MemPool-3D
Boosting Performance and Efficiency of Shared-L1 Memory Many-Core Clusters with 3D Integration

Matheus Cavalcante, Anthony Agnesina, Samuel Riedel, Moritz Brunion,

Alberto García-Ortiz, Dragomir Milojevic, Francky Catthoor, Sung Kyu Lim, Luca Benini

MemPool: PULP's scaled-up shared-L1 system

- Shared-L1 memory is a very common architectural pattern
  - Only scaled-up to a few tens of cores

- **MemPool** takes this to the extreme
  - **256 cores**, sharing **1 MiB of L1**, divided into 1024 SRAM banks, in **5 cycles** of latency
  - **500 MHz** (w.c.) at GlobalFoundries **22FDX** technology
  - And **open-source**, as it should be!
    - pulp-platform/mempool

Connecting 256 cores to 1024 SRAM banks

- It is no easy feat to connect 256 cores and 1024 memory banks

- Hierarchical interconnection to avoid major routing congestion
  - And still, congestion is a major issue

- Low latency target is very constraining
  - We need to cross the whole macro with a single pipeline stage at the middle
  - Wire propagation delay limits the operating frequency
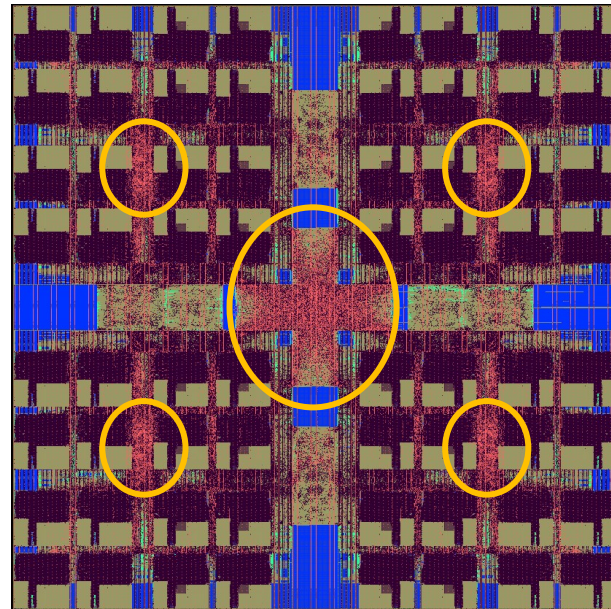

4.6mm × 4.6mm 22FDX MemPool cluster

MemPool: a prime benchmark for F2F 3D ICs

- 3D Integration tackles MemPool's implementation issues
  - MemPool is limited by **wire propagation delay** → **vertical integration reduces** the **design footprint** and **wire length**
  - MemPool is **routing congested** → with **Macro-3D** we can **share the BEOLs of both dies** to avoid congestion bottlenecks

- *How much PPA can we gain from MemPool-3D?*
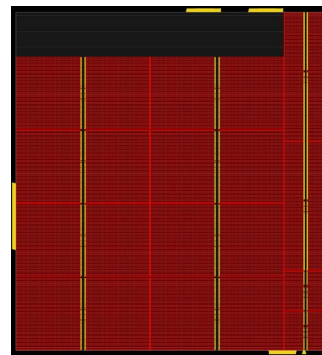
Tile Floorplanning and Partitioning

- MemPool-3D's tile memory die floorplan
  - 16, 32, 64, 128 KiB of L1 SPM
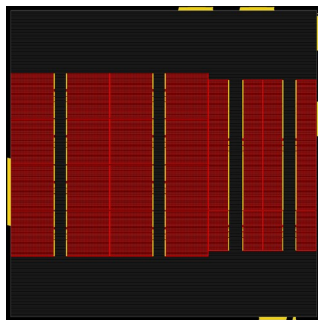
MemPool-3D, 16 KiB

Utilization: 51%
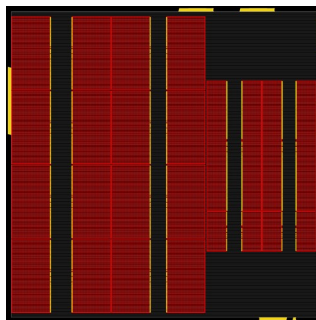
MemPool-3D, 32 KiB

Utilization: 65%

MemPool-3D, 64 KiB
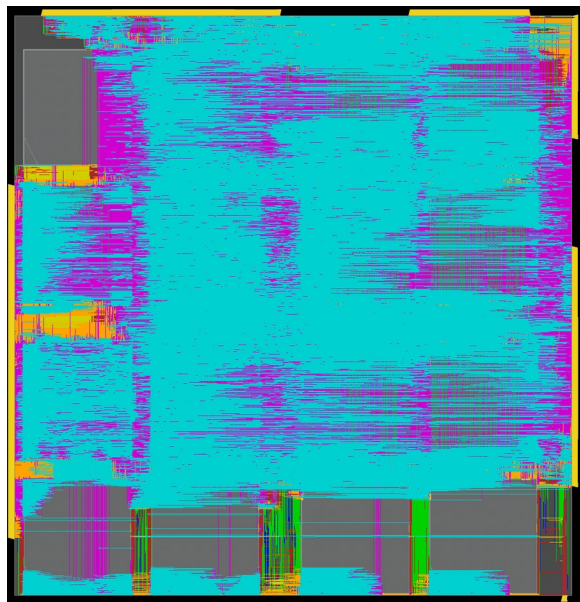
Utilization: 89%

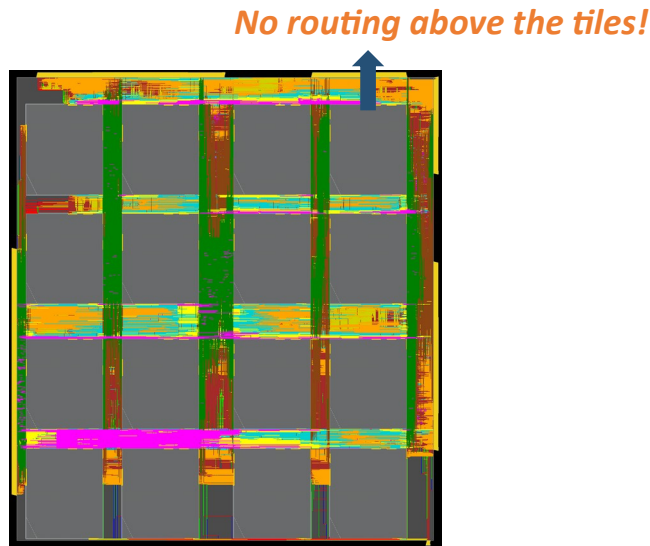MemPool-3D, 128 KiB

Utilization: 100%

Matheus Cavalcante/ETH Zürich

Group implementations in 2D and 3D

128 KiB of L1 per tile (8 MiB total)

No routing above the tiles!

Footprint 46% smaller!

2D

3D

Groups: MemPool-2D vs. MemPool-3D

	MemPool-2D		**MemPool-3D**
1 MiB	*Footprint:* **1.000**		*Footprint:* **0.665**
	↓ *7% larger*		↓ *Same footprint!*
2 MiB	*Footprint:* **1.074**		*Footprint:* **0.665**
4 MiB	*Footprint:* **1.299**		*Footprint:* **0.737**
			14% smaller, despite 8x the L1 capacity!
8 MiB	*Footprint:* **1.572**		*Footprint:* **0.857**

Similar trends for **Wire length** and **#Buffers**

Groups: MemPool-2D vs. MemPool-3D

	MemPool-2D	**MemPool-3D**
1 MiB	*Frequency: **1.000***	*Frequency: **1.040***
2 MiB	*Frequency: **0.930***	*Frequency: **0.979***
4 MiB	*Frequency: **0.875***	*Frequency: **0.955***
8 MiB	*Frequency: **0.885***	*Frequency: **0.930***

4% higher

12% lower

Similar trends for **Power** and **PDP**
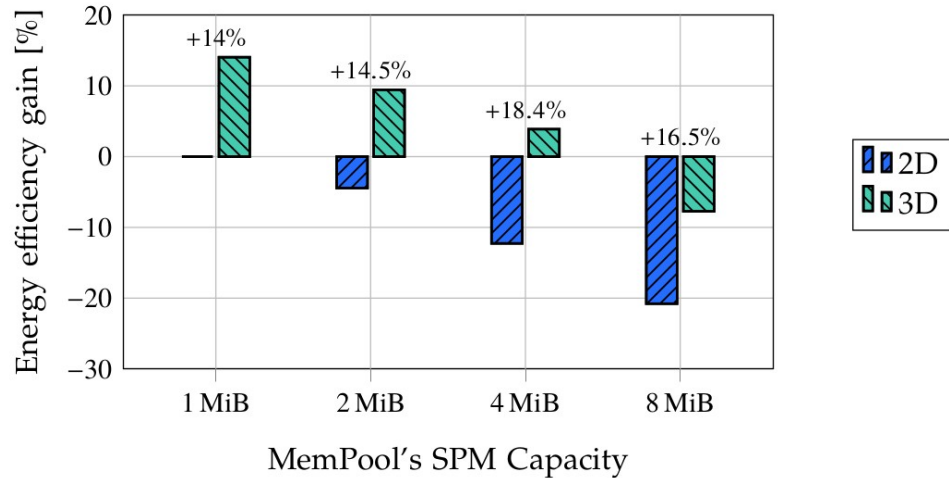
Putting it all together: Energy Efficiency

- *MemPool-3D consistently outperforms MemPool-2D*
  - Smaller footprint leading to fewer buffers, shorter wire length, and smaller power consumption

- Larger L1 capacity → decreased energy efficiency
  - Larger SRAM banks, larger leaking

MemPool-3D

Boosting Performance and Efficiency of Shared-L1 Memory Many-Core Clusters with 3D Integration

Matheus Cavalcante, Anthony Agnesina, Samuel Riedel, Moritz Brunion,

Alberto García-Ortiz, Dragomir Milojevic, Francky Catthoor, Sung Kyu Lim, Luca Benini

ETH zürich   Georgia Tech   Universität Bremen   imec   ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA A.D. 1088   PULP Parallel Ultra Low Power