Gleanmer: A 6 mW SoC for Real-Time 3D Gaussian Occupancy Mapping

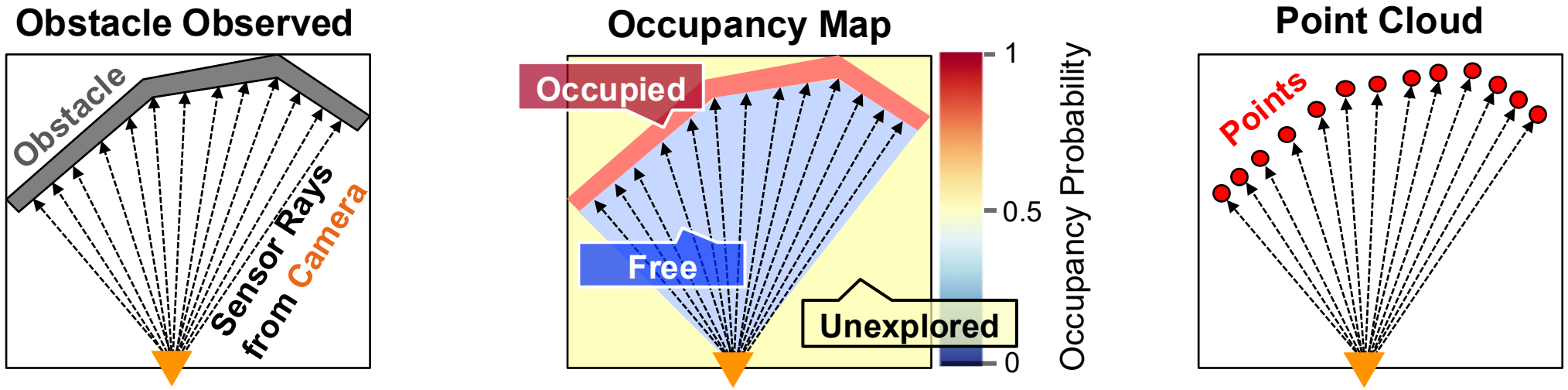
Zih-Sing Fu*, Peter Zhi Xuan Li*, Sertac Karaman, Vivienne Sze

Massachusetts Institute of Technology

* Equal contribution to this work

1 What is Occupancy Mapping?

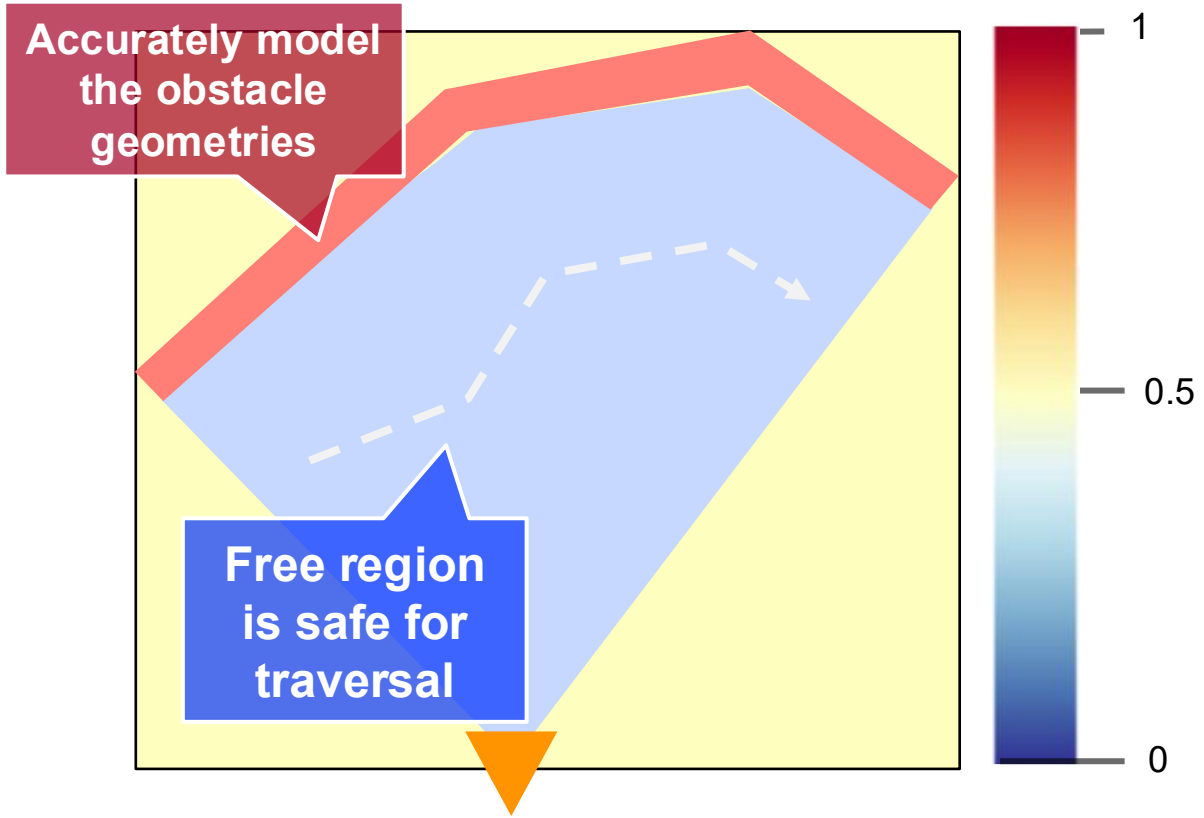
- **Mapping** refers to the process of building a **spatial representation** of an environment from sensor rays.
- **Point clouds** (from depth cameras, LiDAR, or SLAM pipelines) only describe occupied surfaces observed by sensor rays.



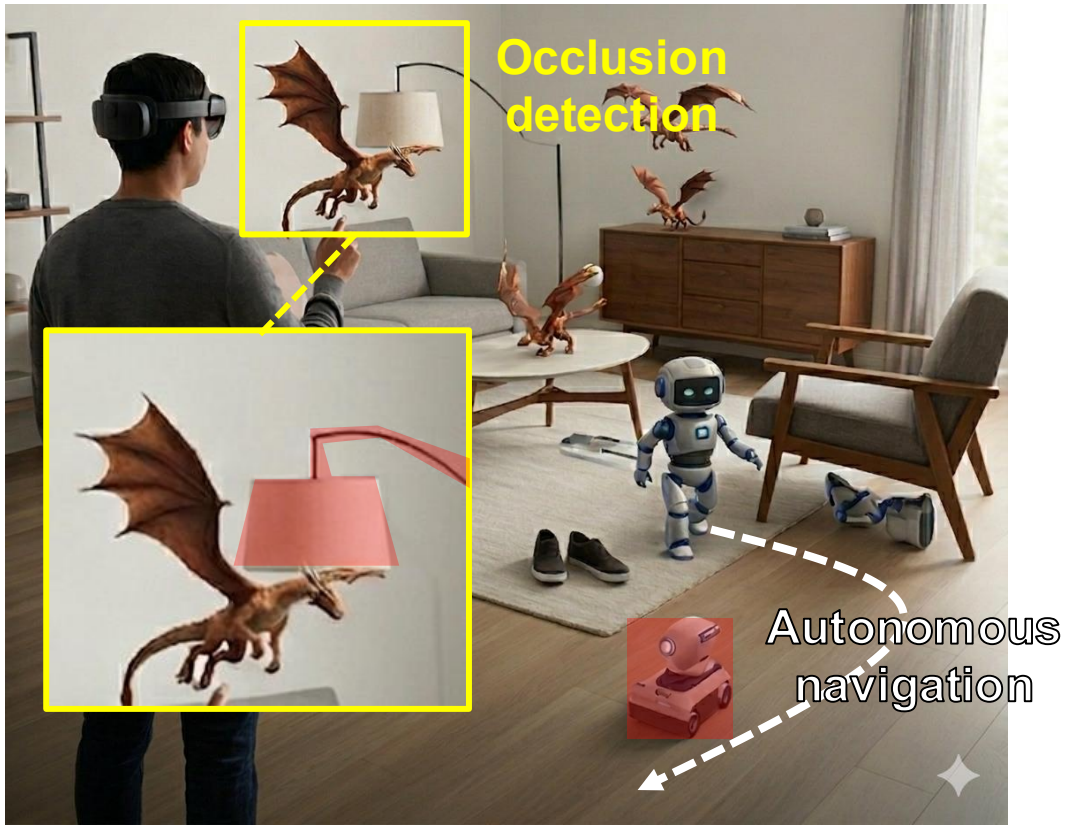
Occupancy maps classify the entire 3D environment into **occupied, free, and unexplored regions**, which are essential for safe navigation.

Occupancy Map Enables Edge Applications

Occupancy Map



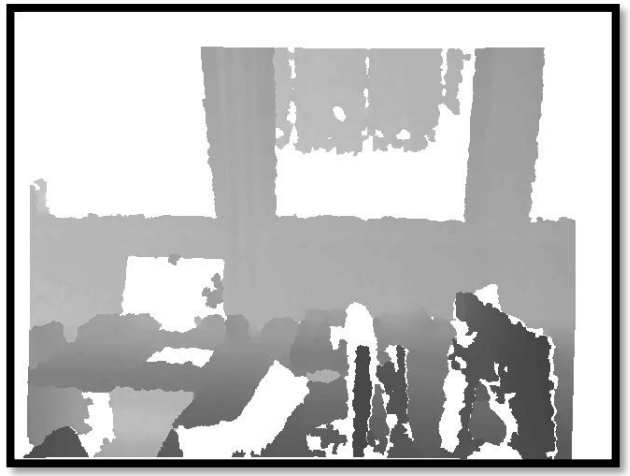
Example Applications



3D Map Representation: Voxels

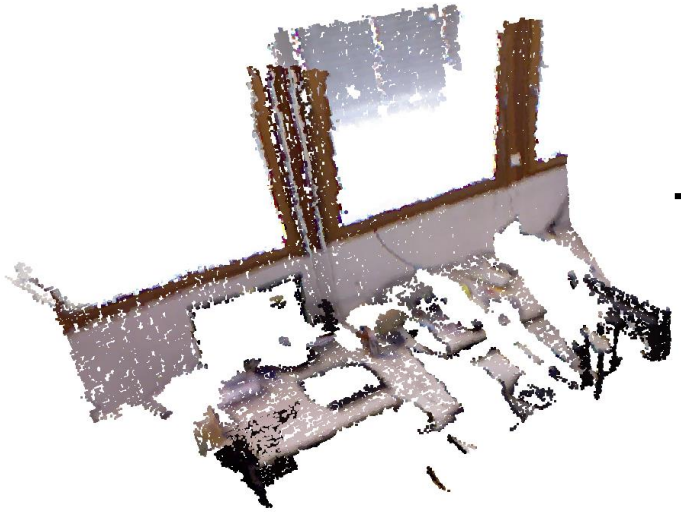
- **Spatial correlation: Adjacent depth pixels usually correspond to 3D points that lie close together on the same surface.**

2D Depth Image



0.6 MB

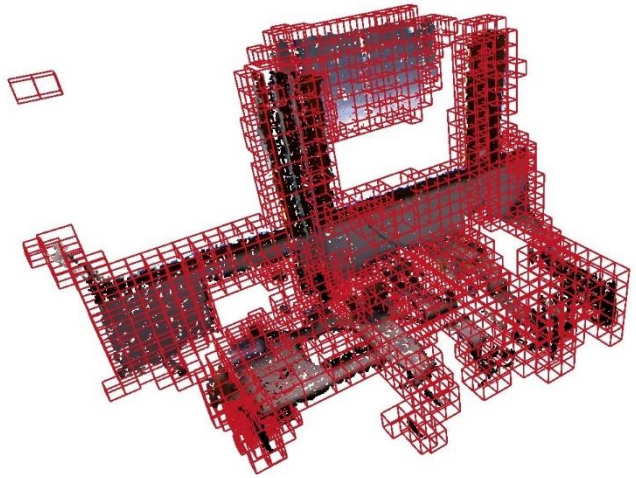
Point Cloud



2MB

Voxels (red)

[Hornung et al., Auton Robot 2013]

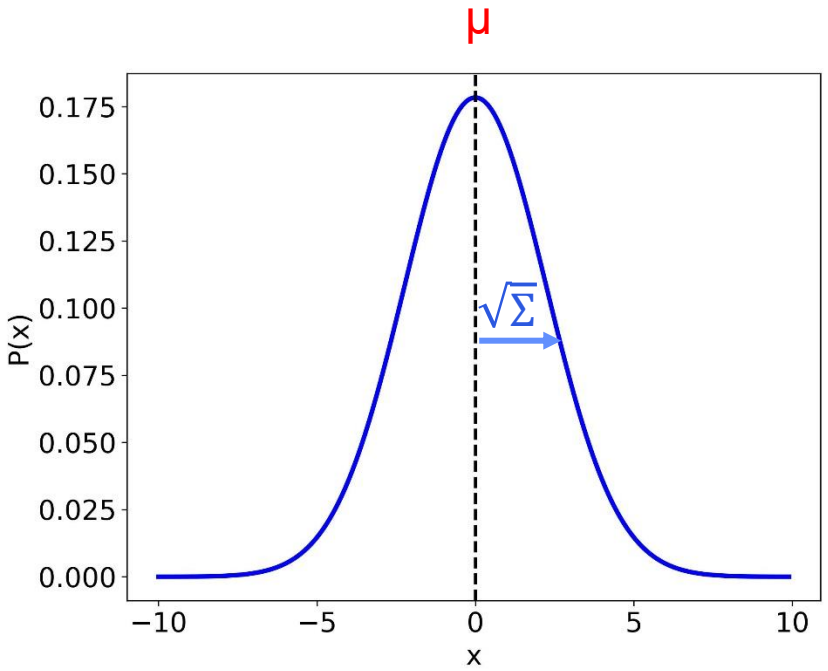


147KB

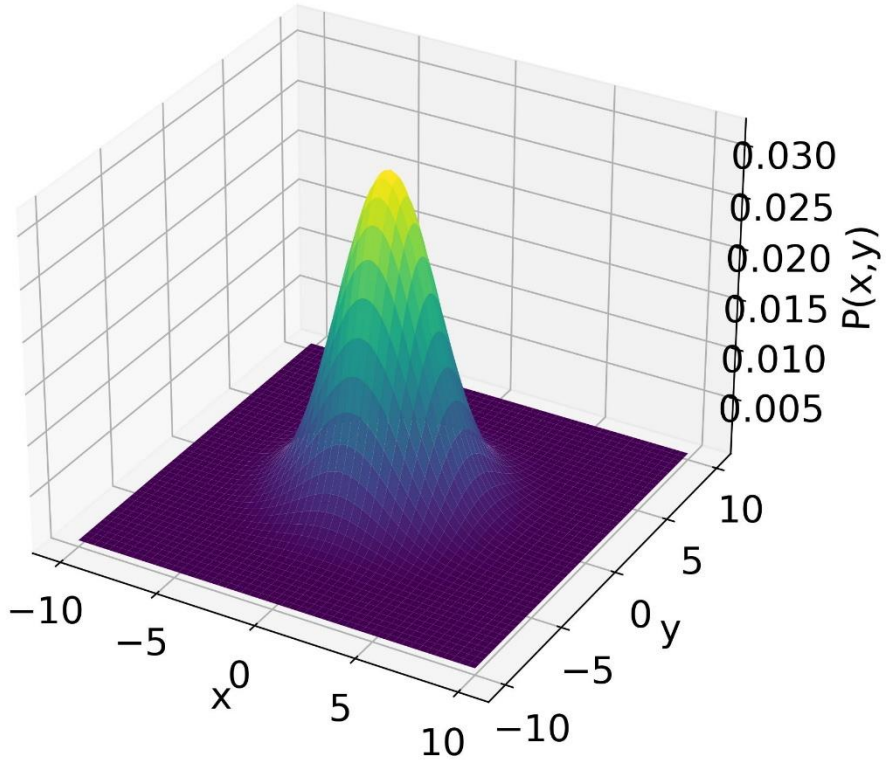
Each voxel exploits **spatial correlation uniformly in each (axis) direction**. Resulting map across **multiple images** can increase quickly to **MBs** and **GBs**.

What is a Gaussian Distribution?

- Probabilistic distribution parametrized by **mean (μ)** and **covariance matrix (Σ)**



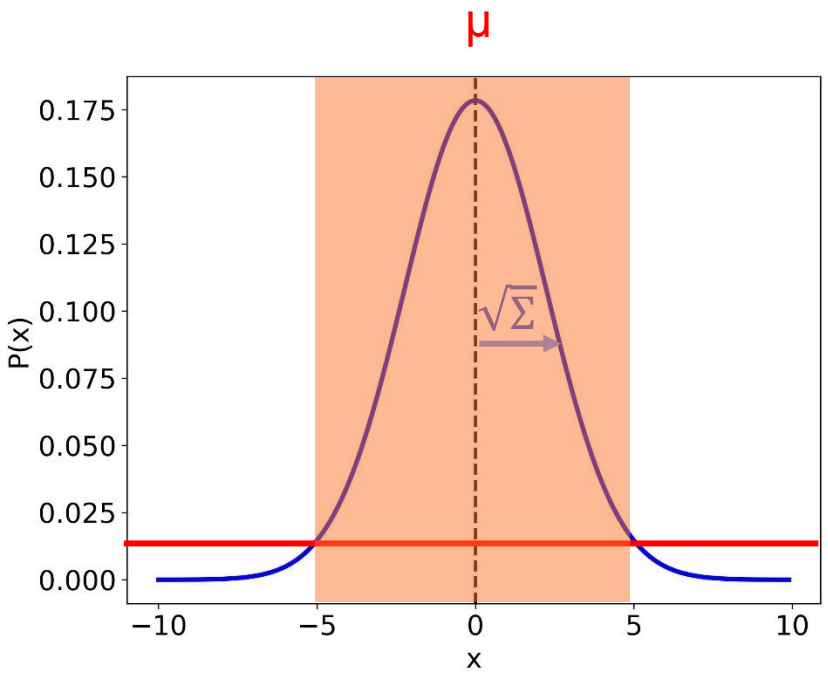
1D Gaussian



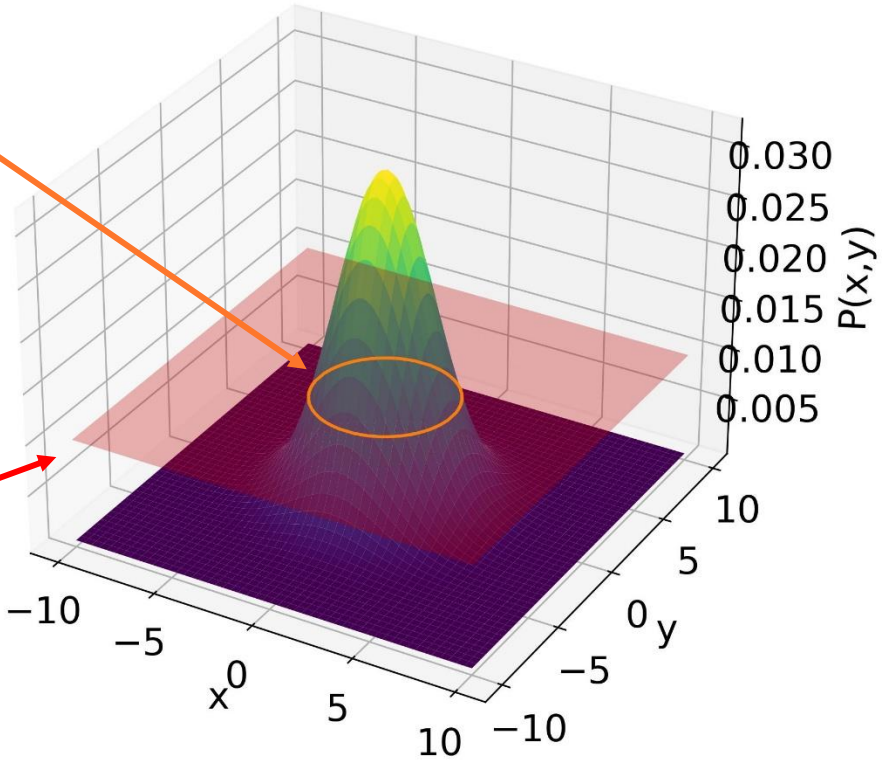
2D Gaussian

What is a Gaussian Distribution?

- Probabilistic distribution parametrized by **mean (μ)** and **covariance matrix (Σ)**
- Iso-surface is visualized as an **ellipsoid**



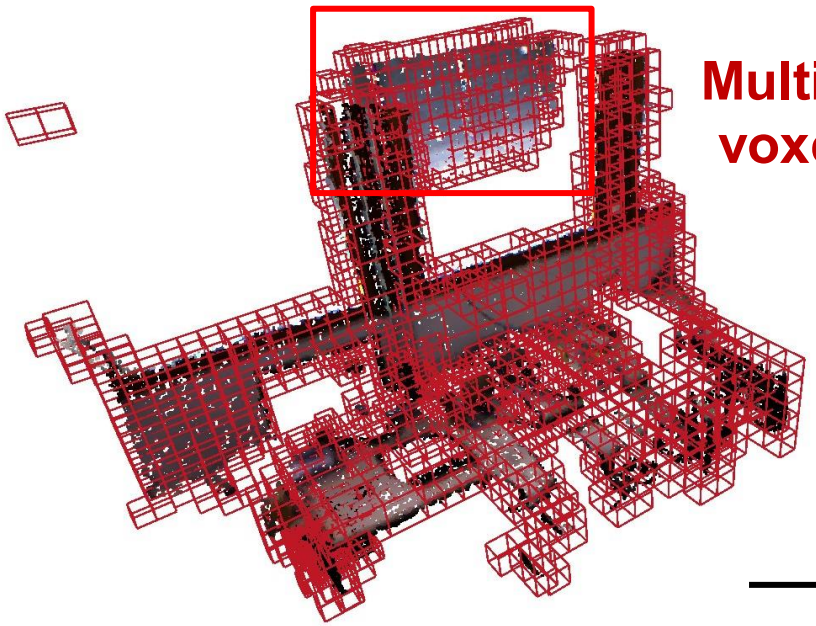
1D Gaussian



2D Gaussian

3D Map Representation: Gaussians

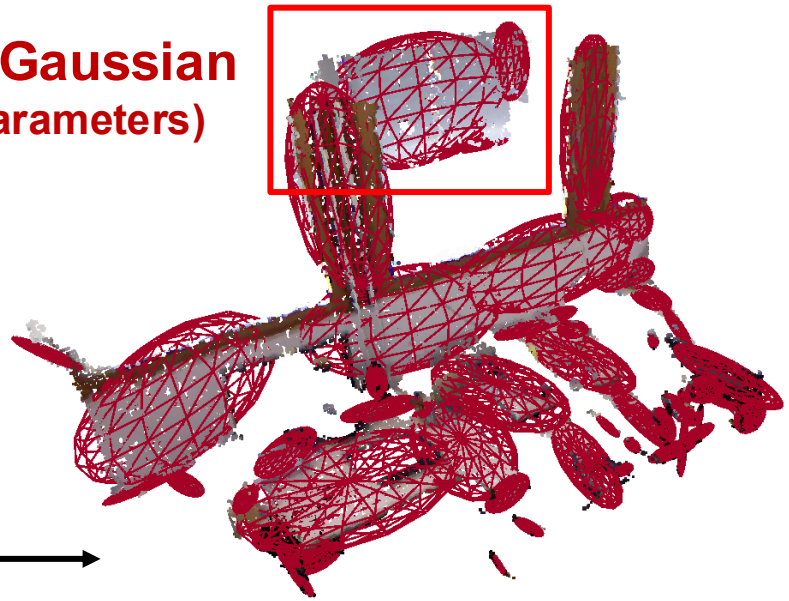
Voxels (red)



Multiple voxels

One Gaussian
(9 parameters)

Gaussians (red)



16x reduction

147KB

9KB

Gaussians capture **spatial correlation flexibly**, adapting their shape and orientation to the underlying geometry.

3D Map Representation: Gaussians

- Gaussian map from multiple images are highly compact.

Voxels (red)

10.5MB

13x reduction

Gaussians (red)

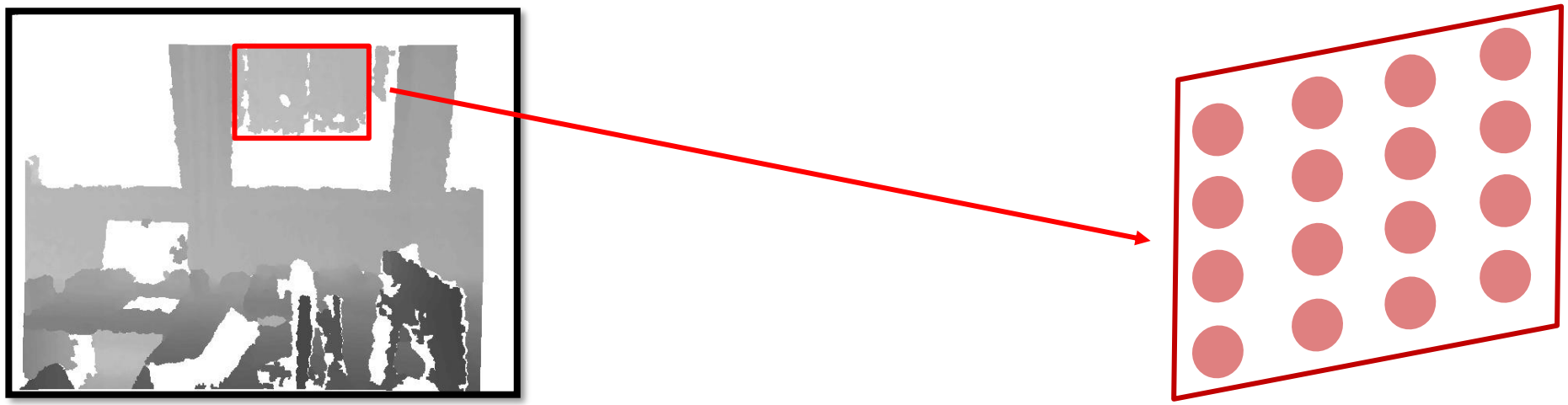
840KB

Background: GMMap [Li et al., T-RO 2024]

- We previously developed **GMMap** for memory-efficient Gaussian map **storage** and **construction**.

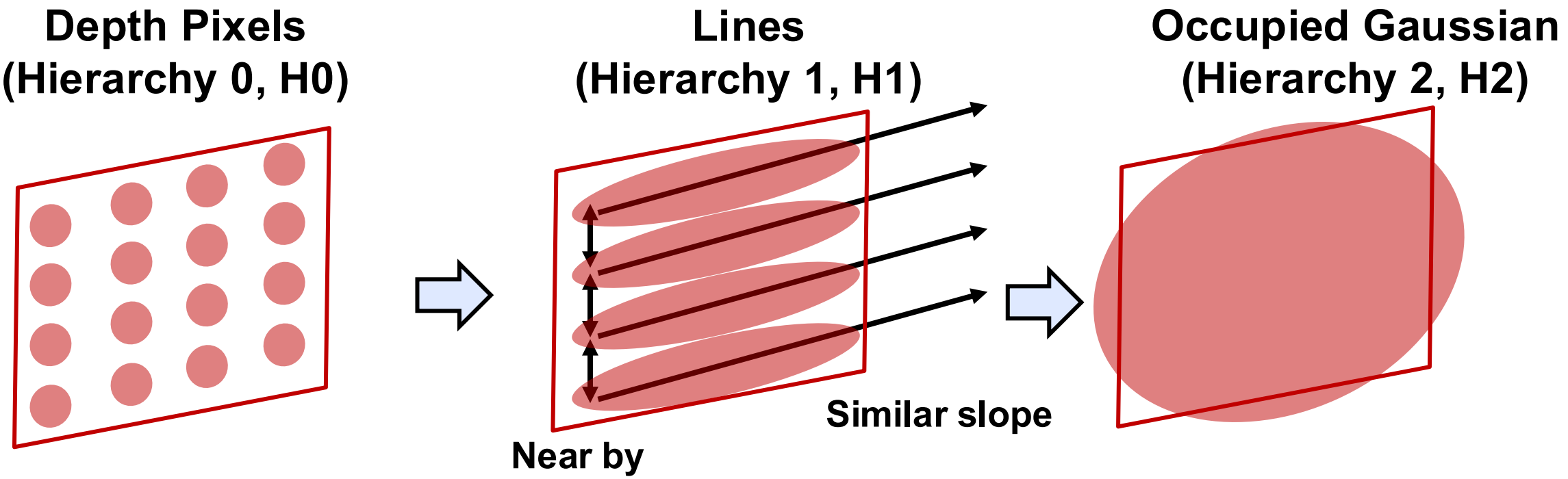
2D Depth Image

Depth Pixels (3D)



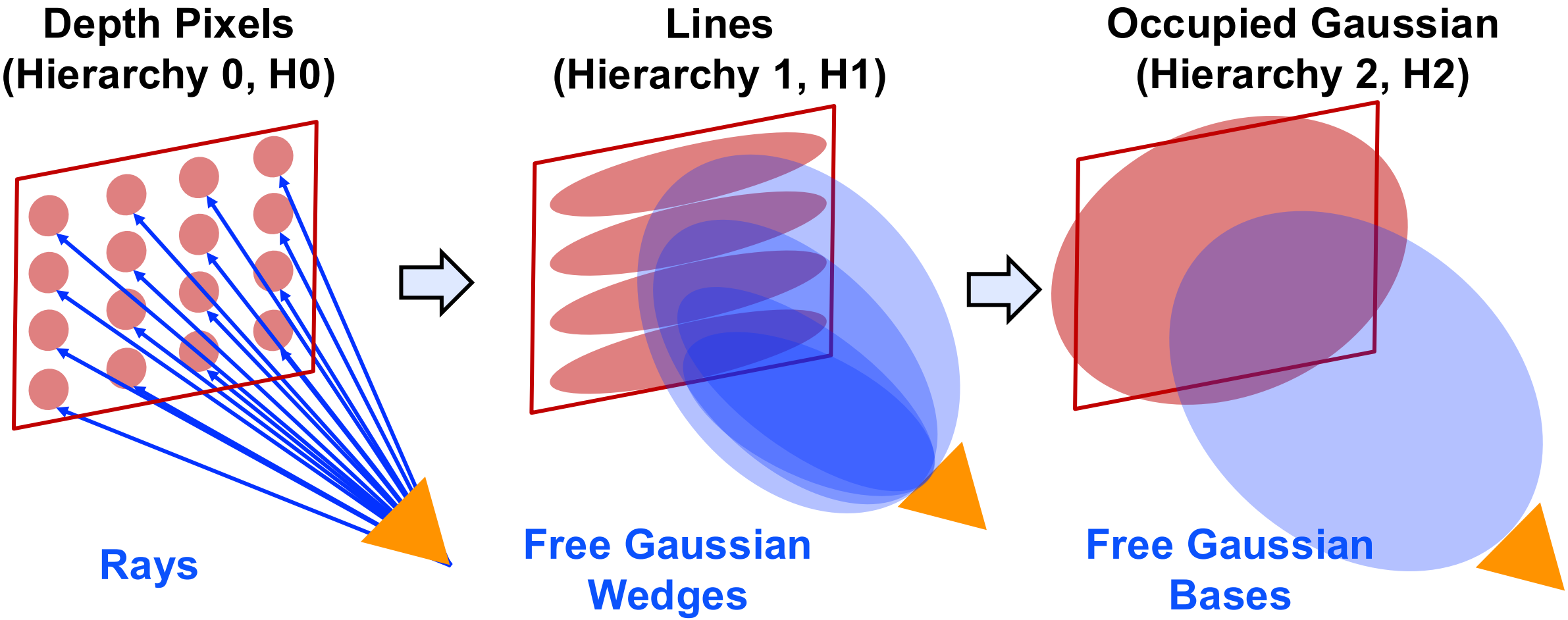
GMMap starts from **depth pixels**: the 3D points where sensor rays terminate on obstacle surfaces.

GMMap: Occupied Gaussian Construction



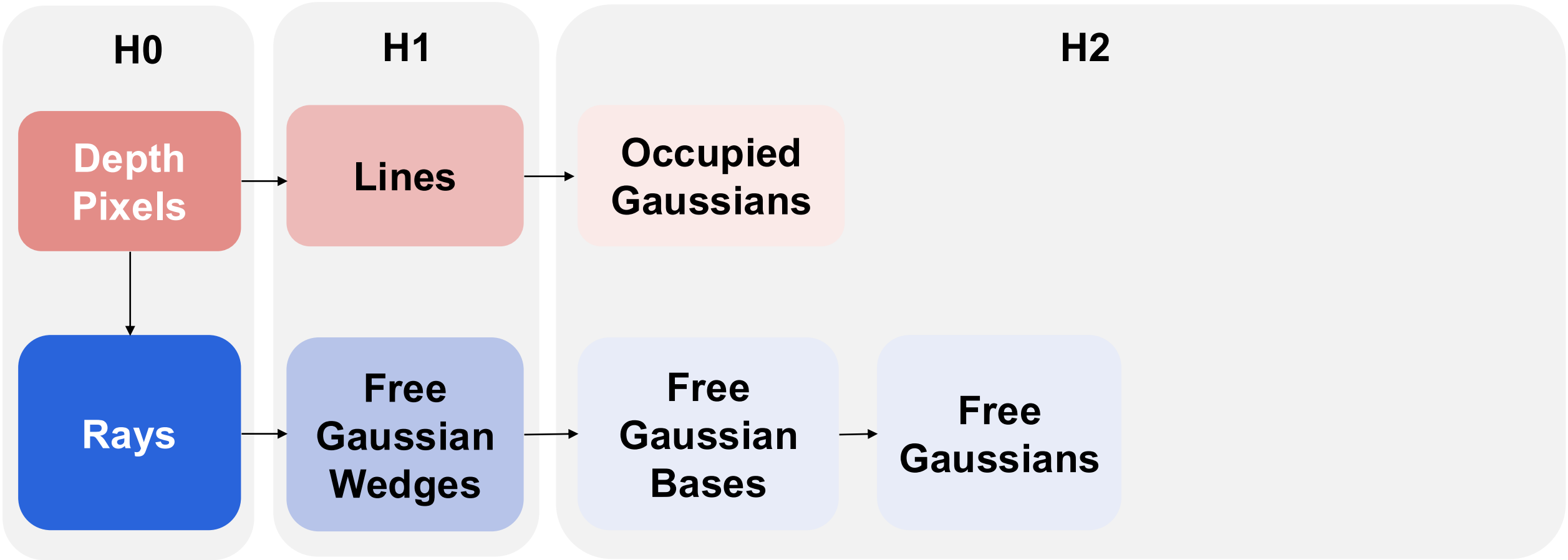
Spatial correlation defines the hierarchy: **pixels (H0)** on the same planar surface collapse into **lines (H1)**, and lines that lie close to each other fuse into **Gaussians (H2)**.

GMMap: Free Gaussian Construction



Spatial correlation defines the **same hierarchy for free Gaussians** as well.

GMMap: Data Volume Reduction from H0 to H2



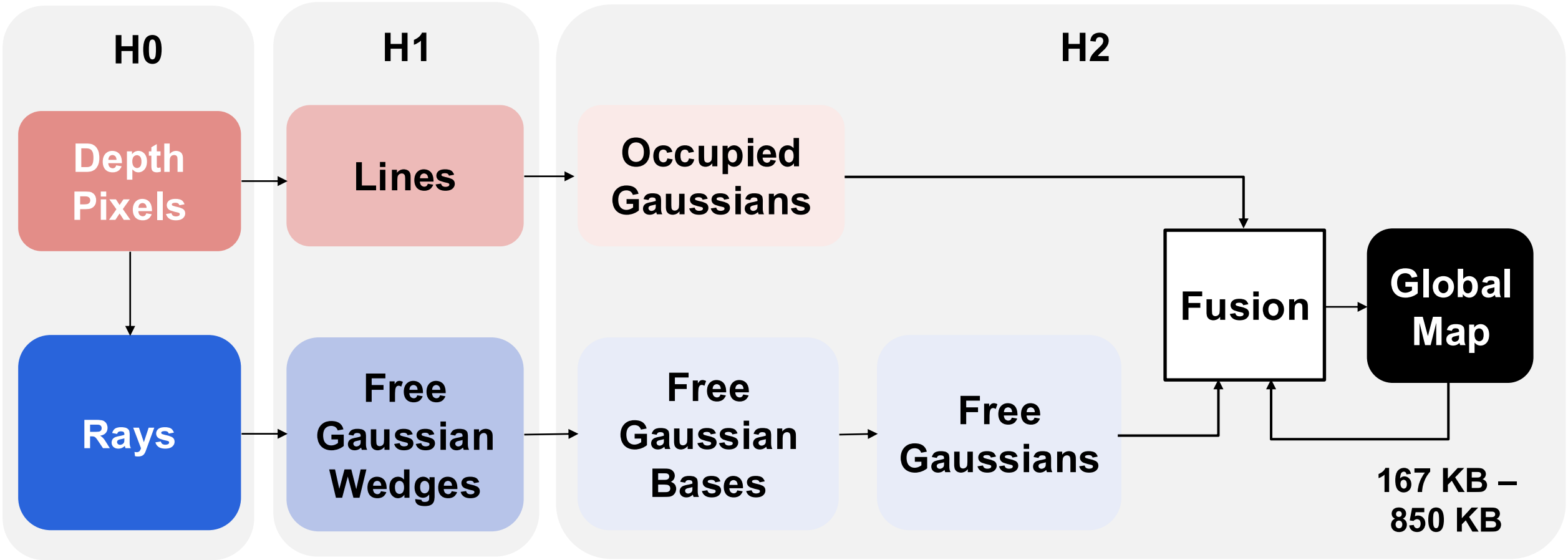
3 MB / image
180 MB – 300 MB / s

300 KB / image

18 KB / image

[Li et al., T-RO 2024]

GMMap: Gaussian Fusion across Images



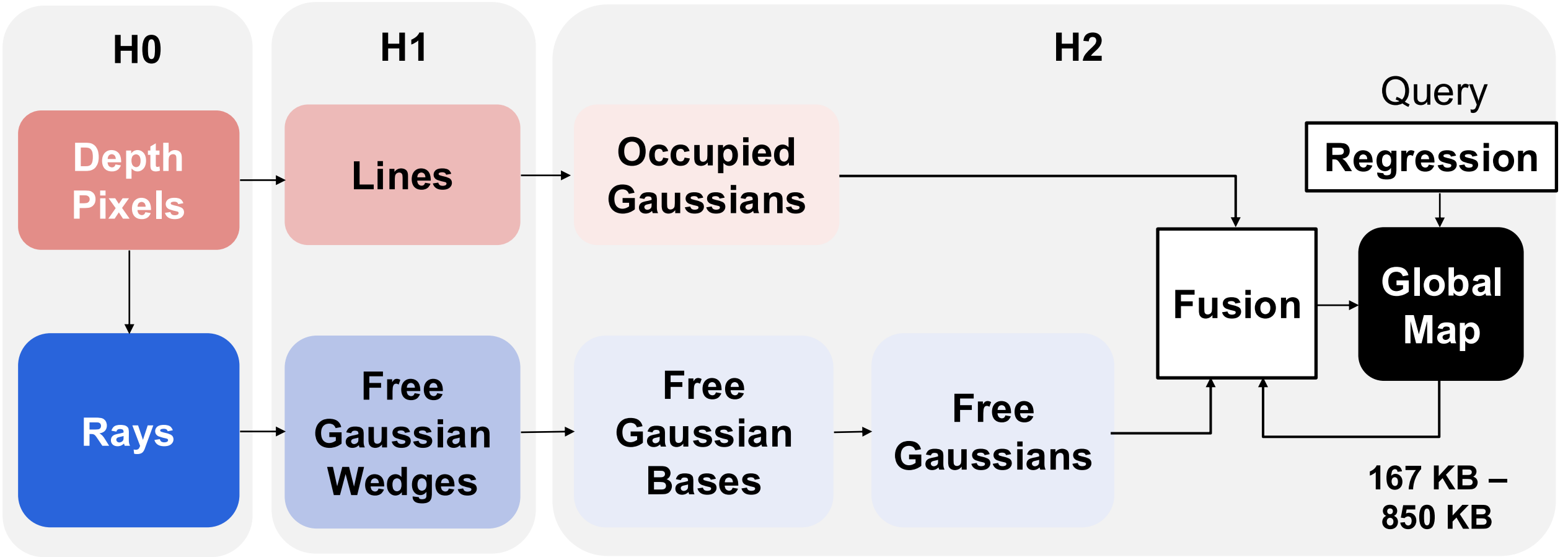
3 MB / image
180 MB – 300 MB / s

300 KB / image

18 KB / image

[Li et al., T-RO 2024]

GMMap: Gaussian Regression for Query



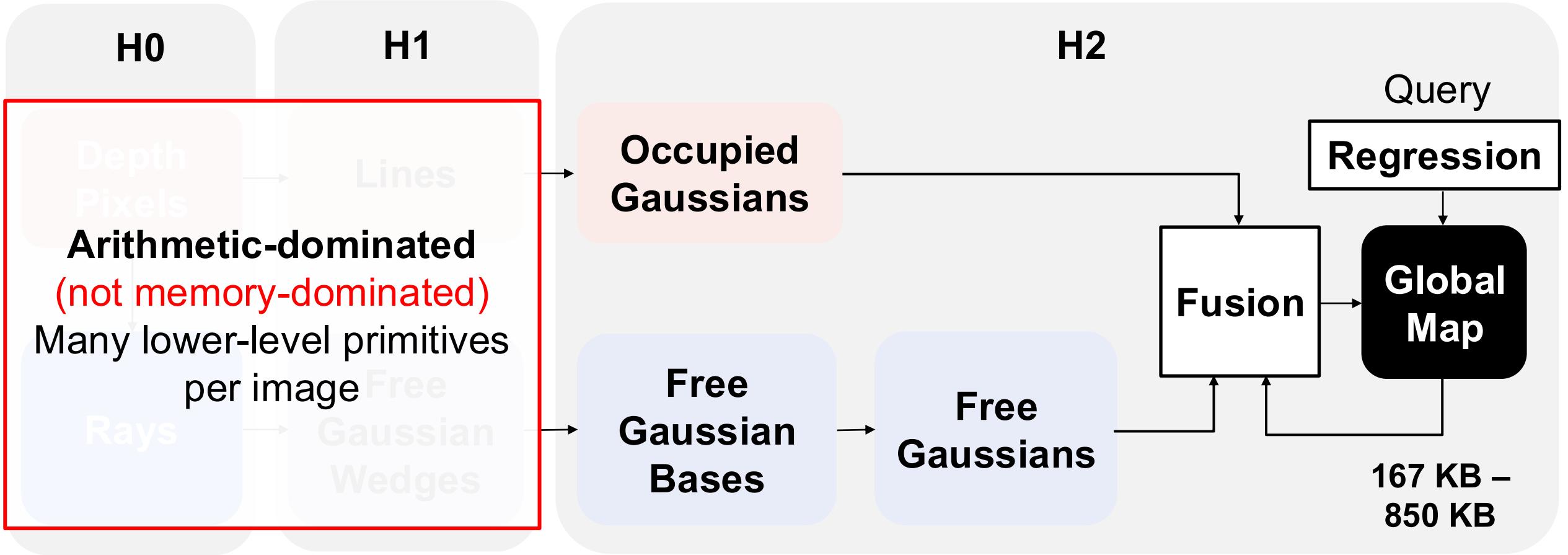
3 MB / image
180 MB – 300 MB / s

300 KB / image

18 KB / image

[Li et al., T-RO 2024]

GMMap: Challenges



3 MB / image
180 MB – 300 MB / s

300 KB / image

18 KB / image

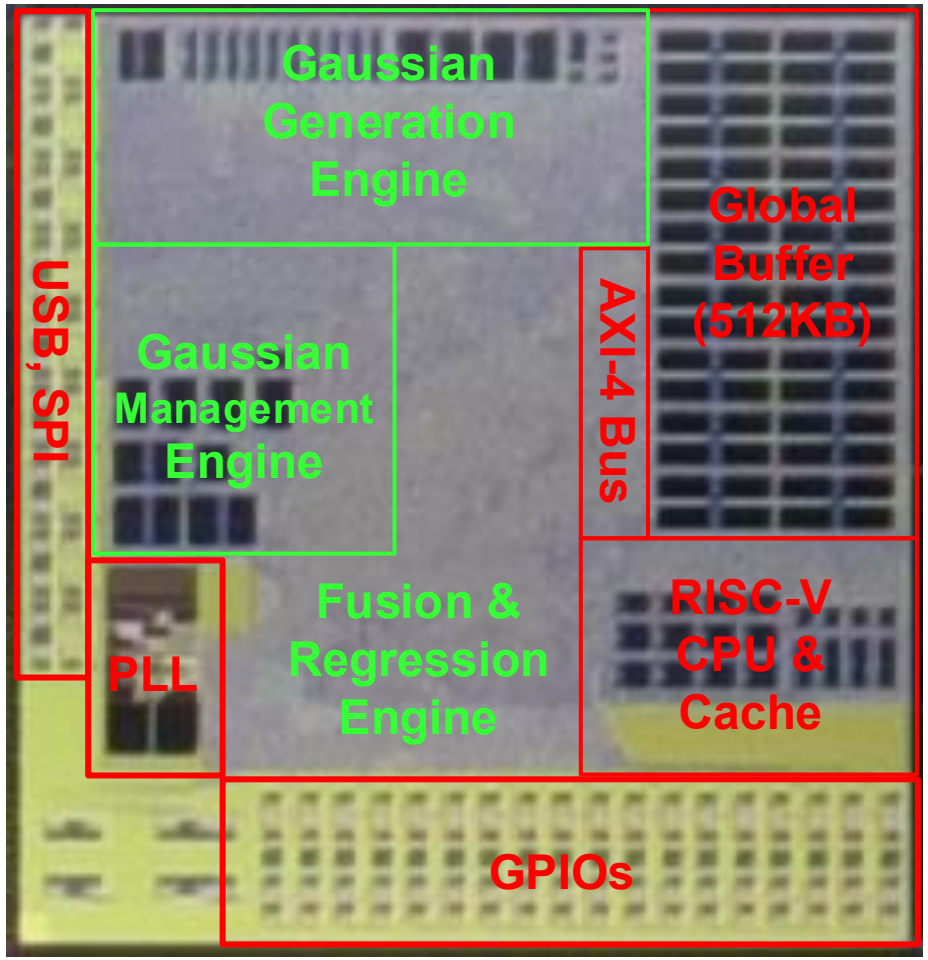
[Li et al., T-RO 2024]

GMMap: Challenges

H0	H1	H2
Arithmetic-dominated (not memory-dominated) Many lower-level primitives per image		**Memory-dominated** Repeated memory access for retrieving overlapping Gaussians in the fused global map. **Area-dominated** Gaussian arithmetic (fusion, regression) demands precise floating-point arithmetic.
3 MB / image 180 MB – 300 MB / s	**300 KB / image**	**18 KB / image**

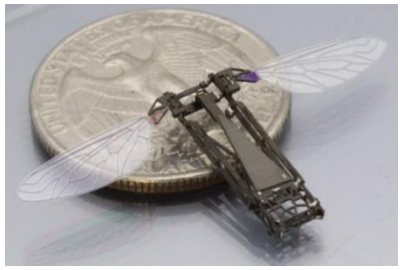
[Li et al., T-RO 2024]

Gleanmer: Energy-Efficient 3D Occupancy Mapping



- **First fabricated SoC** for energy-efficient real-time 3D occupancy mapping (GMMap)
- **Constructs** a map from depth images by at least **88 frames per second (fps)**
- **Queries** the map by up to **1.3M coordinates per second (cps)**
- Average power consumption of **6mW**

Gleanmer Has Wide Range of Edge Applications



Robobee

~28 mW

Blimp

~46 mW

Micro-drone

~88 mW

Cellphone

~2.5 W

AR/VR

~10 W

Drone

~70 W

1 mW | 10 mW | 100 mW | 1 W | 10 W | 100 W

★ Gleanmer (this work)
(GMMap)
[Fu et al., VLSI 2026]
6 mW @ 88+ fps

OMU
(OctoMap)
[Jia et al., DATE 2022]
251 mW @ 64 fps

ARM Cortex A57
(GMMap)
[Li et al., TRO 2024]
4.7 W @ 44 - 81 fps

Gleanmer, at **6 mW**, opens the application spectrum, bringing real-time occupancy mapping within reach for even insect-scale platforms.

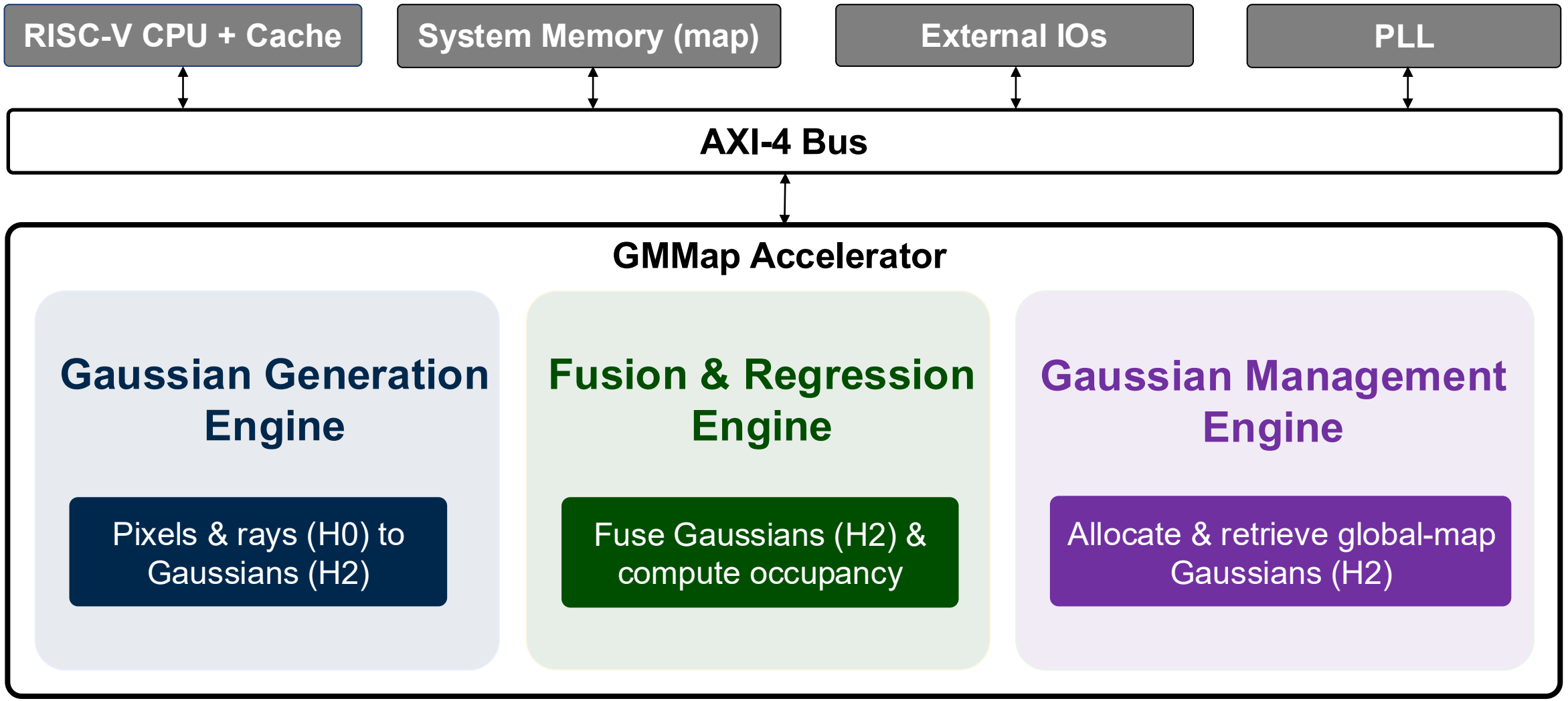
Outline

- Background: 3D Occupancy Mapping
- Chip Architecture and Contributions
- Chip Specifications and Comparisons
- Summary

Outline

- Background: 3D Occupancy Mapping
- **Chip Architecture and Contributions**
- Chip Specifications and Comparisons
- Summary

Gleanmer Chip Architecture

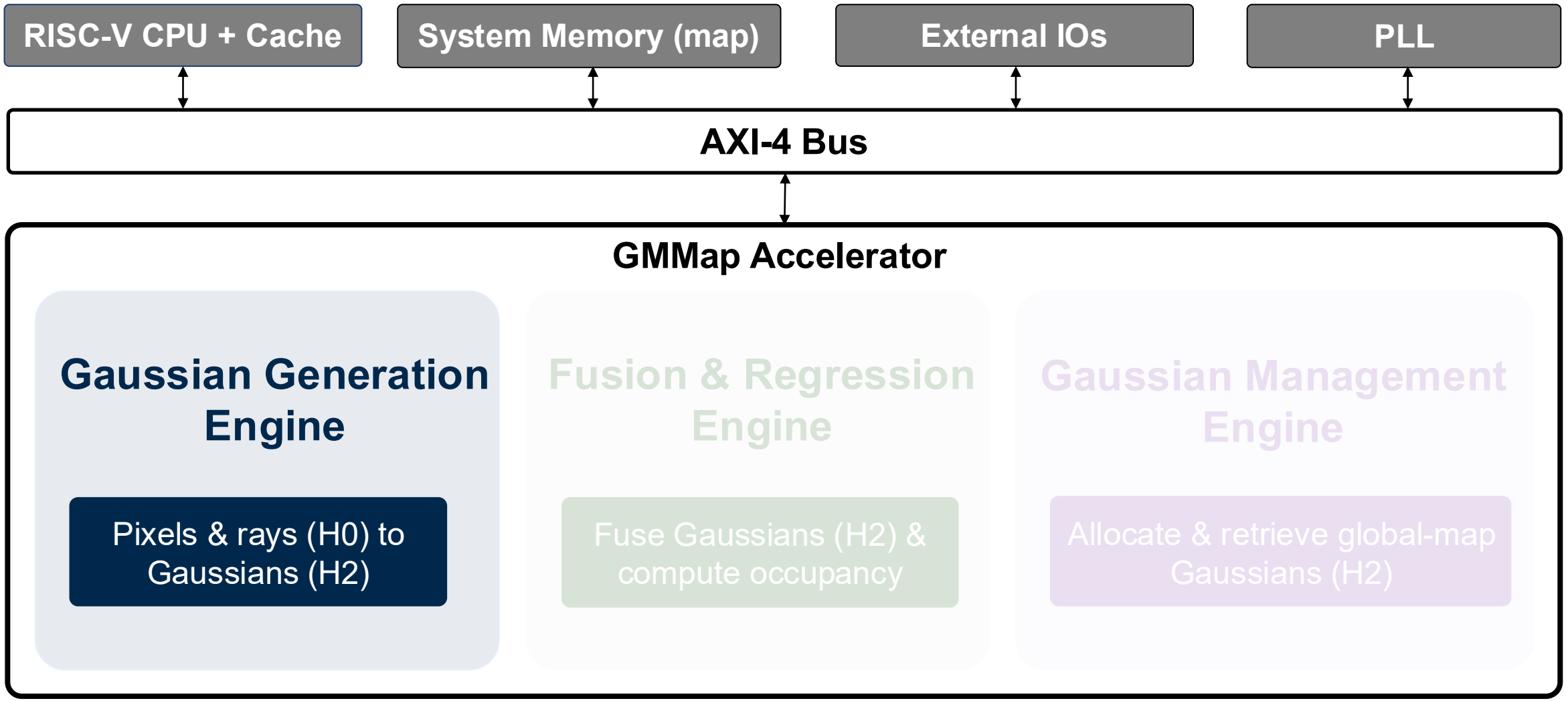


Gleanmer's Two Co-Design Contributions

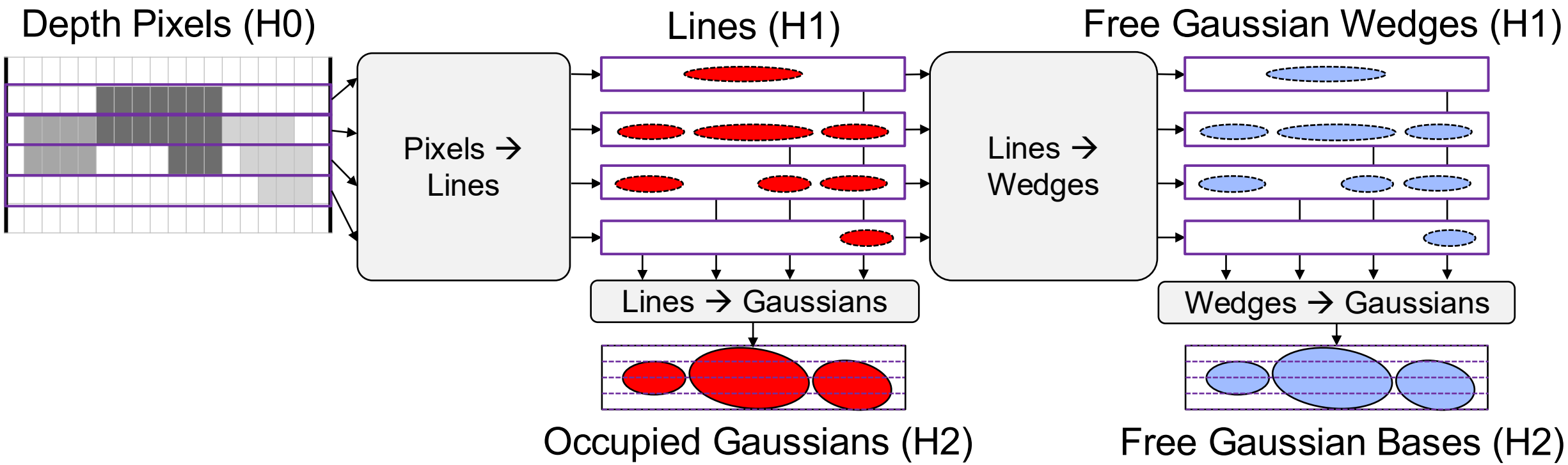
Efficient Hierarchy Conversion (H0 to H2)

Reduce the volume of work by shifting the free Gaussian bases generation from wedges (H1) to occupied Gaussian (H2).

Contribution 1: Efficient Hierarchy Conversion

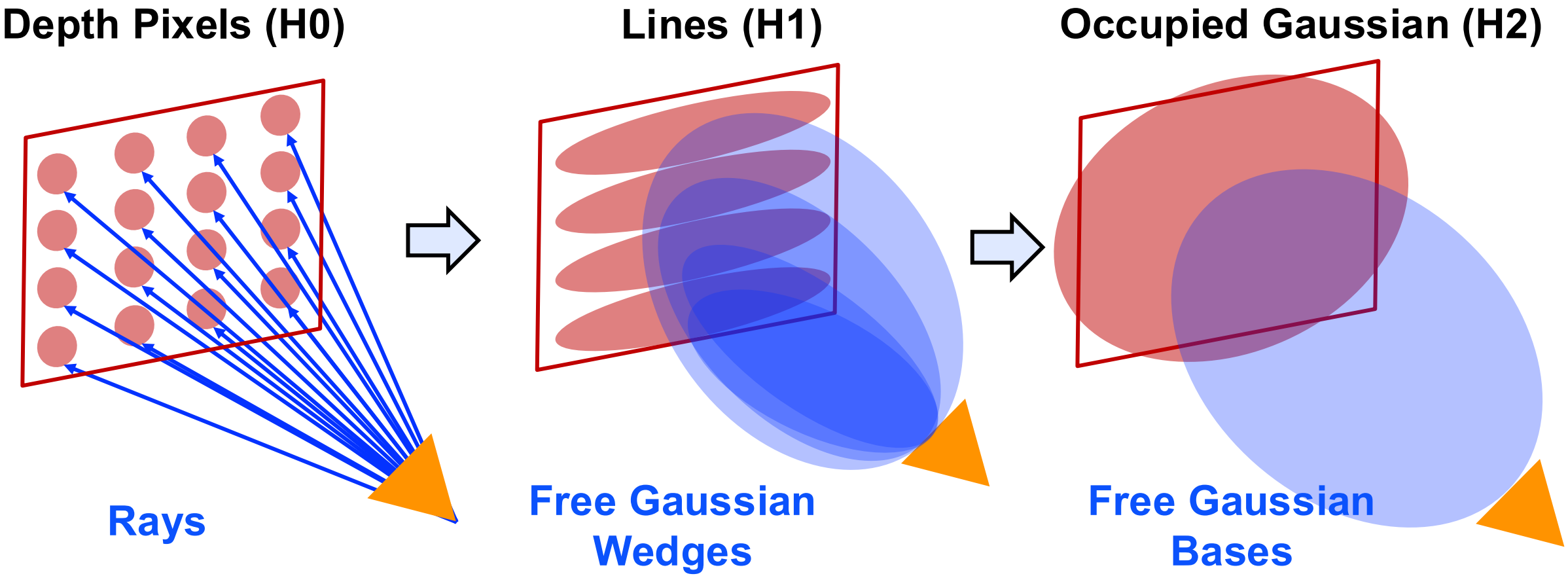


Direct-mapped Hardware Architecture

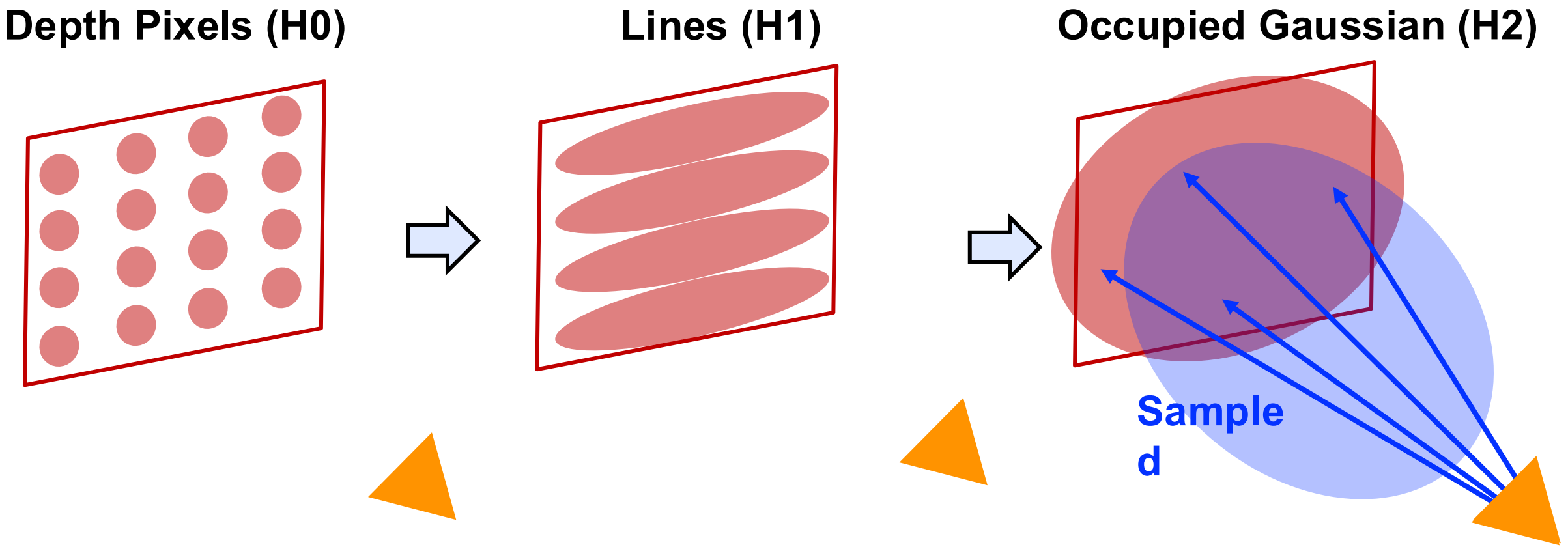


Compute energy scales with **thousands of wedges** and dominates up to **65%** of map construction energy.

Optimized: Occupied Gaussian → Free Gaussian Bases

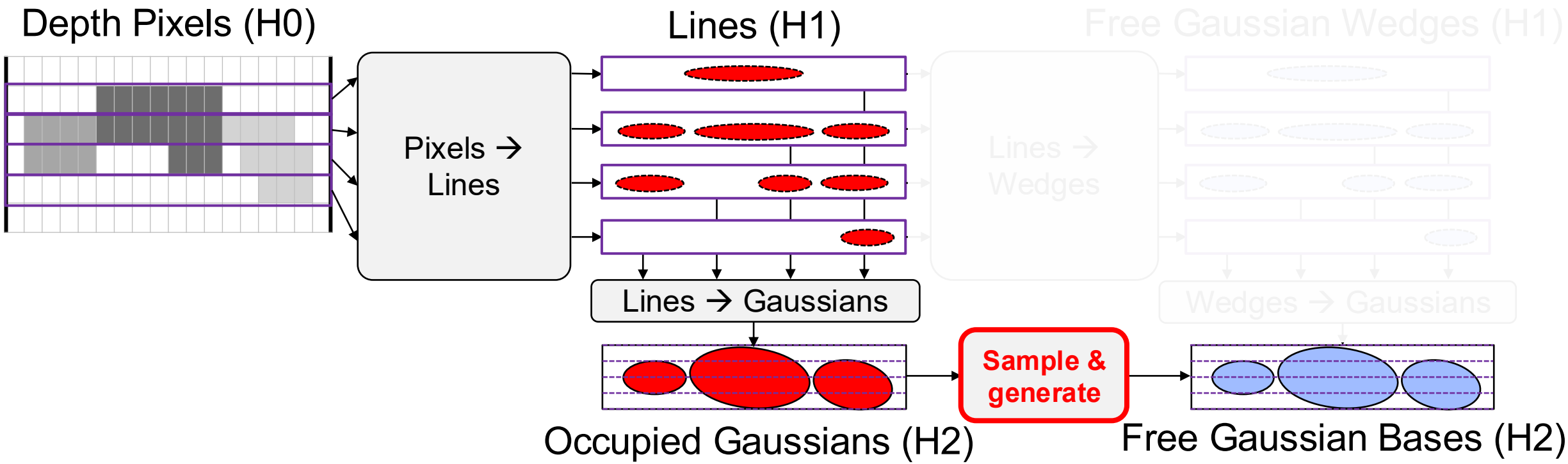


Optimized: Occupied Gaussian → Free Gaussian Bases



The optimized algorithm generates free Gaussian bases from occupied Gaussians (in the **hundreds**) instead.

Optimized Hardware Architecture



Reduce energy by **up to 63%** for map construction while maintaining accuracy.

Gleanmer's Two Co-Design Contributions

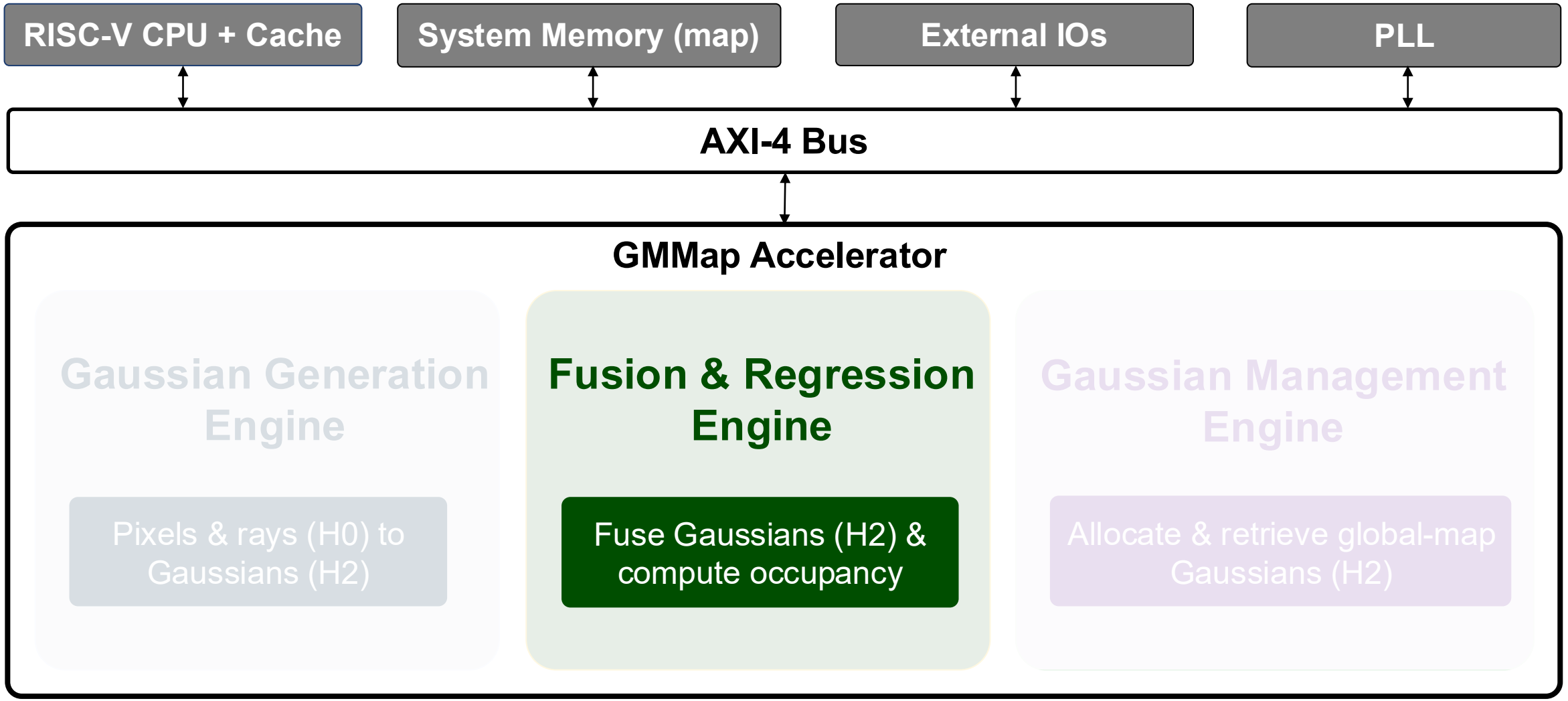
Efficient Hierarchy Conversion (H0 → H2)

Reduce the volume of work by shifting the free Gaussian bases generation from wedges (H1) to occupied Gaussian (H2).

Amplified Optimizations at H2

Apply reduced precision and exploit data reuse on Gaussians (H2) to reduce area and memory-access intensity.

Contribution 2: Amplified Optimizations for Gaussians



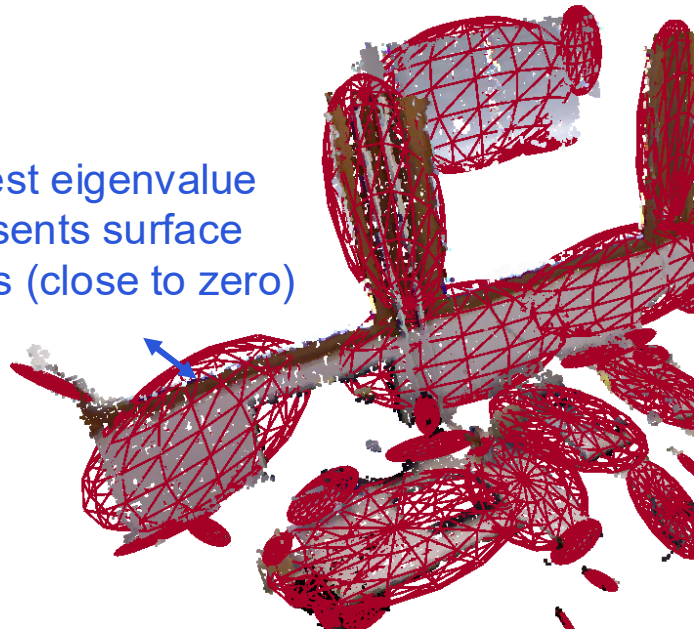
Reducing Precision for Gaussian Arithmetic

- **Reduced precision (19-bit and 26-bit)** are utilized.
- Single (32-bit) precision is required for **updating the covariance matrix**.
  - Failed Gaussian fusion across images if using reduced precision → increase final map size by ~10×.

Must be positive definite (all eigenvalues > 0)

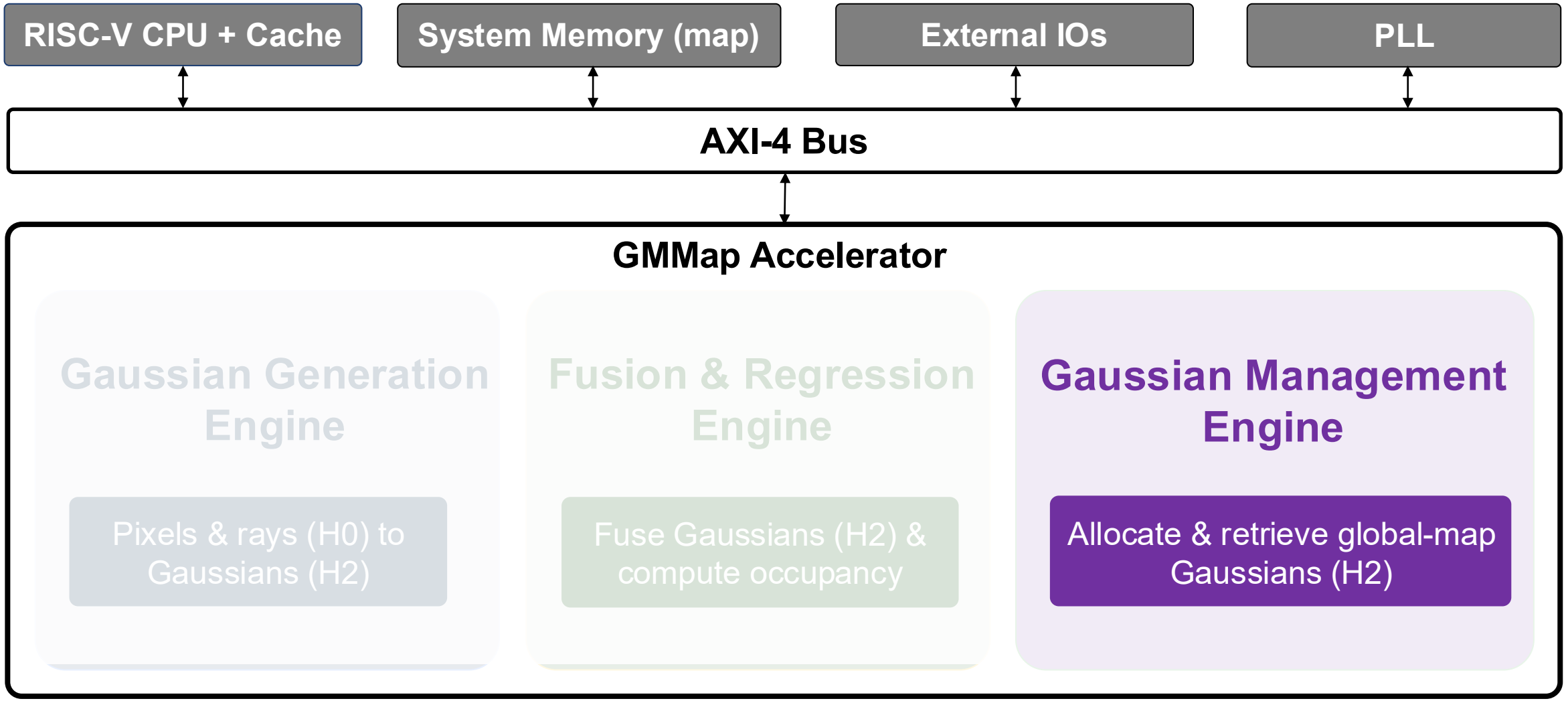
$$\mathcal{N}(p \mid \mu, \Sigma) = \frac{\exp\left(-\frac{1}{2}(p-\mu)^{\top}\Sigma^{-1}(p-\mu)\right)}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}}$$

Smallest eigenvalue represents surface thickness (close to zero)



Reducing the precision reduces the area of fusion and regression engine by **49%.**

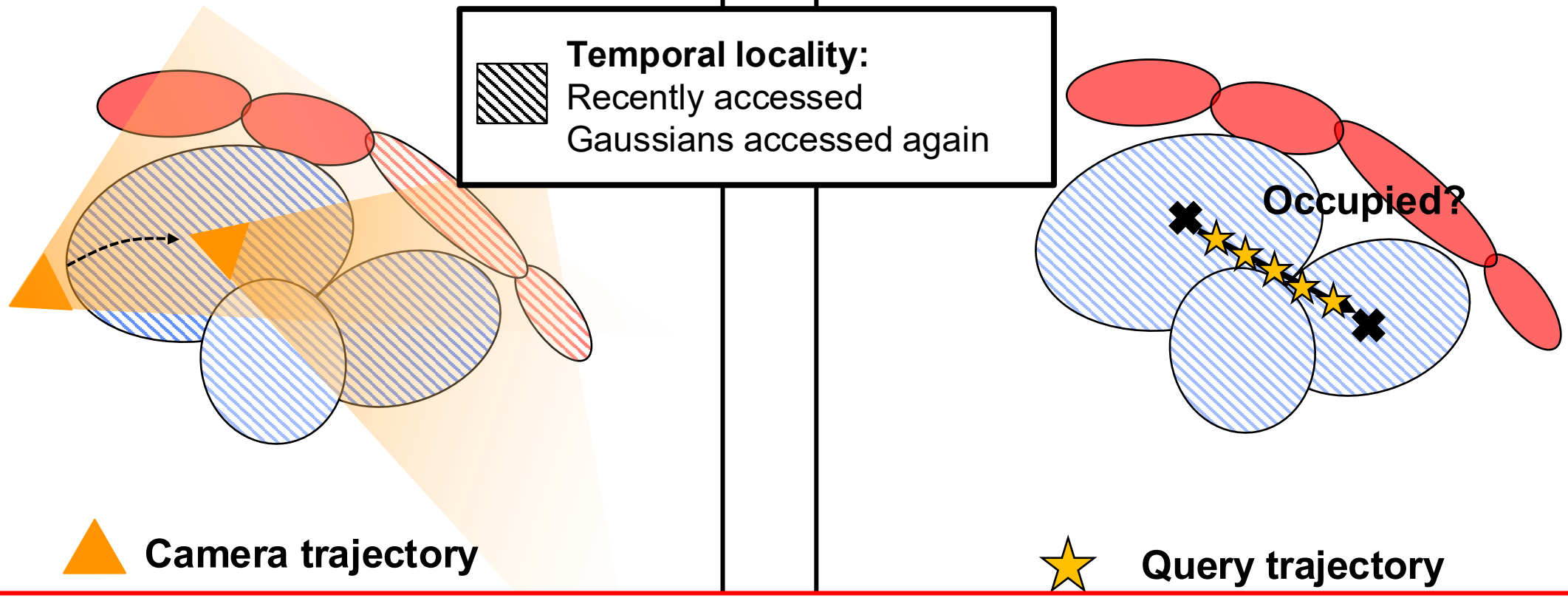
Contribution 2: Amplified Optimizations for Gaussians



Caching Spatially-Close Gaussians

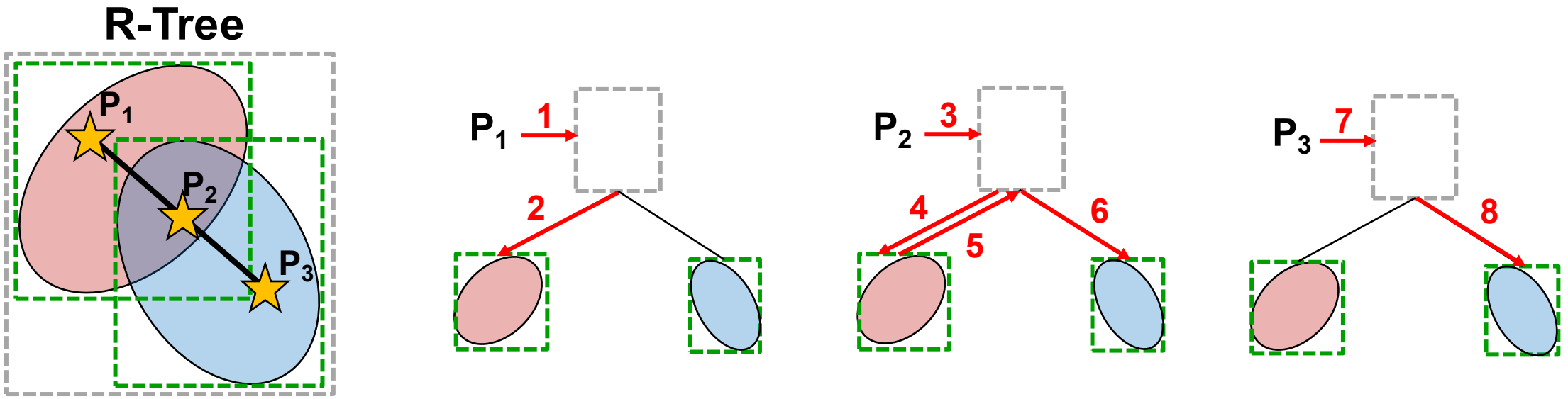
Incremental Map Construction

Map Query Along Trajectory



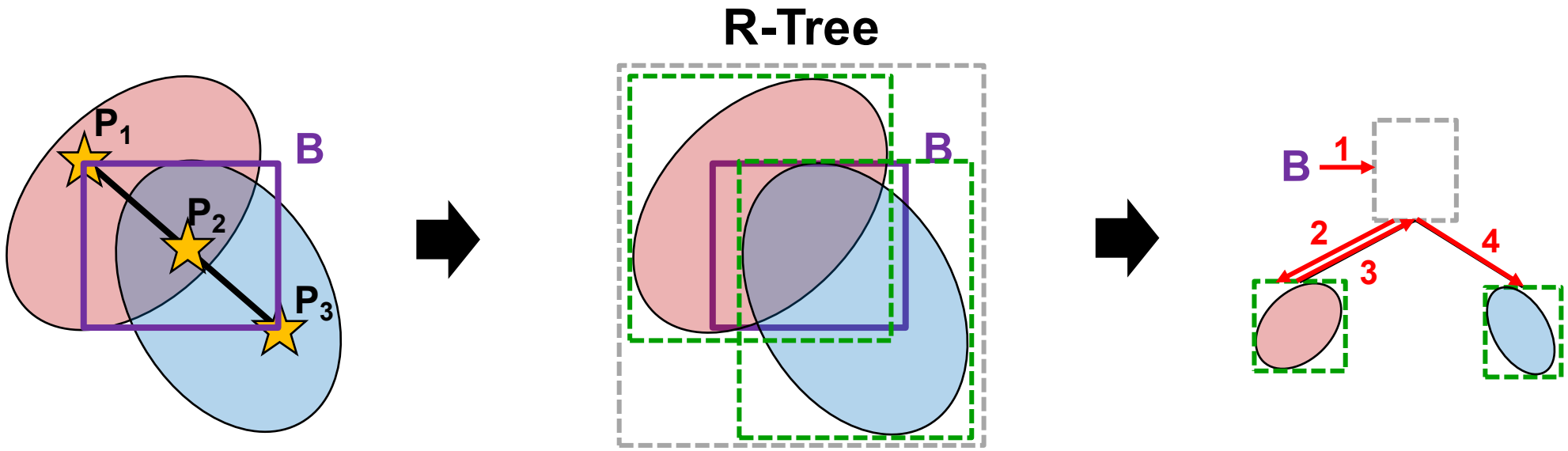
Increase map construction throughput by up to **7**× and map query throughput by up to **9**× via caching spatially-close Gaussians

Redundant Tree Traversal for Single Queries



Each **successive query coordinate** triggers **its own tree traversal** for the **similar set of Gaussians**, causing **largely redundant memory accesses.**

Optimized Tree Traversal for Batch Queries

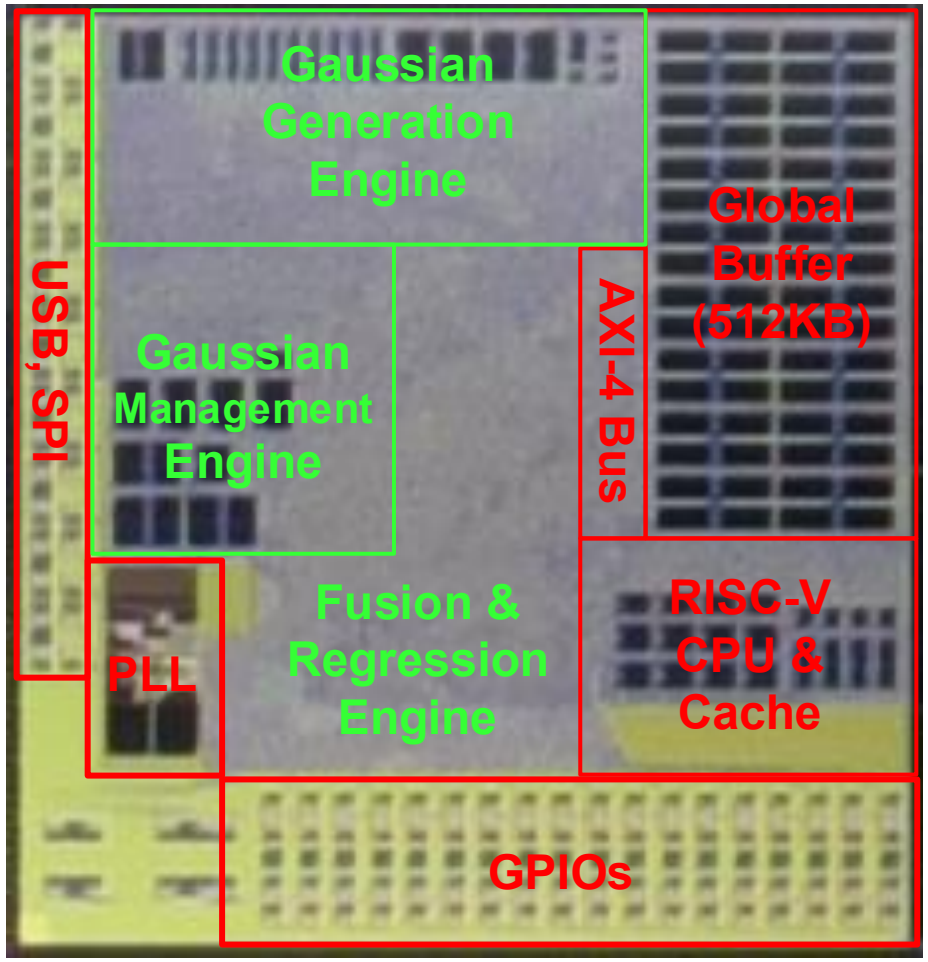


Sharing one tree traversal across 16 nearby query coordinates reduces query energy by **up to 81%** with less than 2% area overhead.

Outline

- Background: 3D Occupancy Mapping
- Chip Architecture and Contributions
- **Chip Specifications and Comparisons**
- Summary

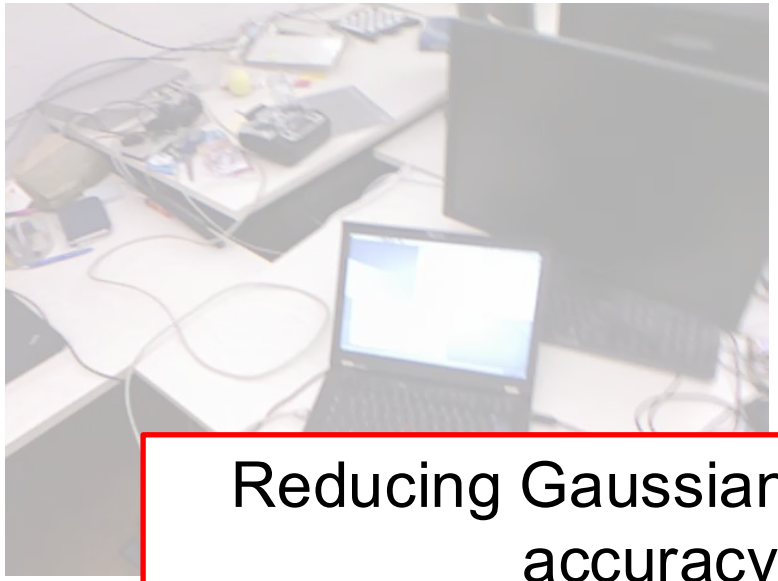
Gleanmer Chip



Technology	16 nm
Area	4mm^2
Logic Gates	3.2 million
SRAM	622 KB
Frequency	125 MHz (construction) 250 MHz (query)
Supply Voltage	0.85 V (construction) 0.95 V (query)
Construction Throughput	88 – 331 fps
Query Throughput	540K – 1320K cps*
Construction Power	4 – 6 mW
Query Power	6 mW

*** cps: coordinate per second**

Evaluation Environments



Reducing Gaussian parameters to **19-bit floating point** preserves map accuracy while shrinking map size by at least **44%**.

Room
(Indoor, $11 \times 12 \times 4$ m^3)

Software	Gleanmer
167 KB	**66 KB**

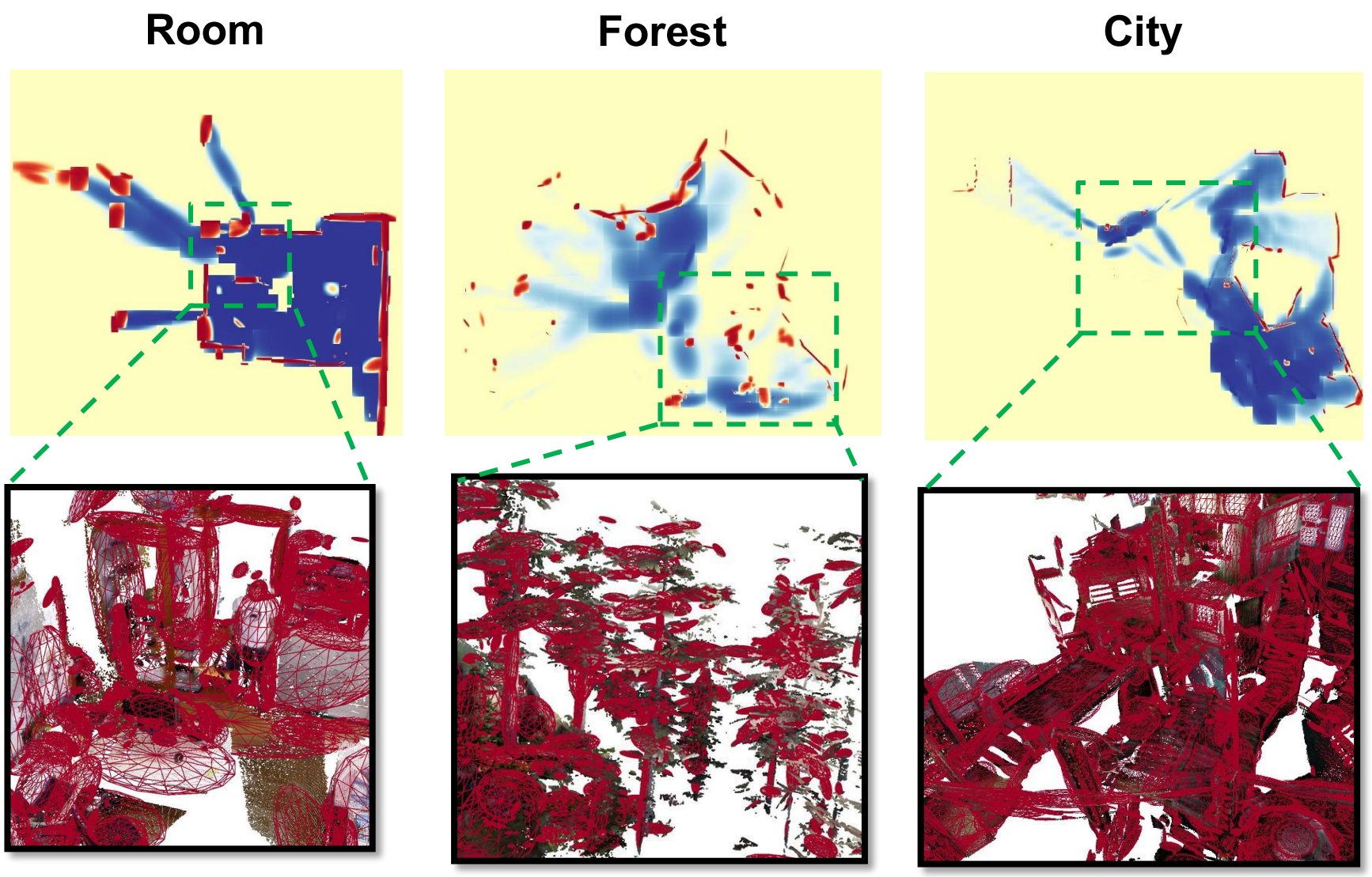
Forest
(Outdoor, $59 \times 53 \times 34$ m^3)

Software	Gleanmer
362 KB	**203 KB**

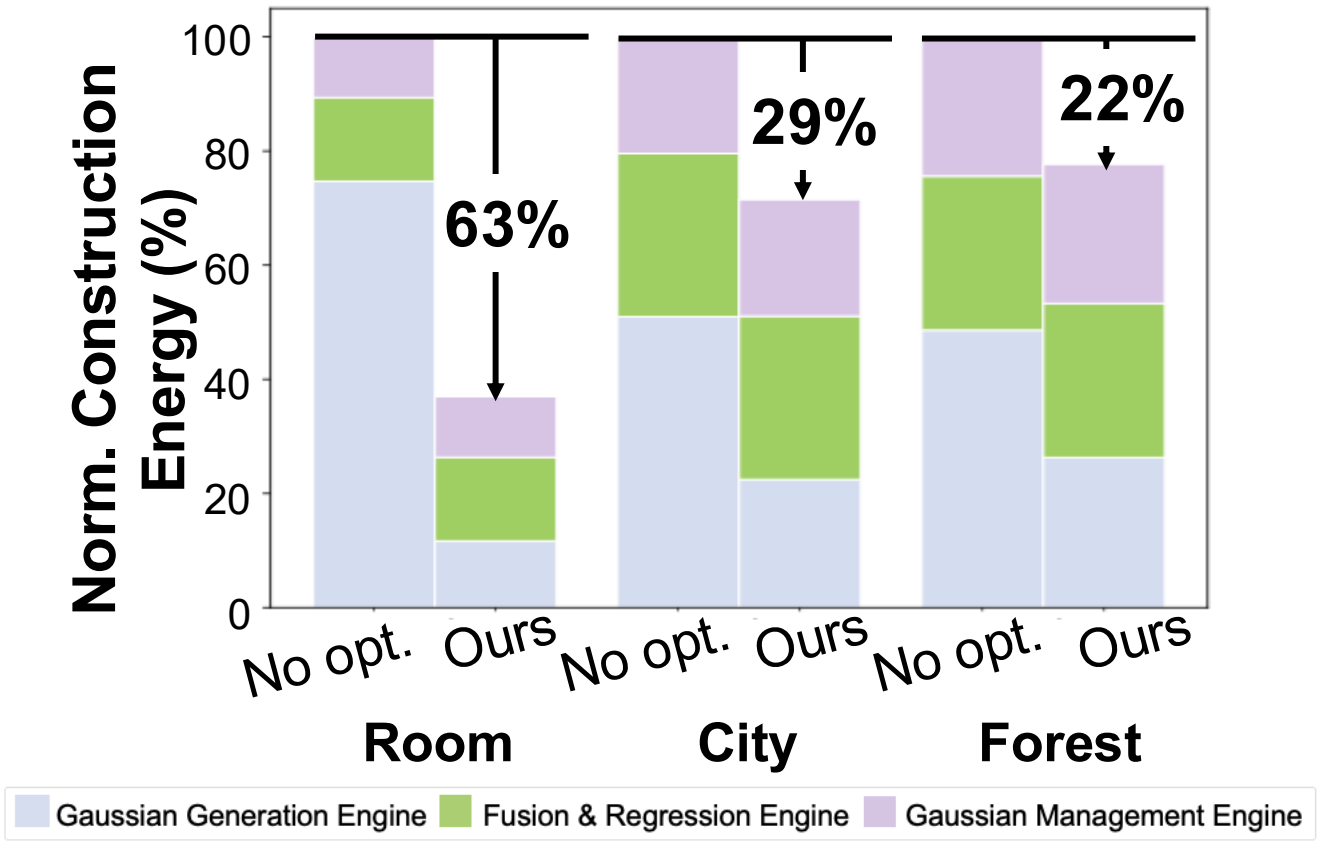
City
(Outdoor, $74 \times 62 \times 43$ m^3)

Software	Gleanmer
850 KB	**356 KB**

Visualization of GMMap

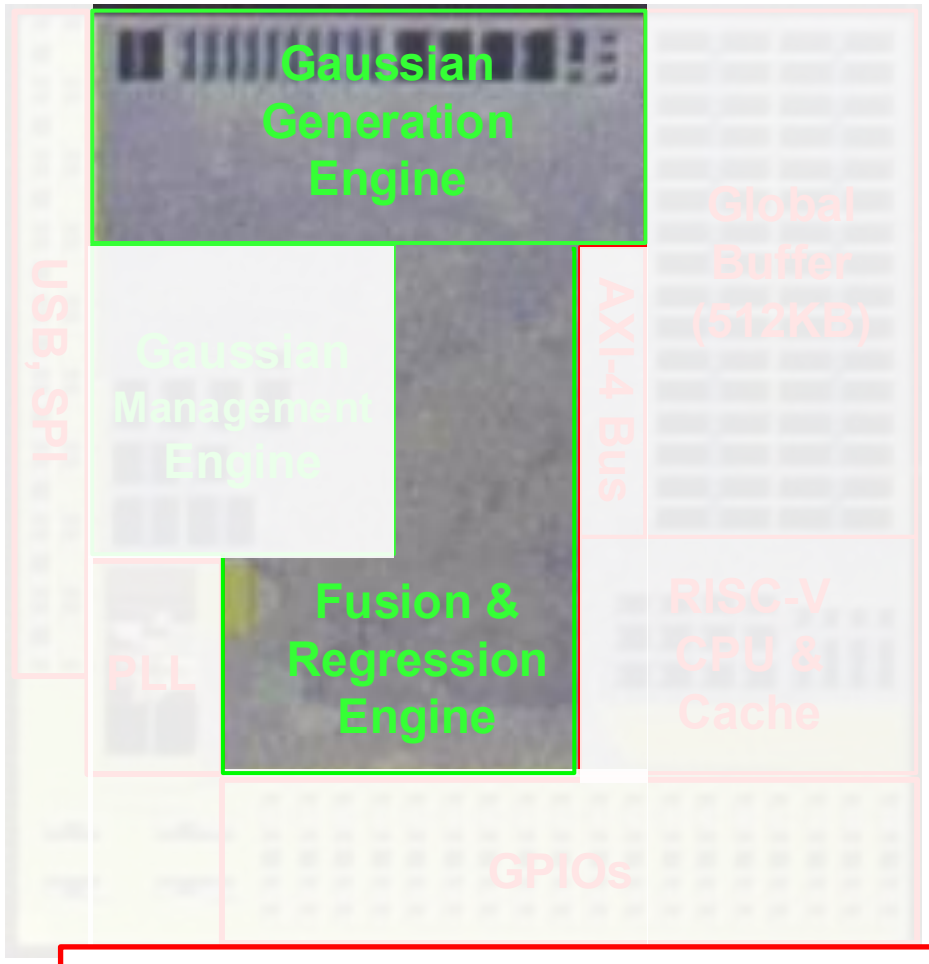


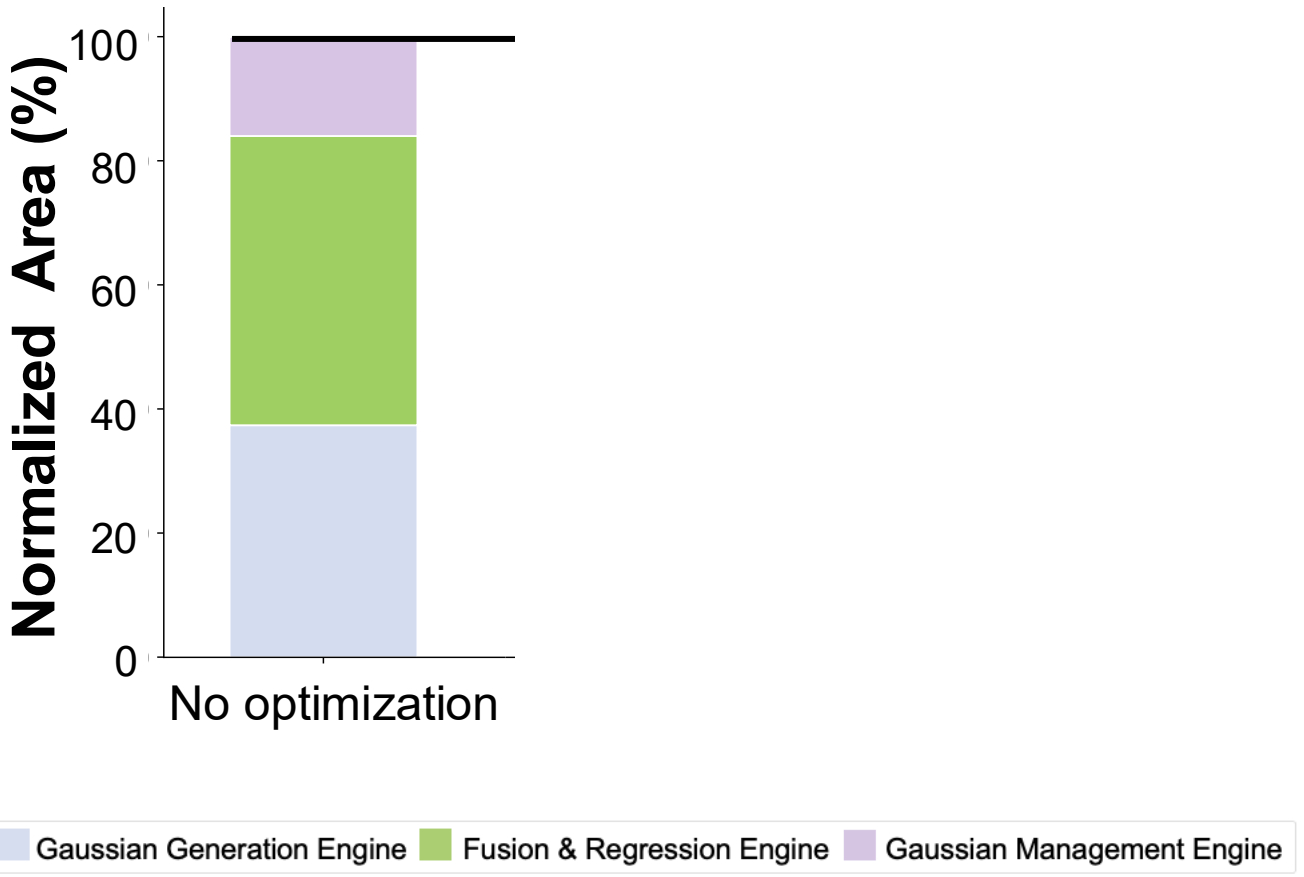
Energy Reduction during Map Construction



Efficient free Gaussian Bases generation from occupied Gaussians reduces map construction energy by **22 - 63%.**

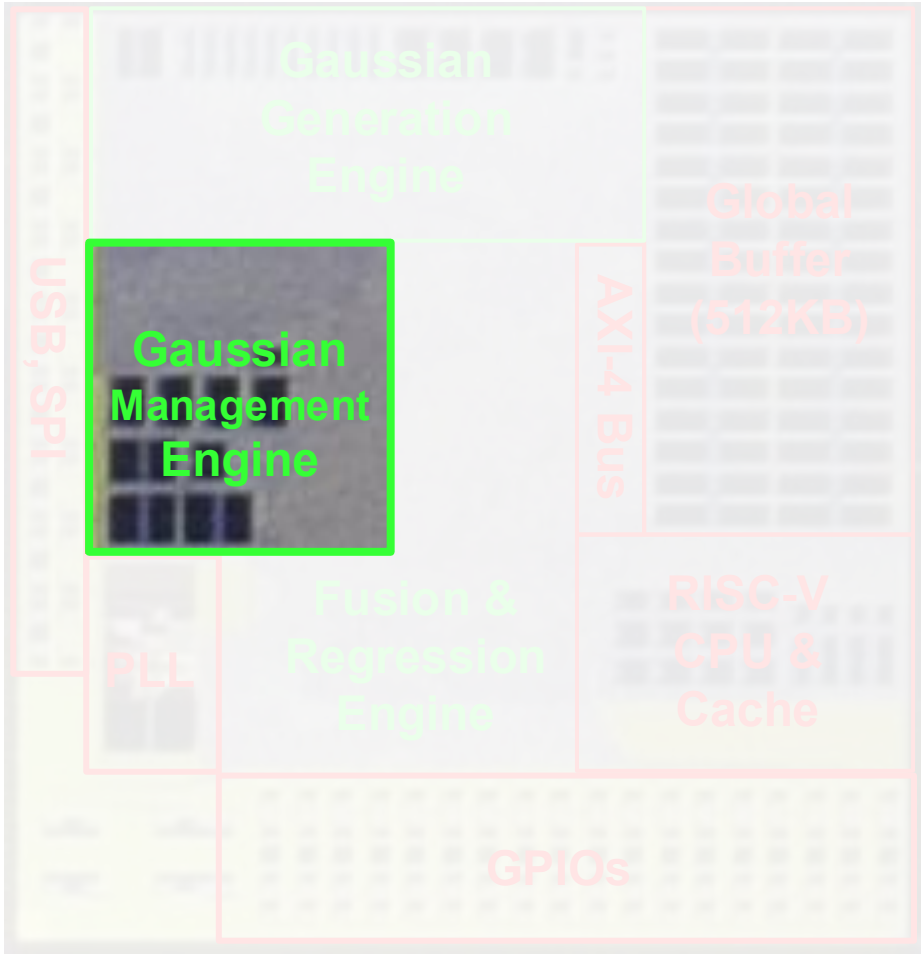
Area Reduction

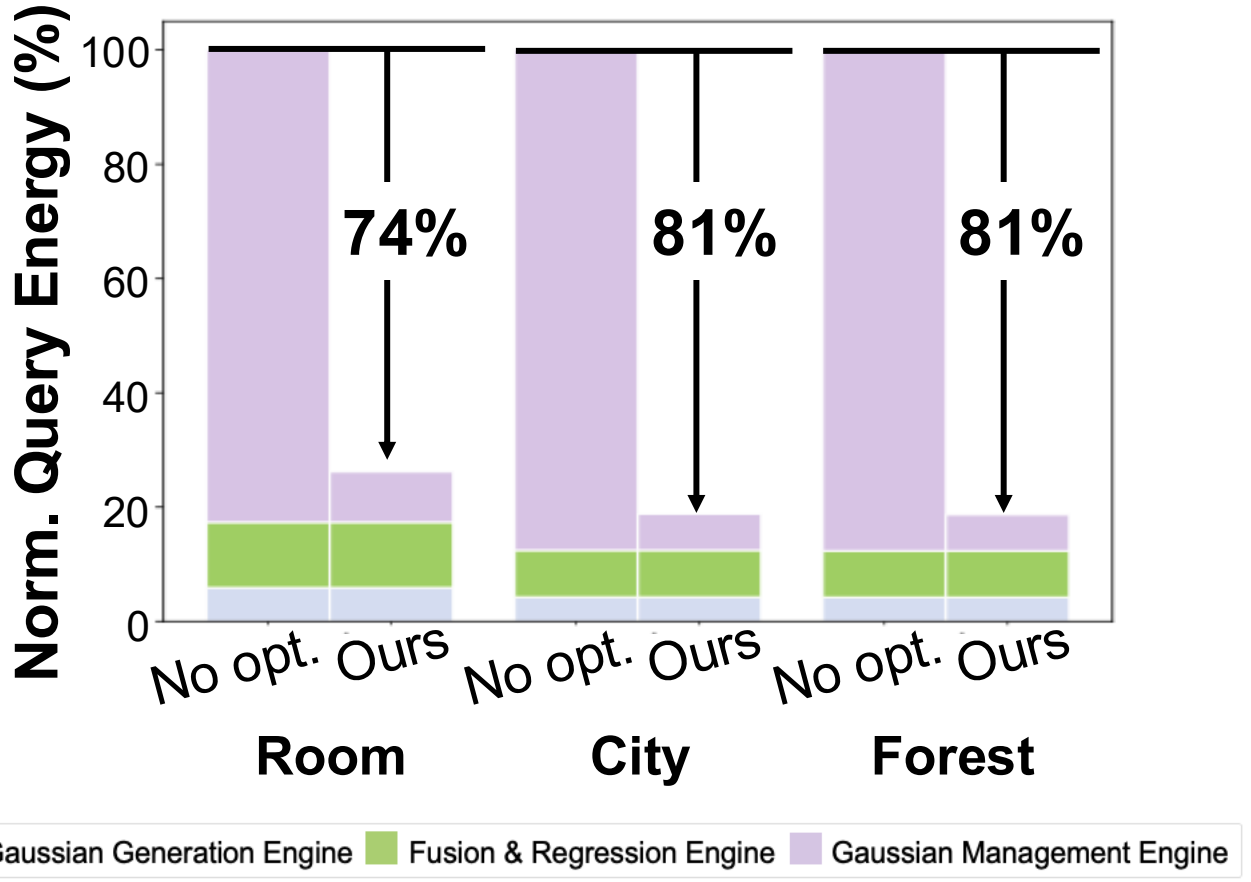




Efficient hierarchy conversion and reducing Gaussian precision reduces accelerator area by **38%**

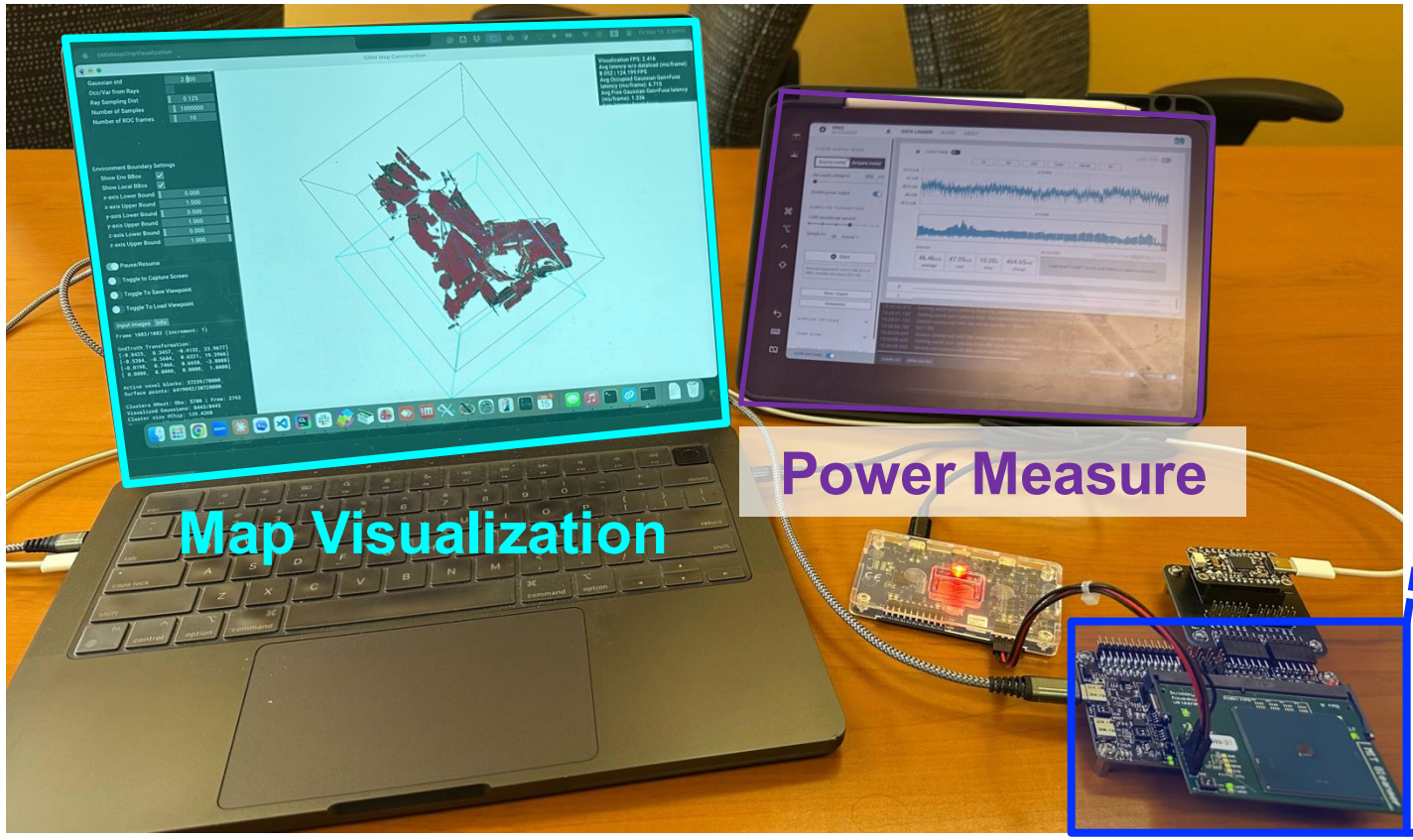
Energy Reduction during Map Query

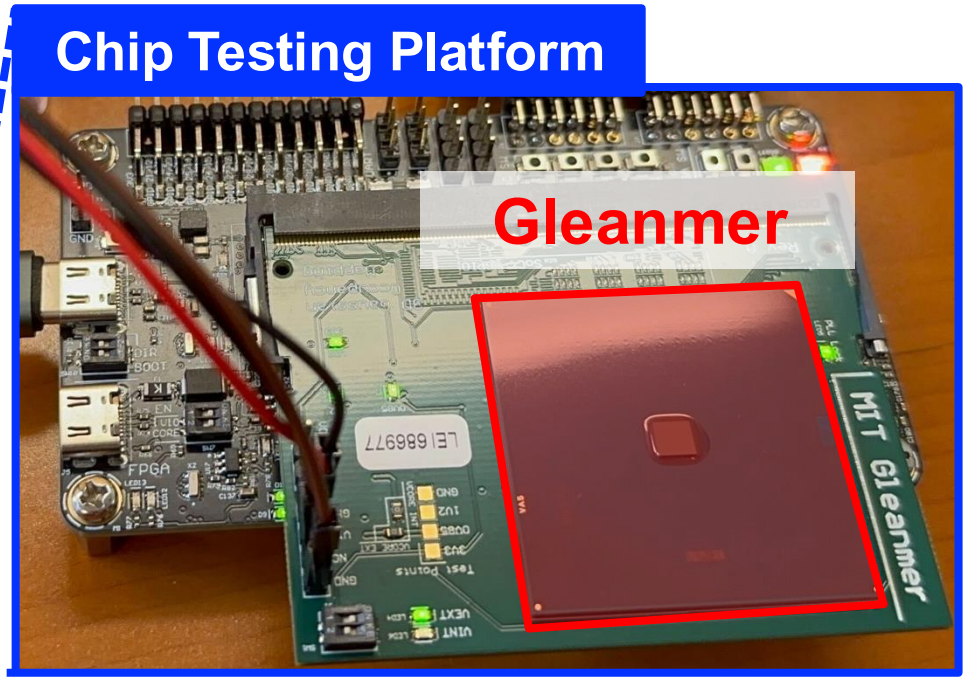




Batch queries across spatially close queries reduces map query energy by **74 - 81%.**

Gleanmer Validation Platform





Gleanmer streams depth images into the chip from the laptop and construct a **real-time 3D occupancy map** on the chip testing platform.

Comparison with Other Hardware

	NVIDIA Jetson TX2 (Cortex A57 + 256-core GPU)	OMU	Gleanmer
Mapping Framework	GMMap	OctoMap	GMMap
Map Accuracy	96% – 99%	93% – 97%	96% – 99%
Map Construction Throughput	44 – 81 fps	61 – 64 fps	**88 – 331 fps**
Map Query Throughput	500 – 800K cps	Not Reported	**540K – 1320K cps**
Average Map Construction Power	4.7 W	251 mW	**4.5 mW**
Average Map Query Power	2.0 W	Not Reported	**5.7 mW**

With comparable accuracy, Gleanmer (16 nm) consumes at least **341× less power than Jetson TX2** (16nm) and **44× less than OMU** (12 nm).

[Jia et al., DATE 2022], [Li et al., T-RO 2024], [Hornung et al., Auton Robot 2013]

Summary

- Gaussian mapping (GMMap) converts data across **three levels of increasing compactness**: depth pixels & rays (H0) → lines & wedges (H1) → Gaussians (H2).
- Gleanmer's **two co-design contributions**:
  - **Efficient hierarchy conversion**: Reduce the volume of work by shifting the free Gaussian bases generation from wedges (H1) to occupied Gaussian (H2).
  - **Amplified optimizations for Gaussians**: Apply reduced precision and exploit data reuse on Gaussians (H2) to reduce area and memory-access intensity.
- **Result**: 63% lower construction energy, 81% lower query energy, 38% smaller area.
- **First fabricated SoC** for real-time 3D Gaussian mapping **under 6 mW**.
- **Sponsors**: MathWorks Fellowship, Amazon, NSF CPS 2400541, Intel USP

Thank you for your attention!

GMMap

Gleanmer