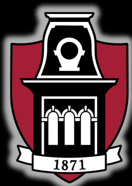
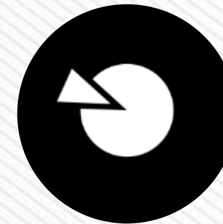



# Master's Thesis



# Design, Extraction, and Optimization Tool Flows and Methodologies for Homogeneous and Heterogeneous Multi-Chip 2.5D Systems

MD Arafat Kabir

**Thesis Committee:**
Prof. Yarui Peng (Chair)
Prof. David Andrews
Prof. Alexander Nelson

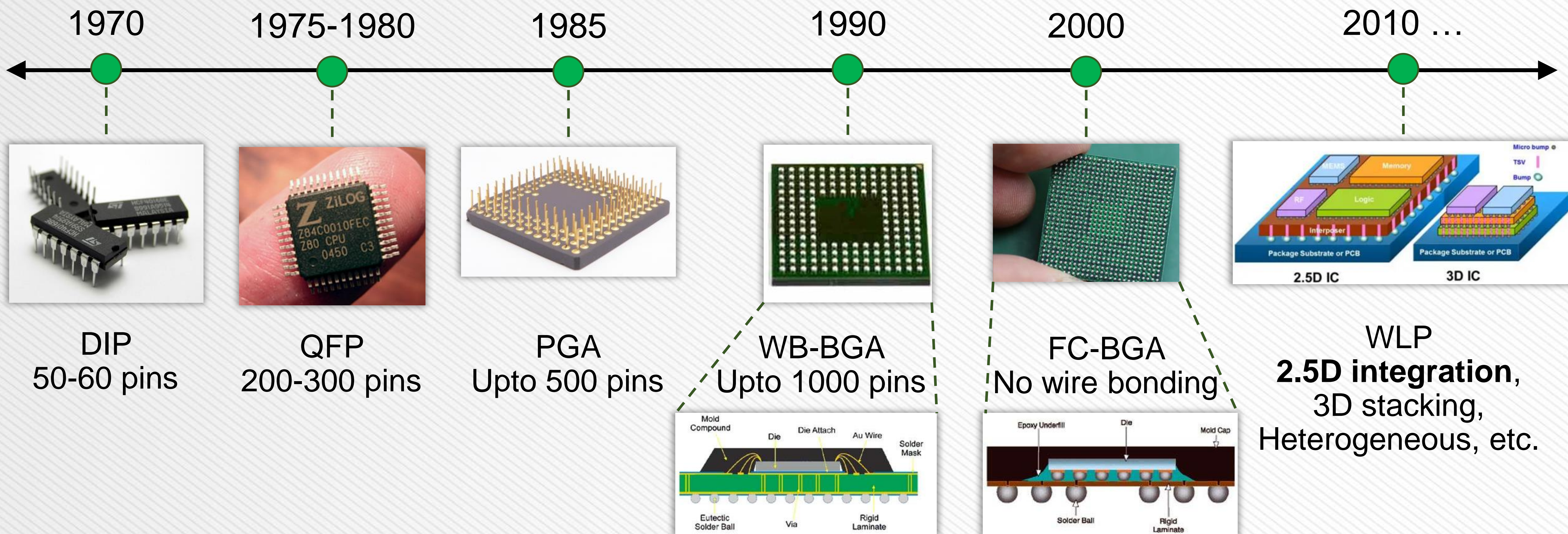
UNIVERSITY OF ARKANSAS

https://e3da.csce.uark.edu | makabir@uark.edu | +1 (479) 301-1293

# Brief History of IC Packages

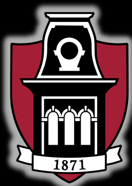
## ❑ Evolution of IC packaging

- Initially, development was driven by pin-count
- Now, driven by performance, power, bandwidth, etc.




Images from public domain: wikipedia.com, eetimes.com, researchgate.com, semiwiki.com

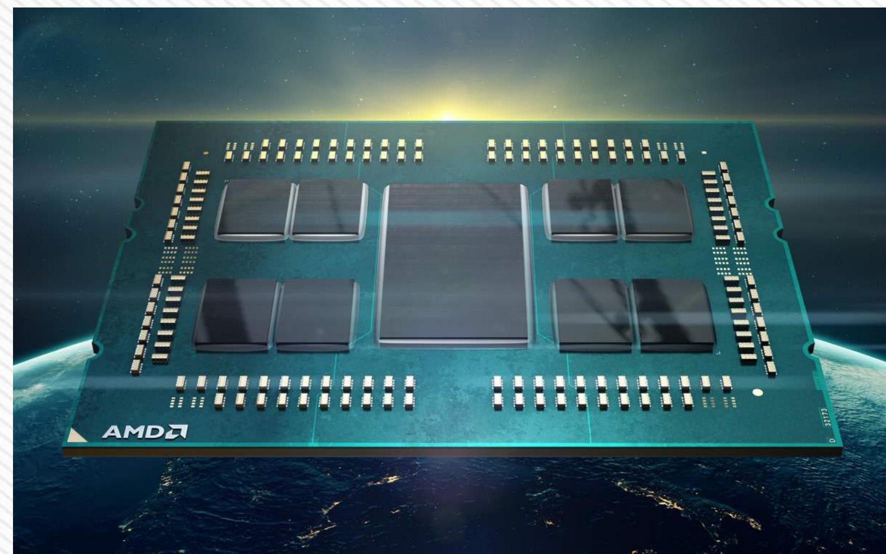
UNIVERSITY OF ARKANSAS



# 2.5D Systems Today

- **2.5D : multiple dies in a single package**
- **Package becomes increasingly critical in post-Moore's Law era**
  - Better performance, bandwidth, power, yield, compact size
  - Novel design techniques
  - Heterogeneous integration capabilities
  - Supports large systems

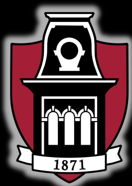
Intel Lakefield Processor*




AMD EPYC Processor*

*From public domain

UNIVERSITY OF ARKANSAS



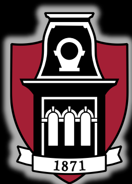
# Existing 2.5D Flows

- No standard tool flow exists

- Existing work
  - Flows for IP-reuse and active interposer
  - RDL routing methodologies
  - PDN and thermally-aware flows
  - Flows for IP security: obfuscation

- Die-by-die flow: chip-package cross-boundary interactions are ignored

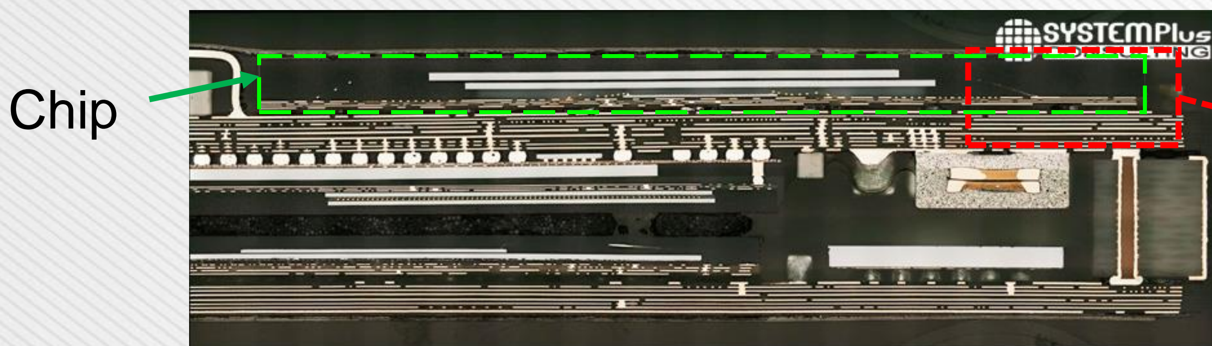
UNIVERSITY OF ARKANSAS



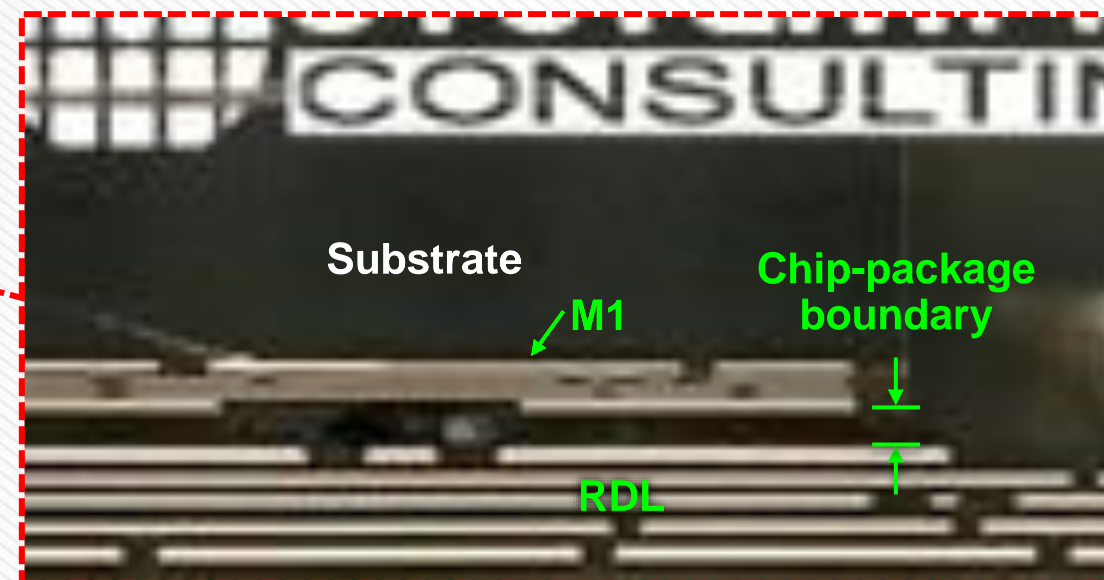
# Thin Dielectric Between Chip and Package

- Significant coupling between chip and package layers are expected

Chip



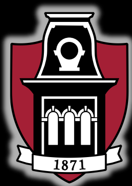

Apple A11 using TSMC's InFO*




This image was published in 2018!

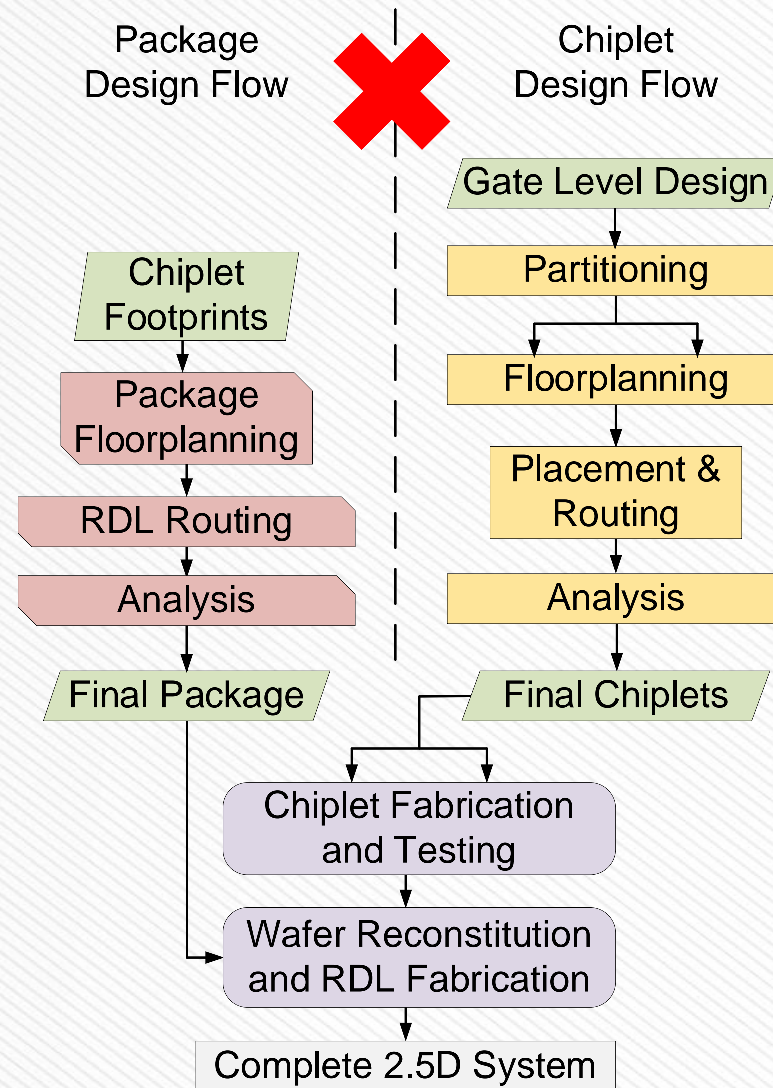
*From public domain

UNIVERSITY OF ARKANSAS



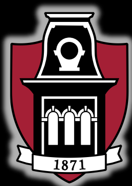
# Need For Cross-Boundary Flow

- ❑ **Chip-package gap decreasing**
  - InFO UHD: 1.5um (approx.)
- ❑ **Mainstream flow: die-by-die**



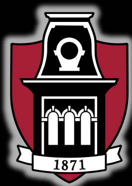

Traditional die-by-die flow

UNIVERSITY OF ARKANSAS



# Contributions of This Thesis

- **ASIC-CAD compatible cross-boundary flow frameworks**
  - Compatible with existing tools
  - Chiplet-package cross-boundary design and optimization
- **System-level iterative optimization**
- **Handling homogeneous and heterogeneous 2.5D systems**
- **Agile customization techniques**



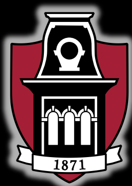

# Proposed Methodologies

- **Holistic flow for homogeneous 2.5D systems**
  - A framework for cross-boundary flow
  - Agile customization techniques
  - Silicon validation

- **In-context flows for heterogeneous systems**
  - A scalable per-chiplet in-context flow
  - A highly accurate per-technology in-context flow
  - A timing-accurate scalable in-context flow

- **Package inductance-aware system-level timing optimization flow**

UNIVERSITY OF ARKANSAS

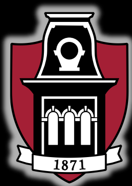


# Holistic Flow for High-Performance Systems

- **Designed for homogeneous systems**
  - Chipletization benefits
  - System-level performance and reliability
  - Better bandwidth, power, form-factor, etc.

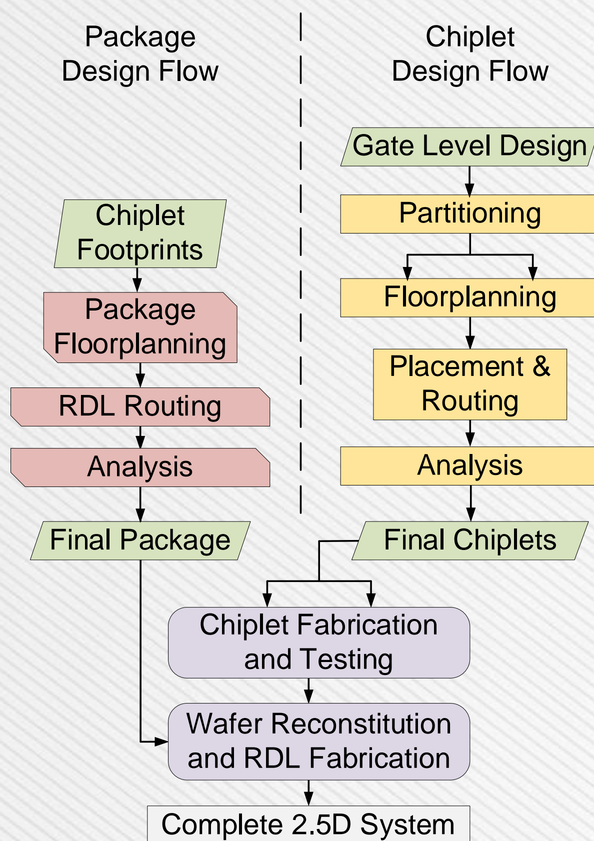
- **Comparable to 2D and traditional 2.5D flows**
  - Provides reference designs

UNIVERSITY OF ARKANSAS

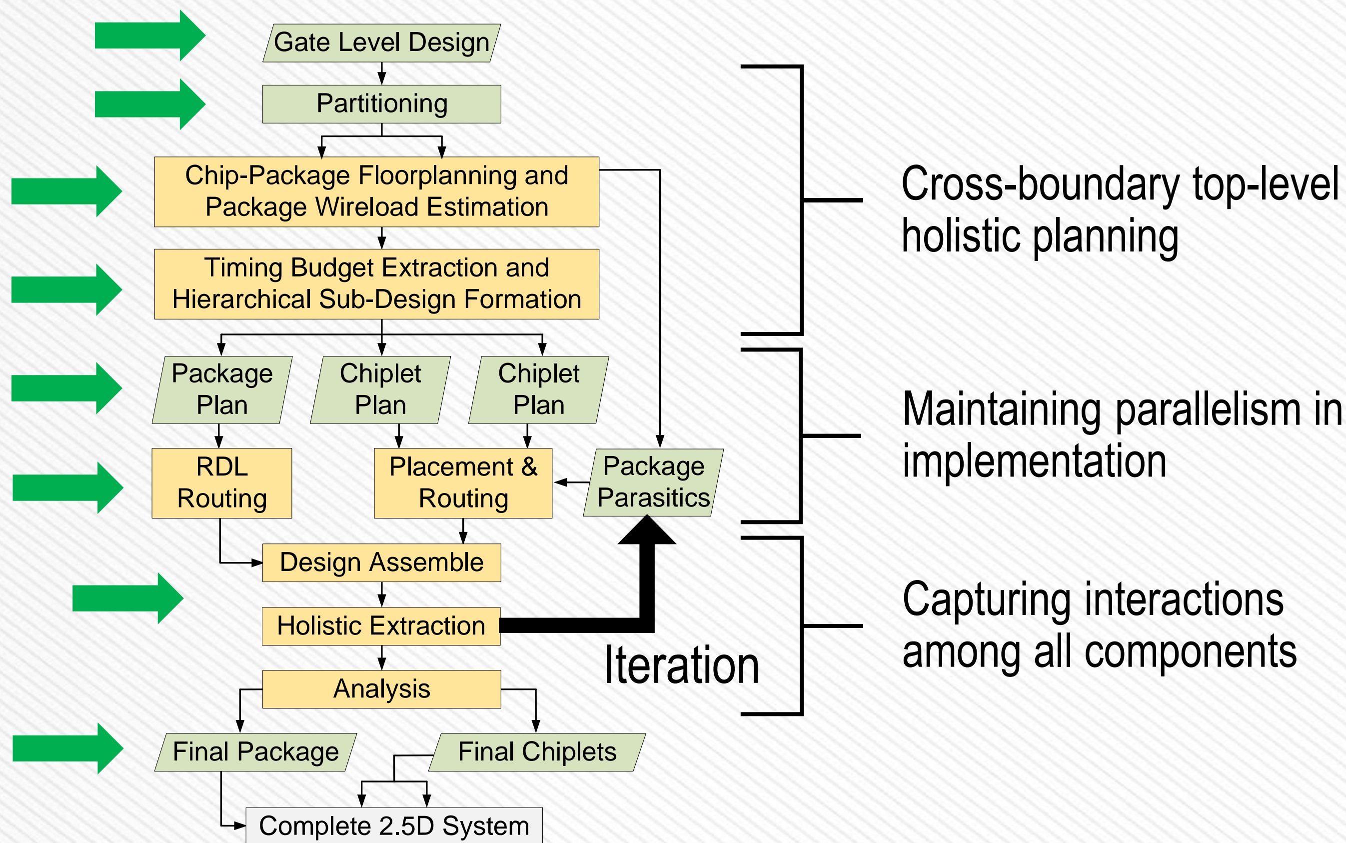


# The Holistic Flow

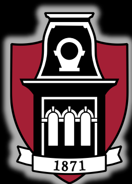
- Exchange of cross-boundary design information in planning, design, analysis, and optimization steps




Traditional die-by-die flow




Holistic Flow



# Experimental Study with MCU

- **Microcontroller system based on ARM Cortex-M0 core**
  - **16KB RAM: 4x4KB banks**
  - **Peripheral devices: GPIO, UART, timers, etc.**

Core-Chiplet

GPIO

AHB Slave Multiplexer

Cortex-M0 Processor

Peripheral Devices

Memory Interface

4KB SRAM

4KB SRAM

All logic + 8KB Memory

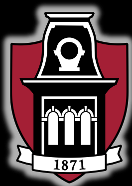
Memory Chiplet

4KB SRAM

4KB SRAM

8KB Extended Memory

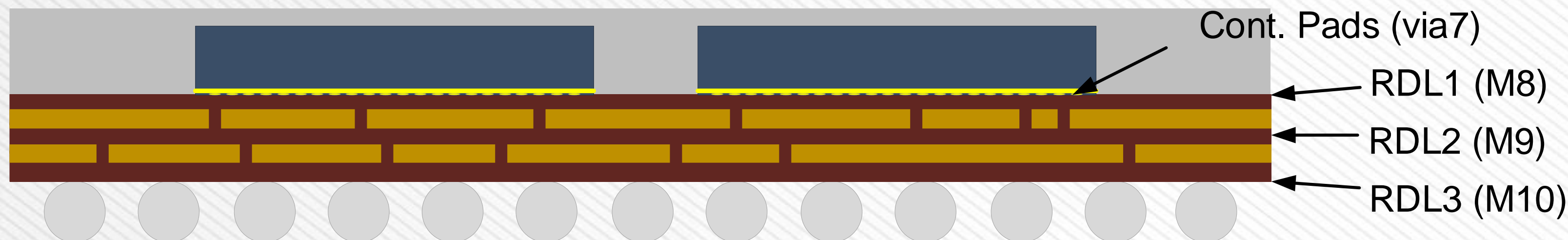
UNIVERSITY OF ARKANSAS



# Chiplet-Package Unified Technology

- **TSMC 65nm used as the PDK**
  - M1-M7 used for chiplet routing
- **Top three layers modified to include 2.5D package RDLs**
  - Similar to the TSMC 2.5D InFO technology

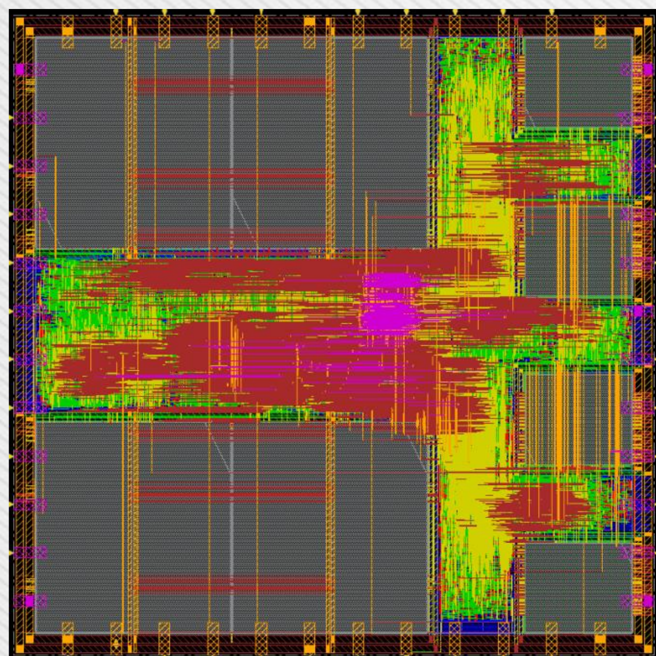
| Layer | Purpose | Width | Spacing | Thickness | Epsilon |
|---|---|---|---|---|---|
| **M1-M7** | Chip Internal Routing | TSMC | TSMC | TSMC | TSMC |
| **ILD7** | Inter-layer Dielectric | - | - | 5 um | 2 |
| **M8** | RDL1 | 10 um | 10 um | 5 um | 2.2 |
| **ILDR1** | Inter-layer Dielectric | - | - | 5 um | 2 |
| **M9** | RDL2 | 10 um | 10 um | 5 um | 2.2 |
| **ILDR2** | Inter-layer Dielectric | - | - | 5 um | 2 |
| **M10** | RDL3 | 10 um | 10 um | 5 um | 2.2 |



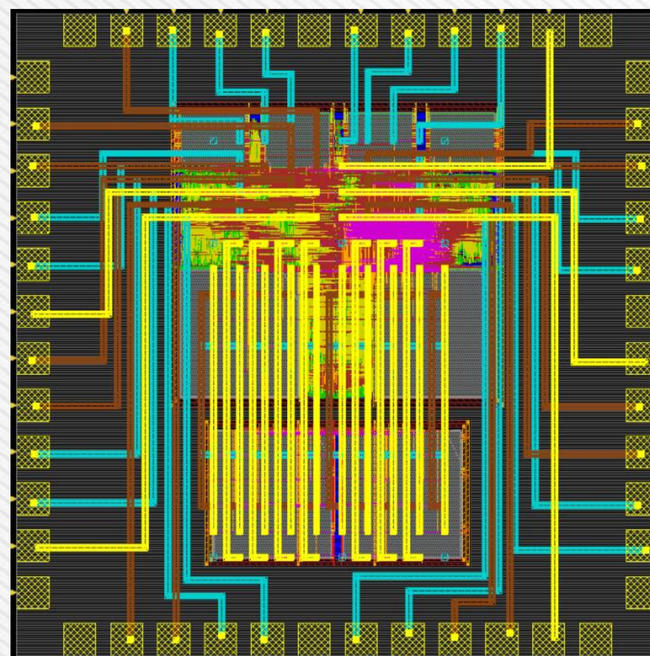

UNIVERSITY OF ARKANSAS

# Physical Design for Case Study

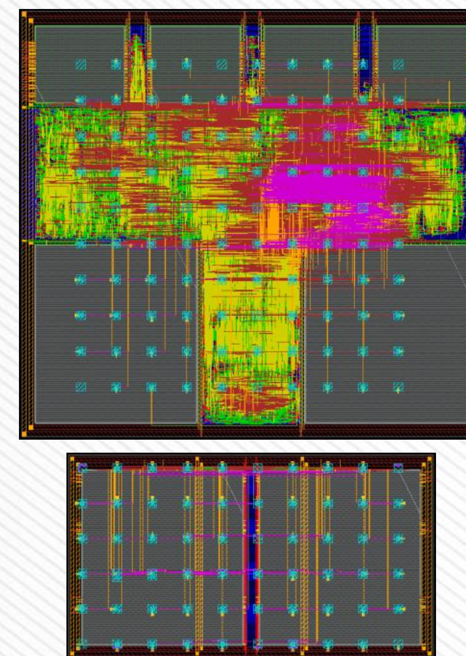
❑ **Different versions of the MCU implemented for comparative study**



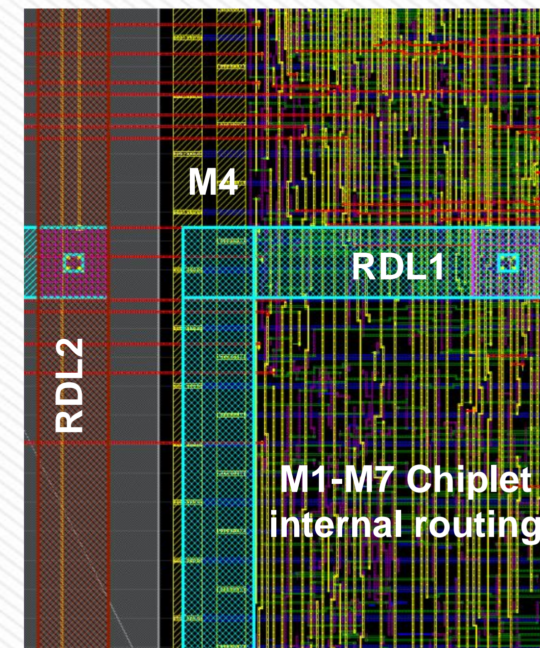
(a) Reference 2D system



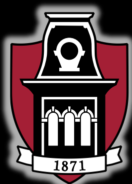
(b) Assembled 2.5D system



(c) Chiplet designs




(d) Zoomed-in view



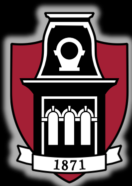
# Holistic Extraction Captures Interactions

- **Detailed chiplet-package coupling capacitance is captured**
  - Chiplet-package coupling not captured in die-by-die flow
  - M7-RDL coupling < M6-RDL coupling
    - less overlap on M7, M6-RDL1 runs in parallel

| Coupling Capacitance (CCAP) | | | | | | |
|---|---|---|---|---|---|---|
| **Metal Layer** | **M1-M5** | **M6** | **M7** | **RDL1** | **RDL2** | **RDL3** |
| **M1-M5** | 16348 | 222.5 | 446.7 | 185.3 | 18.61 | 10.18 |
| **M6** | 222.5 | 137.1 | 32.81 | **51.7** | 4.168 | 2.149 |
| **M7** | 446.7 | 32.81 | 371.1 | **32.43** | 1.459 | 1.891 |
| **RDL1** | 185.3 | 51.70 | 32.43 | 65.67 | 399.3 | 11.19 |
| **RDL2** | 18.61 | 4.168 | 1.459 | 399.3 | 103.3 | 390.5 |
| **RDL3** | 10.18 | 2.149 | 1.891 | 11.19 | 390.5 | 115.3 |

| Ground Capacitance (GCAP) | | | | | | |
|---|---|---|---|---|---|---|
| **Metal Layer** | M1-M5 | M6 | M7 | RDL1 | RDL2 | RDL3 |
| **Capacitance** | 31842 | 1526 | 477 | 853 | 251 | 420 |

VERSITY OF ARKANSAS



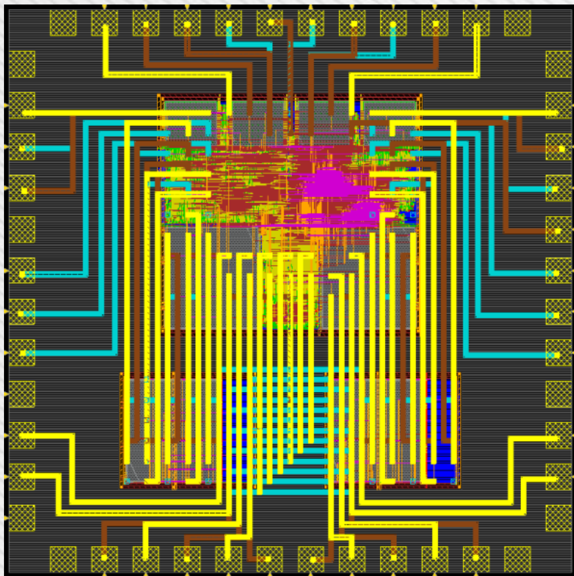
# Iterative Optimization Improves Performance

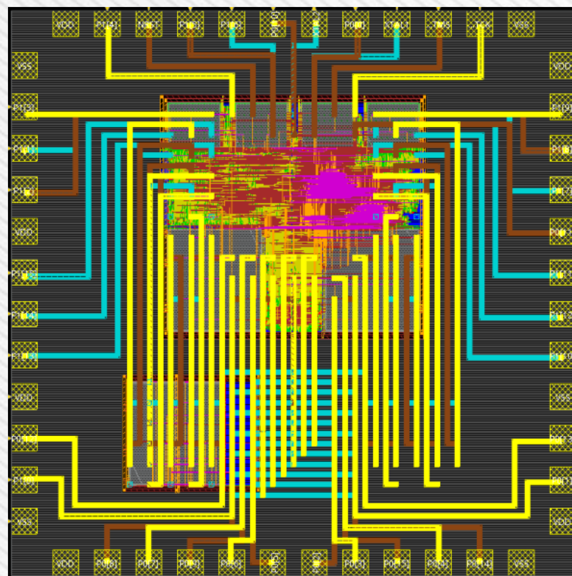
- Package overhead compensated by 85%

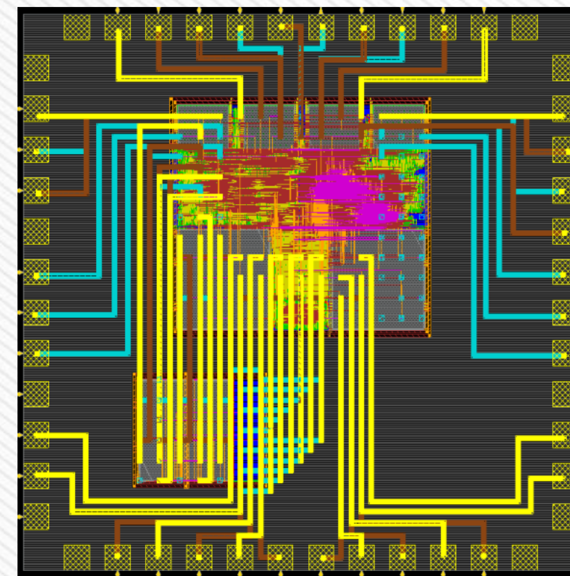
| Design Case | Chiplet Design | Logic Gates# | Buffer/ Inverter# | Die Size (um$^2$) | M6 WL (mm) | M7 WL (mm) | Power (mW) | Freq. (MHz) | Freq. Overhead |
|---|---|---|---|---|---|---|---|---|---|
| **Case-1** | 2D Chip | 24141 | 4760 | 600 x 600 | 15.13 | 8.562 | 20.1 | 400 | **0%** |
| **Case-2** | Core | 23933 | 4684 | 520 x 475 | 12.98 | 19.08 | 18.4 | 366 | **100%** |
| | Mem | 20 | 20 | 415 x 230 | 2.847 | 1.991 | 2.50 | | |
| **Case-3 initial** | Core | 23918 | **4634** | 520 x 475 | **13.6** | **18.12** | 18.2 | 384 | **47.05%** |
| | Mem | 15 | **15** | 415 x 230 | **4.052** | **2.312** | 2.57 | | |
| **Case-3 final** | Core | 23909 | **4653** | 520 x 475 | 11.86 | 17.44 | 18.2 | 395 | **14.70%** |
| | Mem | 0 | **0** | 415 x 230 | 4.579 | 3.264 | 2.57 | | |

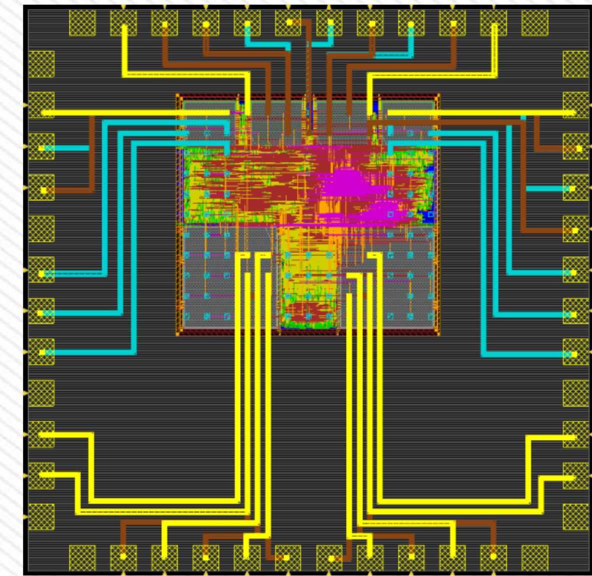
# Agile Customizations

- **Holistic flow offers flexible customizations**
  - Very little design effort


System with 16KB
380 MHz


Drop-In with 12KB
390 MHz

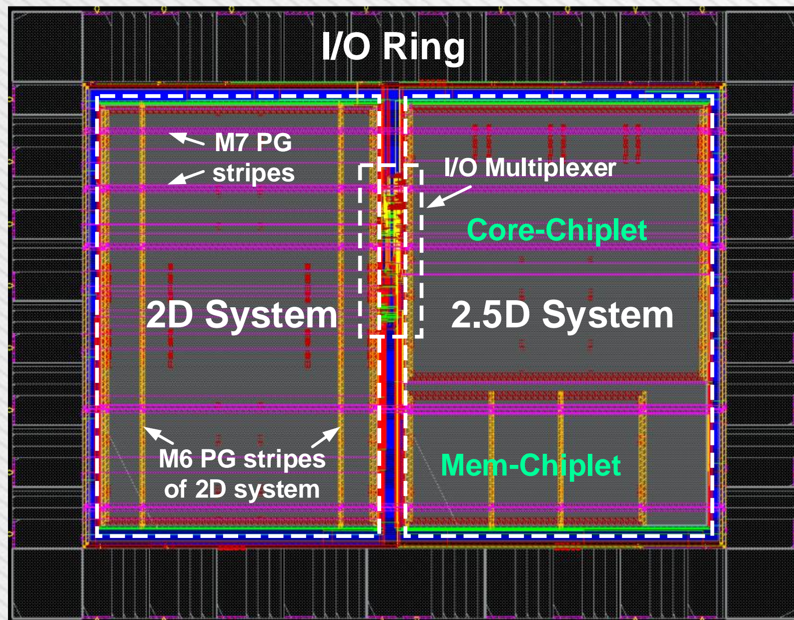

Pay-as-you-use with 12KB
396 MHz


Core-only with 8KB
400 MHz

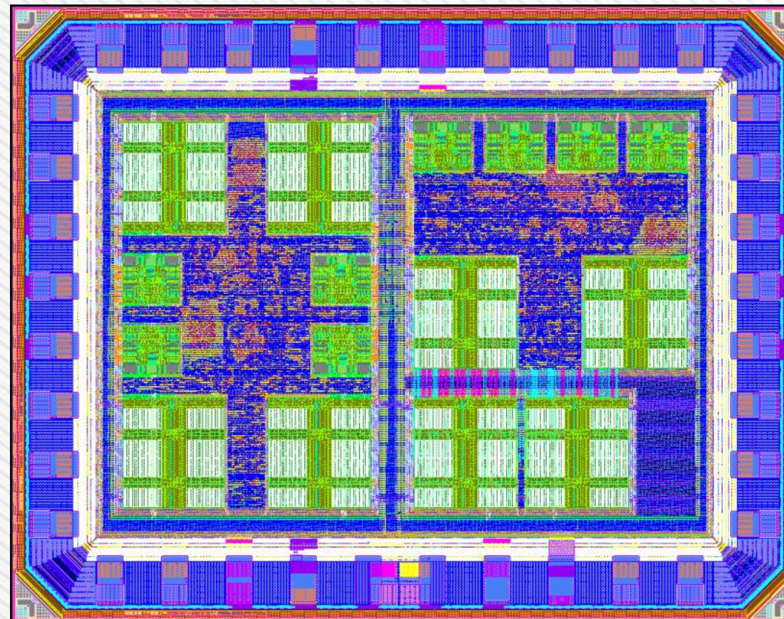
UNIVERSITY OF ARKANSAS

# Silicon Validation

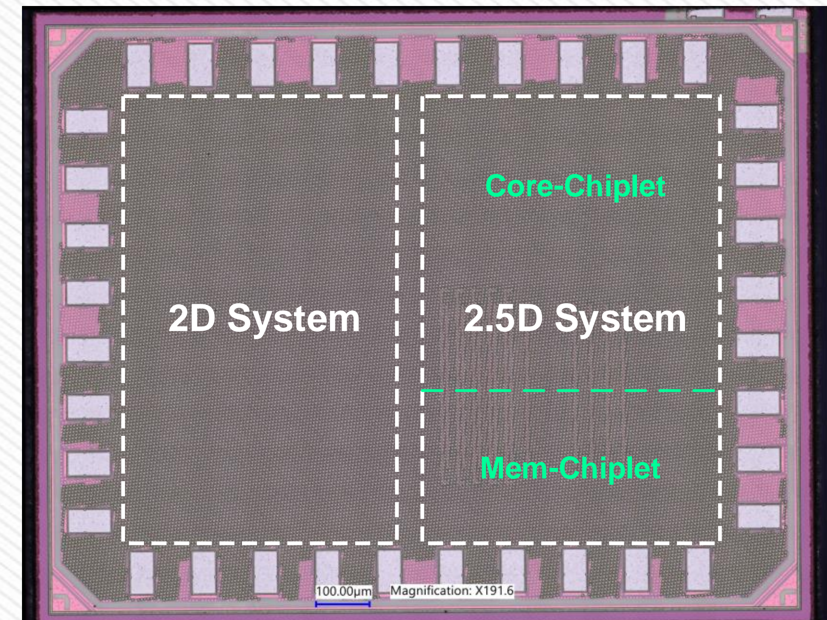
- **Dual system shared-block tape-out in TSMC 65**
  - Shares I/O system: I/O multiplexer module




(a) Die-level design in Innovus



(b) GDS for tapeout




(c) Microscopic die-shot

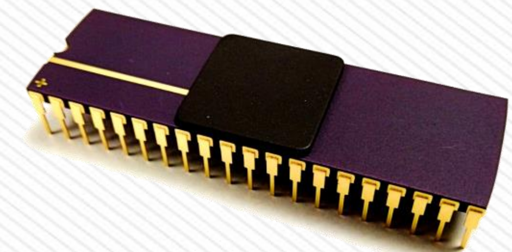
UNIVERSITY OF ARKANSAS

# Functional Verification

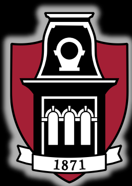
## ❑ Functional verification using logic analyzer



Testing waveforms at the logic analyzer



DIP packaged die



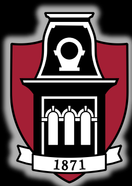
# In-Context Flow for Heterogeneous

- **Heterogeneous: Chips from different technology**

- **Holistic flow cannot handle heterogeneity with existing toolset**
  - Existing tools do not support heterogeneous tech. stack

- **In-Context design and analysis for heterogeneous systems**
  - Package planning with blackbox macros
  - In-context partition
  - Separate tech. stack for each partition

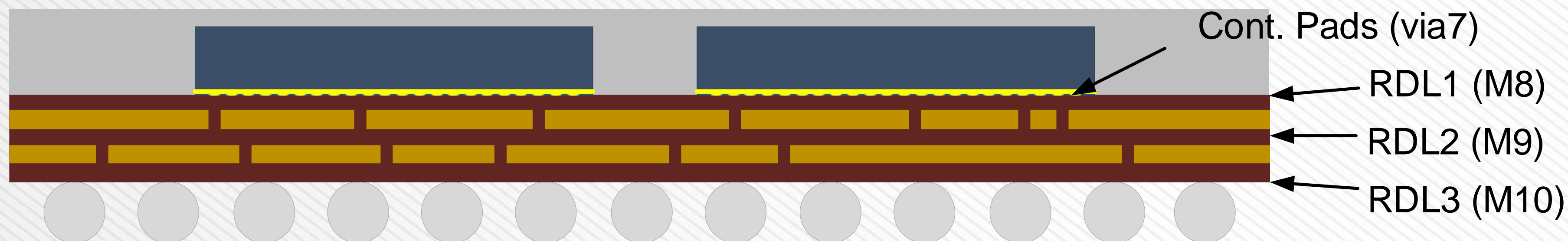
UNIVERSITY OF ARKANSAS

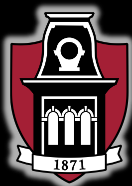


# Technology Setup for Case Study

- **Modified versions of Nangate45nm PDK**
  - 7M3R: 7 chip + 3 package
  - 6M3R: 6 chip + 3 package

| | M6 | via6 | M7 | via7 | RDL1 | viar1 | RDL2 | viar2 | RDL3 |
|---|---|---|---|---|---|---|---|---|---|
| **Height** | 2.28 | 3.08 | 3.9 | 7.5 | 12.5 | 17.5 | 22.5 | 27.5 | 32.5 |
| **Thickness** | 0.8 | 0.82 | 3.6 | 5 | 5 | 5 | 5 | 5 | 5 |
| **Width** | 0.4 | 0.4 | 2 | 5 | 10 | 10 | 10 | 10 | 10 |
| **Spacing** | 0.4 | 0.44 | 2 | 10 | 10 | 20 | 10 | 20 | 10 |

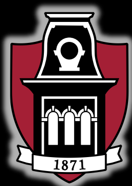




# Reference Holistic Designs in 45nm PDK

- **Holistic designs are re-implemented in the new setup**
  - For direct comparison with in-context designs
  - Using the 7M3R stack
  - Package overhead reduction by 63%

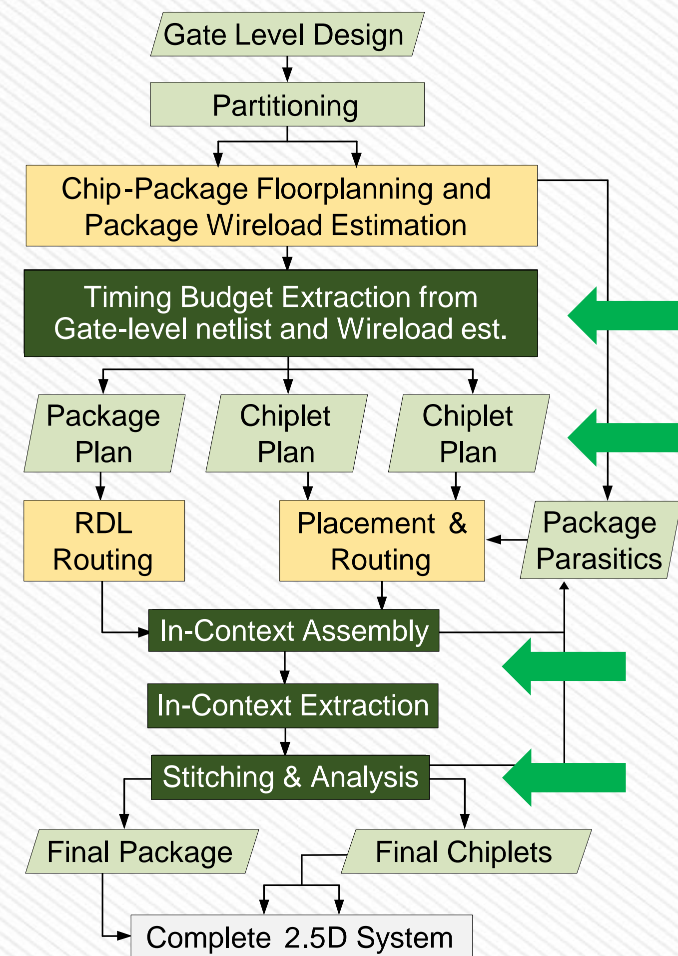
| Design Case | Chiplet Design | Logic Gates# | Buffer/ Inverter# | Die Size ($um^2$) | M6 WL (mm) | M7 WL (mm) | Power (mW) | Freq. (MHz) | Freq. Overhead |
|---|---|---|---|---|---|---|---|---|---|
| **Case-1** | 2D Chip | 17595 | 3700 | 550x550 | 79.94 | 0 | 10.6 | 333 | **0%** |
| **Case-2** | Core | 17783 | 2740 | 390x590 | 30.81 | 1.783 | 7.751 | 245 | **100%** |
| | Mem | 132 | 132 | 350x470 | 5.986 | 0.598 | 0.194 | | |
| **Case-3 initial** | Core | 17915 | 2865 | 390x590 | 31.86 | 1.875 | 9.043 | 280 | **60.23%** |
| | Mem | 148 | 148 | 350x470 | 8.201 | 0.589 | 0.216 | | |
| **Case-3 final** | Core | 18214 | 2955 | 390x590 | 31.42 | 2.02 | 9.840 | **300** | **37.50%** |
| | Mem | 45 | 45 | 350x470 | 8.445 | 0.624 | 0.162 | | |



# Per-Chiplet In-Context Flow

## ❑ Direct modification of the holistic flow

- Timing budgets from gate-level netlist
- Create package contexts
- In-Context assembly
- Extraction on in-context assembly
- Stitching parasitic netlist
- System-level analysis and optimization

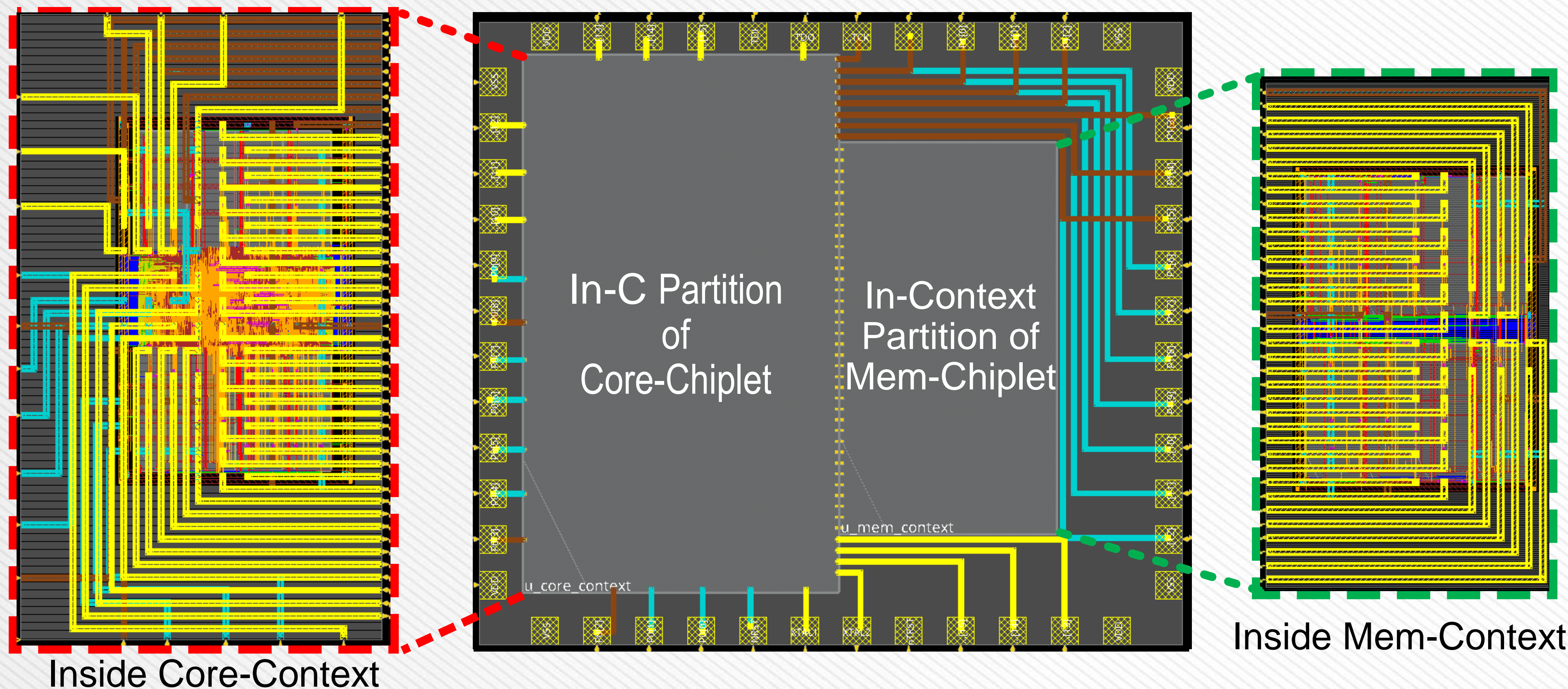



In-Context Flow

UNIVERSITY OF ARKANSAS

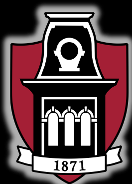
# In-Context Partitions

- An extra level in the design hierarchy for extended partition




Inside Core-Context

Inside Mem-Context



# Captures Cross-Boundary Coupling

## ❑ Extraction comparison

- All coupling captured like holistic
- Reasonable accuracy in coupling
- Overestimated ground cap
  - Fringe caps at cutting edges

Comparison of Extraction result w.r.t. Holistic

| | Metal Layer | M1-M5 | M6 | M7 | R1 | R2 | R3 |
|---|---|---|---|---|---|---|---|
| CCAP | Holi | 9172 | 1263 | 156 | 1544 | 2421 | 1721 |
| CCAP | InC | 9171 | 1265 | 153 | 1563 | 2489 | 1765 |
| CCAP | InC Err | -0.01% | 0.17% | -2.10% | 1.20% | 2.81% | 2.56% |
| GCAP | Holi | 21119 | 2054 | 272 | 1040 | 247 | 636 |
| GCAP | InC | 21119 | 2053 | 273 | 1103 | 306 | 696 |
| GCAP | InC Err | 0.00% | -0.01% | 0.09% | 6.03% | 24.0% | 9.46% |

## ❑ Performance comparison

- Effective iterative optimization
- Performance comparable to holistic implementations

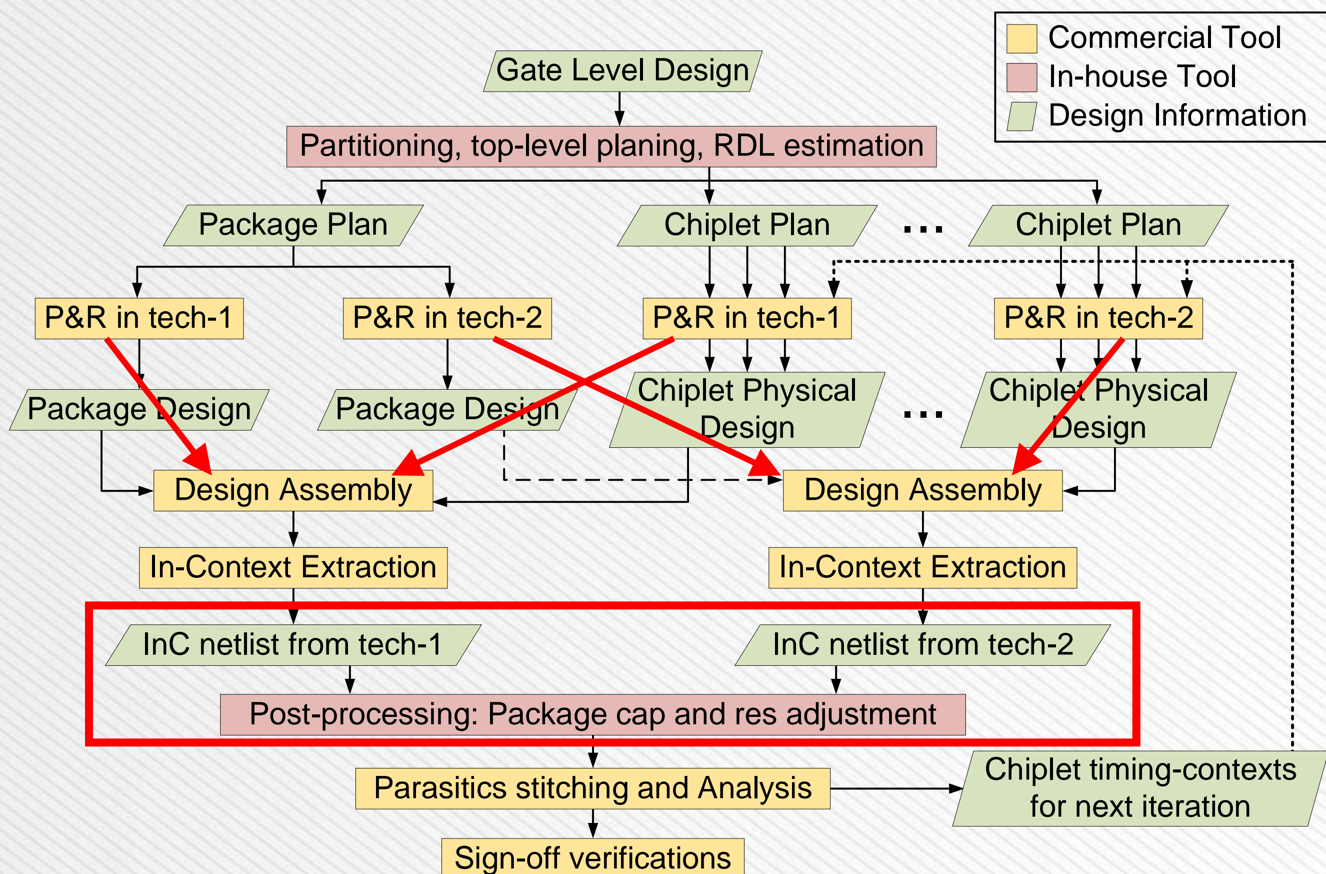
Iterative optimization result

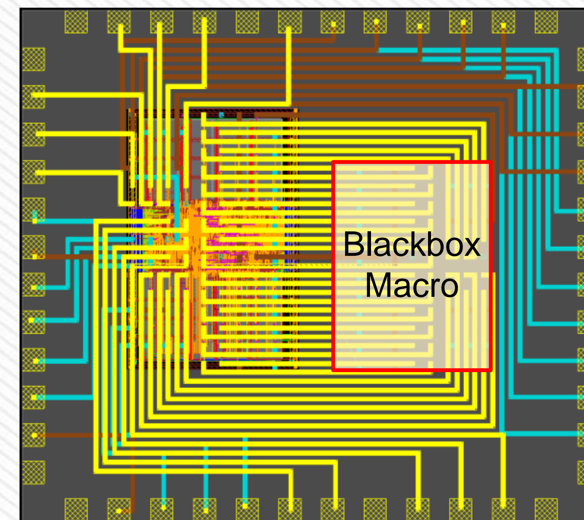
| Design iteration | LPD (ns) | In-C Perf | Holi Perf |
|---|---|---|---|
| with RDL wireload | 3.55 | 281 MHz | 280 MHz |
| In-Context 1st iteration | 3.35 | 298 MHz | - |
| In-Context 2nd/final | 3.35 | 298 MHz | 300 MHz |

UNIVERSITY OF ARKANSAS

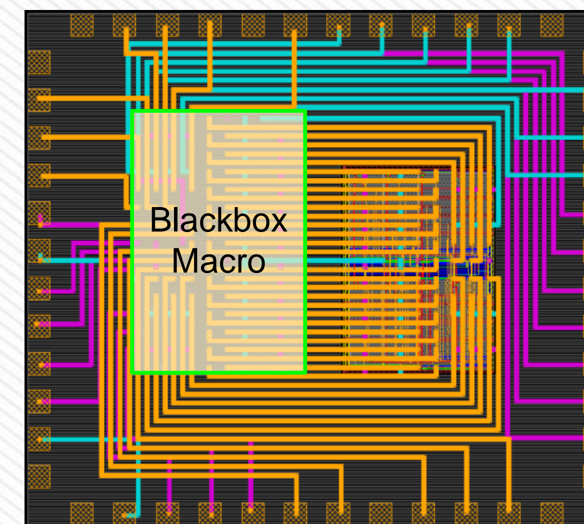
# Per-Technology In-Context Flow

- Avoid cutting the package
  - Assemble all chiplets of same technology
  - Post-processing to fix double-counting



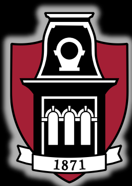




(a) Assembled Core-Context (7M3R)




(b) Assembled Mem-Context (6M3R)

UNIVERSITY OF ARKANSAS



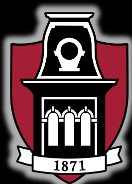
# Post-Processing Methodology

- **Package layer cap is reduced by a fraction of top-only extraction numbers**
  - Top-only extraction: all chiplets as blackbox
  - CapRDL: RDL cap from in-context extraction
  - TCapRDL: Extraction on package only
  - userFact: provided by the designer
  - Cap nodes of a net multiplied with $layerFact_x$ of that layer

$$layerFact_x = \frac{CapRDL_x - userFact \times TCapRDL_x}{CapRDL_x} \quad (1)$$

$$newNodeCap = nodeCap \times layerFact_x \quad (2)$$

UNIVERSITY OF ARKANSAS



# Improved Extraction Accuracy

## ❑ Extraction comparison

- All coupling captured like holistic
- Very high accuracy: 100% approx.
  - GCAP and CCAP

## ❑ Performance comparison

- Homogen: 7M3R + NG
- Heterogen with two stacks and lib
  - Core: 7M3R + NG
  - Mem: 6M3R + GSCL (FreePDK)

## ❑ Major concerns

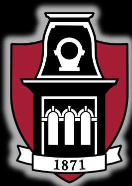
- Scalability
- Empirical param: userFact

Comparison of Extraction result w.r.t. Holistic

| | Metal Layer | M1-M5 | M6 | M7 | R1 | R2 | R3 |
|---|---|---|---|---|---|---|---|
| GCAP | Holi | 21605 | 2161 | 284 | 1032 | 219 | 513 |
| | InC | 21605 | 2162 | 284 | 1034 | 220 | 513 |
| | Err (per-tech) | 0.00% | 0.00% | 0.01% | 0.24% | 0.6% | 0.00% |
| | Err (per-chip) | 0.00% | -0.01% | 0.09% | 6.03% | 24.0% | 9.46% |
| CCAP | Holi | 8988 | 1292 | 203 | 1553 | 2412 | 1648 |
| | InC | 8989 | 1291 | 202 | 1553 | 2412 | 1648 |
| | Err (per-tech) | 0.00% | 0.04% | 0.64% | 0.03% | -0.01% | 0.00% |
| | Err (per-chip) | 0.01% | 0.17% | -2.10% | 1.20% | 2.81% | 2.56% |

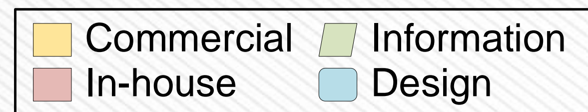
Iterative optimization result

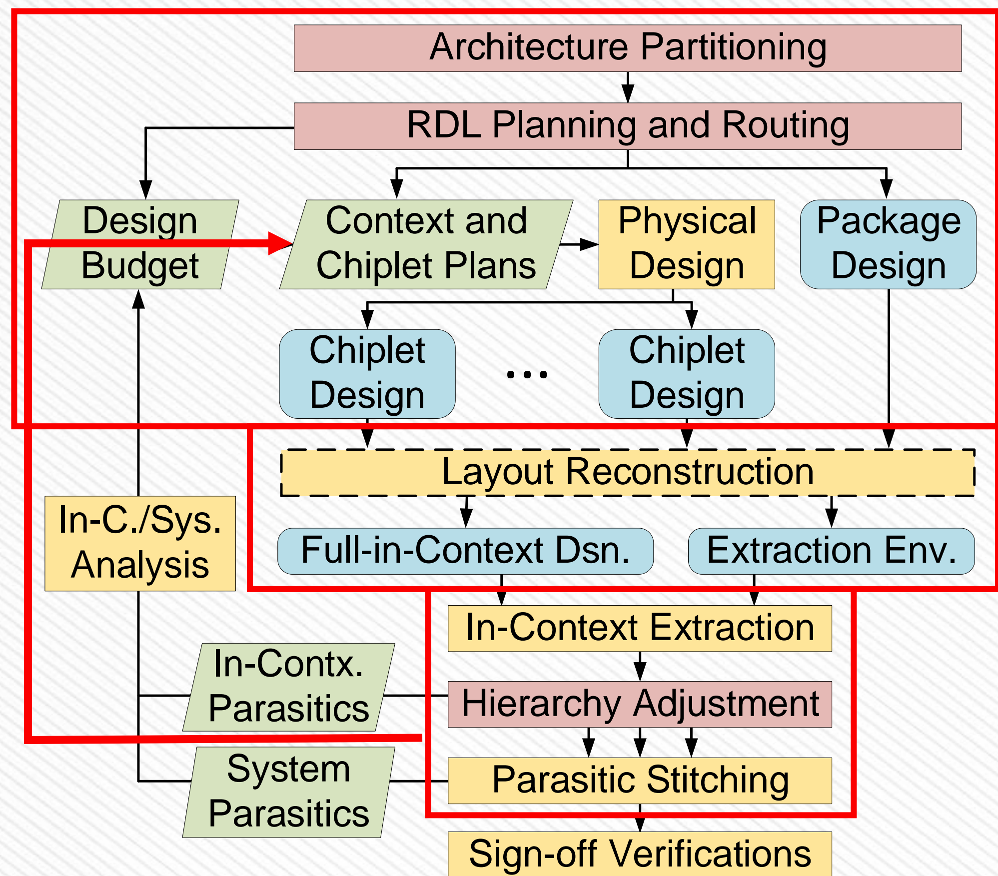
| Design | Homogeneous | | Heterogeneous |
|---|---|---|---|
| Iteration | Holistic | In-Context (per-tech) | |
| Initial | 288 MHz | 288 MHz | 287 MHz |
| 1st iteration | 293 MHz | 294 MHz | 294 MHz |
| 2nd/final iteration | 300 MHz | 300 MHz | 300 MHz |



# Timing-Accurate In-Context Flow

- Takes advantage of the flip-chip extraction flow to perform in-context extraction
  - Planning and physical design: previous flows
  - Layout reconstruction
    - Not cutting the package
    - Not extracting the entire package
  - in-context extraction on each chiplet
  - Hierarchy adjustment before parasitics stitching
  - In-C/Sys. Analysis and verification
  - Iterative optimization
  - Sign-off verifications



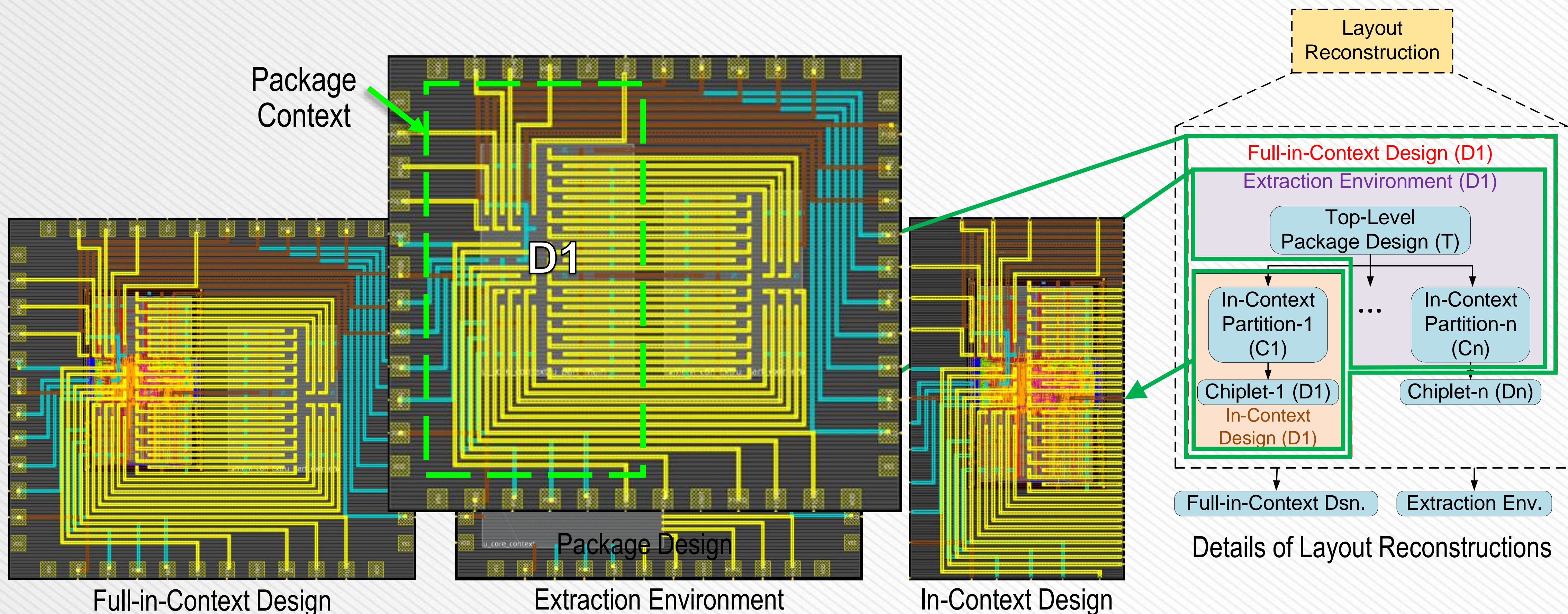




Overall Flow

UNIVERSITY OF ARKANSAS

# Layout Reconstruction for In-Context Extraction

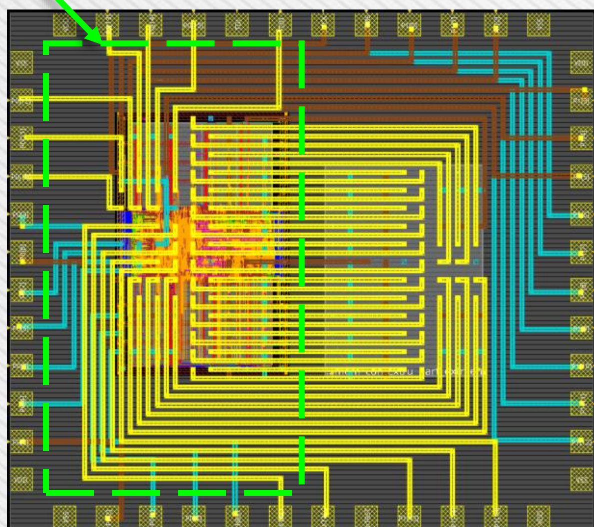
- Generates design files to perform extraction within a chiplet context



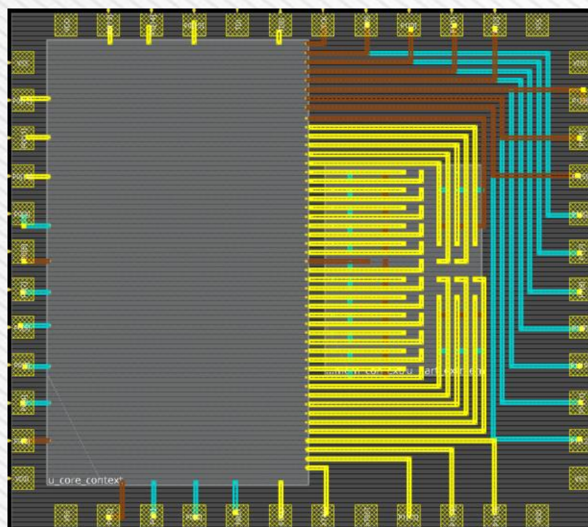
# Layout Reconstruction for In-Context Extraction

- Generates design files to perform extraction within a chiplet context
  - Extraction on the *full-in-context* design
  - Coupling converted to ground caps at the boundary

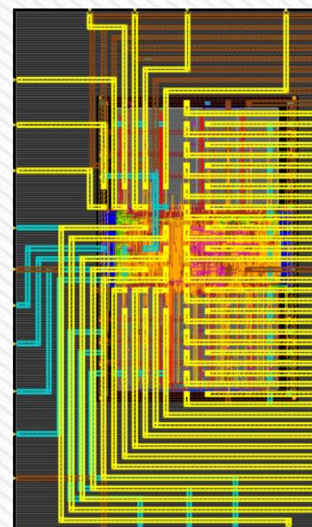
Extraction target



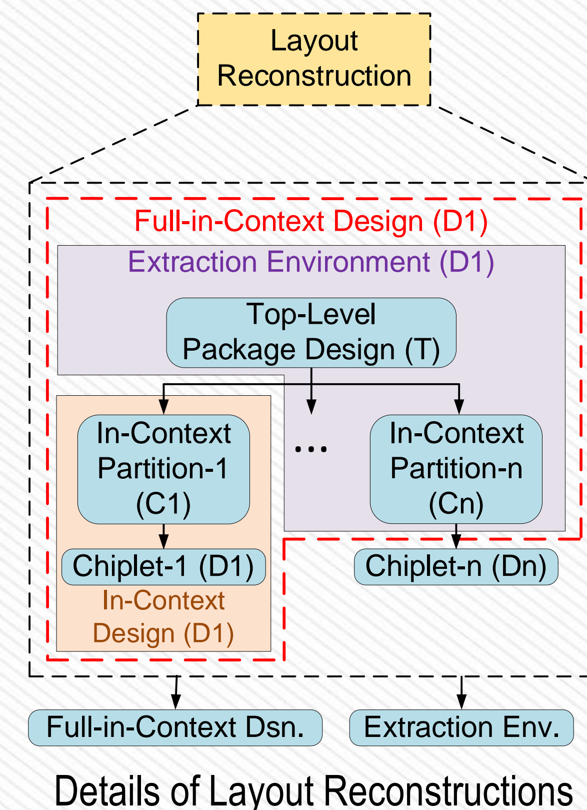
Full-in-Context Design



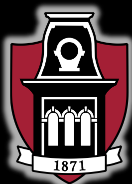

Extraction Environment



In-Context Design



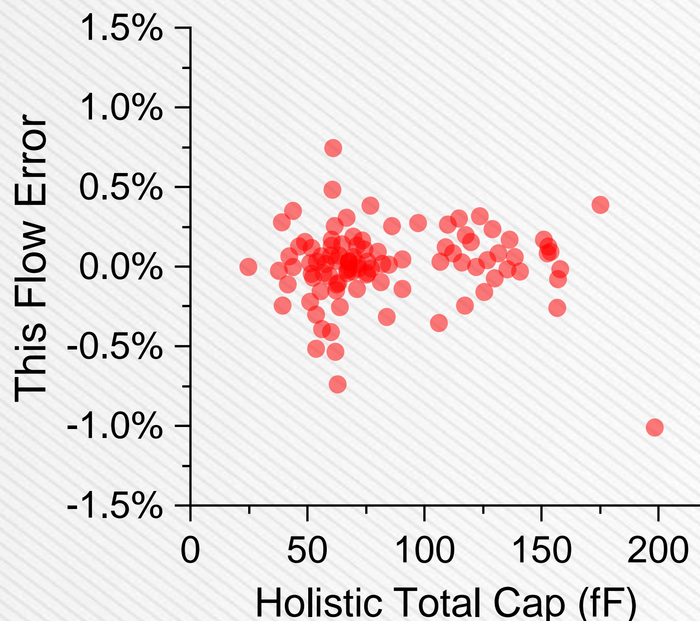

Details of Layout Reconstructions

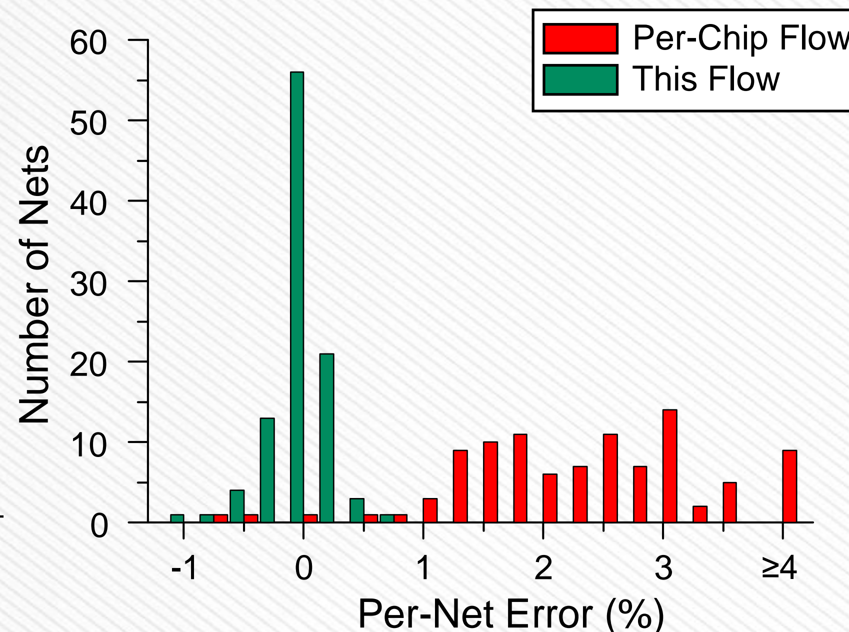


# Accurate Total Capacitance

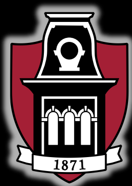
- **Extraction comparison**
  - **Degraded coupling accuracy**
  - **high accuracy in total cap**
    - Within +/-1%
  - **Net delay depends on total cap**

| | Metal Layer | M1-M5 | M6 | M7 | R1 | R2 | R3 |
|---|---|---|---|---|---|---|---|
| CCAP | Holi | 9275 | 1172 | 196 | 1529 | 2441 | 1685 |
| | InC | 8992 | 1203 | 193 | 1517 | 2390 | 1640 |
| | Err (tim-acc) | -3.05% | 2.65% | -1.53% | -0.78% | -2.09% | -2.67% |
| | Err (per-chip) | 0.77% | 0.77% | -4.08% | 2.29% | 1.52% | 0.30% |
| Total CAP | Holi | 31056 | 3307 | 498 | 2547 | 2669 | 2209 |
| | InC | 31238 | 3350 | 495 | 2591 | 2654 | 2192 |
| | Err (tim-acc) | 0.59% | 1.31% | -0.59% | 1.74% | -0.55% | -0.76% |
| | Err (per-chip) | 0.27% | 0.51% | -1.79% | 4.49% | 3.01% | 1.91% |




UNIVERSITY OF ARKANSAS

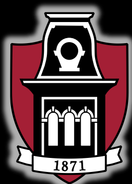


# In-Context Flow Comparison

- Each version has unique strength and weakness

| Flow version | Accuracy | Scalability | Flow Complexity |
|---|---|---|---|
| Per-Chiplet | Worst | High | Simplest |
| Per-Technology | Best | Low | Intermediate |
| Timing Accurate | Good | High | Complex |

- Can be unified into a single framework
  - Per-chiplet: for estimation
  - Timing accurate: distributed design with margin
  - Holistic or per-technology flow: final iteration and sign-off



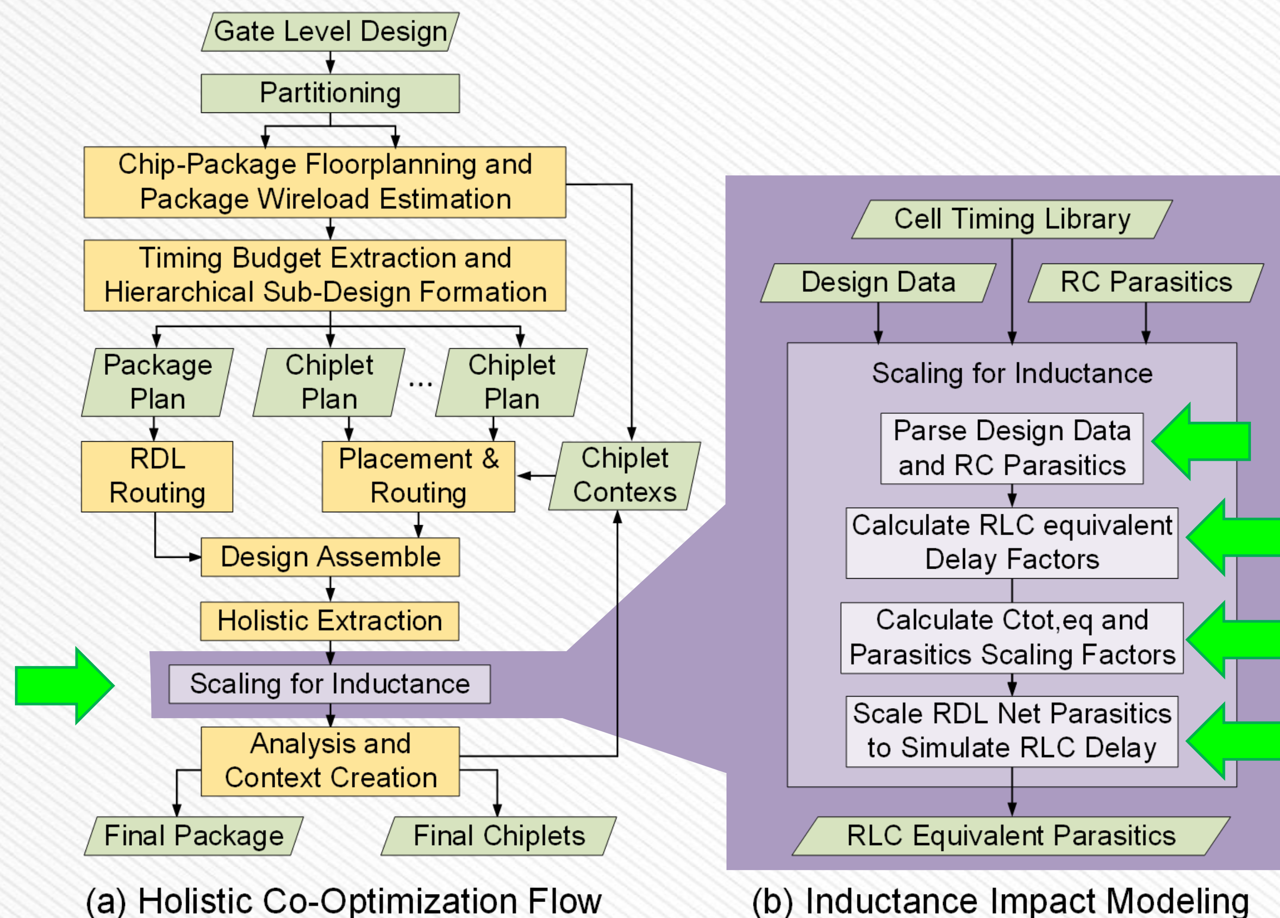
# Inductance-Aware Flow

## ❑ Represent RLC equivalent delay using RC parasitics

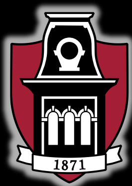
- RC scaling
- STA tools don't support inductance

## ❑ RC scaling flow

- Read design info
- Calculate RLC delay
- Scaling factor = RLC-delay / RC-delay
- Net caps scaled




(a) Holistic Co-Optimization Flow
(b) Inductance Impact Modeling



# RC Parasitic Scaling for Inductance

- **RLC equivalent parasitics is computed using equation (3)**
  - Cell delay: input transition, total output capacitance
  - Net delay: Elmore delay model

$C_{tot,eq} / C_{tot}$

$$\begin{aligned} RLC\ delay &= cell\ delay + net\ delay \\ &= LUT\ (C_{tot,eq}\ ,\ t_r\ ) + scalePar \times (RC\ net\ delay) \end{aligned} \quad (3)$$

Where,

$C_{tot}$ : Total Capacitance in the RC network,

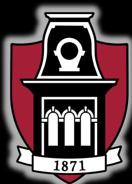
$t_r$ : Input transition time of the driver cell,

$C_{tot,eq}$ : Equivalent total capacitance required to simulate RLC delay,

$LUT$ : Cell timing library look-up table

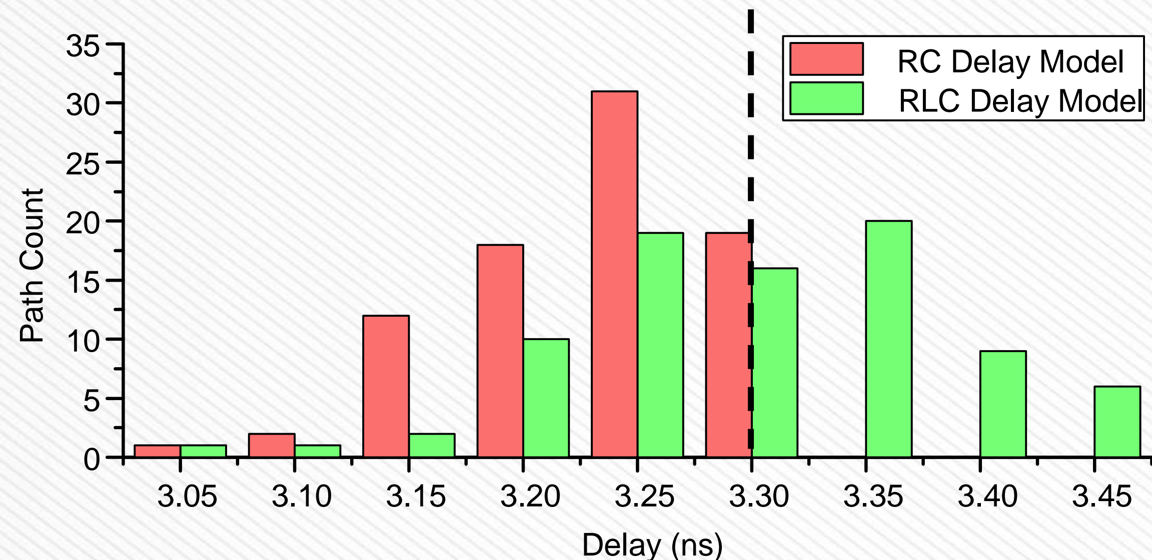
$scalePar$ : $C_{tot,eq} / C_{tot}$

UNIVERSITY OF ARKANSAS

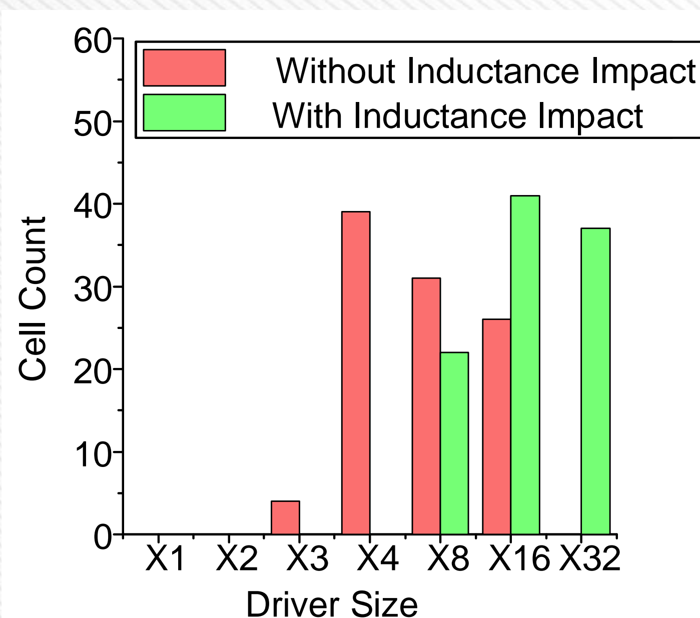


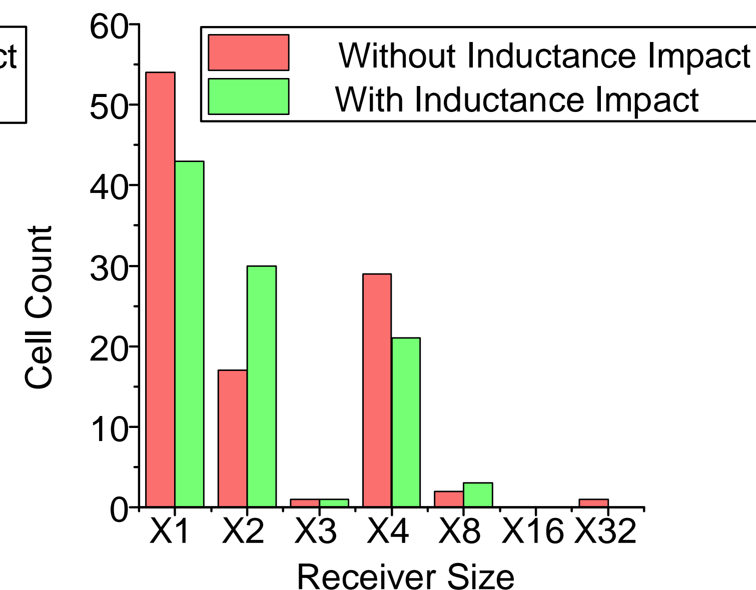
# Automatic Driver and Receiver Optimization

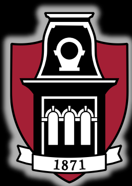
- In RC analysis violations goes undetected
  - 35% of the paths in timing violation
  - The worst violation is by 0.15 ns




- Automatic optimization
  - Upsized drivers
  - Downsized receiver load



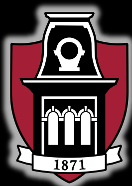






# Conclusions

- **Chiplet-Package interactions are significant in 2.5D systems**
- **Presented flows effectively captures the interactions in analysis & optimization**
  - Enables holistic planning and optimizations
  - Can be used as reference flows
- **Inductance-aware system-level optimization is necessary**
  - RC scaling is one way to achieve it

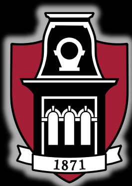
UNIVERSITY OF ARKANSAS




# Future Work

- ❑ Study the impact of these flows on advanced and/or diverse technologies
- ❑ Unify all of their unique feature into a single framework
- ❑ Study signal and power integrity with all RCLM elements
- ❑ Chiplet-Package co-placement, routing, and optimizations
- ❑ System performance and SI-aware package design

UNIVERSITY OF ARKANSAS



# Publications

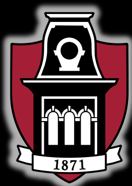
## ❑ Journal

1. **MD Arafat Kabir** and Yarui Peng, "Holistic Chiplet-Package Co-Optimization for Agile Custom 2.5D Design", IEEE Transactions on Components, Packaging, and Manufacturing Technology (TCPMT), 2021. (**IF: 2.04**)

## ❑ Conferences

1. **MD Arafat Kabir** and Yarui Peng, "Chiplet-Package Co-Design For 2.5D Systems Using Standard ASIC CAD Tools", Asia and South Pacific Design Automation Conference (**ASPDAC**), 2020. (**Acc. Rate: 32.6%**)
2. **MD Arafat Kabir** and Yarui Peng, "Holistic 2.5D Chiplet Design Flow: A 65nm Shared-Block Microcontroller Case Study", IEEE International System-on-Chip Conference (**SoCC**), 2020. (**Acc. Rate: 30.1%**)
3. **MD Arafat Kabir,** Dusan Petranovic, and Yarui Peng, "Coupling Extraction and Optimization for Heterogeneous 2.5D Chiplet-Package Co-Design", International Conference on Computer-Aided Design (**ICCAD**), 2020. (**Acc. Rate: 27%**)
4. **MD Arafat Kabir**, Dusan Petranovic, and Yarui Peng, "Cross-Boundary Inductive Timing Optimization for 2.5D Chiplet-Package Co-Design", ACM Great Lakes Symposium on VLSI (**GLSVLSI**), 2021. (**Acc. Rate: 27%**)
5. **MD Arafat Kabir**, Weishiun Hung, Tsung-Yi Ho, and Yarui Peng, "Holistic and In-Context Design Flow for 2.5D Chiplet-Package Interaction Co-Optimization", International Symposium on VLSI Design, Automation and Test (**VLSI-DAT**), 2021, **Invited Paper**.
6. **MD Arafat Kabir**, Dusan Petranovic, and Yarui Peng, "A Scalable In-Context Design and Extraction Flow for Heterogeneous 2.5D Chiplet-Package Co-Optimization", (**accepted**) IEEE Conference on Electrical Performance of Electronic Packaging and Systems (**EPEPS**), 2021.

UNIVERSITY OF ARKANSAS



# Thank You

## Questions?

UNIVERSITY OF ARKANSAS