Exploiting the Advantages of 3D Integration: A Benefit and Limit Study

Students: Mitchelle Rasquinha, Kwanyeob Chae, Minki Cho,,
Syed Minhaj Hassan, William Song,

PI’s: Saibal Mukhopadhyay, Sudhakar Yalamanchili

Introduction

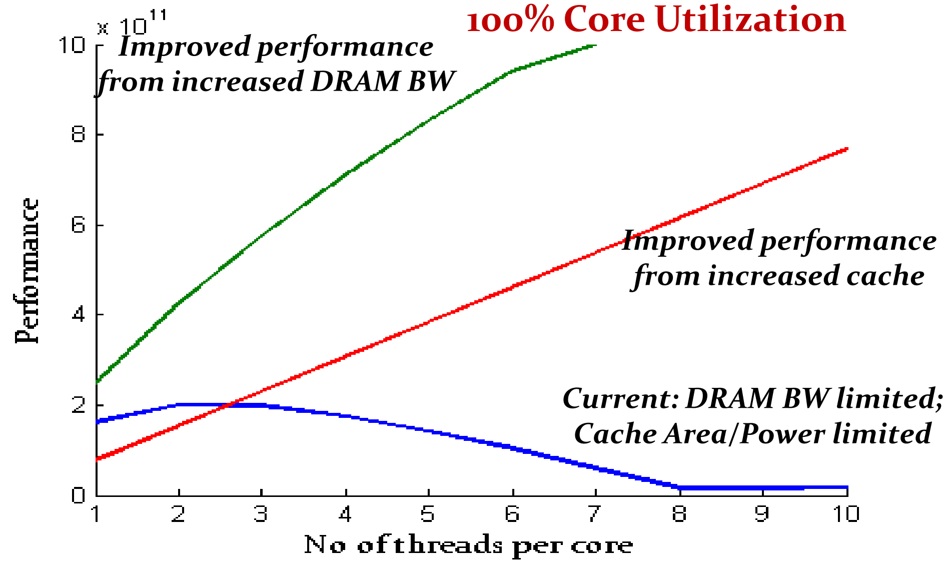
- 3D system integration is a key enabling technology
  - Abundance of on-chip bandwidth
- Understand and quantify the impact of 3D bandwidth on multi-core performance and power scaling
  - How can bandwidth better modulate the trade-offs between parallelism, speed, and power?
- Identify specific system components that need to be redesigned.

$$Optimize\ for\ \frac{Performance}{Power}\ for\ a\ fixed\ area.$$

Performance Scaling

$$Performance = N_{core} \cdot f_{core} \cdot \eta \cdot IPC \qquad \eta = f(N_{tpc}, \$C, T_{mem\_ss})$$

$$\eta = \min\left(1, \frac{N_{TPC}}{1 + T_{mem\_ss} \cdot r \cdot IPC}\right) \qquad T_{mem\_ss} = t_{cache} + m\left[L \times (N_{tsv}, N_{core}, \beta) + t_{mem}\right]$$



100% Core Utilization

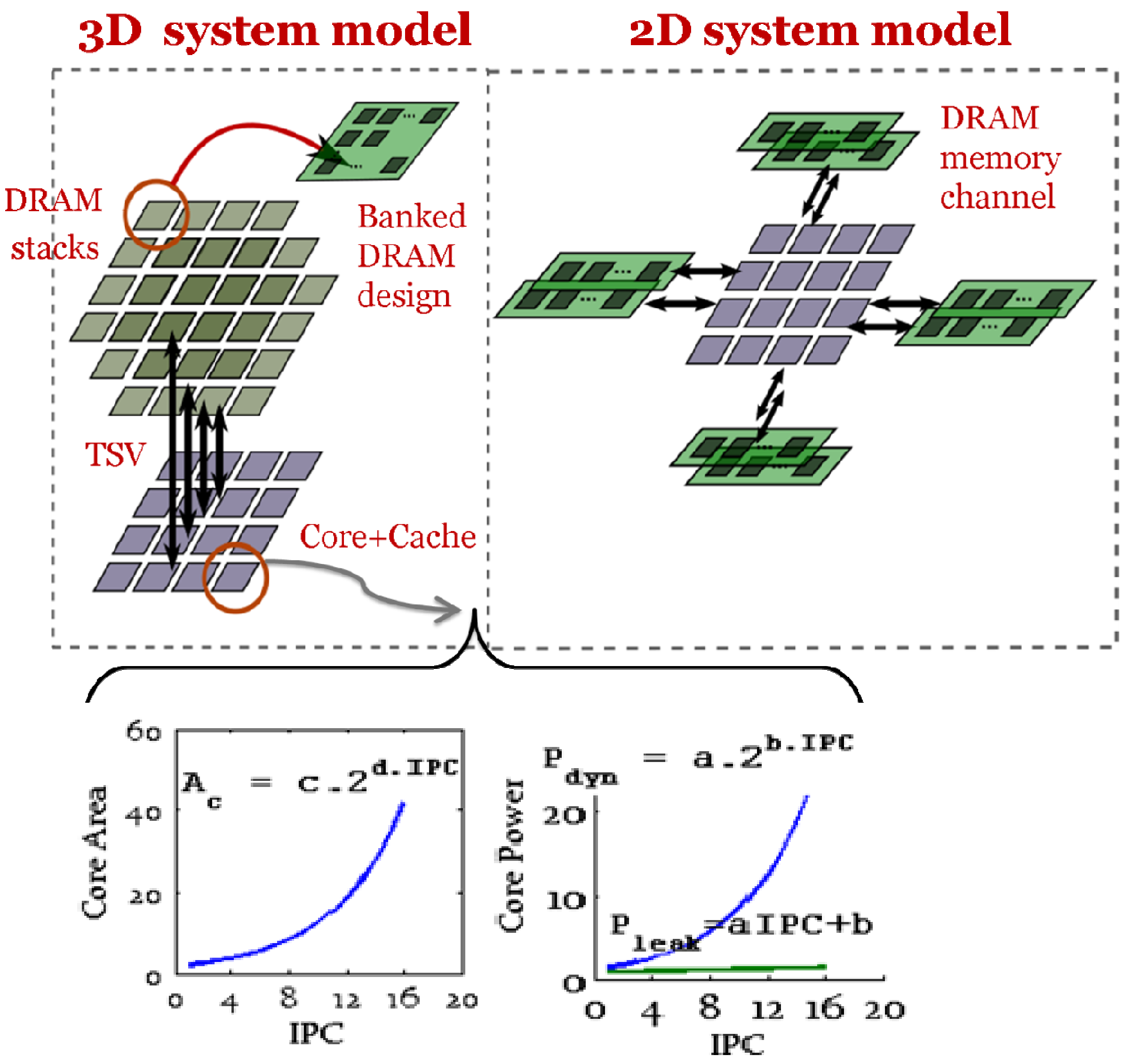
Improved performance from increased DRAM BW

Improved performance from increased cache

Current: DRAM BW limited; Cache Area/Power limited

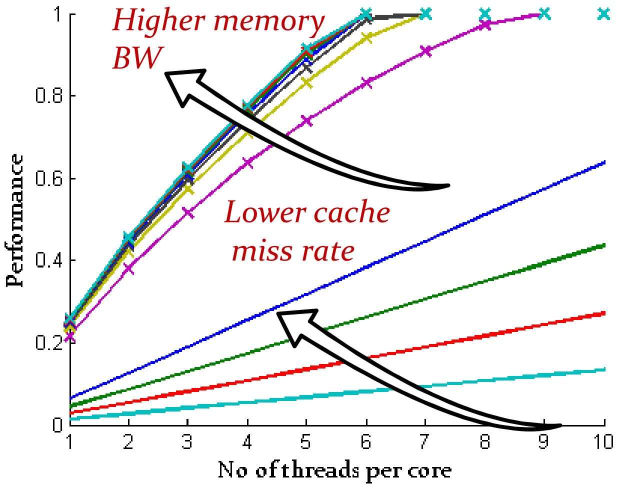
N_{core}=50; Technology=16nm; f_{core}=4GHz

System Model


3D system model

2D system model

Performance Scaling



Higher memory BW

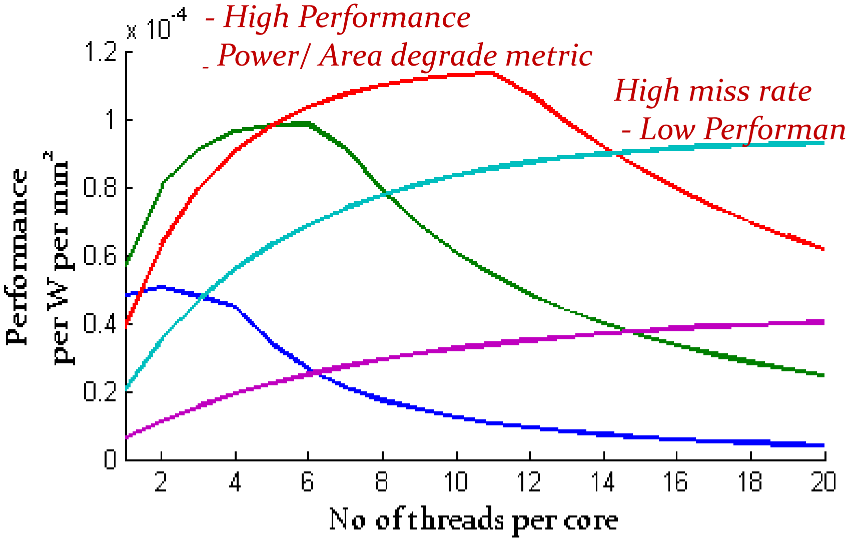
Lower cache miss rate

Higher core utilization via larger caches

$$T_{mem\_ss} = t_{cache} + mK$$

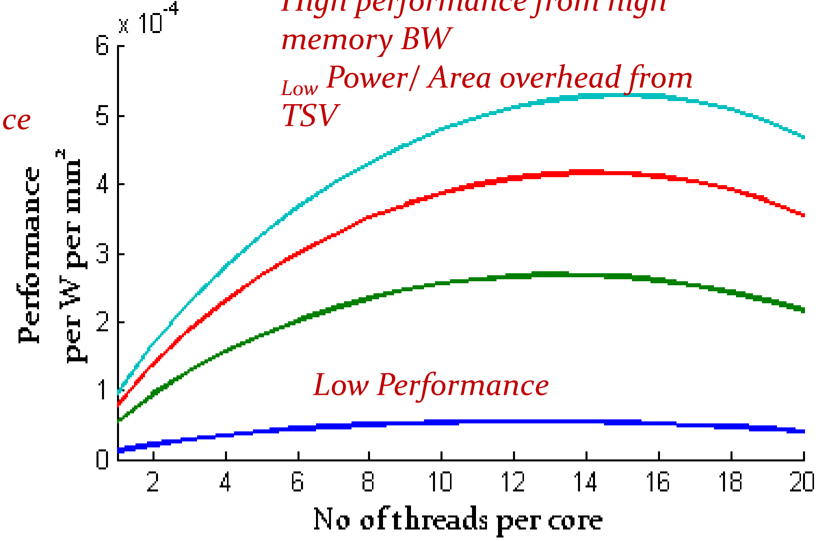
Higher core utilization with more TSVs

$$T_{mem\_ss} = t_{cache} + \frac{K_1}{N_{tsv}} + K_2$$



Low miss rate
- High Performance
- Power/ Area degrade metric

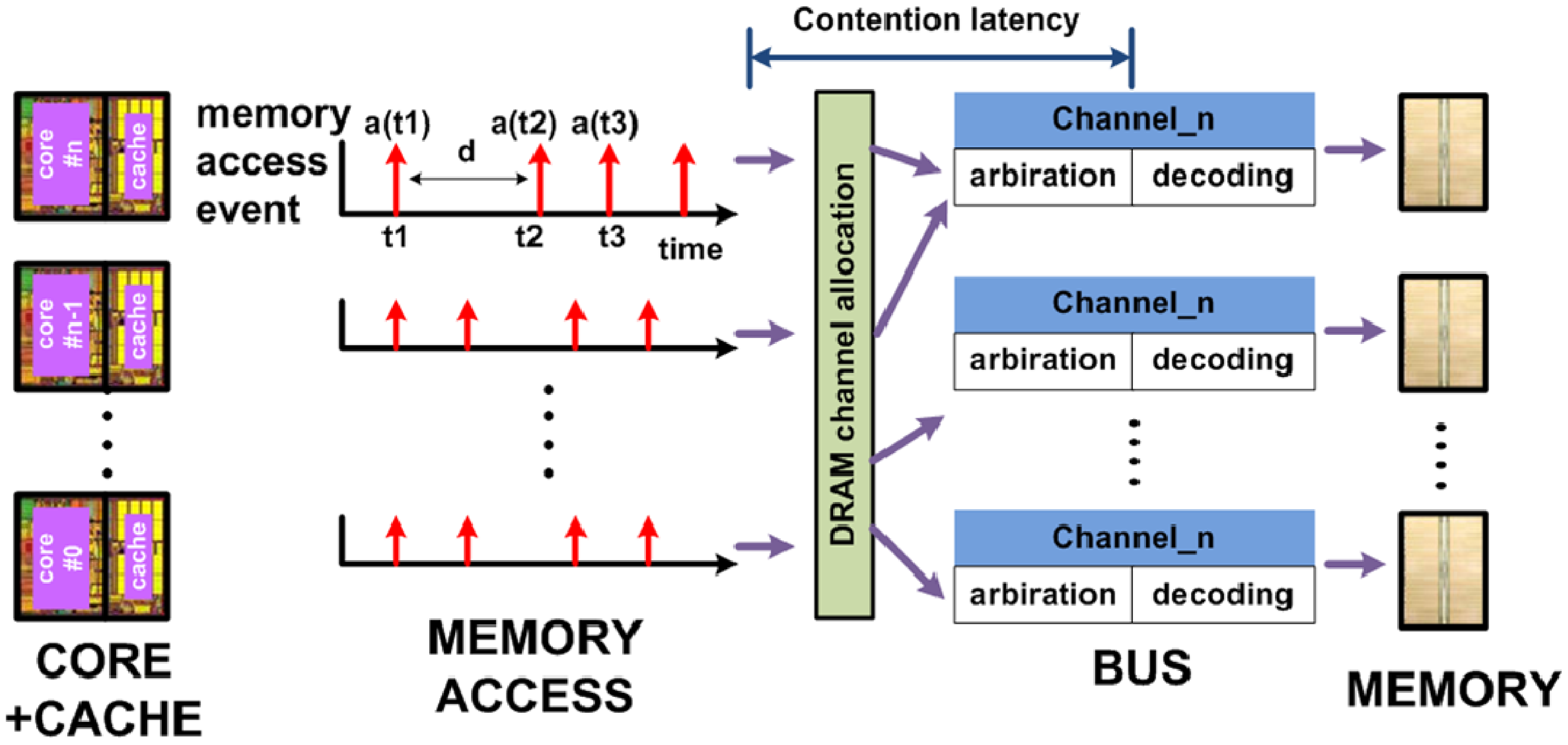
High miss rate
- Low Performance



High performance from high memory BW
Low Power/ Area overhead from TSV

Low Performance

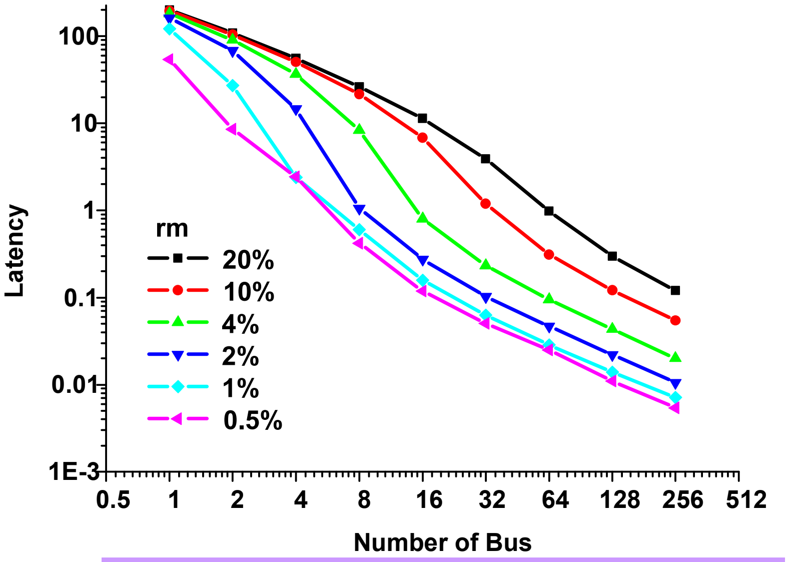
Bus Latency Modeling

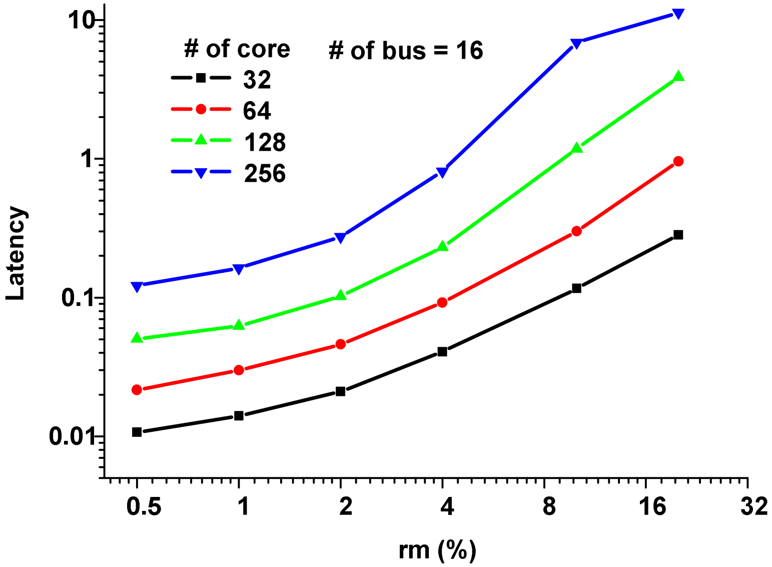


- Core generates transactions (d: Poisson distribution, a: Uniform distribution)
- Bus allocation block assigns transactions to a bus corresponding to transaction address.
- Transactions tothe same bus are allocated one by one at every clock cycle with round-robin arbitration scheme.
- Not allocated transaction should wait until it is assigned → increase latency

$$contention \propto f(N_{tsv}, N_{core}, r \cdot m, t_{mem})$$

r = (memory instruction)/(total instruction)
m = cache miss rate



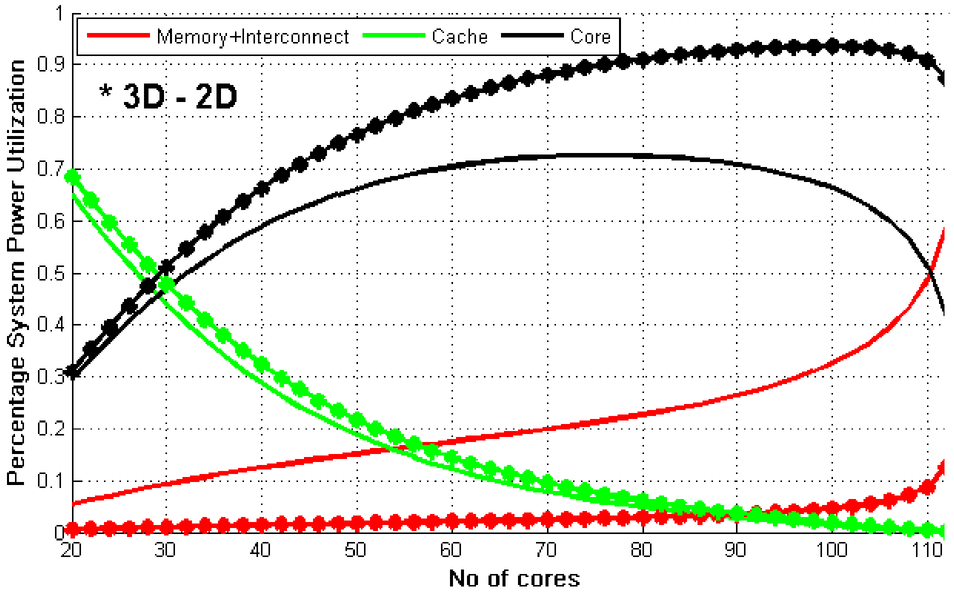


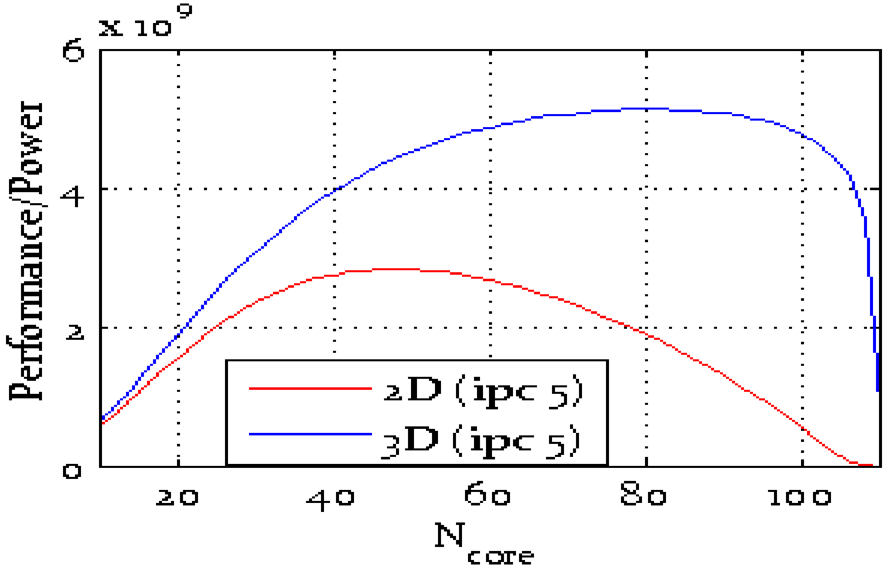
- r·m = 1/d = 0.05% ~ 20%
- # of core = 256
- # of bus = 2^n (n=0,1 ··· 8)

- r·m = 1/d = 0.05% ~ 20%
- # of core = 32 ~ 256
- # of bus = 16

The contention factor determines the effectiveness of TSV utilization

System Power Utilization





With increasing number of cores a greater percentage of the power

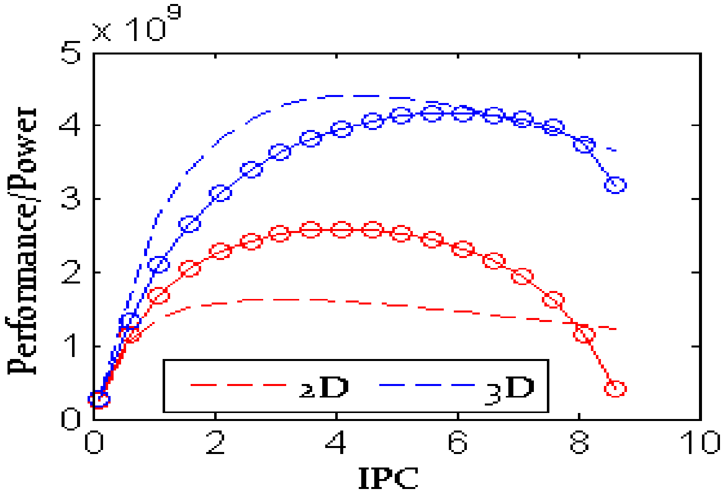
- **is utilized by the cores in a 3D system**
- **is utilized by the interconnect and DRAM subsystem in a 2D system**

Optimal Area Utilization

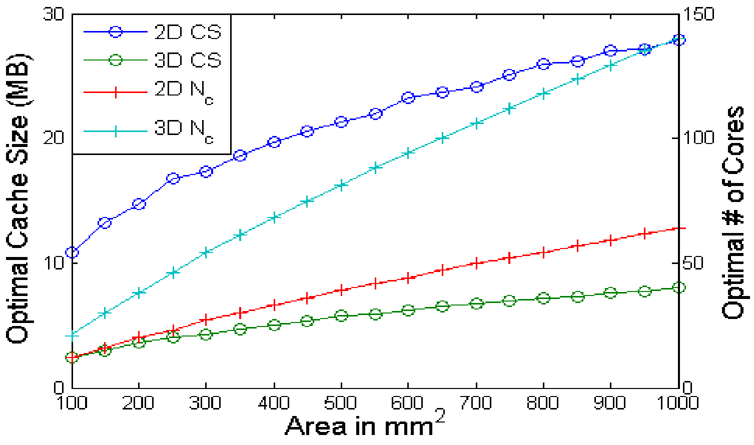
$$DieArea = N_{core} \bullet \{f(IPC) + f(\$C)\}$$

Core power and area for varying Issue and Execution widths was analyzed using McPat.

- Both Power and Area of the core increases exponentially with increase in IPC
- Reduces core count or area available for caches



Die size determines the optimal number of cores and cache size



Optimal design point for # of cores, cache size and IPC shifts in 3D to higher # of cores and commit widths.

Cores can increase at a much faster rate with increase in die area

Conclusions

- To fully utilize the 3D interconnect, the system should be redesigned to have smaller caches and more cores
    - Increased TSV bandwidth favors higher IPC cores
    - Maximizing TSV utilization requires new highly concurrent memory hierarchies

- More study is necessary to
    - Characterize the impact of traffic contention for the TSVs
    - Assess thermal consequences

- A major challenge is the real time management of TSVs to deliver the available bandwidth to concurrent cache miss events

Questions?

Contact info: mitchelle.rasquinha@gatech.edu
http://www.ece.gatech.edu/research/labs/casl/
http://www.ece.gatech.edu/research/labs/GREEN/