Designing Vertical Processors in Monolithic 3D

Bhargava Gopireddy, Josep Torrellas
University of Illinois at Urbana-Champaign

http://iacoma.cs.uiuc.edu

ISCA 2019
What is Monolithic 3D (M3D) ?

- Fabrication method that sequentially grows one or more Si layers on top of a base layer
- As opposed to TSV based 3D stacking (TSV3D), that bonds pre-fabricated dies together
What is Monolithic 3D (M3D) ?

• Fabrication method that sequentially grows one or more Si layers on top of a base layer
  • As opposed to TSV based 3D stacking (TSV3D), that bonds pre-fabricated dies together
Monolithic 3D vs TSV based 3D Stacking (TSV3D)

• Monolithic Inter-layer Via (MIV) diameter is much smaller
  • Diameter: 50 nm (MIV) vs 1.3 um (TSV)

• Layers are close to each other
  • Better vertical thermal conduction

✓ Fine-grained integration
✓ High bandwidth and low latency
✓ Good vertical thermal conduction
Challenges of Monolithic 3D Integration

• Manufacturing the top layer without compromising the base layer
  • Conventional approach: High temperature steps in top layer fabrication
  • This hurts the bottom layer (metal and active silicon)

• Alternative: Use lower temperature process for the top layer

✔ Doesn’t hurt the bottom layer
✗ Performance degradation in top layer
Contribution: Designing Vertical Processors in Monolithic 3D

• Design vertical cores by partitioning the pipeline stages into two layers -- with and without top layer performance degradation

• First, partition the core assuming no slowdown in top layer

• Next, mitigate the impact of the top layer slowdown
  • Critical Path Aware Partitioning for logic and storage stages.

• Our evaluation shows an M3D core can significantly improve performance over a 2D or TSV3D design while reducing energy consumption
Partitioning Storage Stages in M3D

• SRAM/CAM structures within the core:
  • *Multi-ported*: Register File, Register Alias Table, Issue Queue etc.
  • *Single-ported*: Branch predictors, BTB, L1/L2 etc.

• 3D SRAM/CAM array partitioning schemes [1]
  • Bit Partitioning
  • Word Partitioning
  • Port Partitioning

Bit Partitioning

- Distribute half the bits of every word in each layer
- Each word requires an inter-layer via

- Multiported structure (e.g. Register File):
  - Bigger bitcells $\rightarrow$ higher gains

- Singleported structure (e.g. Branch Predictor):
  - Smaller bitcells $\rightarrow$ lower gains
Word Partitioning

- Distribute half of the words in each layer
- Requires one inter-layer via per bit
- Similar behavior as Bit partitioning

![Diagram showing word partitioning](image)

![Bar chart showing percentage latency reduction over 2D](chart)

- Register File: M3D 27, TSV3D 24
- Branch Prediction: M3D 14, TSV3D -6
Port Partitioning

- Distribute half the ports in each layer
- Quadratic reduction in area footprint
- Two vias per bitcell
- Applicable only for multiported structures
- TSVs simply have too much area overhead

![Diagram showing port partitioning and vias](image)

<table>
<thead>
<tr>
<th>Percentage Latency Reduction over 2D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Register File</td>
</tr>
<tr>
<td>M3D</td>
</tr>
<tr>
<td>-361</td>
</tr>
<tr>
<td>41</td>
</tr>
<tr>
<td>0</td>
</tr>
</tbody>
</table>

---

i-acoma group
Port Partitioning

- Distribute half the ports in each layer
- Quadratic reduction in area footprint
- Two vias per bitcell
- Applicable only for multiported structures
- TSVs simply have too much area overhead

Multiported structures → Port Partitioning is best
Singleported structures → Bit Partitioning is good
Partitioning Logic Stages in M3D

- Benefit: Reduction in wire delay

- Semi-global wires are significantly shorter
  - Critical paths within a core such as load-to-use, branch misprediction, ALU+Bypass are shorter
  - Improves both core frequency and IPC
Hetero Layer Partitioning (M3D-Hetero)

• Low temperature processing $\rightarrow$ performance degradation in the top layer
  • Top layer inverter delay: 17% lower
  • Overall frequency: 9% lower

• Propose: Mitigate the impact through Critical Path Aware Partitioning
M3D-Hetero Logic Stages: Execution Unit

• ALU has many paths that are not critical to stage delay
  • For example, in a carry skip adder, only shaded parts are critical.
  • Propose: Place the critical paths in bottom layer; rest in the top layer
M3D-Hetero Storage Structures

• Port Partitioned Structures:
  • Area slack exists in top layer, as the bitcell is only in the bottom layer
  • Propose: *Asymmetric Port Partitioning*
    • Increase the size of access transistors in the top layer (for speed)
    • Place more ports in bottom layer than in top layer
M3D-Hetero Storage Structures

• Port Partitioned Structures:
  • Area slack exists in top layer, as the bitcell is only in the bottom layer
  • Propose: *Asymmetric Port Partitioning*
    • Increase the size of access transistors in the top layer (for speed)
    • Place more ports in bottom layer than in top layer

• Bit/Word Partitioned Structures:
  • Propose: *Asymmetric Array Partitioning*
    • Asymmetrically partition bits/words across layers
    • Assign smaller section to the top layer
    • Increase the size of transistors in top layer (for speed)
M3D-Hetero Stages with both Logic and SRAM Structures

• Rename consists of Register Alias Table (RAT) and other logic:
  • Register Alias Table (multi-ported) → Perform asymmetric port partitioning
  • Dependency check → Place in top layer (Not critical)
  • Shadow RAT tables → Place in top layer (Not critical)
  • Decoder to RAT and peripheral logic → Place in bottom layer (Critical)
## Summary: Hetero Layer Partitioning in M3D

<table>
<thead>
<tr>
<th>Stage Type</th>
<th>Mitigation Technique</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logic Stages</td>
<td>Critical paths in bottom layer; non-critical paths in top</td>
</tr>
<tr>
<td>Storage Stages</td>
<td>Port Partitioning</td>
</tr>
<tr>
<td></td>
<td>Asymmetric partitioning of ports; and larger access transistors in top layer</td>
</tr>
<tr>
<td></td>
<td>Bit/Word Partitioning</td>
</tr>
<tr>
<td></td>
<td>Asymmetric partitioning of array; and larger bitcells in top layer</td>
</tr>
<tr>
<td>Mixed Stages</td>
<td>Combination of above techniques</td>
</tr>
</tbody>
</table>
How to Exploit Stage Delay Reduction in M3D?

- Increase frequency

- Increase the size or the number of ports for key structures
  - Register File, Issue Queue etc.
  - Increase issue width at same frequency

- Keep the same frequency, but reduce the voltage
  - Operate more cores within the same power budget
Evaluation: Configurations & Frequency

• **2D Base**: Four 6-wide OOO cores
  - Caches: DL1 & IL1 32KB ; L2 256 KB ; Entries: RF(160) ; ROB(192) ; IQ(84) ; LQ(72) ; SQ(56)
  - Frequency = 3.3 GHz

• **M3D-Het**: M3D with top layer slowdown and our modifications
  - Stage with lowest delay reduction: SQ/BPT at 13%
  - Frequency = 3.79 GHz

• **TSV3D**: Traditional TSV3D
  - Stage with lowest delay reduction: BTB at -6% (negative)
  - Frequency = 3.3 GHz

• **M3D-Het-2X**: Increase cores within same power budget
  - 2X as many cores at F = 3.3 GHz
Evaluation of M3D Design Choices

![Bar chart showing evaluation of M3D design choices. The chart compares the normalized execution times and energy consumption of 2DBase, TSV3D, M3D-Het, and M3D-Het-2X. The x-axis represents Avg Execution Time and Avg Energy, while the y-axis represents the normalized values.](image)
Evaluation of M3D Design Choices

<table>
<thead>
<tr>
<th></th>
<th>Avg Execution Time</th>
<th>Avg Energy</th>
</tr>
</thead>
<tbody>
<tr>
<td>2DBase</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>TSV3D</td>
<td>0.8</td>
<td>0.8</td>
</tr>
<tr>
<td>M3D-Het</td>
<td>0.8</td>
<td>0.8</td>
</tr>
<tr>
<td>M3D-Het-2X</td>
<td>0.8</td>
<td>0.8</td>
</tr>
</tbody>
</table>
Evaluation of M3D Design Choices

M3D executes faster while consuming lower energy.
Evaluation of M3D Design Choices

Can operate twice as many M3D cores as 2DBase in a similar power budget

### More Results

<table>
<thead>
<tr>
<th>Design Variant</th>
<th>Avg Execution Time</th>
<th>Avg Energy</th>
<th>Normalized to 2DBase</th>
</tr>
</thead>
<tbody>
<tr>
<td>2DBase</td>
<td>1.00</td>
<td>0.50</td>
<td>1.00</td>
</tr>
<tr>
<td>TSV3D</td>
<td>0.54</td>
<td>0.46</td>
<td>0.88</td>
</tr>
<tr>
<td>M3D-Het</td>
<td>0.46</td>
<td>0.39</td>
<td>0.95</td>
</tr>
<tr>
<td>M3D-Het-2X</td>
<td>0.44</td>
<td>0.39</td>
<td>0.93</td>
</tr>
</tbody>
</table>

- **Avg Execution Time**: 46% improvement over 2DBase
- **Avg Energy**: 39% improvement over 2DBase

Normalized to 2DBase
Thermal Behavior: Peak Temperature

- M3D has good thermal behavior

Bar chart showing peak temperature comparisons:
- Peak Temperature (Deal II)
- Avg Peak Temperature (Across applications)

Legend:
- 2DBase
- TSV3D
- M3D-Het
Summary: Designing Vertical Processors in Monolithic 3D

• First work to partition a core in an M3D stack

• Proposed Critical Path Aware Partitioning strategies to mitigate the performance impact of a degraded top layer.

• Overall, M3D can operate twice as many cores as a 2D design within a similar power budget
  • Improves execution time by 46% while consuming 39% lower energy.
Designing Vertical Processors in Monolithic 3D

Bhargava Gopireddy, Josep Torrellas
University of Illinois at Urbana-Champaign

http://iacoma.cs.uiuc.edu

ISCA 2019
Backup
Caches and NoC

• Shared L2 Cache, and NoC Router Stop across two adjacent cores
Stages with Logic and SRAM Structures

• Fetch and Branch Prediction:
  • Consists of 4 parallel paths: Increment PC, RAS, BTB and Branch Prediction
  • RAS, increment PC are non-critical and are in top tier
  • BTB access and Branch Prediction are both critical and use asymmetric array partition
Stages with Logic and CAM Structures

• Issue Stage consists of Wakeup and Select logic:
  • Wakeup consists of IQ which is a CAM structure \(\rightarrow\) Asymmetric port partitioning
  • Select consists of multi-level arbitration
  • Within each arbiter:
    • First phase: Request \(\rightarrow\) critical and is in bottom layer
    • Second phase: Grant \(\rightarrow\) has slack and is in top layer.
Stages with Logic and CAM Structures

• Load Store Unit:
  • Critical path: Store forwarding to a younger load, on a hit in Store Queue
    • Includes SQ access, Priority encoder and Store buffer access
  • LQ, SQ CAMs \(\rightarrow\) Asymmetric Port partitioning
  • Priority encoder after SQ is placed in bottom layer.
  • Balance area between LQ/SQ across layers.
Evaluation: Frequency of Operation for M3D

- **2D Base**: 6-wide OOO core
  - F = 3.3 GHz
- **IsoM3D**: M3D without top layer degradation:
  - Stage with least delay reduction: SQ/BPT at 14% → F = 3.83 GHz
- **HetM3D–Naïve**: 9% lower frequency due to slow top layer
  - F = 3.5 GHz
- **HetM3D**: M3D with top layer slowdown and our modifications
  - Stage with least delay reduction: SQ/BPT at 13% → F = 3.76 GHz
- **HetM3D-Agg**: An aggressive HetM3D with traditional frequency critical structures
  - IQ (24%) and ALU+Bypass (28%) are the only limiters → F = 4.34 GHz
Single Thread Results: SPEC Benchmark Suite

- Avg Speedup
- Avg Energy

Normalized to 2DBase

Graph shows normalized performance data for different benchmarks.

- 2DBase
- TSV3D
- IsoM3D
- HetM3D-Naïve
- HetM3D
- HetM3D-Agg
Single Thread Results: SPEC Benchmark Suite

- Avg Speedup
- Avg Energy

Normalized to 2DBase

- 2DBase
- TSV3D
- IsoM3D
- HetM3D-Naïve
- HetM3D
- HetM3D-Agg
Single Thread Results: SPEC Benchmark Suite

![Graph showing average speedup and average energy normalized to 2DBase for different benchmarks.]

- 2DBase
- TSV3D
- M3D-Iso
- M3D-HetNaïve
- M3D-Het
- M3D-HetAgg
Single Thread Results: SPEC Benchmark Suite

- Avg Speedup
- Avg Energy

Normalized to 2DBase

- 2DBase
- TSV3D
- M3D-Iso
- M3D-HetNaïve
- M3D-Het
- M3D-HetAgg
Single Thread Results: SPEC Benchmark Suite

![Bar Chart](image)

- Avg Speedup
- Avg Energy

Normalized to 2DBase

- 2DBase
- TSV3D
- M3D-Iso
- M3D-HetNaïve
- M3D-Het
- M3D-HetAgg
Single Thread Results: SPEC Benchmark Suite

Normalized to 2DBase

- Avg Speedup
- Avg Energy

Bar chart showing performance metrics for different benchmarks:

- 2DBase
- TSV3D
- M3D-Iso
- M3D-HetNaïve
- M3D-Het
- M3D-HetAgg
Evaluation of M3D Design Choices

![Bar chart showing normalized to 2DBase for Avg Execution Time and Avg Energy. The x-axis is labeled Avg Execution Time on the left and Avg Energy on the right. The y-axis is labeled Normalized to 2DBase with values ranging from 0 to 1.4. The chart compares 2DBase, TSV3D, M3D-Het, M3D-Het-Wide, and M3D-Het-2X.]
Evaluation of M3D Design Choices

![Bar graph showing normalized execution time and energy consumption for different design choices normalized to 2DBase.]

- **2DBase**
- **TSV3D**
- **M3D-Het**
- **M3D-Het-Wide**
- **M3D-Het-2X**

**Axes:**
- Y-axis: Normalized to 2DBase
- X-axis 1: Avg Execution Time
- X-axis 2: Avg Energy
Evaluation of M3D Design Choices

M3D executes faster while consuming lower energy.

- 2DBase
- TSV3D
- M3D-Het
- M3D-Het-Wide
- M3D-Het-2X

20% increase in execution time
33% decrease in energy consumption
Evaluation of M3D Design Choices

Increasing issue width provides similar results

- Increasing issue width provides similar results
Evaluation of M3D Design Choices

Can operate twice as many M3D cores as 2DBase in a similar power budget

- 2DBase
- TSV3D
- M3D-Het
- M3D-Het-Wide
- M3D-Het-2X

Can operate twice as many M3D cores as 2DBase in a similar power budget.
Monolithic 3D Core -- Partitioning Storage Stages

• Storage structures i.e. SRAM/CAM delays are proportional to height and width of an array

\[
\text{Width} = N\downarrow\text{bits} \times (\text{BitcellWidth} + K \times (N\downarrow r + N\downarrow w))
\]

\[
\text{Height} = N\downarrow\text{words} \times (\text{BitcellHeight} + K \times (N\downarrow r + N\downarrow w))
\]

\[
\text{Area} = \text{Width} \times \text{Height} \propto (N\downarrow r + N\downarrow w)^2
\]

Area of multiported cell is quadratic on the number of ports