Designing Vertical Processors in Monolithic 3D

Bhargava Gopireddy, Josep Torrellas

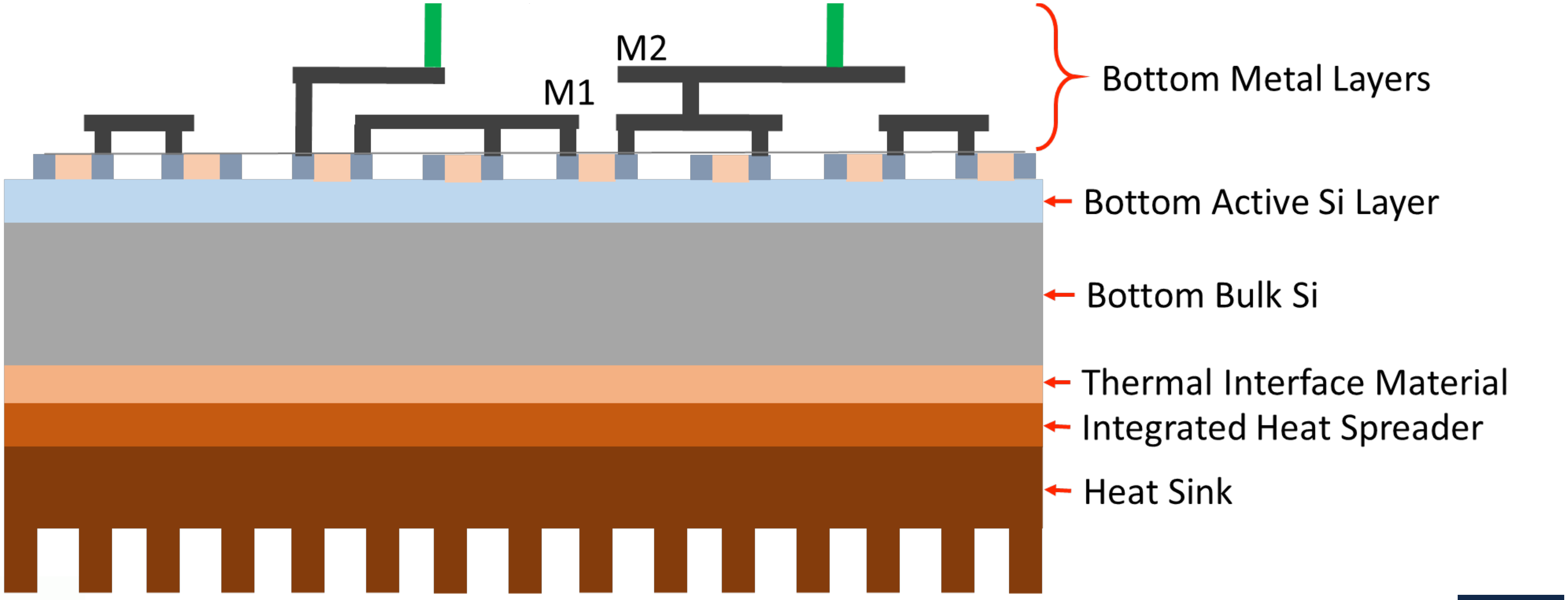
University of Illinois at Urbana-Champaign

http://iacoma.cs.uiuc.edu

ISCA 2019

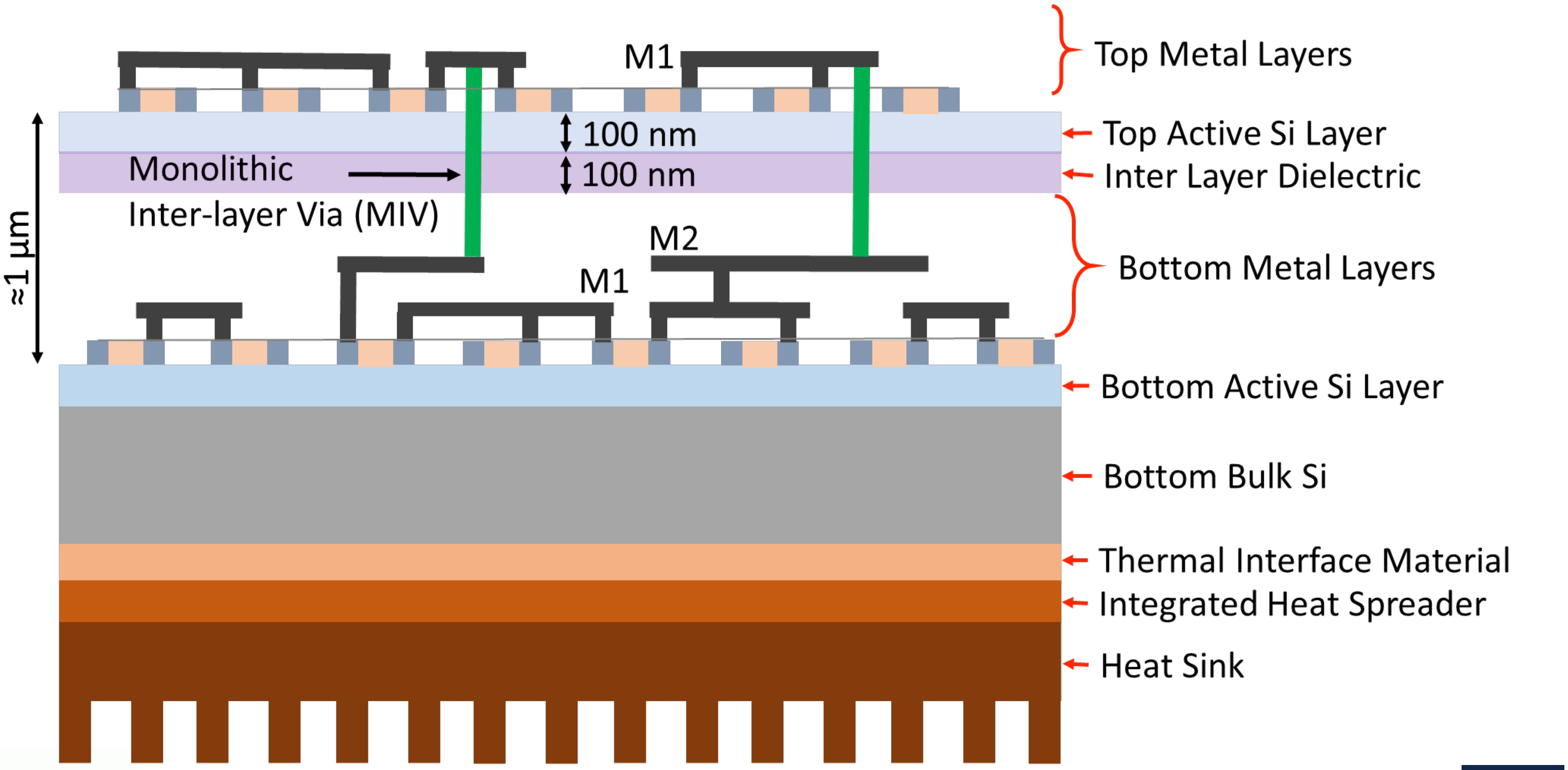
What is Monolithic 3D (M3D) ?

- 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) ?

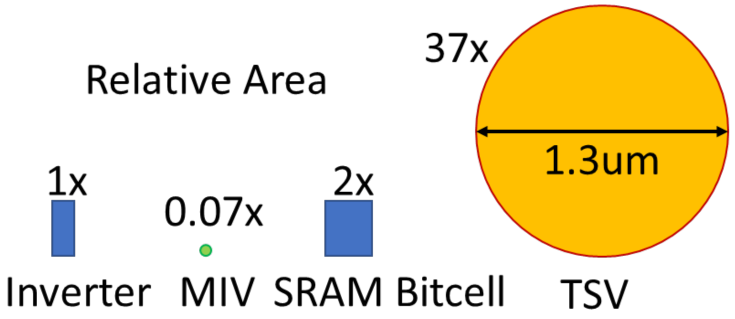
- Fabrication method that sequentially grows one or more Si layers on top of a base layer
  - As opposed to TSV based 3D stacking (TSV3D), that bonds pre-fabricated dies together



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

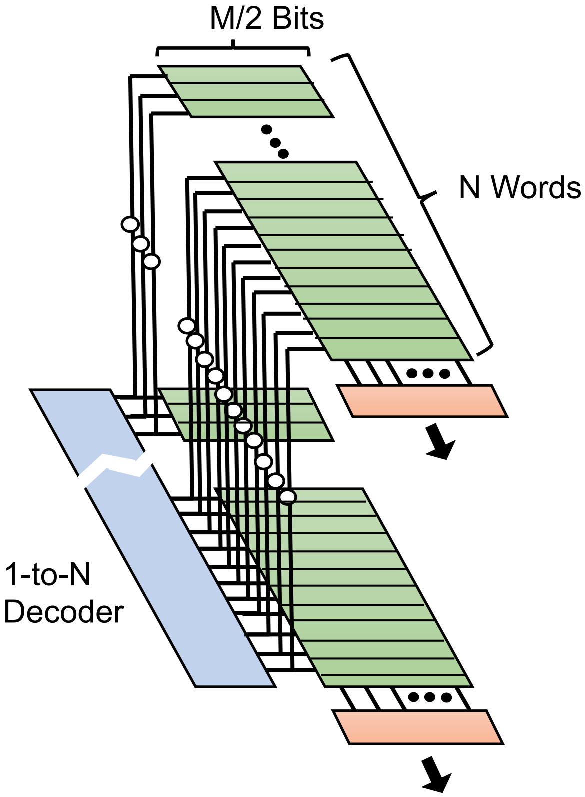
Partitioning Storage Stages in M3D

- 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

[1] K. Puttaswamy and G. H. Loh. 2009. 3D-Integrated SRAM Components for High-Performance Microprocessors. IEEE Trans. Comput. (Oct 2009).

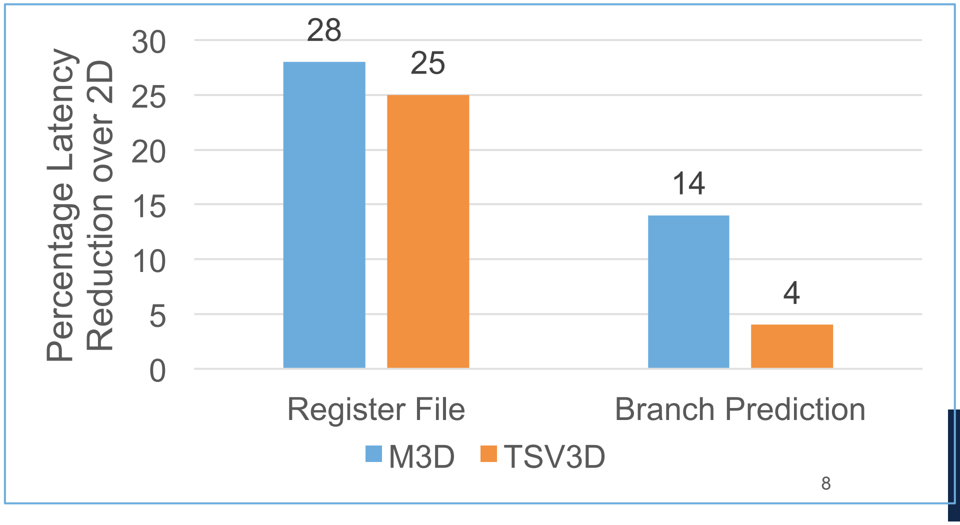
Bit Partitioning



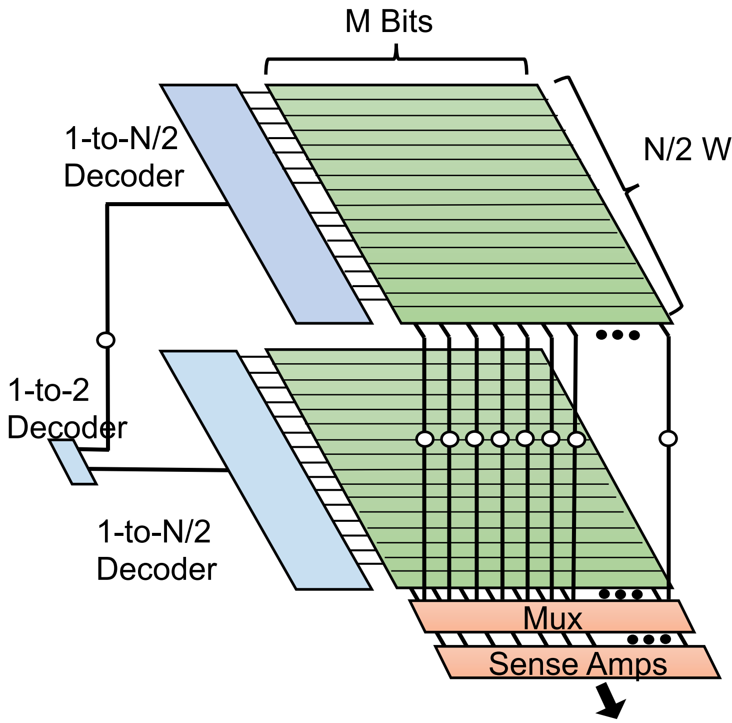
- Distribute half the bits of every word in each layer
- Each word requires an inter-layer via

- Multiported structure (e.g. Register File):
  - Bigger bitcells → higher gains

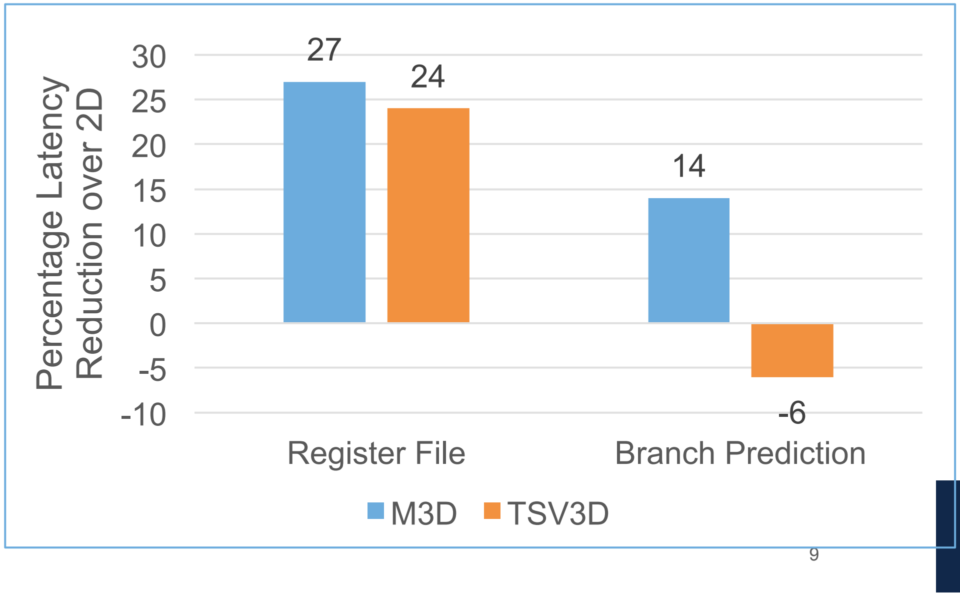
- Singleported structure (e.g. Branch Predictor):
  - Smaller bitcells → lower gains



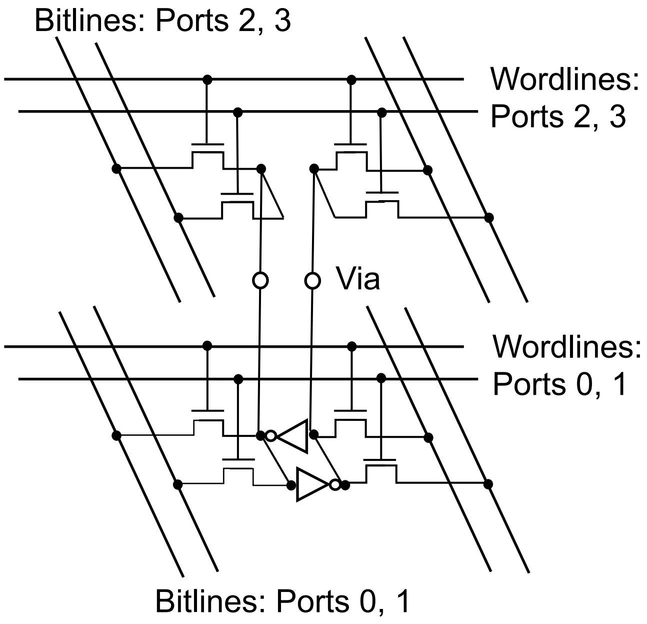
Word Partitioning



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



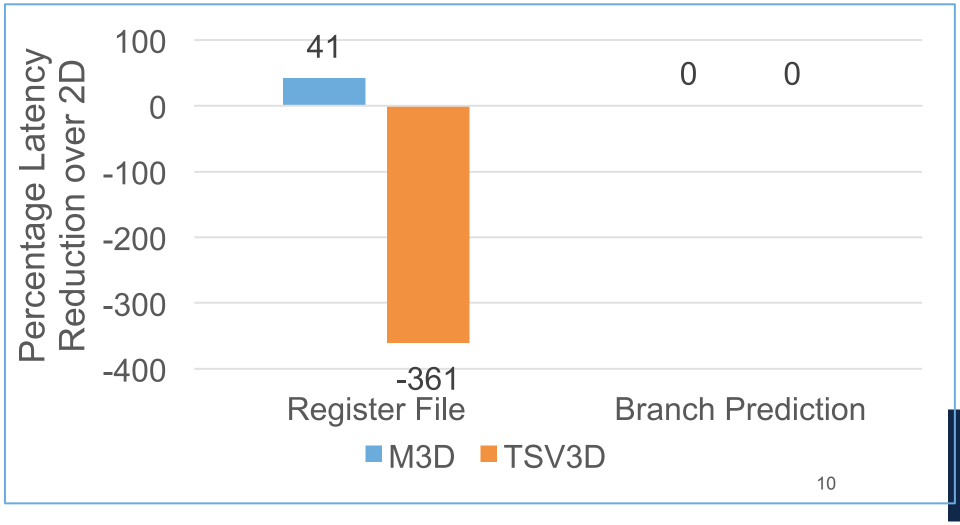
Port Partitioning



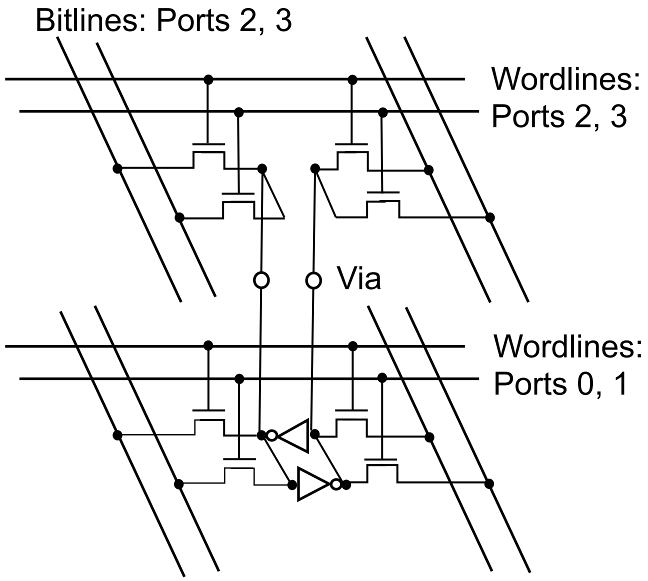
- Distribute half the ports in each layer
- Quadratic reduction in area footprint

- Two vias per bitcell
- Applicable only for multiported structures

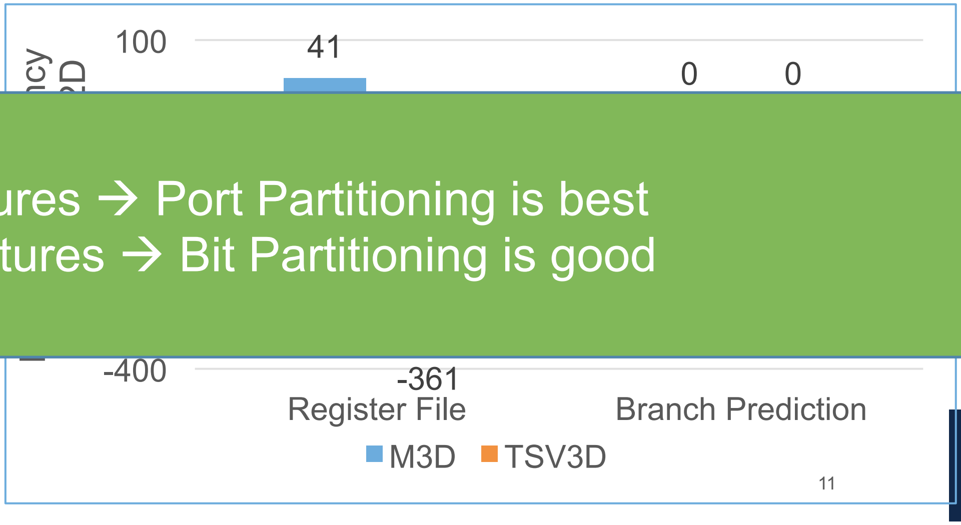
- TSVs simply have too much area overhead



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 with in a core such as load-to-use, branch misprediction, ALU+Bypass are shorter
  - Improves both core frequency and IPC

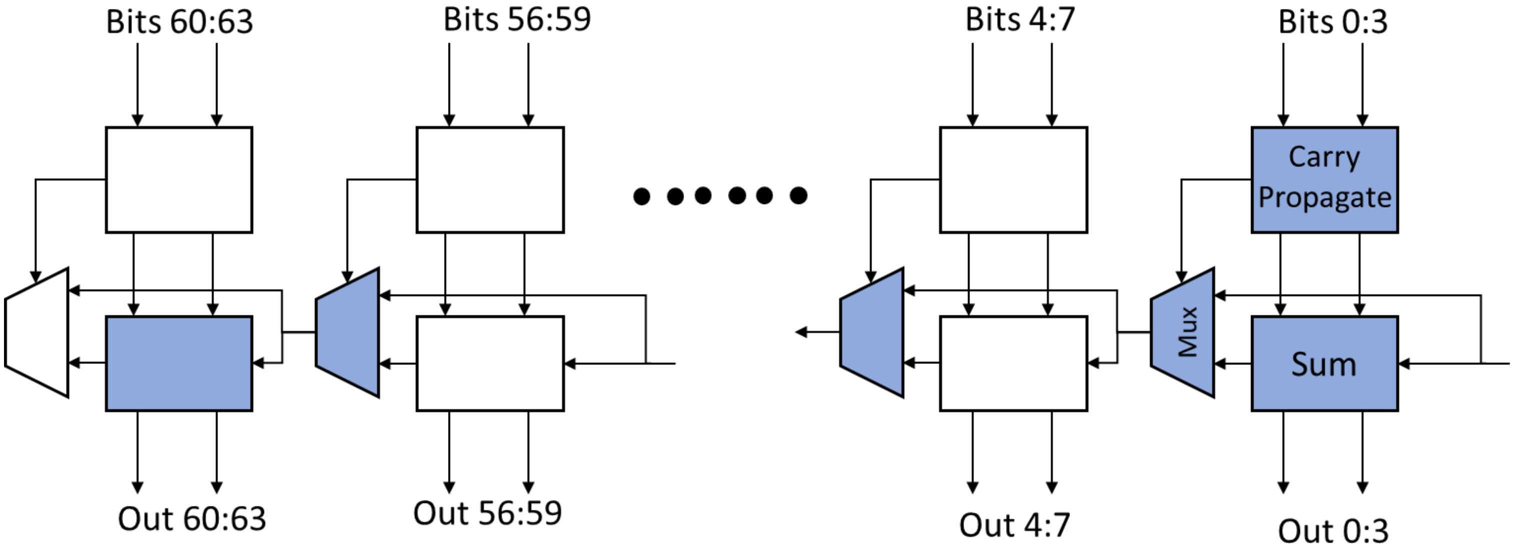
Hetero Layer Partitioning (M3D-Hetero)

- Low temperature processing → 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

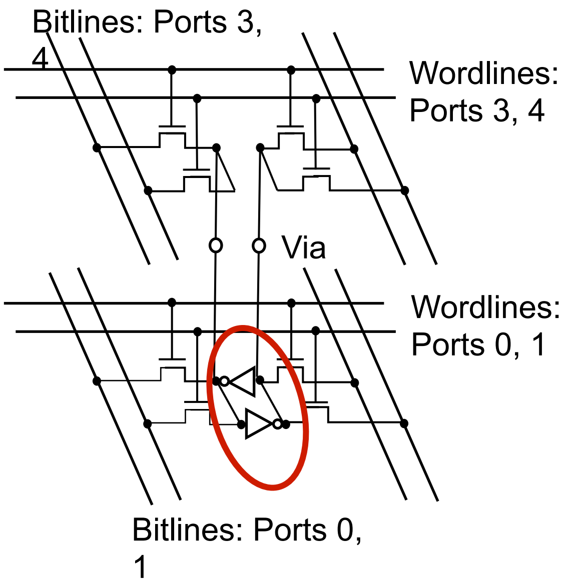
M3D-Hetero Logic Stages: Execution Unit

- 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



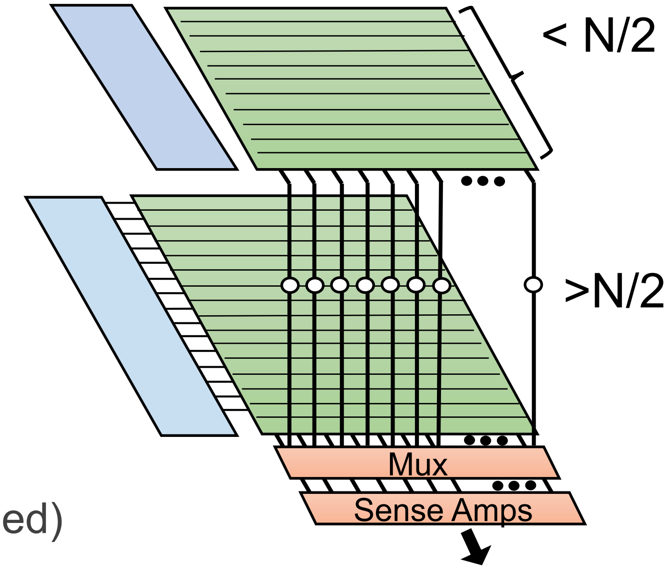
M3D-Hetero Storage Structures

- 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

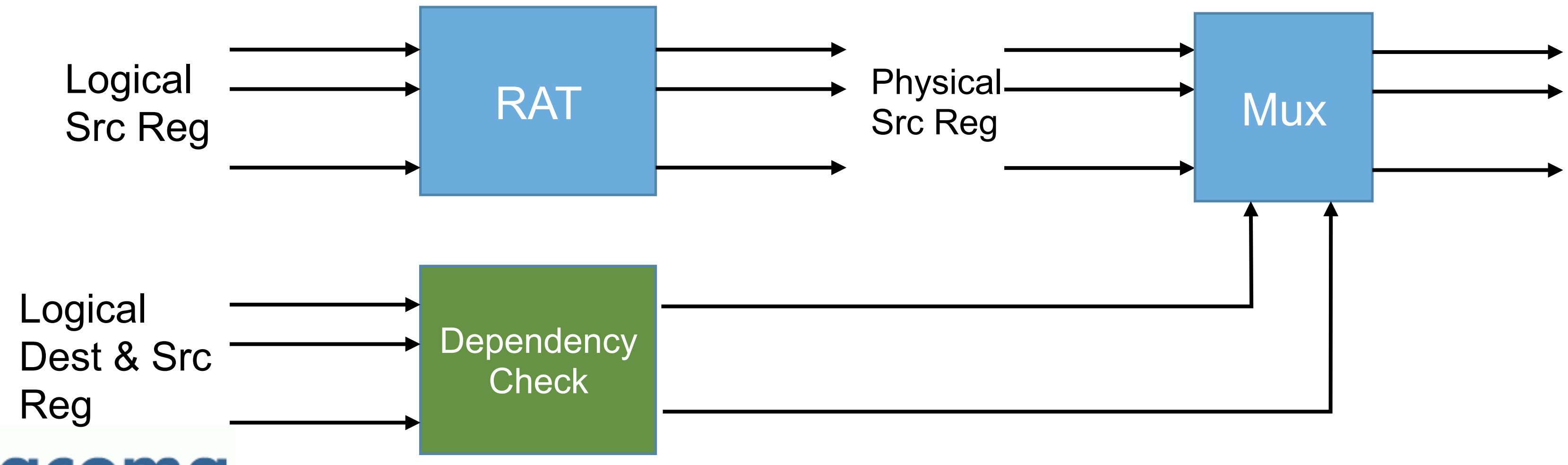
- Port Partitioned Structures:
  - Area slack exists in top layer, as the bitcell is only in the bottom layer
  - Propose: *Asymmetric Port Partitioning*
    - Increase the size of access transistors in the top layer (for speed)
    - Place more ports in bottom layer than in top layer



- 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

Stage Type	Mitigation Technique	
Logic Stages	Critical paths in bottom layer; non-critical paths in top	
Storage Stages	Port Partitioning	Asymmetric partitioning of ports; and larger access transistors in top layer
	Bit/Word Partitioning	Asymmetric partitioning of array; and larger bitcells in top layer
Mixed Stages	Combination of above techniques	

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

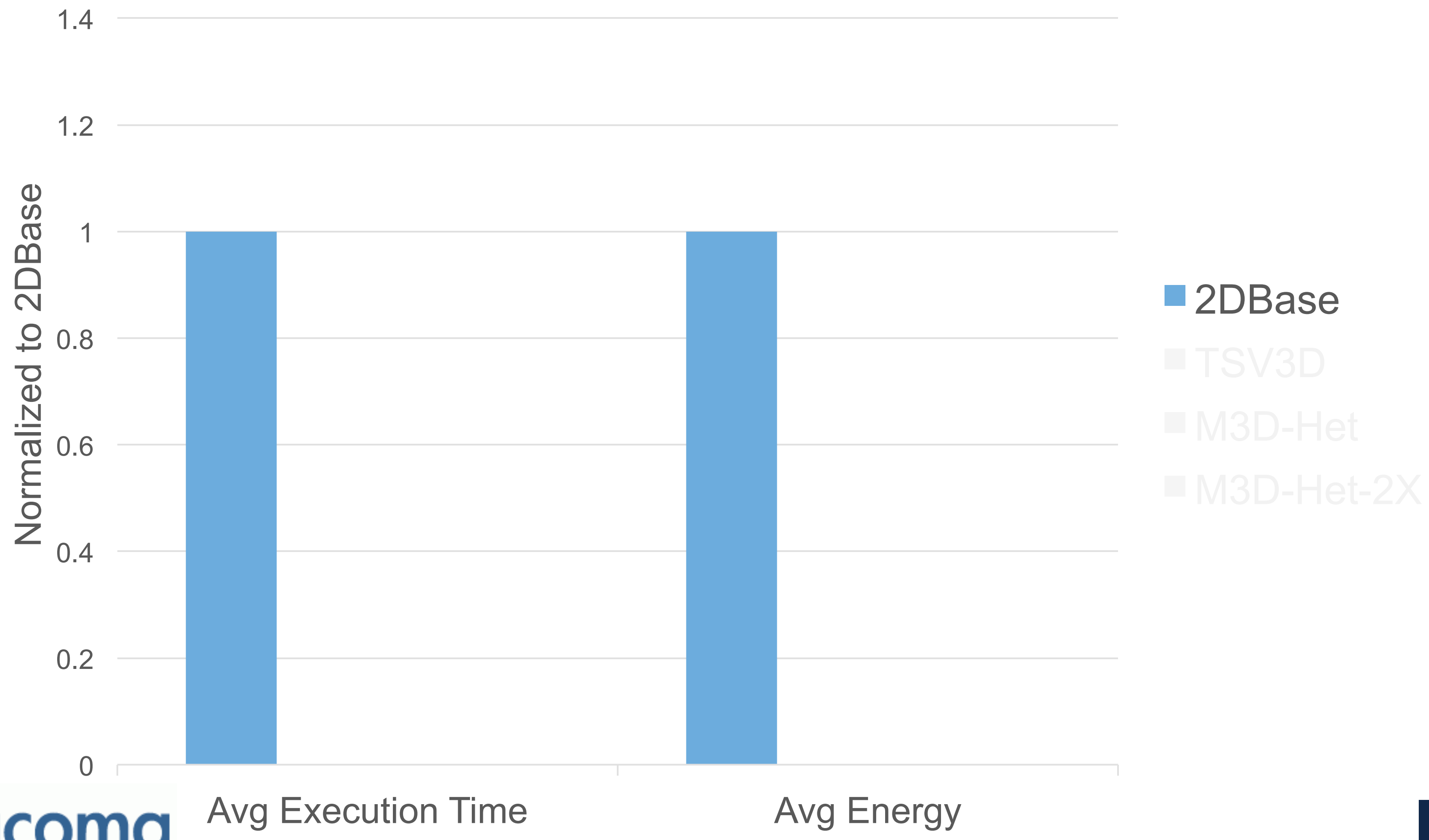
- ✓ Keep the same frequency, but reduce the voltage
  - Operate more cores within the same power budget

Evaluation: Configurations & Frequency

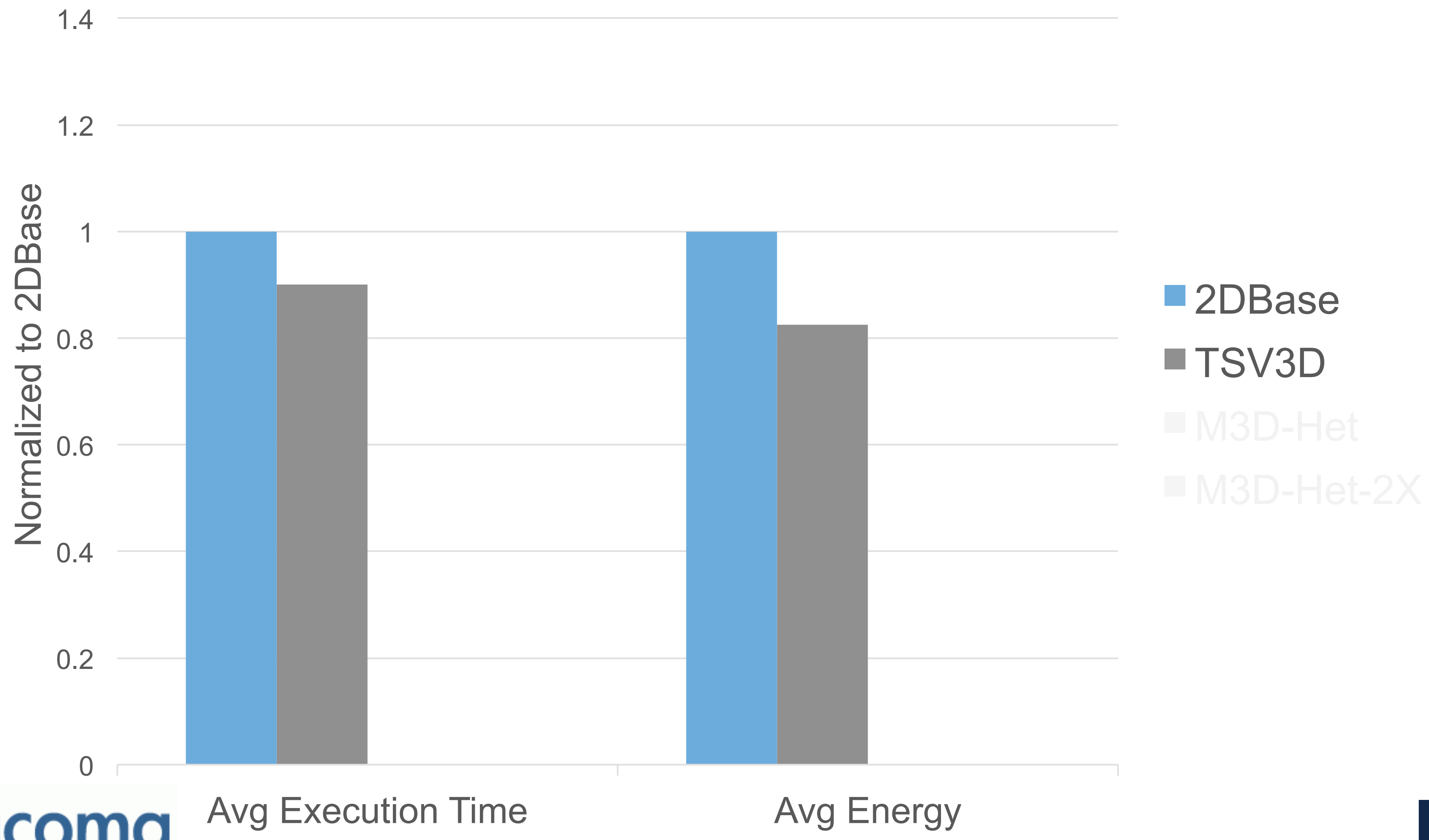
- *2D Base:* Four 6-wide OOO cores
  - Caches: DL1 & IL1 32KB ; L2 256 KB ; Entries: RF(160) ; ROB(192) ; IQ(84) ; LQ(72); SQ(56)
  - Frequency = 3.3 GHz
- *M3D-Het* : M3D with top layer slowdown and our modifications
  - Stage with lowest delay reduction: SQ/BPT at 13%
  - Frequency = 3.79 GHz
- *TSV3D:* 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

i-acoma group

Evaluation of M3D Design Choices

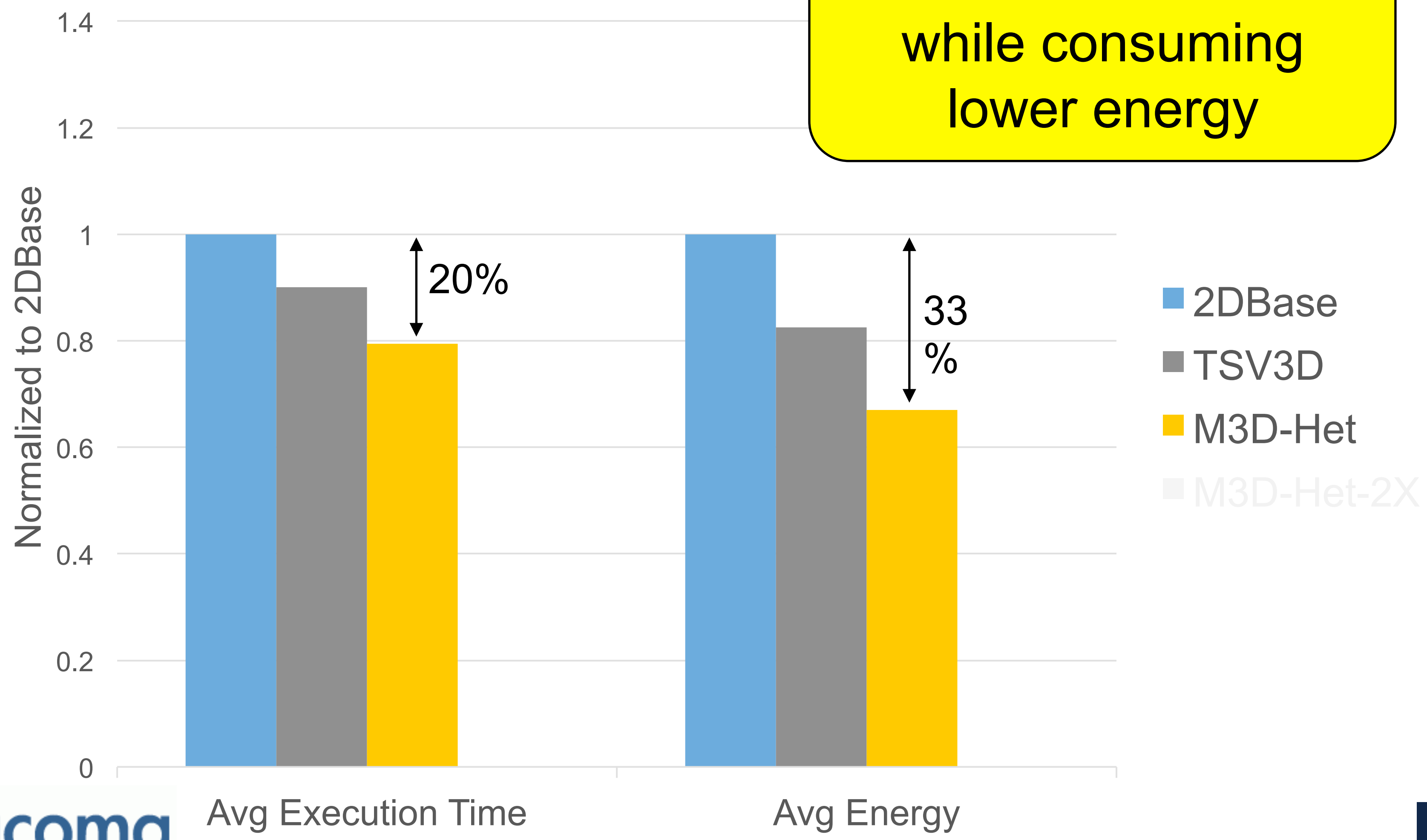


Evaluation of M3D Design Choices



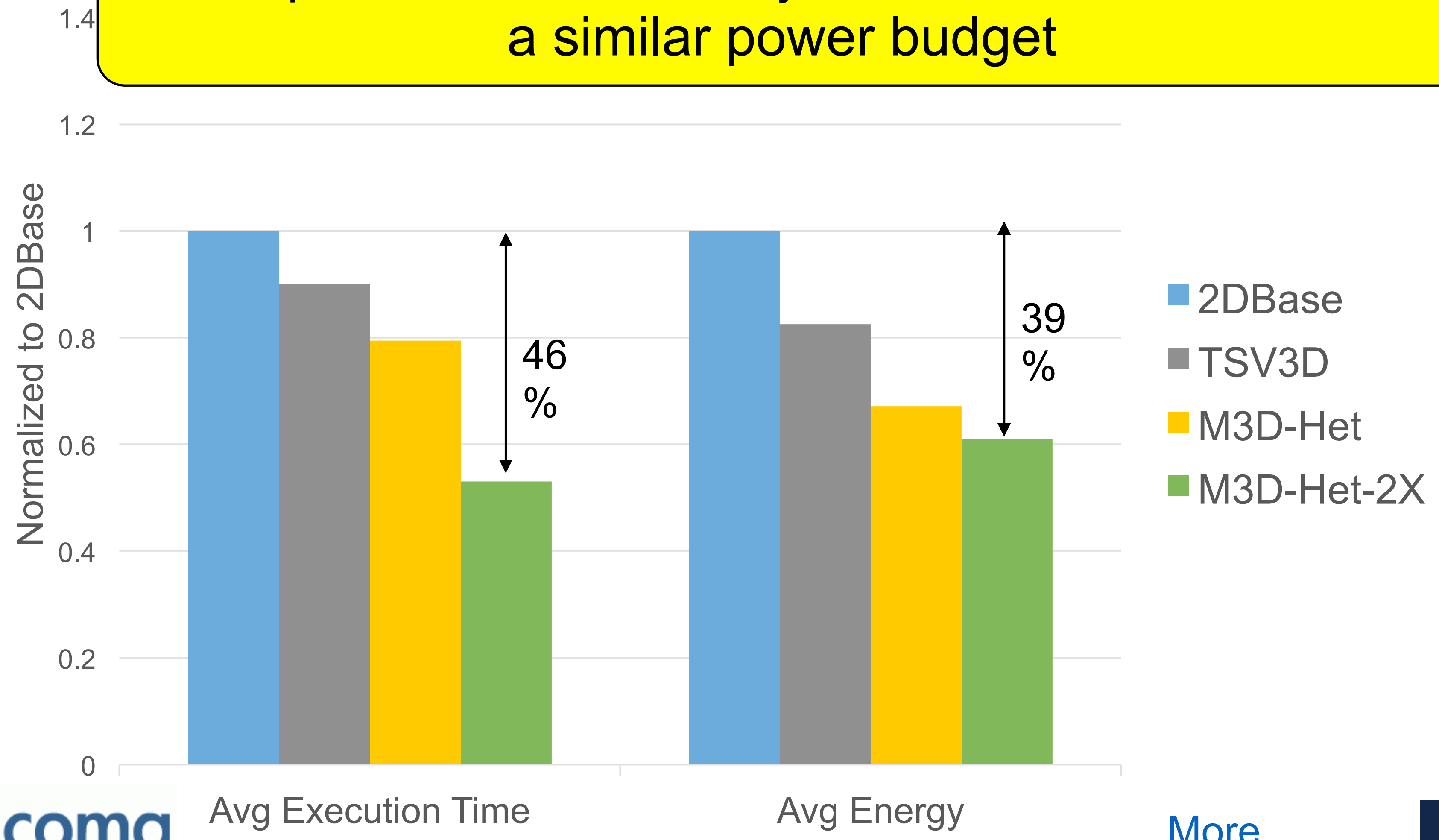
Evaluation of M3D Design Choices

M3D executes faster while consuming lower energy



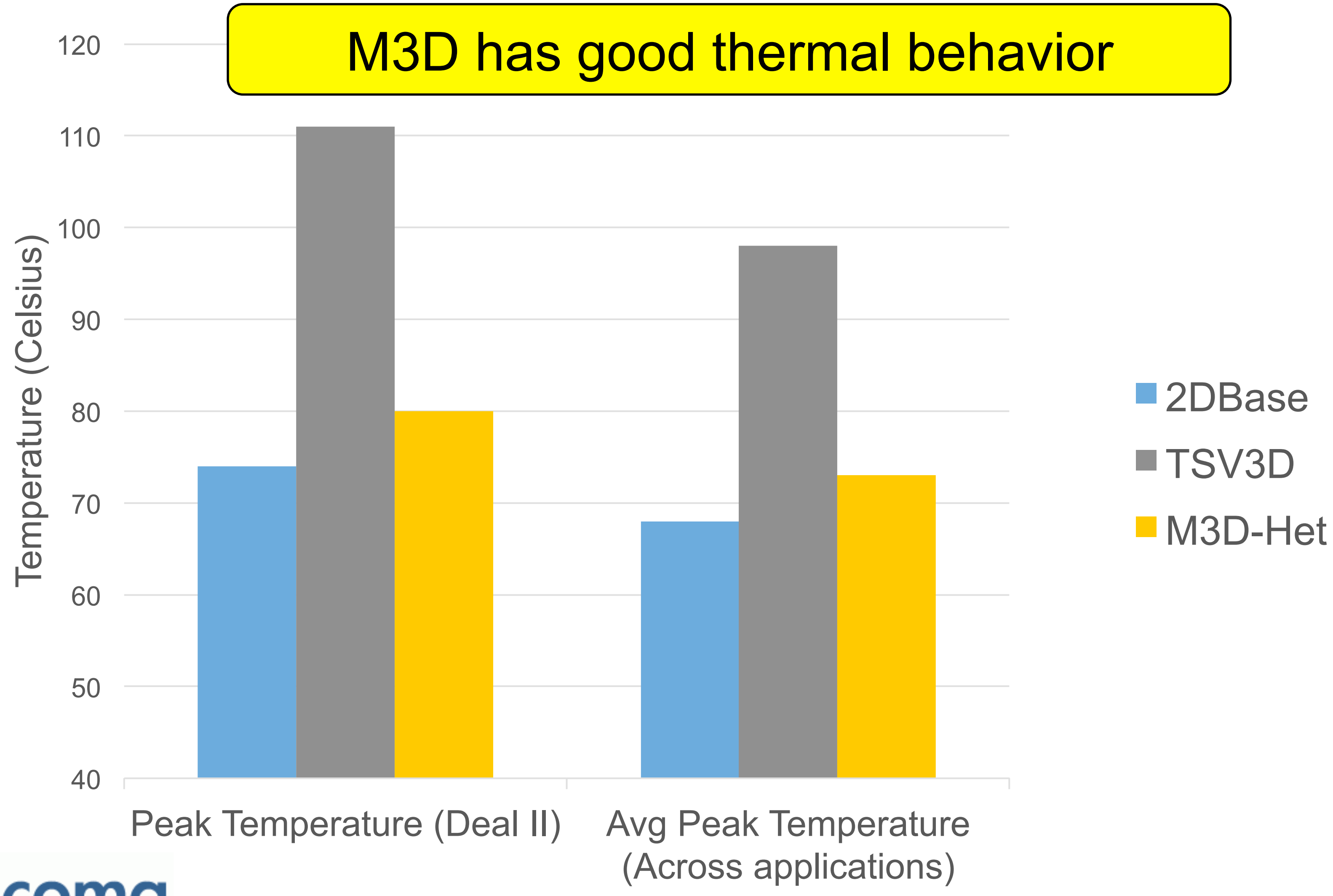
Evaluation of M3D Design Choices

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



More Results

Thermal Behavior: Peak Temperature



M3D has good thermal behavior

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

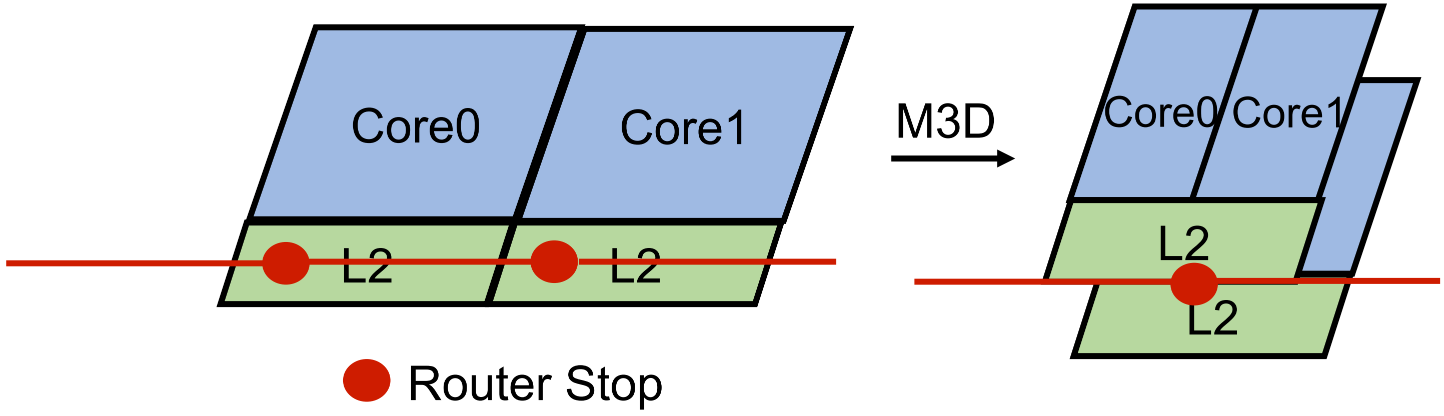
http://iacoma.cs.uiuc.edu

ISCA 2019

Backup

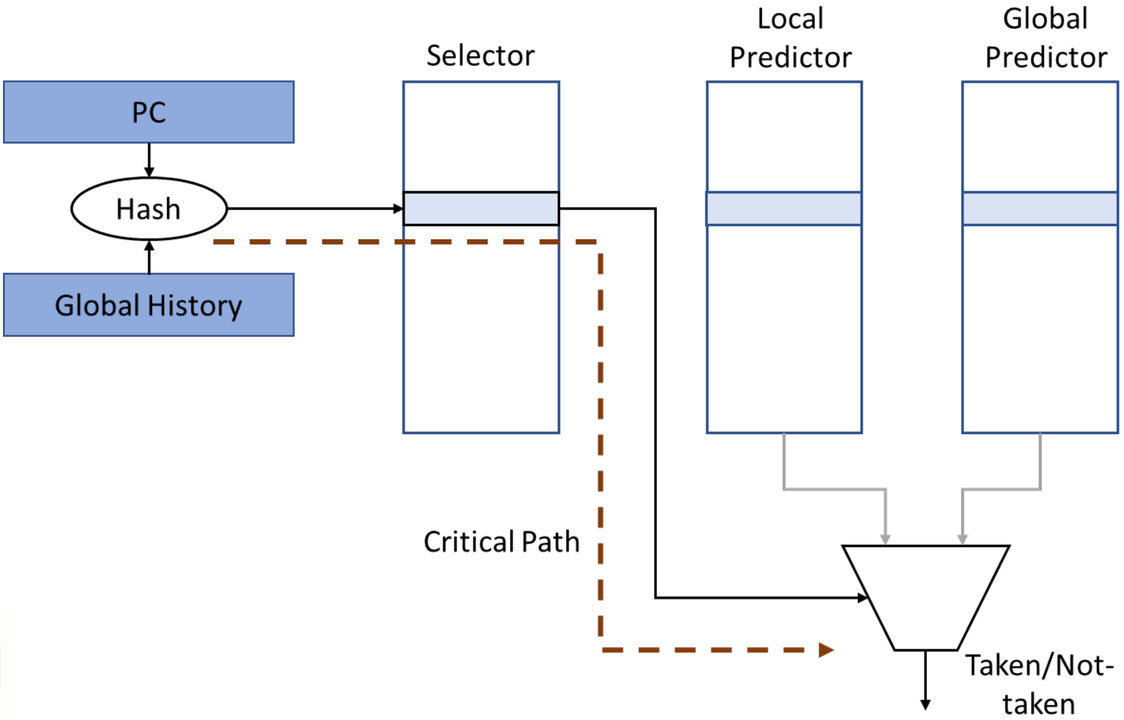
Caches and NoC

- Shared L2 Cache, and NoC Router Stop across two adjacent cores



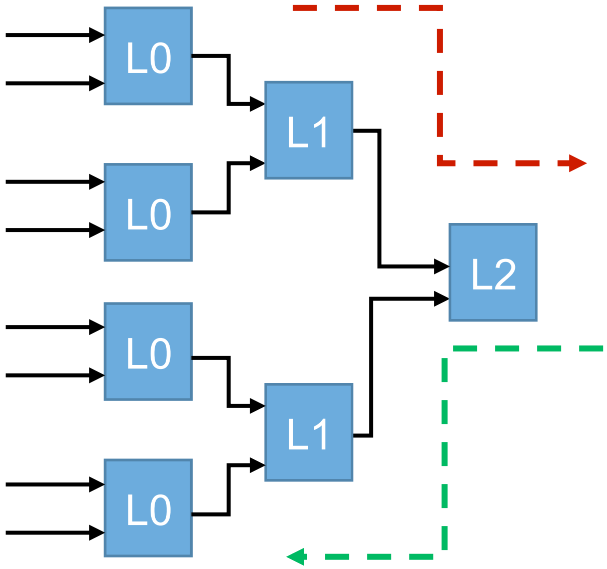
Stages with Logic and SRAM Structures

- 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 → Asymmetric port partitioning
  - Select consists of multi-level arbitration
  - Within each arbiter:
    - First phase: Request → critical and is in bottom layer
    - Second phase: Grant → has slack and is in top layer.



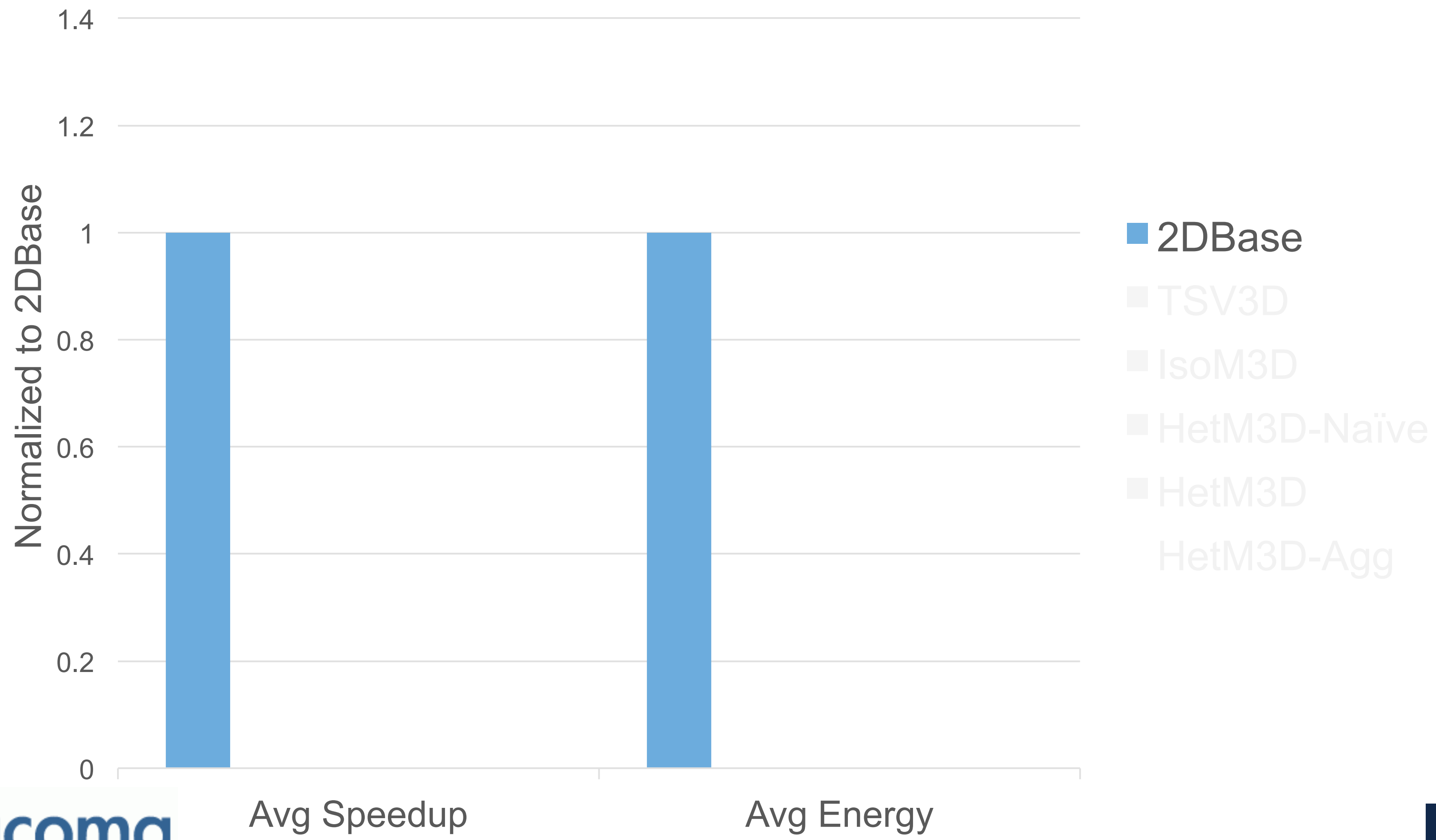
Stages with Logic and CAM Structures

- 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 → Asymmetric Port partitioning
  - Priority encoder after SQ is placed in bottom layer.
  - Balance area between LQ/SQ across layers.

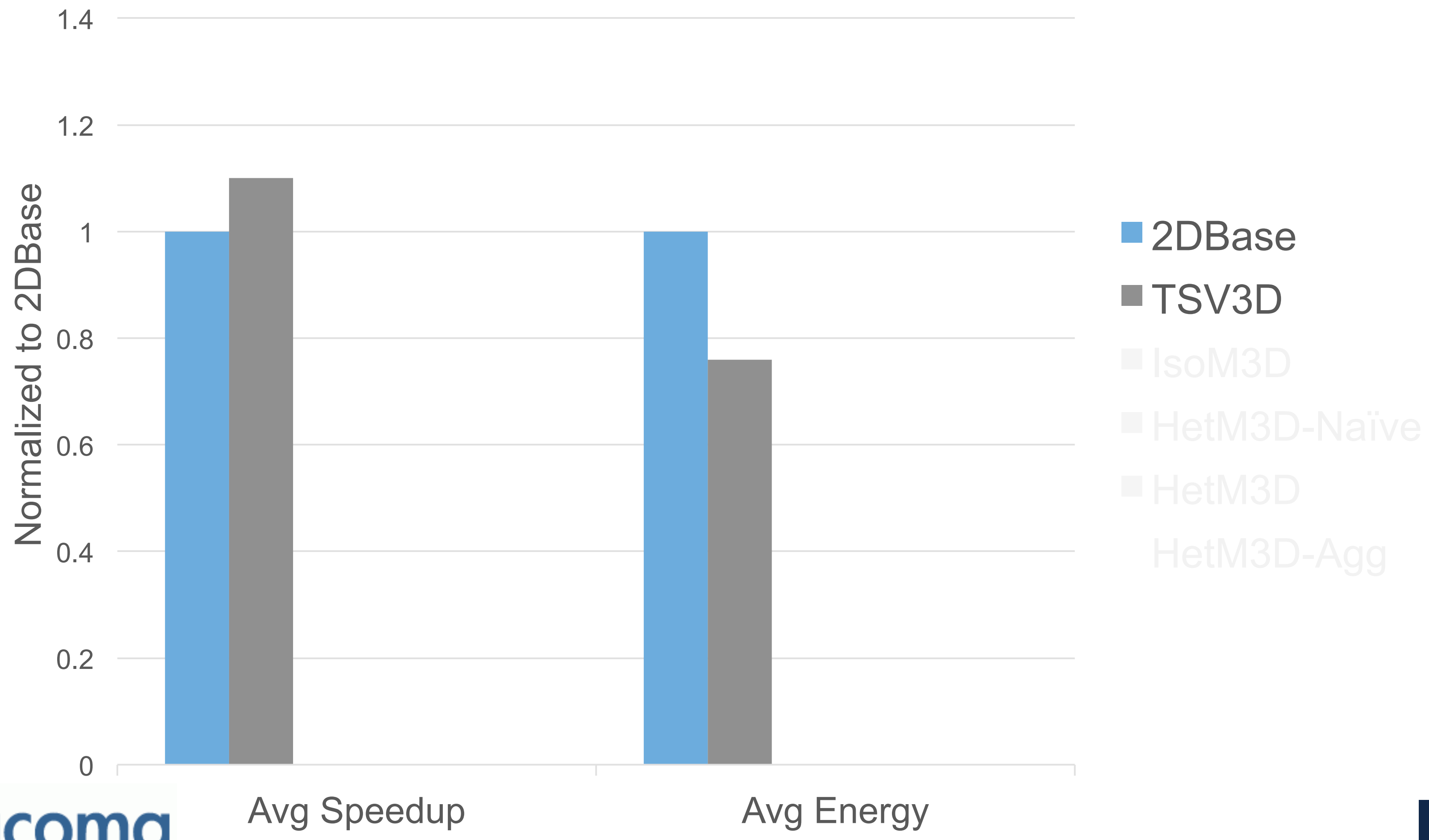
Evaluation: Frequency of Operation for M3D

- *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

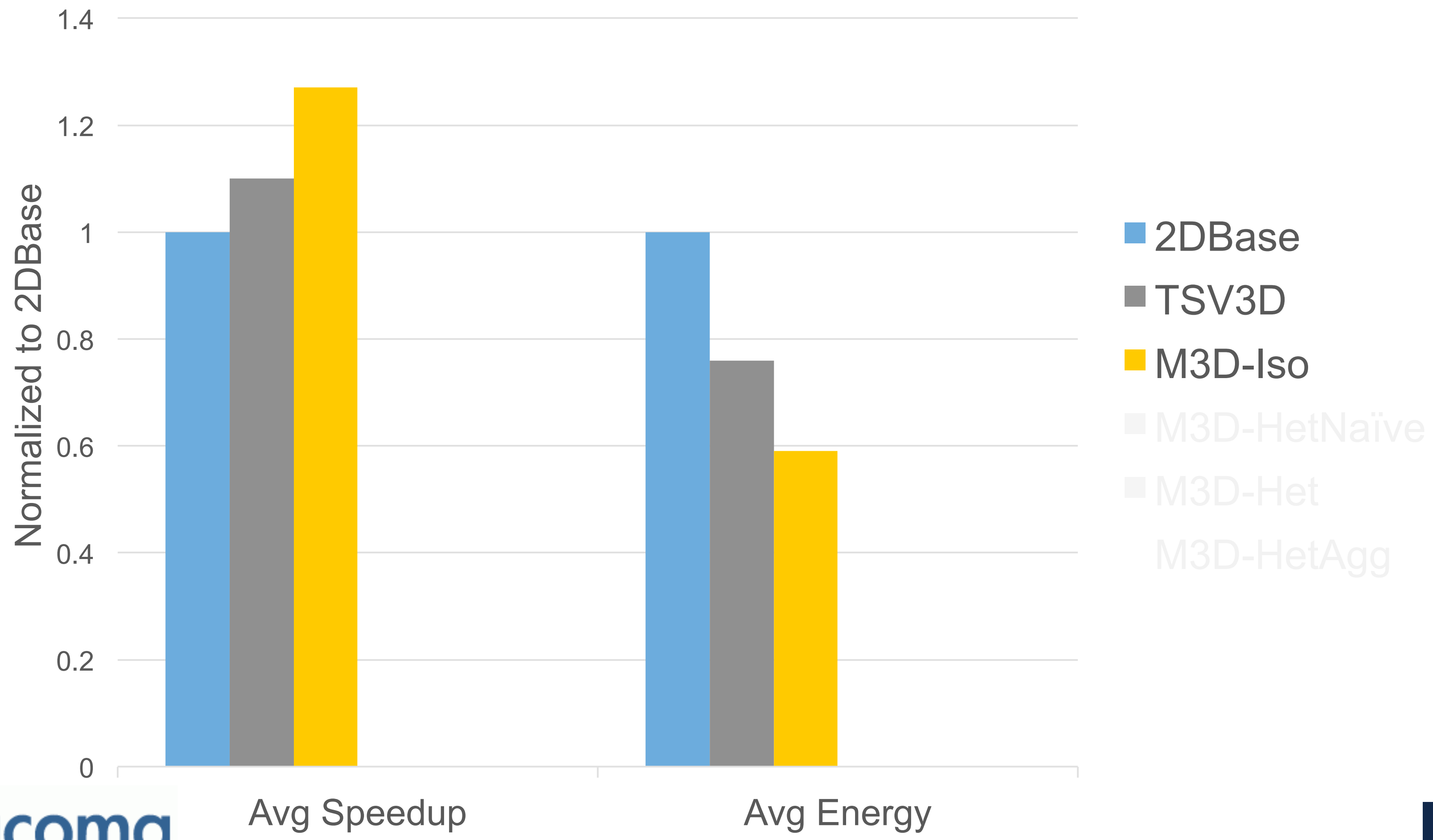
Single Thread Results: SPEC Benchmark Suite



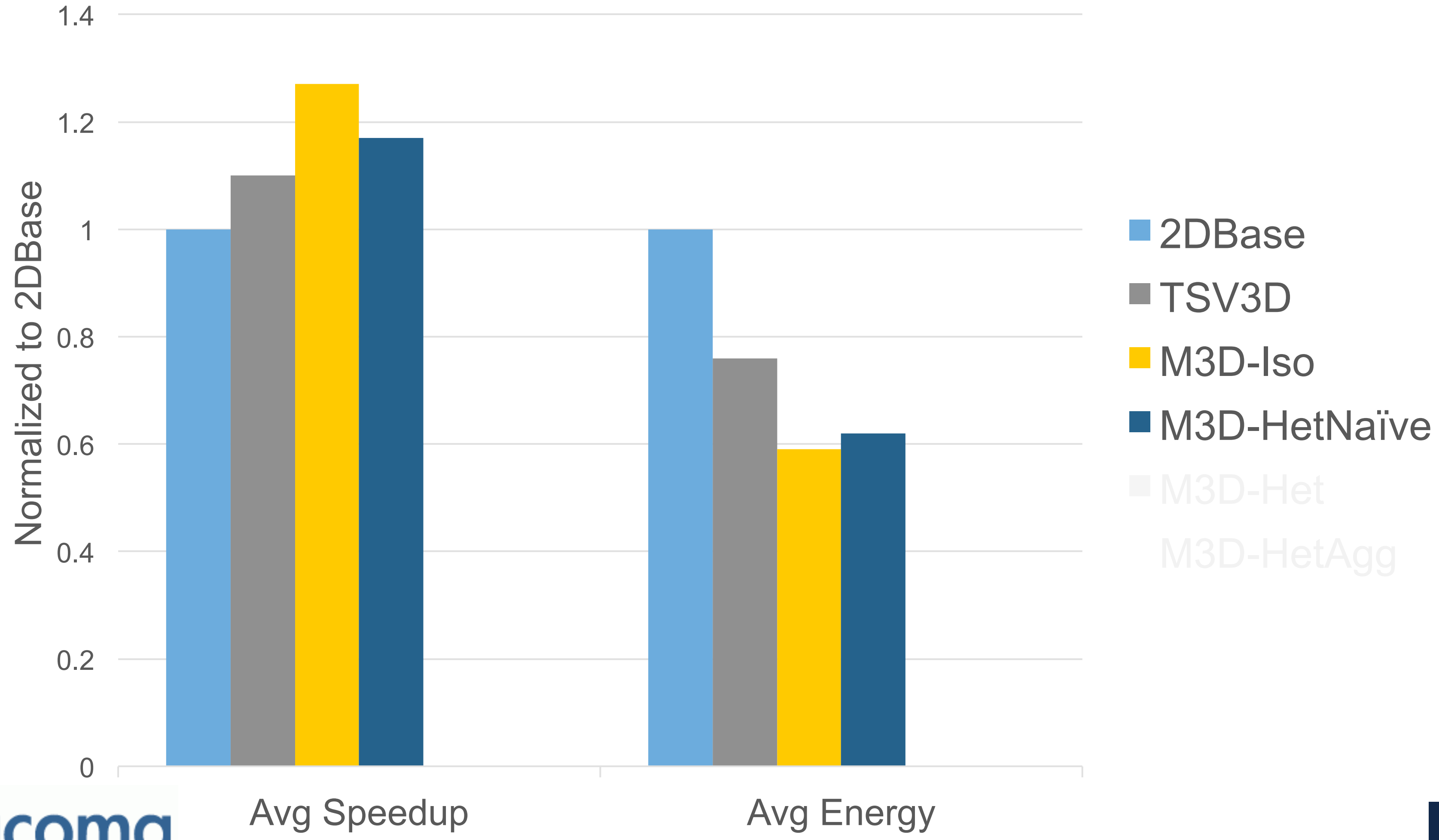
Single Thread Results: SPEC Benchmark Suite



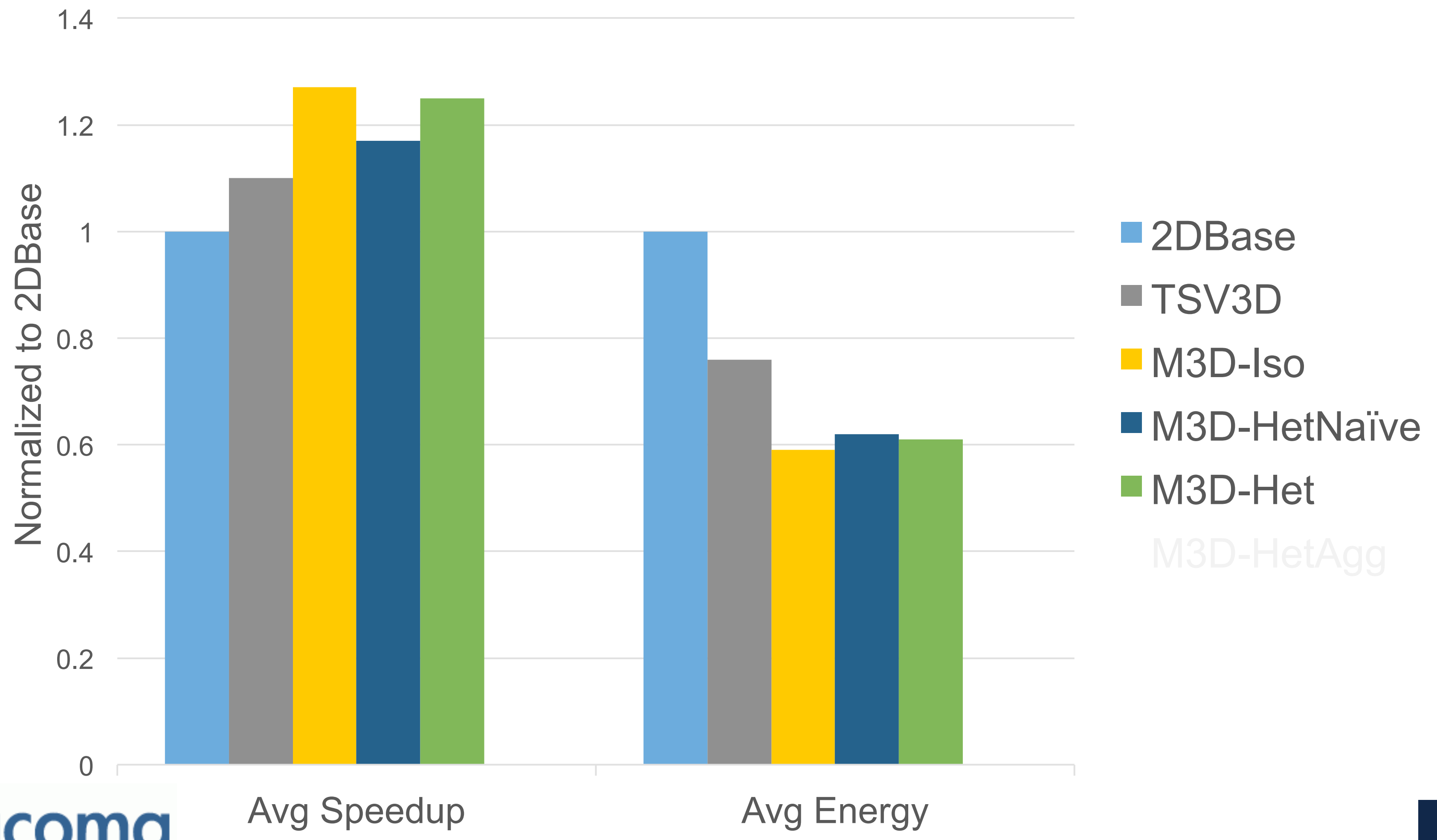
Single Thread Results: SPEC Benchmark Suite



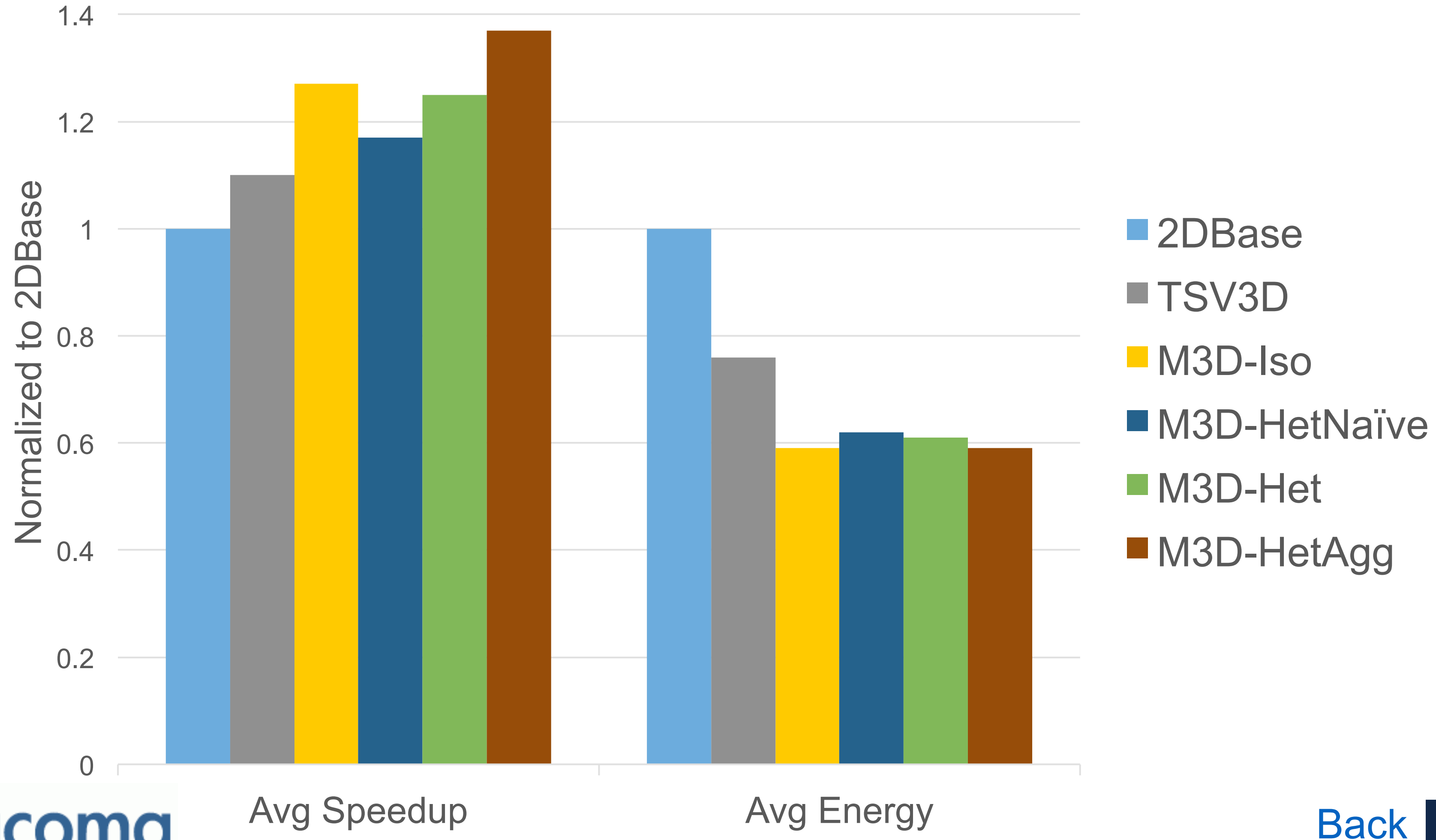
Single Thread Results: SPEC Benchmark Suite



Single Thread Results: SPEC Benchmark Suite

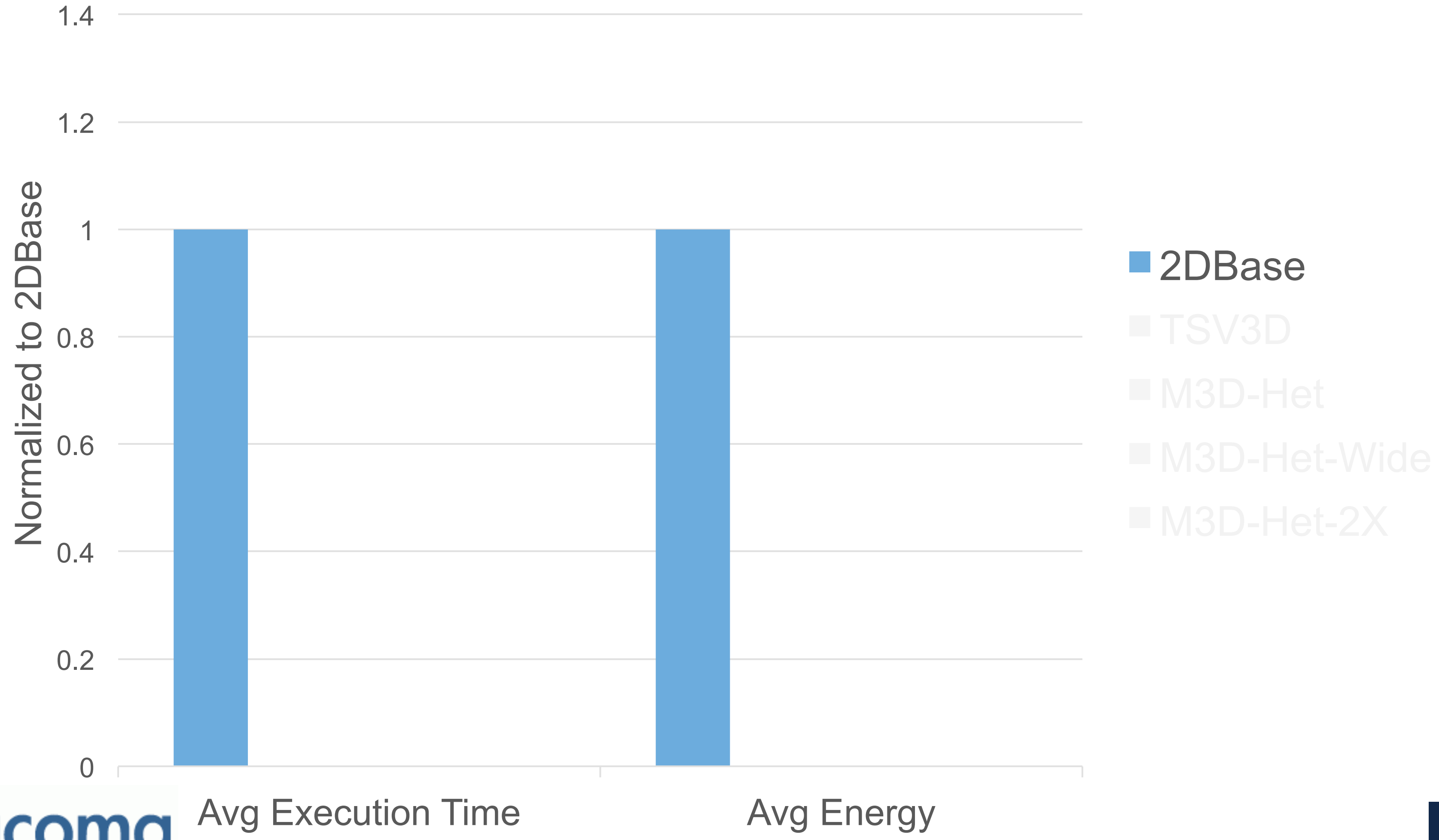


Single Thread Results: SPEC Benchmark Suite

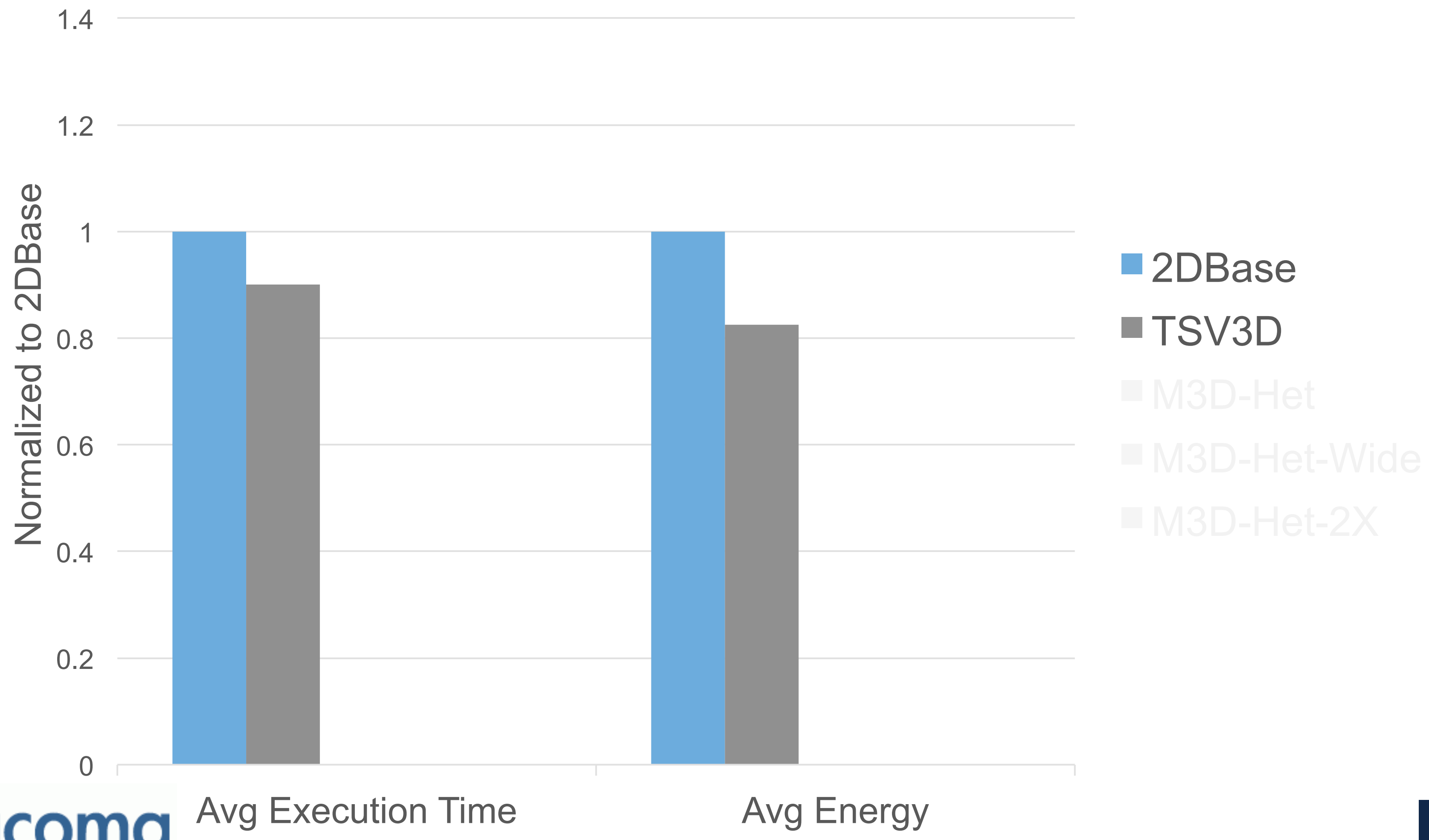


Back

Evaluation of M3D Design Choices

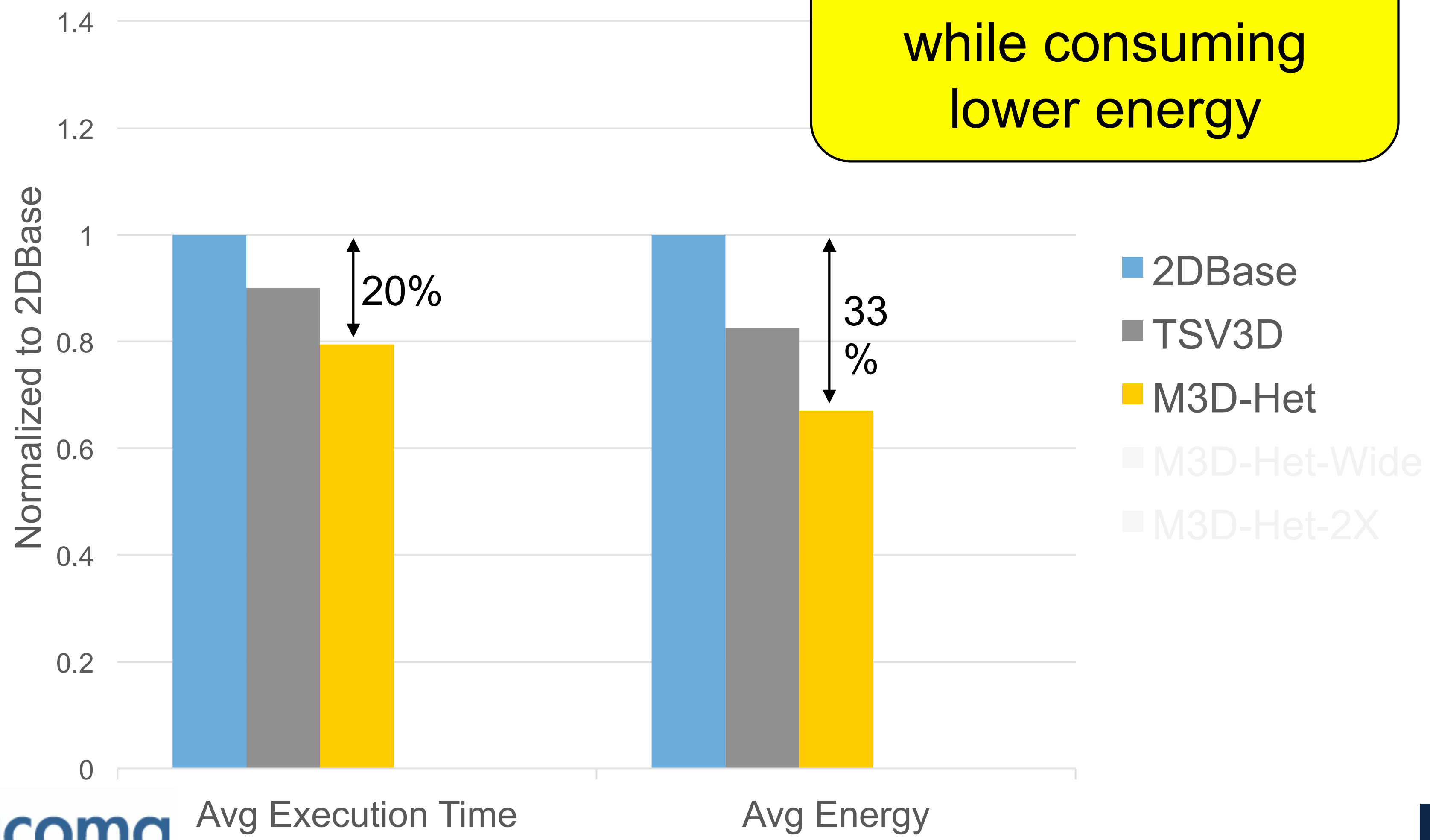


Evaluation of M3D Design Choices

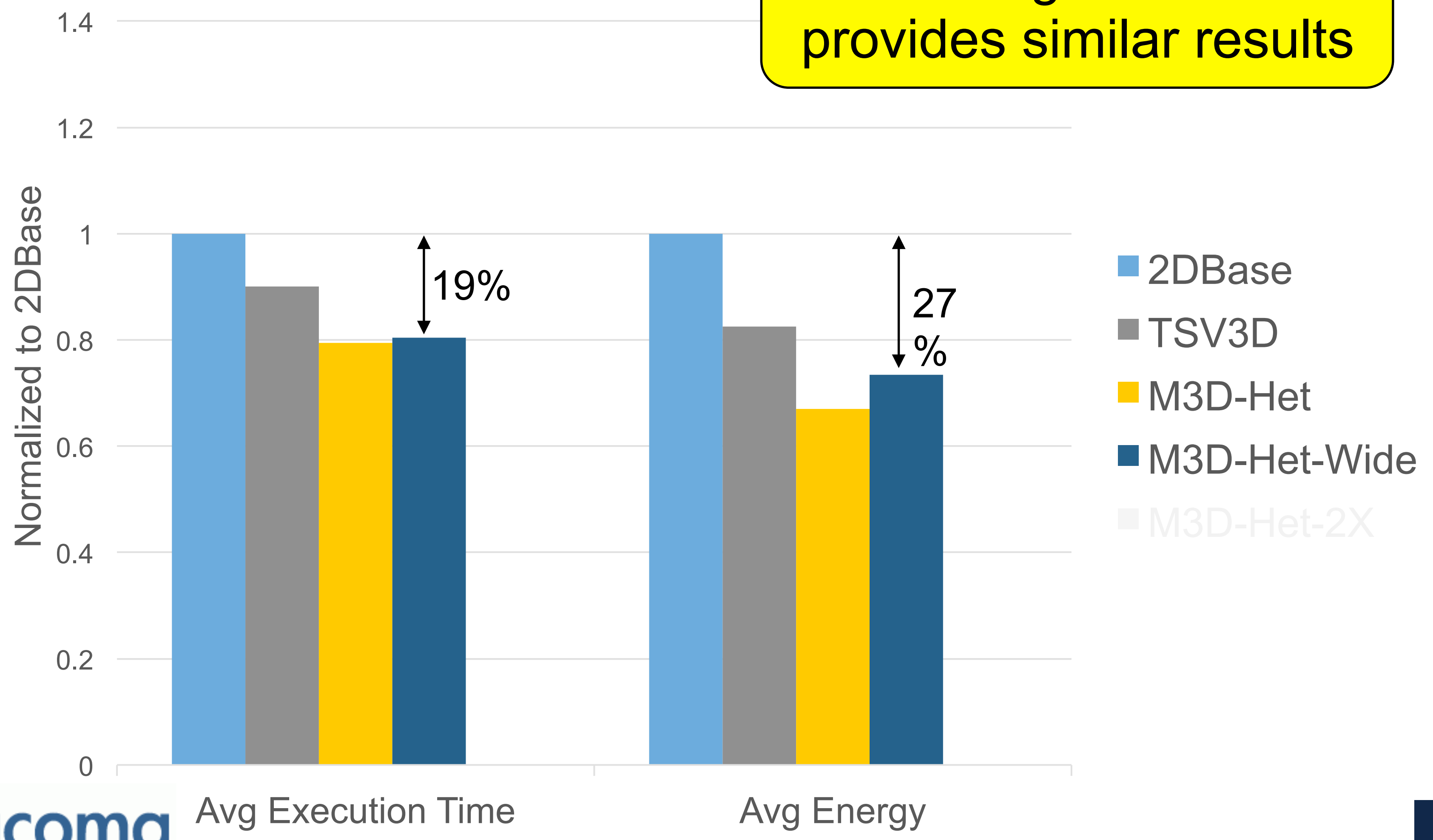


Evaluation of M3D Design Choices

M3D executes faster while consuming lower energy

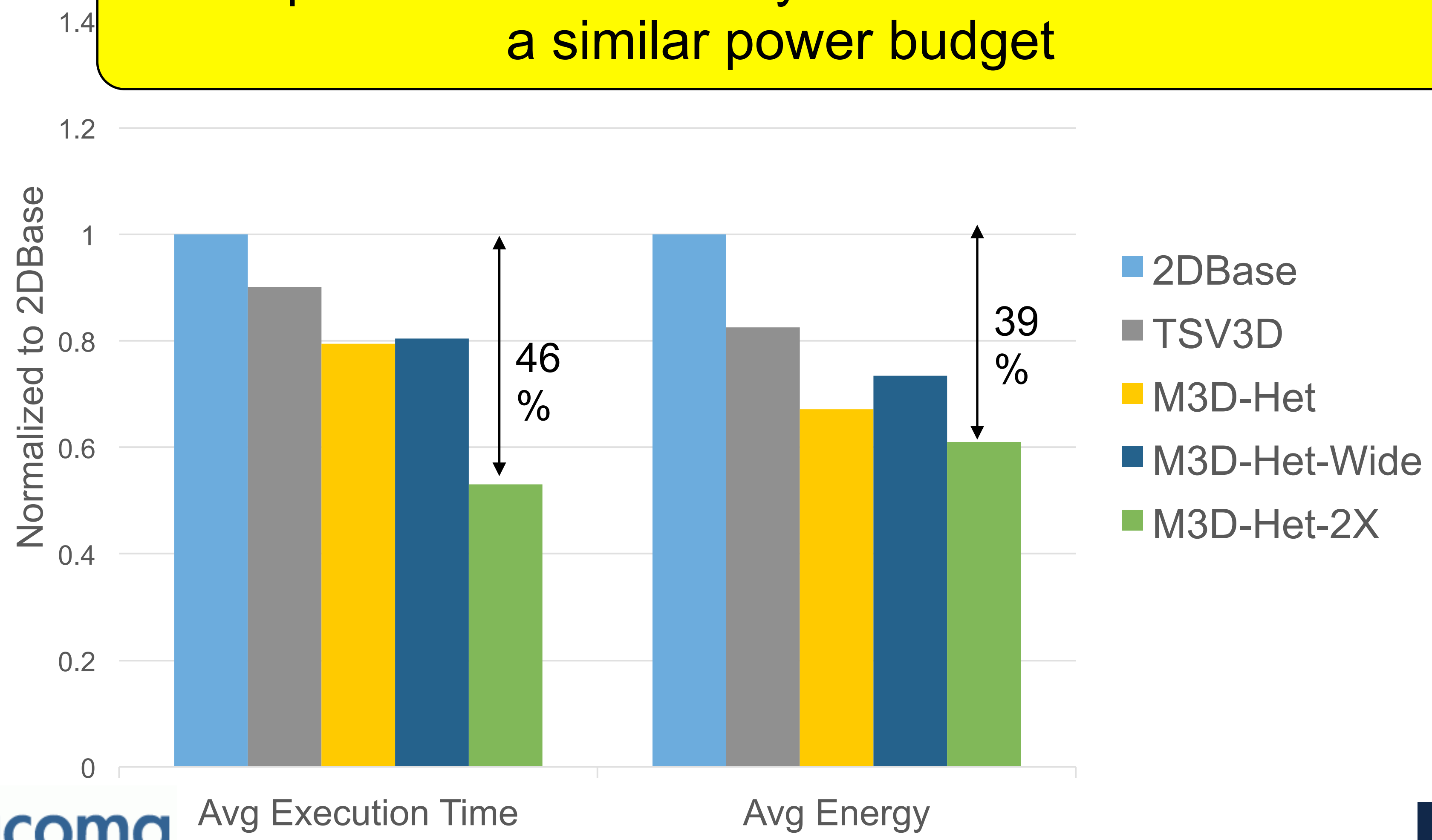


Evaluation of M3D Design Choices



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

Monolithic 3D Core -- Partitioning Storage Stages

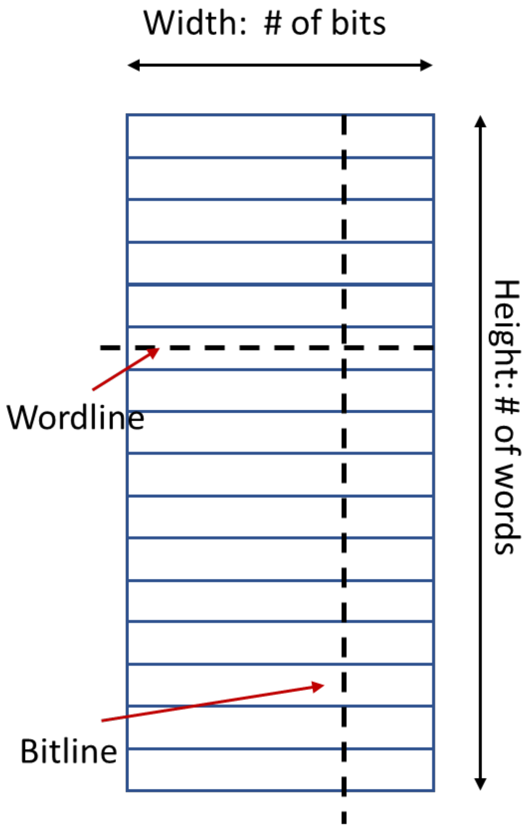
- Storage structures i.e. SRAM/CAM delays are proportional to height and width of an array

$$Width = N_{bits} * (BitcellWidth + K * (N_{r} + N_{w}))$$

$$Height = N_{words} * (BitcellHeight + K * (N_{r} + N_{w}))$$

$$Area = Width * Height \propto (N_{r} + N_{w})^{2}$$

Area of multiported cell is quadratic on the number of ports


Width: # of bits
Height: # of words
Wordline
Bitline