Semiconductor
Research
Corporation®

Paper

H3DFact: Heterogeneous 3D Integrated CIM for Factorization with Holographic Perceptual Representations

Zishen Wan*, Che-Kai Liu*, Mohamed Ibrahim, Hanchen Yang, Samuel Spetalnick, Tushar Krishna, Arijit Raychowdhury (*Equal Contributions)

Georgia Institute of Technology

SRC TECHCON 09/09/2024, Session 8.1
zishenwan@gatech.edu

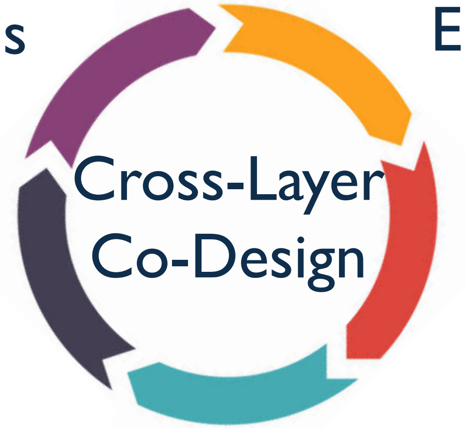
JUMP 2.0-CoCoSys-3131.005

Presenter

Presenter: Zishen Wan
- PhD Student at Georgia Tech
- Advisors: Prof. Arijit Raychowdhury and Prof. Tushar Krishna
- SRC Research Scholar (CBRIC, CoCoSys)

Webpage: https://zishenwan.github.io

Research Interest



Autonomous Systems

Neuro-Symbolic AI, Embodied AI

Cross-Layer Co-Design

Computer Architecture

Chip Design, Compute-In-Memory

Reliability, Safety, Sustainability

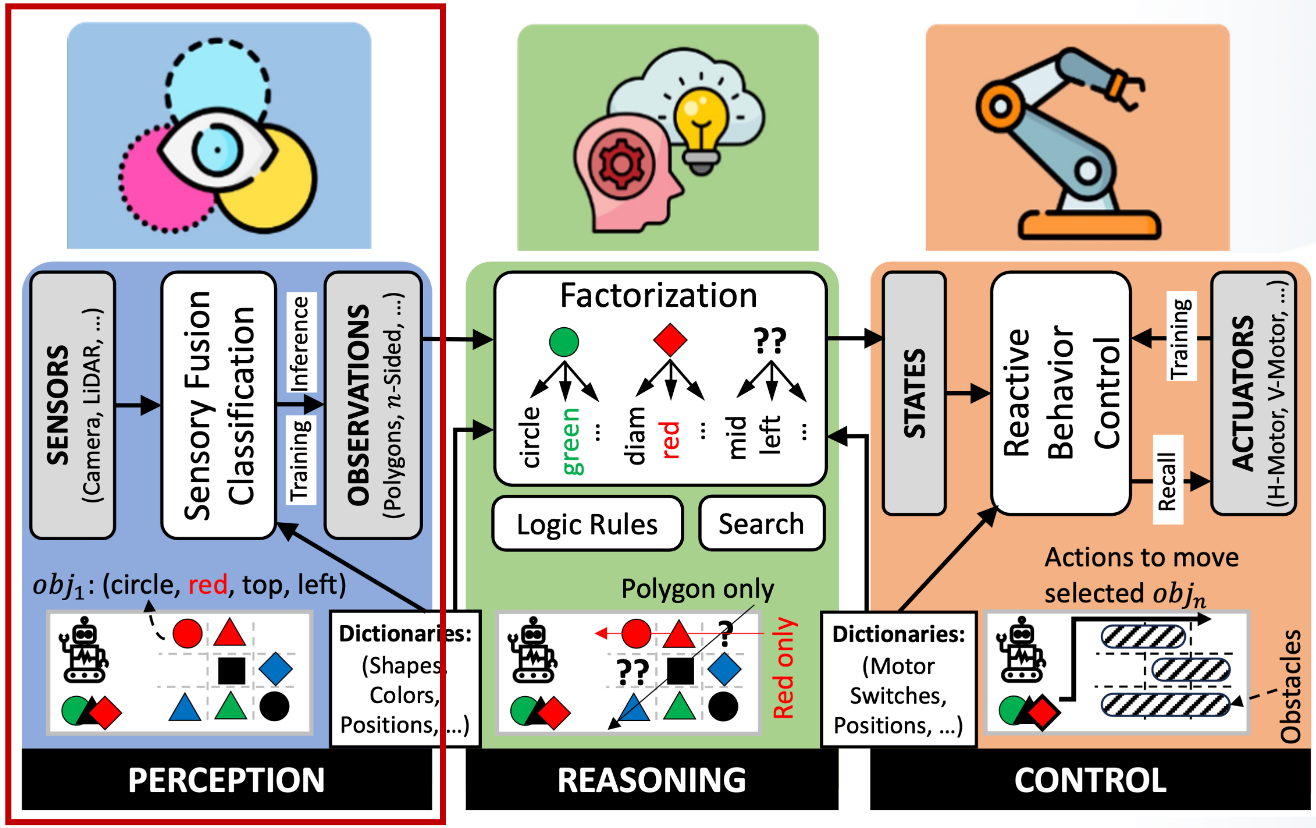
Outline

- Hierarchical Cognition
- Background – Holographic Vector Factorization
- H3DFact
    - Architecture
    - Floorplan
    - Interconnect
    - Circuitry
- Evaluation Results
- Conclusion

Outline

- Hierarchical Cognition
- Background – Holographic Vector Factorization
- H3DFact
  - Architecture
  - Floorplan
  - Interconnect
  - Circuitry
- Evaluation Results
- Conclusion

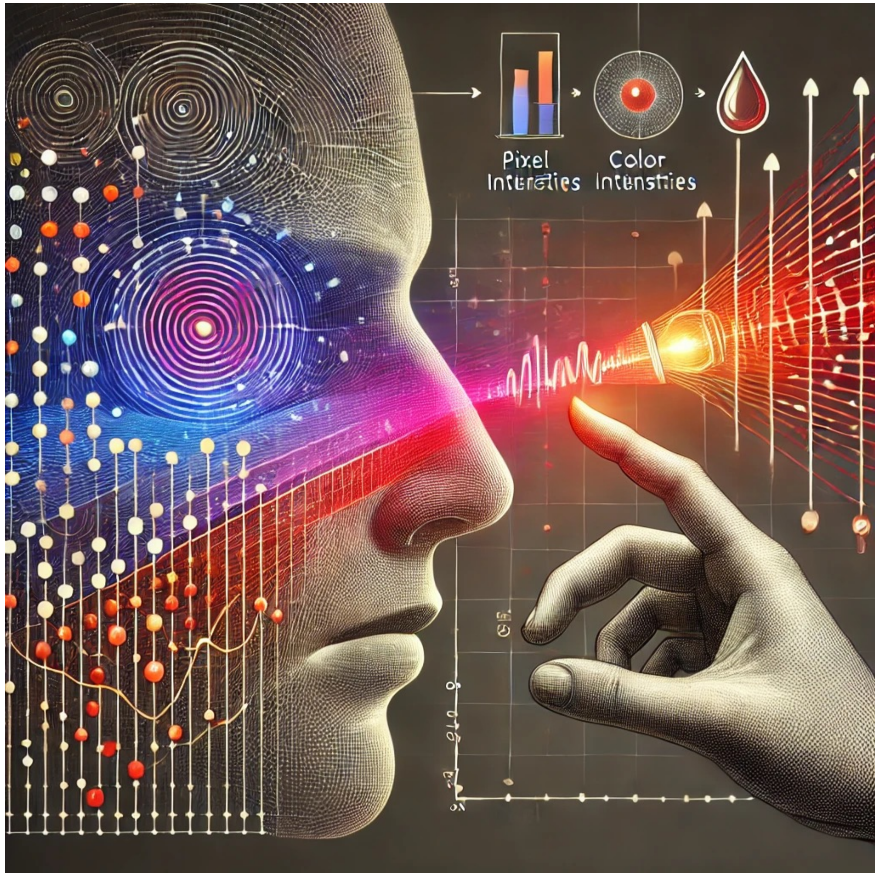
Human-like Hierarchical Cognition



- **Hierarchical Cognition Procedure**: Perception – Reasoning – Control.
- **Perception** is the foundation for high-order cognition, like problem thinking and reasoning.
- **Disentangling the attributes of sensory signal** is central to sensory perception and cognition, hence a critical task for future AI and neuro-symbolic systems.

Perception Problem – Disentangle Sensory Attributes

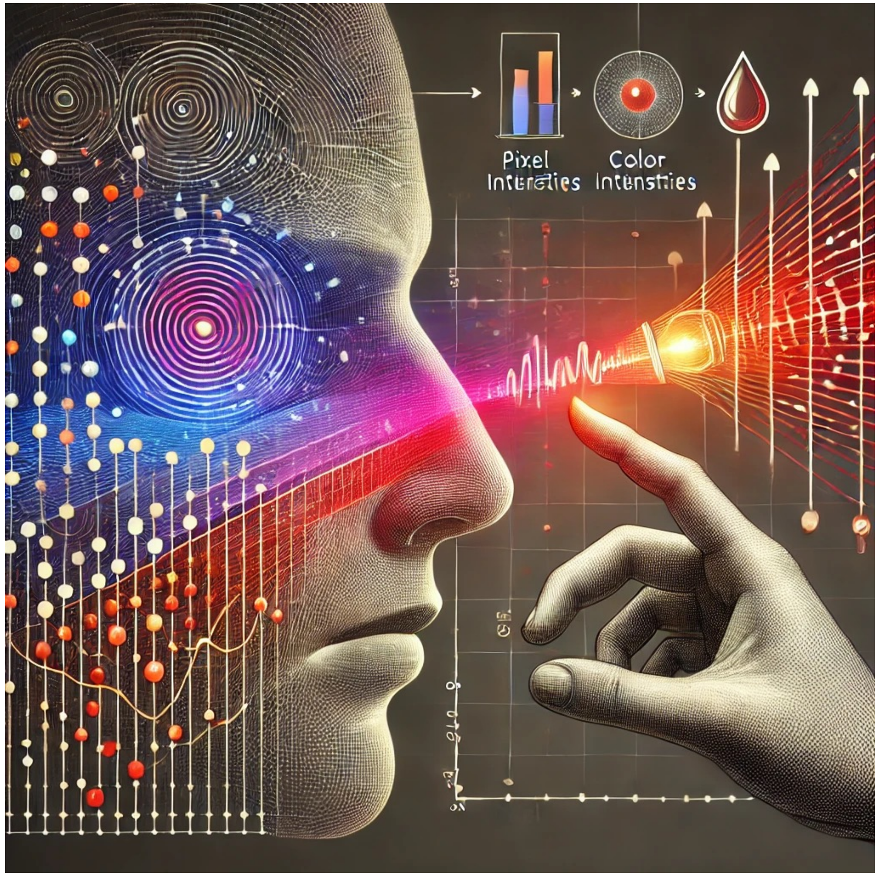
Perception Problem – Disentangle Sensory Attributes



(figure generated by DALL.E)

Foundational Unbinding problem: separate causes of a raw sensory signal that contain multiple attributes.

Perception Problem – Disentangle Sensory Attributes



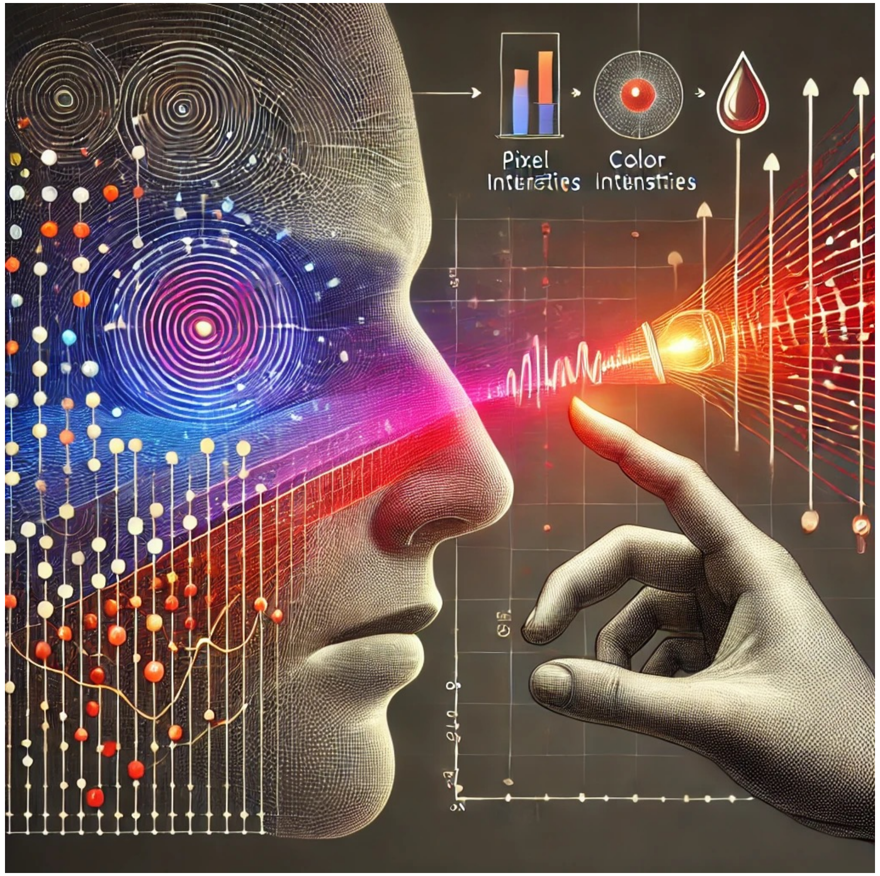
(figure generated by DALL.E)

Foundational Unbinding problem: separate causes of a raw sensory signal that contain multiple attributes.

Examples:

- Pixel intensities sensed by photoreceptors: from the combination of different physical attributes.
- Observed luminance at a point: from a multiplicative combination of reflectance and shading

Perception Problem – Disentangle Sensory Attributes



(figure generated by DALL.E)

Foundational Unbinding problem: separate causes of a raw sensory signal that contain multiple attributes.

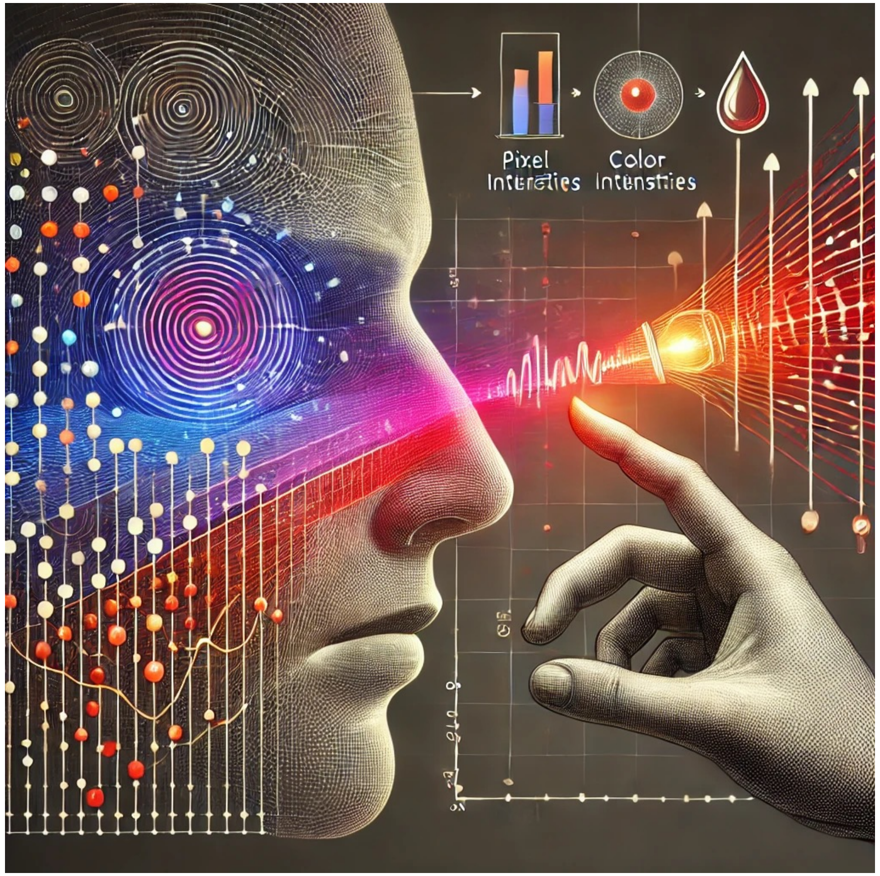
Examples:

- Pixel intensities sensed by photoreceptors: from the combination of different physical attributes.
- Observed luminance at a point: from a multiplicative combination of reflectance and shading

↓

Factorization Problem

Perception Problem – Disentangle Sensory Attributes



(figure generated by DALL.E)

Foundational Unbinding problem: separate causes of a raw sensory signal that contain multiple attributes.

Examples:

- Pixel intensities sensed by photoreceptors: from the combination of different physical attributes.
- Observed luminance at a point: from a multiplicative combination of reflectance and shading

↓

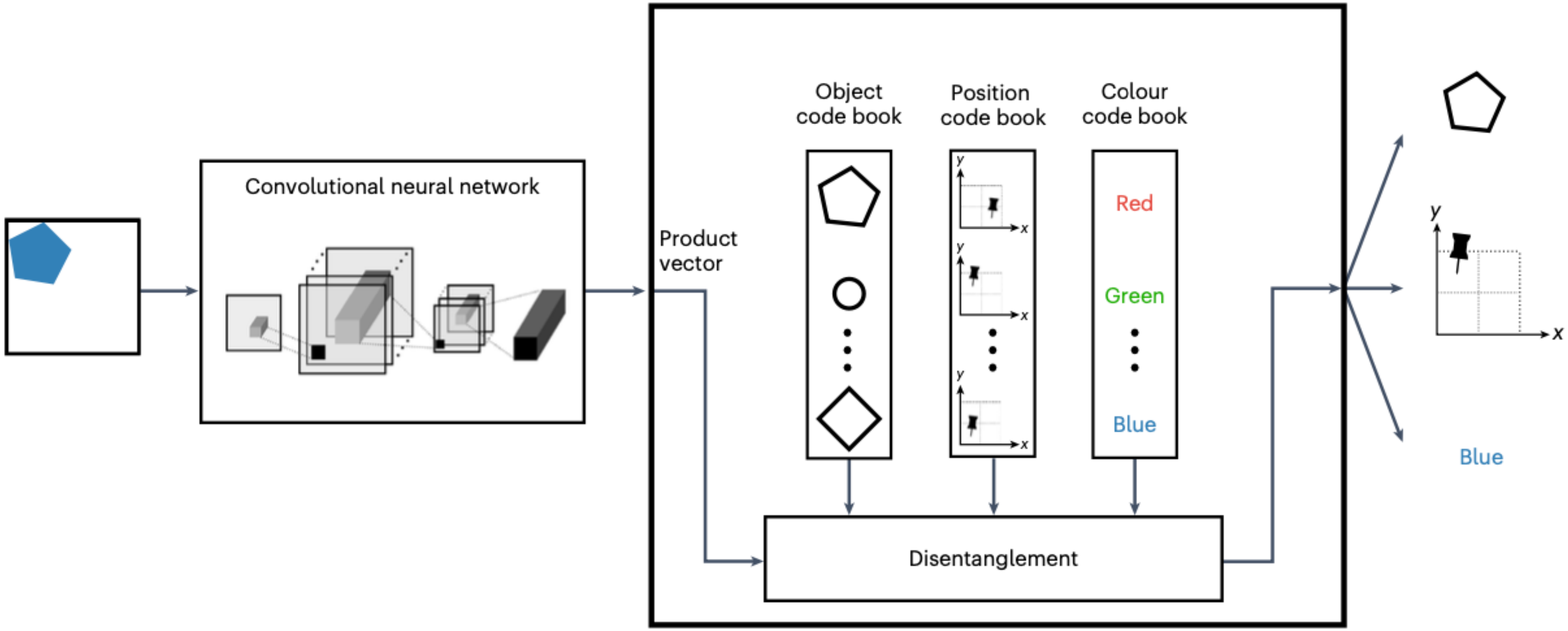
Factorization Problem

- Factoring scene pixels into persistent and dynamic components
- Factoring sentence structure into roles and fillers
- Factoring cognitive analogical reasoning

Outline

- Hierarchical Cognition
- Background – Holographic Vector Factorization
- H3DFact
  - Architecture
  - Floorplan
  - Interconnect
  - Circuitry
- Evaluation Results
- Conclusion

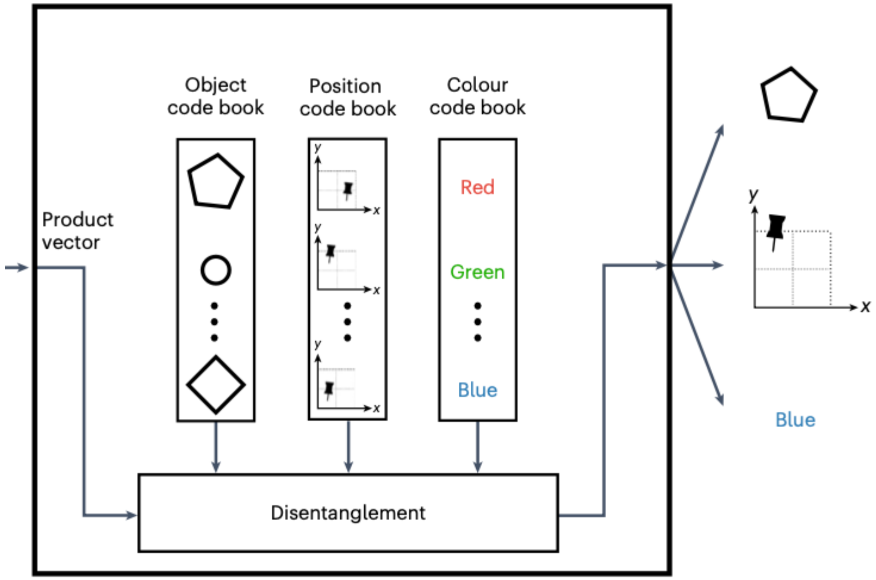
Holographic Vector Factorization

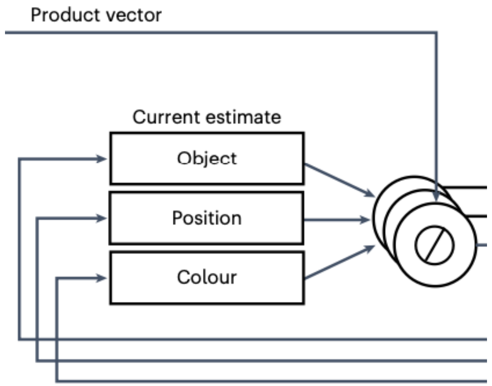


- **Holographic Vector Factorization**: brain-inspired vector-symbolic architecture.
- Each sensory attribute is **encoded** and **processed** using a unique holographic vector, thereby creating distinct and separable representations.

Langenegger et al, "In-memory factorization of holographic perceptual representations", Nature Nanotechnology, 2024

Holographic Vector Factorization



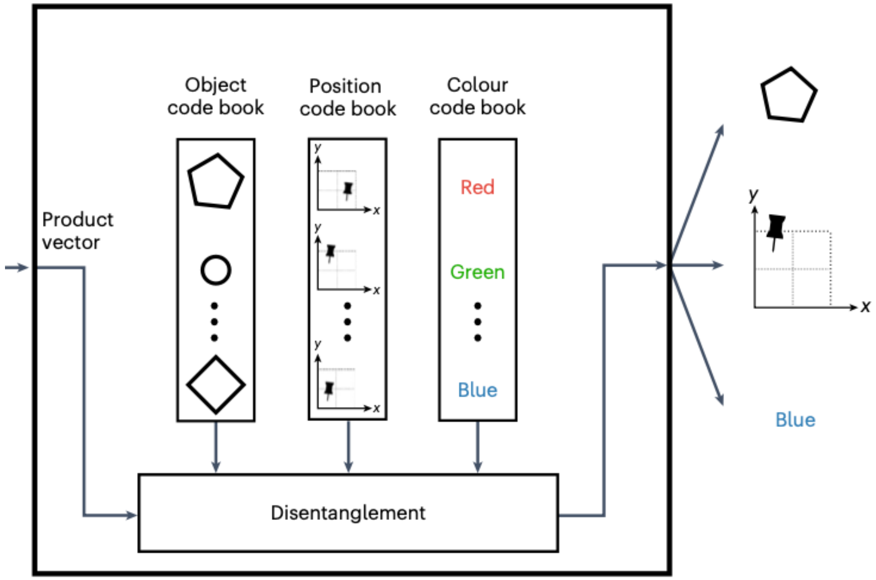


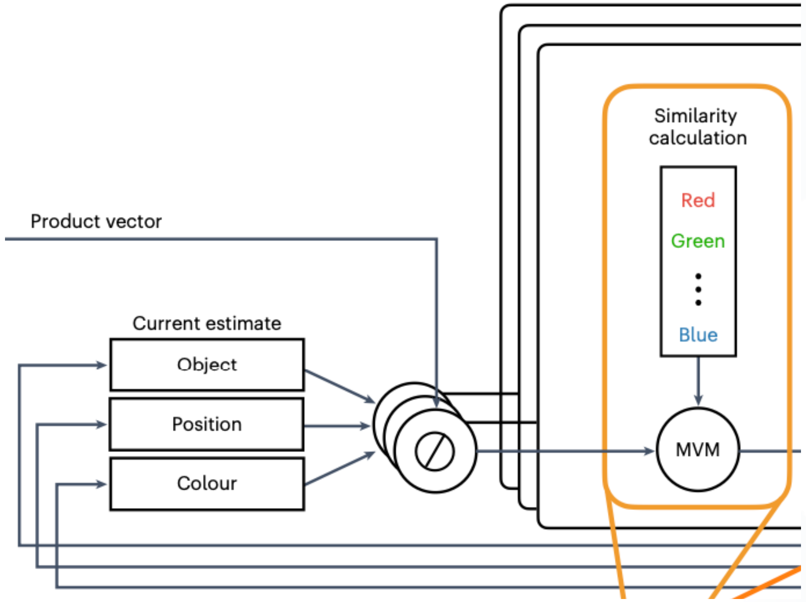
Step 1:
Unbinding

- **Step 1 Unbinding:** unbinding the contribution of the other factors from product vector

Langenegger et al, "In-memory factorization of holographic perceptual representations", Nature Nanotechnology, 2024

Holographic Vector Factorization





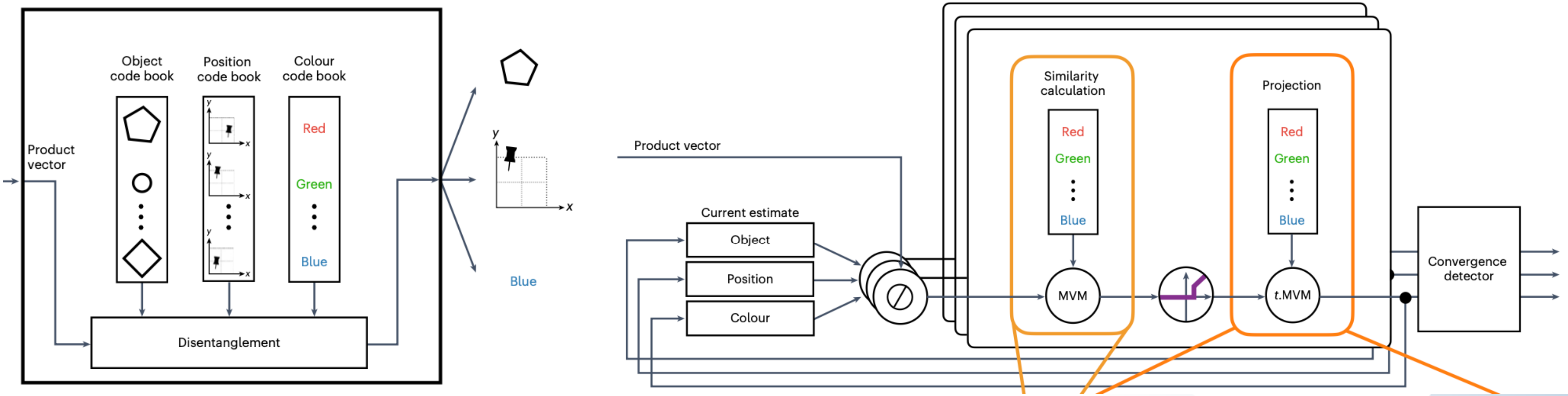
Step 1:
Unbinding

Step 2:
Similarity

Step 3:
Projection

- **Step 1 Unbinding:** unbinding the contribution of the other factors from product vector
- **Step 2 Similarity:** compute similarity values for each unbound estimate

Holographic Vector Factorization



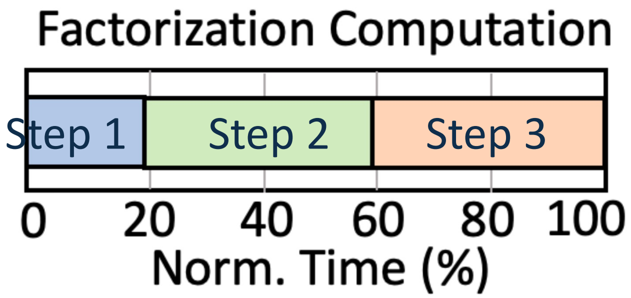
Step 1:
Unbinding

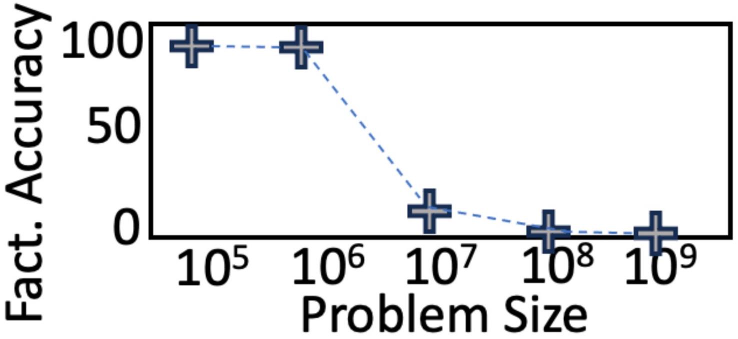
Step 2:
Similarity

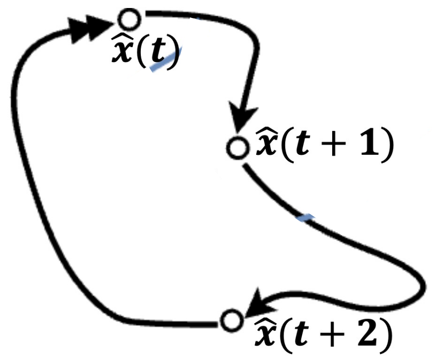
Step 3:
Projection

- **Step 1 Unbinding:** unbinding the contribution of the other factors from product vector
- **Step 2 Similarity:** compute similarity values for each unbound estimate
- **Step 3 Projection:** compute the factors for the subsequent time step

Challenges







Challenge 1:
Intensive computation
Dominated by matrix-vector multiplication operations

Challenge 2:
Limited scalability
Factorization accuracy drops greatly with increasing the problem size

Challenge 3:
CPU/GPU stuck in limited cycle
Factorization constantly end up checking the same solutions

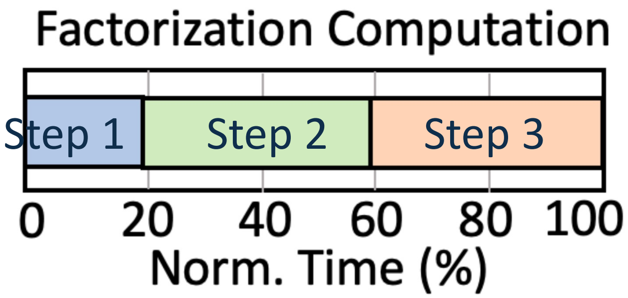
H3DFact

~~How to~~ enable **efficient** and **scalable** factorization of holographic vector representations for human-like sensory cognitive perception?

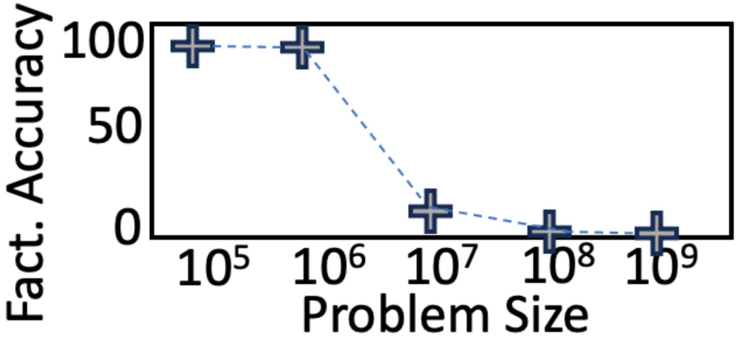
Outline

- Hierarchical Cognition
- Background – Holographic Vector Factorization
- H3DFact
  - Architecture
  - Floorplan
  - Interconnect
  - Circuitry
- Evaluation Results
- Conclusion

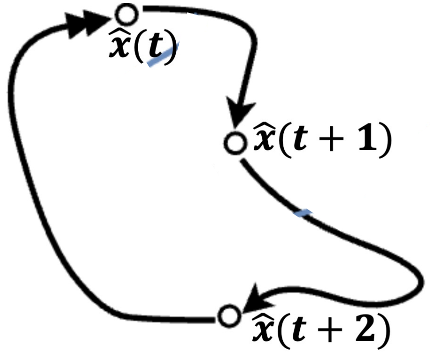
H3DFact Features



Challenge 1:
Intensive computation
Dominated by matrix-vector multiplication operations



Challenge 2:
Limited scalability
Factorization accuracy drops greatly with increasing the problem size



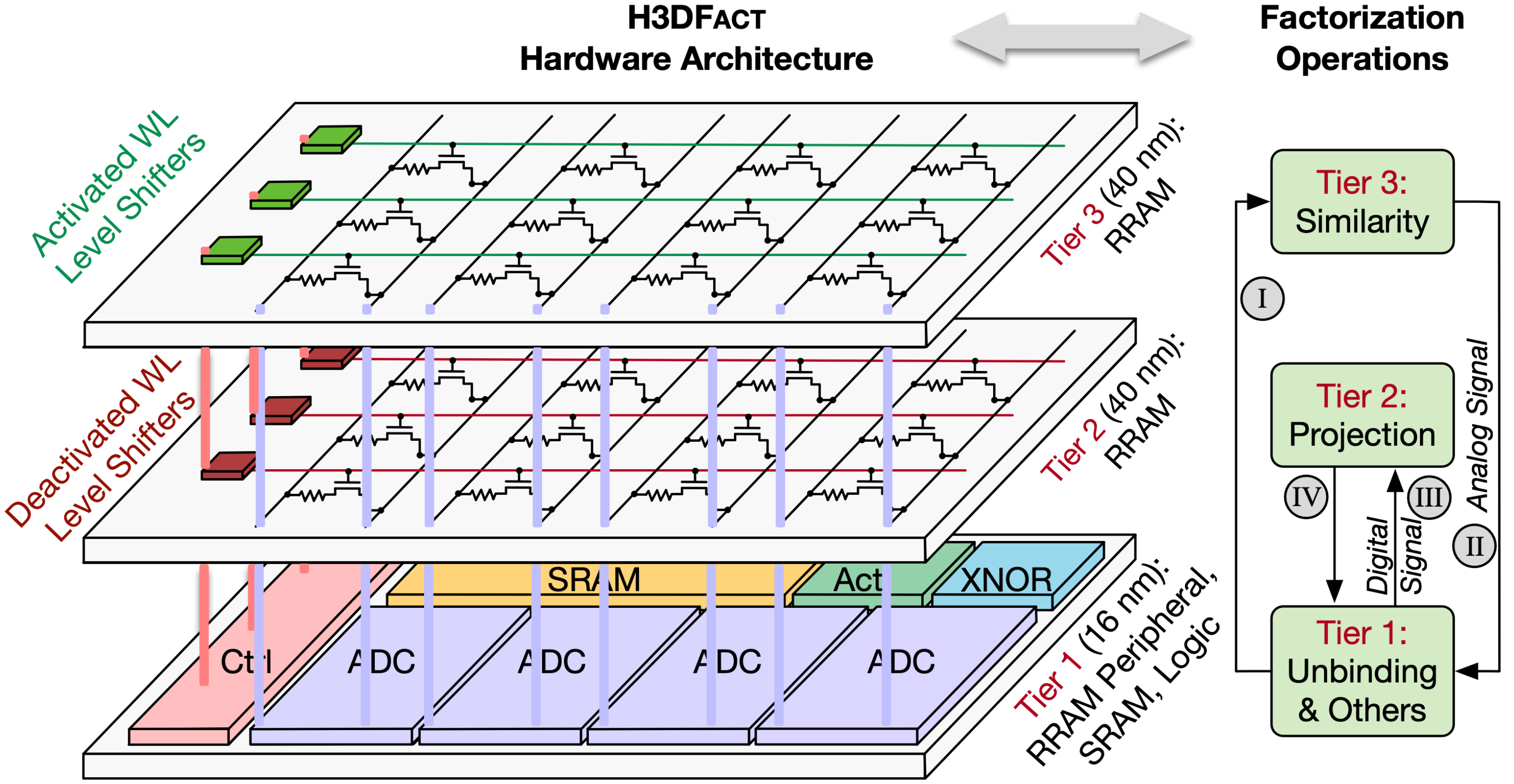
Challenge 3:
CPU/GPU stuck in limited cycle
Factorization constantly end up checking the same solutions

Feature 1:
Computation-in-superposition
CIM paradigm for **efficient** factorization computation

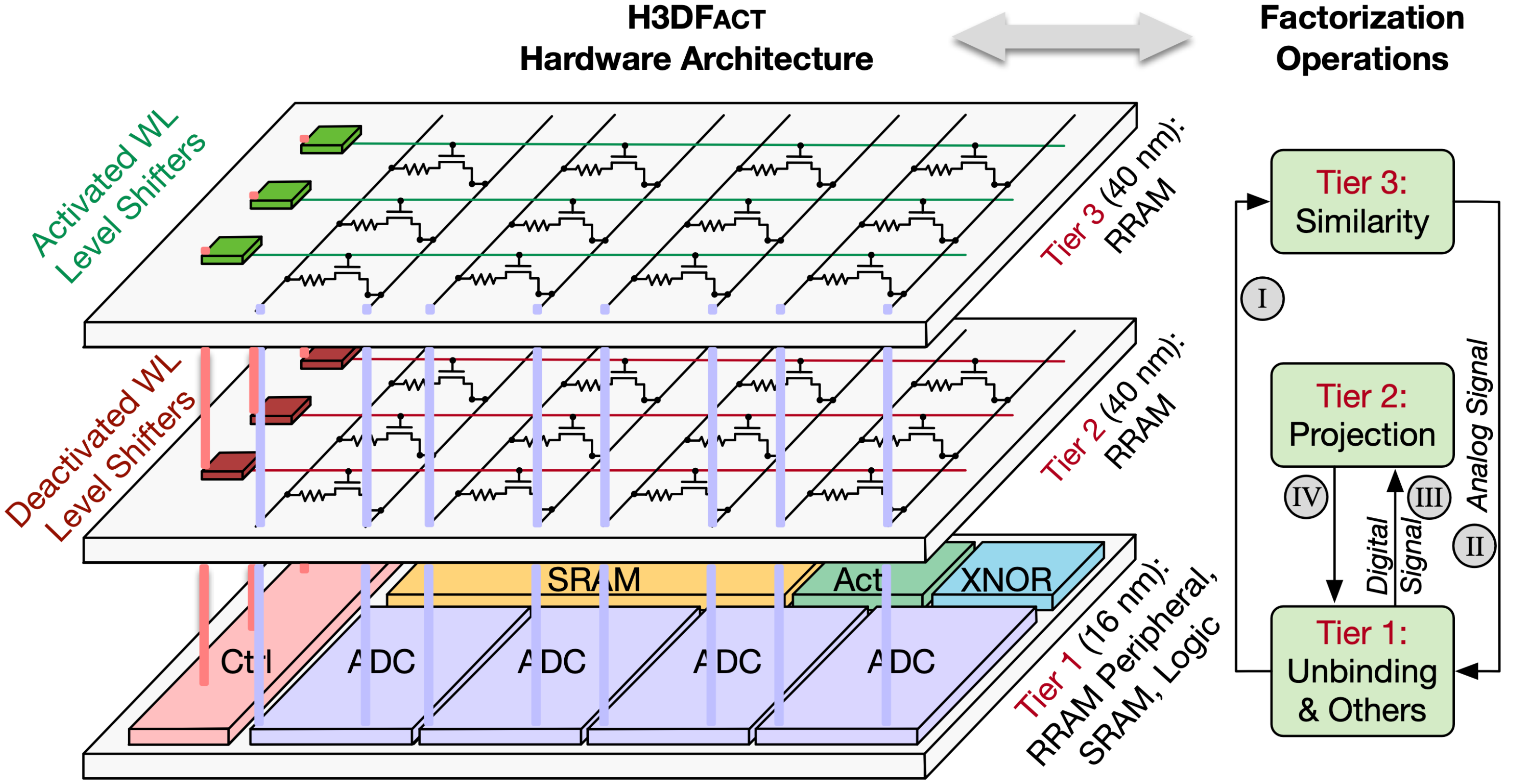
Feature 2:
Heterogeneous 3D integration
system design for **scalable** factorization computation

Feature 3:
Nanoscale memristive devices
intrinsic **stochasticity** to break the factorization limited cycles

H3DFact Architecture - Overview



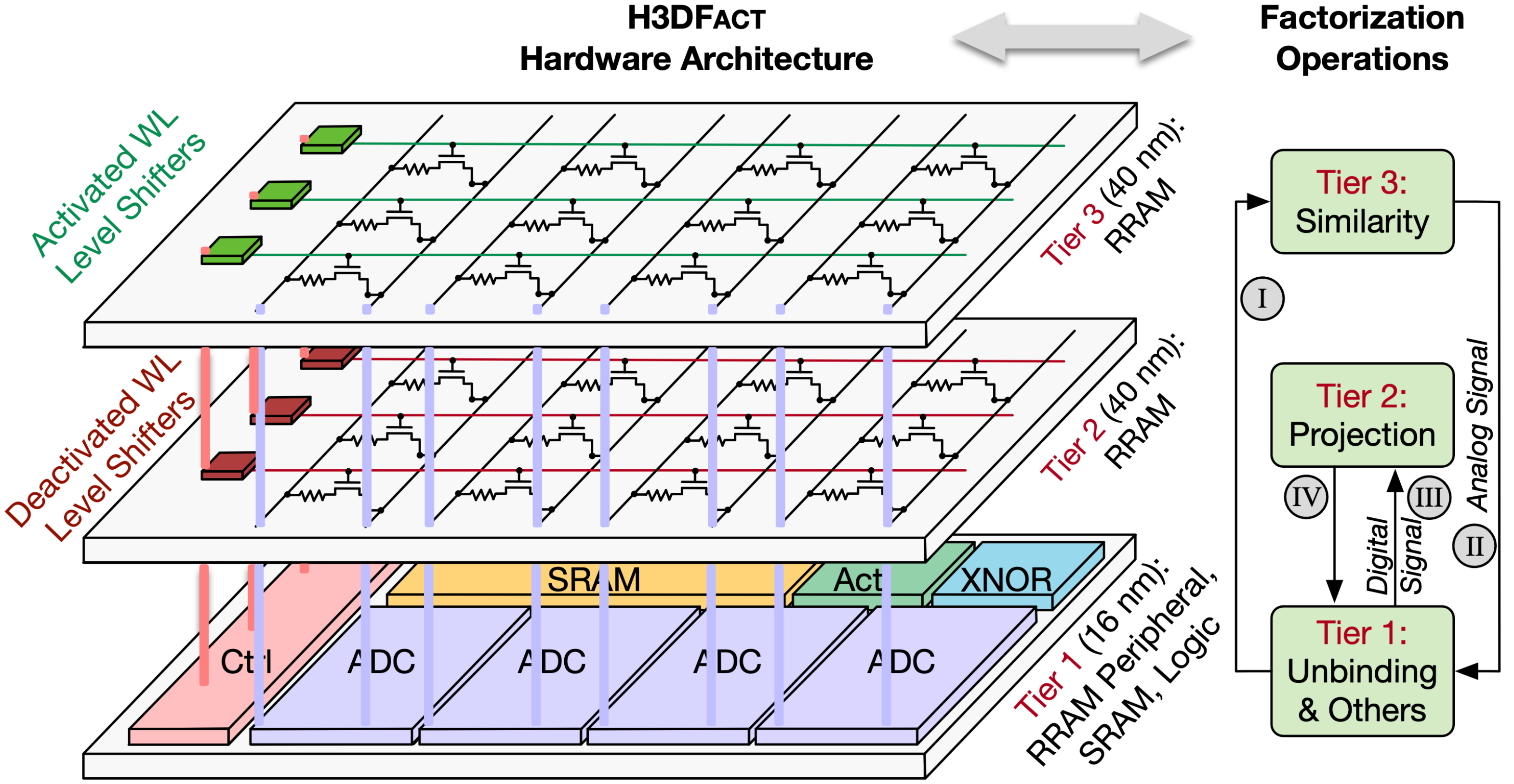
H3DFact Architecture - Overview



Three-Tier architecture:

- Tier 3 (top):
- Tier 1 (bottom):
  - Technology: 16nm SRAM, peripheral, logic
  - Operations: unbinding, others

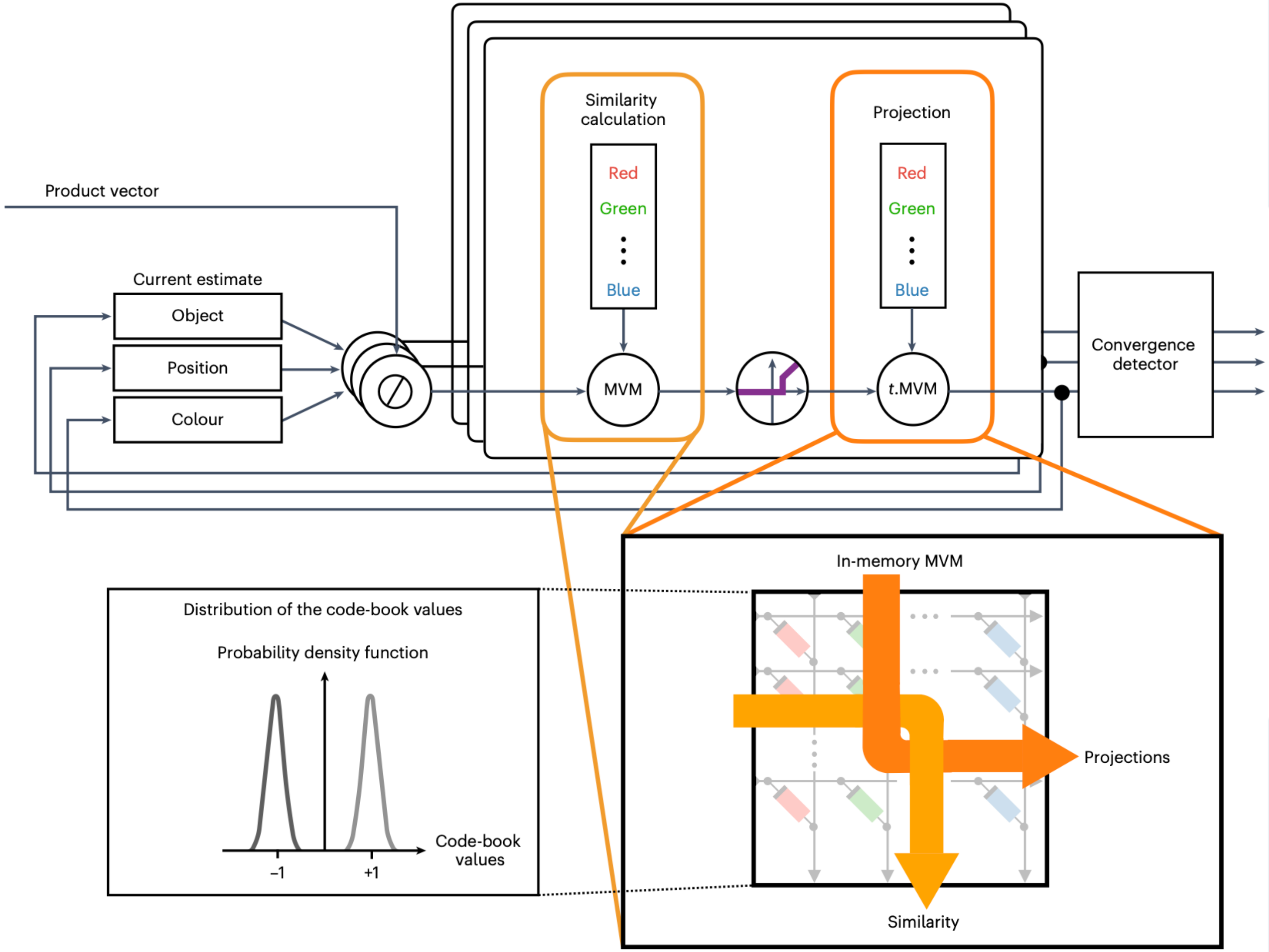
H3DFact Architecture - Overview



Three-Tier architecture:

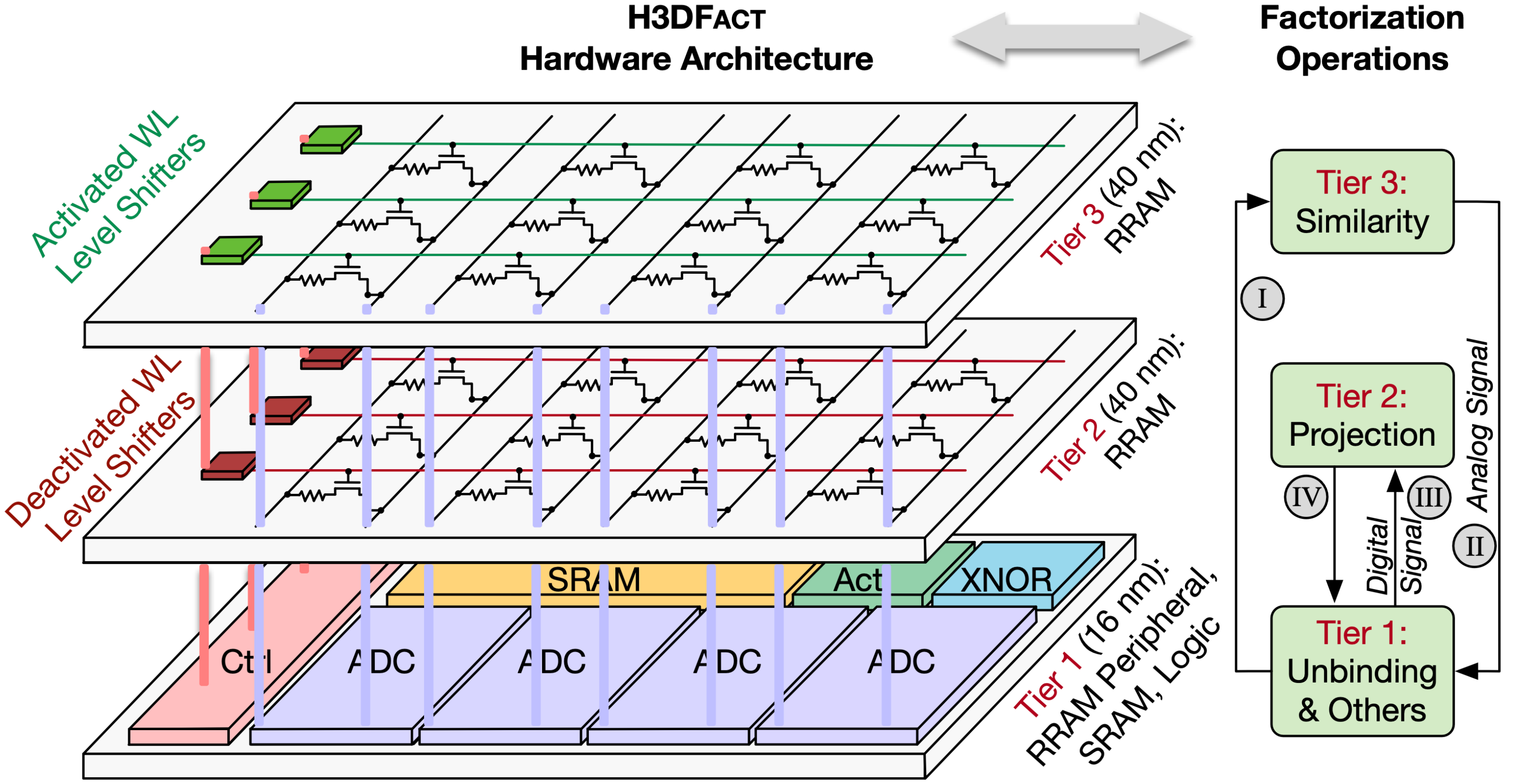
- Tier 3 (top):
  - Techn[illegible]AM
  - [illegible]
- Tier 2 (middle):
  - Technology: 40nm RRAM
  - Operations: projection
- Tier 1 (bottom):
  - Technology: 16nm SRAM, peripheral, logic
  - Operations: unbinding, others

Compute-In-Memory for Projection and Similarity



Langenegger et al, "In-memory factorization of holographic perceptual representations", Nature Nanotechnology, 2024

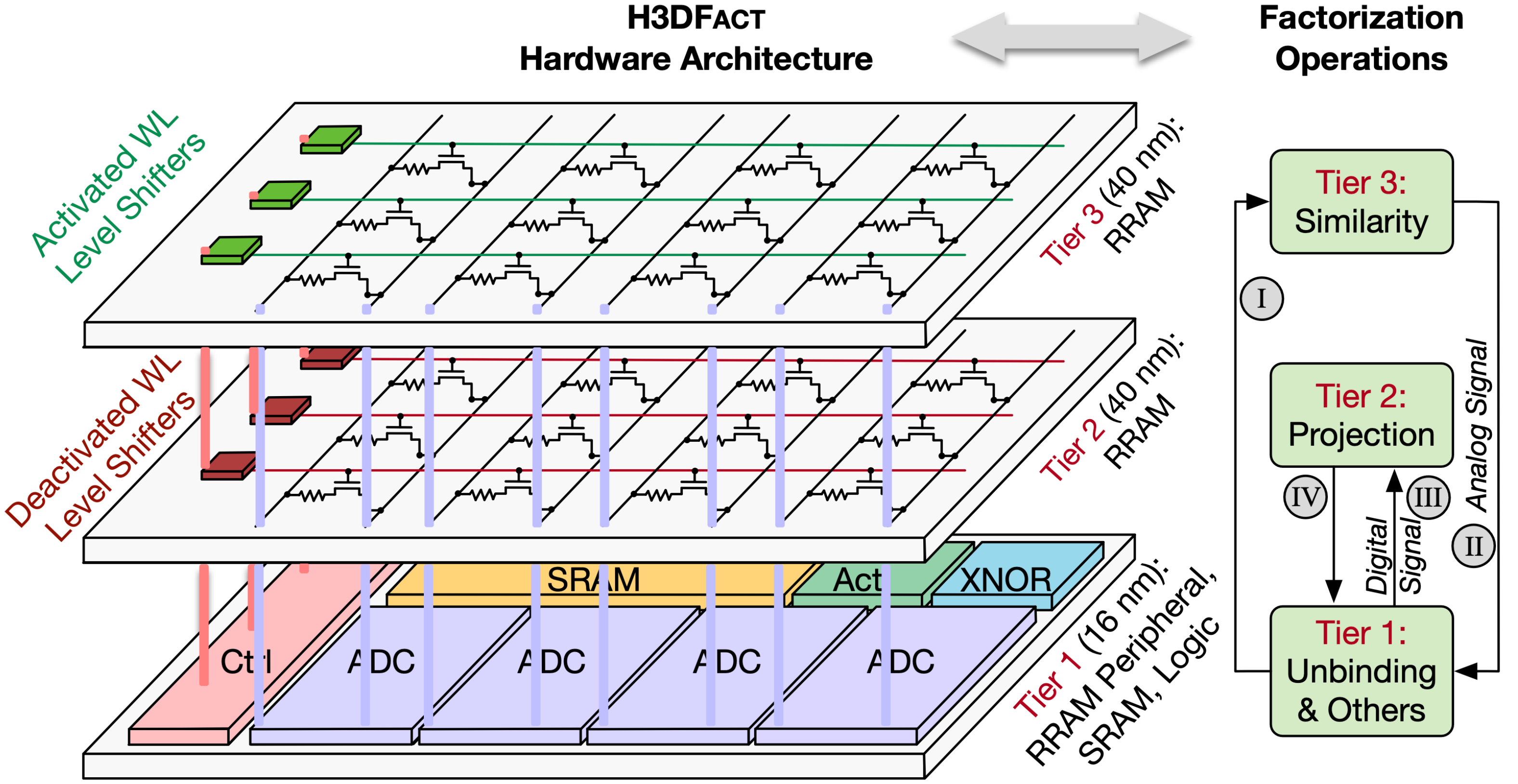
H3DFact Architecture - Overview



Three-Tier architecture:

- Tier 3 (top):
  - Techn [illegible] RAM
- Tier 2 (middle):
  - Technology: 40nm RRAM
  - Operations: projection
- Tier 1 (bottom):
  - Technology: 16nm SRAM, peripheral, logic
  - Operations: unbinding, others

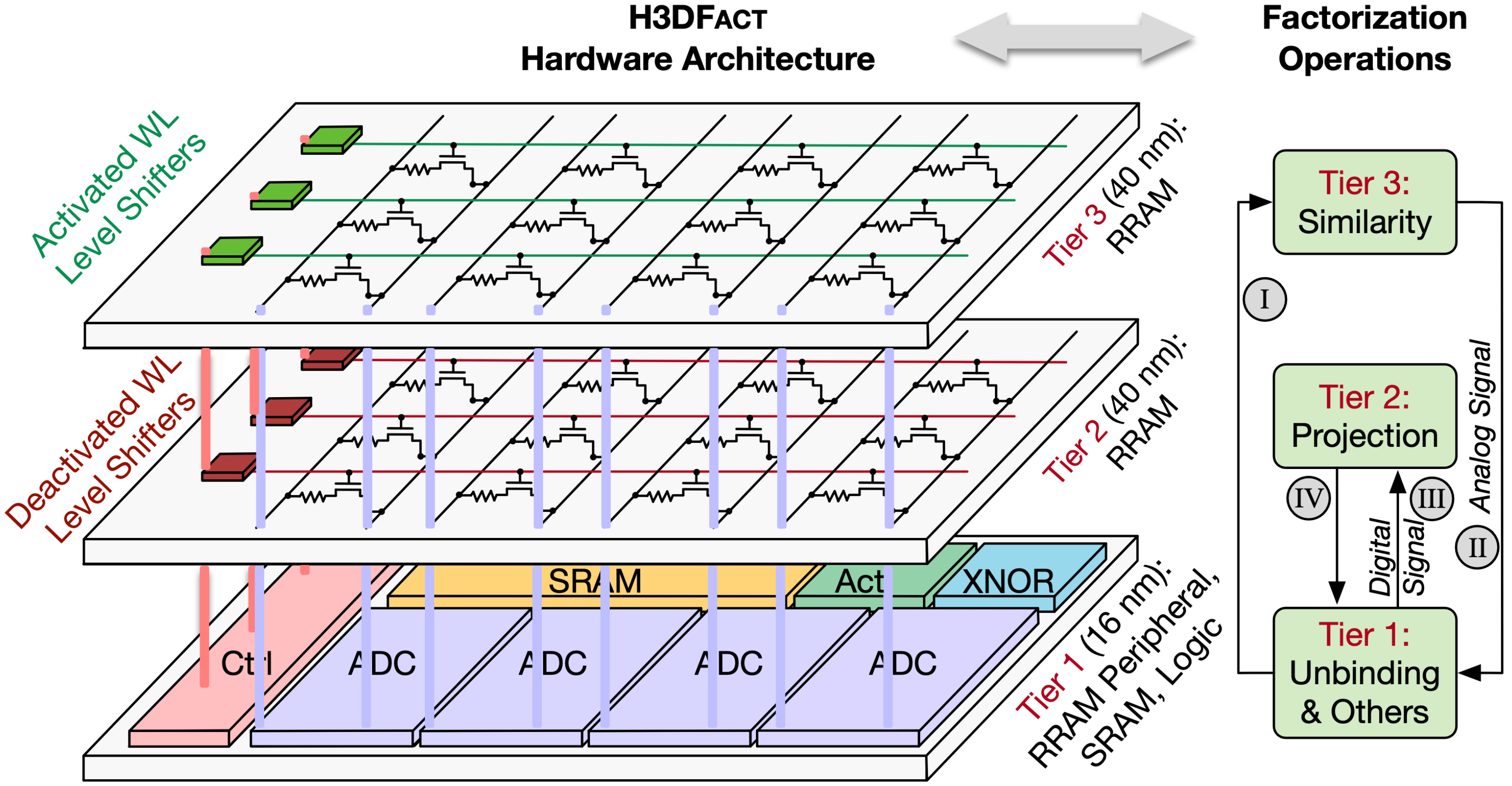
H3DFact Architecture - Overview



Three-Tier architecture:

- Tier 3 (top):
  - Technology: 40nm RRAM
  - Operations: similarity
- Tier 2 (middle):
  - Technology: 40nm RRAM
  - Operations: projection
- Tier 1 (bottom):
  - Technology: 16nm SRAM, peripheral, logic
  - Operations: unbinding, others

H3DFact Architecture - Overview

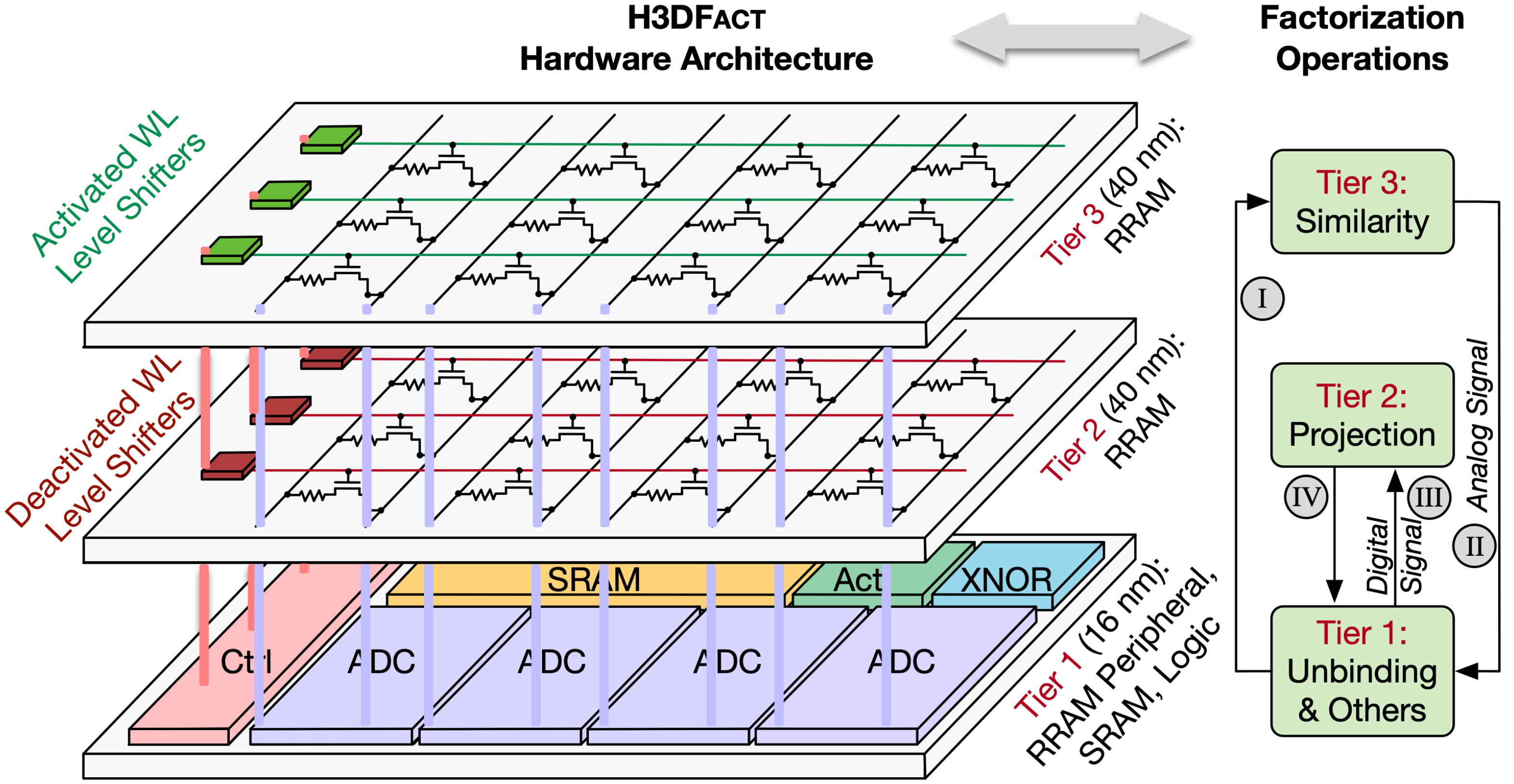


Three-Tier architecture:

- Tier 3 (top):
  - Technology: 40nm RRAM
  - Operations: similarity
- Tier 2 (middle):
  - Technology: 40nm RRAM
  - Operations: projection
- Tier 1 (bottom):
  - Technology: 16nm SRAM, peripheral, logic
  - Operations: unbinding, others

Advantage of heterogeneous 3D integration: enable (1) different technology nodes, (2) hybrid memories, (3) high density

H3DFact Architecture - Overview



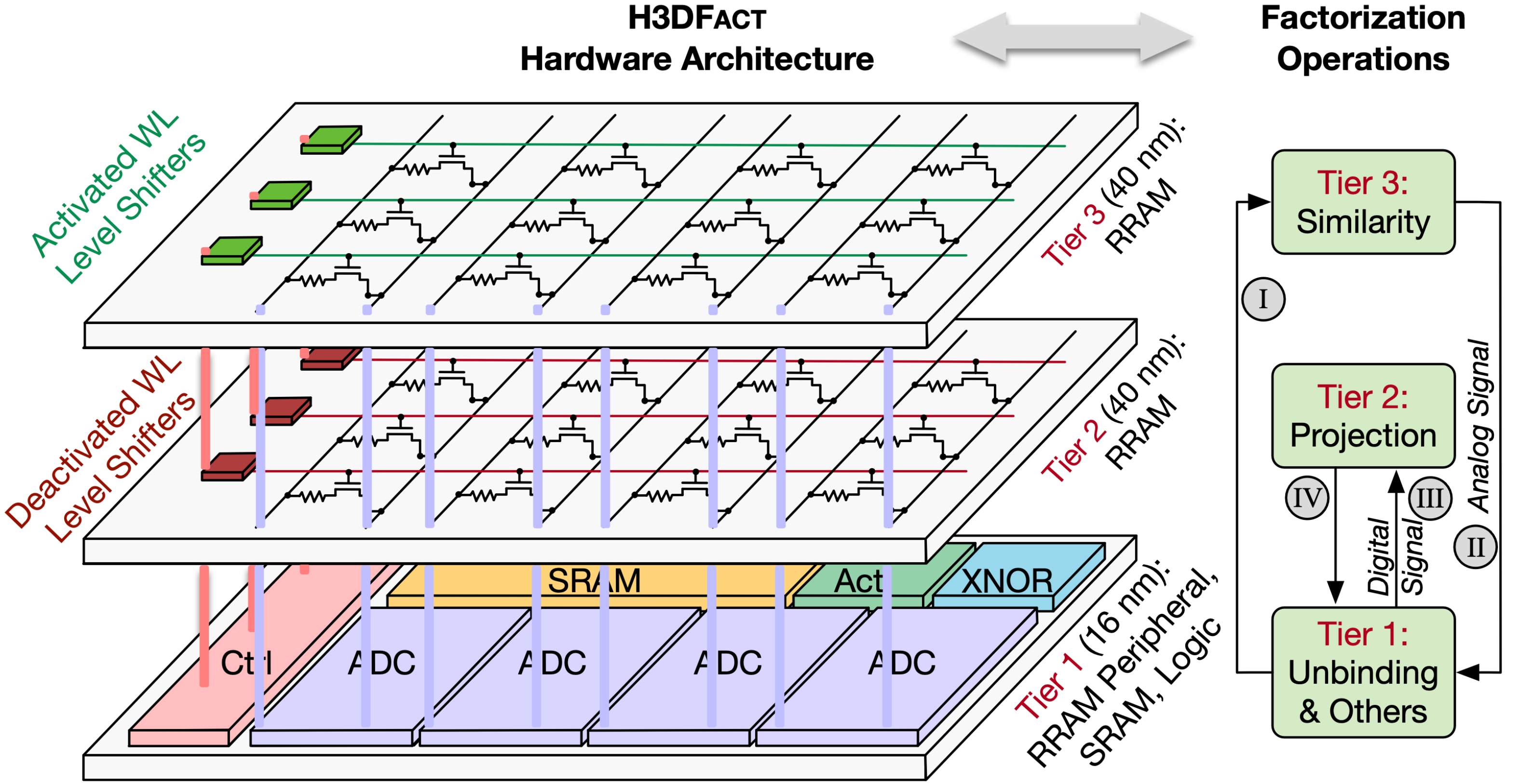
Three-Tier architecture:

- Tier 3 (top):
  - Technology: 40nm RRAM
  - Operations: similarity
- Tier 2 (middle):
  - Technology: 40nm RRAM
  - Operations: projection
- Tier 1 (bottom):
  - Technology: 16nm SRAM, peripheral, logic
  - Operations: unbinding, others

Advantage of heterogeneous 3D integration: enable (1) different technology nodes, (2) hybrid memories, (3) high density

Advantages of compute-in-memory: enable (1) efficient factorization, (2) break stuck cycle with device stochasticity

H3DFact Architecture - Overview

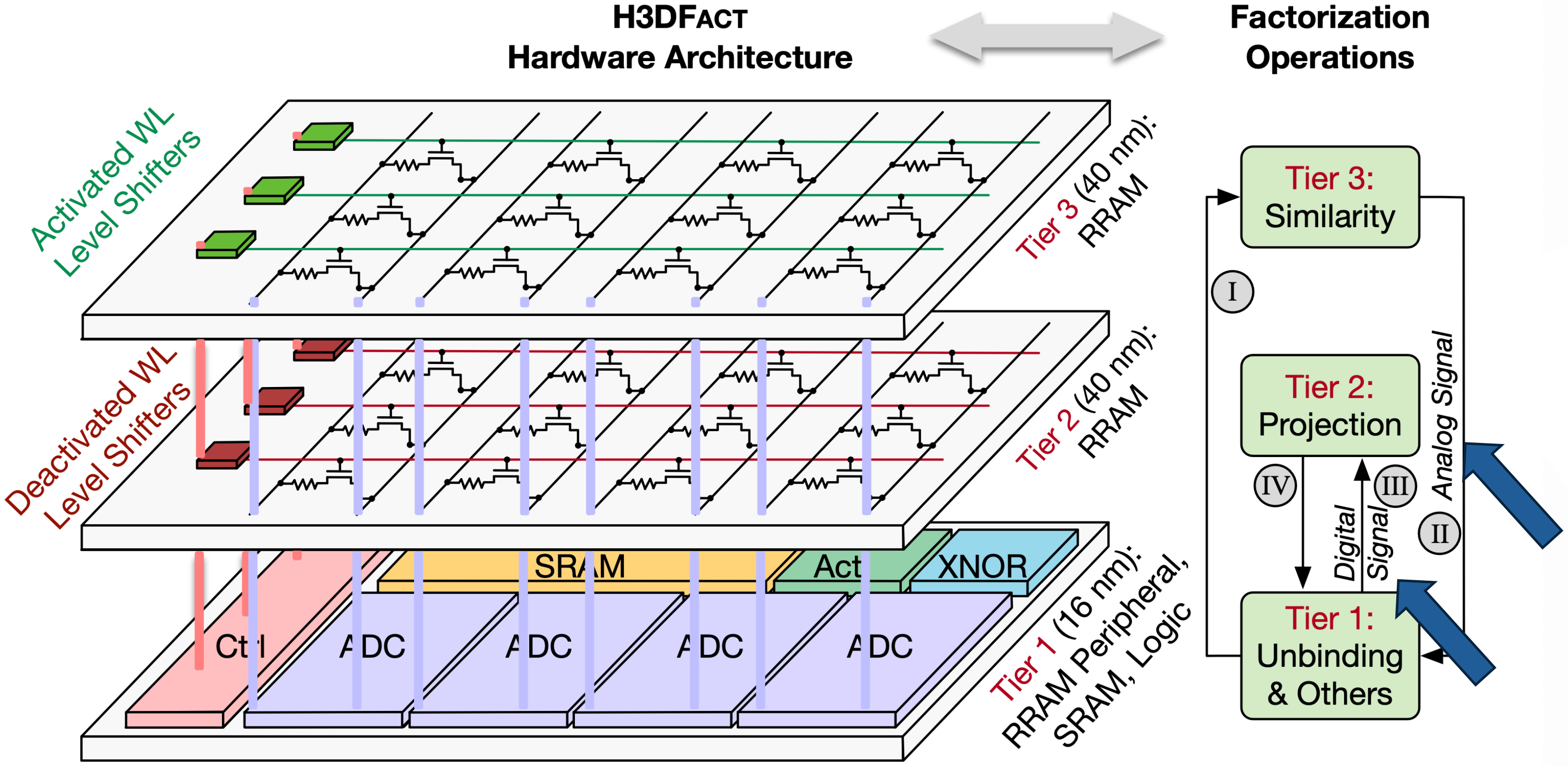


Three-Tier architecture:

- Tier 3 (top):
  - Technology: 40nm RRAM
  - Operations: similarity
- Tier 2 (middle):
  - Technology: 40nm RRAM
  - Operations: projection
- Tier 1 (bottom):
  - Technology: 16nm SRAM, peripheral, logic
  - Operations: unbinding, others

H3DFact features a 3-tier architecture,

H3DFact Architecture - Overview

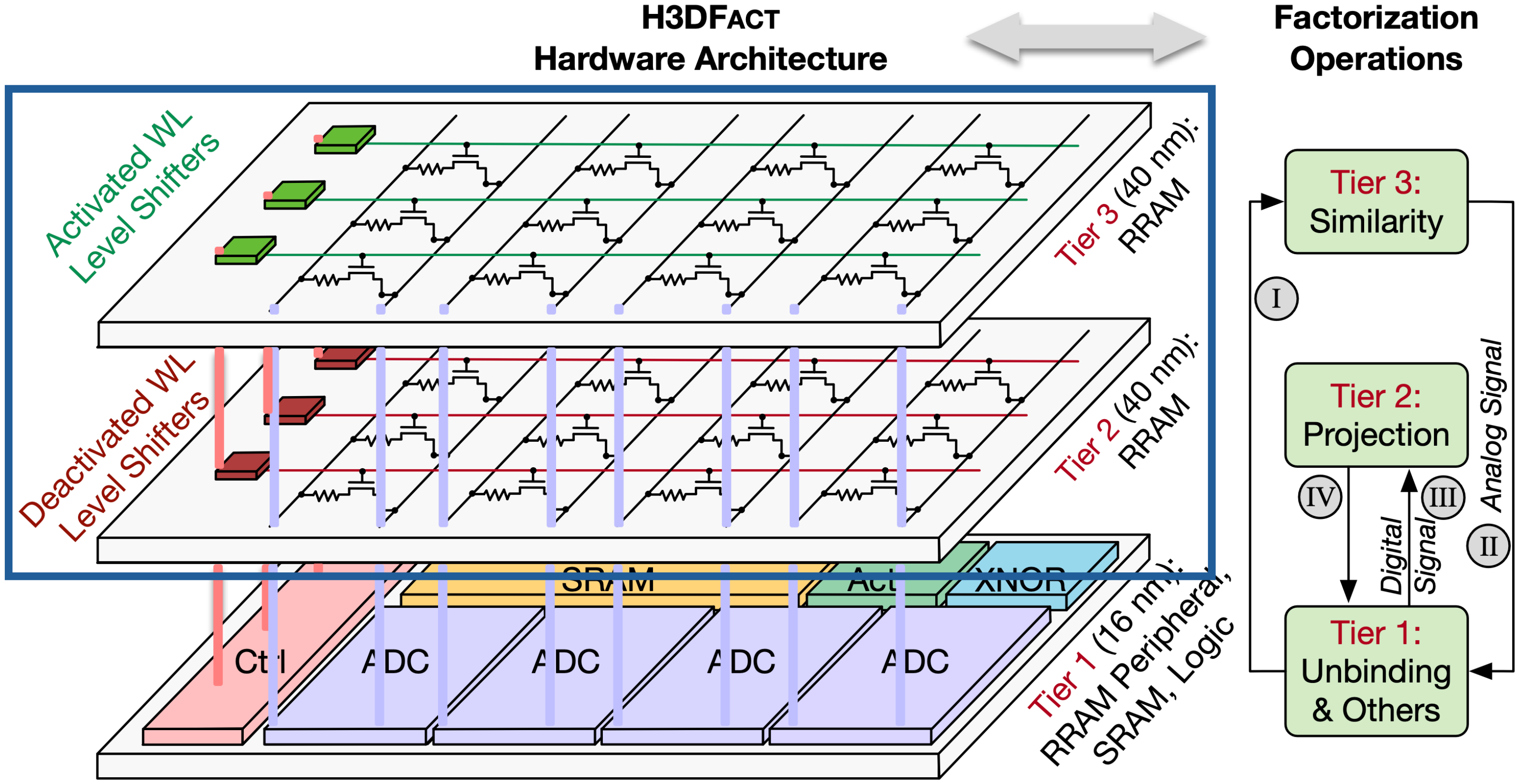


Three-Tier architecture:

- Tier 3 (top):
  - Technology: 40nm RRAM
  - Operations: similarity
- Tier 2 (middle):
  - Technology: 40nm RRAM
  - Operations: projection
- Tier 1 (bottom):
  - Technology: 16nm SRAM, peripheral, logic
  - Operations: unbinding, others

H3DFact features a 3-tier architecture, considering of data traversing format (analog/digital),

H3DFact Architecture - Overview



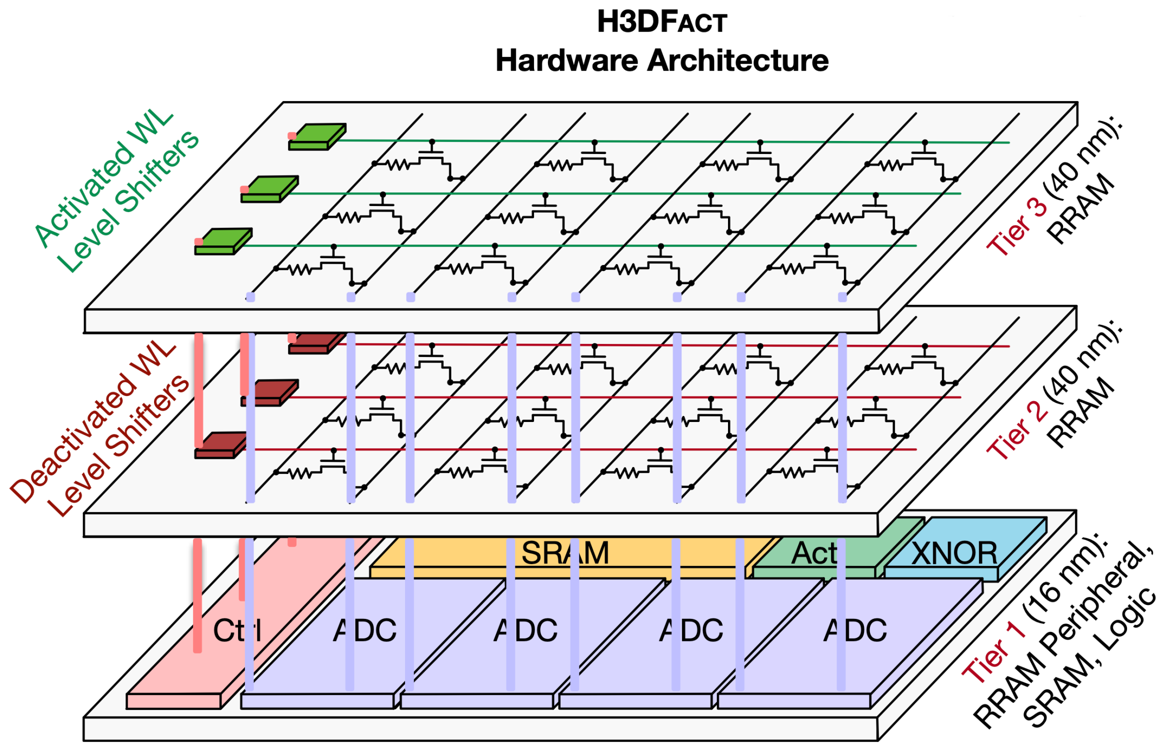
Three-Tier architecture:

- Tier 3 (top):
  - Technology: 40nm RRAM
  - Operations: similarity
- Tier 2 (middle):
  - Technology: 40nm RRAM
  - Operations: projection
- Tier 1 (bottom):
  - Technology: 16nm SRAM, peripheral, logic
  - Operations: unbinding, others

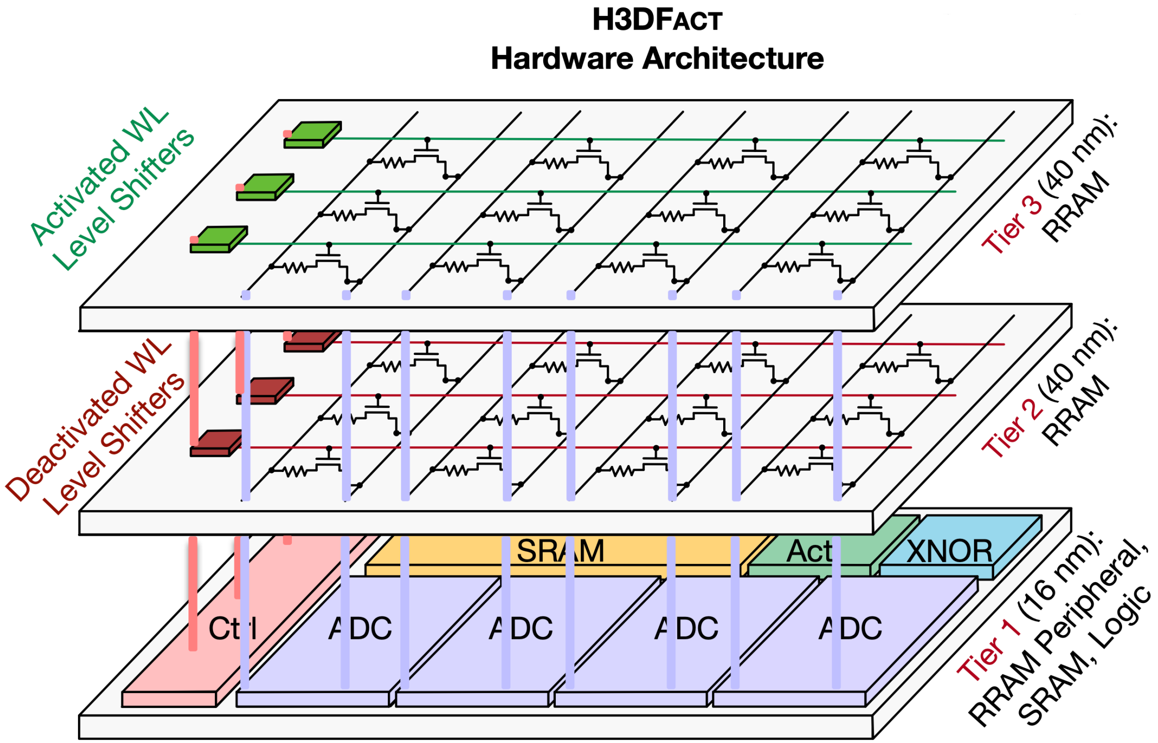
H3DFact features a 3-tier architecture, considering of data traversing format (analog/digital), one RRAM tier is activated at any given time

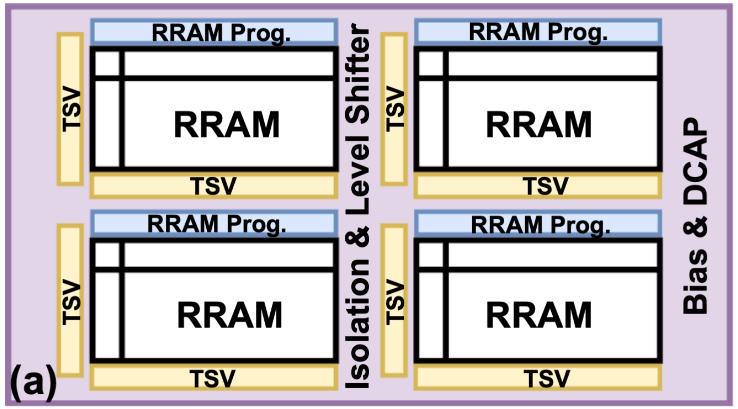
H3DFact Architecture – Floorplan and Interconnect

H3DFact Architecture – Floorplan and Interconnect



H3DFact Architecture – Floorplan and Interconnect

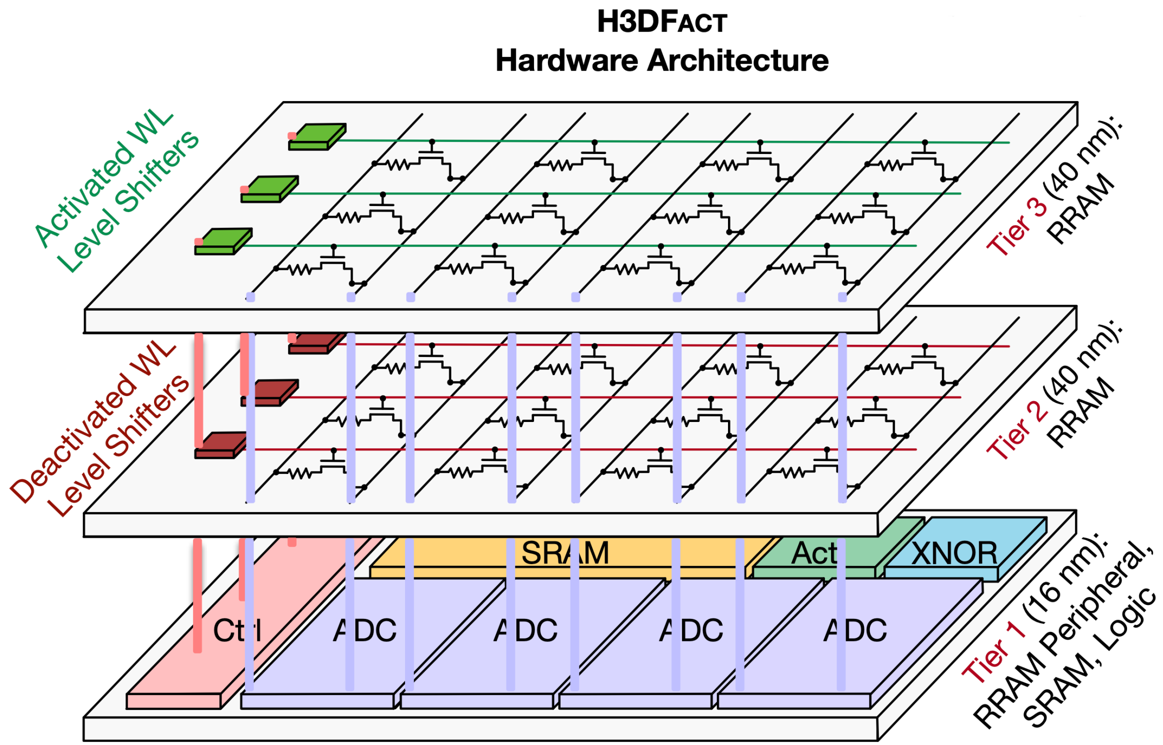


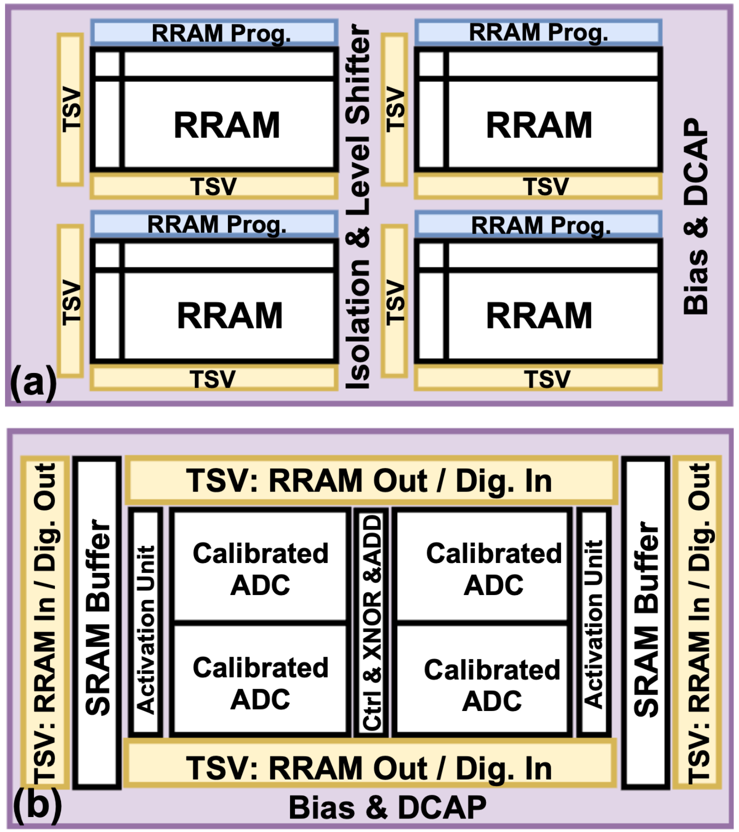


Floorplan:

- Tier-2/3 RRAM: each tier has four RRAM subarrays, each RRAM subarray has 256x256 size

H3DFact Architecture – Floorplan and Interconnect

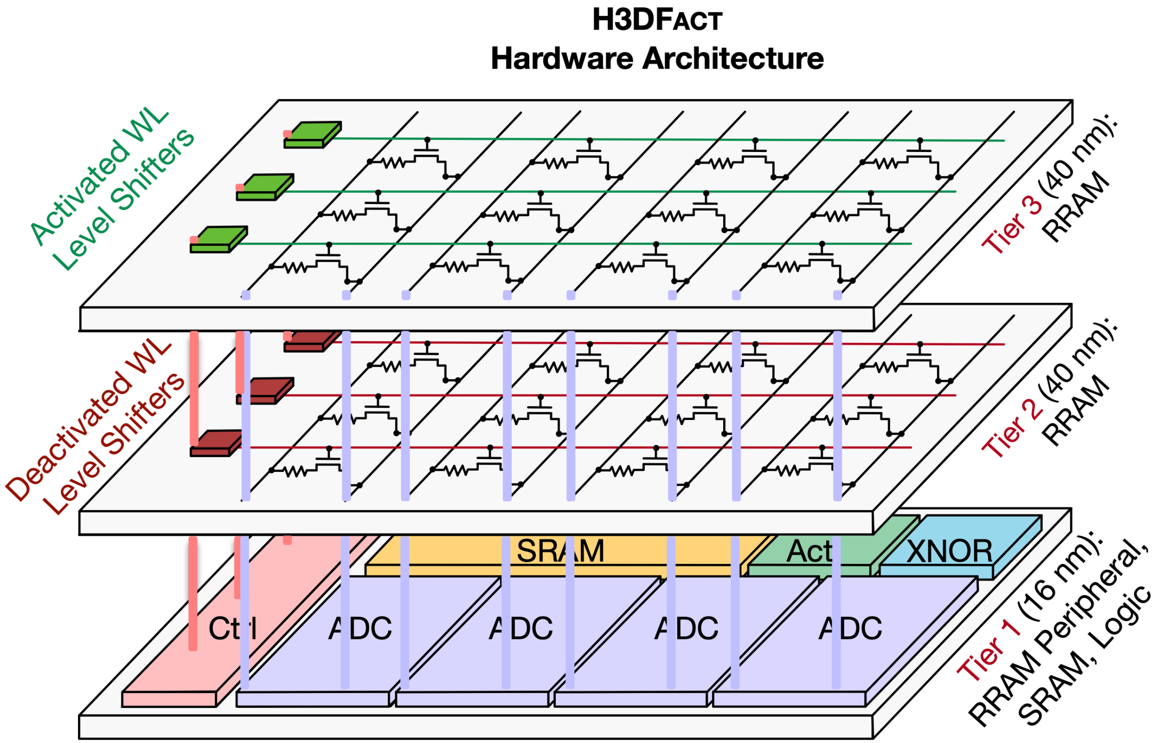


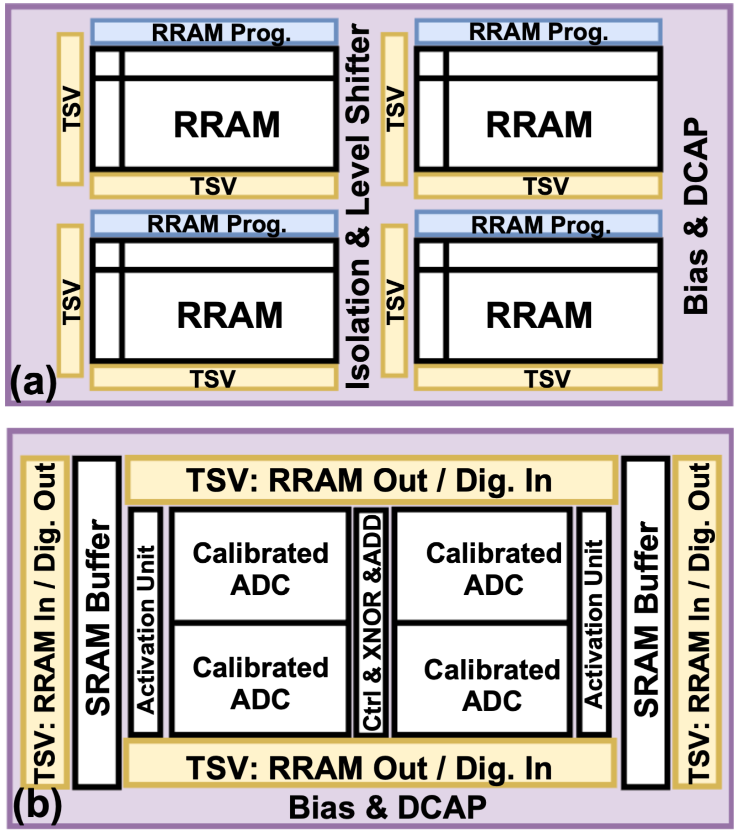


Floorplan:

- Tier-2/3 RRAM: each tier has four RRAM subarrays, each RRAM subarray has 256x256 size
- Tier-1 SRAM, digital, and peripherals: External pins and bumps

H3DFact Architecture – Floorplan and Interconnect

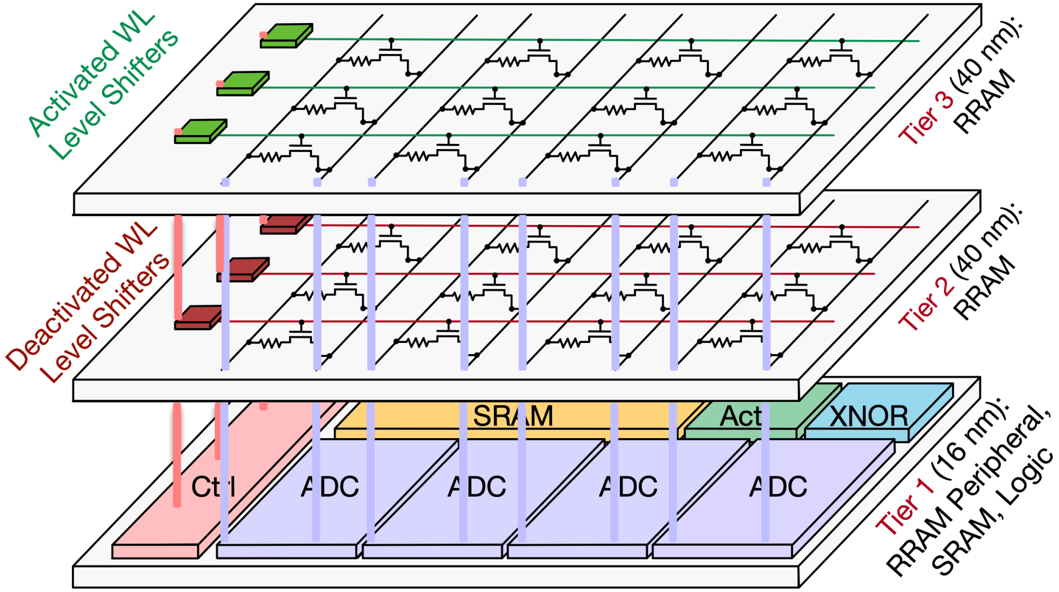




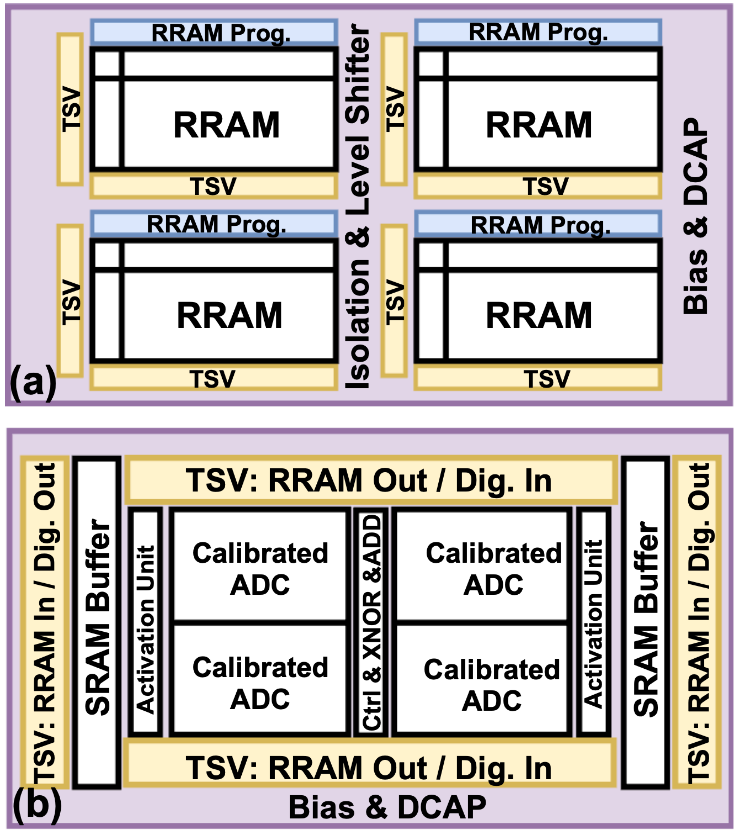
Floorplan:

- Tier-2/3 RRAM: each tier has four RRAM subarrays, each RRAM subarray has 256x256 size
- Tier-1 SRAM, digital, and peripherals: External pins and bumps
- Generalized design method to determine hardware configurations

H3DFact Architecture – Floorplan and Interconnect



TSV Diameter	TSV Pitch	TSV Oxide Thickness	TSV Height	Hybrid Bonding Pitch	Hybrid Bonding Thickness
2 μm	4 μm	100 nm	10 μm	10 μm	3 μm

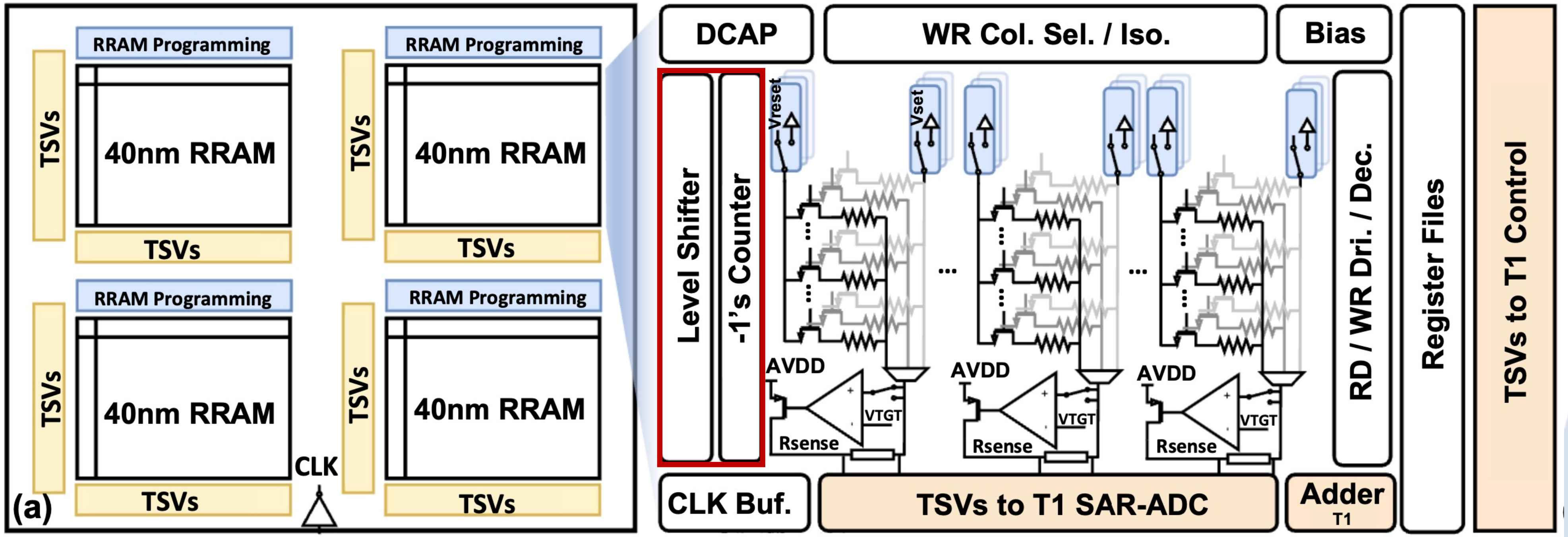


Interconnect & Bonding:

- Interconnect: through-silicon vias (TSVs). One (MxN) RRAM subarray needs (M+N+N/2) TSVs
- Bonding: mix of face-to-face (F2F) and face-to-back (F2B) bonding

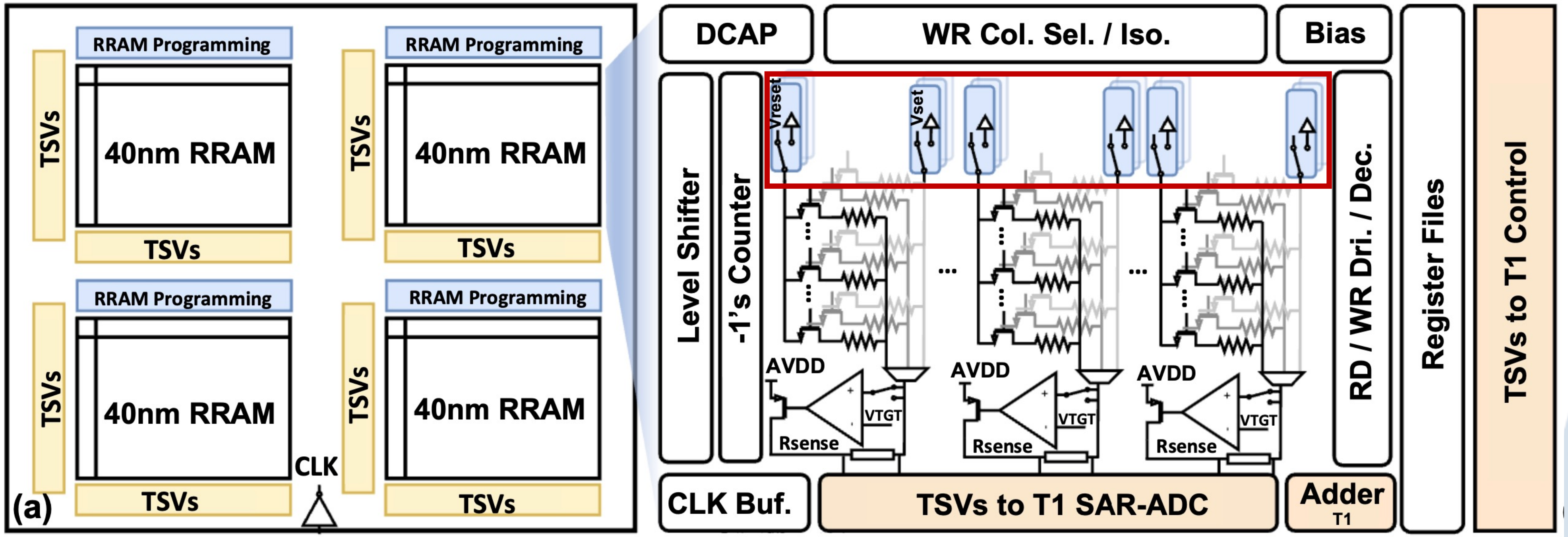
H3DFact Architecture – Circuit Details

H3DFact Architecture – Circuit Details



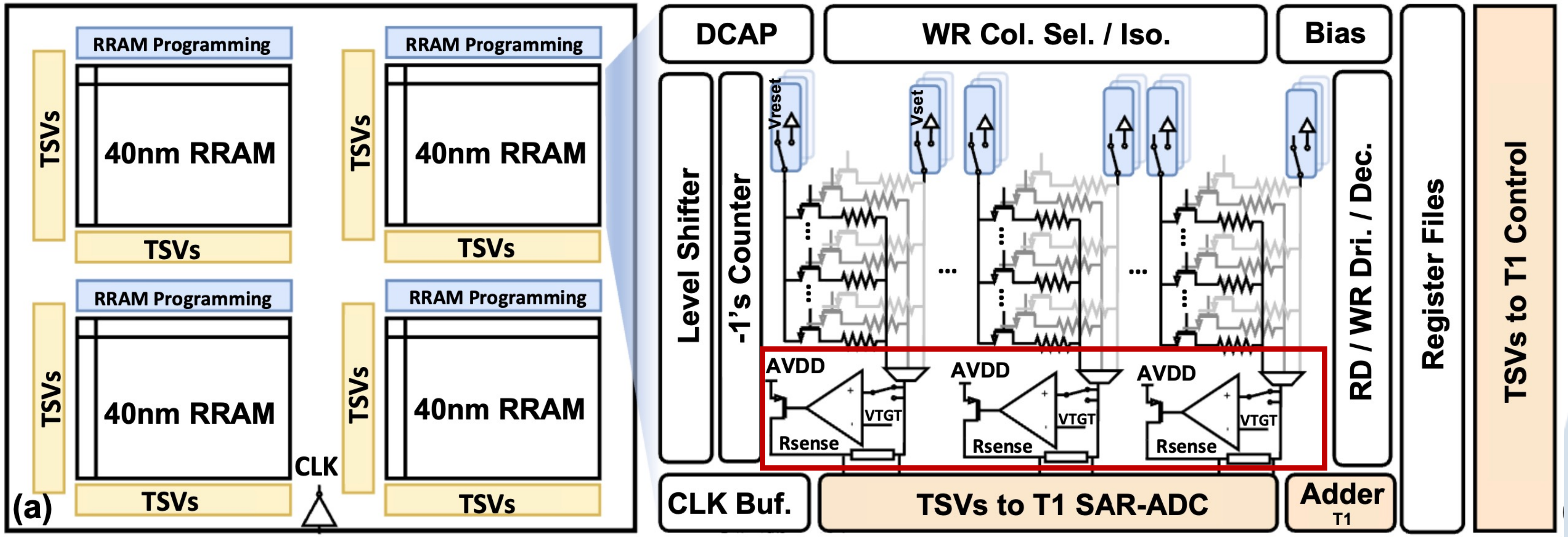
- **Circuitry**: capable of executing high-dimensional bipolar space $(\{-1, +1\})^D$
  - **-1's counter and adder**: process bipolar quantities

H3DFact Architecture – Circuit Details



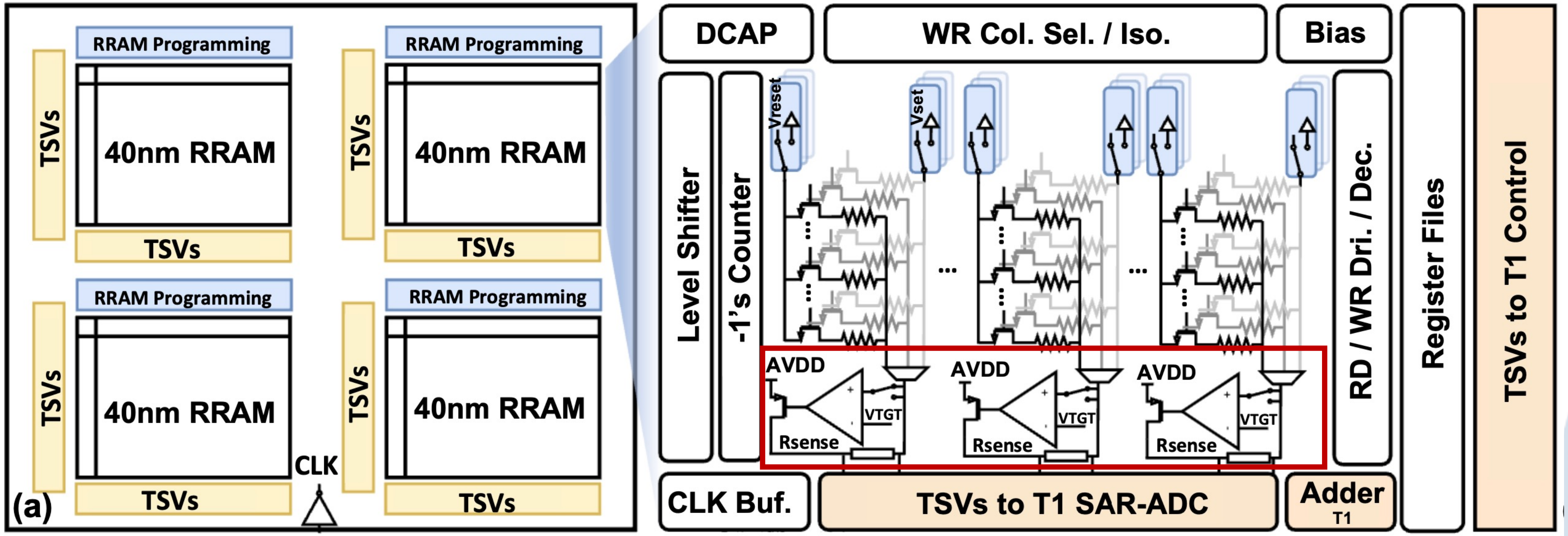
- **Circuitry**: capable of executing high-dimensional bipolar space $(\{-1, +1\})^D$
- **Isolated switches**: protect peripherals against high-voltages for RRAM setting and resetting

H3DFact Architecture – Circuit Details



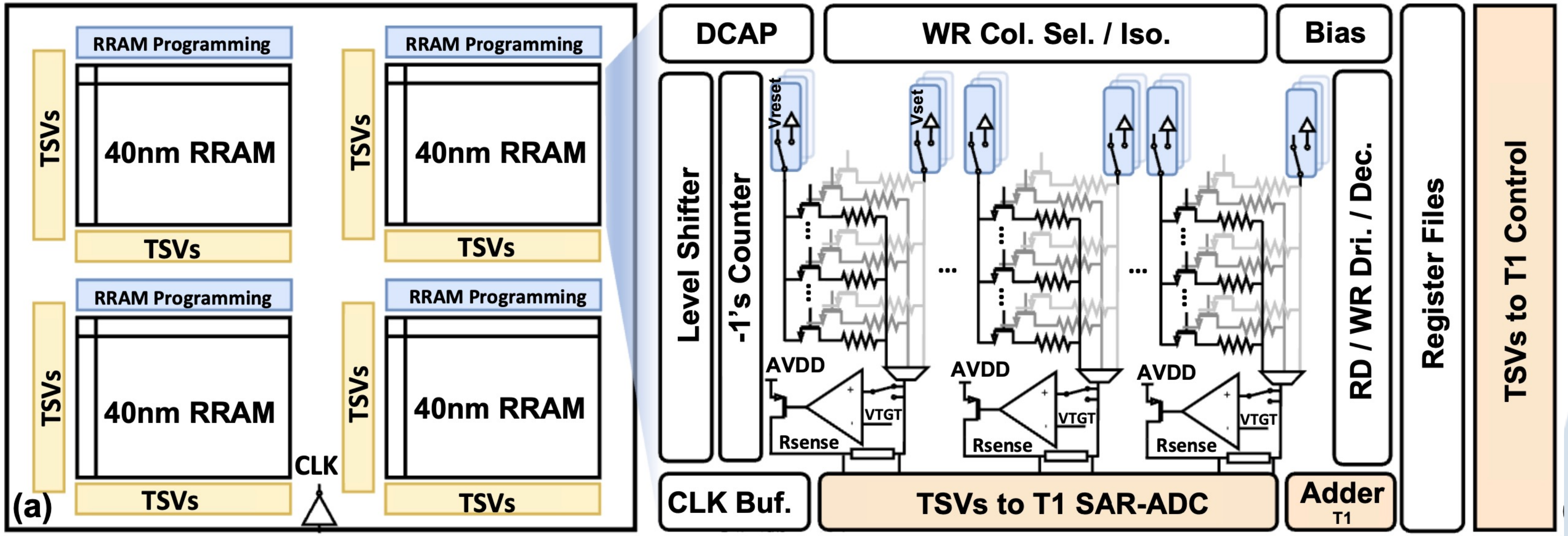
- **Circuitry**: capable of executing high-dimensional bipolar space $(\{-1, +1\})^D$
- **Isolated switches**: protect peripherals against high-voltages for RRAM setting and resetting
- **Voltage regulation**: power supply (AVDD) with operational amplifier

H3DFact Architecture – Circuit Details



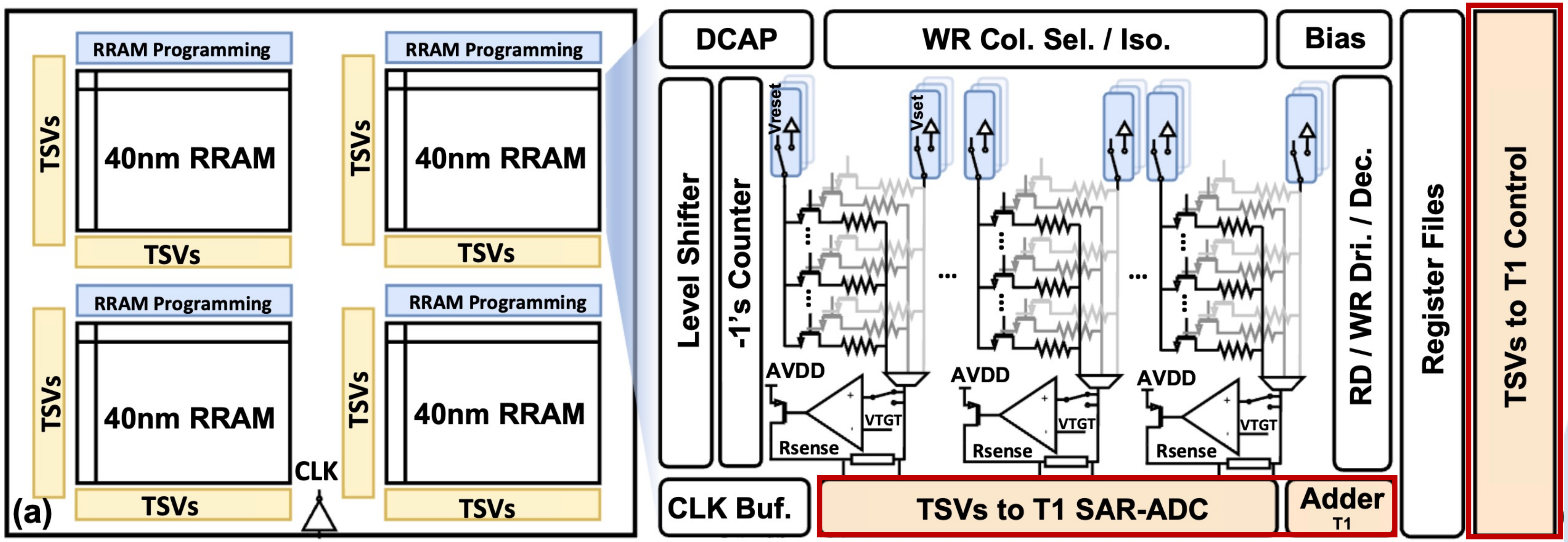
- **Circuitry**: capable of executing high-dimensional bipolar space ({-1, +1})^D
- **Isolated switches**: protect peripherals against high-voltages for RRAM setting and resetting
- **Voltage regulation**: power supply (AVDD) with operational amplifier
- **Current-sensing resistor (Rsense)**: enhance the process-voltage-temperature (PVT) immunity

H3DFact Architecture – Circuit Details



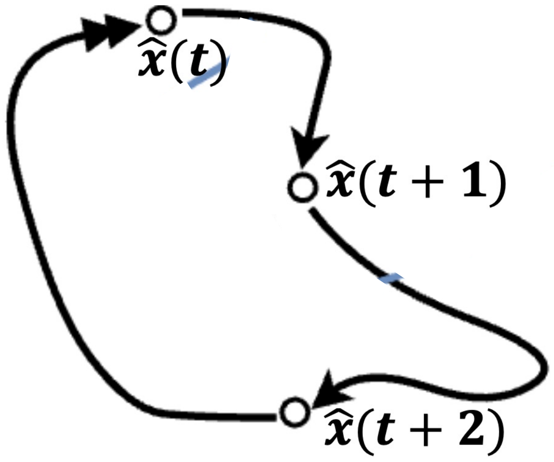
- **Circuitry**: capable of executing high-dimensional bipolar space $(\{-1, +1\})^D$
- **Isolated switches**: protect peripherals against high-voltages for RRAM setting and resetting
- **Voltage regulation**: power supply (AVDD) with operational amplifier
- **Current-sensing resistor (Rsense)**: enhance the process-voltage-temperature (PVT) immunity
- **Power mode**: allow for different power-off models when enabling other tiers to remain active

H3DFact Architecture – Circuit Details



- **Circuitry**: capable of executing high-dimensional bipolar space $(\{-1, +1\})^D$
- **Isolated switches**: protect peripherals against high-voltages for RRAM setting and resetting
- **Voltage regulation**: power supply (AVDD) with operational amplifier
- **Current-sensing resistor (Rsense)**: enhance the process-voltage-temperature (PVT) immunity
- **Power mode**: allow for different power-off models when enabling other tiers to remain active
- **Hybrid memory**: RRAM for read-intensive operations, SRAM for write-intensive operations

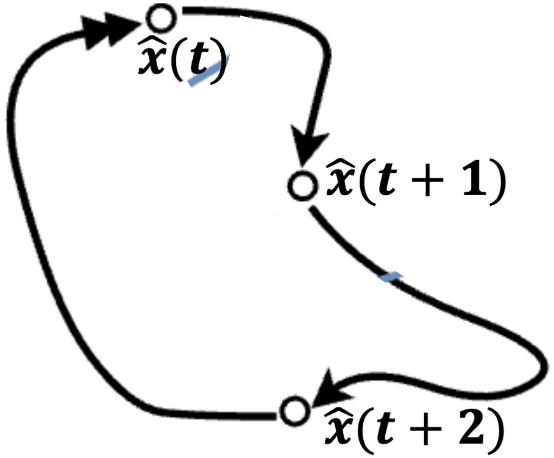
H3DFact Architecture – Stochastic Factorizer

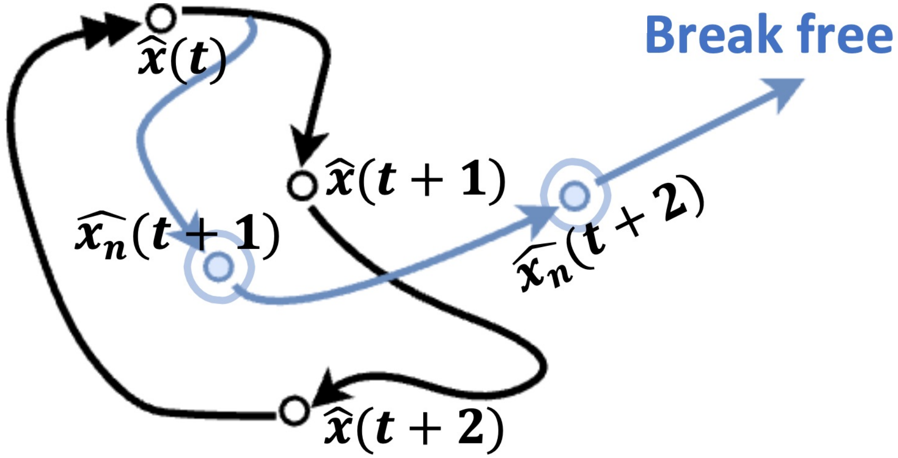


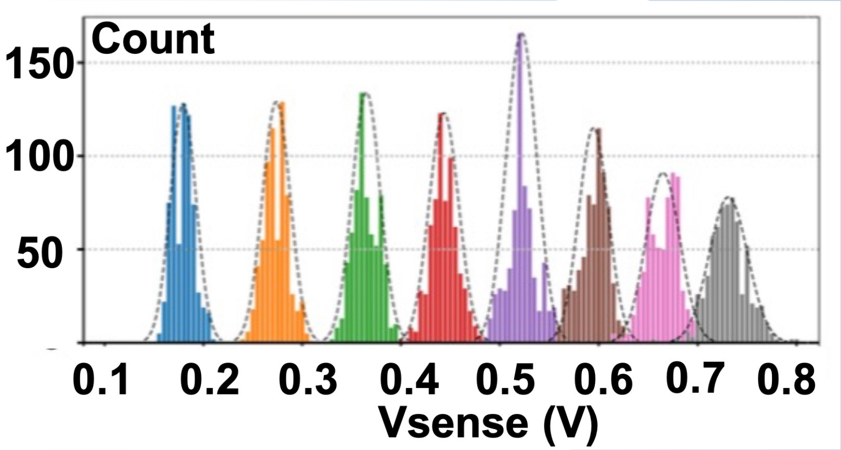
Deterministic factorization

Stuck in the local minima
and long convergence time

H3DFact Architecture – Stochastic Factorizer







Deterministic factorization

Stuck in the local minima and long convergence time

H3D RRAM/SRAM-based stochastic factorization

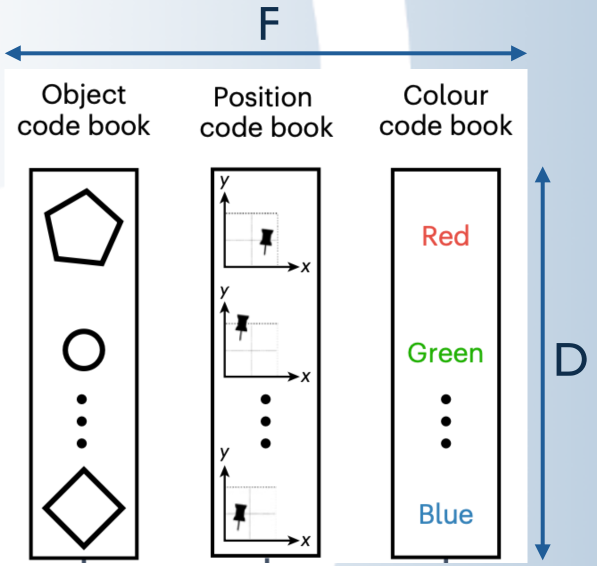
Intrinsic stochasticity of memristive devices can break being stuck at limited cycles, enabling ability to explore larger space

Outline

- Hierarchical Cognition
- Background – Holographic Vector Factorization
- H3DFact
  - Architecture
  - Floorplan
  - Interconnect
  - Circuitry
- **Evaluation Results**
- Conclusion

Evaluation – Accuracy and Operational Capacity

	Factorization Accuracy (%)			
	F=3		F=4	
	Baseline	H3D	Baseline	H3D
D=16	99.4	99.3	99.2	99.2
D=32	99.3	99.3	99.1	99.2
D=64	99.1	99.3	89.9	99.2
D=128	96.9	99.3	0	99.2
D=256	10.8	99.2	0	99.2
D=512	0.2	99.2	0	99.2



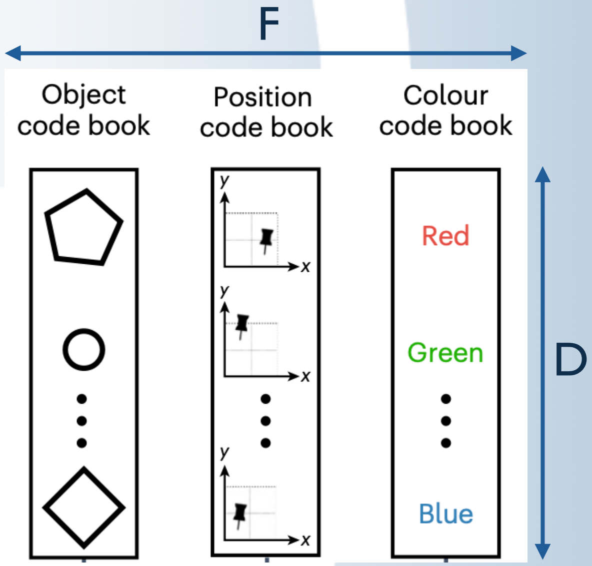
F: number of attributes
D: vector dimensions

H3DFact enhances and maintains accuracy under high dimensionality -> improved scalability and operational capacity

Evaluation – Accuracy and Operational Capacity

	Factorization Accuracy (%)				Number of Iterations[*]			
	F=3		F=4		F=3		F=4	
	Baseline	H3D	Baseline	H3D	Baseline	H3D	Baseline	H3D
D=16	99.4	99.3	99.2	99.2	4	5	31	33
D=32	99.3	99.3	99.1	99.2	13	15	234	140
D=64	99.1	99.3	89.9	99.2	43	39	Fail	1347
D=128	96.9	99.3	0	99.2	Fail	108	Fail	17529
D=256	10.8	99.2	0	99.2	Fail	443	Fail	269931
D=512	0.2	99.2	0	99.2	Fail	1685	Fail	2824079

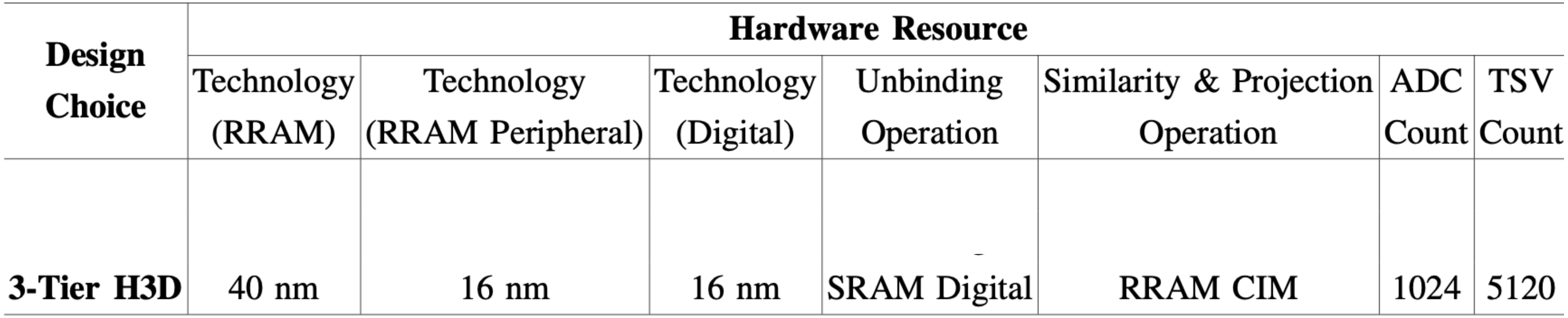
[*] Number of iterations required to reach at least 99% accuracy under different problem sizes.



F: number of attributes
D: vector dimensions

H3DFact enhances and maintains accuracy under high dimensionality -> improved scalability and operational capacity
H3DFACT enables faster convergence and solves larger problem -> lowering computational cost

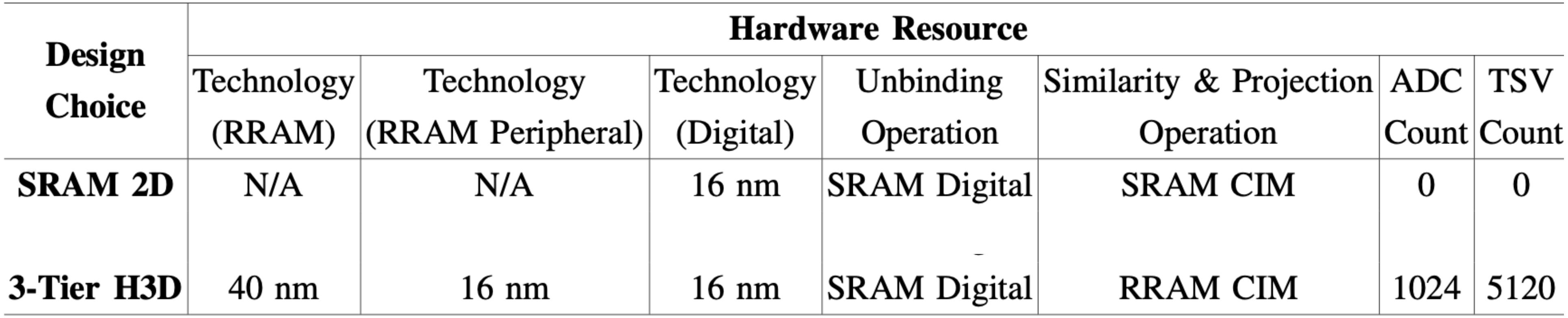
Evaluation – Hardware Efficiency



Ours

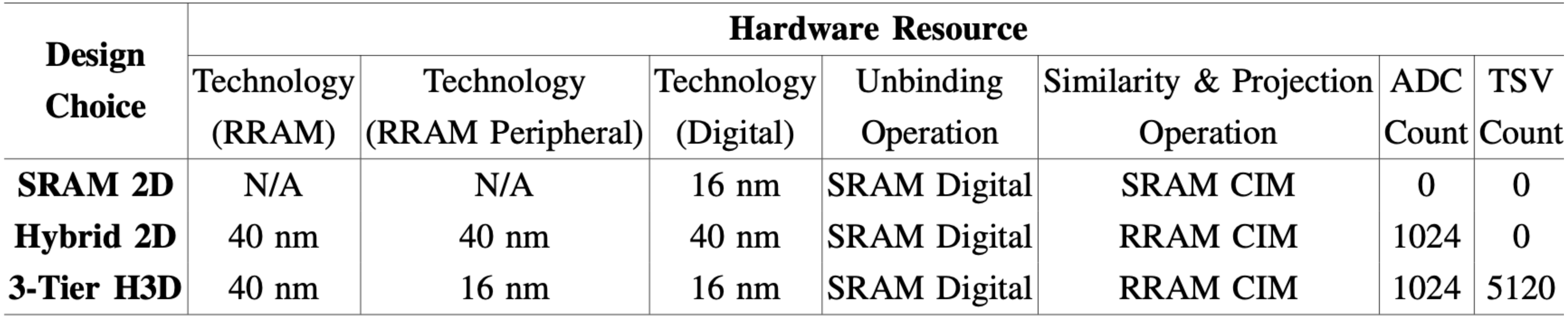
Design Choice	Hardware Resource						
	Technology (RRAM)	Technology (RRAM Peripheral)	Technology (Digital)	Unbinding Operation	Similarity & Projection Operation	ADC Count	TSV Count
3-Tier H3D	40 nm	16 nm	16 nm	SRAM Digital	RRAM CIM	1024	5120

Evaluation – Hardware Efficiency



	Design Choice	**Hardware Resource**						
		Technology (RRAM)	Technology (RRAM Peripheral)	Technology (Digital)	Unbinding Operation	Similarity & Projection Operation	ADC Count	TSV Count
Baseline	**SRAM 2D**	N/A	N/A	16 nm	SRAM Digital	SRAM CIM	0	0
Ours	**3-Tier H3D**	40 nm	16 nm	16 nm	SRAM Digital	RRAM CIM	1024	5120

Evaluation – Hardware Efficiency



	Design Choice	Hardware Resource						
		Technology (RRAM)	Technology (RRAM Peripheral)	Technology (Digital)	Unbinding Operation	Similarity & Projection Operation	ADC Count	TSV Count
Baseline	**SRAM 2D**	N/A	N/A	16 nm	SRAM Digital	SRAM CIM	0	0
Baseline	**Hybrid 2D**	40 nm	40 nm	40 nm	SRAM Digital	RRAM CIM	1024	0
Ours	**3-Tier H3D**	40 nm	16 nm	16 nm	SRAM Digital	RRAM CIM	1024	5120

Evaluation – Hardware Efficiency

	Design Choice	Hardware Resource						
		Technology (RRAM)	Technology (RRAM Peripheral)	Technology (Digital)	Unbinding Operation	Similarity & Projection Operation	ADC Count	TSV Count
Baseline	**SRAM 2D**	N/A	N/A	16 nm	SRAM Digital	SRAM CIM	0	0
Baseline	**Hybrid 2D**	40 nm	40 nm	40 nm	SRAM Digital	RRAM CIM	1024	0
Ours	**3-Tier H3D**	40 nm	16 nm	16 nm	SRAM Digital	RRAM CIM	1024	5120

	Design Choice	Performance					
		Area	Frequency	Throughput	Compute Density	Energy Efficiency	Accuracy
Baseline	**SRAM 2D**	0.114 mm^2					
Baseline	**Hybrid 2D**	0.544 mm^2					
Ours	**3-Tier H3D**	0.091 mm^2					

Compared to fully SRAM 2D and hybrid SRAM/RRAM 2D design, H3DFact achieves more compact silicon footprint,

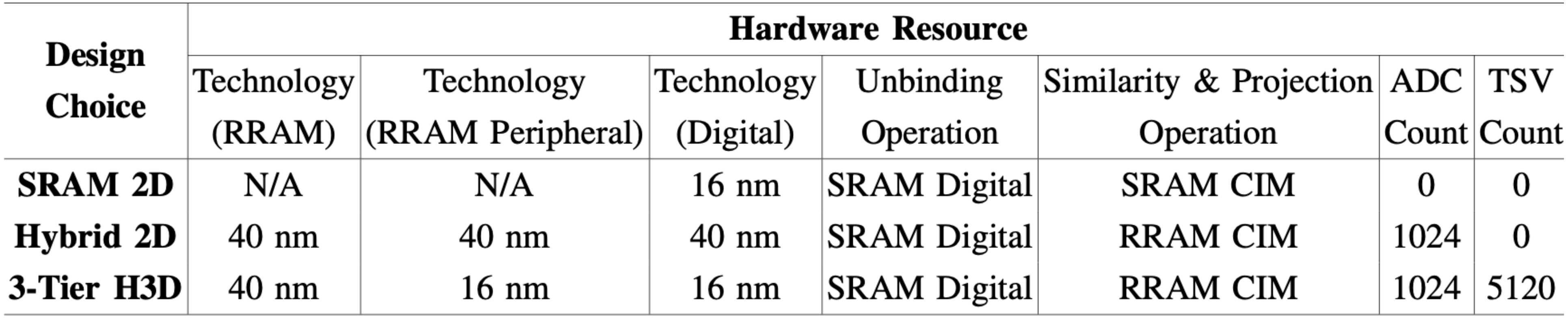
Evaluation – Hardware Efficiency

	Design Choice	**Hardware Resource**						
		Technology (RRAM)	Technology (RRAM Peripheral)	Technology (Digital)	Unbinding Operation	Similarity & Projection Operation	ADC Count	TSV Count
Baseline	**SRAM 2D**	N/A	N/A	16 nm	SRAM Digital	SRAM CIM	0	0
Baseline	**Hybrid 2D**	40 nm	40 nm	40 nm	SRAM Digital	RRAM CIM	1024	0
Ours	**3-Tier H3D**	40 nm	16 nm	16 nm	SRAM Digital	RRAM CIM	1024	5120

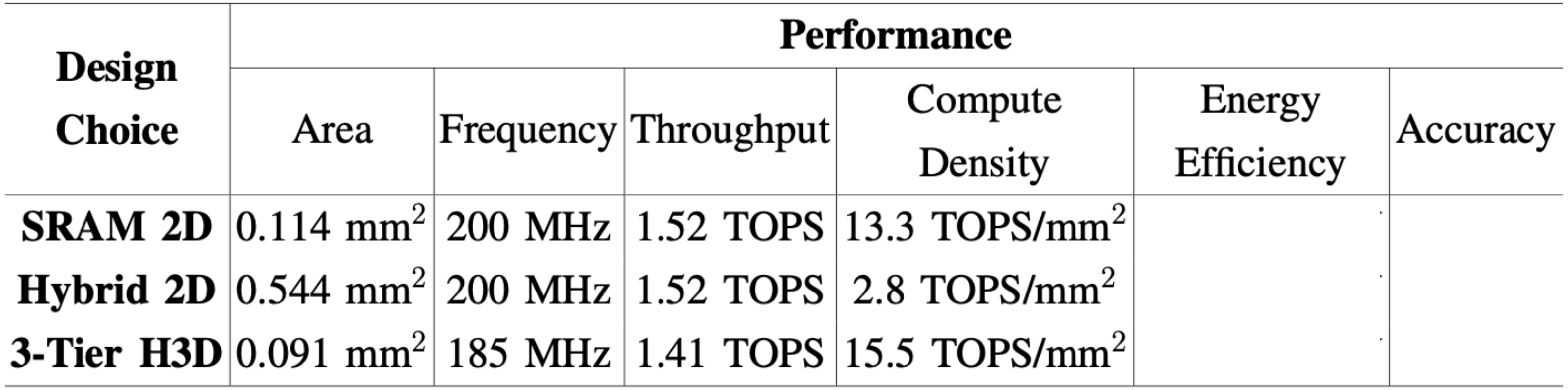
	Design Choice	**Performance**					
		Area	Frequency	Throughput	Compute Density	Energy Efficiency	Accuracy
Baseline	**SRAM 2D**	0.114 mm^2	200 MHz	1.52 TOPS			
Baseline	**Hybrid 2D**	0.544 mm^2	200 MHz	1.52 TOPS			
Ours	**3-Tier H3D**	0.091 mm^2	185 MHz	1.41 TOPS			

Compared to fully SRAM 2D and hybrid SRAM/RRAM 2D design, H3DFact achieves more compact silicon footprint,

Evaluation – Hardware Efficiency



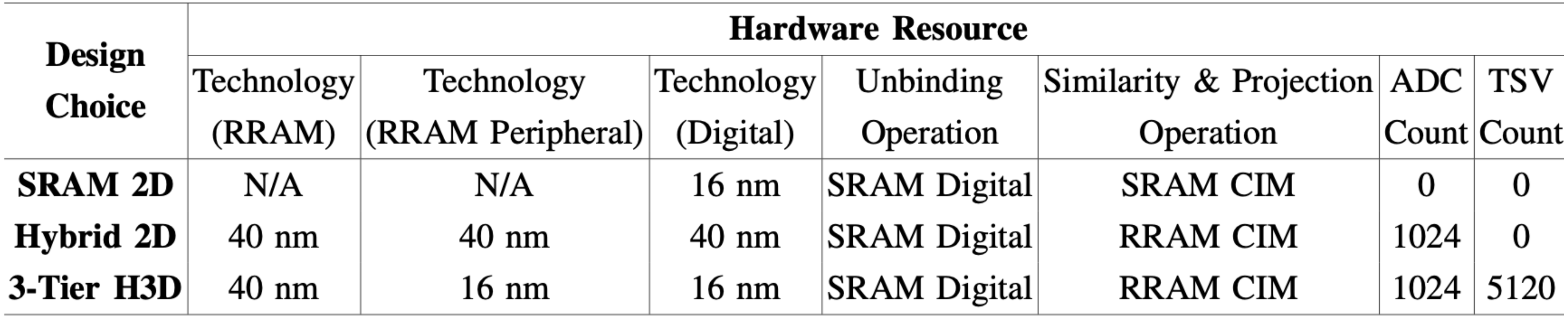
	Design Choice	Hardware Resource						
		Technology (RRAM)	Technology (RRAM Peripheral)	Technology (Digital)	Unbinding Operation	Similarity & Projection Operation	ADC Count	TSV Count
Baseline	**SRAM 2D**	N/A	N/A	16 nm	SRAM Digital	SRAM CIM	0	0
Baseline	**Hybrid 2D**	40 nm	40 nm	40 nm	SRAM Digital	RRAM CIM	1024	0
Ours	**3-Tier H3D**	40 nm	16 nm	16 nm	SRAM Digital	RRAM CIM	1024	5120



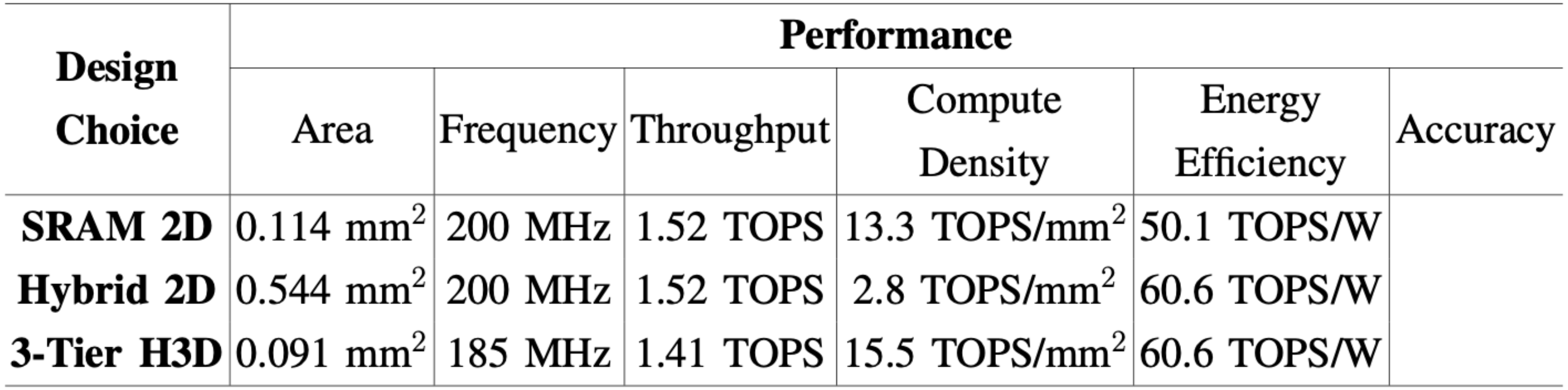
	Design Choice	Performance					
		Area	Frequency	Throughput	Compute Density	Energy Efficiency	Accuracy
Baseline	**SRAM 2D**	0.114 mm^2	200 MHz	1.52 TOPS	13.3 $TOPS/mm^2$		
Baseline	**Hybrid 2D**	0.544 mm^2	200 MHz	1.52 TOPS	2.8 $TOPS/mm^2$		
Ours	**3-Tier H3D**	0.091 mm^2	185 MHz	1.41 TOPS	15.5 $TOPS/mm^2$		

Compared to fully SRAM 2D and hybrid SRAM/RRAM 2D design, H3DFact achieves more compact silicon footprint, higher compute density,

Evaluation – Hardware Efficiency



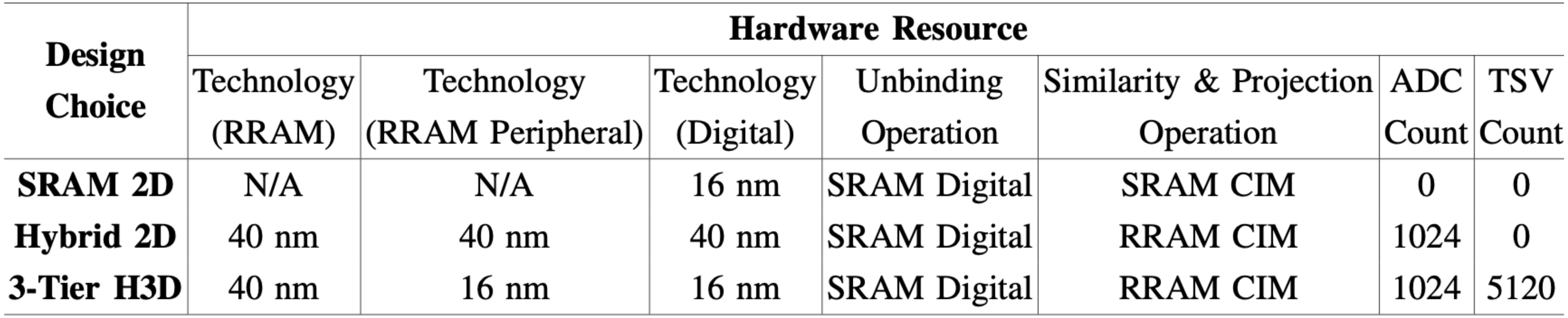
	Design Choice	**Hardware Resource**						
		Technology (RRAM)	Technology (RRAM Peripheral)	Technology (Digital)	Unbinding Operation	Similarity & Projection Operation	ADC Count	TSV Count
Baseline	**SRAM 2D**	N/A	N/A	16 nm	SRAM Digital	SRAM CIM	0	0
Baseline	**Hybrid 2D**	40 nm	40 nm	40 nm	SRAM Digital	RRAM CIM	1024	0
Ours	**3-Tier H3D**	40 nm	16 nm	16 nm	SRAM Digital	RRAM CIM	1024	5120



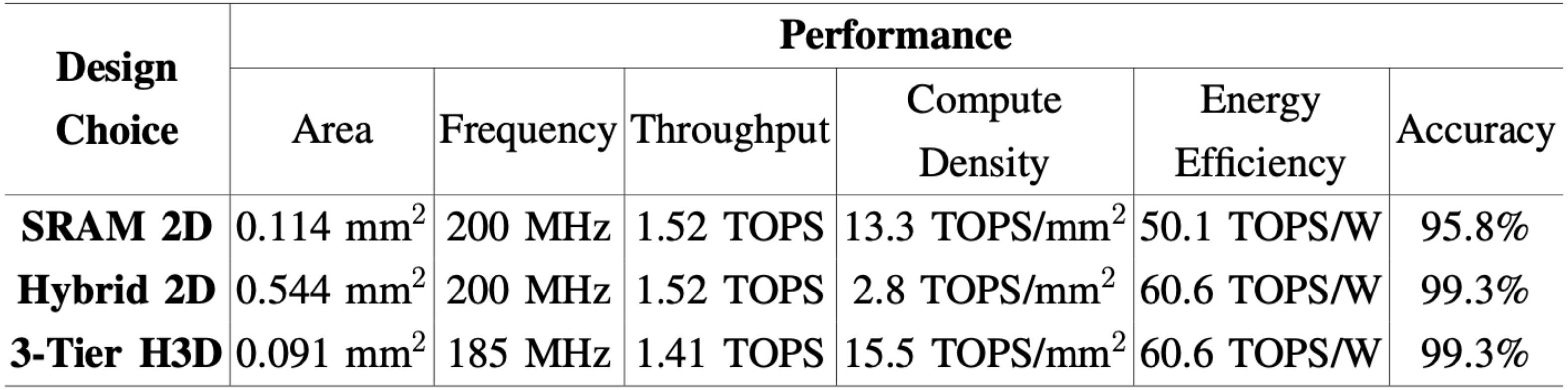
	Design Choice	**Performance**					
		Area	Frequency	Throughput	Compute Density	Energy Efficiency	Accuracy
Baseline	**SRAM 2D**	0.114 mm^2	200 MHz	1.52 TOPS	13.3 $TOPS/mm^2$	50.1 TOPS/W	
Baseline	**Hybrid 2D**	0.544 mm^2	200 MHz	1.52 TOPS	2.8 $TOPS/mm^2$	60.6 TOPS/W	
Ours	**3-Tier H3D**	0.091 mm^2	185 MHz	1.41 TOPS	15.5 $TOPS/mm^2$	60.6 TOPS/W	

Compared to fully SRAM 2D and hybrid SRAM/RRAM 2D design, H3DFact achieves more compact silicon footprint, higher compute density, and higher energy efficiency

Evaluation – Hardware Efficiency



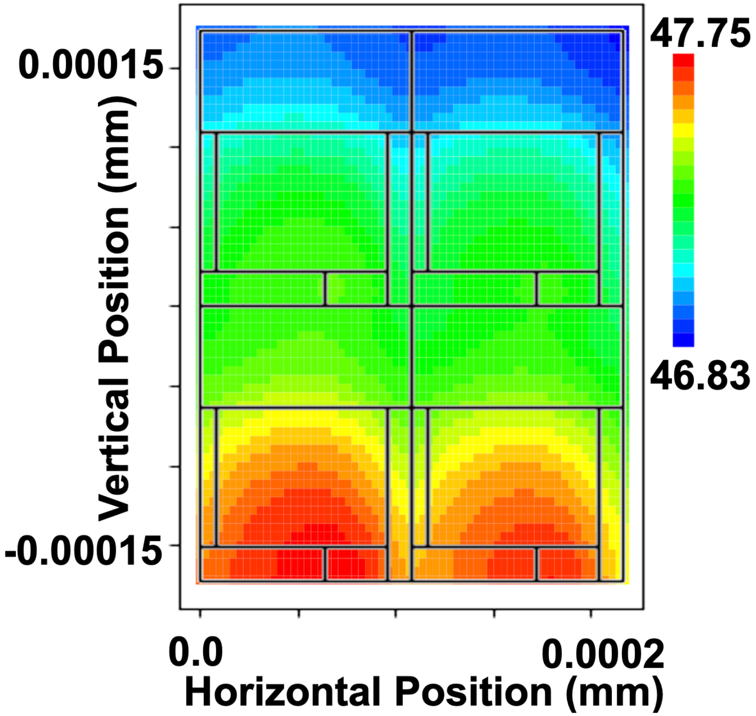
	Design Choice	Hardware Resource						
		Technology (RRAM)	Technology (RRAM Peripheral)	Technology (Digital)	Unbinding Operation	Similarity & Projection Operation	ADC Count	TSV Count
Baseline	**SRAM 2D**	N/A	N/A	16 nm	SRAM Digital	SRAM CIM	0	0
Baseline	**Hybrid 2D**	40 nm	40 nm	40 nm	SRAM Digital	RRAM CIM	1024	0
Ours	**3-Tier H3D**	40 nm	16 nm	16 nm	SRAM Digital	RRAM CIM	1024	5120



	Design Choice	Performance					
		Area	Frequency	Throughput	Compute Density	Energy Efficiency	Accuracy
Baseline	**SRAM 2D**	0.114 mm^2	200 MHz	1.52 TOPS	13.3 TOPS/mm^2	50.1 TOPS/W	95.8%
Baseline	**Hybrid 2D**	0.544 mm^2	200 MHz	1.52 TOPS	2.8 TOPS/mm^2	60.6 TOPS/W	99.3%
Ours	**3-Tier H3D**	0.091 mm^2	185 MHz	1.41 TOPS	15.5 TOPS/mm^2	60.6 TOPS/W	99.3%

Compared to fully SRAM 2D and hybrid SRAM/RRAM 2D design, H3DFact achieves more compact silicon footprint, higher compute density, and higher energy efficiency

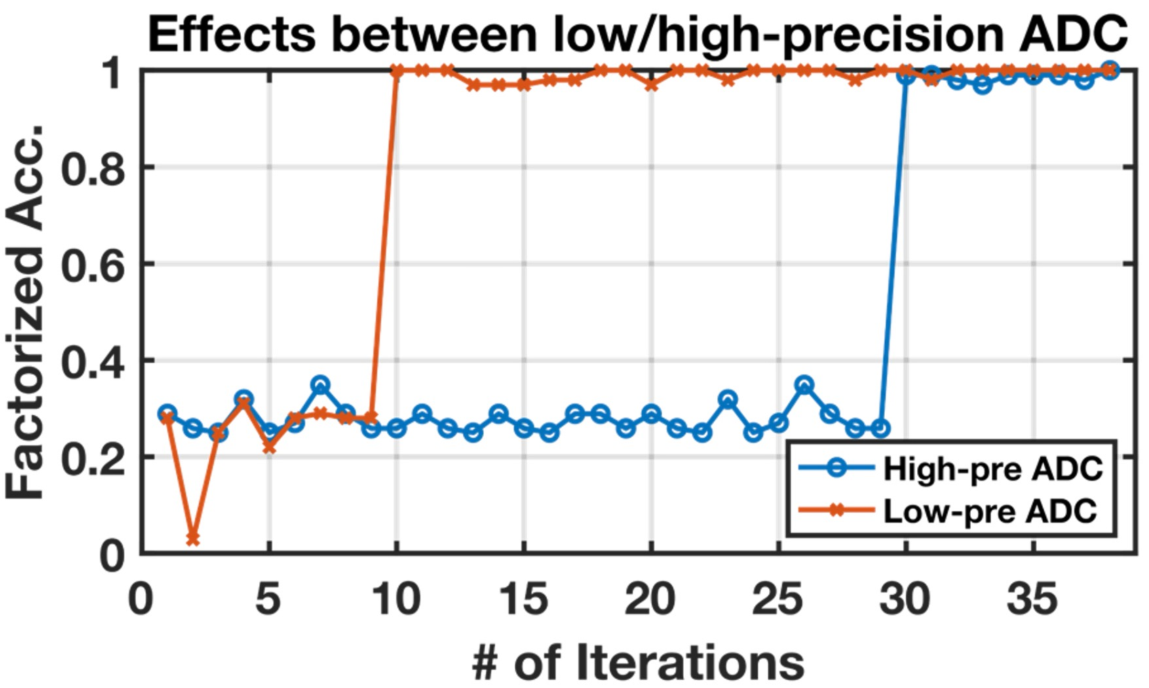
Evaluation – Thermal Analysis



Attribute	Value
Number of tiers	3
PCB thickness	2 mm
Bumping thickness	100 μm
Package thickness	1 mm
TIM thickness	TIM1: 20 μm TIM2: 20 μm
Heat transfer coefficient	1000 (W/m^2℃)
Ambient temperature	25℃

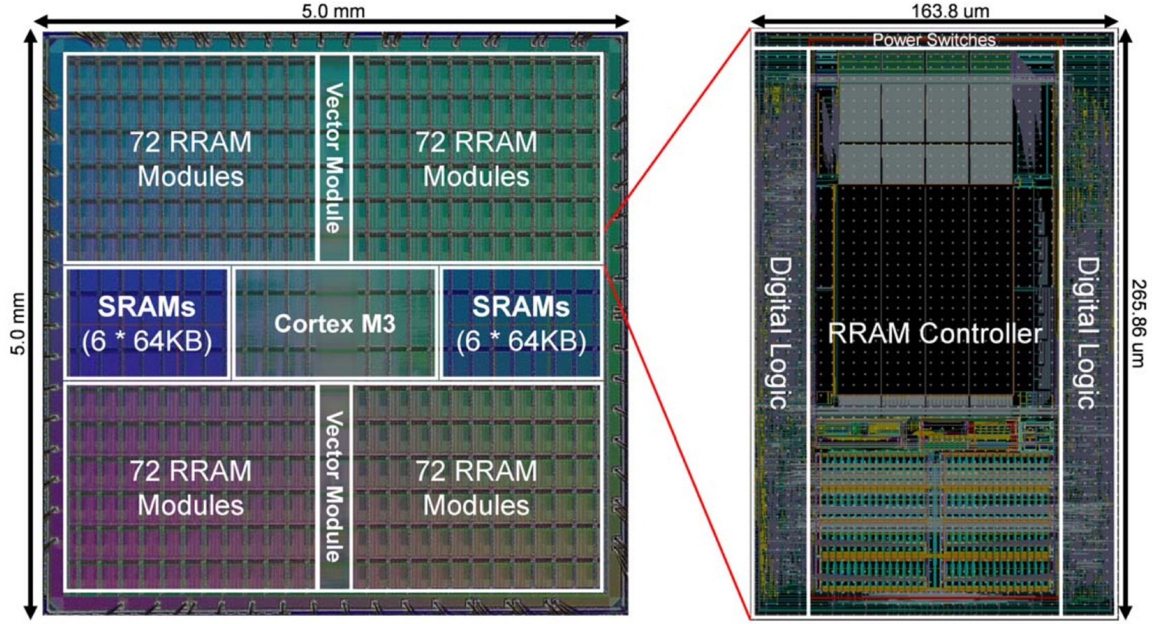
H3DFact tier temperature ranges from 46.8-47.8℃ (2D design 44℃), within SRAM/RRAM thermal limits

Evaluation – Robustness and Convergence Speedup

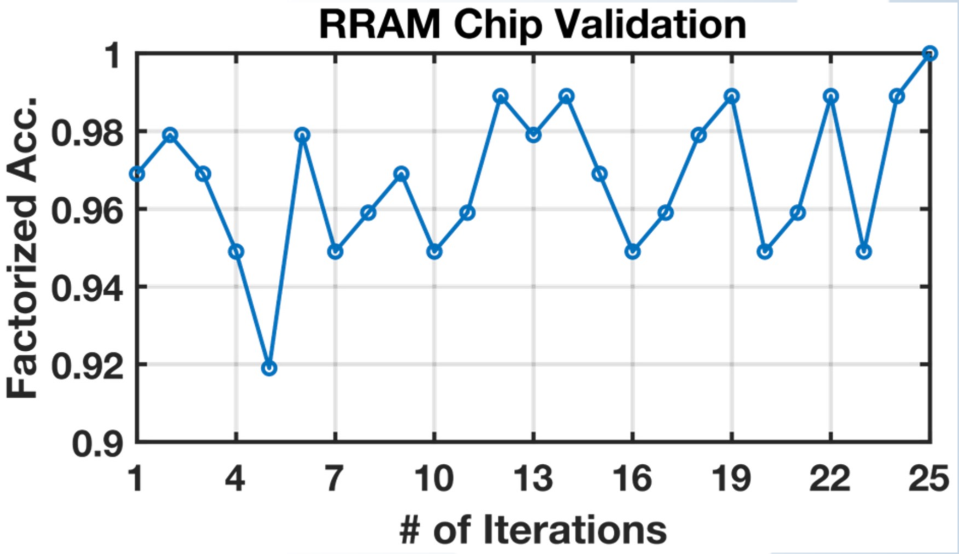


Lowering ADC precision can reduce hardware costs and enable faster convergence of holographic perceptual factorization with similar accuracy.

Evaluation – 2D RRAM Chip Validation



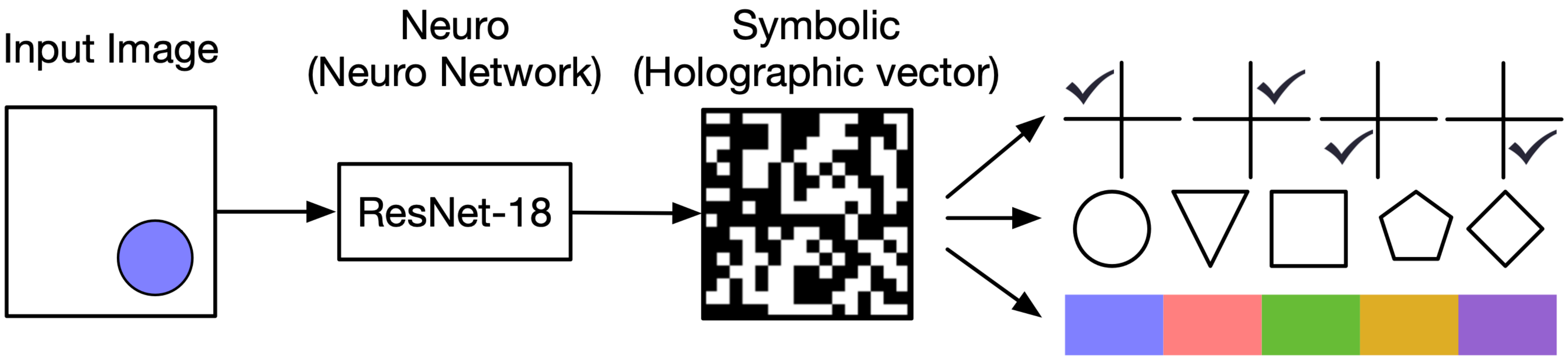
Technology	TSMC 40nm ULP	**General Purpose IOs**	32 (16*2)
Chip Size	5 mm x 5 mm	**Debug Interface**	JTAG / Serial Wire(SW)
Package	QFN7x7-60	**Voltage Domains**	6 + VSS
Embedded Microprocessor	Cortex M3	**Low Power Design Technique**	Clock gating / Power gating
Number of RRAM Module	288	**Clock Source / Max. Clock Rate**	Crystal or External / 200 MHz
On-Chip RRAM / SRAM	2.25 MB / 768 KB	**Core / IO Supply Voltage**	0.9 V / 3.3 V



Chang et al, "A 40nm RRAM/SRAM system with embedded cortex M3 microprocessor for edge recommendation systems", ISSCC, 2022

Fabricated TSMC 40nm RRAM testchip validated H3DFACT achieves > 96% factorization accuracy at one-shot and reaches 99% accuracy after 25 iterations

Evaluation – Holographic Perception Task



Evaluated on the relational and analogical visual reasoning (RAVEN) dataset, H3DFACT achieves 99.4% accuracy of attributes estimation

Outline

- Hierarchical Cognition
- Background – Holographic Vector Factorization
- H3DFact
  - Architecture
  - Floorplan
  - Interconnect
  - Circuitry
- Evaluation Results
- **Conclusion**

Conclusion

- Needs to factorize holographic perceptual representation efficiently & scalable
  - Fundamental to human-like hierarchical cognitive progress
  - Factorization is challenging: intensive compute, limited scalability, suboptimal stuck
- **H3DFact: towards efficient and scalable holographic factorization**
  - Heterogeneous 3D architecture
  - Hybrid SRAM/RRAM compute-in-memory
  - Intrinsic stochasticity for improved convergence

Reach to me at: zishenwan@gatech.edu
Learn more about our work at: http://zishenwan.github.io

H3DFact Paper

Semiconductor
Research
Corporation®