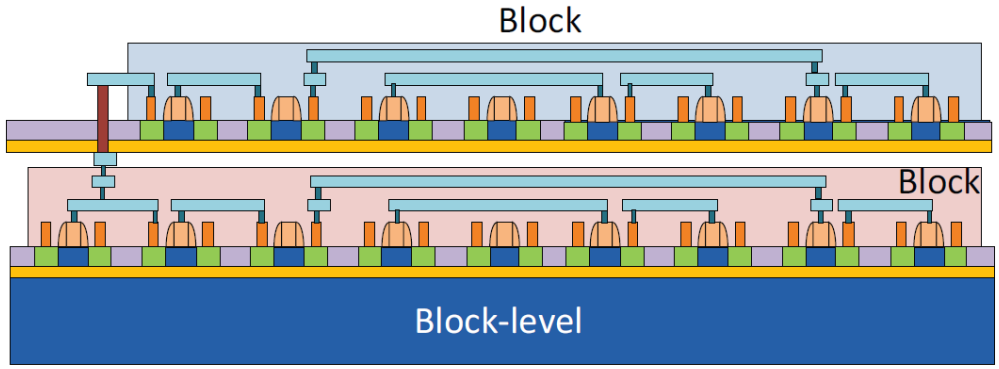
TSV vs MIV Comparison

	TSV	MIV
Dimensions	5-10um diameter	2-3x size of min. via (40n-100n)
Keep-outs	Dedicated keep-out required	Can be placed within whitespace
Connectivity	100-1000 (Die-to-Die)	1K-1000K (Block-to-block)
Parasitics	Significant – Major impact on timing and power	Relatively minor impact on timing and power
Current Density	High	Relatively low

Outline

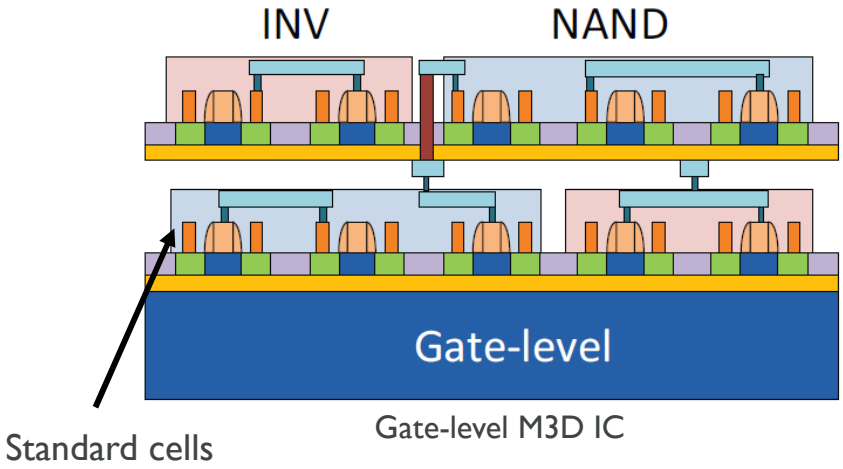
- Background and Motivation
- **Monolithic 3D IC Flavours**
- Cascade2D – Using 2D-tools for 3D Integration
- Power Delivery for Monolithic 3D
- Summary

Monolithic 3D Flavors



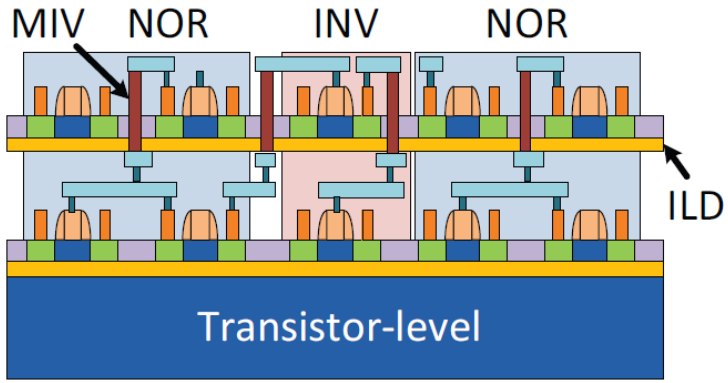
Block -level M3D IC

- Multiple flavours of monolithic 3D possible
- Trade-offs associated with each configuration



Gate-level M3D IC

MIV: Monolithic Inter-tier Via



Transistor-level M3D IC

Shreepad Panth, Kambiz Samadi, Yang Du, and Sung Kyu Lim, "Design and CAD Methodologies for Low Power Gate-level Monolithic 3D ICs," ISLPED 2014

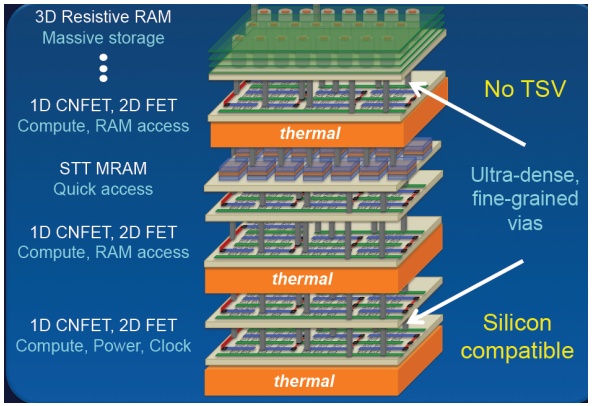
ARM

Monolithic 3D: Fabrication Challenges

- **Challenge:** Getting high performance transistors by low-temperature processing on the top-tier to preserve bottom tier interconnects and devices
  - Active research by CEA-LETI to overcome integration challenges
  - Stanford team (Shulakar, Mitra et al.) integrates CNT computer with RRAM in monolithic 3D configuration

Process step	Planar technology		Top FET process	
Dopant activation	Thermal activation	>1000°C spike	Solid Phase Epitaxy (SPE)	600°C 30s
			Nanosecond laser anneal (λ=308nm, 150ns)	<700°C, 100ns
Epitaxy	Si ($SiCl_2$+HCl)	750°C, min	Si Cyclic Deposition Etch Growth:Si_2H_6 Etch : HCl + GeH_4	600°C, min
	SiGe ($SiCl_2$+GeH_4+HCl)	650°C, min	SiGe Deposition Etch Growth:Si_2H_6 +GeH_4 Etch : HCl + GeH_4	500°C, min
Offset nitride spacer	ALD	600°C, h	Low k dielectrics	500°C, min
Substrate bonding	NA	NA	SOI bonding and substrate etch	200°C, h (ox/ ox bonding)
			SMART CUT ™	200°C, h (ox/ ox bonding) 400°C, 1h (SMART CUT ™)

VLSI Technology, 2015



Shulakar, Mitra et al.

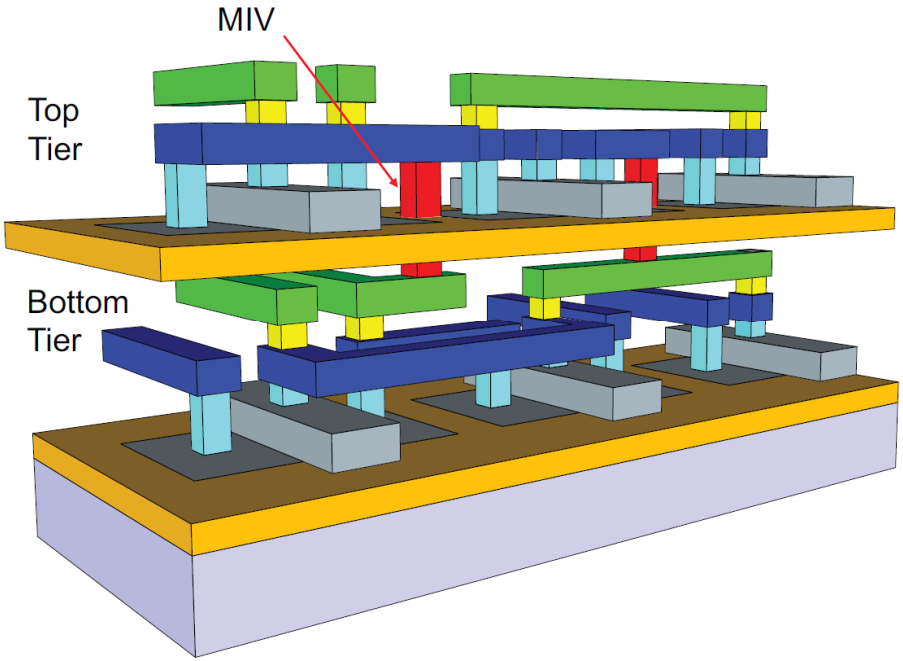
- **Our Goal:** Quantify the power, performance and area savings assuming manufacturing challenges can be met.
  - Gives us an upper bound on the improvements possible with M3D

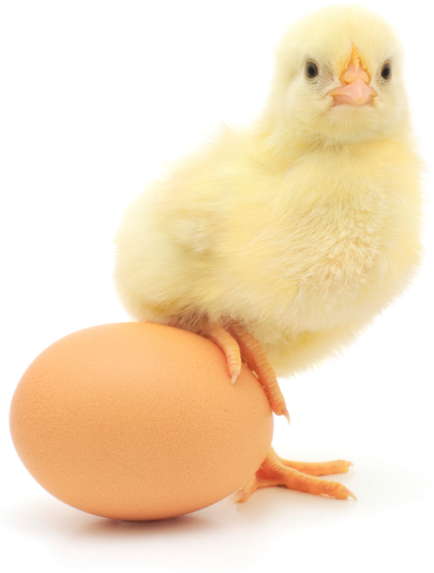
ARM

How do we implement a design in 3D?

- There are no EDA tools that support monolithic 3D designs
  - Chicken and an Egg Problem: Prove the benefits of 3D without the 'egg', i.e., EDA tools

Need tricks to implement M3D with 2D tools





ARM

Georgia Tech GTCAD Collaborators



Kwang Yook Chang



Sun Kyu Lim

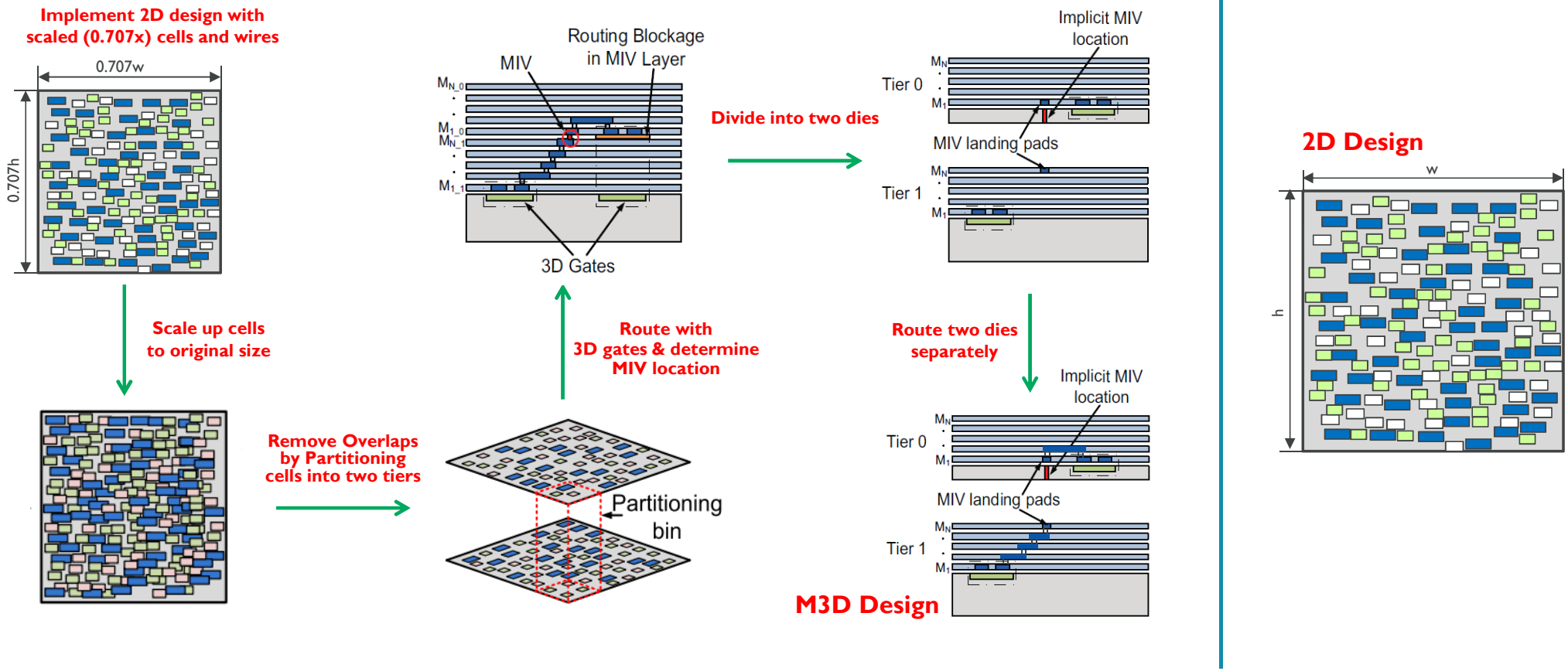
- Match-making for monolithic 3D IC: finding the right technology node, DAC 2016
- Cascade2D: A Design-Aware Partitioning Approach to Monolithic 3D IC with 2D Commercial Tools, ICCAD 2016
- Frequency and Time Domain Analysis of Power Delivery Network for Monolithic 3D ICs, ISLPED 2017 (Accepted)



GTCAD Lab

ARM

M3D Implementation using Shrunk 2D flow

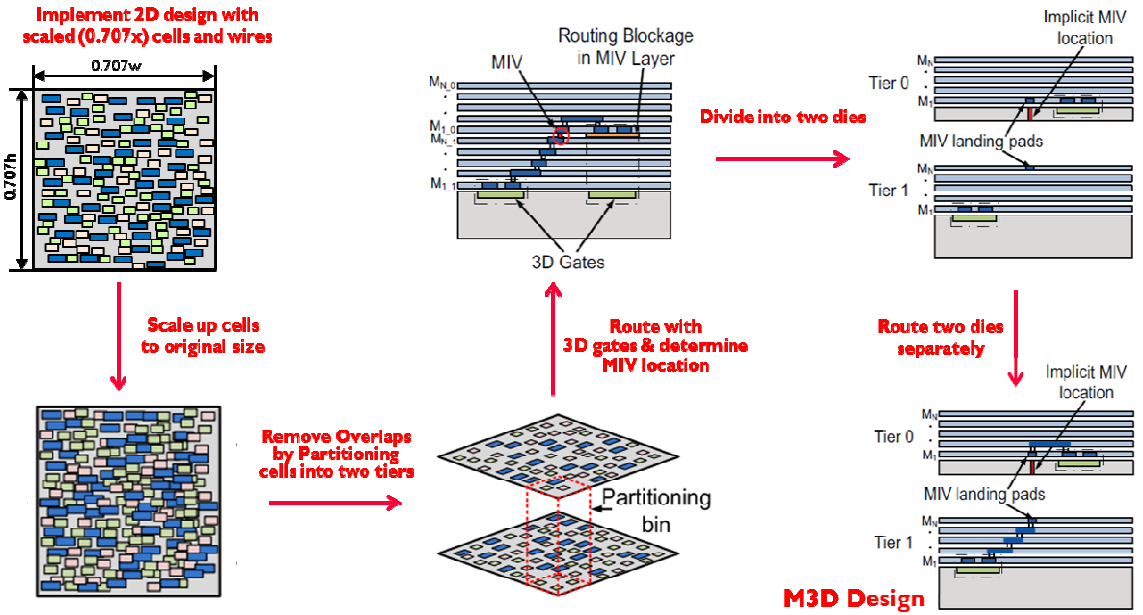


Shreepad Panth, Kambiz Samadi, Yang Du, and Sung Kyu Lim, "Placement-Driven Partitioning for Congestion Mitigation in Monolithic 3D IC Designs," ACM International Symposium on Physical Design, 2014

ARM

M3D Implementation using Shrunk 2D flow

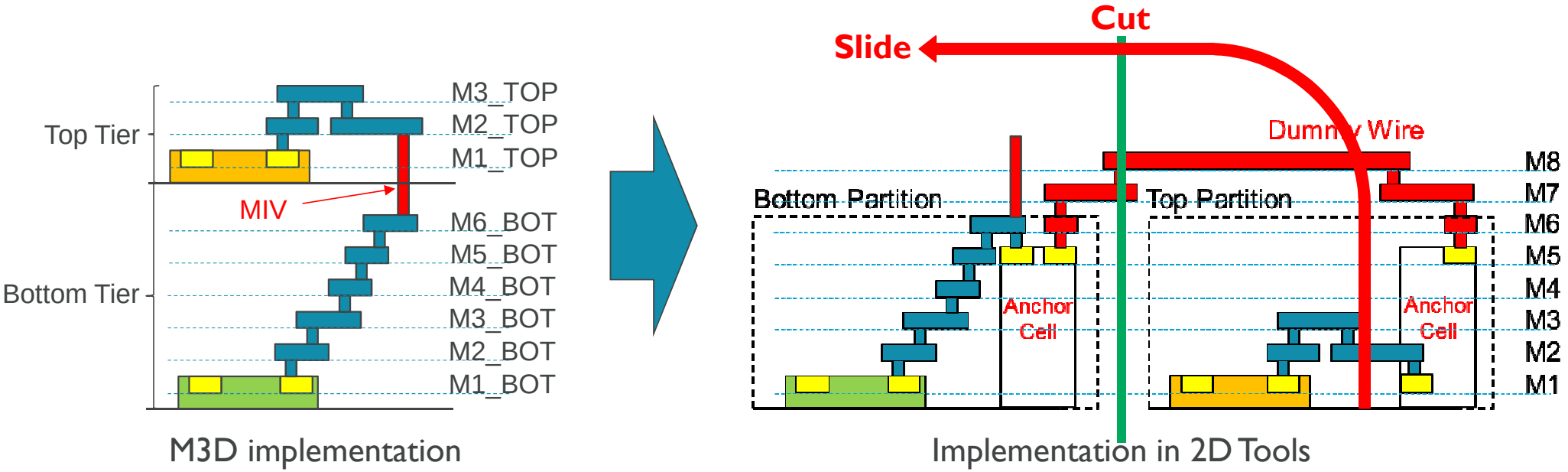
- Designer has no control over tier-assignment of cells
  - Implements gate-level M3D
  - **CANNOT** handle RTL-level constraints
- Implements top and bottom tier **separately**
  - → Problem with meeting timing
- Buffer insertion based on **shrunk** technology parameters
  - → Inaccurate buffer insertion



ARM

Proposed Method: Cascade 2D Flow

- Placing cells in 3D



M3D implementation

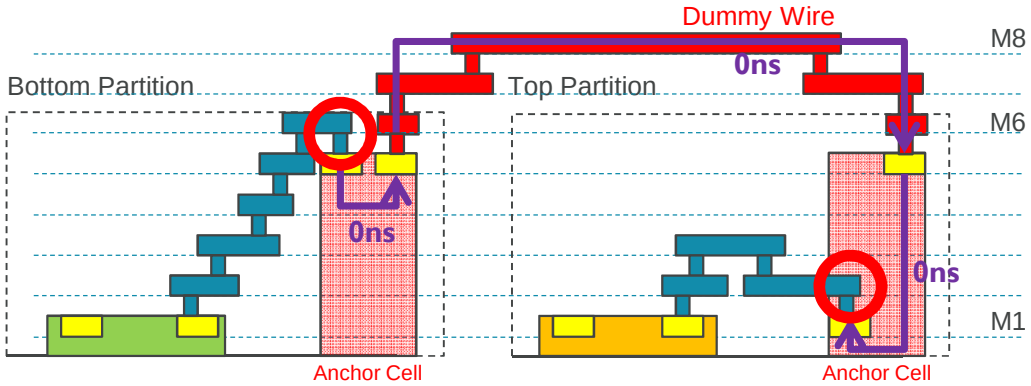
Implementation in 2D Tools

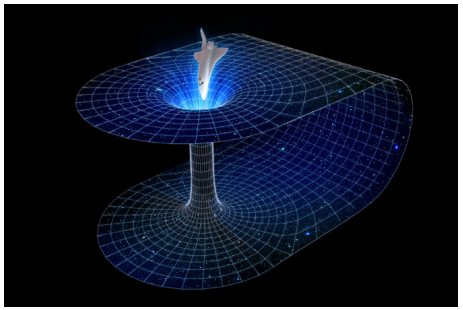
K. Chang et. al. , ICCAD 2016

ARM

Cascade2D Stage

- Implement top and bottom tier simultaneously adjacent each other
- Create hard partition between tiers, so that cells (modules) get placed at their assigned tiers
- Model MIVs with sets of anchor cells and dummy wires
  - These anchor cells and dummy wires act as a **wormhole** connecting top metal layer of bottom partition and bottom metal layer of top partition





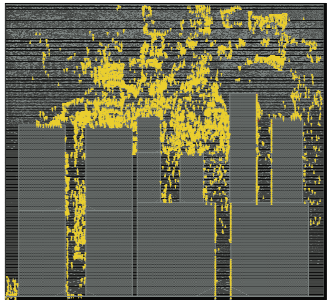
ARM

Experimental Setup

- Benchmark:
  - Commercial 32-bit in-order microprocessor
- Process Nodes and Design Libraries

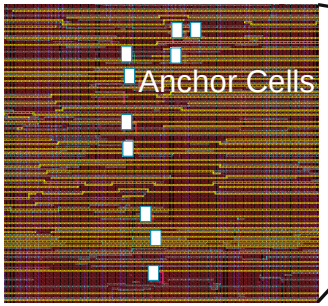
Parameters	28HPM	16FF	PTM 7nm
Transistor Type	Planar	FinFET	FinFET
Supply Voltage	0.9V	0.8V	0.7V
CPP (representative)	120nm	90nm	50nm
M1 Pitch (representative)	90nm	64nm	36nm
PDK	Foundry	Foundry	Predictive
Library	Production	Production	Prototype
MIV cross-section	80x80nm	40x40nm	32x32nm
MIV Height	140nm	170nm	170nm

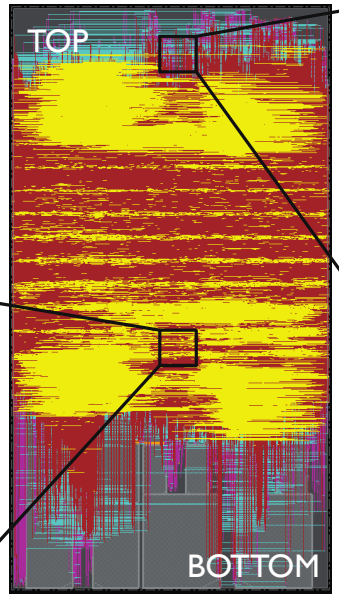
Die-Images of Cascade2D steps

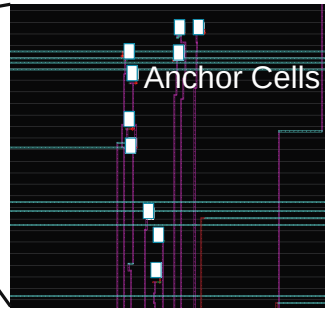


MIV-locations

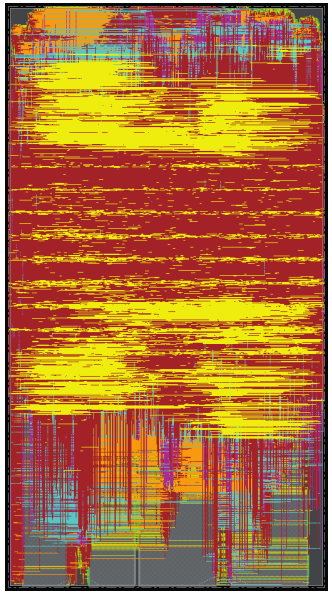
Assembling top and bottom view





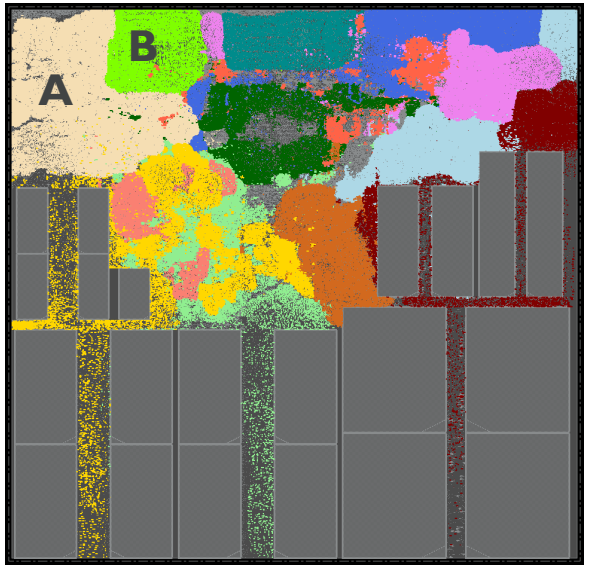


Cascade2D Design

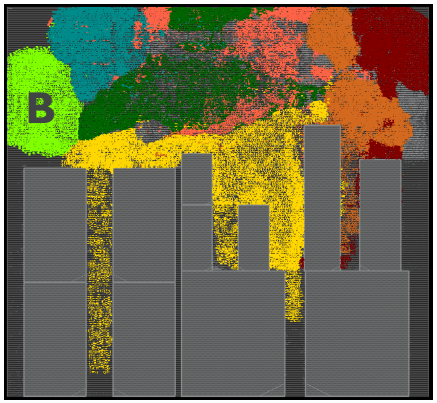


ARM

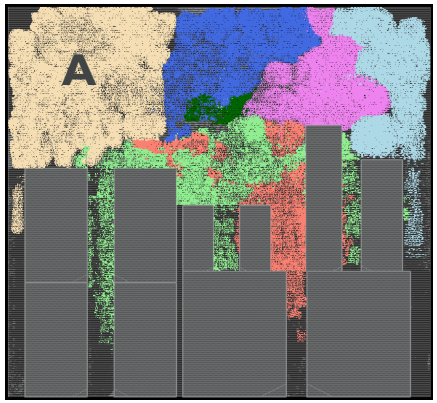
Floorplan Comparison: 2D vs. Cascade2D M3D



2D implementation



Top Tier

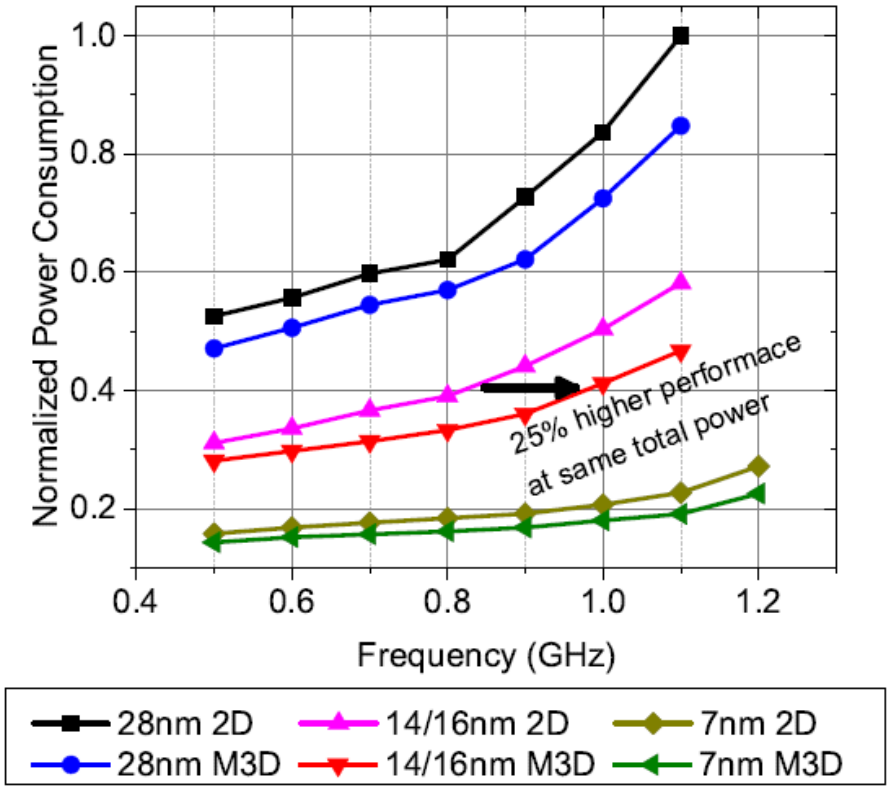


Bottom Tier

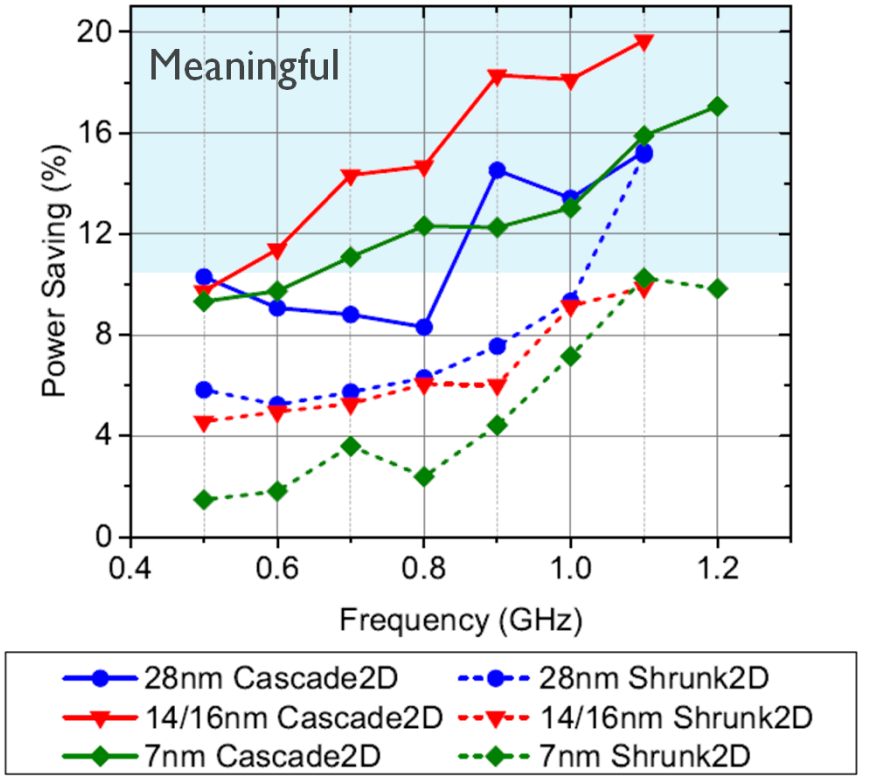
ARM

Power Comparison

2D vs. Cascade 2D M3D

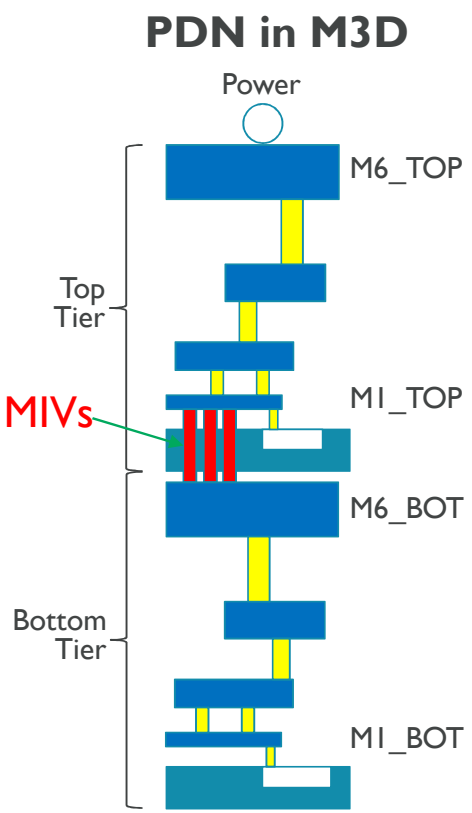


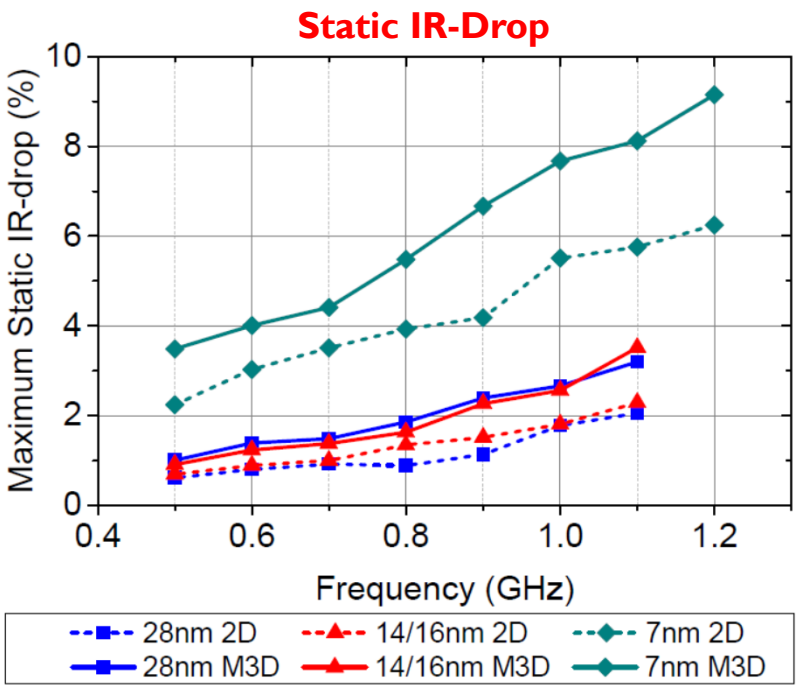
Cascade 2D M3D vs Shrunk 2D M3D

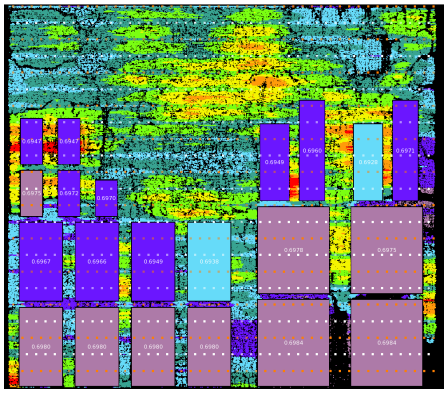


ARM

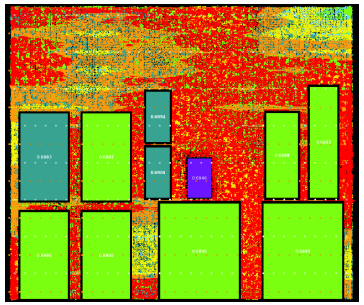
M3D PDN Analysis: Static IR drop



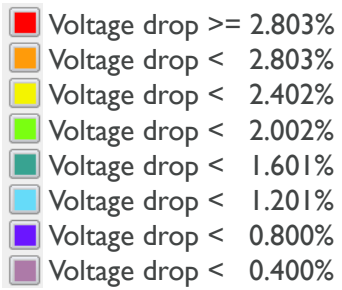




Static IR-drop on 2D



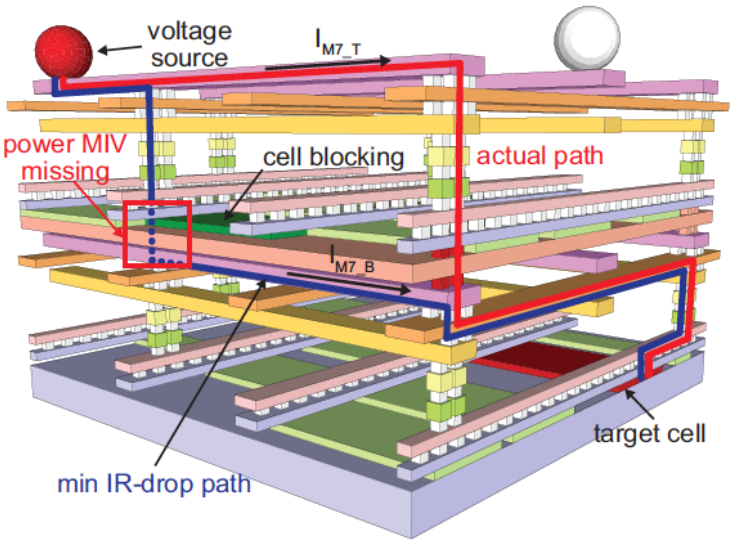
Static IR-drop on M3D Bottom Tier

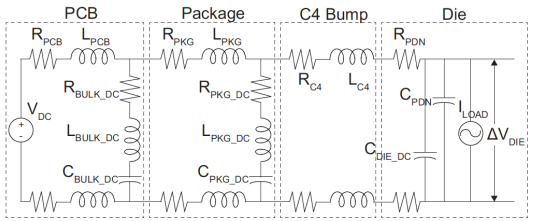


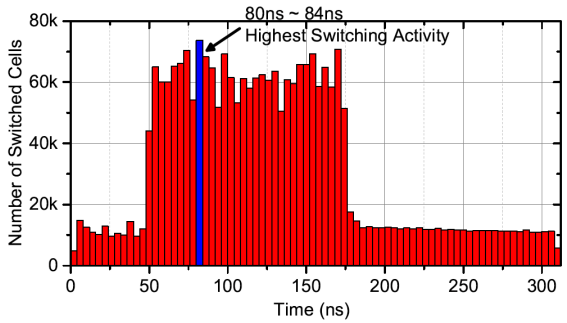
ARM

M3D PDN Analysis: AC Response

- System level PDN study of both resistive (IR) and inductive (Ldi/dt) components of power-supply integrity in monolithic 3D designs.
- Benchmark Circuits:
  - DCT
  - AES-128
  - JPEG Encoder
- Technology node: 45nm OpenPDK
- Detailed modeling of PCB, package, C4 bumps and M3D PDN
- Vector based analysis





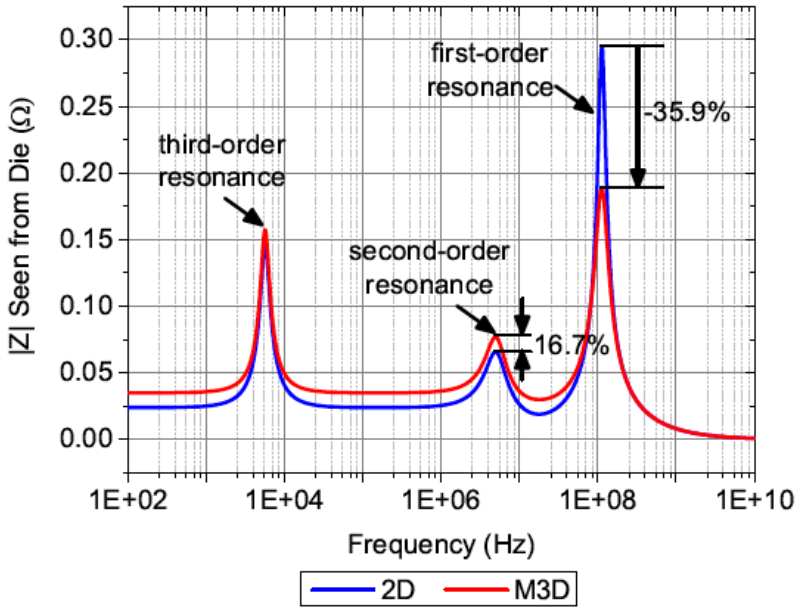


K. Chang, et. al. , ISLPED 2017 (to be published)

ARM

M3D dynamic PDN study: Observations

- 2X higher static IR drop in M3D designs compared to 2D; however voltage drop due to IR + Ldi/dt is only 11.3% worse.
- M3D designs utilize decap cells placed in the same plane (x-y) as well as the different tier (z-direction).
- Additional resistance in the PDN network makes it more resilient to overall current noise due to worst-case resonance oscillations, showing 36% peak impedance reduction.



K. Chang, et. al. , ISLPED 2017 (to be published)

ARM

Summary

- Developed a new design-aware partitioning and M3D implementation flow called "Cascade2D flow"

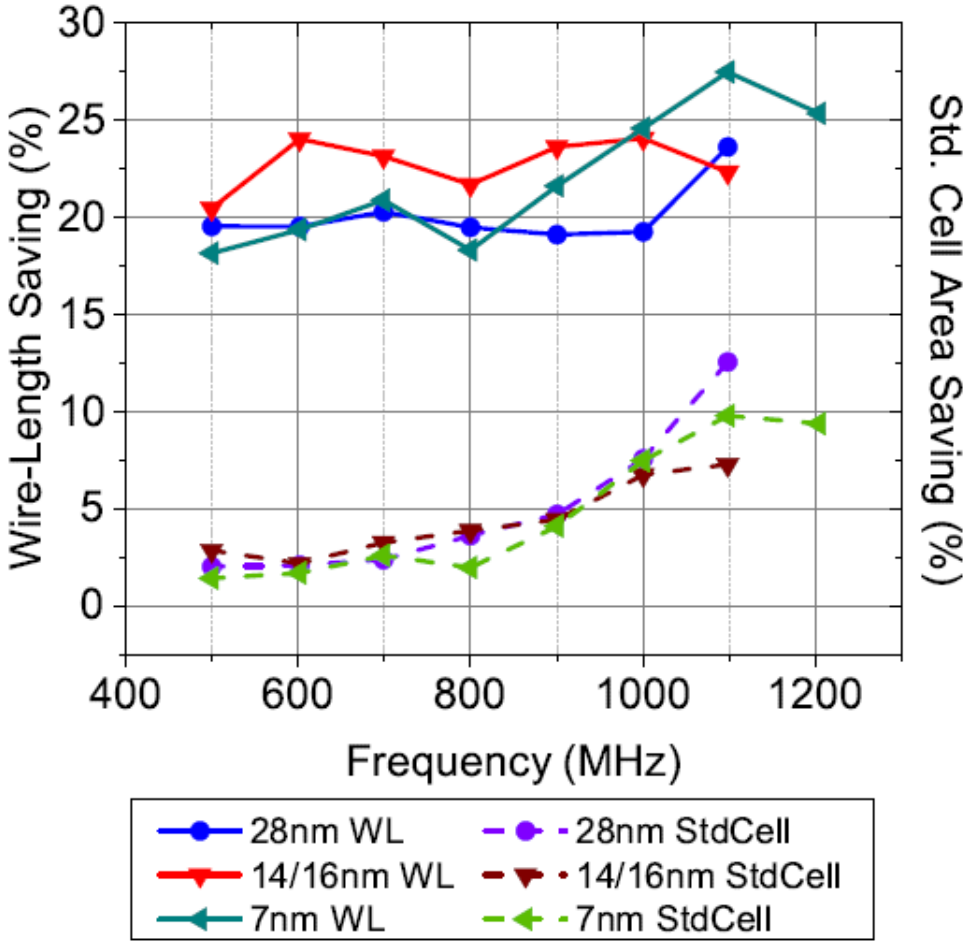
- Implemented M3D designs of a commercial 32-bit application processor at 28nm, 14/16nm and Predictive 7nm nodes using the Cascade2D flow
  - Cascade2D is flexible, handles RTL-level constraints and utilizes the optimization capabilities of commercial 2D tools.
  - Up to **25% higher performance** at iso-power, **20% lower power** at iso-performance and **10% reduced area** with Cascade2D flow.

- Explored power delivery network (PDN) design and ran static and dynamic analysis with a comprehensive PCB, package, bump and PDN model.
  - Ldi/dt voltage drop dominates the designs showing up to **11.3% worse** drops for M3D designs.
  - Additional resistance in M3D PDN can benefit by allowing better Ldi/dt noise immunity.

Future Work

- **GOAL:** Monolithic 3D demonstrator
  - Submitted funded research project proposal in collaboration with CEA-LETI to build a monolithic 3D prototype chip
  - ARM supplying processor IP to build demonstrator; 3D integration by CEA-LETI and other partners
  - Demonstrator to be followed by simulation studies exploring the impact of monolithic 3D on large scale designs
- Need to accelerate the process through improved collaboration between technology integrators, design houses and EDA support.
- Open research problems:
  - Thermal implications of high-density 3D integration.
  - Robust clock and power delivery network design.
  - Handling process variation across tiers.
  - Microarchitecture implications of 3D designs.
  - Heterogeneous integration: how do we simulate and evaluate a system, especially with different technology choices across tiers.
  - Cost?

Power Saving in M3D: Analysis

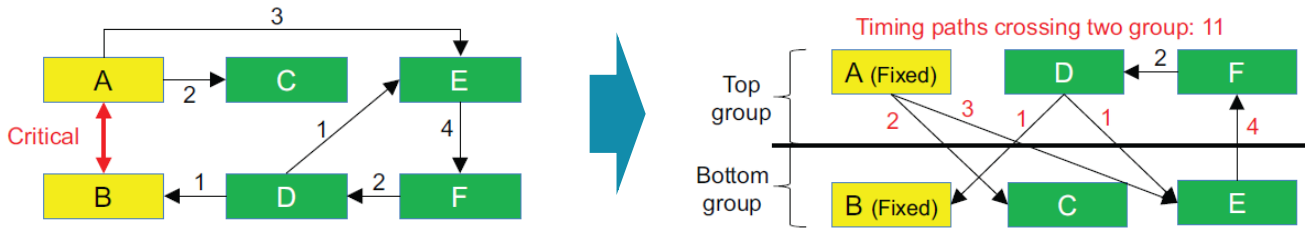


K. Chang et. al. , DAC 2016

ARM

Design-Aware Partitioning Stage

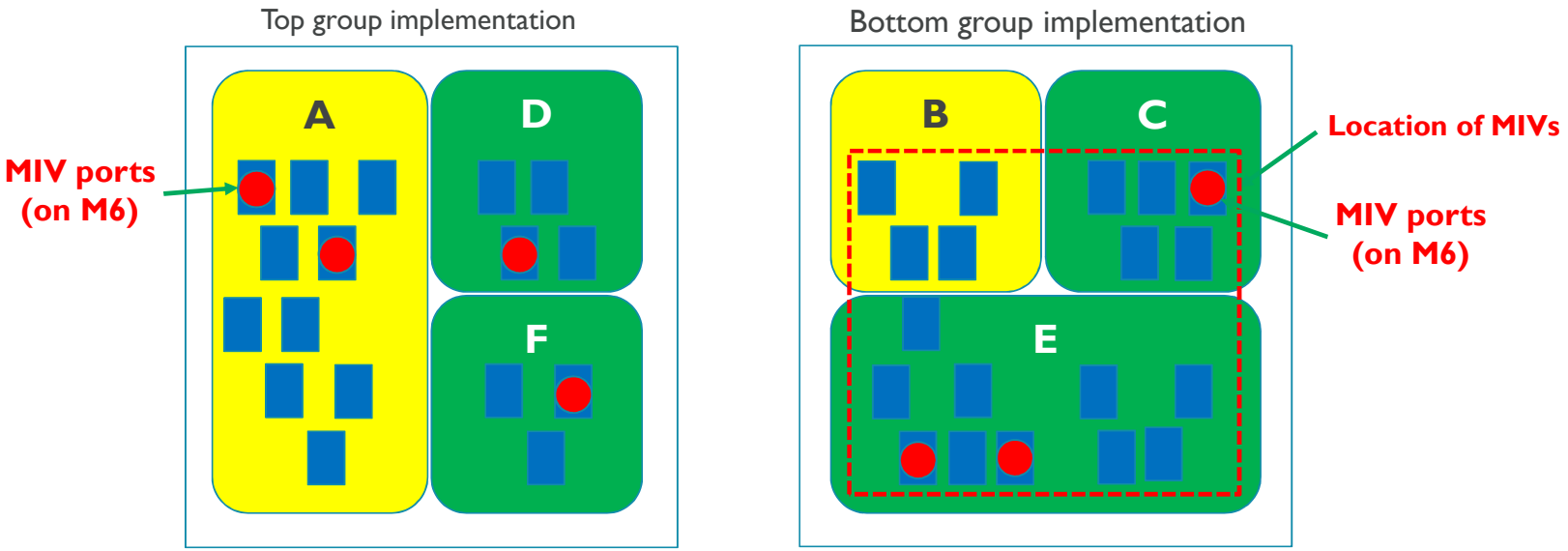
- Microarchitecture Organization
  - Two functional blocks are timing critical → Floorplan them into different tiers to benefit from z-dimension
- Information from 2D implementation
  - Extract cell area of functional modules
  - Extract all timing-paths and count the number of timing-paths crossing each pair of functional modules
- Partition functional modules with the following criteria:
  - Balance area between top and bottom tiers
  - Maximize the number of timing-path crossing two tiers



ARM

MIV Planning Stage

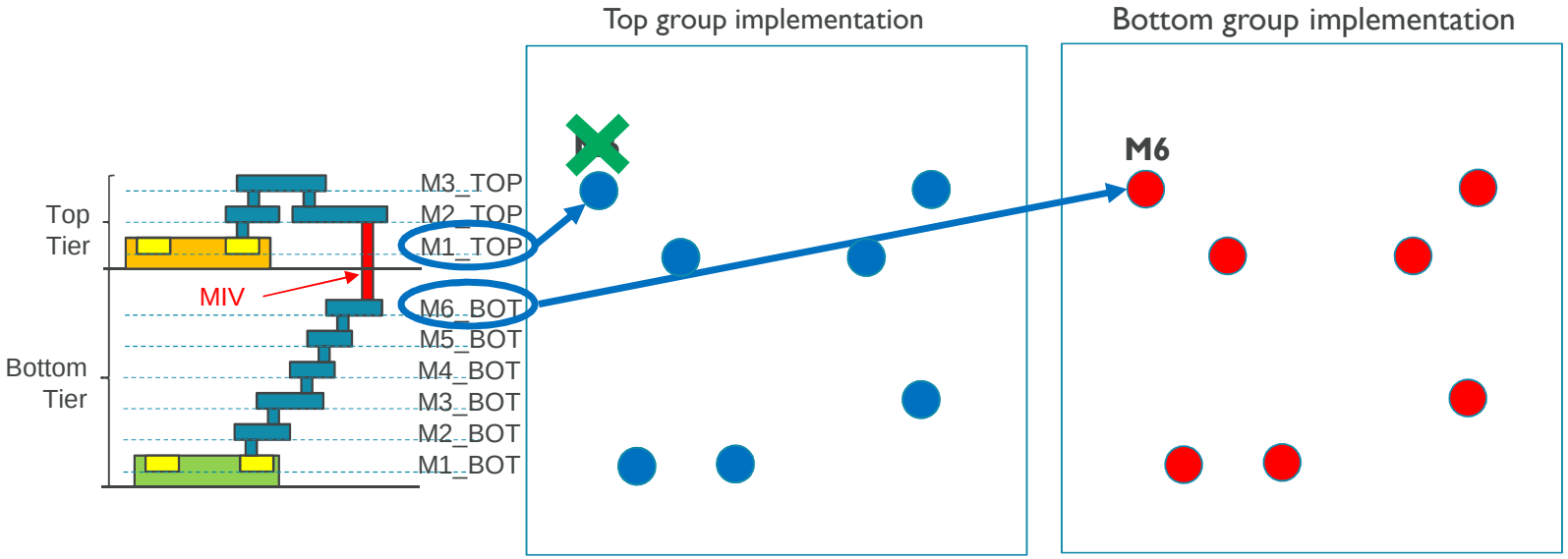
- Determine the location of MIVs in M3D design
- Objective: Minimize the distance between modules with large number of timing paths



ARM

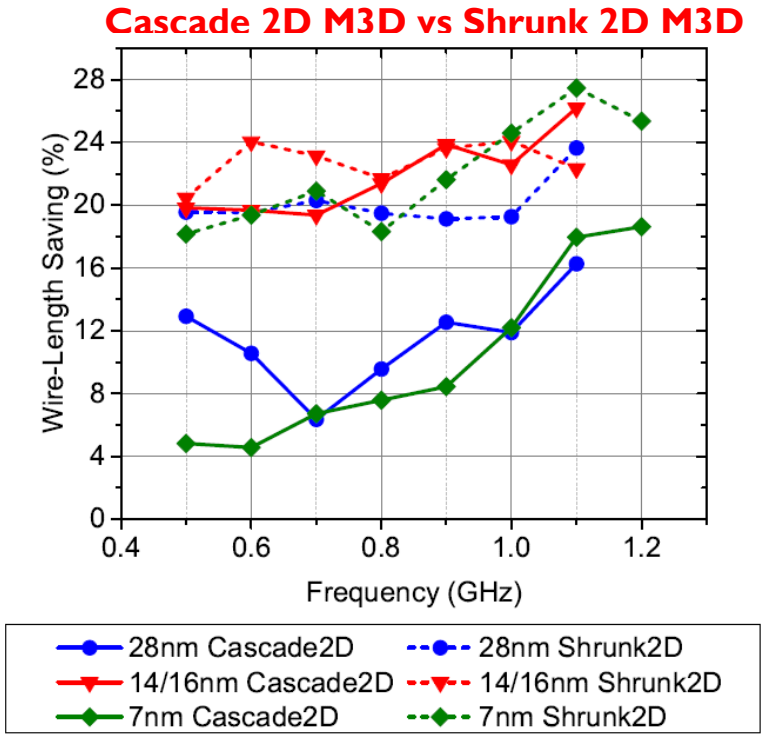
MIV Planning Stage

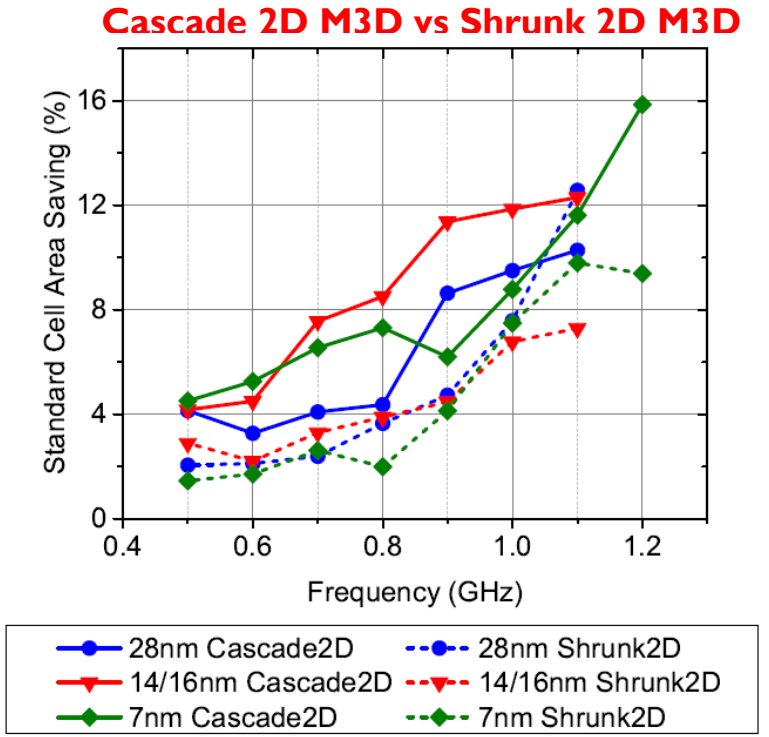
- Actual metal layer of MIV ports on top group is M1



ARM

Power Saving Analysis: Cascade2D Flow





Number of MIVs @ 1.0 GHz

	CLN28HPM	CLN16FF+	PTM7nm
Cascade 2D	7,545	7,545	7,545
Shrunk 2D	164,553	120,770	99,587

ARM

Results: Power Savings Breakdown

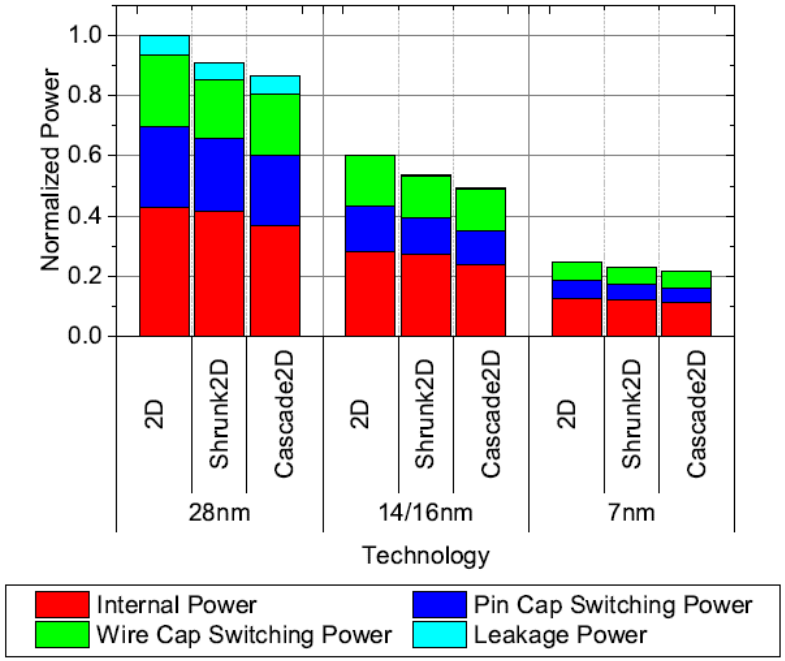
Total Power = Internal Power + Switching Power

= Internal Power + (Pin Cap Switching Power + Wire Cap Switching Power)

Internal Power + (Pin Cap Switching Power: Depends on cell count, drive-strength distribution

Wire Cap Switching Power): Depends on wire-length

80% 20%



ARM

Design metrics, area savings, run-time, etc.

	Normalized 2D			Shrunk2D percentage change from 2D			Cascade2D percentage change from 2D		
Parameters	28nm	14/16nm	7nm	28nm	14/16nm	7nm	28nm	14/16nm	7nm
Std. cell area	1	0.331	0.077	-7.6 %	-6.8 %	-7.5 %	-9.5 %	-11.9 %	-8.8 %
Wire-length	1	0.728	0.404	-19.3 %	-24.1 %	-24.6 %	-11.9 %	-22.6 %	-12.2 %
Wire cap	0.531	0.375	0.205	-18.1 %	-14.2 %	-13.7 %	-9.5 %	-19.7 %	-19.2 %
Pin cap	0.469	0.422	0.203	-12.1 %	-6.3 %	-9.7 %	-11.1 %	-13.2 %	-7.9 %
Total cap	**1**	**0.797**	**0.408**	**-15.5** %	**-10.1** %	**-11.7** %	**-9.6** %	**-15.2** %	**-12.9** %
Internal power	0.428	0.282	0.128	-4.8 %	-7.6 %	-4.7 %	-14.5 %	-15.2 %	-11.1 %
Switching power	0.505	0.318	0.119	-13.4 %	-10.6 %	-10.1 %	-13.0 %	-20.8 %	-15.1 %
Leakage power	0.066	0.002	0.000	-7.7 %	-4.0 %	-2.0 %	-9.5 %	-7.7 %	-2.8 %
Total power	**1**	**0.602**	**0.247**	**-9.3** %	**-9.1** %	**-7.2** %	**-13.4** %	**-18.1** %	**-13** %

Normalized design metrics at 1GHz

Parameters	2D	Cascade2D	
Die-area	1	1	0.9
Density	69.7%	63.2%	71.1%
Total power	1	0.841	0.871

Area Savings

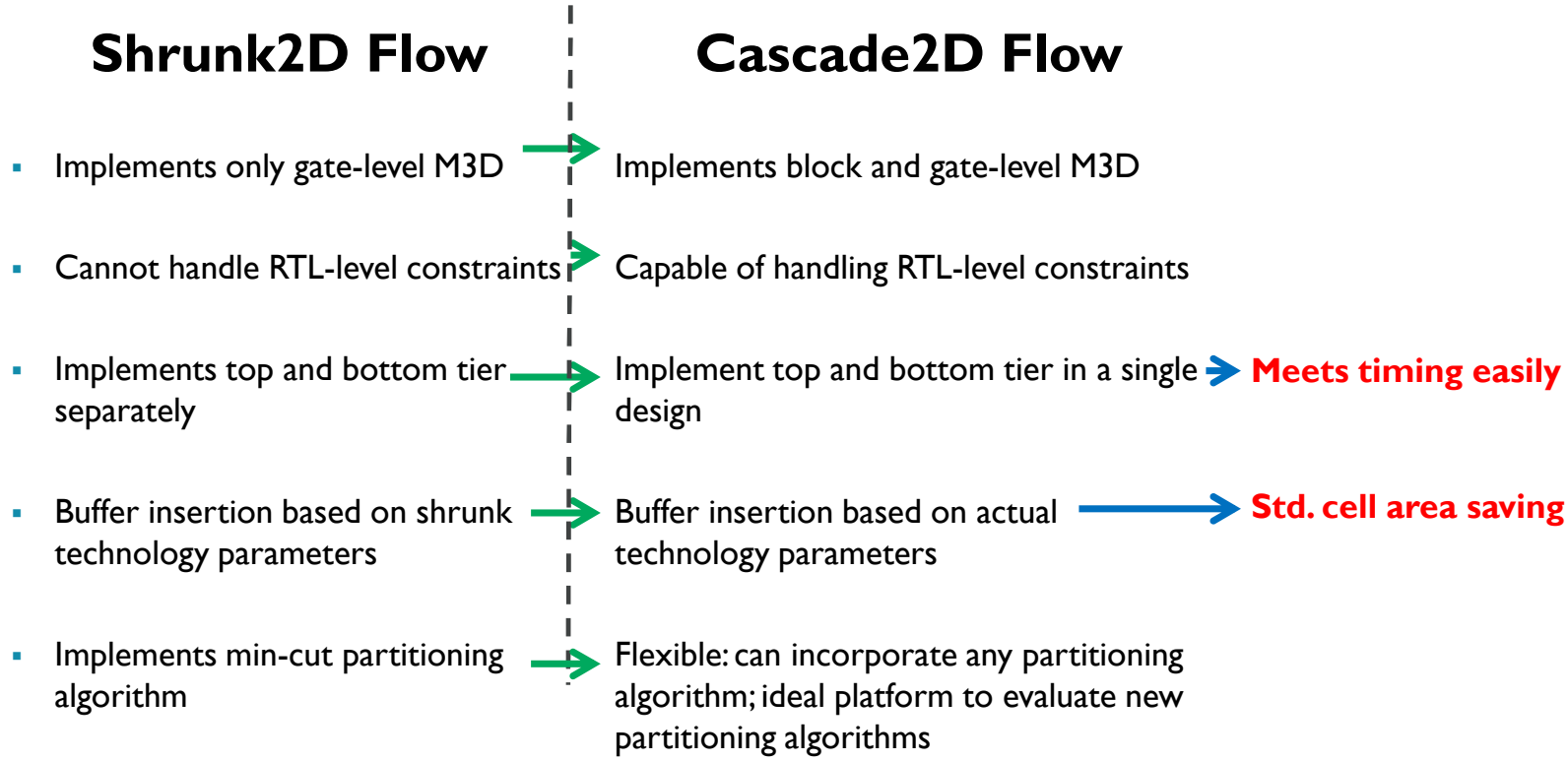
Shrunk2D flow		Cascade2D flow	
Step	Run-time	Step	Run-time
1. Shrunk2D impl.	5hr	1. Design-aware part.	0.5hr
2. Gate-level part.	0.5hr	2. MIV plan	2hr
3. MIV plan	0.5hr	3. Cascade2D impl.	4.5hr
4. Top/bottom tier impl.	1.5hr	-	-
Total	7.5hr	Total	7hr

Run-time

K. Chang et. al. , ICCAD 2016

ARM

Cascade2D vs. Shrunk2D: Qualitative Comparison



ARM