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
• 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](https://pulp-platform/mempool)
• 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
• 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%
Group implementations in 2D and 3D

128 KiB of L1 per tile (8 MiB total)

2D

3D

No routing above the tiles!

Footprint 46% smaller!
Groups: MemPool-2D vs. MemPool-3D

<table>
<thead>
<tr>
<th>MemPool-2D</th>
<th>MemPool-3D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Footprint: 1.000</td>
<td>Footprint: 0.665</td>
</tr>
<tr>
<td>7% larger</td>
<td>Same footprint!</td>
</tr>
<tr>
<td>Footprint: 1.074</td>
<td>Footprint: 0.665</td>
</tr>
<tr>
<td>Footprint: 1.299</td>
<td>Footprint: 0.737</td>
</tr>
<tr>
<td>14% smaller, despite 8x the L1 capacity!</td>
<td>Footprint: 0.857</td>
</tr>
<tr>
<td>Footprint: 1.572</td>
<td></td>
</tr>
</tbody>
</table>

Similar trends for Wire length and #Buffers
Groups: MemPool-2D vs. MemPool-3D

<table>
<thead>
<tr>
<th>Size</th>
<th>MemPool-2D Frequency</th>
<th>MemPool-3D Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 MiB</td>
<td>1.000</td>
<td>1.040</td>
</tr>
<tr>
<td>2 MiB</td>
<td>0.930</td>
<td>0.979</td>
</tr>
<tr>
<td>4 MiB</td>
<td>0.875</td>
<td>0.955</td>
</tr>
<tr>
<td>8 MiB</td>
<td>0.885</td>
<td>0.930</td>
</tr>
</tbody>
</table>

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