From Vision To Reality : SiMa.ai's Physical Design Journey with Synopsys

Shibashish Patel, Jignesh Shah, Ajit Kumar, Vishal Katba
SiMa.ai

SiMa.ai at a Glance

We are a **software** company that is *building our own* ***silicon***

Amazing team: First silicon to production with off-the shelf boards.

Industry's **best ML software** in production for desktop and cloud

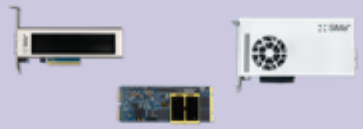
Goal: **No code visual development** ML environment

Scale ML at the Embedded Edge

Effortless ML Customer Experience

ANY. 10X. Pushbutton

Palette™ Software



Boards

MLSoC Silicon

SiMa.ai at a Glance

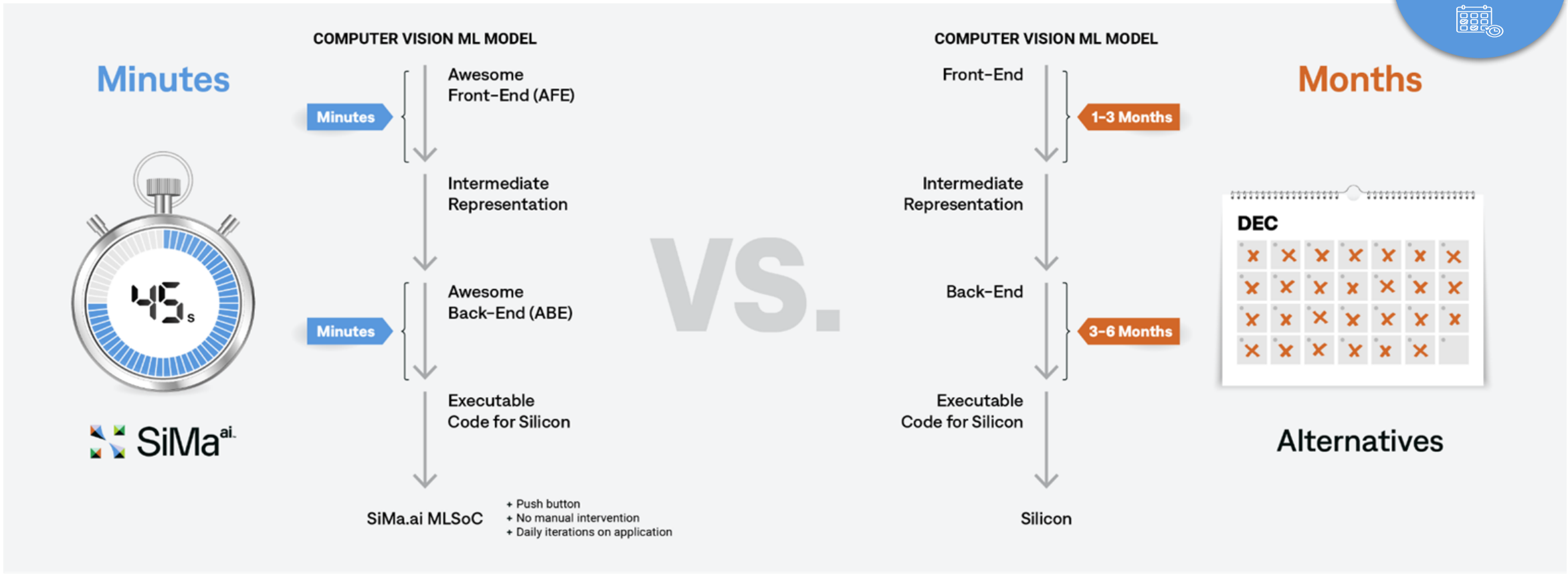
SiMa.ai focus: Embedded edge market

SiMa.ai at a Glance

Software schedule value proposition

Customer challenge: *I must accelerate my design to meet product schedule*

SiMa.ai's key differentiator: Model compiler, pushbutton build, pushbutton deploy



SiMa.ai at a Glance

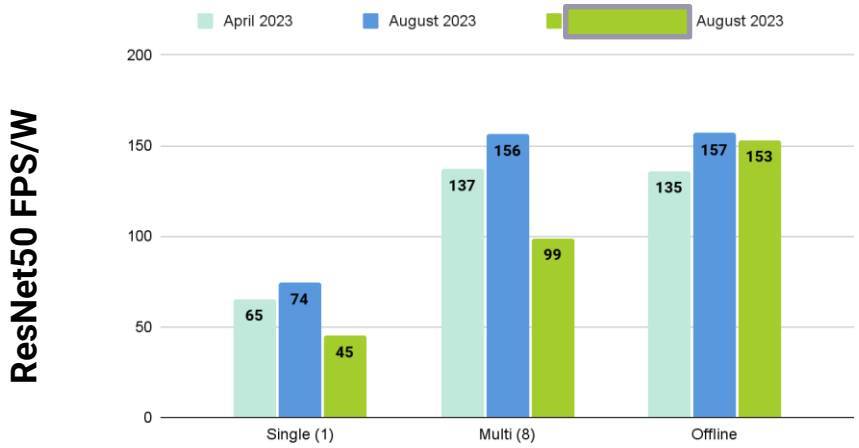
Power value proposition

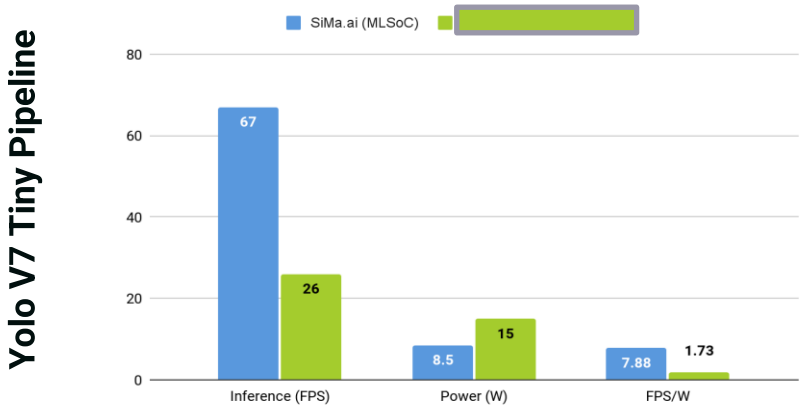
Customer challenge: *I have size, weight or power constraints in my design*

Power

SiMa.ai's key differentiator:

- **40% better FPS/watt** on MLPerf benchmark than hand coded *competitor*
- **2-4x better FPS/watt** than *competitor* compiled ML models (450% YoloV7 tiny)





SiMa.ai at a Glance

Performance value proposition

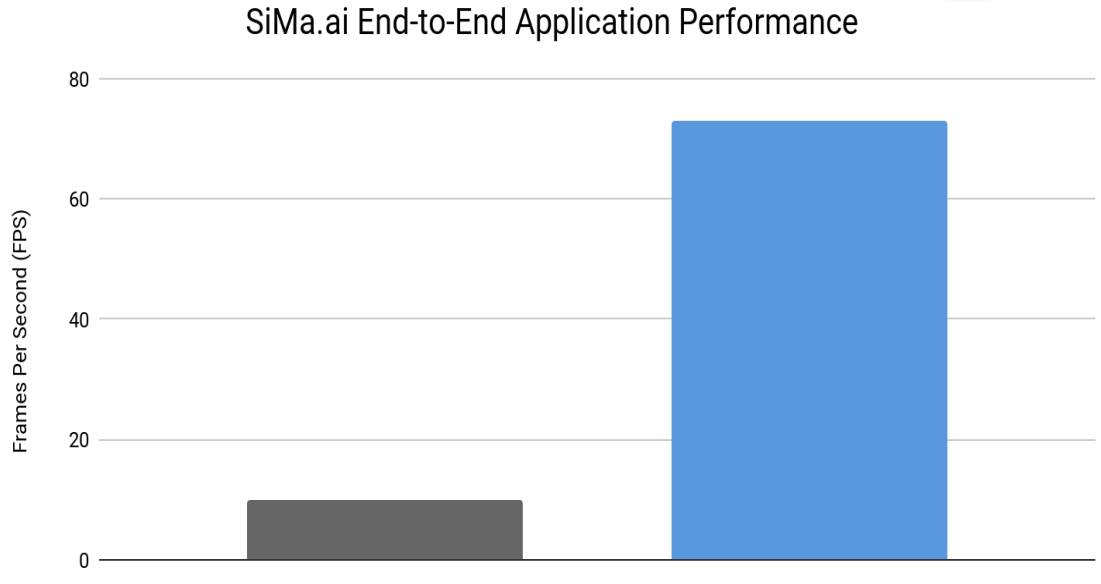
Customer challenge: *I have high FPS and low latency ML pipeline in my design*

SiMa.ai's key differentiator:

- **10x** faster response time
- **12x** faster end-to-end pipeline FPS than PCIe ML accelerator

FPS + latency

10x faster control loop latency in microseconds, not milliseconds



Previous ML Accelerator Design (competitor) PCIe+FPGA + x86 server (ML ~25 watts)

MLSoC 8.5 watts

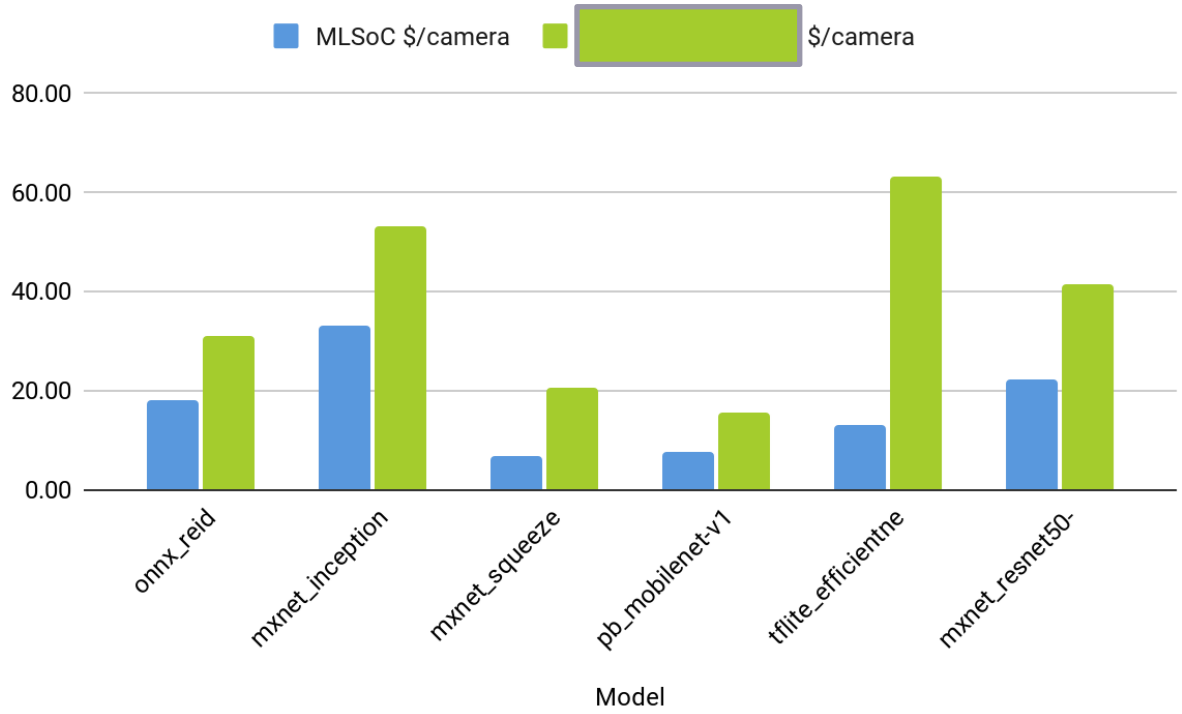
SiMa.ai at a Glance

Cost value proposition

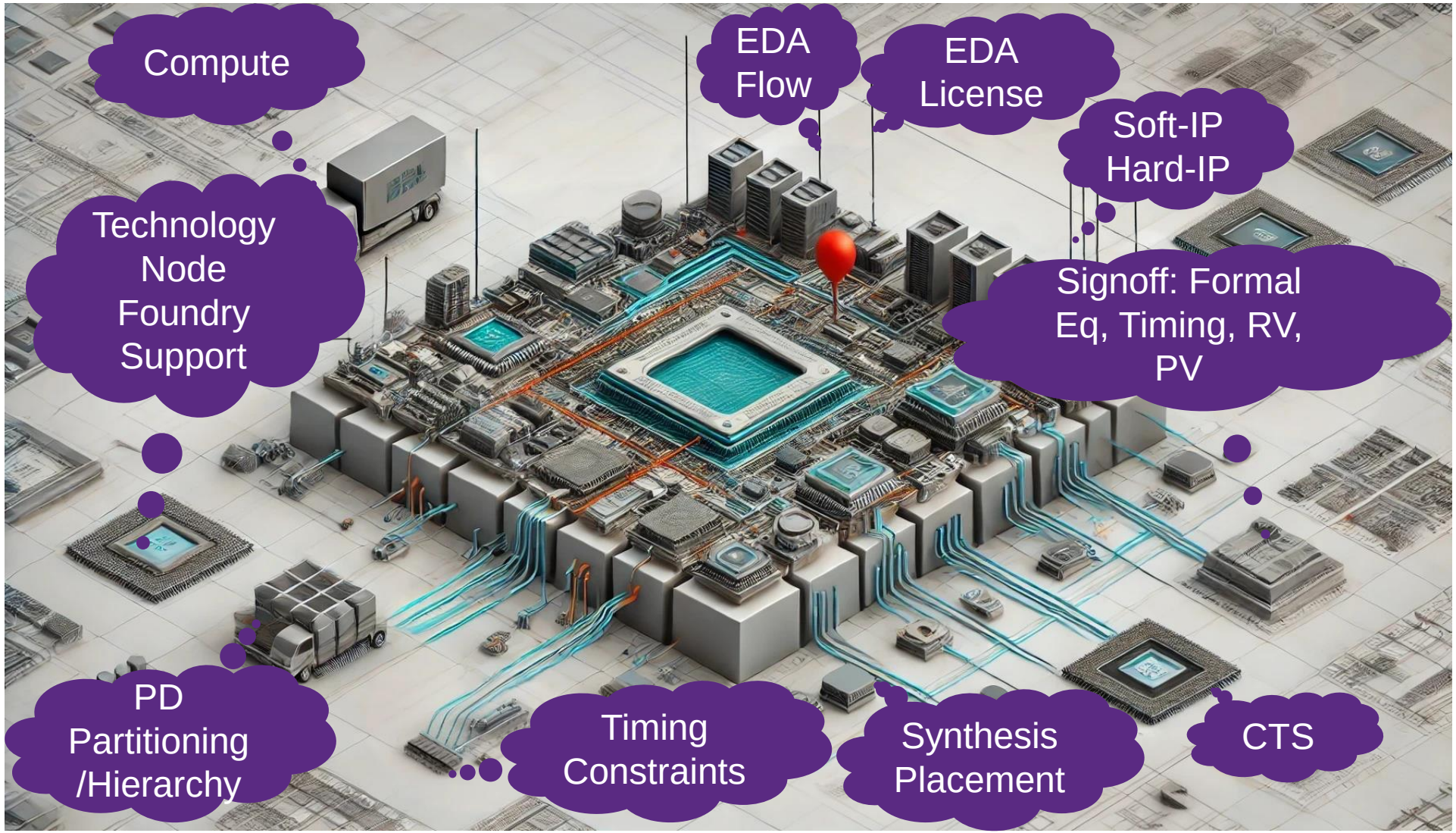
Customer challenge: *I have a low cost target in my design*

SiMa.ai's key differentiator: **50%** lower $/camera stream than *competitor*

MLSoC Dual M.2 16GB card
Competitor 16GB SOM

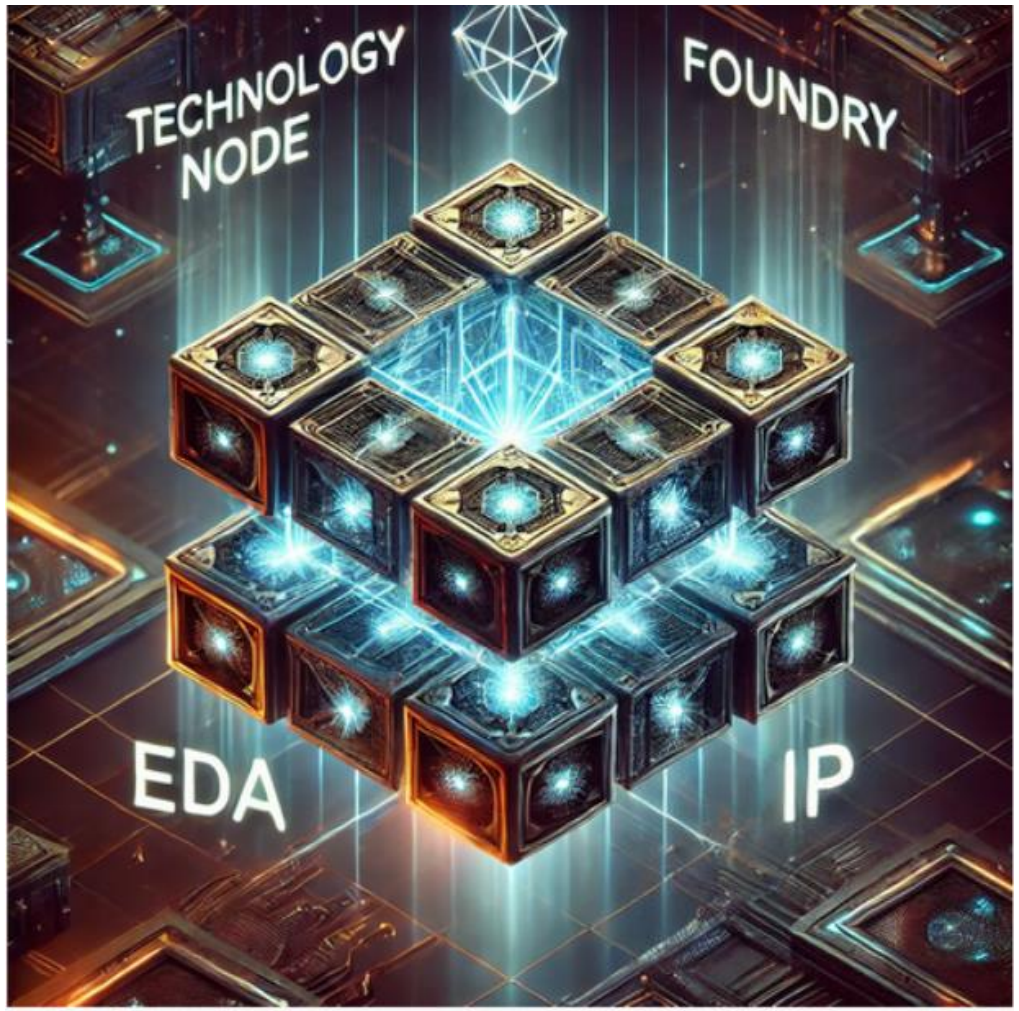


Challenges As A Startup In R2G Space



Complex Landscape Needs: Robust Solutions + Excellent Support Systems

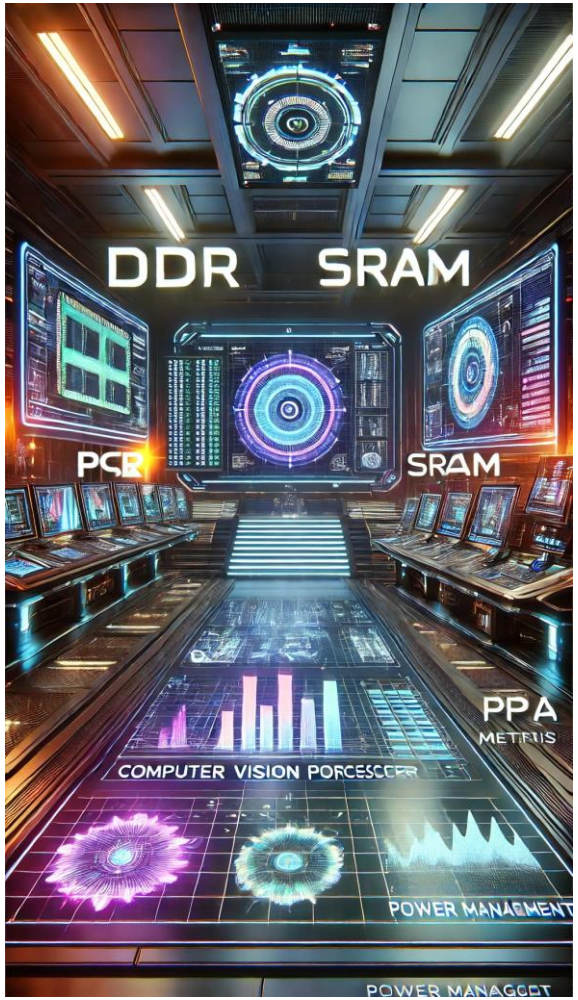
Technology Node ⇔ Foundry ⇔ IP ⇔ EDA



- Choosing the right technology node
- Foundry-qualified EDA tool versions & signoff settings, known limitations/workarounds
- IP availability
- IP qualification metrics

- Bottom line
  - Simultaneous decision-making needed with multiple variables

Soft-IP & Hard-IP

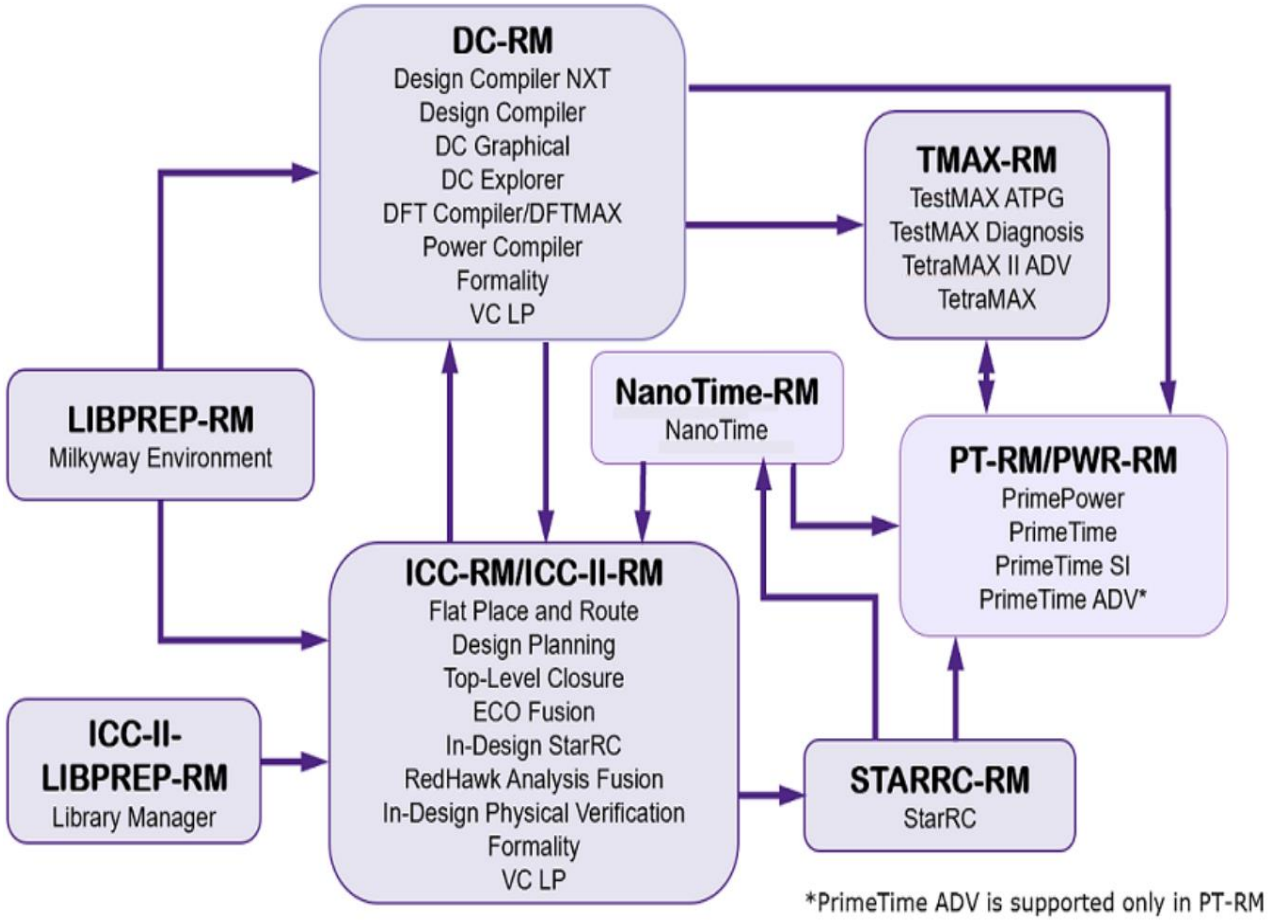


- Key Decision Making
  - Soft/Hard IP Licensing, IP Hardening, In-house development
  - Feature support like power management etc.
- Validation & Integration Know-Hows
  - IP PD Integration Guide
  - PPA Metrics details: Synopsys EV74 IP had complete among all IPs
- Watch out for
  - Cost vs Schedule
  - Feasibility: IP integration with different features
  - PPA metrics provided by IP vendor

EDA License & Compute Requirements

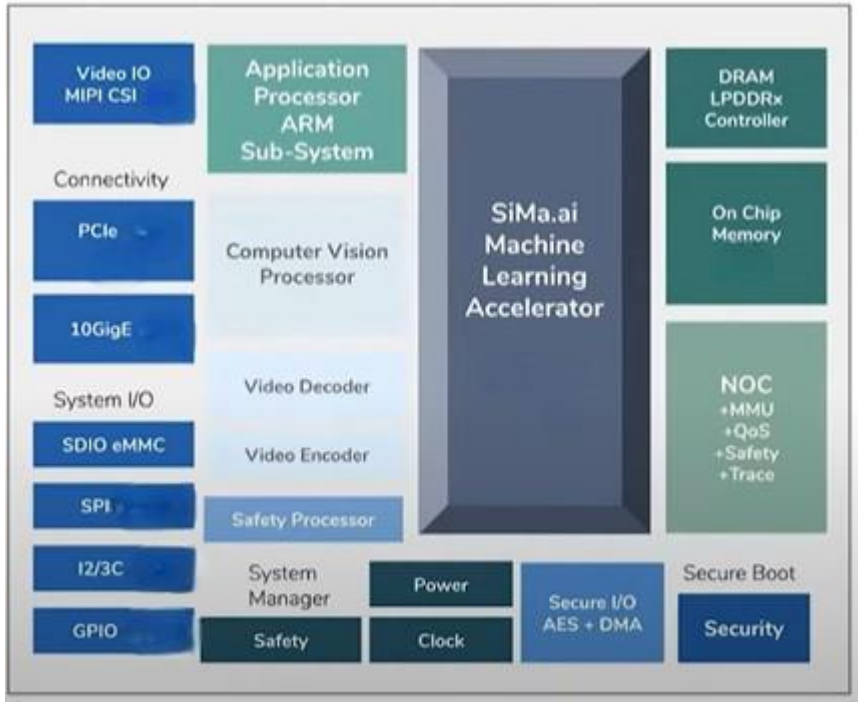
- Deciding Factors
  - Schedule vs Compute vs License
- Tool Readiness checks
  - EDA tool features & required licenses
  - Compute cores support per license
  - License hold & release during staggered execution
- Solution:
  - Focussed approach backed by analysis *to limit the trials*
  - *Hiring* & *To-Be-Hired* mindset

EDA Methodology



- Characteristics of good EDA flow:
  - Efficiency
  - Accuracy
  - Repeatability
  - Scalability
  - Integration
  - Flexibility
- RMgen adaption from Synopsys
  - Easy to download & configure
  - Customizable
  - Worked out-of-the-box for most of the blocks
- Important Implication : Next slide

Physical Design Partitioning/Hierarchy



- Limiting & Deciding parameters
  - Schedule/runtimes, Licenses, Compute
  - Constraints Management
  - Additional interface timing closure
  - Clock budgeting/balancing
- Design X flat implementation for faster TAT
  - set_app_options -name extract.starrc_mode with -val none instead of -val fusion_adv
- Remember this
  - Good EDA flow helps in quicker decision-making
  - Tool app options : Correlation tradeoff vs schedule

Constraining the Timing Constraints

- Key considerations:
  - Constraints development: Bottom-up or top-down
  - Constraints coding styles & integration.
    - A combination of TCL format and SDC
  - Constraints quality signoff
    - Combination of GCA & custom quality checks
  - Paranoia checks
    - -from , -through , -to switch usage for every exception definition
    - RTL design integrator and/or IP vendor review

Synthesis/Placement QoR vs Runtime

- Synopsys team actively worked with SiMa.ai PD teams on runtime & QoR improvements. Reference design snapshot:
  - Overall better convergence with new switches with minor impact on runtime
  - With additional 30 minutes of route_detail –incremental benefits in DRCs 196 (20 shorts) without impacting timing (-0.17/-48.34/604)
  - RM flow switch: High effort congestion switch at placement was increasing runtime

Stage: Route_opt	High Effort True	High Effort False	High Effort False + Opt
Shorts/DRCs	8/369	601/2176	87/686 → PRDI 20/196
Total Power	21.18	20.827	20.225
WNS	-0.1687	-0.197	-0.1706
TNS	-48.6245	-54.4816	-47.9718
NUM	801	782	463
Total R2G Runtime	94.4 hrs	77 hrs	84.8 hrs

Formal Verification using Formality

- Follow IP Physical Integration guidelines
  - Logic preservation and formality-related recommended switches
- Multiple challenges of
  - Hard verification, SVF guide rejection, bad logic optimization & wrong guide merging
- Solutions
  - SVF hacks & workaround
  - Close collaboration with Synopsys R&D to provide native fixes for corner cases in Fusion Compiler & Formality
- Pro Tip: Engage early, and parallelize solution finding.

MSCTS Methodology Development

- Synopsys successfully demonstrated MSCTS technology improving overall Latency, Timing & overall TAT
  - **Design A** to reduce latency from **1.3ns to 0.9ns** without impact on timing
  - Guided for **Design B** to build custom MSCTS
  - **Design C to improve latency, TNS & FEP**

- Game Changer :
  - Interface timing closure as well as internal timing improvement by latency reduction
  - 30% latency reduction
  - 2x TNS & FEP reduction

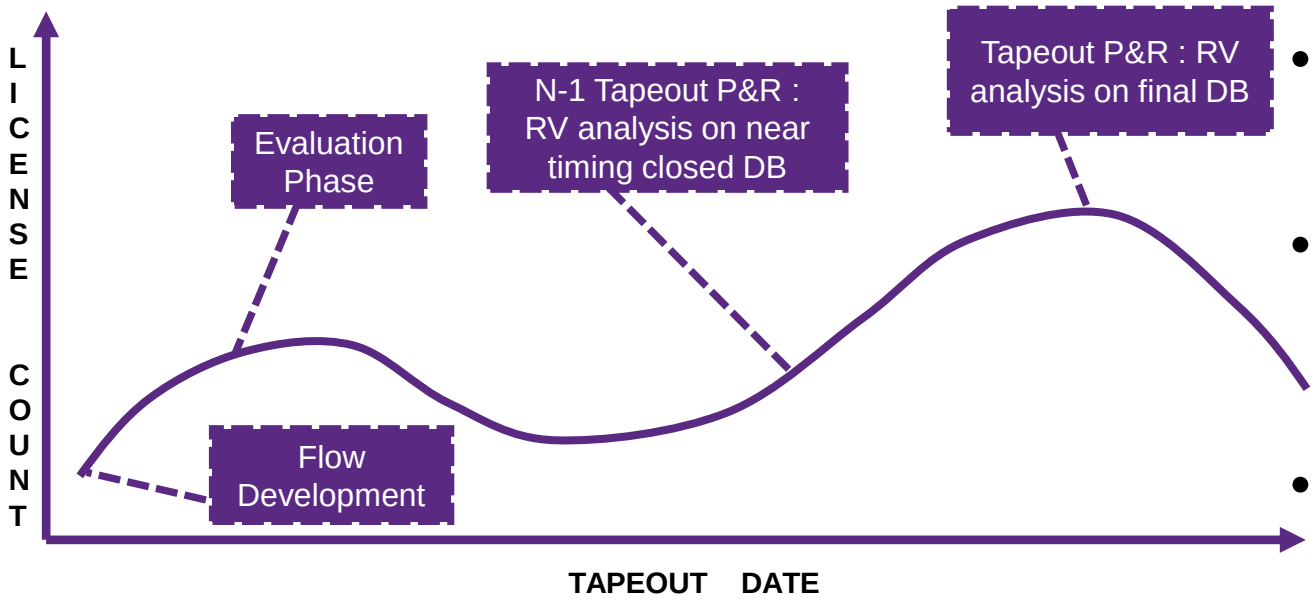
Setup Violations			Normal CTS	
	Total	R2R	I2R	R2O
WNS	-0.831	-0.43	-0.802	-0.831
TNS	-2216.71	-748.807	-795.395	-672.507
NUM	18239	11052	4727	2460
Hold Violations				
	Total	R2R	I2R	R2O
WNS	-0.269	-0.269	0	0
TNS	-97.059	-97.059	0	0
NUM	6779	6779	0	0

Setup Violations			MSCTS	
	Total	R2R	I2R	R2O
WNS	-1.1602	-0.4949	-1.1602	-0.64
TNS	-6458.43	-205.918	-6230.78	-21.73
NUM	14457	4164	10176	117
Hold Violations				
	Total	R2R	I2R	R2O
WNS	-0.0832	-0.0832	-0.0288	0
TNS	-55.48	-54.6	-0.884	0
NUM	3580	3283	297	0

Extraction & Timing Signoff

- Spef-stitch methodology adopted
  - With marginal miscorrelation (< -30ps Setup) on interface timing, leveraged faster TAT
  - Miscorrelation was mainly identified on nets that had not adhered to a custom interface dmz rule
- Beyond the regular timing signoff
  - Special clocking structure (MESH-based clock network/MSCTS)
  - Voltage scaling requirement, Multi-voltage signoff
  - Hard-IP specific considerations: Aging margins
  - Judicious use of DMSA & Primetime license management
- Food for thought
  - Hierarchical-Flat Correlation vs Signoff Closure vs Schedule

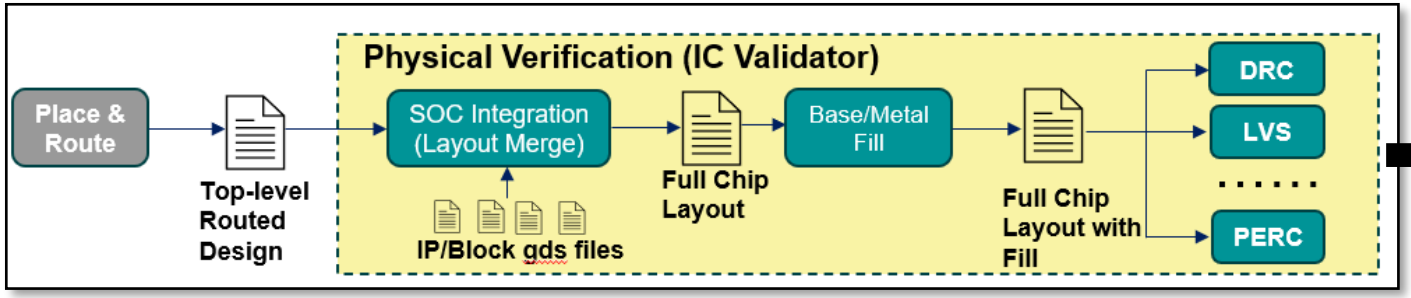
Reliability Verification



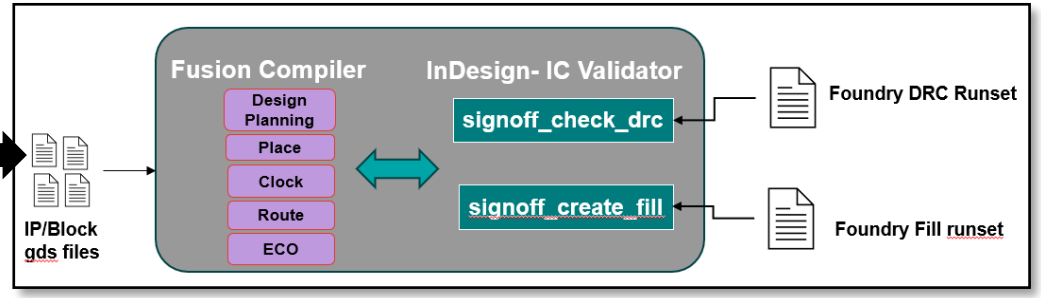
- Custom PGA solution was used in FC
- Redhawk-SC Fusion Evaluation:
  - TAT ⬇ reduced with IR-Fix+Implementation loop.
- Actionable insights
  - Start RV early
    - Could reset P&R or longer closure loop
  - Views/models/methodology vs Accuracy
  - End-to-end closure: At least once before final run
    - IR-Analysis:IR-Fix:Timing-DRC:IR-Analysis Loop
    - On near timing closed DB

- Views/models & signoff criteria
  - Without AVM : 0 , With AVM and avm_read:avm_write:stand_by ~10 to 20mV more IR drop
  - Methodology : Grid robustness, IR & ESD

Physical Verification (IC-Validator)



Traditional Physical Verification Flow



Shift Left Physical Verification within Fusion Compiler (InDesign IC-Validator)

- Seamless integration into Fusion Compiler
  - Efficient Execution, Error viewing and fixing within PnR tool
- In-design ICV is scalable to multiple CPUs/Hosts
  - SLURM setup enabled for multi-CPU/Host for faster TAT (~40% improvement)
- Be Curious & Possibilities
  - ICV features (explorer, hotspot/cluster analysis)
  - PV closure is possible without an army of PV augmentation/fire-fighters.

Startup: Onwards and upwards!

- The Big Picture
  - Startup Opportunities
    - 30,000 foot view & granular knowledge
    - Observational learning & situational understanding
  - Synergistic
    - Tech Node, EDA, IP, Compute, License, Schedule, Implementation, Signoff & “Hit the ground running”
  - Synopsys
    - Robust solutions
    - Excellent support systems

- Excited to explore the new while maintaining momentum on current
  - RTL-A : For shift-left PPA, blur the lines between RTL & PD, which makes perfect sense for a startup
  - TCM : Constraints management – A silicon Savior
  - DSO.ai : TAT reduction & PPA improvement

Acknowledgements

Thanks to Synopsys Team

Azhar Imam
Mukunda Nakkana
Jin Wang
Bhavani Prasad
Vijay Sivalingam
Karan Shah
Charu Khosla
Irfan Shaik

SiMa.ai snug

THANK YOU

Our
Technology,
Your
Innovation™