# Interconnect Meets Architecture: On-Chip Communication in the Age of Heterogeneity


**Partha Pratim Pande**

**School of Electrical Engineering and Computer Science, Washington State University**

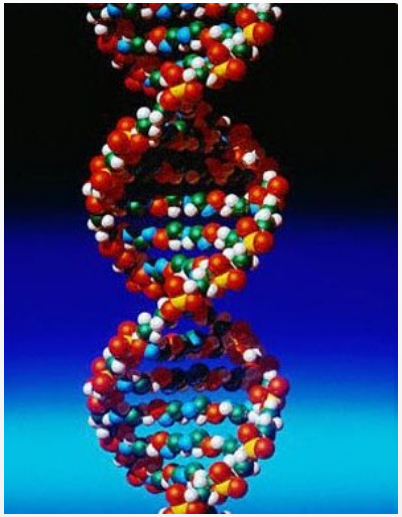

WASHINGTON STATE UNIVERSITY
*World Class. Face to Face.*

# Outline

- **Motivation**
  - Deployment Scenarios and Applications require high performance but constrained HW
  - Heterogeneous manycore systems are complex
- **Heterogeneous NoC**
  - Wireless, 3D and Photonics
  - EDA challenges
  - Scalable ML-Based EDA for NoC
- **Heterogeneous architectures for deep learning**
  - ReRAM
  - GPU-ReRAM heterogeneous architecture
  - Monolithic 3D NoC designs
  - Pros and Cons of M3D
  - Additional Design Considerations

# Why Many-Core Chips?

- Explosive computational power
  - Scientific applications
    - Weather prediction, Astrophysics
    - Bioinformatics, forensics
    - Language processing
  - Consumer electronics
    - Graphics, Animation

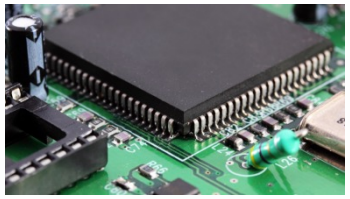



**The era of single processor systems is over**

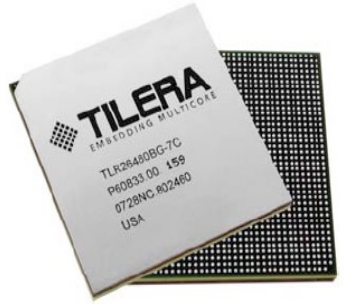
# The era of Many-Core systems

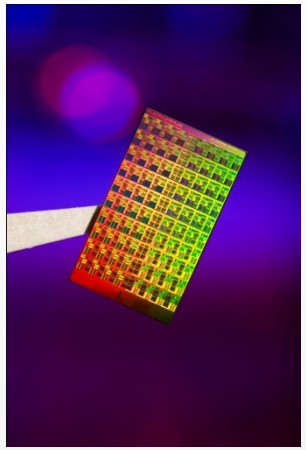
How to keep up with demands on computational power?

- Can not scale clock frequency
- Solution: Increase number of cores parallelism
  - Mass Market production of Intel, AMD dual-core and quad-core CPUs
  - Custom Systems-on-Chip (SoCs)
- Many Core chips from Tilera for networking, cloud computing and multimedia applications.

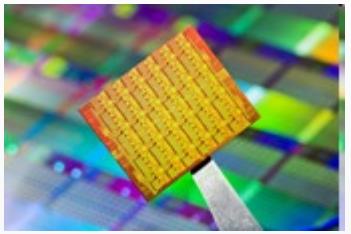


Adapteva's Epiphany



Intel 80 core processor

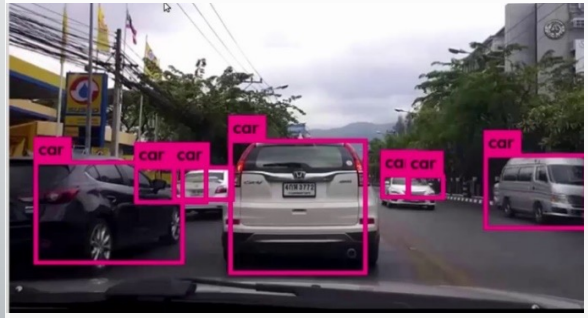


Single-chip Cloud Computer

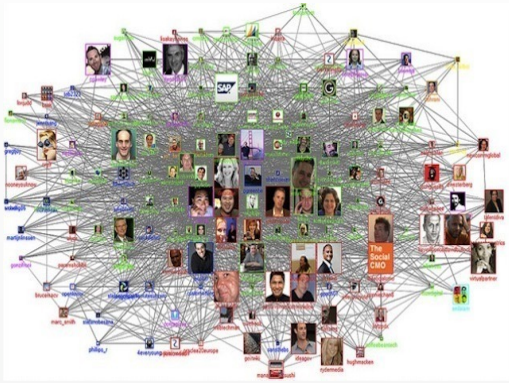
**'Number of cores will double every 18 months'**
**- Prof. A. Agarwal, MIT, founder of Tilera Corporation**

# Big Data Revolution

**Machine Learning**



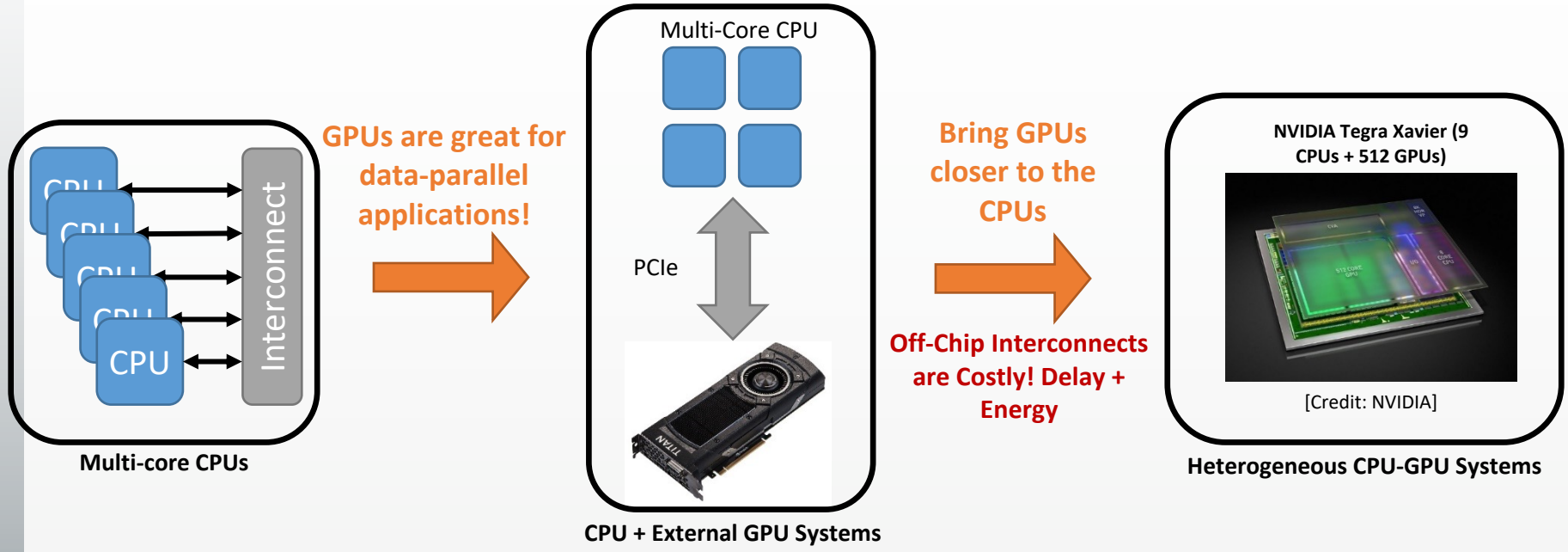

**Graph Analytics**



**Genomics**

**Highly compute- and data-intensive!**
**High performance and low-power platforms are needed to enable them**

# Traditional Hardware for Big Data Applications




- Move the GPUs on to the chip!
- Higher data compute + lower power using Network-on-Chip (NoC)

# Motivation: Communication Backbone

- Massive multicore processors are enablers for ICT innovations
- Need for holistic power optimization and management
- Energy across the layers

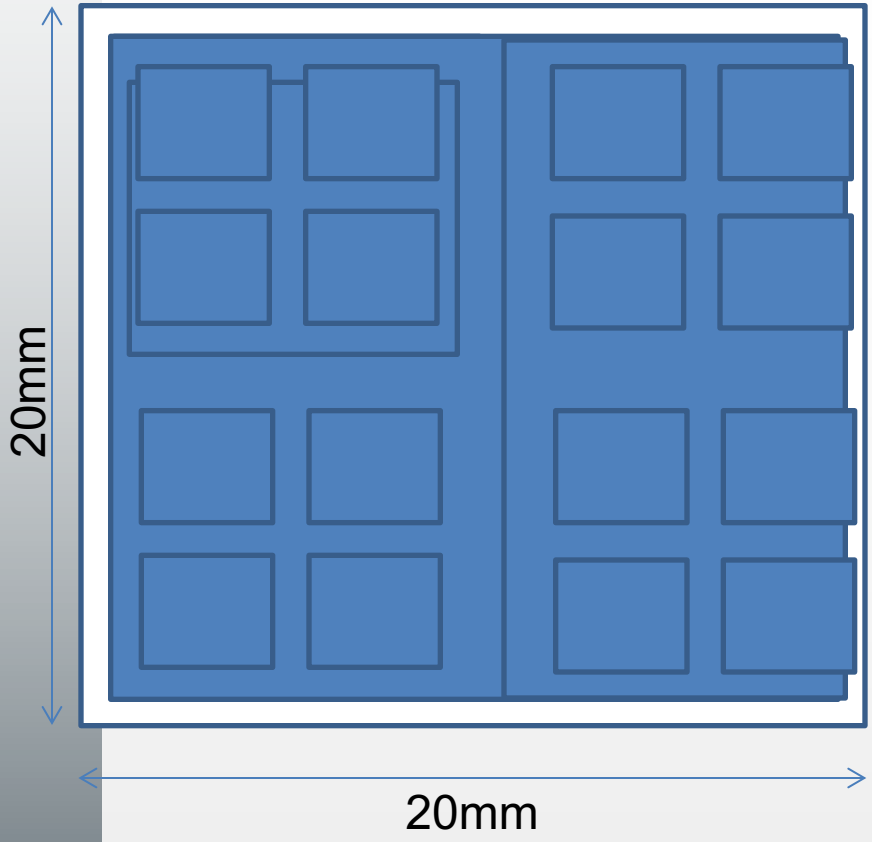
**“We need research on how to minimize communication, since energy is largely spent in moving data”**

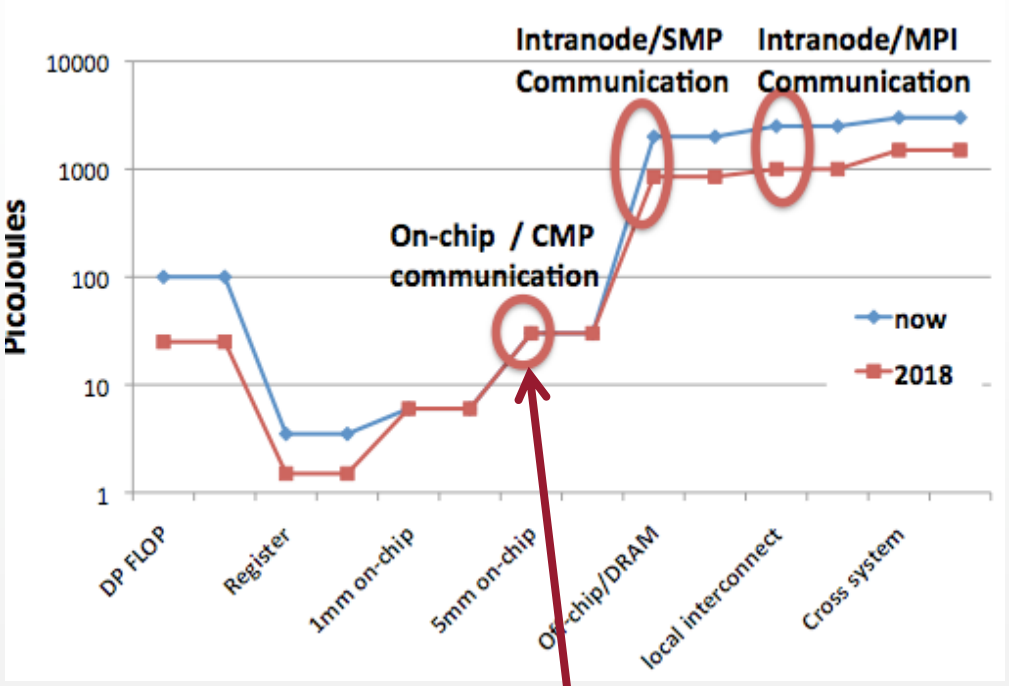
**“21st Century Computer Architecture" commissioned by the Computing Community Consortium**

**NSF Workshop on Cross-Layer Power Optimization and Management (Feb ‘12)**
**NSF Workshop on achieving ultra-low latencies in wireless networks (March ‘15)**

# Moving a bit across die

**Moore's Law:**



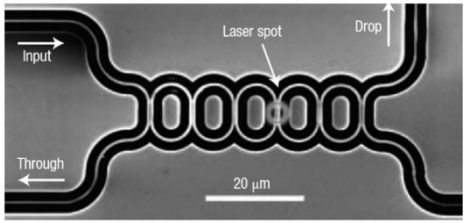




Source: Exascale Roadmap Meeting, Dec. 2009

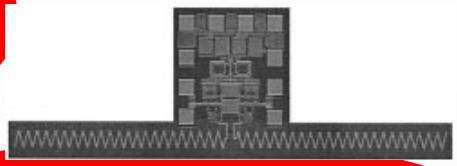
**On-Chip: Not Scaling**

**Global on-chip movement of data is very power hungry!!**

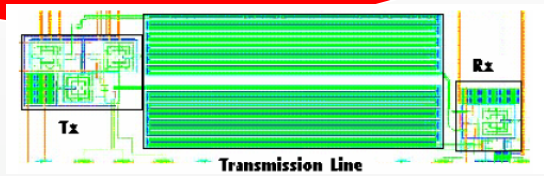
# Novel Interconnect Paradigms for Multicore designs



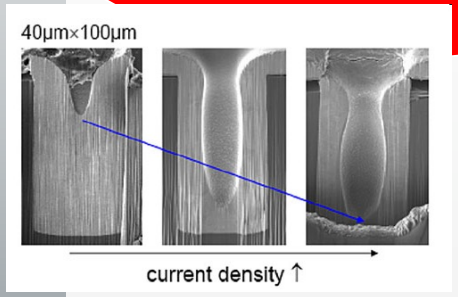

**Optical Interconnects**



**Wireless/RF Interconnects**



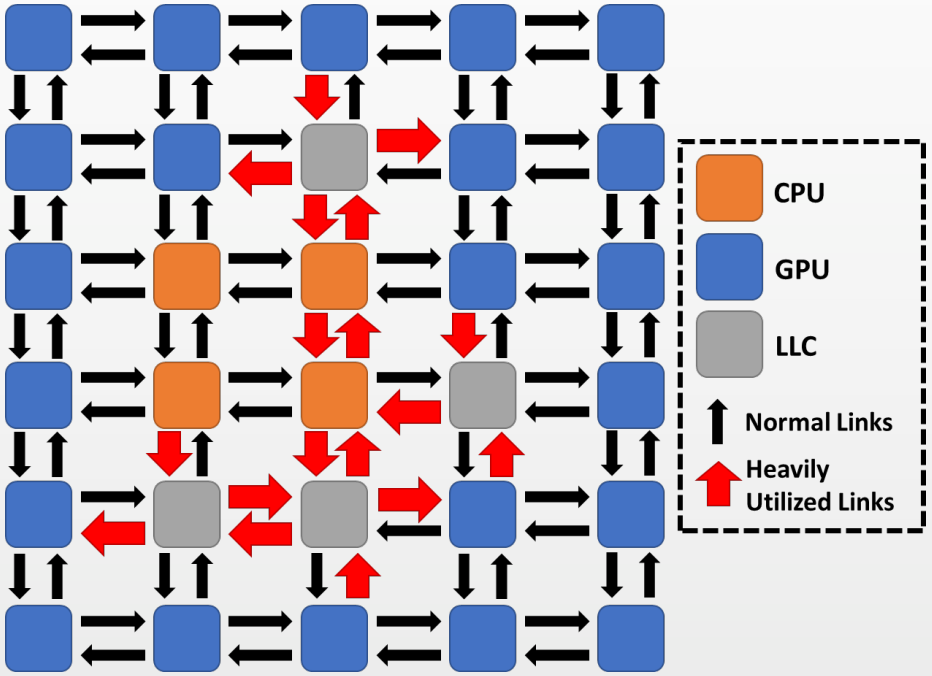

**Three Dimensional Integration**




**High Bandwidth and Low Energy Dissipation**

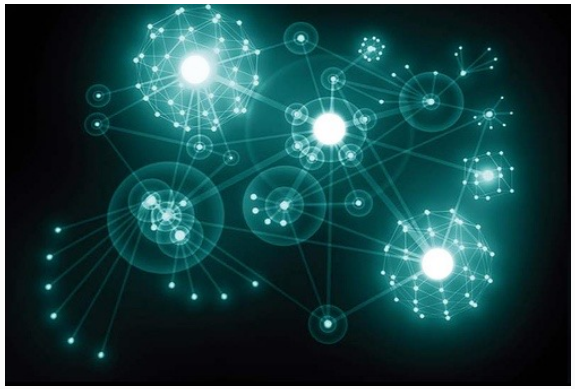
# Why not Mesh?

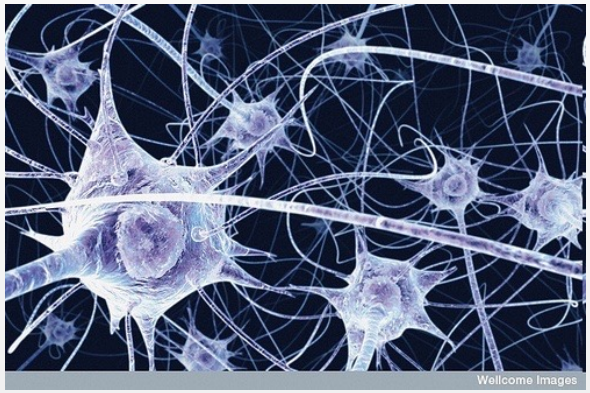
- Mesh has multi-hop nature
  - Higher latency and energy for bigger system sizes
- Skewed traffic in heterogeneous architectures with CPU + GPU
  - Many-to-few-many communication around the LLCs
  - Few links become traffic hotspots and create bandwidth bottlenecks in Mesh for heterogeneous architectures
- This necessitates the investigation of more complex NoC designs
  - Needs to account for specific traffic characteristics and system requirements



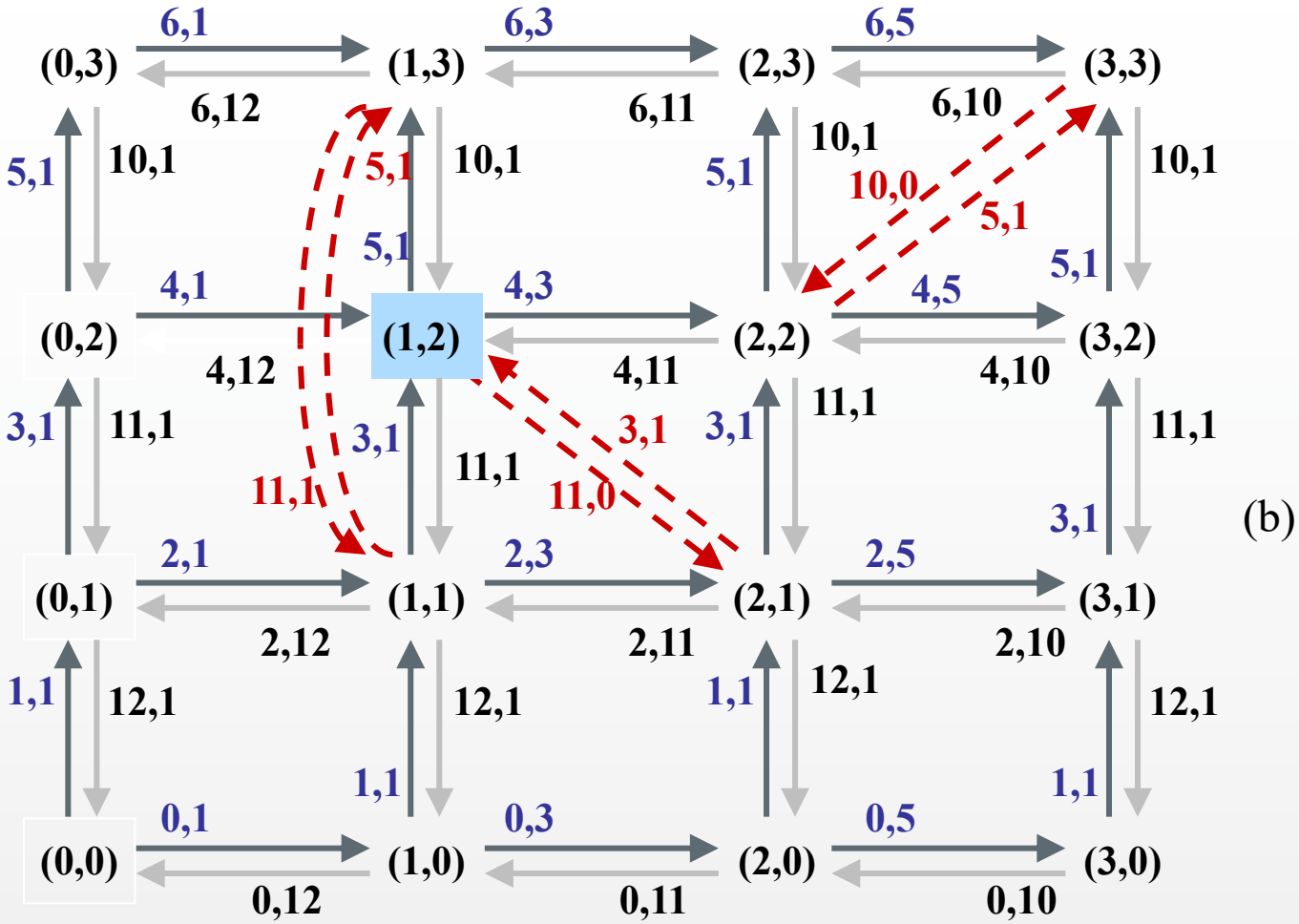
# Designing better NoC: Natural Complex Networks

- Natural Complex Networks
  - Brain
  - Microbes
  - Social Networks
- Small-World/Exponential graphs
  - Attacks
- Scale-free graphs
  - Random Failures





# Small World and NoC




Umit Ogras and Radu Marculescu, "It's a Small World After All": NoC Performance Optimization Via Long-Range Link Insertion", TVLSI, vol. 14, No. 7, July 2006

# Power-Law based Small World Network

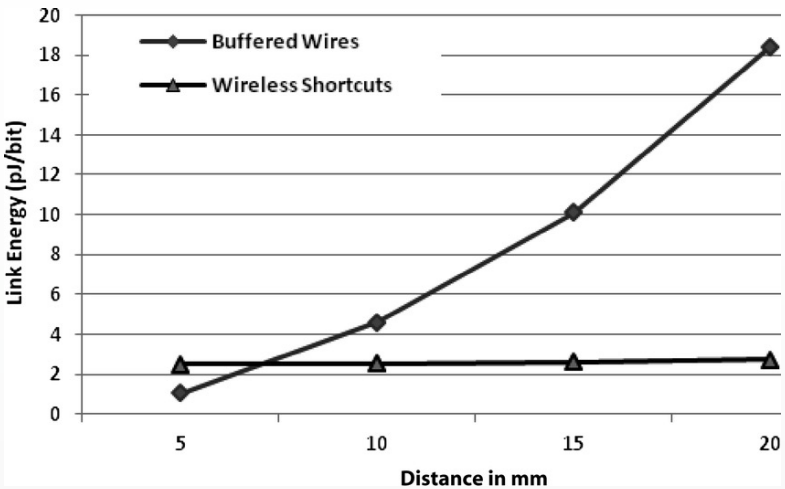
- Power-law based connectivity

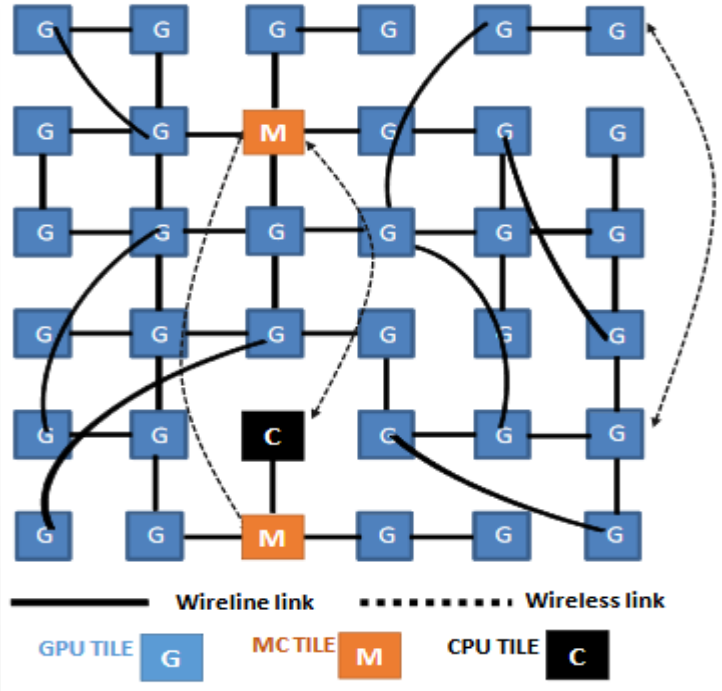
$$P(i,j) = \frac{l_{ij}^{-\alpha} f_{ij}^{\beta}}{\Sigma_{\forall i}\Sigma_{\forall j} l_{ij}^{-\alpha} f_{ij}^{\beta}}$$

  - Many short-range local links
    - Conventional Wireline links
  - A few long-range shortcuts
    - Utilize emerging interconnects

**How to efficiently distribute the shortcuts?**

# WiNoC Configuration

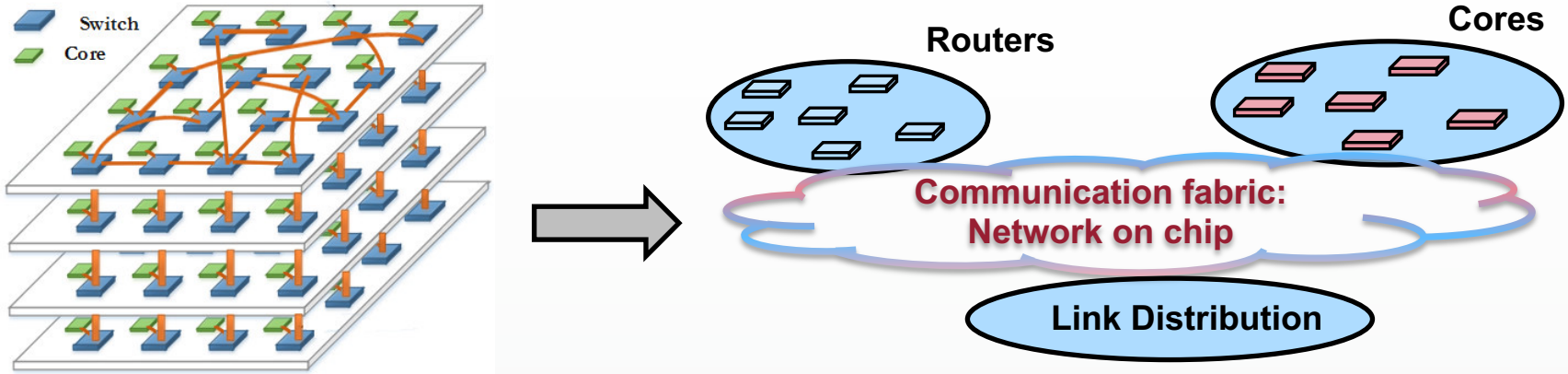



- Wireless links more efficient for long range communication
- Add few long-range shortcuts
  - Hybrid wired-wireless NoC design

S. Deb, A. Ganguly, P. P. Pande, B. Belzer and D. Heo, "Wireless NoC as Interconnection Backbone for Multicore Chips: Promises and Challenges," in IEEE Journal on Emerging and Selected Topics in Circuits and Systems, vol. 2, no. 2, pp. 228-239, June 2012.

# How does it look in 3D?



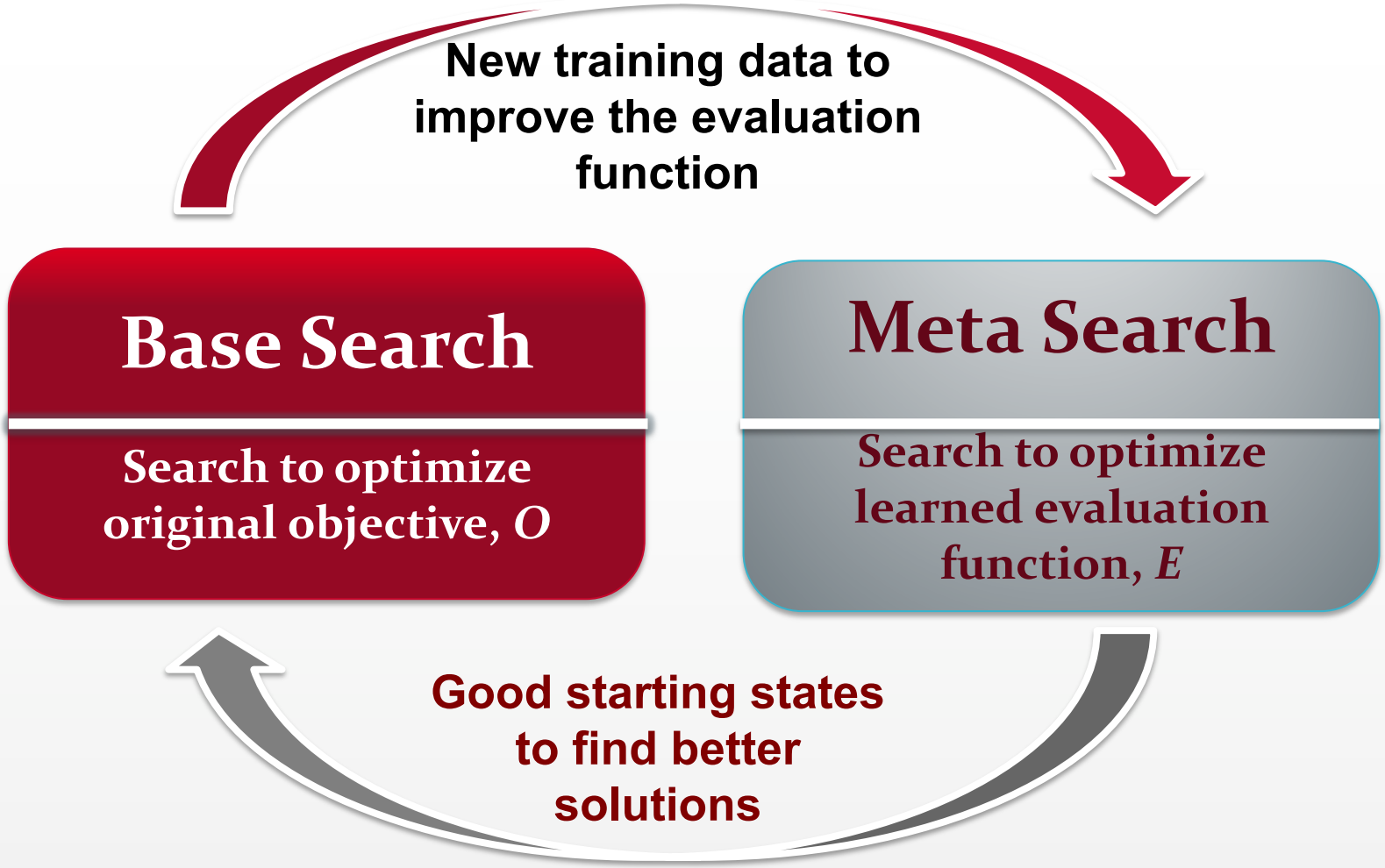

- Physical design space is combinatorial by nature
- Placement of routers, cores and links ensures NoC performance level

Challenges: 1. How to place the routers, cores and links ?
2. How to map the tasks?
3. How to explore optimized link placement efficiently?

# Machine learning in Design Optimization

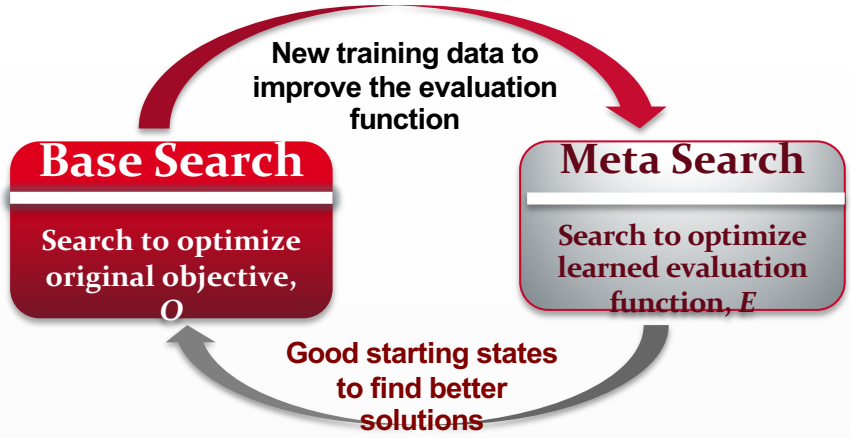
- **Space of feasible NoC Designs is combinatorial**
  - Cannot perform exhaustive search, specifically for bigger system size
- **Key Insight:** Intelligent exploration of the design space to quickly find (near)-optimal SW-NoC design
- **How to explore intelligently?** via Machine Learning

# Design Optimization: STAGE Algorithm



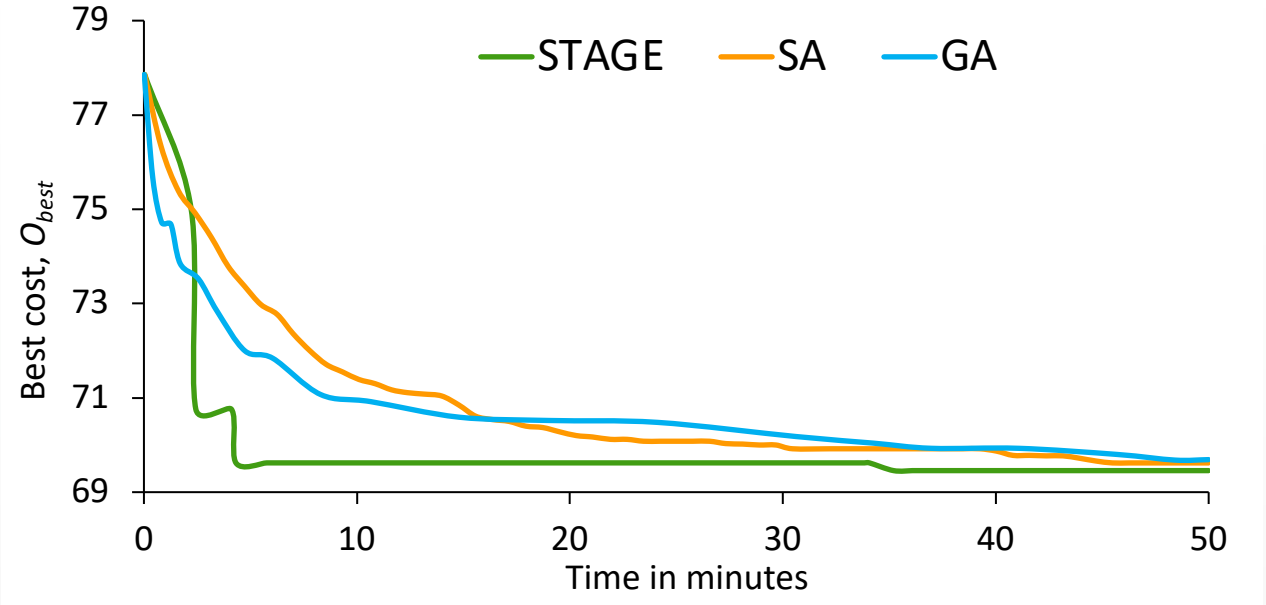

***E*** is the search control knowledge that will improve based on the search experience via machine learning

# 3D NoC Optimization: STAGE Instantiation




- **Regression learning algorithm**
  - Need fast training time and testing time -- they contribute to the overall optimization time
- We employed Regression Tree Learner (via WEKA package)
  - Regression tree training is fast; and allows us to learn accurate predictors

# Performance Evaluation: STAGE vs. SA and GA

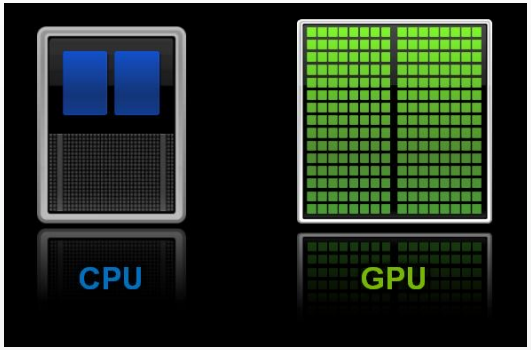



- STAGE converges faster than Simulated Annealing (SA) and Genetic Algorithm (GA)
- For a given time budget, solution quality of STAGE is better than SA or GA

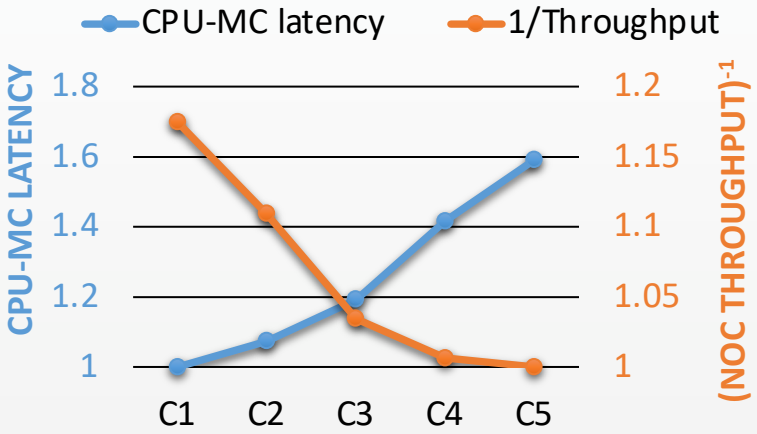
S. Das; J. R. Doppa; P. P. Pande; K. Chakrabarty, "Design-Space Exploration and Optimization of an Energy-Efficient and Reliable 3D Small-world Network-on-Chip," TCAD, 2016.

# Design Challenges: Heterogeneity

- Multiple Quality of Service (QoS) requirements
- Disparate natures of CPU and GPU architectures
  - Conflict with one another
  - CPU: Latency sensitive
  - GPU: Throughput sensitive

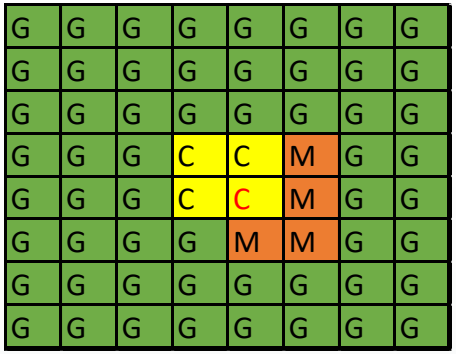


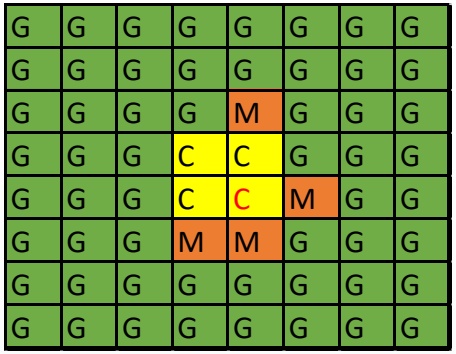

Source: blogs.nvidia.com

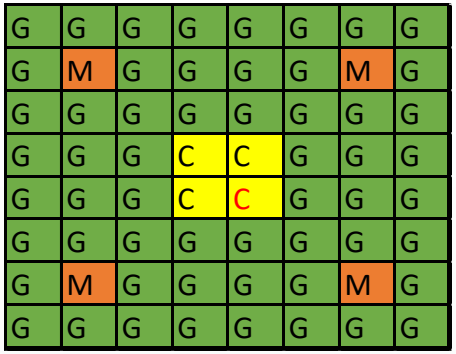



W. Choi et al., On-Chip Communication Network for Efficient Training of Deep Convolutional Networks on Heterogeneous Manycore Systems, in IEEE TC, vol. 67, no. 5, pp. 672-686, 2018.

# Mesh: Candidate Solution Tradeoffs

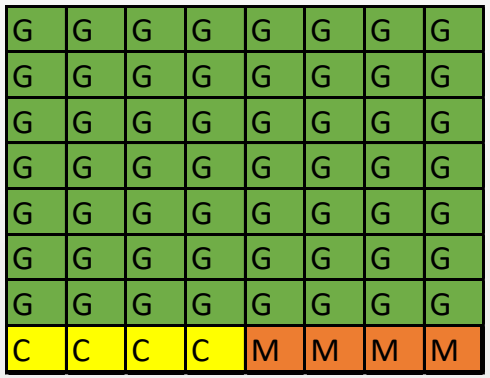


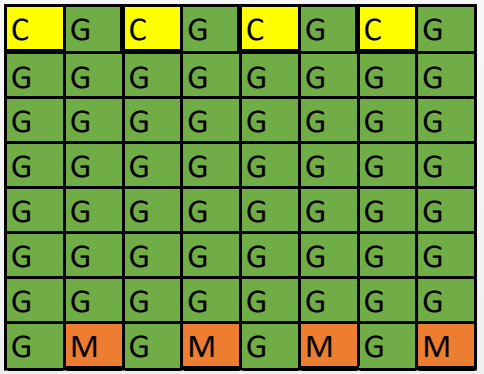




**Closer MC and CPU Placement**
***Low CPU Latency for Master core***

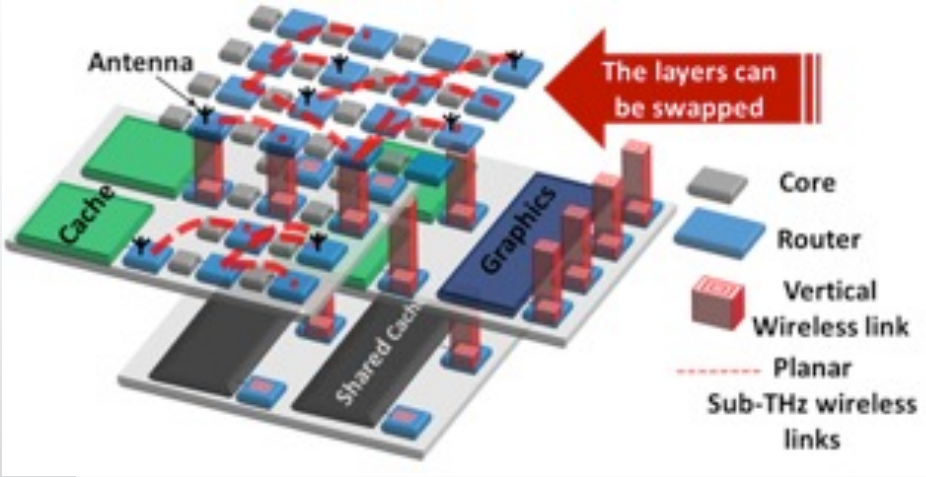
**Well distributed MC**
***High-Throughput***

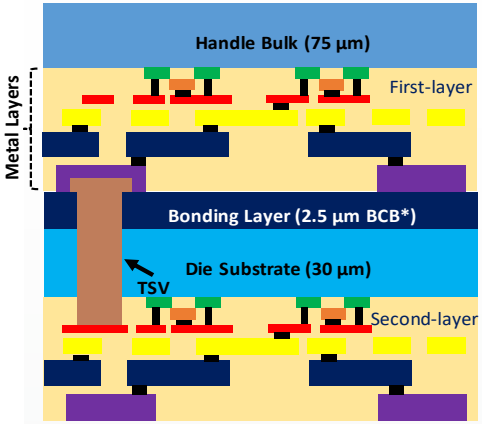


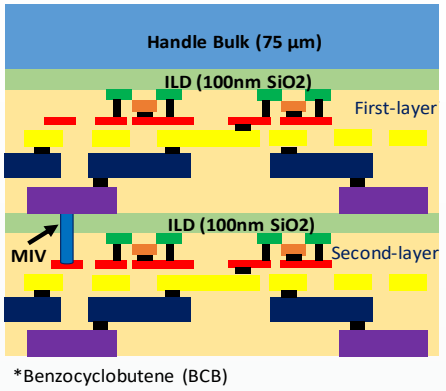

**Unoptimized Placement (Random)**

# Heterogeneity: Wireless, M3D




TSV-based 3D IC

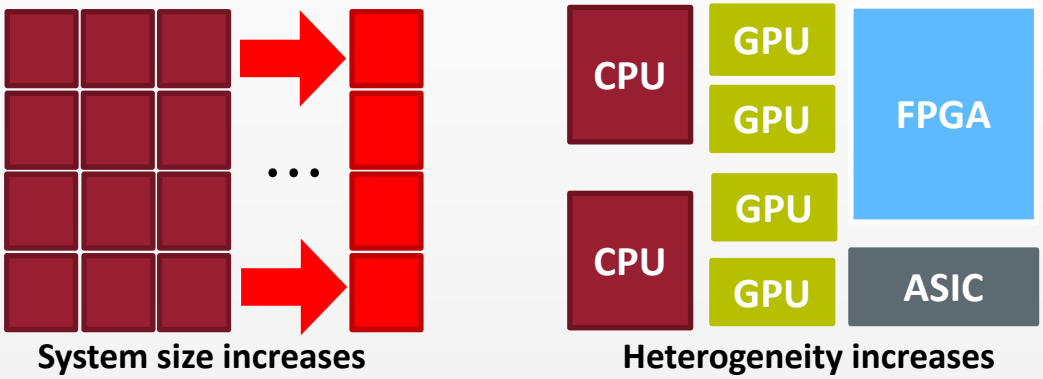



Monolithic 3D IC

- Emerging interconnects like wireless and Monolithic 3D (M3D)
  - More dimensions to NoC design in future platforms
- Similar to increased heterogeneity in compute, increased heterogeneity in communication
  - Wireless: Different data rates, power consumption than conventional wires; Heterogeneity among links
  - M3D: Process Variations between layers; Heterogeneity among layers

# Heterogeneous NoC Design Optimization: Challenges

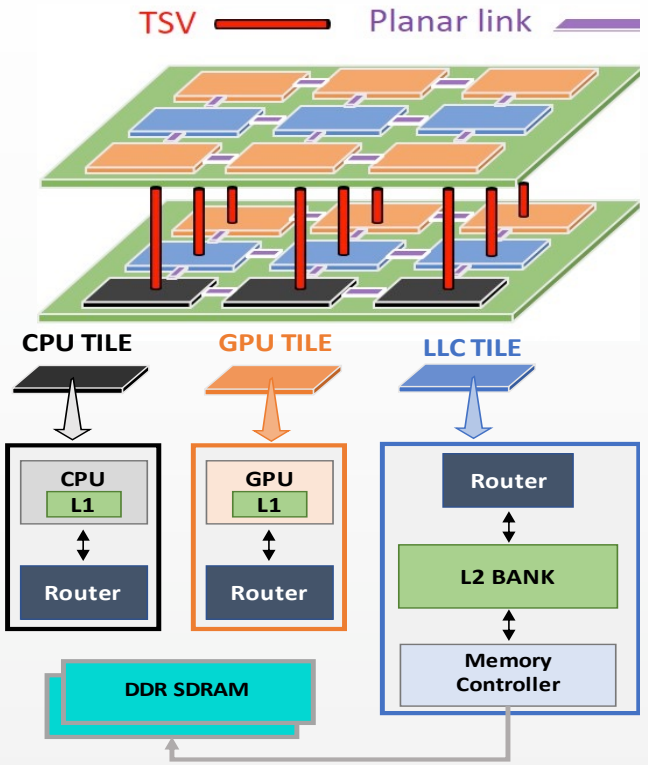
- Design complexity will increase in future architectures
  - Increasing system size
  - More heterogeneity
- Bigger search space
  - Difficult to find good solutions



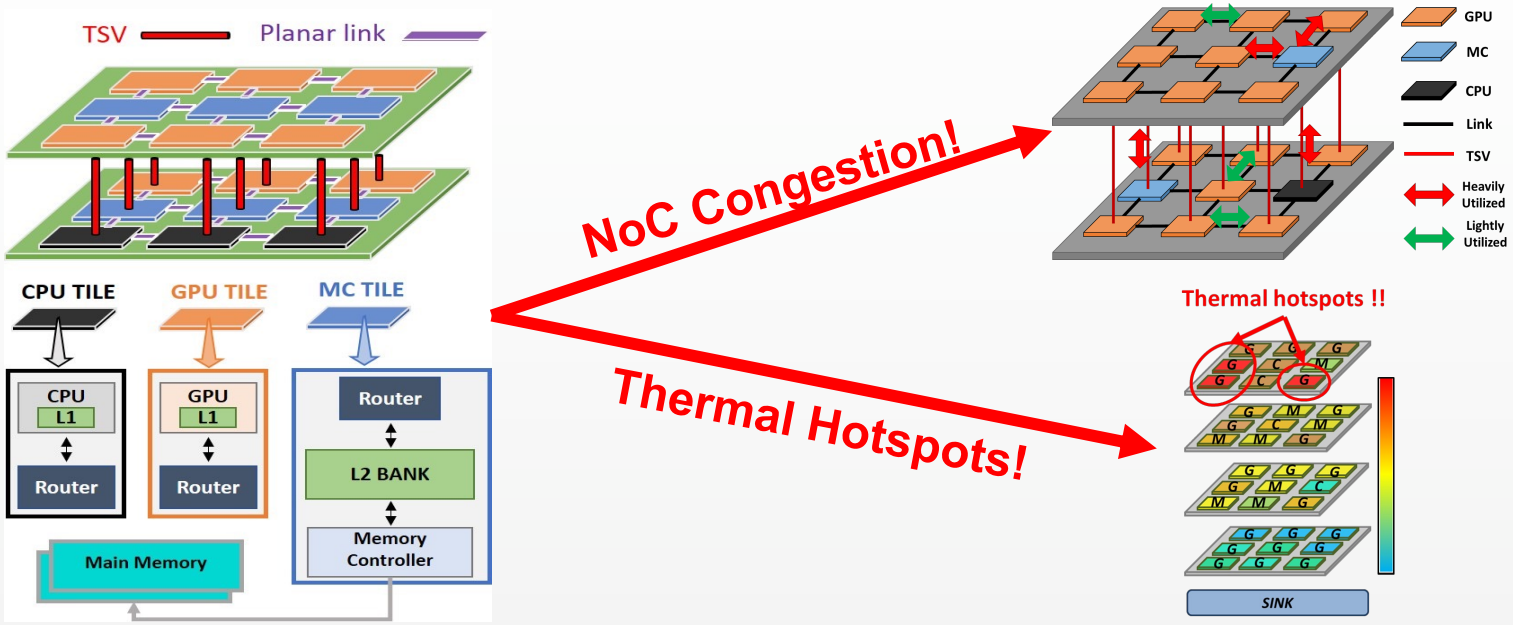

S. Das et al., "Monolithic 3D-enabled High Performance and Energy Efficient Network-on-Chip," *ICCD,* 2017.
R.G. Kim et al., "Machine Learning and Manycore Systems Design: A Serendipitous Symbiosis," *IEEE Computer*, 2018.

# 3D Heterogeneous Systems

- We can incorporate multiple types of cores or tiles into a 3D heterogeneous system
  - CPU (latency centric)
  - GPU (throughput centric)
  - Last level cache (LLC): consists of L2 cache slice and access to main memory
- Connect tiles on the same die using a normal NoC
- Connect dies using TSVs
- Increase scalability and reduce network diameter



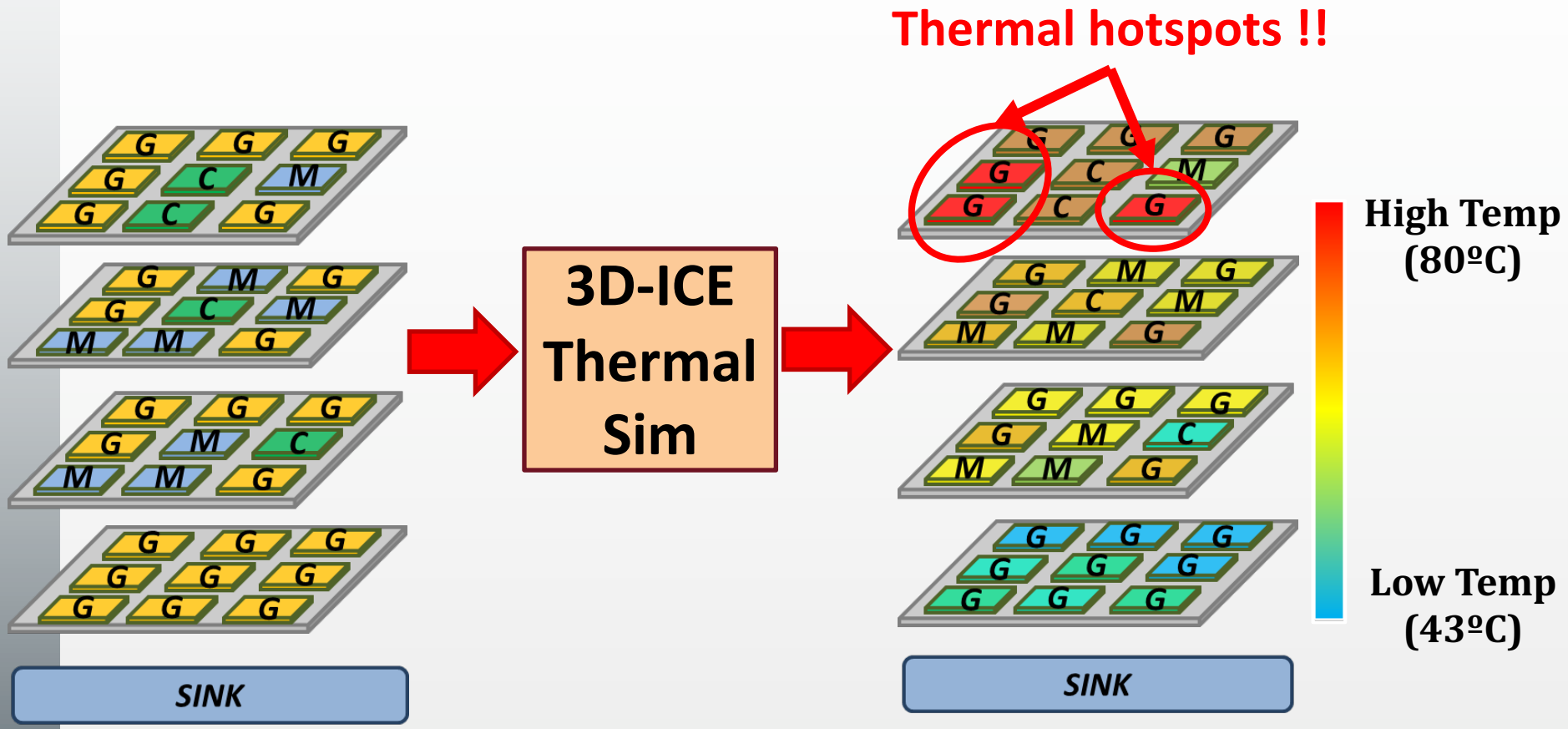
# CNN Training using 3D Heterogeneous Architectures




- **We can utilize these 3D heterogeneous systems to speed up CNN training**
  - **30% less hops (2D Mesh vs 3D Mesh)**

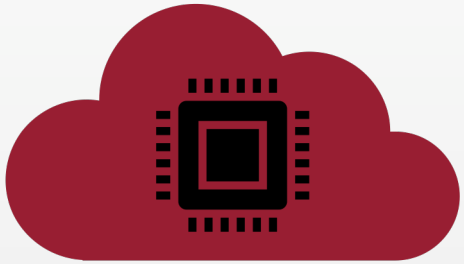
# $3DHet_{perf}$: Thermal Issues

- Ignoring thermal constraints leads to significant temperature hotspots



# Multi-Objective Optimization (MOO) is Necessary!

- Machine learning and big-data analytics can be deployed anywhere
  - Cloud, automotive embedded system, IoT, etc.
  - Each of these have drastically different design constraints
  - Demands for computing platforms with higher performance in highly constrained scenarios have led to more specialized systems
- **NEED** to abandon design solutions that consider only a single aspect and pursue designs that jointly consider power, reliability, and the performance of individual components

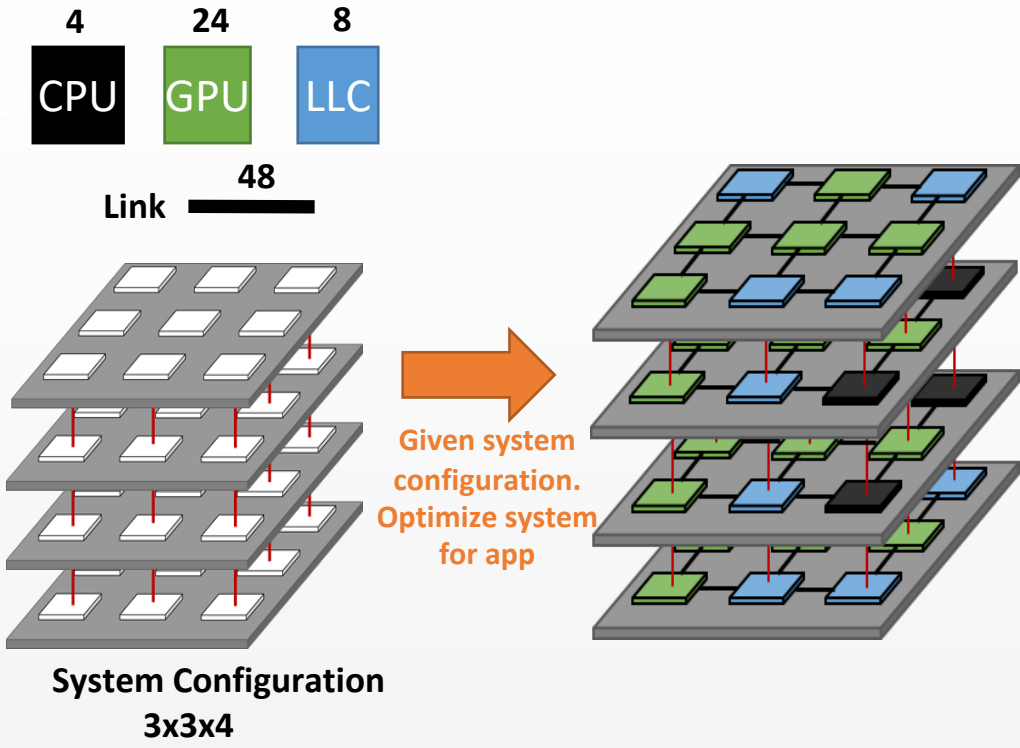


**Cloud**: Performance and Thermals

**Embedded System**: Real-time constraints and Reliability

# General Problem Formulation (MOO)

- Given:
  - The mix of CPU/GPU/LLC tiles
  - The number of planar links
  - System configuration
- Optimize the placement of:
  - CPU/GPU/MC tiles
  - Planar Links
- Three objectives (Example)
  - *CPU Communication Objective*
  - *GPU Communication Objective*
  - *Thermal Objective*



# Problem Formulation

- GPU Communication Objective:
  - GPU communication is more reliant on network throughput
  - minimize $\overline{U}$ (Avg. link util.) and $\sigma$ (Std. dev. Link util.)
  - This also helps balance the MC-Core traffic!
- CPU Communication Objective:
  - Observation: CPUs communicate with MCs and among themselves
  - CPUs are sensitive to communication latency: minimize traffic-weighted CPU-MC and CPU-CPU hop-count ($H$)
- Thermal Objective:
  - 3D systems have higher power densities that need to be accounted for.
  - Estimate thermal effects using power and thermal resistivity (T)

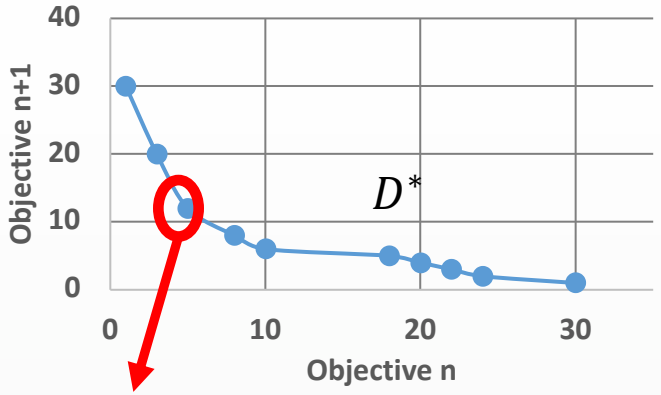
# 3DHet: Optimizing 3D NoCs for Accelerating CNN Training

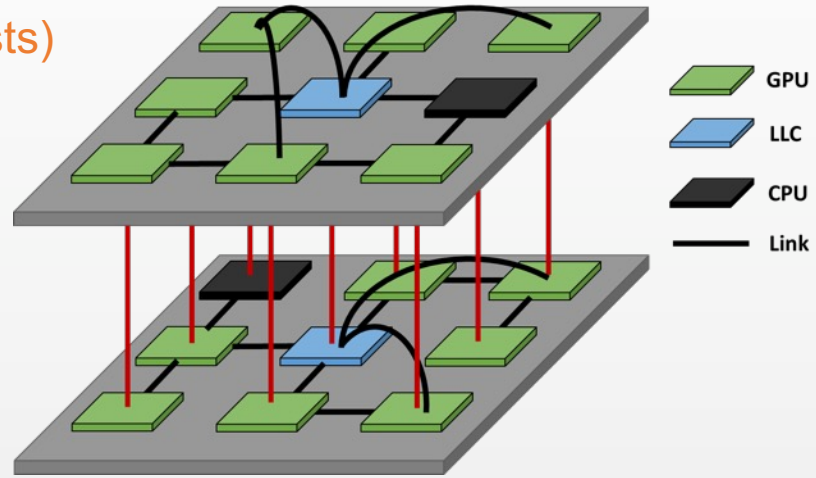
- Want to design the system while simultaneously looking at ***GPU***, ***CPU***, *and* ***Thermal*** objectives
- AMOSA (Archived Multi-Objective Simulated Annealing)
  - Creates set of candidate solutions $D^*$
- $D^* = AMOSA(D, OBJ = \{\bar{U}(d), \sigma(d), H(d), T(d)\})$

  s.t. $\forall i: L_i \leq k_{max}$ (max # of links for a router)

  $\forall i, j \in d: Path(i, j) = 1$ (Comm. Path exists)

- Choose best solution based on detailed simulation
- $\hat{d} = argmin_{d \in D^*} EDP(d)$
  - s.t. $T(\hat{d}) \leq T'$ (minimize EDP within temp. constraint)

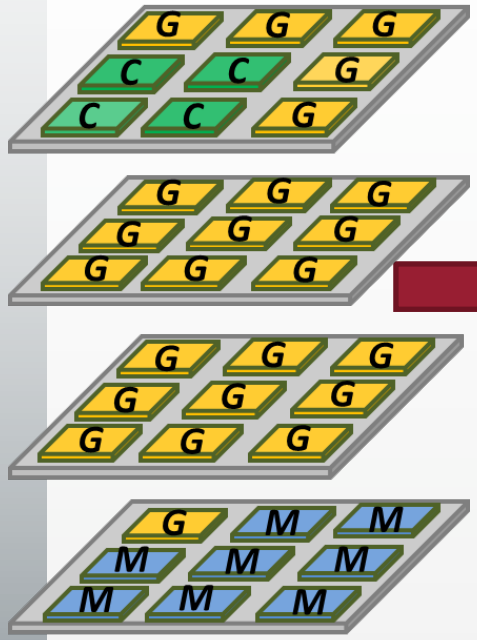



**BK Joardar et al., "3D NoC-Enabled Heterogeneous Manycore Architectures for Accelerating CNN Training: Performance and Thermal Trade-offs," in NOCS, 2017.**

# $3DHet_{perf}$: Optimizing performance

Random starting network



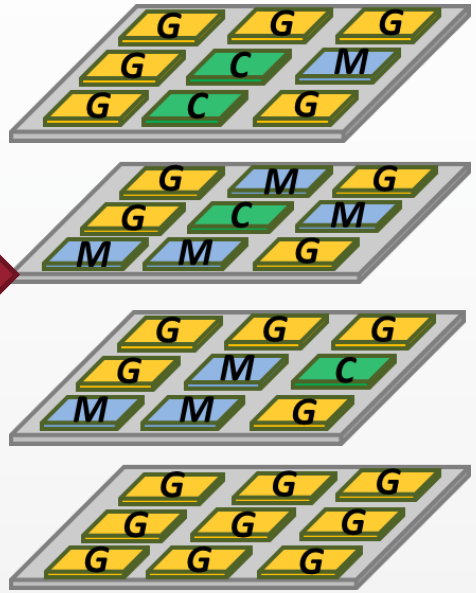

Constraints
e.g., $k_{max}$

**MOO Solver**

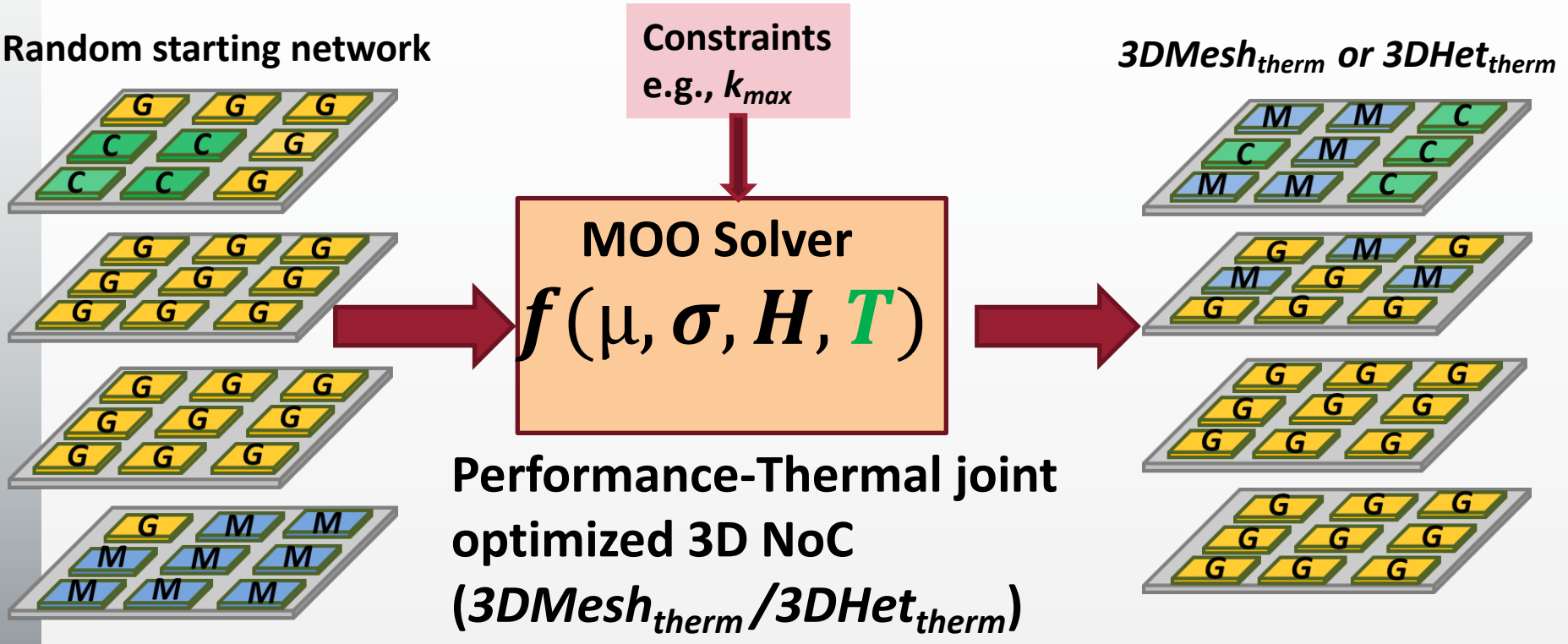
$$f(\mu, \sigma, H)$$

**Performance-only optimized 3D NoCs ($3DMesh_{perf}$/$3DHet_{perf}$)**

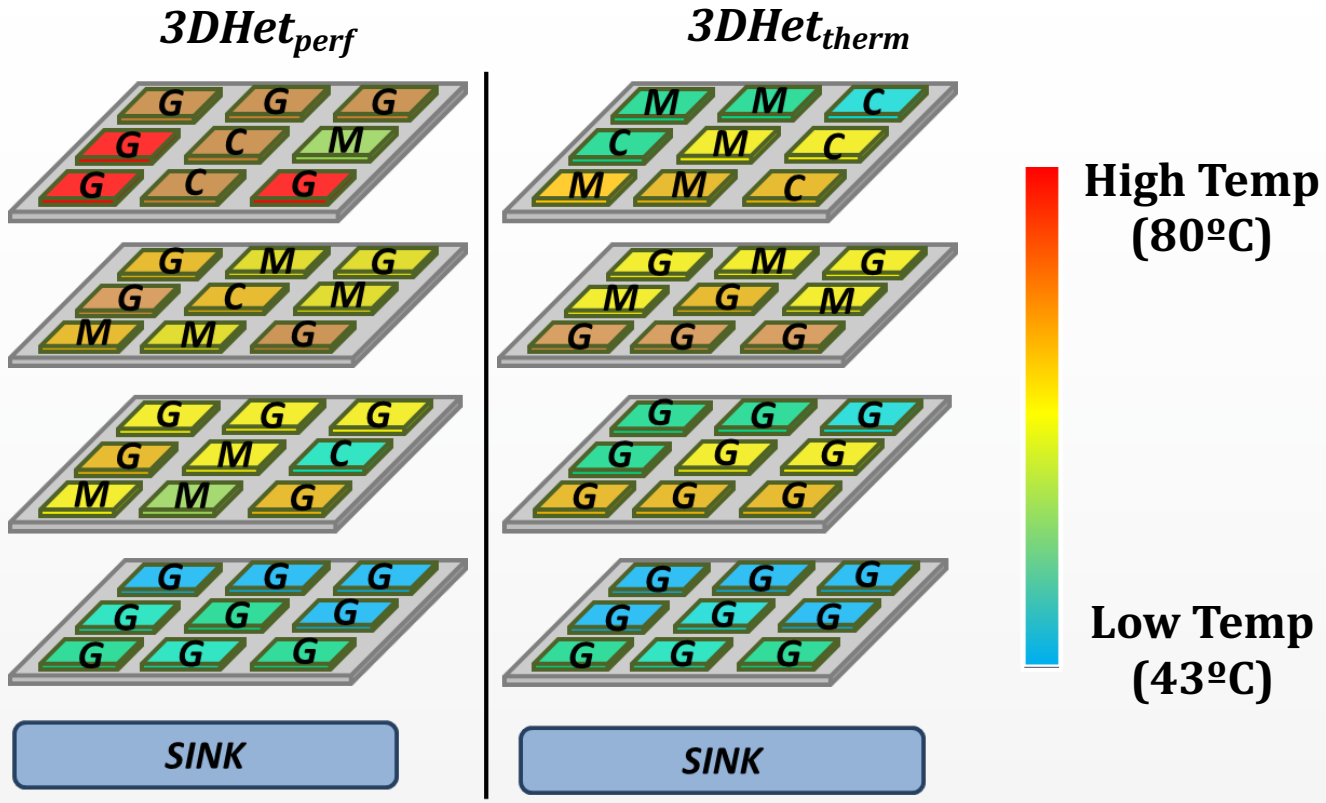
$3DMesh_{perf}$ or $3DHet_{perf}$



# $3DHet_{therm}$: Balancing performance and thermal

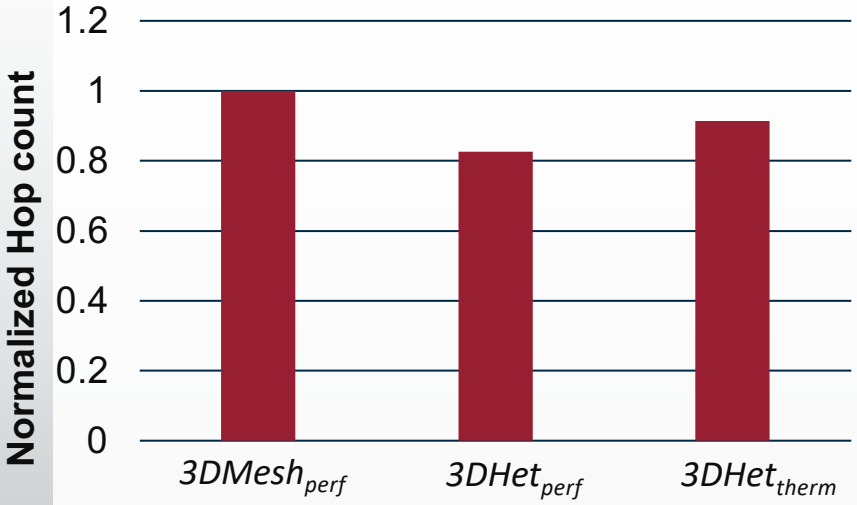


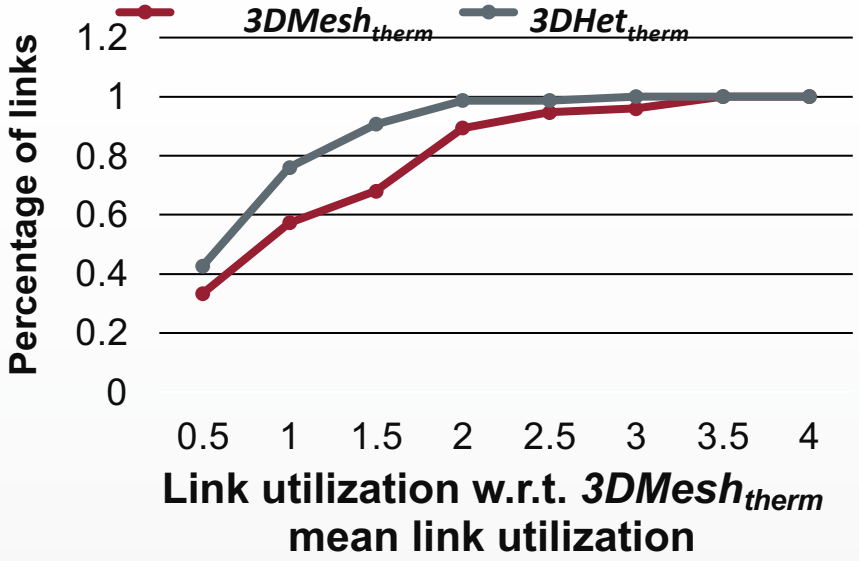
# $3DHet_{perf}$ vs. $3DHet_{therm}$: Thermal




- Much lower hotspots observed
- GPUs pushed down towards the sink
- Approximately 18°C improvement in maximum temperature compared to $3DHet_{perf}$

# *perf* vs. *therm*: NoC Hop-Count and Utilization



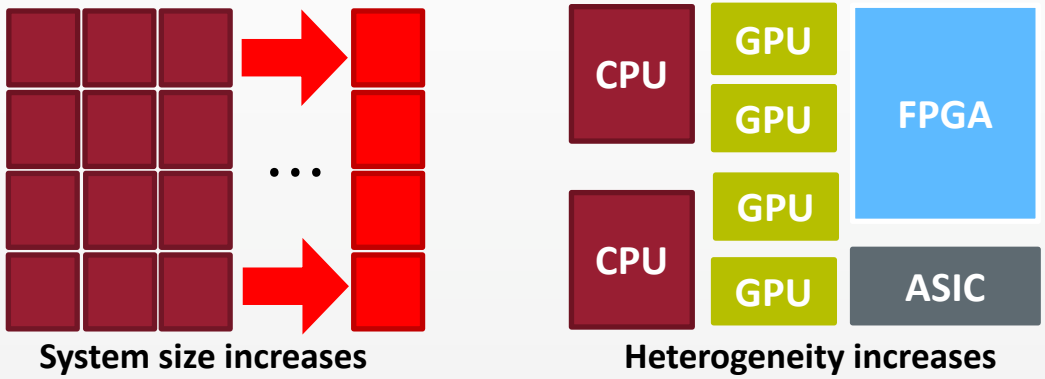




- Average inter-router Hop count
  - *3DHet* reduces average inter-router hop count compared to 3D Mesh
- Link utilization
  - More than 10% links in *3DMesh*$_{therm}$ have 2X higher link utilization than the mean while some carry greater than 3X the mean link utilization.
  - Traffic more evenly distributed in proposed *3DHet*$_{therm}$ *compared to* *3DMesh*$_{therm}$, hence higher throughput

**Possible bottlenecks during execution**

# 3D Heterogeneous NoC Design: ML revisited?

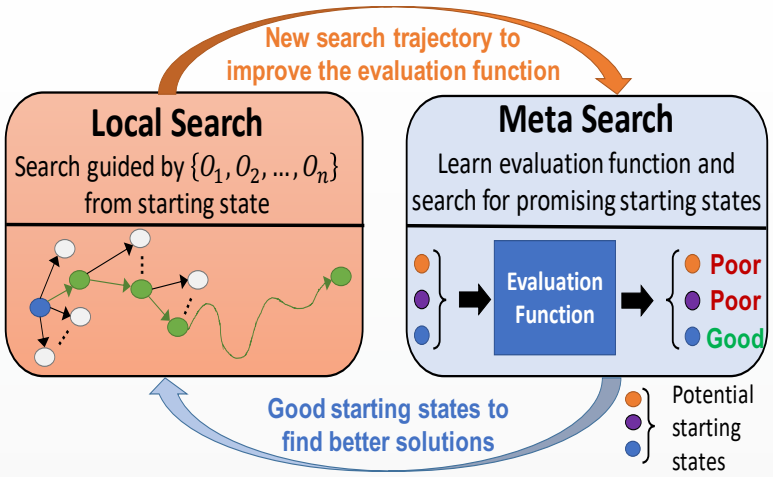
- Popular MOO algorithms like NSGA-II and AMOSA can find Pareto front
- **HOWEVER,** as design complexity increases, they will take longer and longer to find near-optimal solutions
  - Increasing system size
  - More heterogeneity

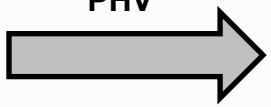


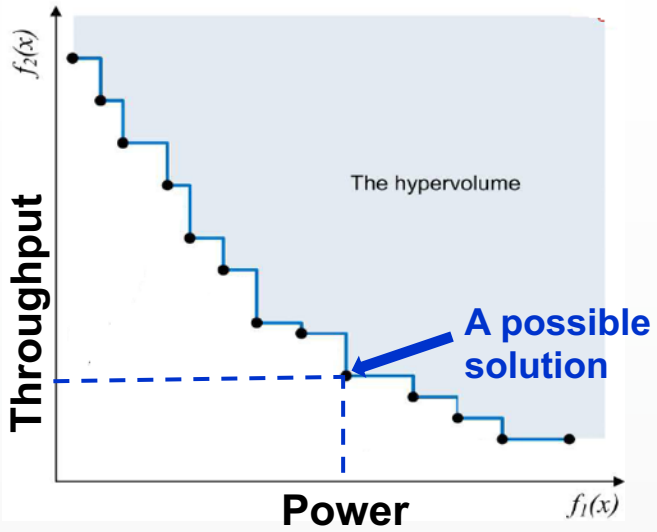

**ML inspired techniques should be used**

S. Das et al., "Monolithic 3D-enabled High Performance and Energy Efficient Network-on-Chip," *ICCD,* 2017.
R.G. Kim et al., "Machine Learning and Manycore Systems Design: A Serendipitous Symbiosis," *IEEE Computer*, 2018.

# MOO-STAGE: ML to solve MOO problems



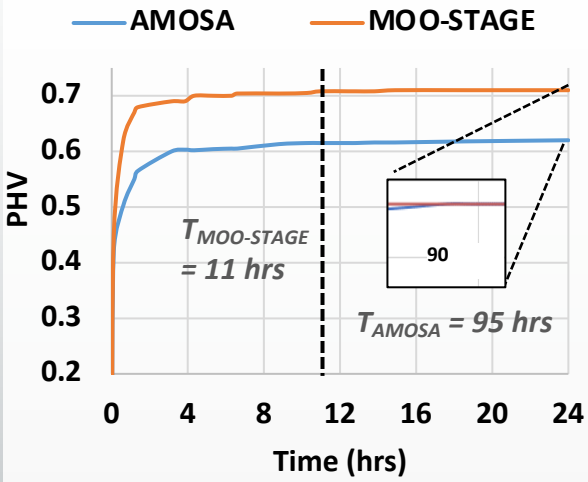

**Pareto Hypervolume (PHV)**

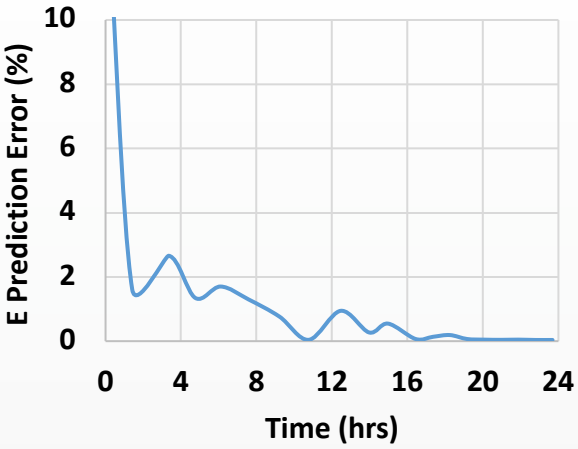
- Pareto Hyper Volume (PHV) measures solution quality
- Key Idea: *Learn* evaluation function to find better solutions in less time

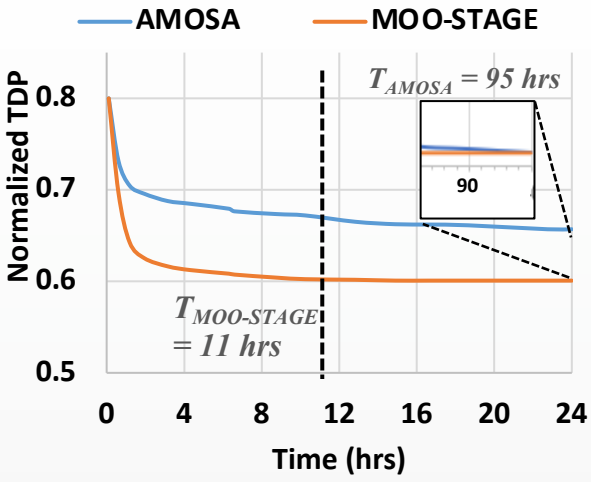

B. K. Joardar et. al., "Learning-Based Application-Agnostic 3D NoC Design for Heterogeneous Manycore Systems," in IEEE TC, vol. 68, no. 6, pp. 852-866, 1 June 2019

J.A. Boyan and A.W. Moore, "Learning Evaluation Functions to Improve Optimization by Local Search," *JMLR*, 2000.

# MOO-STAGE: Performance




- 9x Speedup in optimization time
- Evaluation function (E) prediction error quickly moves towards 0
- Better Pareto Hypervolume (PHV) [higher] results in lower thermal delay product (TDP)

B.K. Joardar, "Learning-based Application-Agnostic 3D NoC Design for Heterogeneous Manycore Systems, *IEEE TC*, 2019.

# MOO-STAGE: Performance (Cont.)

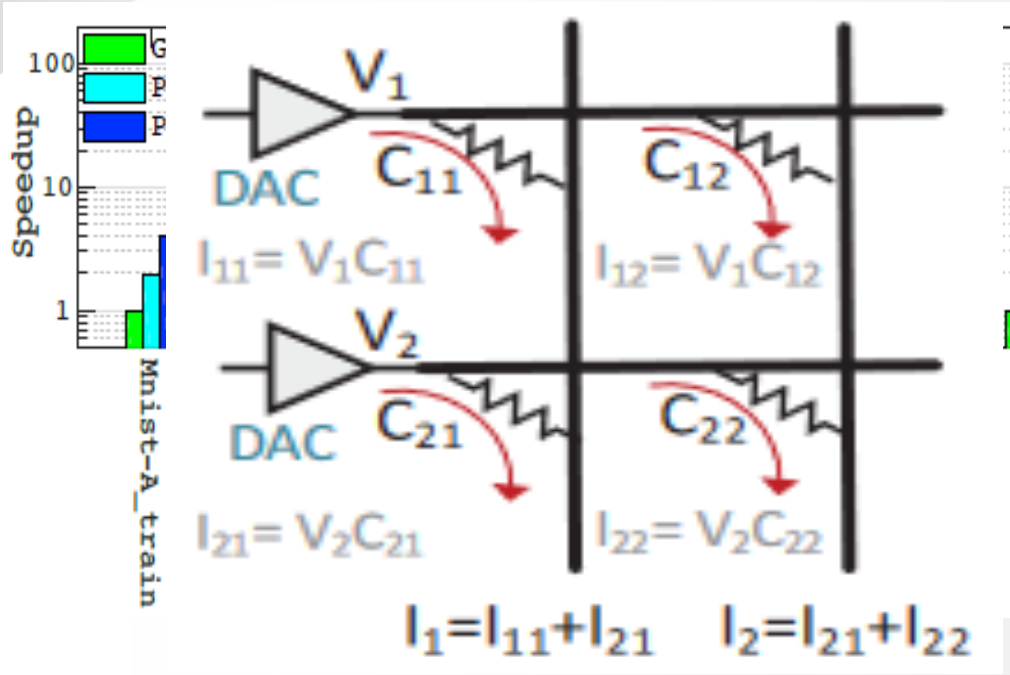
**MOO-STAGE speed-up over AMOSA**

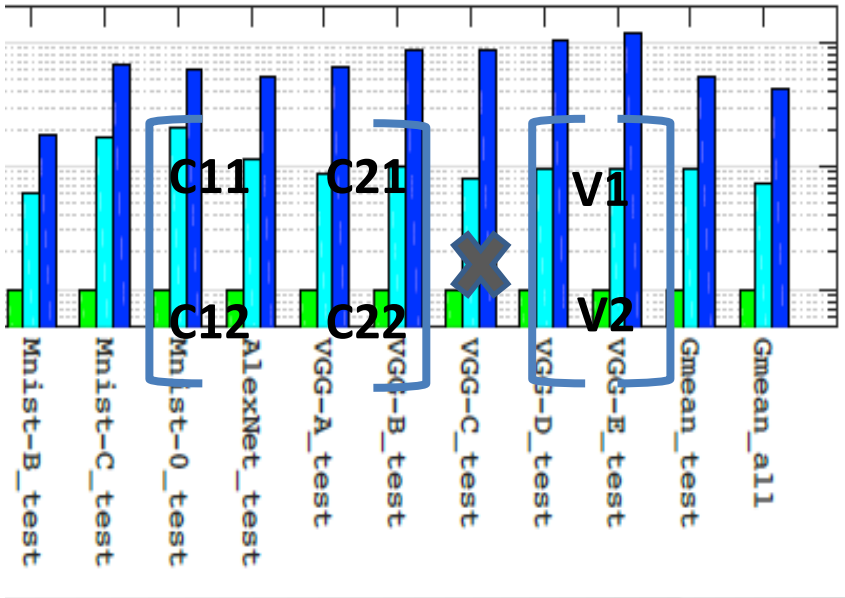
| Applications | Two-obj | Three-obj | Four-obj |
|---|---|---|---|
| BP | 1.5 | 6.4 | 12.5 |
| BFS | 2 | 5 | 9.4 |
| CDN | 1.5 | 5.8 | 13.7 |
| GAU | 1.3 | 6 | 7.2 |
| HS | 1.5 | 8 | 10 |
| LEN | 2 | 5.8 | 14.2 |
| LUD | 1.3 | 5 | 10 |
| NW | 1.5 | 5 | 11.4 |
| KNN | 1.2 | 6.4 | 7.5 |
| PF | 1.2 | 5 | 11.4 |
| Average | 1.5 | 5.84 | 10.7 |

Rodinia Benchmark Suite

- MOO-STAGE speed-up over AMOSA improves as objectives increase
- AMOSA *never* found the best solution like MOO-STAGE for three- and four-objectives

# ReRAM vs GPU for deep learning



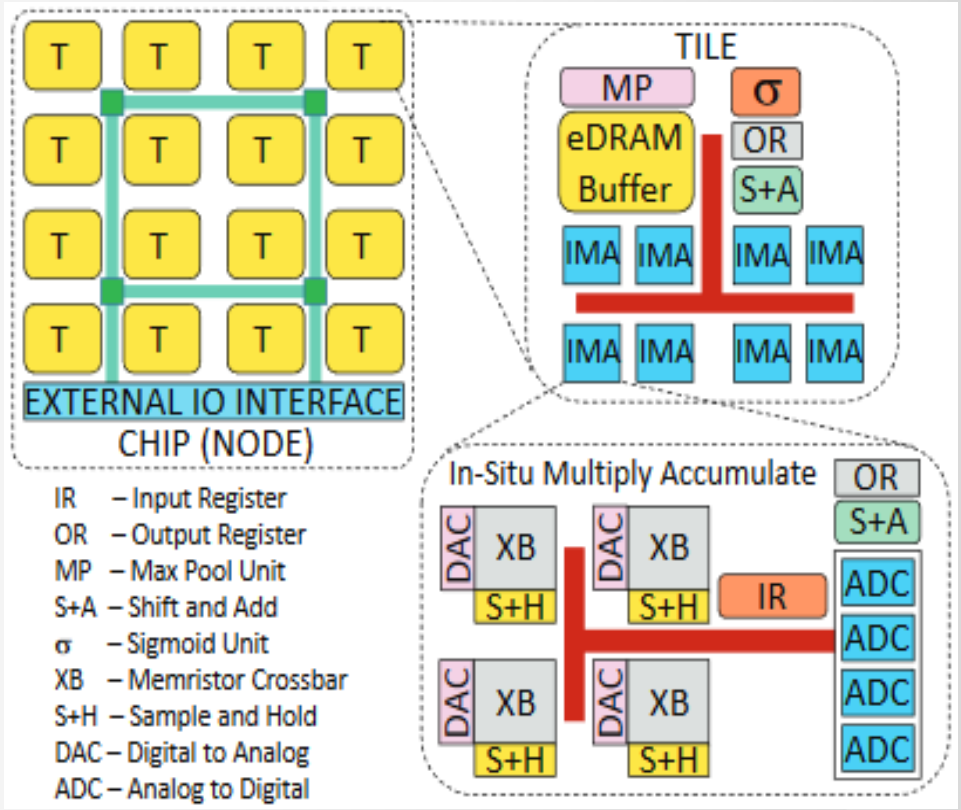

- ReRAMs more efficient for matrix multiplications
  - High performance
    - O(1) time complexity
  - Energy efficient
  - Low area


L. Song et al., "PipeLayer: A pipelined ReRAM-based accelerator for deep learning", HPCA, 2017.

D. Fujiki, S. Mahlke, and R. Das, “In-Memory Data Parallel Processor,”. In Proc. of ASPLOS, 2018. NY, USA, 1-14.

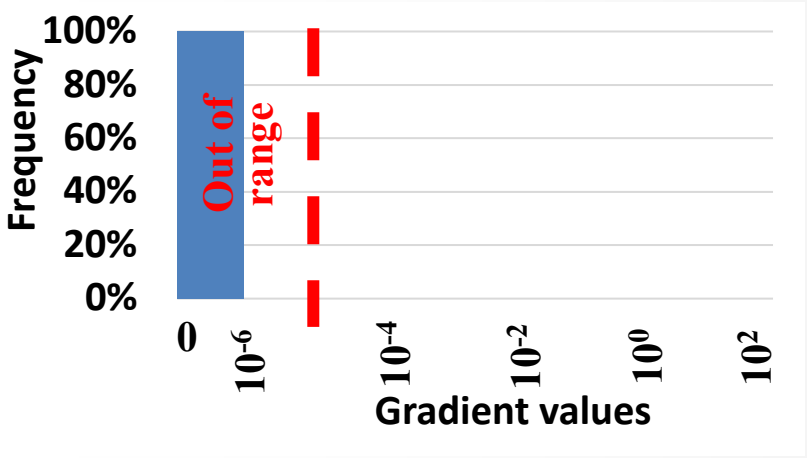
# Challenges with existing ReRAM architectures

- Low Precision
  - Accuracy Loss
  - Unstable training
- Lack of Normalization
  - Requires full-precision
- Temperature dependence
  - Change in resistance
  - Thermal noise
  - Low noise margin

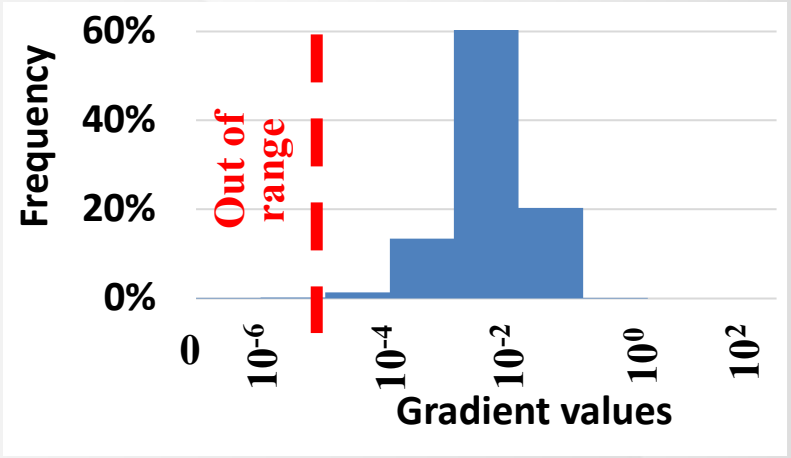



A. Shafiee et al., "ISAAC: A Convolutional Neural Network Accelerator with In-Situ Analog Arithmetic in Crossbars," 2016 ACM/IEEE 43rd ISCA, Seoul, 2016, pp. 14-26.

# Low precision and lack of Normalization




*No Normalization support*

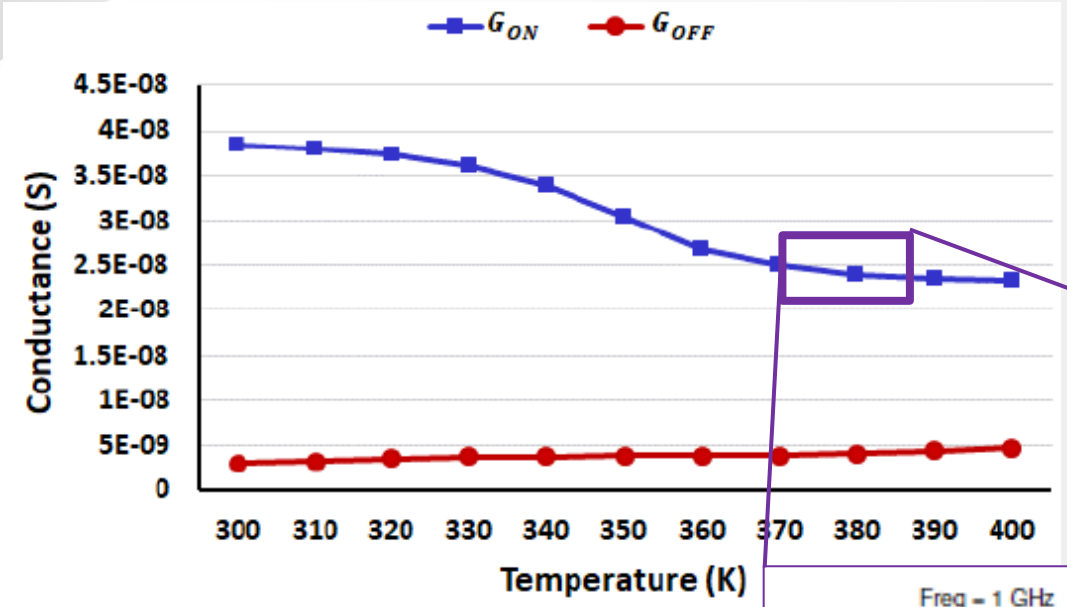



*With Normalization support*

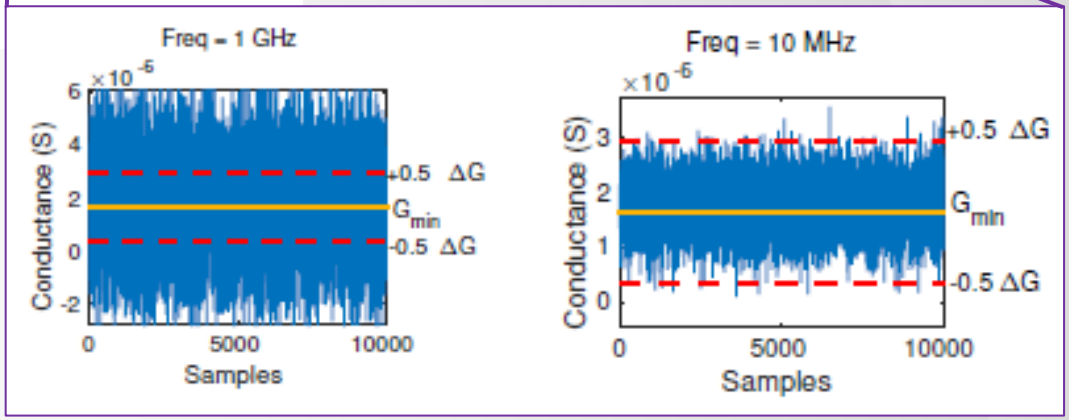
$$w_{new} = w_{old} - \alpha * \Delta w$$

- Gradients too small without normalization
- Low precision cannot represent too small/large values
  - Gradients rounded to zero
  - No (or minimal) weight update
    - No meaningful learning

# Thermal challenges




- Resistance changes with temperature
- Reduces noise margin

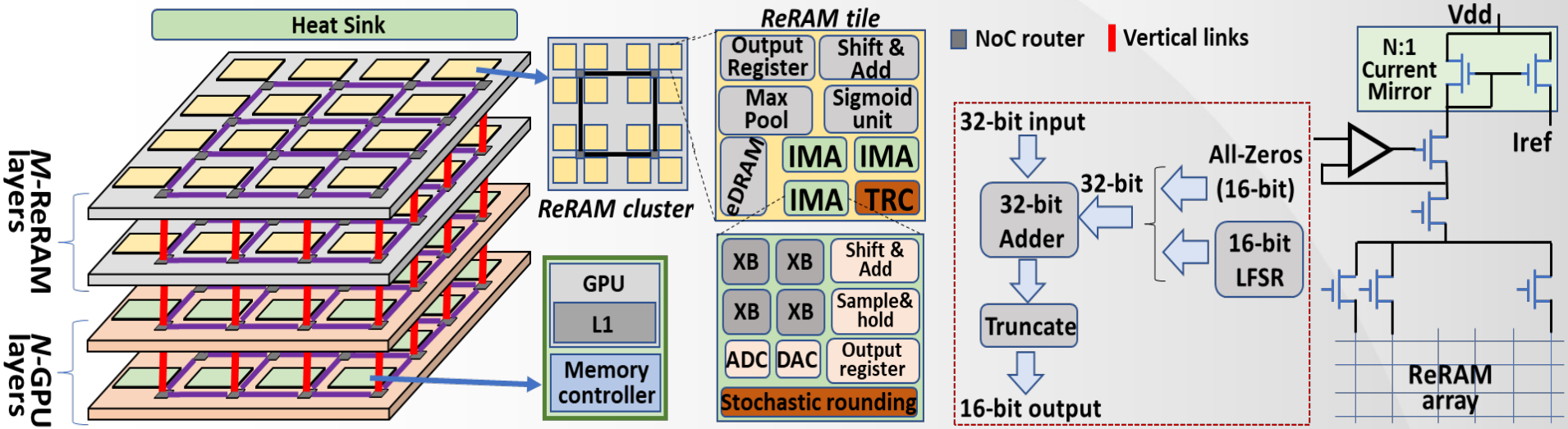



- Thermal noise can lead to misinterpretation
- Noise increases with frequency

$$g_{rms} = \frac{i_{rms}}{V} = \sqrt{\frac{Gf(4K_BT + 2qV)}{V^2} + \left(\frac{\Delta G}{3}\right)^2}$$


Z. He et. al., "Noise Injection Adaption: End-to-End ReRAM Crossbar Non-ideal Effect Adaption for Neural Network Mapping," 2019 DAC, Las Vegas, NV, USA, 2019, pp. 1-6

# AccuReD: ReRAM/GPU-based heterogeneous architecture

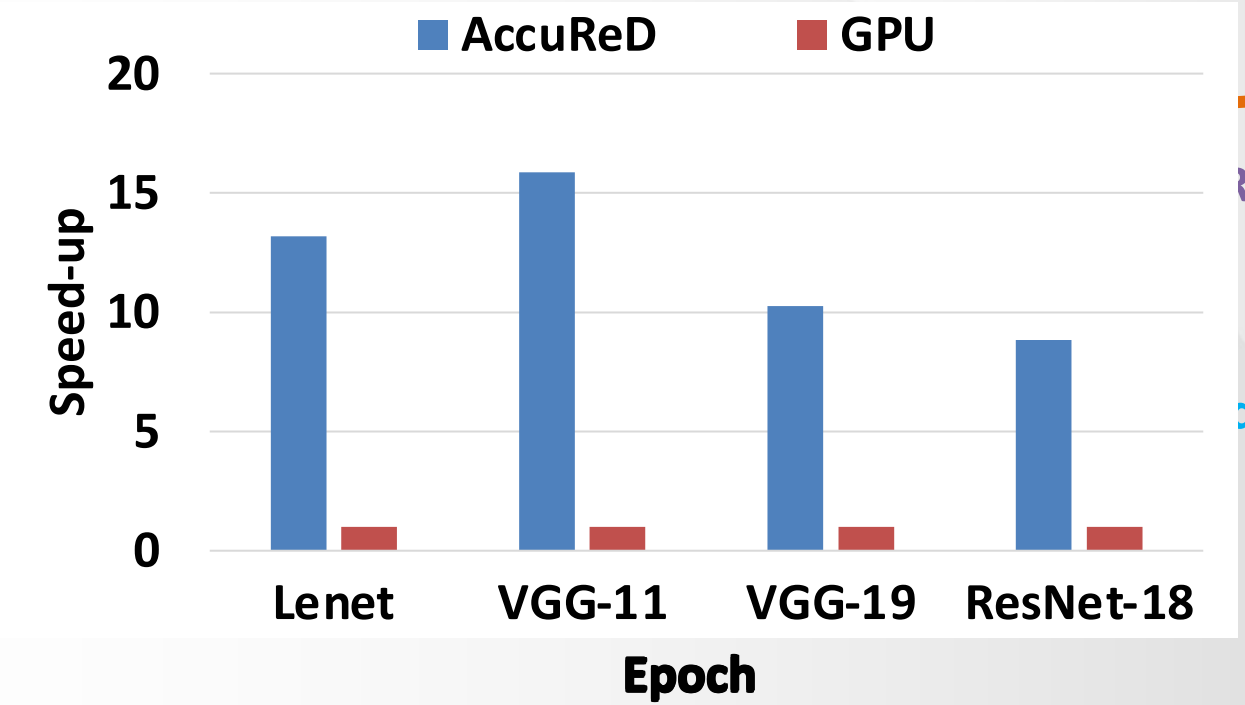



- Normalization:
  - Full precision GPUs
- Low Precision:
  - Stochastic rounding
- Temperature:
  - Thermal reference cell
  - Performance-Thermal aware mapping
  - M3D integration


- B. K. Joardar, J. R. Doppa, P. P. Pande, H. Li and K. Chakrabarty, "AccuReD: High Accuracy Training of CNNs on ReRAM/GPU Heterogeneous 3-D Architecture," in *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 40, no. 5, pp. 971-984, May 2021

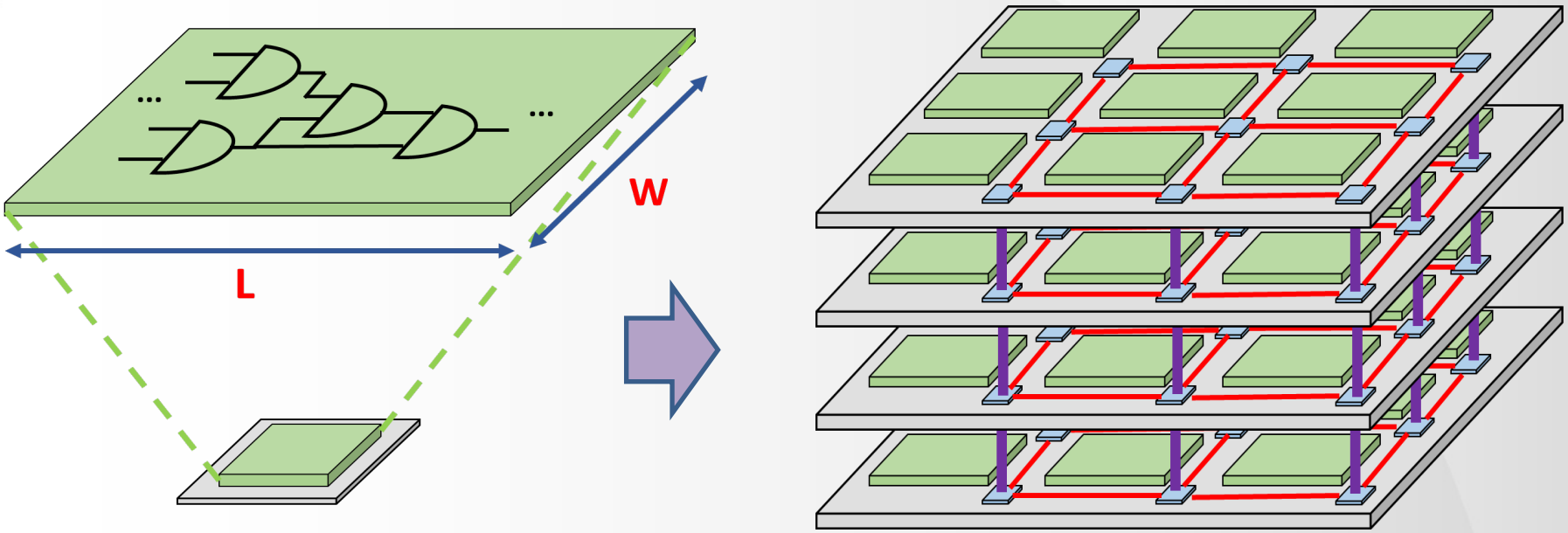
- B. K. Joardar, A. Deshwal, J. R. Doppa, P. P. Pande and K. Chakrabarty, "High-Throughput Training of Deep CNNs on ReRAM-Based Heterogeneous Architectures via Optimized Normalization Layers," in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 41, no. 5, pp. 1537-1549, May 2022

# Performance and Accuracy



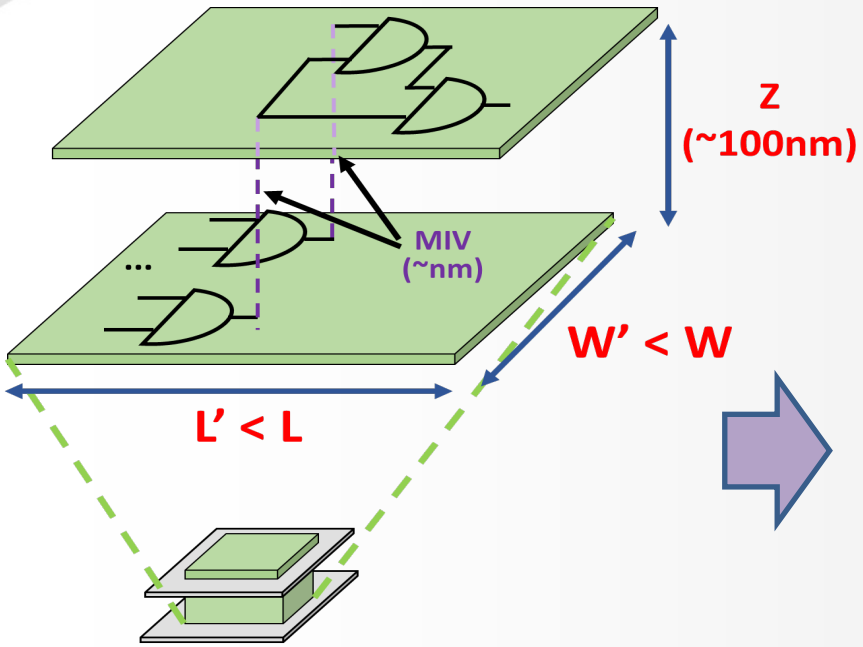

- AccuReD achieves GPU-level accuracy
  - Normalization, performance-thermal aware mapping, M3D
- Up to 15X speed-up compared to GPUs
  - ReRAMs are more efficient as dot-product accelerators

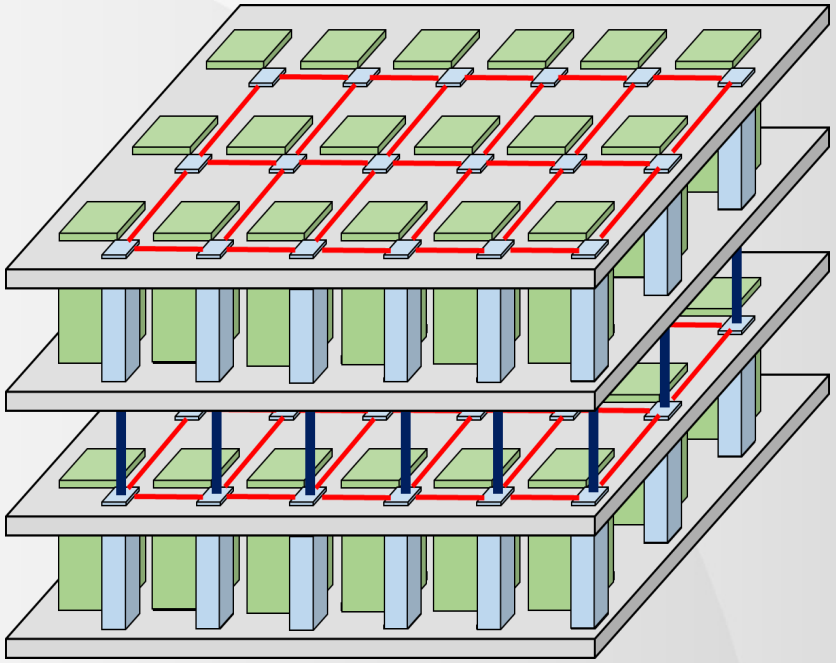
# Manycore design using TSV




- Conventional hardware design is planar
  - Sub-optimal power-performance
- Planar logic blocks stacked physically to create 3D

# Manycore design using M3D

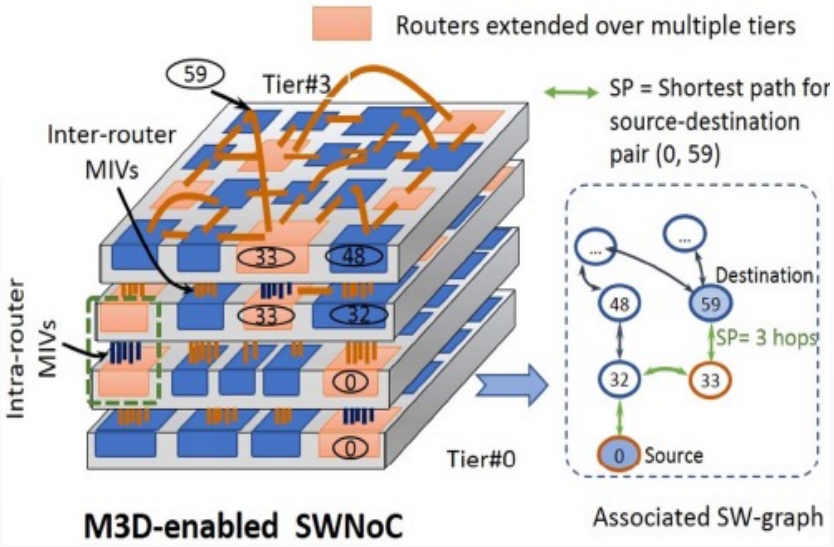


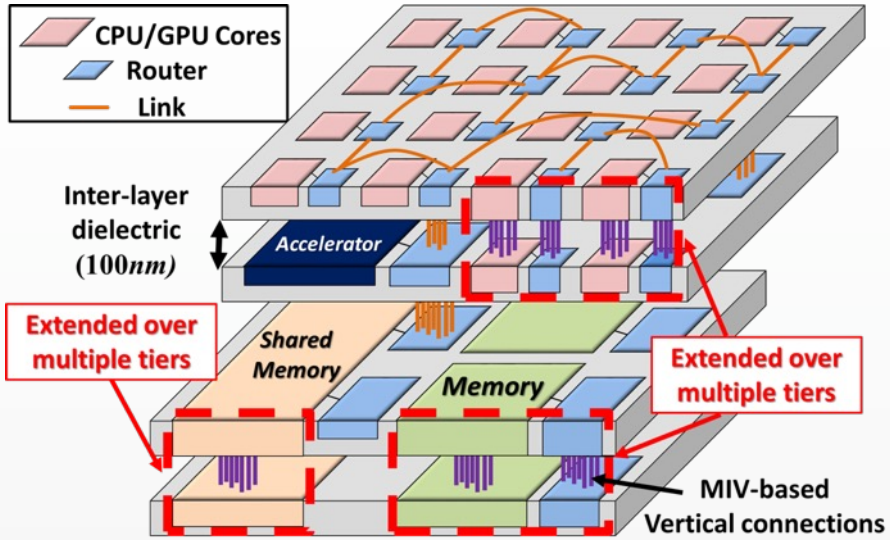



- M3D enables 3D hardware blocks
  - Less area, power
  - Better performance
- Logic blocks can span multiple tiers
- How to optimize the placements?

B. Gopireddy and J. Torrellas, "Designing vertical processors in monolithic 3D," in Proc. 46th Int. Symp. Computer Architecture (ISCA), Phoenix, AZ, USA, Jun. 2019, pp. 643–656.

# NoC Design Using M3D

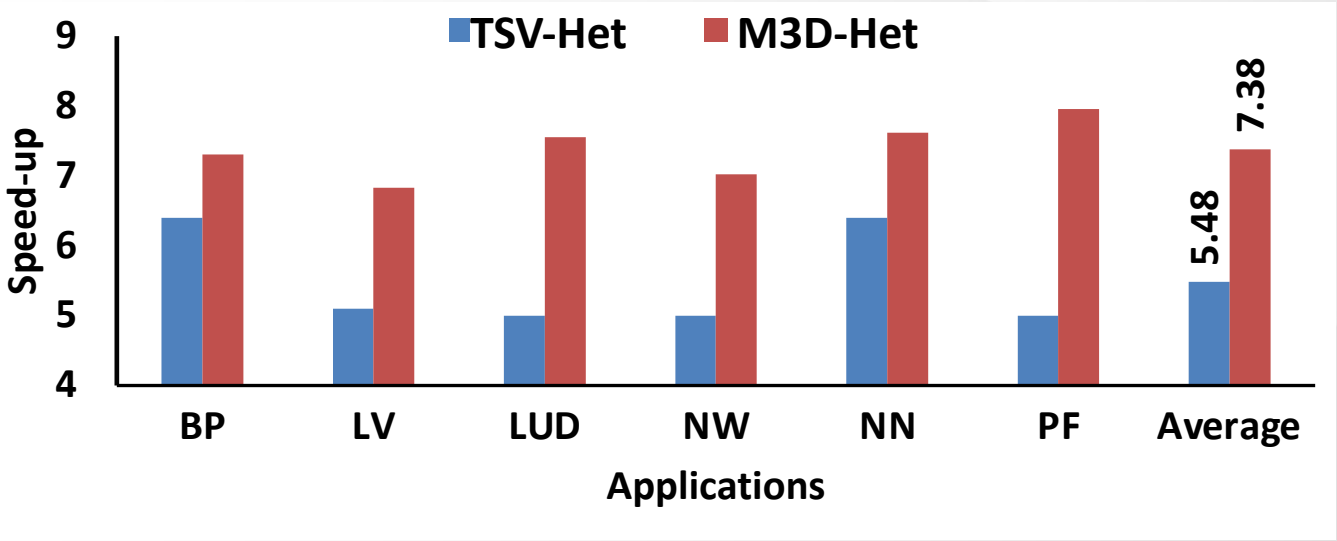



- Routers extended over multi-tiers
- Lower average hop count, shorter communication paths
- MIVs more energy-efficient than TSVs

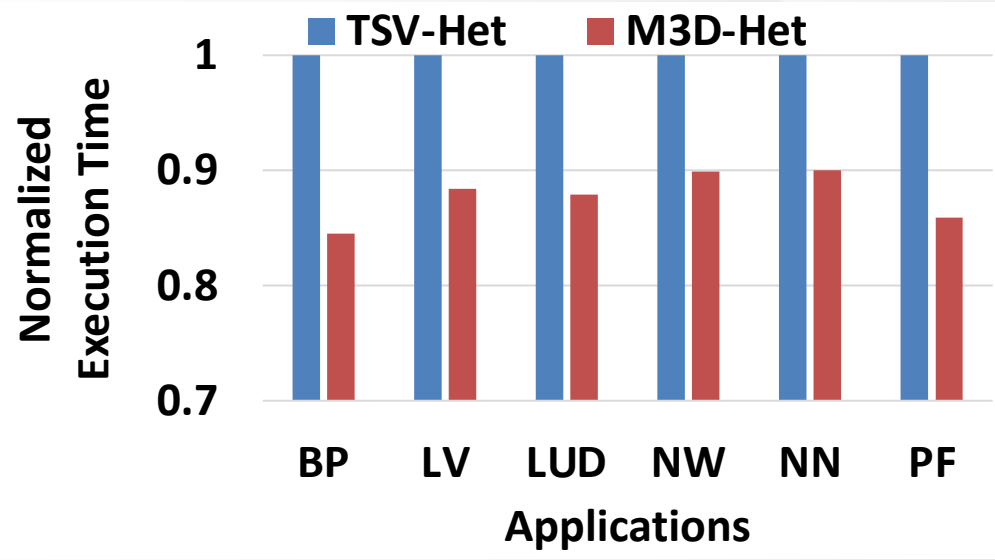
D. Lee et al., "Performance and Thermal Tradeoffs for Energy-Efficient Monolithic 3D Network-on-Chip," *ACM TODAES*, 2018.

# MOO-STAGE for M3D




- Design space of M3D larger than TSV
  - More design options
- MOO-STAGE performs even better
  - 5.48X speed-up in TSV
  - 7.38X speed-up in M3D
  - AMOSA needs even more time to find good solutions

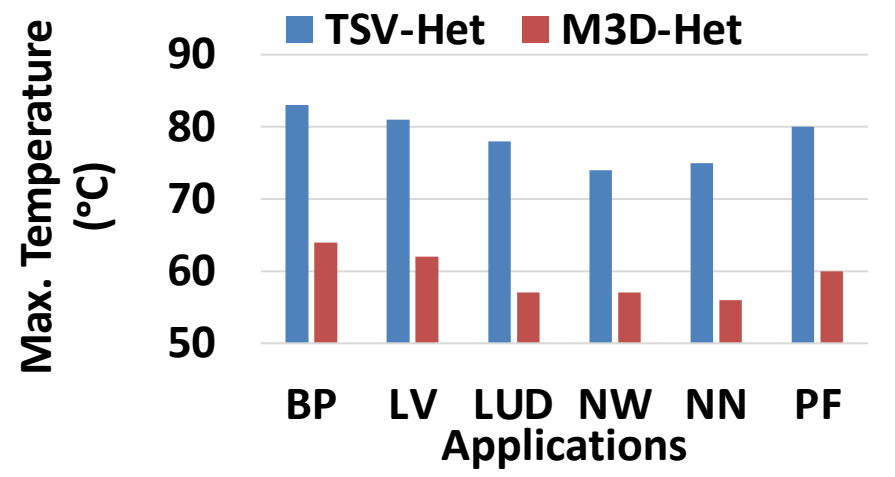
# M3D vs TSV: Performance




- M3D-Het is 12.3% faster than TSV-Het
  - M3D-based designs have lower wirelengths
    - Lower critical path delays
- Higher clock frequency can be used

B. Gopireddy and J. Torrellas, "Designing vertical processors in monolithic 3D," in Proc. 46th Int. Symp. Comput. Archit., Phoenix, AZ, USA, Jun. 2019, pp. 643–656.

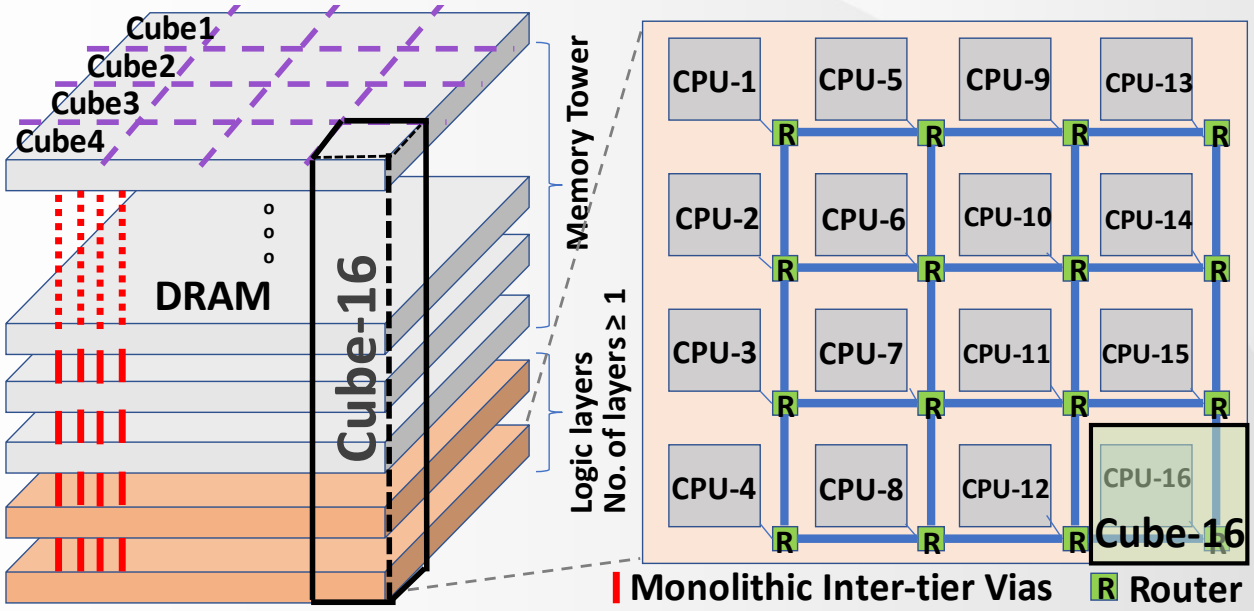
# M3D vs TSV: Temperature




- M3D-Het is 19°C cooler than TSV-Het
  - M3D-based designs are power efficient
    - Lower wirelength
    - Fewer number of buffers
  - Absence of bonding material
  - Smaller dimensions

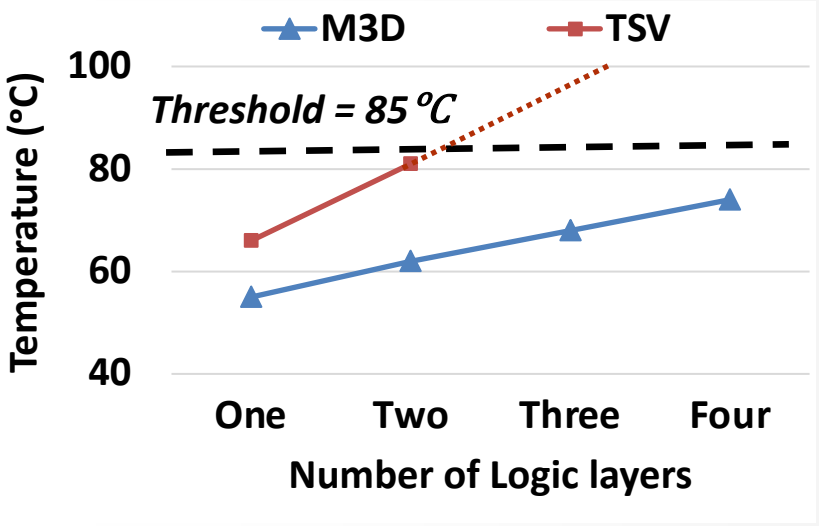
Aqeeb Iqbal Arka, Biresh Kumar Joardar, Ryan Gary Kim, Dae Hyun Kim, Janardhan Rao Doppa, and Partha Pratim Pande. 2021. HeM3D: Heterogeneous Manycore Architecture Based on Monolithic 3D Vertical Integration. ACM Trans. Des. Autom. Electron. Syst. 26, 2, Article 16 (March 2021), 21 pages.

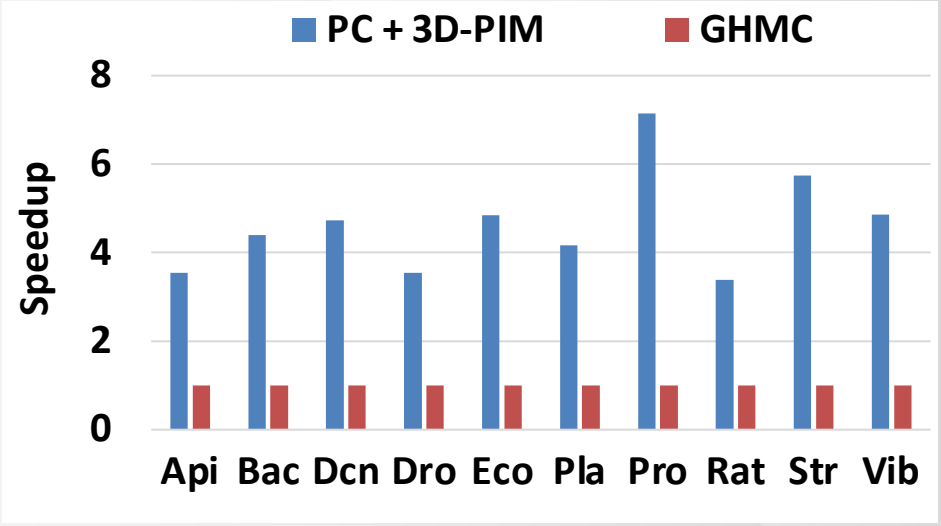
# M3D-based PIM




- Good for memory intensive applications
  - Memory closer to computation
- Conventional TSV-PIM limited to one logic layer
  - DRAM retention falls drastically beyond 85°C
    - Higher refresh rates offset any benefit
- M3D allows stacking multiple logic layers

# M3D-PIM: Thermal and Performance

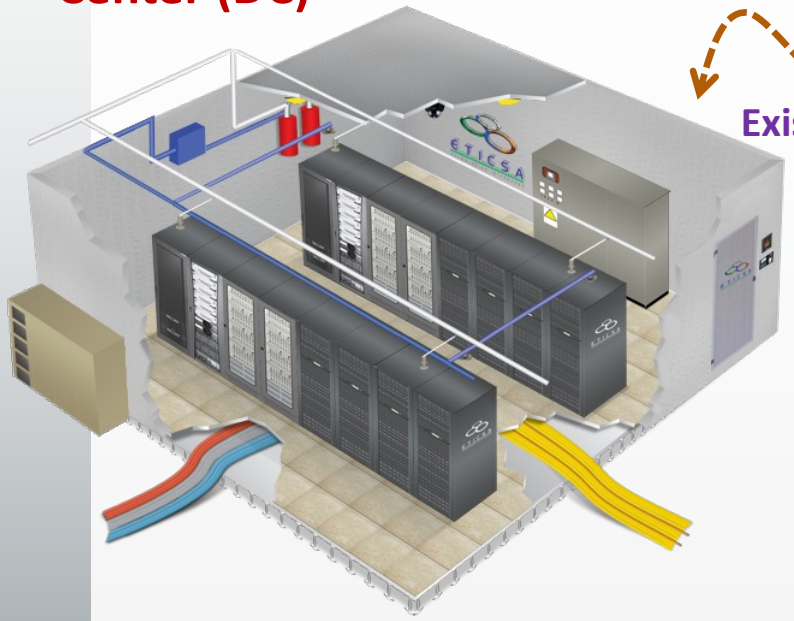



- M3D allows up to four logic layers
  - Not possible using TSV
- Up to 7X speed-up for *k-mer counting*
  - Evaluated using real-world gene sequences

# Evolution of NoC to DoC

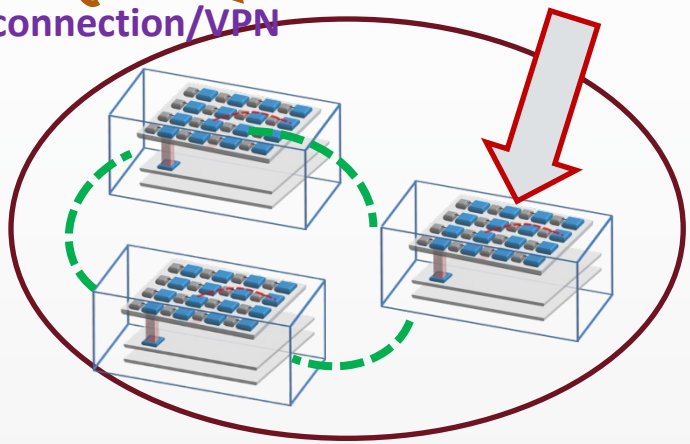
**Conventional Data Center (DC)**



**In the future, the DoC can, in principle, interact w/ conventional DC**

**3D WiDoC**

**Existing Internet connection/VPN**



**Network of WiDoCs (future, based on proposed work)**

**Size: tens of thousands sq. feet**
**Power: Mega Watts, Cooling cost: Huge**
**Users: millions (mostly HPC)**
**Maintenance Cost (wiring, utility service providers)**

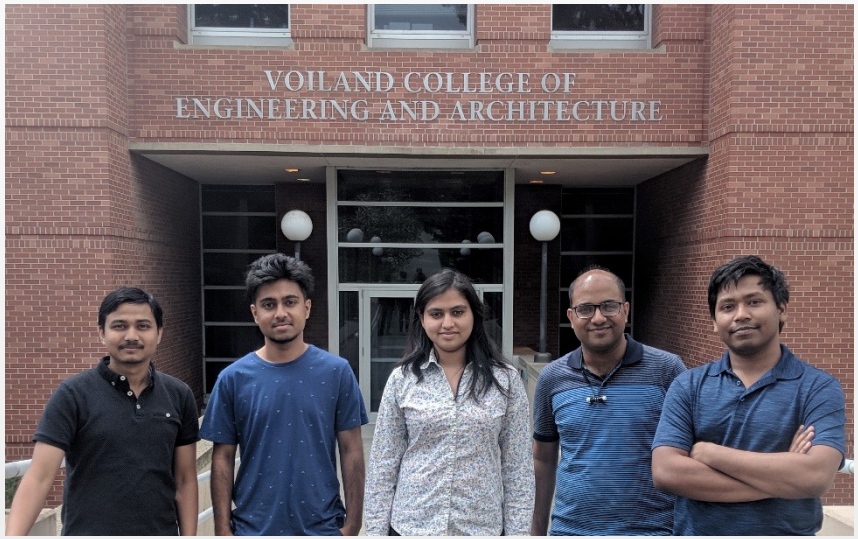
**Size: tens of $cm^2$**
**Power: 100 Watts, Cooling cost: minimal**
**Users: 1-10 (mostly personal/mobile computing)**

# Conclusion

- Finding optimal designs is getting more difficult due to greater system complexities:
  - More cores (System size)
  - More core types (Heterogeneity)
  - Emerging interconnects (3D: TSV, M3D, Wireless, NFIC)
  - Application-specific HW and deployment scenarios (IoT, datacenter, etc.)
- Need ML-based EDA techniques
  - Conventional techniques can’t keep up
- Altogether, emerging interconnect-based heterogeneous manycore systems show promise
  - Higher performance, Lower Thermals
- Evolution of DoC

# Acknowledgements


- Collaborators
  - Radu Marculescu, CMU
  - Diana Marculescu, CMU
  - Deukhyoun Heo, WSU
  - Janardhan Rao (Jana) Doppa, WSU
  - Krishnendu Chakrabarty, Duke University
  - Paul Bogdan, USC
- Graduate students
- Funding from NSF and DOD and Boeing


# Thank you


감사합니다.


GRAZIE
Grazie
Grazie
Grazie
Grazie
Grazie
Grazie
Grazie
Grazie
Grazie
Grazie


ধন্যবাদ।


merci


謝謝