AuxiliarySRAM: Exploring Elastic On-Chip Memory in 2.5D Chiplet Systems Design

Zichao Ling, Lin Li, Yi Huang, Yixin Xuan, Jianwang Zhai* ,Kang Zhao

Beijing University of Posts and Telecommunications

* Corresponding author [zhaijw@bupt.edu.cn]

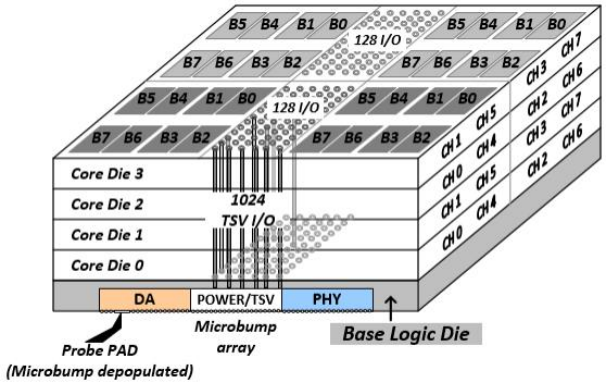
I. Intro. & Motivation

“Memory Wall”

“The growth rate of hardware peak floating-point operations per second (FLOPS) has outpaced DRAM bandwidth by approximately 100× over the past two decades”[1].

Bandwidth Bound

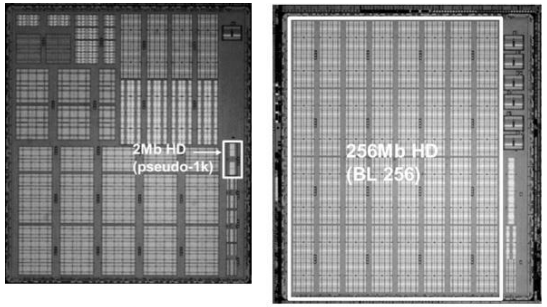
- e.g. graphics computing, AI acceleration
- High Bandwidth Memory (HBM)



(a) Structure of HBM[2]

Latency Sensitive

- e.g., autonomous driving decision-making, high frequency trading systems
- **still** rely on high-speed yet **costly** on-chip SRAM



Technology	2nm nanosheet
Metal scheme	1P7M
Supply voltage	0.75V
Bit cell size	HD: 0.021μm²
SRAM macro configuration	4096x145 MUX4 (2Mb) 4096x32 MUX16 (256Mb)
SRAM capacity	2Mb and 256Mb
Design Features	Redundancy Programmable E-fuse NBL write assist option

(b) Micrograph of 2nm SRAM array[3]

[1] Miao Liu et al. 2025. Processing-Near-Memory with Chip Level 3D-IC. In Proc. ASPDAC. 302–307.

[2]H. Jun et al., "HBM (High Bandwidth Memory) DRAM Technology and Architecture," 2017 IEEE International Memory Workshop (IMW), Monterey, CA, USA, 2017, pp. 1-4, doi: 10.1109/IMW.2017.7939084.

[3]T. -Y. J. Chang et al., "A 38.1Mb/mm2 SRAM in a 2nm-CMOS-Nanosheet Technology for High-Density and Energy-Efficient Compute," 2025 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 2025, pp. 492-494, doi: 10.1109/ISSCC49661.2025.10904759.

Limitations

***Why* can't we keep scaling on-chip SRAM limitlessly?**

- **Physical Device Limitations**

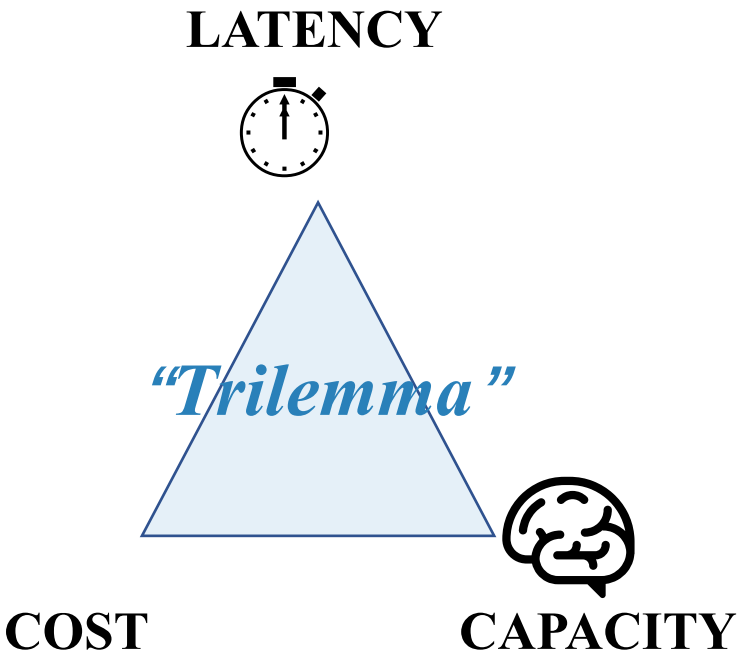
  longer addressing paths

  increased signal propagation delays

  .....

- **Manufacturing costs**

  escalate exponentially below 7nm.[*]

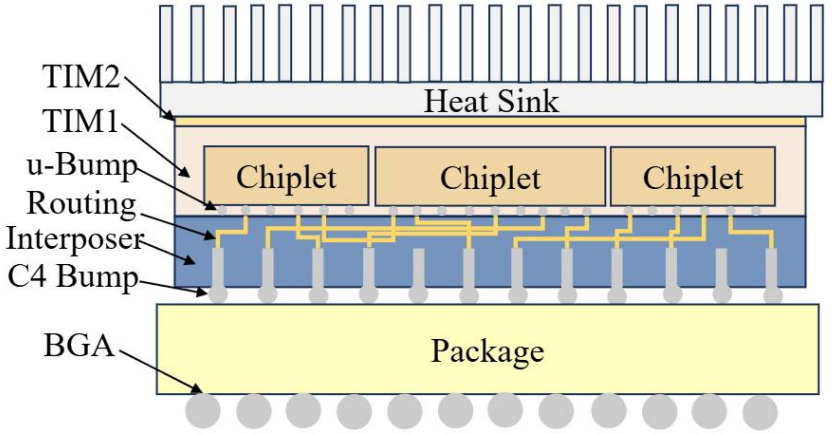


[*] Bon Woong Ku et al. 2016. How much cost reduction justifies the adoption of monolithic 3D ICs at 7nm node?. In Proc. ICCAD. 1–7.

Solution in 2.5D Chiplet Systems

2.5D Chiplet

- A heterogeneous integration technology
- Employs silicon **interposers** and through-silicon via (**TSV**) to enable modular system assembly
- Significantly alleviates the **cost overhead** of substantial design, verification, and manufacturing
- Compared to 3D ICs, **thermal** and **physical design** are more manageable.



(a)The architecture of chiplet-based 2.5D IC[1].

Achieving Higher On-Chip SRAM Capacity at Equivalent Cost

Presenting new possibilities for mitigating this *Trilemma*

[1] Shixin Chen et al. 2025. The Survey of 2.5 D Integrated Architecture: An EDA perspective. In Proc. ASPDAC. 285–293.

Why it works?

But how to deal with the inherent latency introduced by inter-Chiplet communication?

- **Elastic Design**
- **Asynchronous**

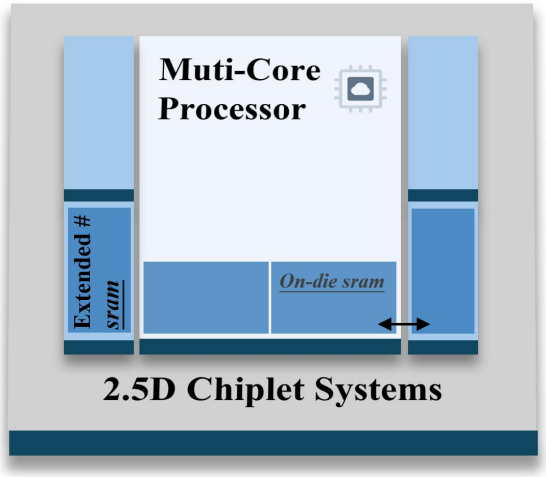
Key: *Execution Frequency within Workload*

1. Specialized SRAM for specific apps underperforms generalized cache expansion, which can maximize the memory utilization.
2. Fully offloading SRAM to chiplets incurs fully latency penalties in general-purpose scenarios.
3. Scenario-specified elastic scaling maximizes processor adaptability

Maybe it's something only *Chiplet systems* can do better ...

Concept of AuxiliarySRAM

- **On-die SRAM :**
  High-speed, low-latency local cache.
- **Extended SRAM Chiplet:**
  High-capacity remote cache incurring additional latency.
- **Die-to-Die Interface:**
  Multiple high-speed channels connecting on-die and extended SRAM.



(a) Schematic diagram

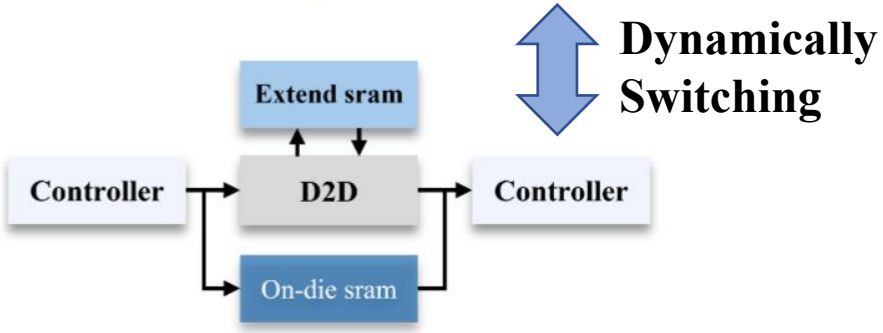
Latency First Mode

- Data processing primarily completed within on-chip SRAM
- Suitable for scenarios extremely sensitive to latency

Capacity First Mode

- Data requires additional access to extended SRAM Chiplets
- Suitable for application scenarios demanding large-scale data caching

(b) Latency first mode



(c) Capacity first mode

AuxiliarySRAM Blueprints

Blueprint A: Elastic Capacity-Latency Scaling

- Decouple SRAM resources into on-die retention and extended chiplets.
- Enable elastic switch between latency-sensitive and capacity-prioritized modes
- Dynamic scaling across multiple capacity-latency tiers

Blueprint B: Reusable Chiplet Ecosystem

- Design a multi-design compatible SRAM chiplet shared library
- Alleviate substantial design/verification/manufacturing overhead of monolithic SoCs

Main Contributions

1. **Exploring High Performance SRAM Chiplet Implementation**
   - Lightweight network-on-chip (NoC) with simplified crossbars, dual local ports, and address prediction.
   - Reduces average latency by 49.29% and boosts bandwidth by 79.35%.

2. **Fast, Automatic Design Space Exploration and Evaluation framework**
   - Evaluation framework integrated with Bayesian optimization (BO) to resolve Pareto-optimal on/off-die capacity ratios.
   - Pruning strategies achieve 1.93× speedup.
   - Provides Pareto frontier-based design guidelines.

II. Implementation

SRAM Chiplet Design：A Lightweight NoC

Why is the lightweight NoC?

- **Scalability :** Enables flexible multi-specification library design through modular bank replication.
- **Simplicity :** Reduces physical design complexity and NRE costs with streamlined crossbar arbitration.
- **Fast :** Achieves 49.29% latency reduction gain via architectural minimalism.

Core Features:

① Thin Crossbar

Simplify the crossbar based on data access frequency to reduce complexity.

② Dual Local Ports

Optimize local access and cross-border access latency.

③ Address Prediction

Reduce cross-border transmission overhead and accelerate routing.

④ Grouped Addressing

Supports flexible address space configuration, enhancing system flexibility.

Network Design

- Router
  - virtual channel allocators,
  - crossbar switches,
  - buffers, etc.
- 2D Mesh
- XY strategy

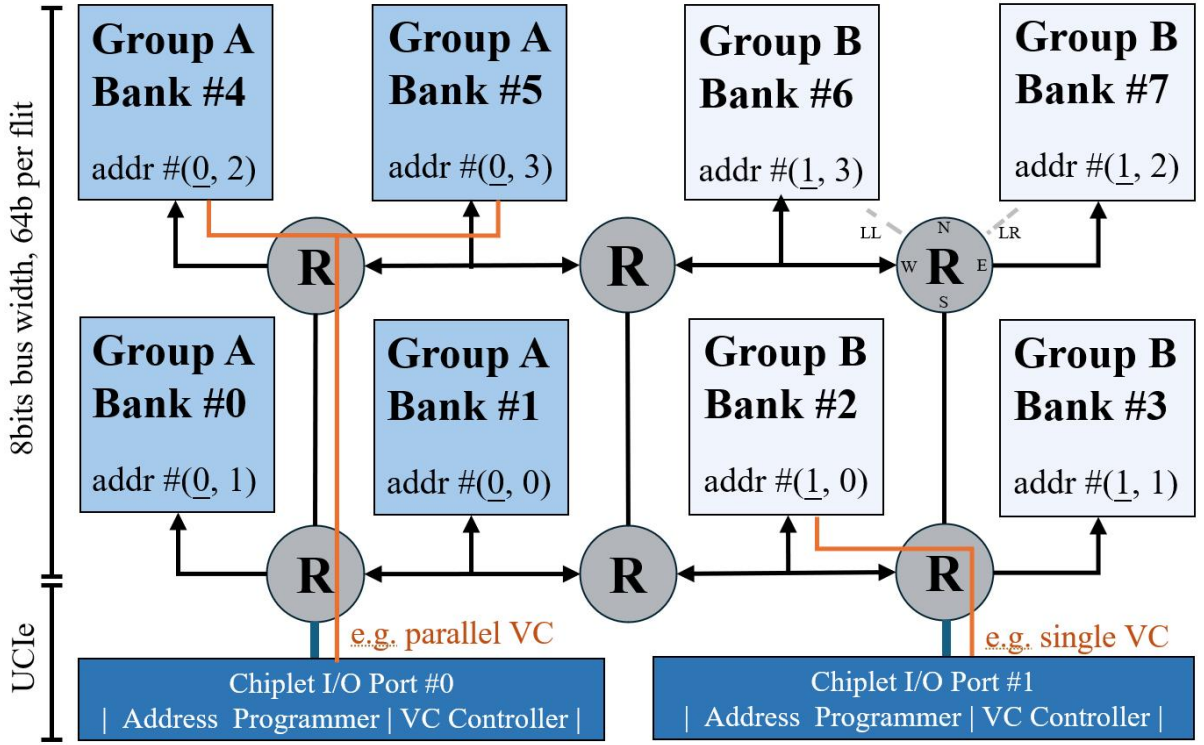


Figure: 2×4 SRAM Bank Mesh Chiplet with dual D2D-ports.
Each memory bank is configured with an 8 KB capacity, 64-bit
data width, and 1 read/write port shared by up to 2 routers.

Optimization I & II : Thin Crossbar & Dual Local

Why

- Optimization for Stream Request
- Avoid Crossbar Bloat
- Traffic Asymmetry Exploitation

  (blue vs red)

How

- Passthrough for N/S, E/W Pairs
- Mediation via Local Port
- Distributed Arbitration

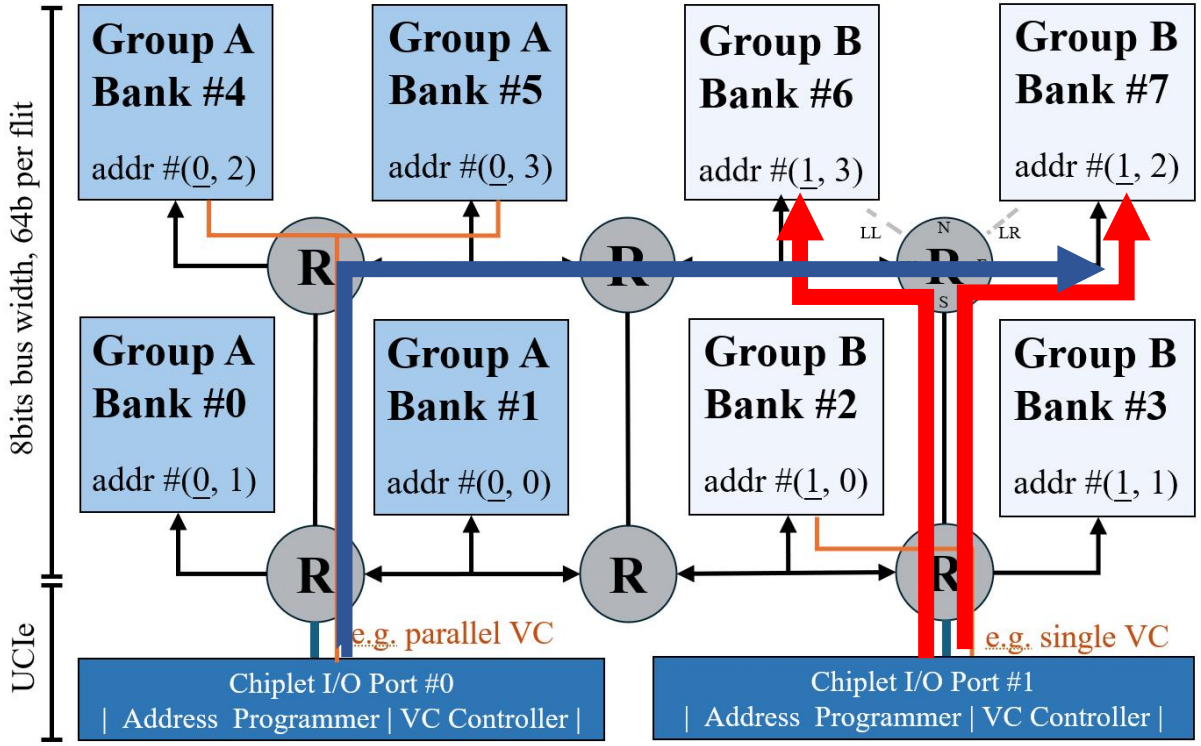


Figure: 2×4 SRAM Bank Mesh Chiplet with dual D2D-ports.
Each memory bank is configured with an 8 KB capacity, 64-bit data width, and 1 read/write port shared by up to 2 routers.

Optimization III：Address Prediction

Why

- Mitigate cross-boundary latency for burst access patterns

Key Features

- Sliding Window Register (W)
- Gradient-based ΔA_i Detection
- Pre-allocate VCs via Priority Mask
- Distributed Power-gating Synergy

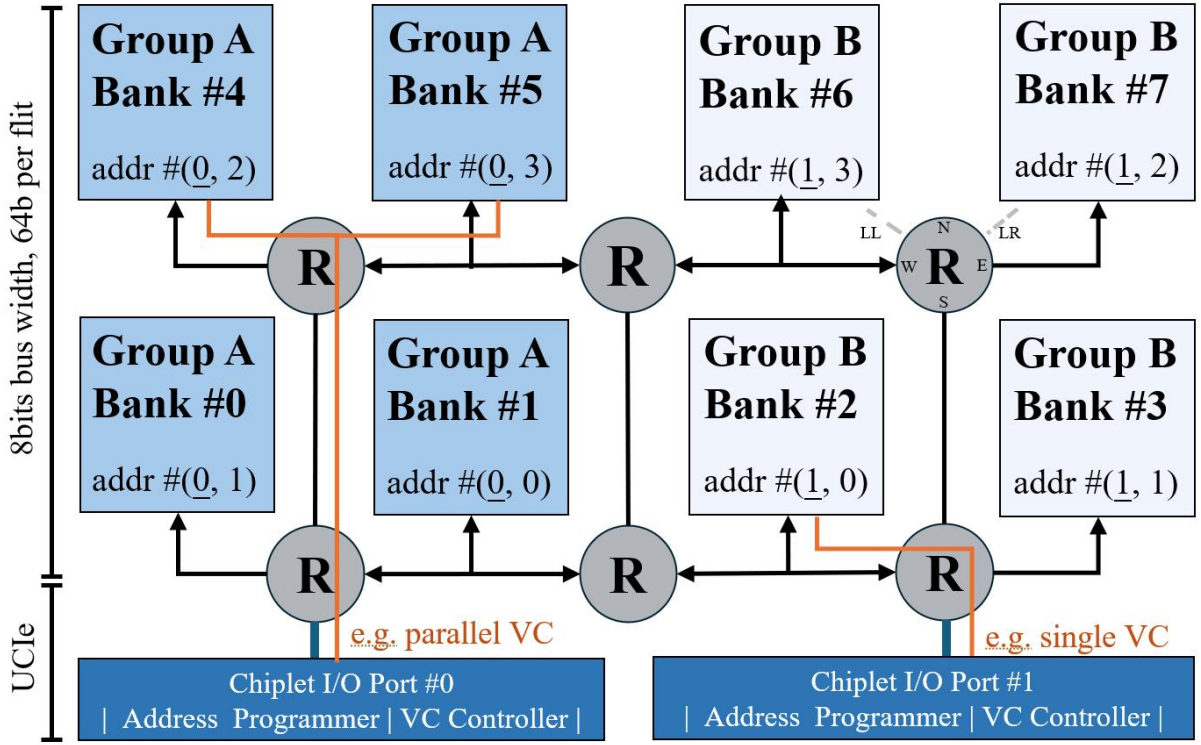


Figure: 2×4 SRAM Bank Mesh Chiplet with dual D2D-ports.
Each memory bank is configured with an 8 KB capacity, 64-bit
data width, and 1 read/write port shared by up to 2 routers.

Optimization IV: Grouped Addressing

Why

- Eliminate rigid physical-logical mapping for workload-aware bank allocation

How

- Decouple physical banks from logical space
- Δ-offset redistributes hot-zone addresses remapping

$$\mathcal{M}(A) = (A \gg (n-k)) \mod B, \quad A_C = A \mod C. \tag{1}$$

$$\mathcal{M}'(A) = ((A \oplus \Delta) \gg (n-k)) \mod B. \tag{2}$$

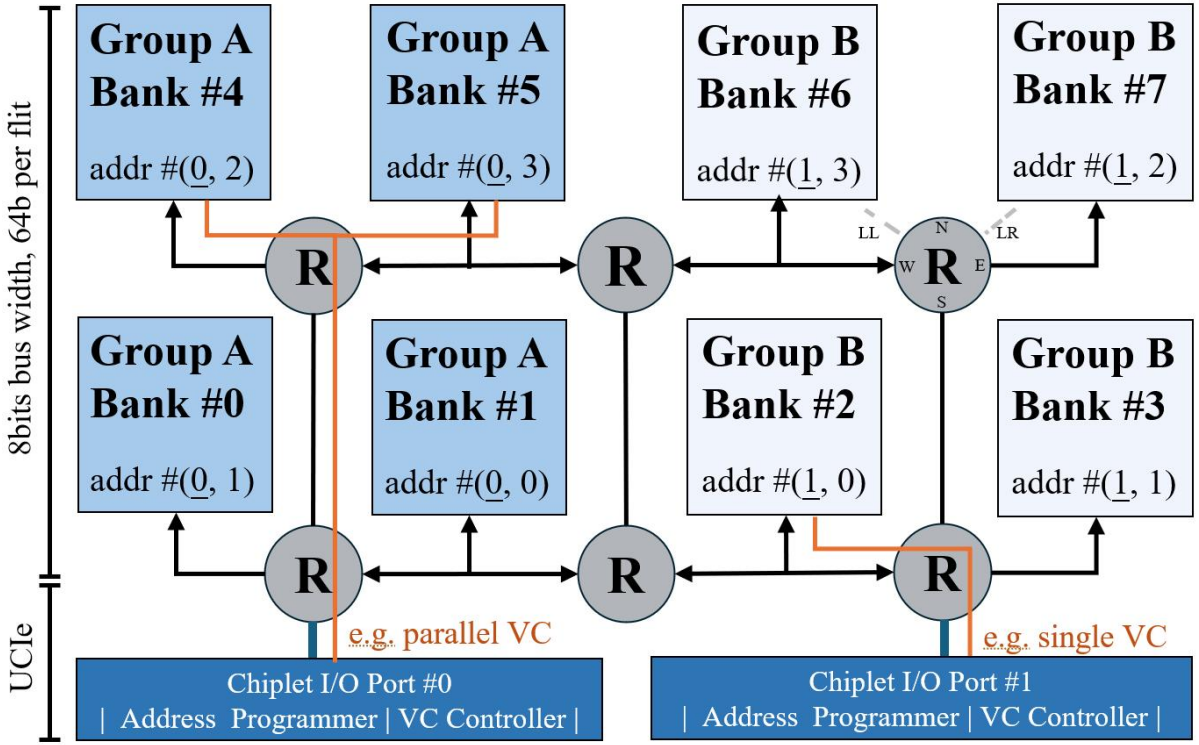


Figure: 2×4 SRAM Bank Mesh Chiplet with dual D2D-ports.
Each memory bank is configured with an 8 KB capacity, 64-bit
data width, and 1 read/write port shared by up to 2 routers.

III. Evaluation Framework

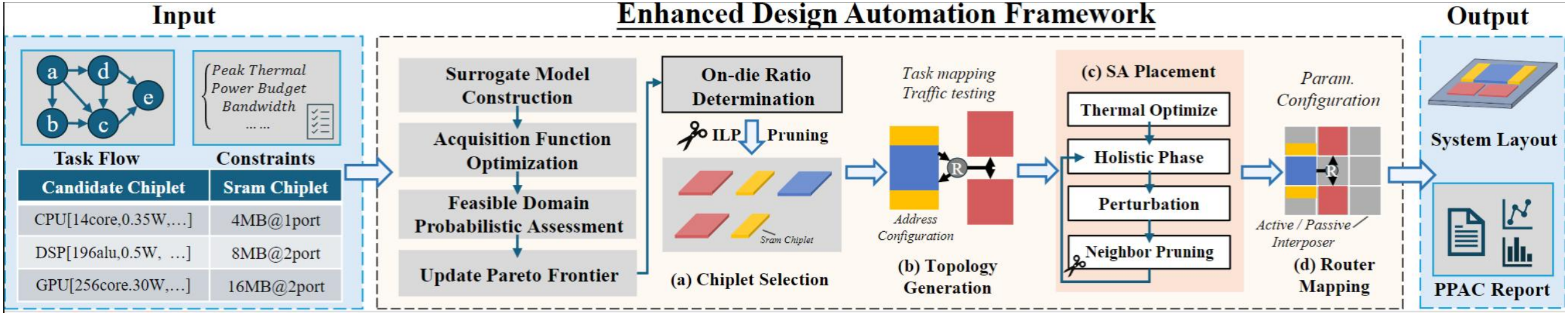
GIA[*] + BO-GP + Pruning

Basic GIA Framework:

- **(a) Chiplet Selection**
- **(b) Topology Generation**
- **(c) Chiplet Placement**
- **(d) Interposer Mapping**

Augmented Part :

- **BO-GP (Bayesian Optimization with Gaussian Processes) :**
  - Resolving Pareto-optimal on/off-die capacity ratios
  - Quantifies Pareto improvement
  - Balances multi-objective conflicts
- **Pruning in ILP and SA Placement for Acceleration**



[*]F. Li et al., "GIA: A Reusable General Interposer Architecture for Agile Chiplet Integration," 2022 IEEE/ACM International Conference On Computer Aided Design (ICCAD), San Diego, CA, USA, 2022, pp. 1-9.

Quantitative Modeling

1. Latency Model (f_{latency})

$$f_{\text{latency}} = \underbrace{\alpha_1 H_{\text{on}}}_{\text{On-chip}} + \underbrace{\gamma_1 \frac{D_{\text{data}}}{B_{\text{link}}}}_{\text{Interconnect}} + \underbrace{\alpha_2 (1 - H_{\text{on}}) \cdot \left(\beta_1 + \beta_2 \frac{C_{\text{off}}}{D_{\text{block}}} \right)}_{\text{Chiplet Access}}$$

2. Power Model (f_{power})

$$f_{\text{power}} = P_{\text{leakage}} + P_{\text{dynamic}}$$

Leakage Power

$$P_{\text{leakage}} = I_{\text{leak}} \cdot V_{dd} \cdot A_{\text{sram}} + I_{tsv\_leak} \cdot V_{tsv}$$

Dynamic Power:

$$P_{\text{dynamic}} = \frac{1}{T_{\text{task}}} \sum_{i \in \{\text{on, off, link}\}} E_i \cdot N_i$$

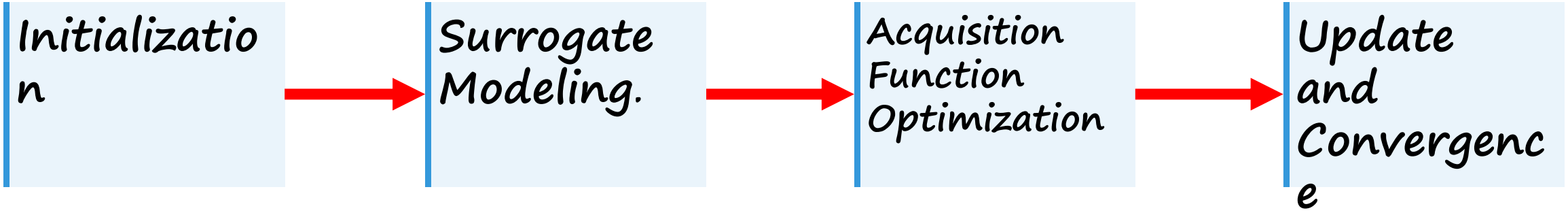
3. Cost Model (f_{cost})

$$f_{\text{cost}} = \frac{1}{Y_{\text{assembly}}} \left(\sum Co_{\text{die},i} + Co_{\text{assembly}} \right)$$

Yield Calculation:

$$Y_{\text{assembly}} = Y_{\text{align}}^{N_c} \times Y_{\text{bond}}^{N_p}$$

BO-GP



- **Efficient Search:**

  Rapid identification of optimal solutions in high-dimensional, non-convex design spaces.

- **Probabilistic Modeling:**

  Evaluating design feasibility and performance uncertainty

- **Accelerated Convergence:**

  Significantly fewer evaluations required for convergence

- **Architecture/Process Agnosticism**

Pruning Optimization

Why

- **GIA Framework Efficiency Enhancement:**
  Accelerate chiplet-based system design workflows
- **SRAM Scaling Constraint Resolution:**
  Guaranteeing SRAM chiplets placement in peripheral silicon regions

How

- ILP-driven optimization of SRAM instance selection with multi-objective constraints (monetary cost, power budgets, latency targets).

$$k_{\text{Mem}} \cdot \sum s_k \text{Cost}(s_k) + k_P \cdot P + k_{FT} \cdot FT \qquad (16)$$
$$(k_{\text{Mem}} + k_P + k_{FT} = 1).$$

- Enforcing SRAM chiplets placement in peripheral regions during SA to eliminate computationally infeasible layouts.

IV. Experiment

Lightweight NoC Improvement

Table 2: Comparison Across Different Mesh and Router Configuration.

Configuration	2×2@1port		2×4@1port		2×4@2ports		4×4@2ports		8×4@3ports	
	τ_{avg}	B_{max}	τ_{avg}	B_{max}	τ_{avg}	B_{max}	τ_{avg}	B_{max}	τ_{avg}	B_{max}
Baseline	**15.07**	4.71	**20.45**	4.37	**18.01**	5.47	**26.51**	5.18	**36.15**	6.48
+ Thin Crossbar	**13.43**	5.03	**18.97**	4.69	**16.45**	5.92	**25.33**	5.53	**35.12**	6.91
+ Dual Localport	**10.59**	6.48	**14.59**	6.04	**13.87**	7.97	**21.08**	7.55	**31.18**	8.60
+ Address Predictor	**9.48**	7.34	**13.32**	6.98	**12.86**	9.46	**19.76**	8.76	**28.77**	9.53
+ Grouped Address	**9.67**	7.33	**13.18**	6.99	**9.78**	10.65	**15.70**	10.68	**17.12**	11.64

- **The progressive optimizations achieve maximum 62.6% latency reduction and 79.6% bandwidth improvement in largest configurations.**
- **Synergistic optimizations maintain performance scalability across diverse topologies while preserving design modularity.**

Bandwidth Heatmap

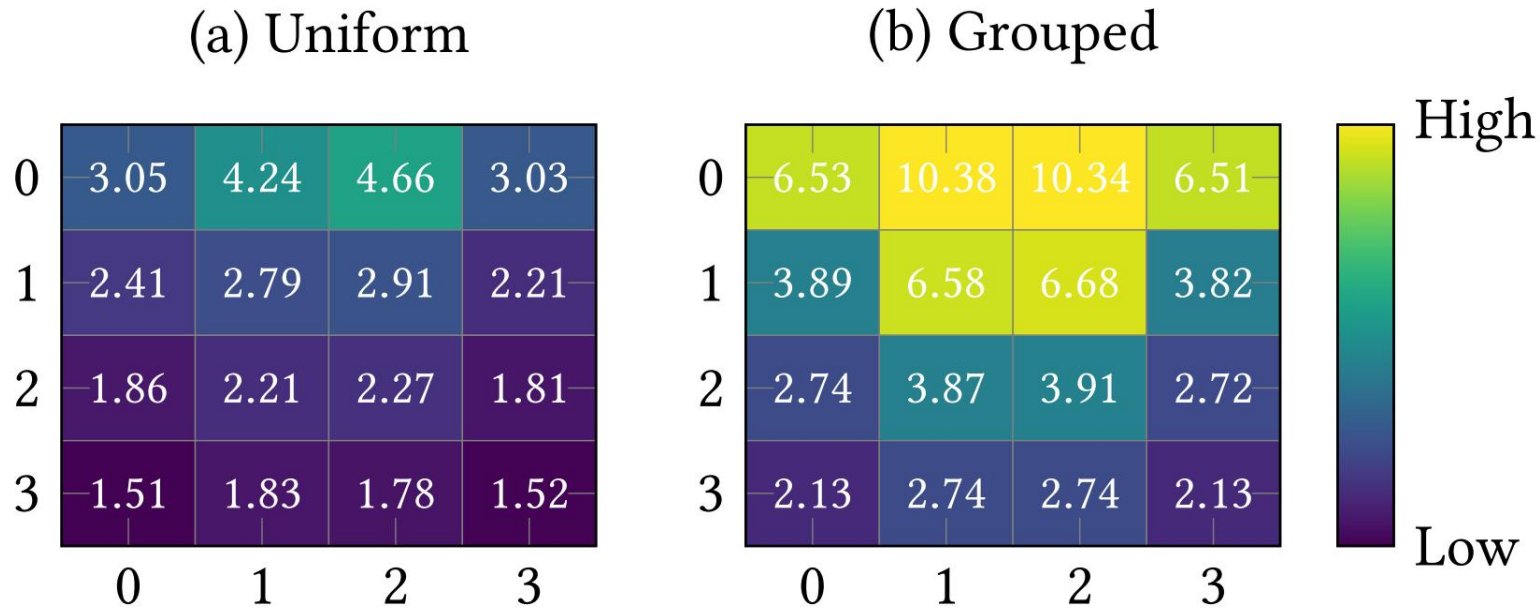


Figure 5: 4×4@2ports with vertical bisection (8 nodes/group) bandwidth (Gbps) heatmap. (1, 0) (2, 0) are IO routers.

- Multi-level heat zones demonstrate 41% bandwidth variance across groups, enabling dynamic resource reallocation for burst workloads.
- Heatmap-guided bank placement cuts average access latency for hotspot-adjacent nodes

Pareto Frontier

System-Level Evaluation

- Determining optimal on/off-die Capacity ratio κ via constrained multi-objective optimization.
- Driving architectural decision-making through power/area/resilience tradeoff characterization.

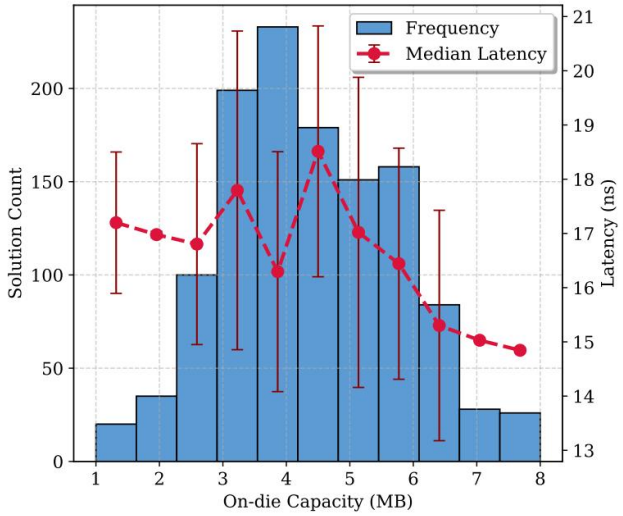


Figure 7: C_{on} Distribution.

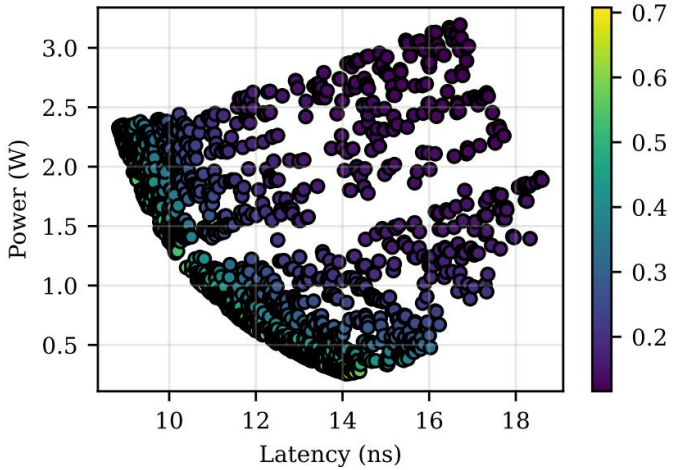


Figure 8: Pareto Frontier.

Optimal on-die ratio κ trades off NRE/RE costs:

- κ=0.40 for cost-constrained edge devices,
- κ=0.64 for latency-critical systems.
- κ=0.53 establishes a verified sweet spot

Cost Analysis & Speedup

Table 3: Total On-die Ratio, System Cost (normalized) Reductions compared to Monolithic with Manufacturing Quantities of 500K and 10M and Solve Time(sec) Reduction.

Design	On-die Ratio	Cost Save		Solve Time	
		500K	10M	GIA [10] (s)	Ours (s)
CPUs-1	40%	24.5%	7.8%	481.65	243.26
CPUs-2	35%	22.1%	6.9%	337.27	240.90
CPUs-3	25%	26.3%	8.5%	67.62	32.21
CPUs-4	40%	18.7%	5.2%	52.76	25.73
GPUs-1	55%	20.9%	6.3%	755.87	354.86
Avg.	-	22.5%	5.94%	(Speedup) 1.93×	

- Neighbor pruning and ILP pre-screening achieve 1.93× average speedup in GIA framework, reducing invalid evaluations by 68.7%.

V. Conclusion

Conclusion

1. Core Innovation & Validation

- AuxiliarySRAM: A lightweight on-chip memory architecture featuring a latency-optimized NoC-based Chiplet implementation
- 49.29% lower average latency & 79.35% higher peak bandwidth vs. baseline

2. System-Level Cost Efficiency

- Demonstrates cost-saving advantages: Achieves 26.3% system cost reduction under equivalent capacity
- Multi-design compatible SRAM chiplet shared library minimizes main die area

3. Framework Contribution

- Integrated analytical models into enhanced GIA framework for system-level evaluation
- BO-GP resolves Pareto-optimal on/off-die ratios with 1.93× pruning speedup

Thanks for your time!