Hybrid Network on Chip (HNoC): Local Buses with a Global Mesh Architecture

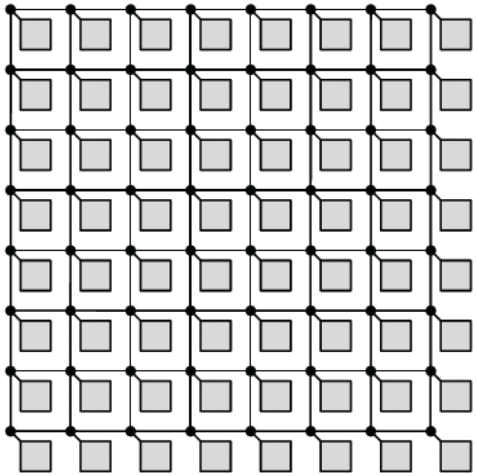
Payman Zarkesh-Ha, George B. P. Bezerra*, Stephanie Forrest*, and Melanie Moses*

University of New Mexico
Department of Electrical and Computer Engineering
** Department of Computer Science*
Albuquerque, NM 87131

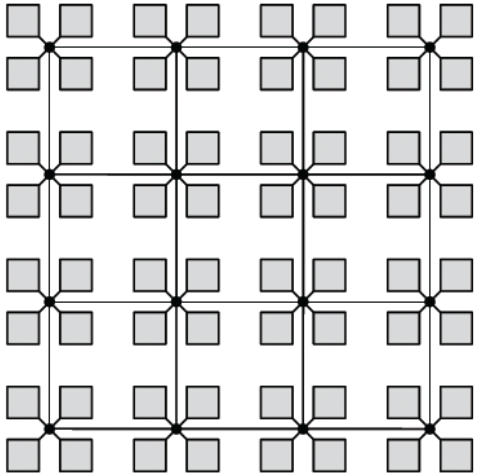
Outline

1. *Overview of Hybrid NoC Architectures*
2. *Proposed HNoC: Local Buses with Global Mesh*
3. *HNoC Throughput Assessment using Rent's Rule*
4. *HNoC Energy Assessment using Rent's Rule*
5. *Simulation Results*
6. *Discussions*

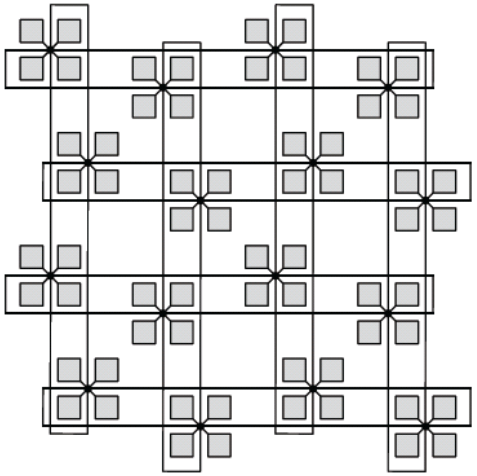
Some Network on a Chip Configurations



Mesh



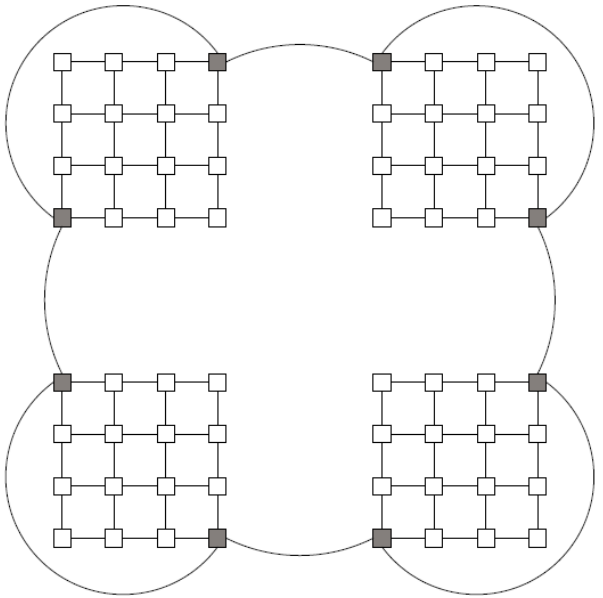
Concentrated Mesh



Concentrated Torus

Existing Hybrid NoC Configuration Chip 1

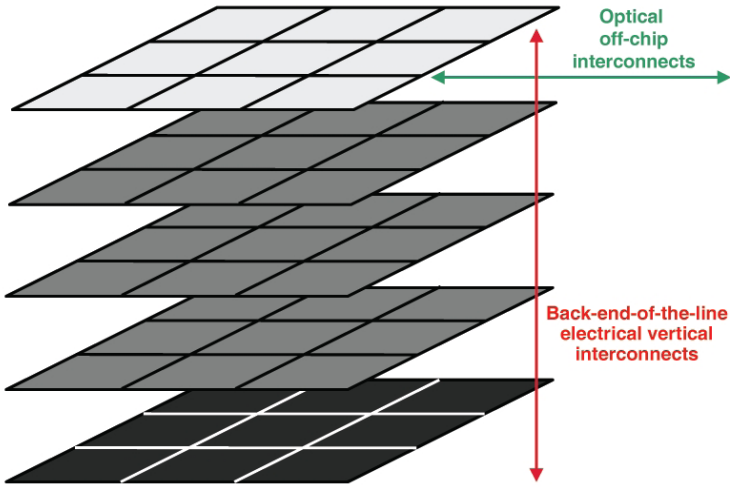
- Hybrid Mesh-Ring Topology
- Mesh is used for local network
- Ring is used for global network
- Improves throughput and reduces delay



V. Rantala et. al. [DAC ‘08]

Existing Hybrid NoC Configuration Chip 2

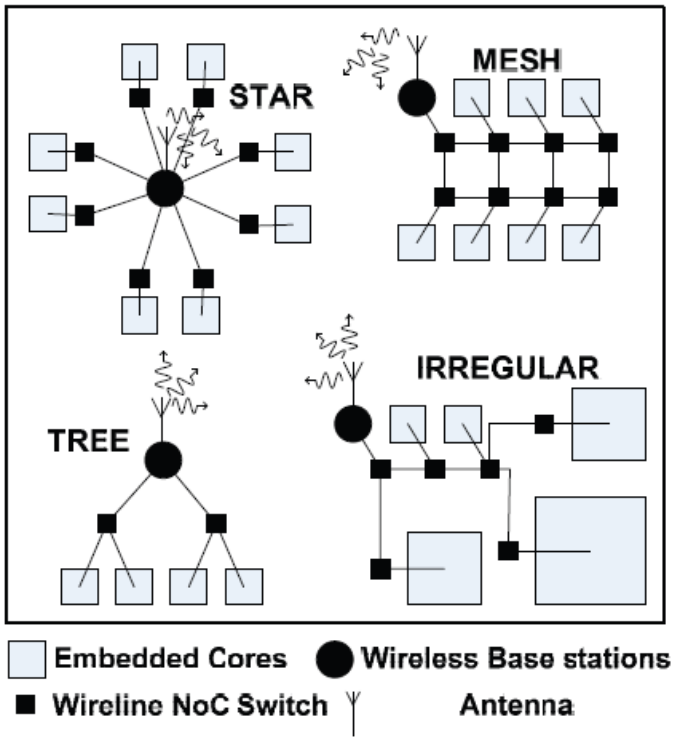
- Hybrid Photonic NoC
- Electronic network carries While the electronic network carries small-size control (and data) packets
- The photonic network transfers large-size data messages
- Can support energy-efficient high-bandwidth data transfers in 3D chips



A. Shacham et. al. [NOCS '07]

Existing Hybrid NoC Configuration Chip 3

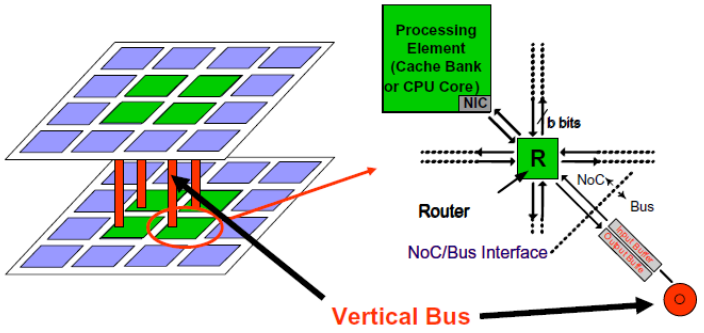
- Hybrid Wireless-Wired NoC Topology
- RF interconnect acts as an information ”highway” enabling fast data transport across longer distances on the chip
- Can potentially reduce overall communication power and latency



Y. Wang et. al. [ICCAD ‘07]

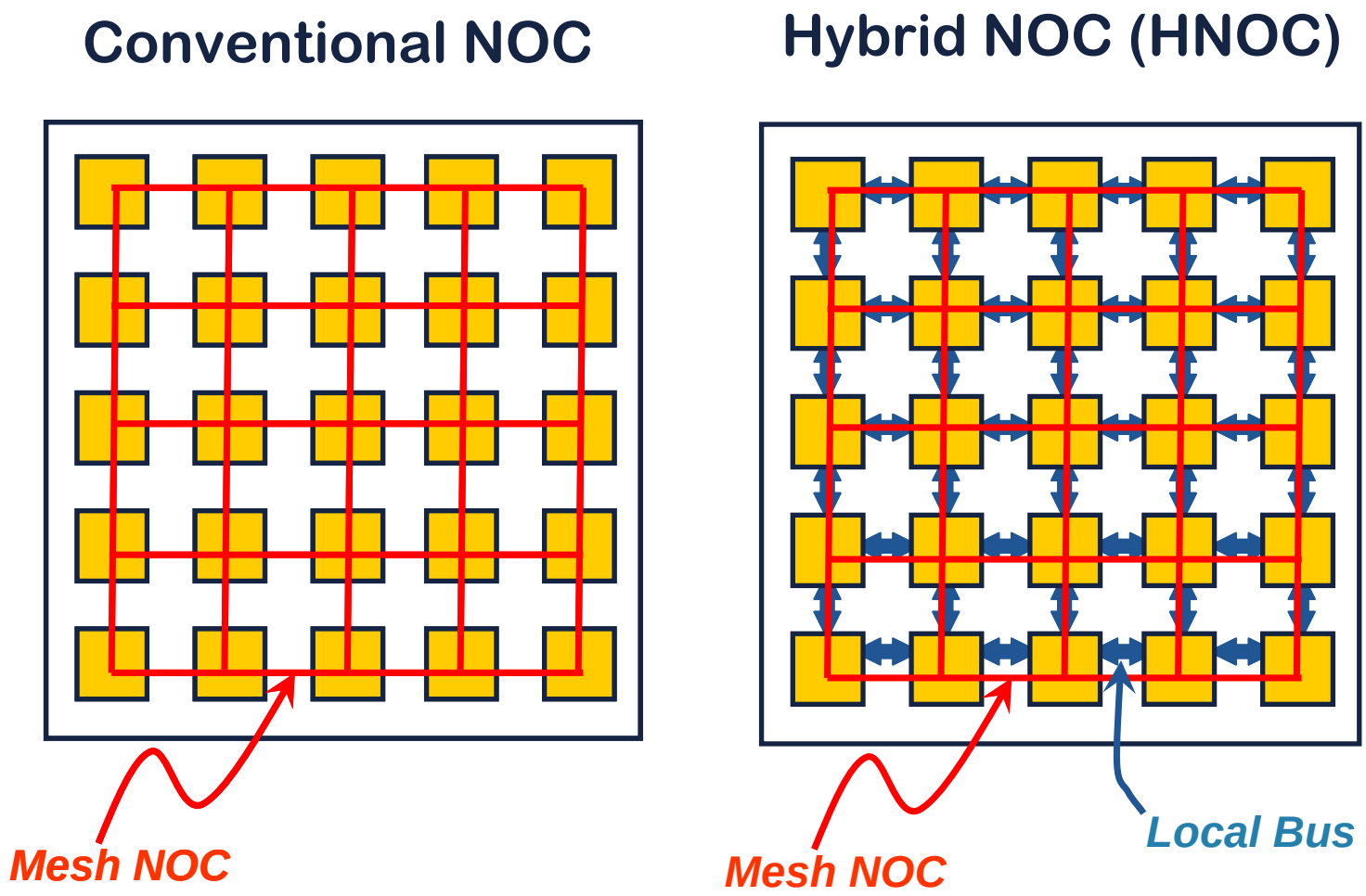
Existing Hybrid NoC Configuration Chip 4

- Hybrid 3D NoC-Bus Topology
- NoC on each stack
- Dedicated bus between stacks
- Provides performance and area benefits
- Good for two planes, but not efficient for multiple stacks



L. Carloni et. al. [NOCS '09]

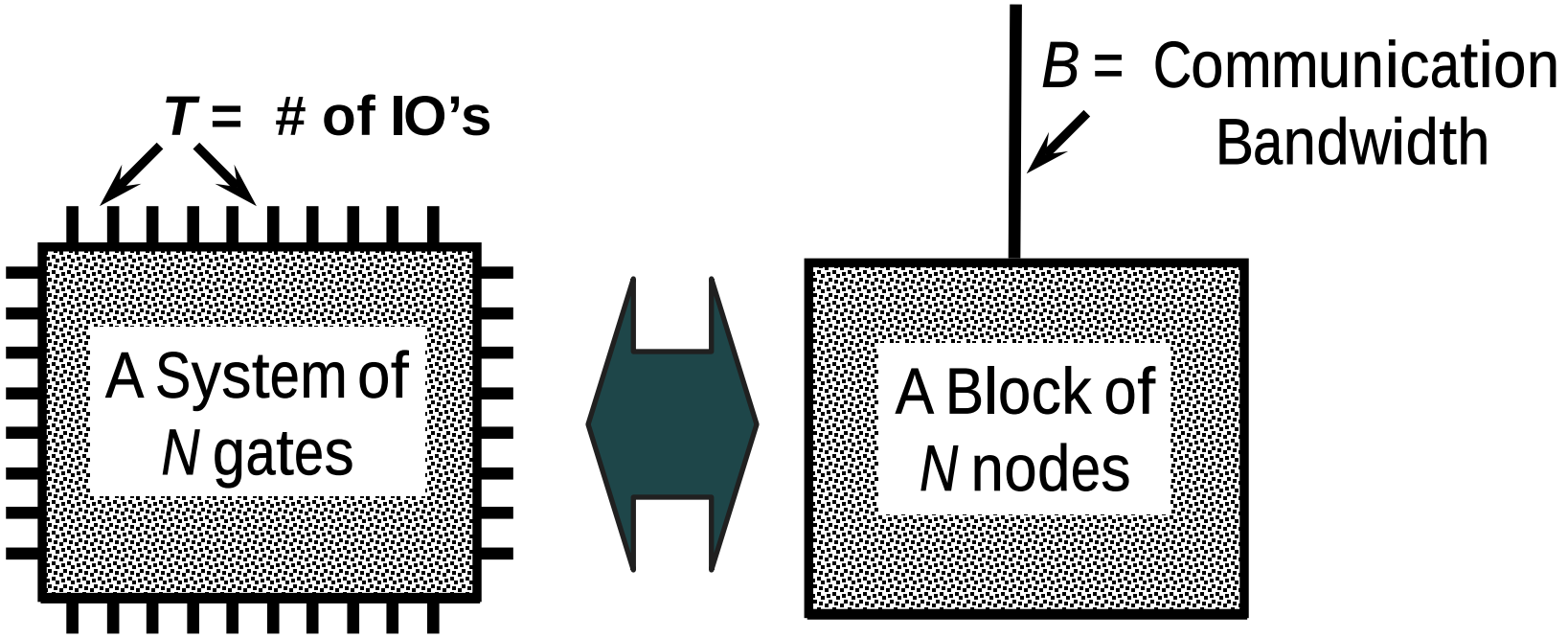
Proposed Hybrid Network on a Chip (HNoC)



Proposed Hybrid Network on a Chip (HNoC)

- Standard NoC topology (e.g., mesh) is used for packet-based global interconnections and local buses for nearest-neighbor communications
- HNoC uses local buses to transmit data directly to the nearest neighbors in a parallel fashion, which eliminates the need for serializer, router, and deserializer
- Since the local bus interconnects are short, they inherently exhibit lower loss and therefore can provide higher bandwidth and consume less power.

How to use Rent's Rule to Evaluate HNoC?



$$T = kN^{p}$$

k and *p* are empirical constants

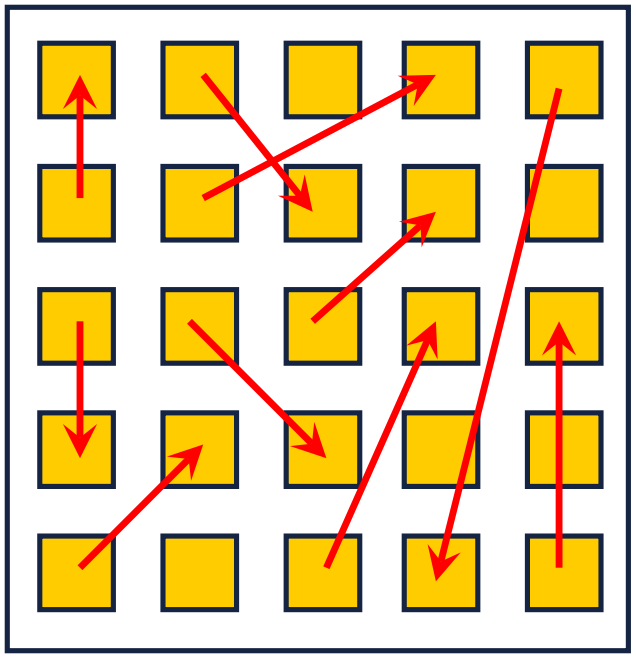
$$B = bN^{p}$$

b and *p* are empirical constants

Probability of Communication

- Probability of communication represents how probable is a block to communicate with another block within a certain distance

- Similar to wire length distribution model, communication probability distribution (CPD) can be derived using Rent's Rule



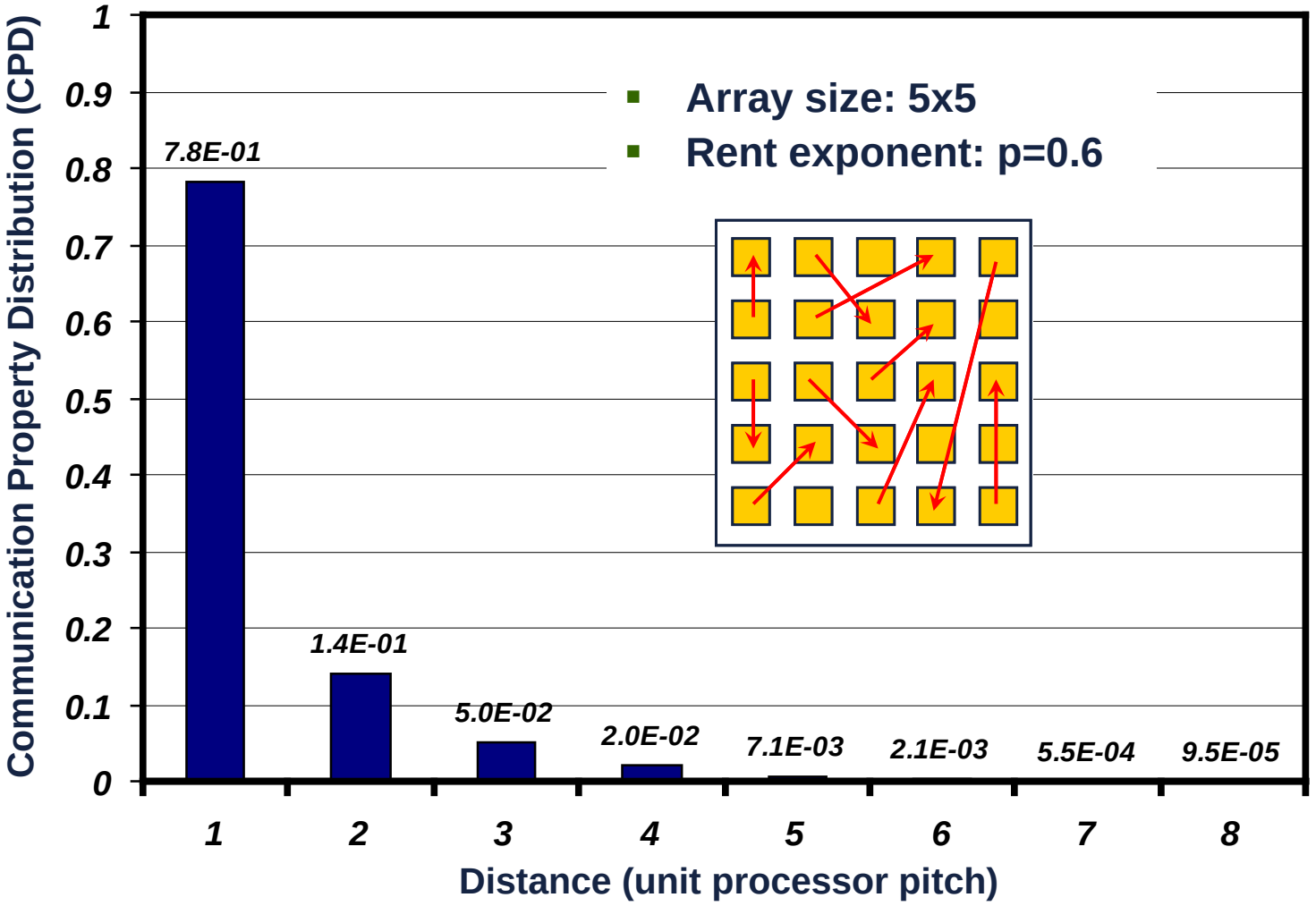
Communication Probability Distribution (CPD)

$$CPD(d) = \frac{\Gamma f(d)}{d}\left\{\left[1+d(d-1)\right]^p - \left[d(d-1)\right]^p + \left[d(d+1)\right]^p - \left[1+d(d+1)\right]^p\right\}$$

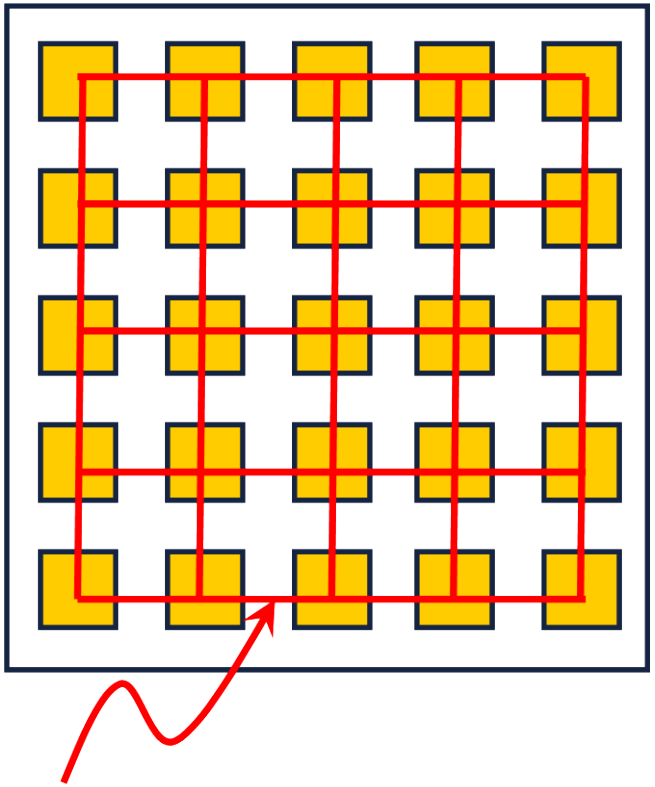
$$f(d) = \begin{cases} \frac{d^3}{3} - 2d^2N + \frac{d}{3}\left(6N^2-1\right), & 1 \le d < N \\ -\frac{d^3}{3} + 2d^2N - \frac{d}{3}\left(12N^2-1\right) + \frac{2}{3}N\left(4N^2-1\right), & N \le d \le 2N-2 \end{cases}$$

- CPD is the communication probability of distance d in an N×N multiprocessor system, p is the Rent's exponent, and Γ is the normalization coefficient.

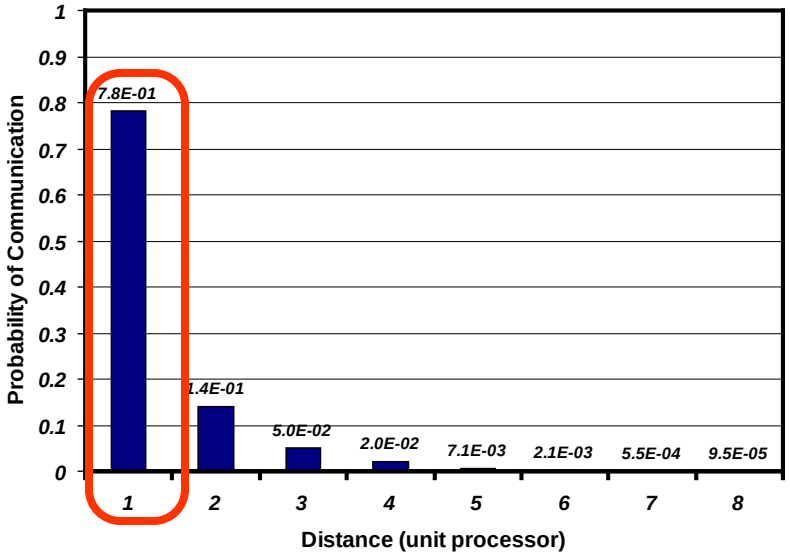
Example: Probability of Communication



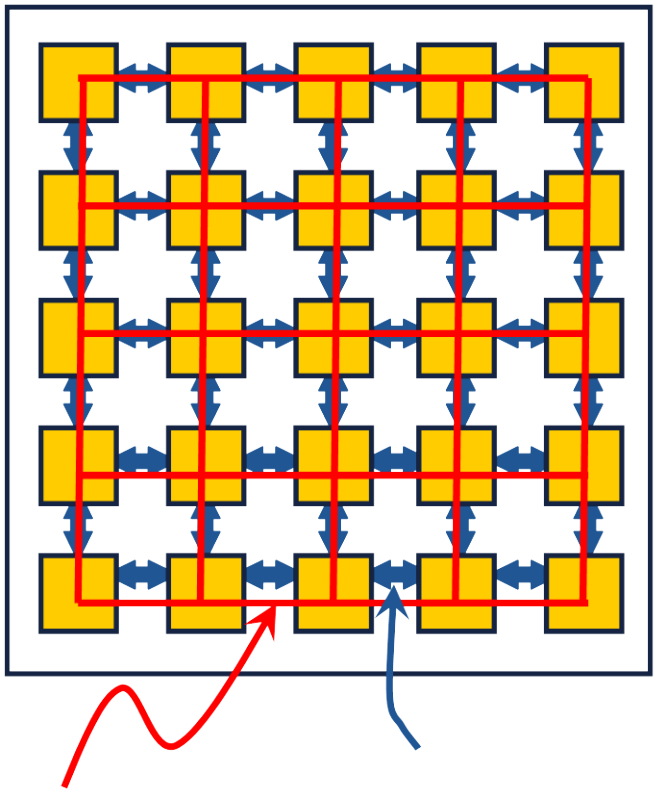
Observation from CPD



- **In an standard NOC architecture of 5x5 processor array, about 78% of communications are with the nearest neighbors!**
- **If the nearest neighbor communication can be lifted from the network traffic, it can tremendously improve the NOC bandwidth utilization and its throughput**

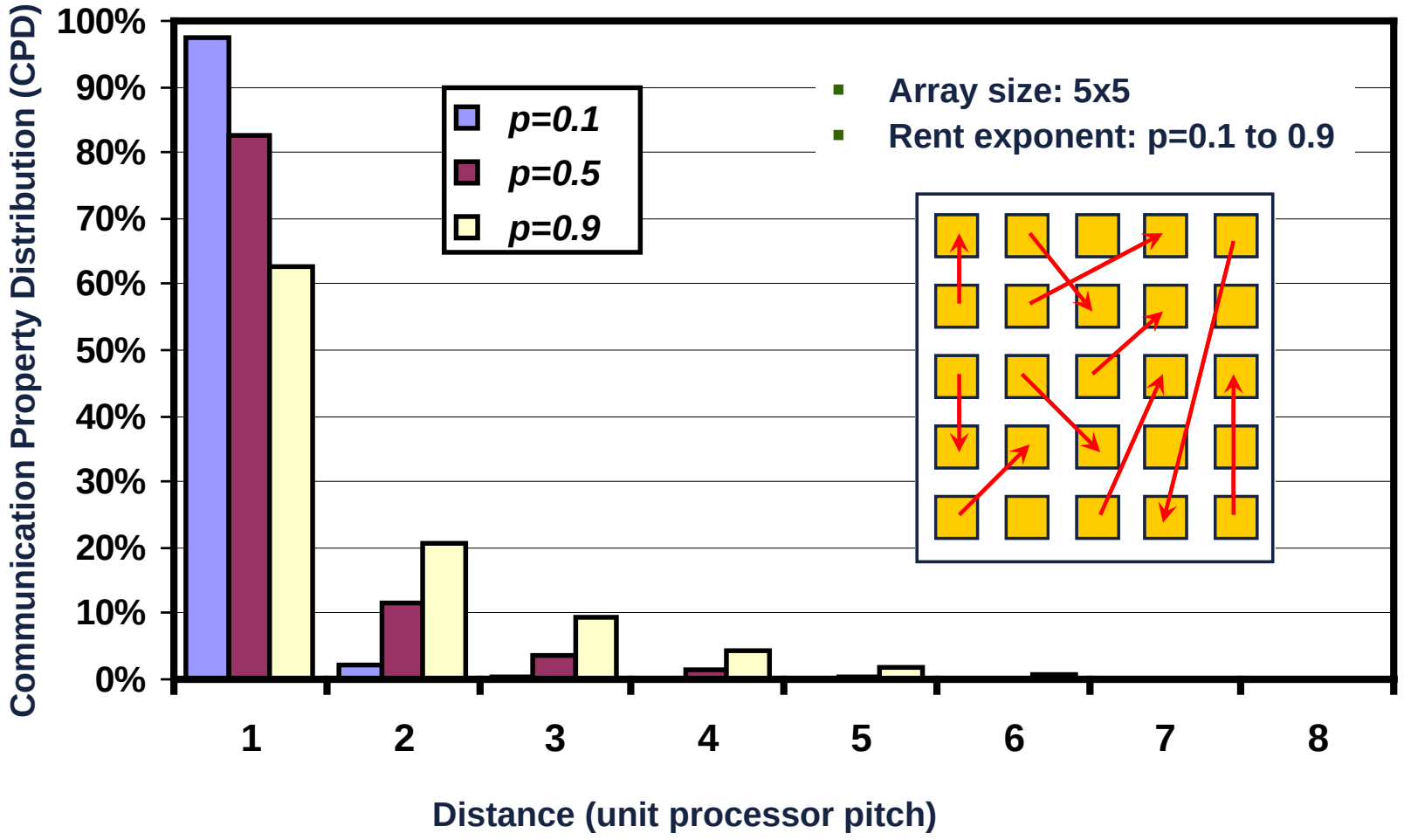


Observation from CPD

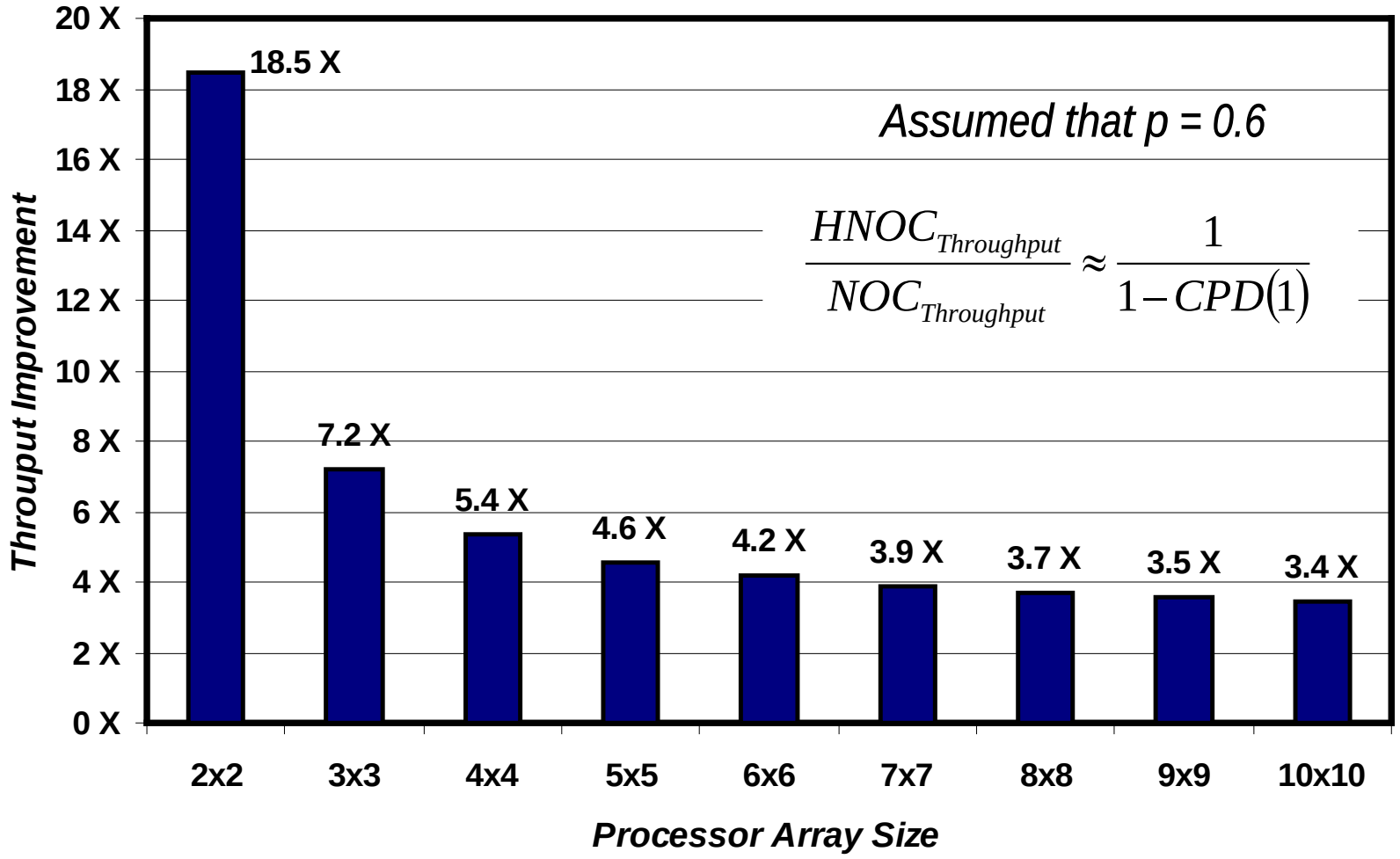


- The proposed HNoC uses a “local bus” for the nearest neighbor communications and the global NOC for the longer communication only
- Therefore, the system will consist of two hybrid networks; 1) local bus (like in systolic arrays), and 2) global mesh NOC
- This is similar to the concept of “local wires” and “global wires” in standard VLSI designs
- Assuming that the throughput is still limited by the global NOC, the hybrid NOC can potentially improve throughput by a factor of 4.6X in a system of 5x5 processors

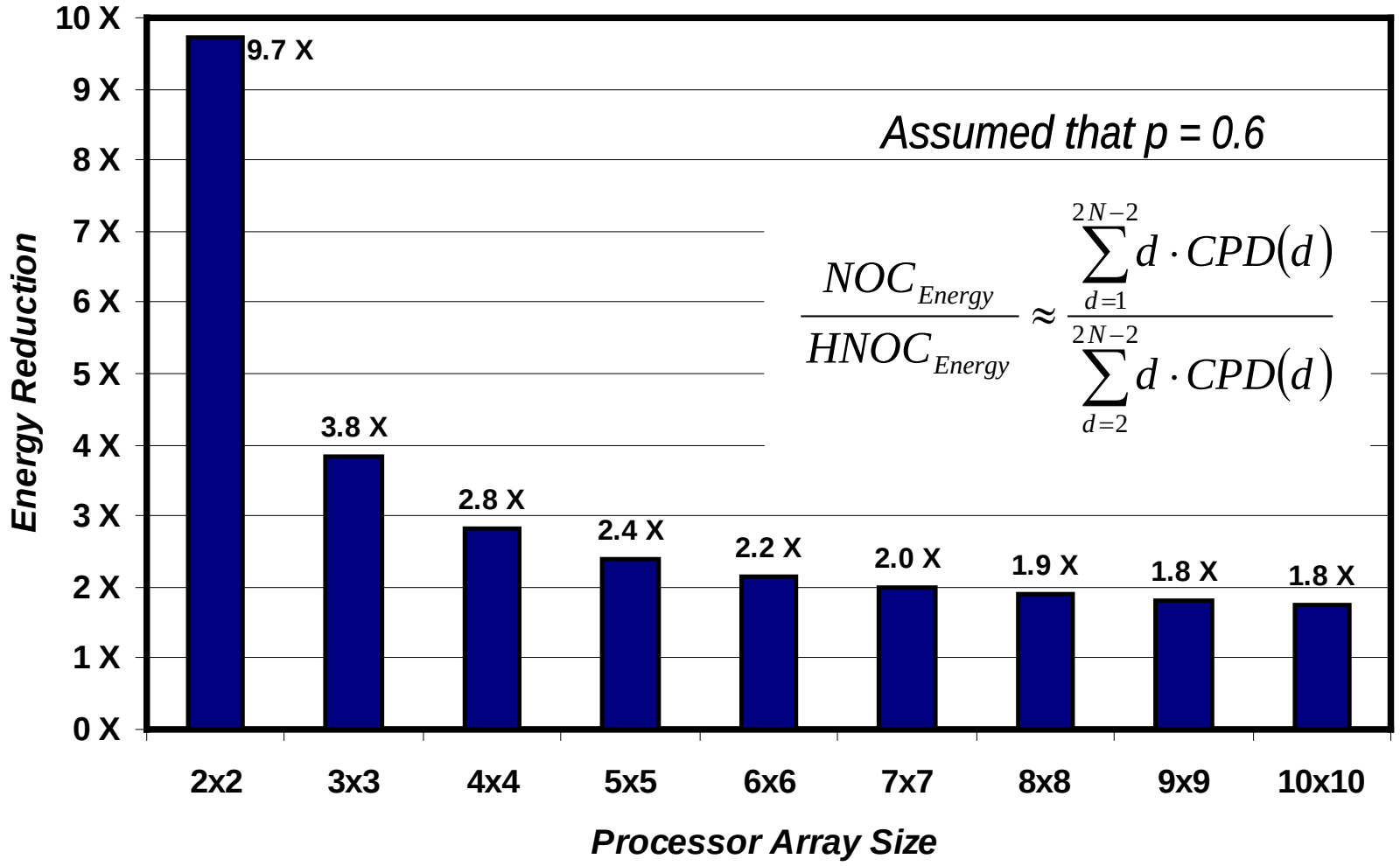
CPD for Various Rent's Exponents



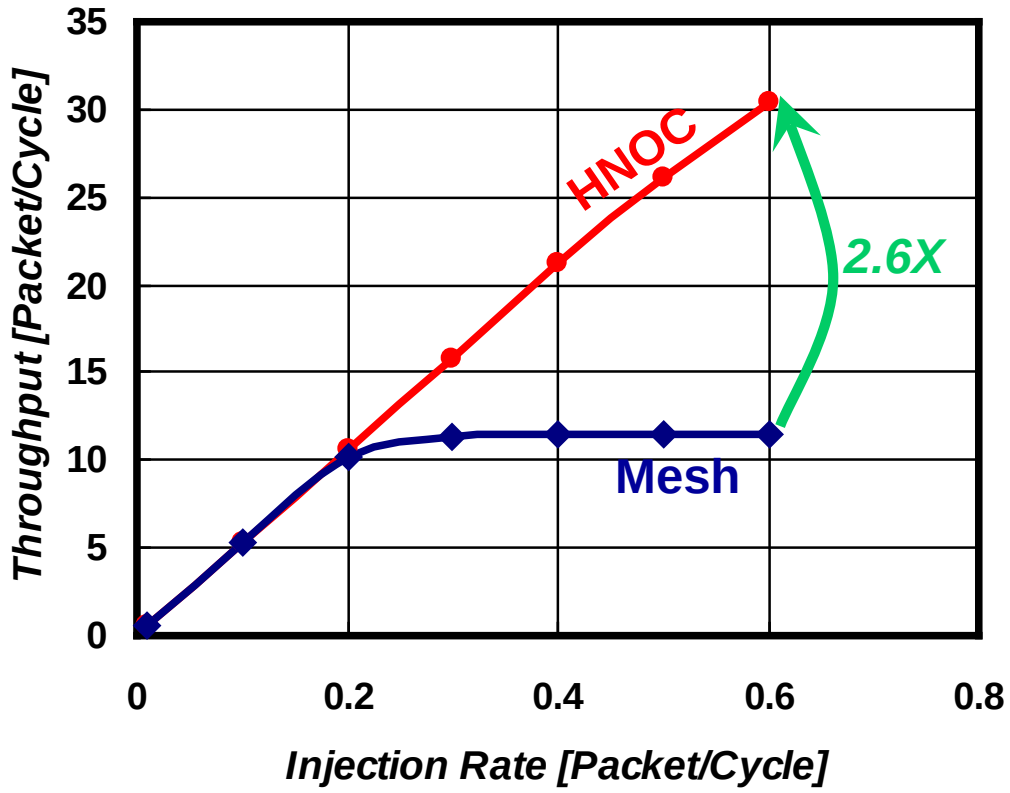
Maximum Throughput Improvement in HNoC



Maximum Energy Improvement in HNoC

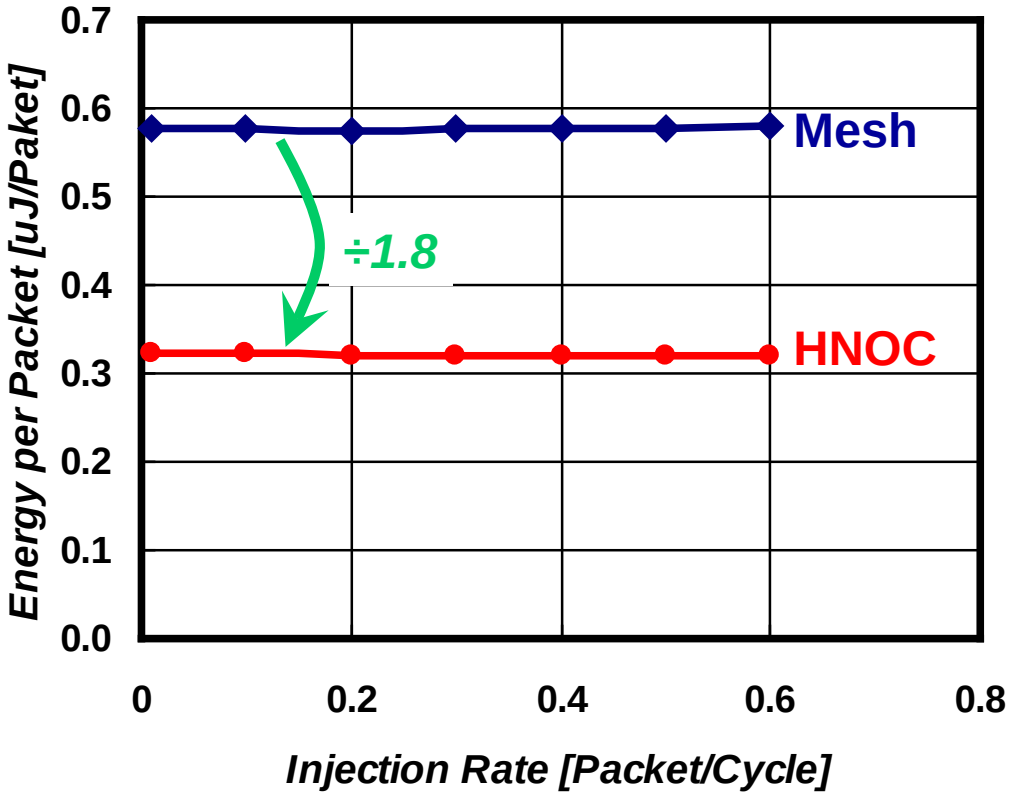


Throughput Simulation Results using Orion



System Parameters	Values
Number of Cores	64
Die Size	1cm x 1cm
Technology Node	45 nm
Clock Frequency	1 GHz
Flit Size	64 bits
Packet Size	5 flits
Rent's Exponent, *p*	0.60

Energy Simulation Results using Orion



System Parameters	Values
Number of Cores	64
Die Size	1cm x 1cm
Technology Node	45 nm
Clock Frequency	1 GHz
Flit Size	64 bits
Packet Size	5 flits
Rent's Exponent, *p*	0.60

Discussions

- Similar to EDA tools, compilers must take locality into account to achieve traffic localization with optimized program mapping and task assignment
- The proposed HNoC architecture can significantly improve the energy usage and performance of the system by directing the local communications through the low-latency, high-bandwidth, and low-power local buses and leaving the global communications to the standard NoC topology
- In practice, however, achieving this locality may be challenging. The compiler needs to be able to map the program such that the neighboring threads are mapped onto neighboring network cores. Moreover, runtime re-mapping and significant data movement maybe required, when the local buses in HNoC may not be able to provide significant support
- However, on average HNoC will indirectly support long-distance communication by removing the local communication traffic from the mesh NoC, leaving the mesh NoC fully dedicated to long-distance traffic