LNMIIT
The LNM Institute of
Information Technology

Network-on-Chip

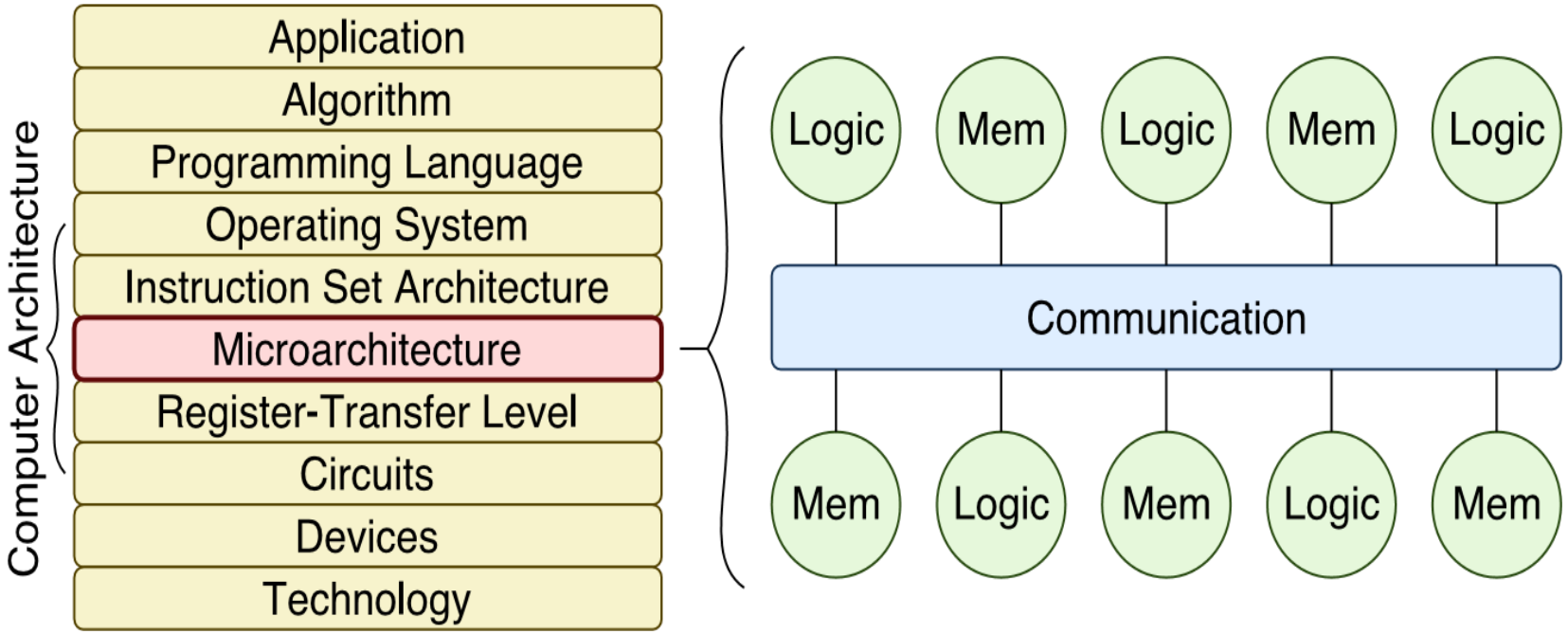
Sunil Kumar

The course will enable you to:

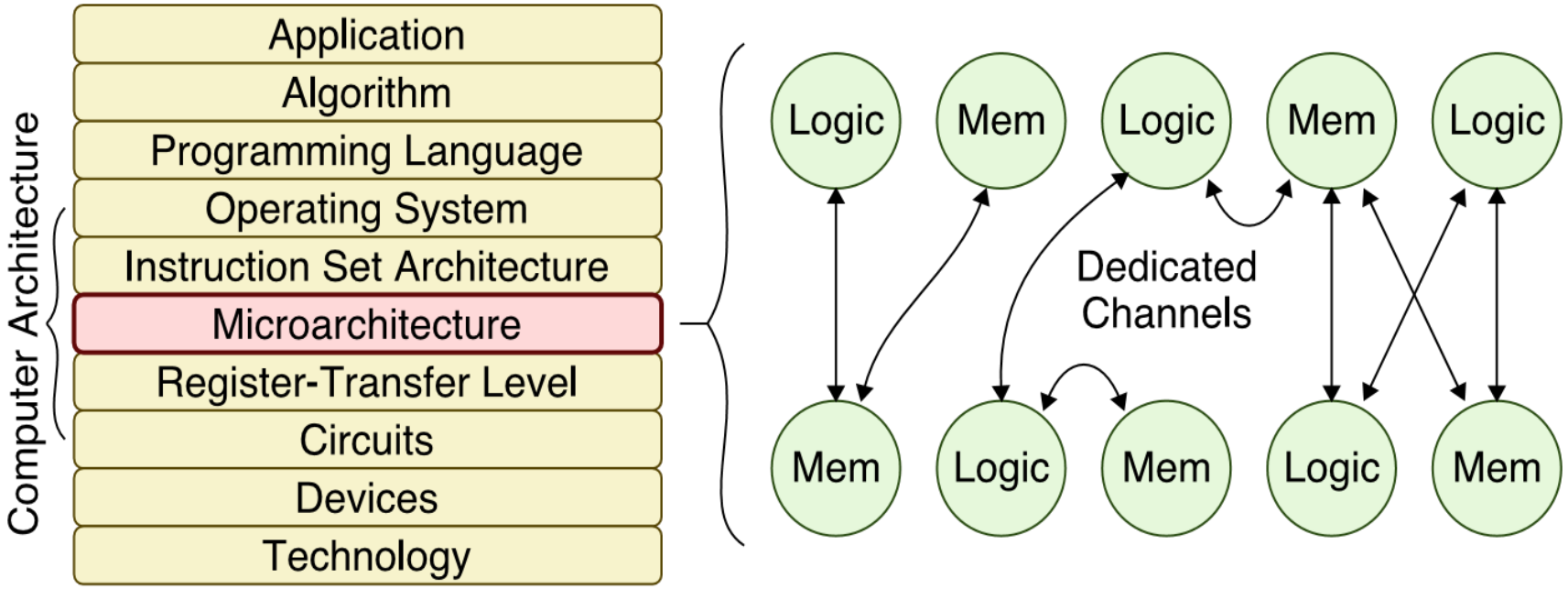
LNMIIT
The LNM Institute of Information Technology

- Understand the basic principles of Network-on-Chip.
- Understand the various techniques of NoC topologies.
- Understand the switching, flow control, routing techniques and router microarchitecture.
- Carry out experiments, analyze results and to make necessary conclusions using NoC simulator.

Introduction



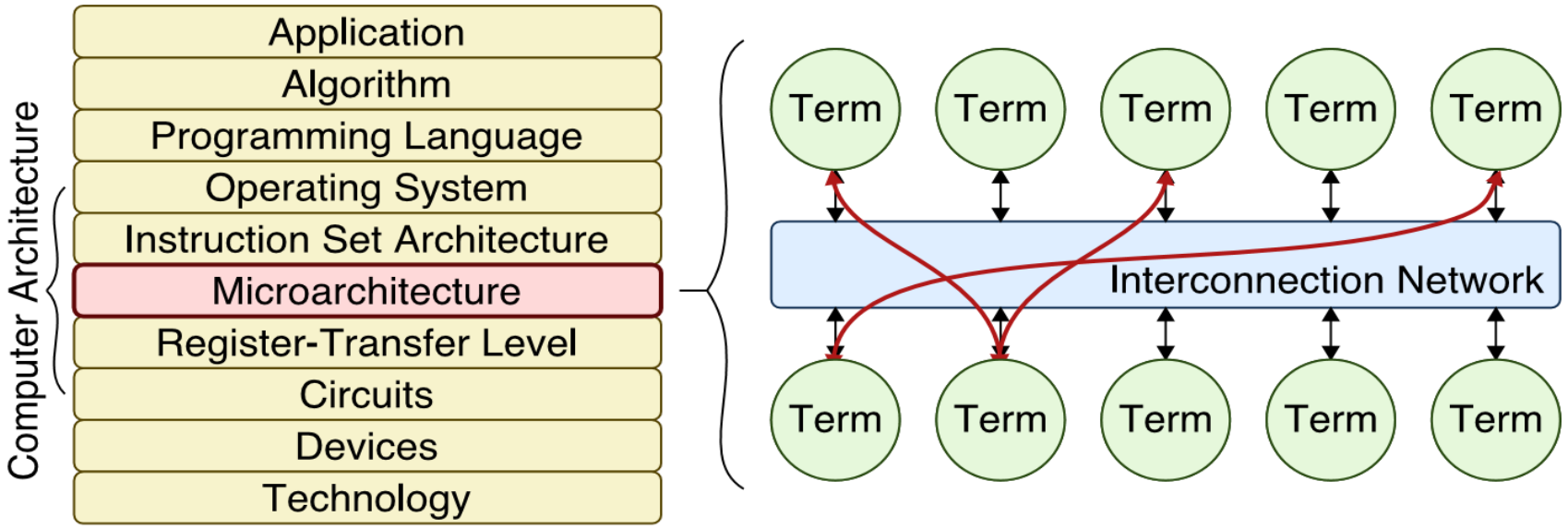
Introduction



Application: Ideally wants low-latency, high-bandwidth, dedicated channels between logic and memory

Technology: Dedicated channels too expensive in terms of area and power

Introduction

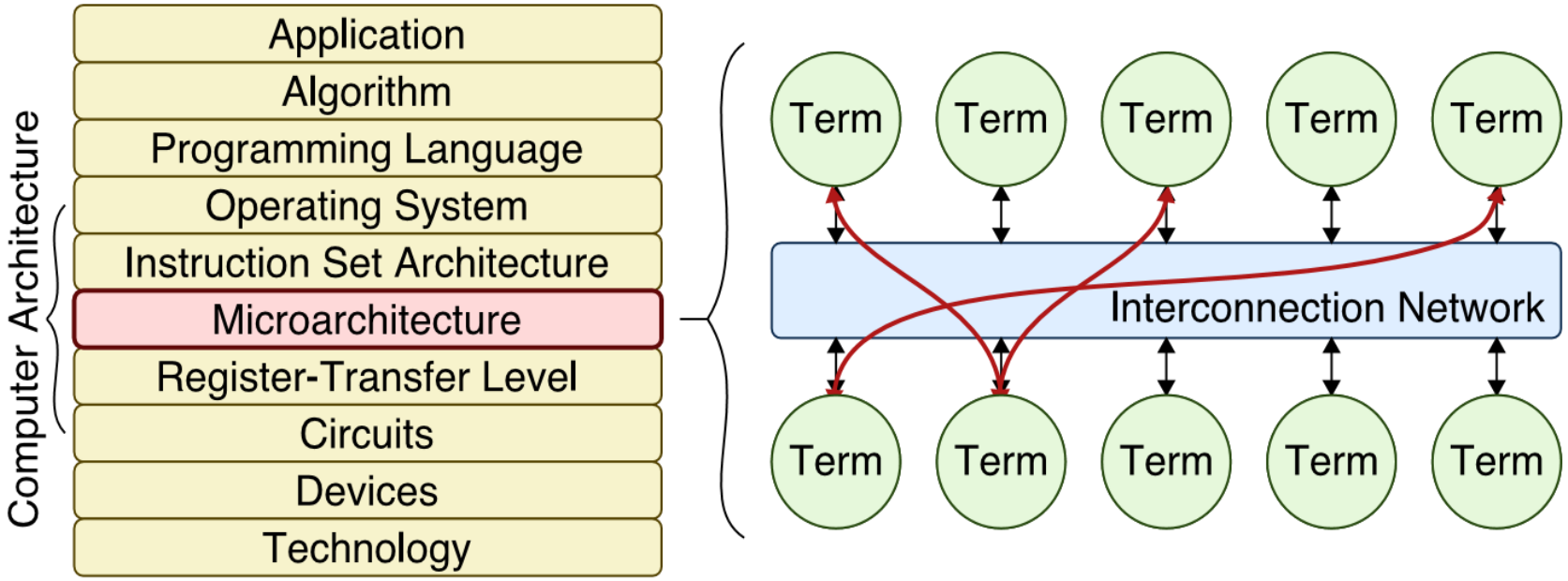


An Interconnection Network is a programmable system that transports data between terminals

Technology: Interconnection network helps efficiently utilize scare resources such as area and power

Application: Managing interconnection network can be critical to achieving good performance

Introduction



One Theme for Course:
Interplay between application requirements, technology constraints, and interconnection networks

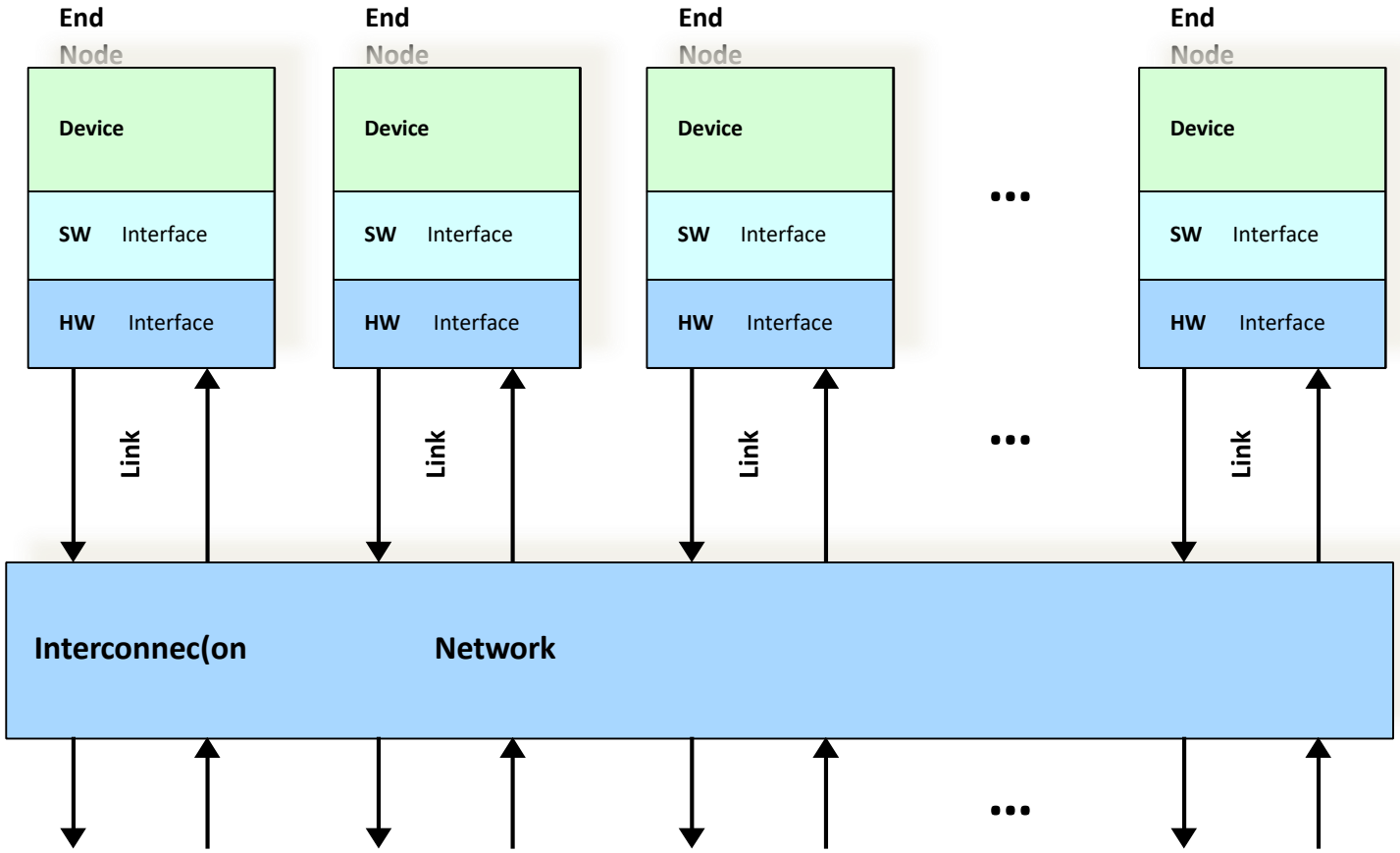
Interconnection Networks: Introduction

LNMIIT
The LNM Institute of Information Technology

- How to connect individual devices together into a community of communicating devices?
- Device:
  - Component within a computer
  - Single computer
  - System of computers
- Types of elements:
  - end nodes (device + interface)
  - Links
  - interconnection network
- Internetworking: interconnection of multiple networks

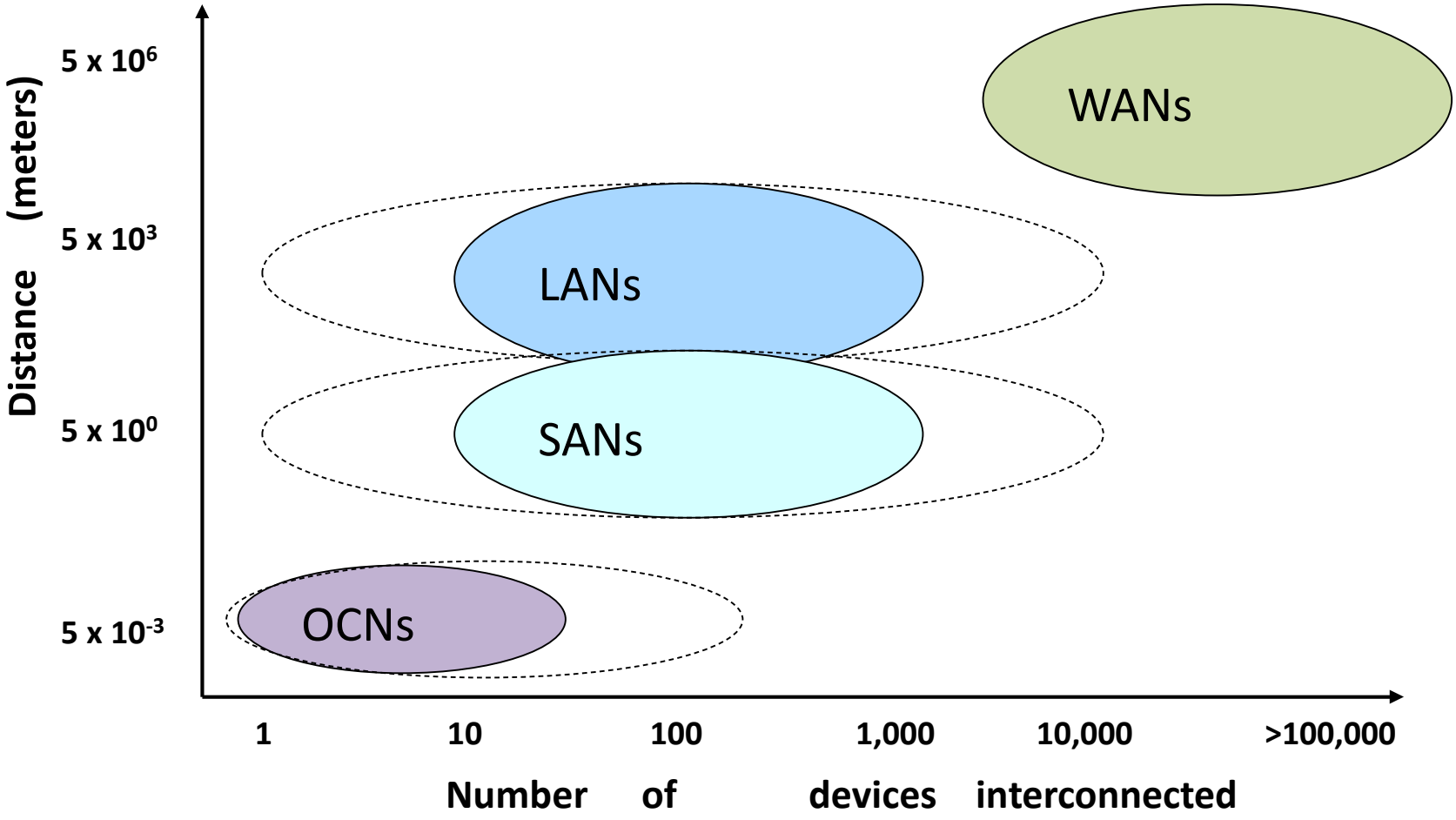
Interconnect network

LNMIIT
The LNM Institute of Information Technology



Interconnection Network Domains

LNMIIT
The LNM Institute of Information Technology



Interconnection Network

LNMIIT
The LNM Institute of Information Technology

- Key Design Principles
  - Transfer maximum amount of information (high bandwidth) within the least amount of time (low latency) so as to not bottleneck the system
  - Efficiently utilize shared but scarce resources (buffers, links, logic) to reduce area and power.

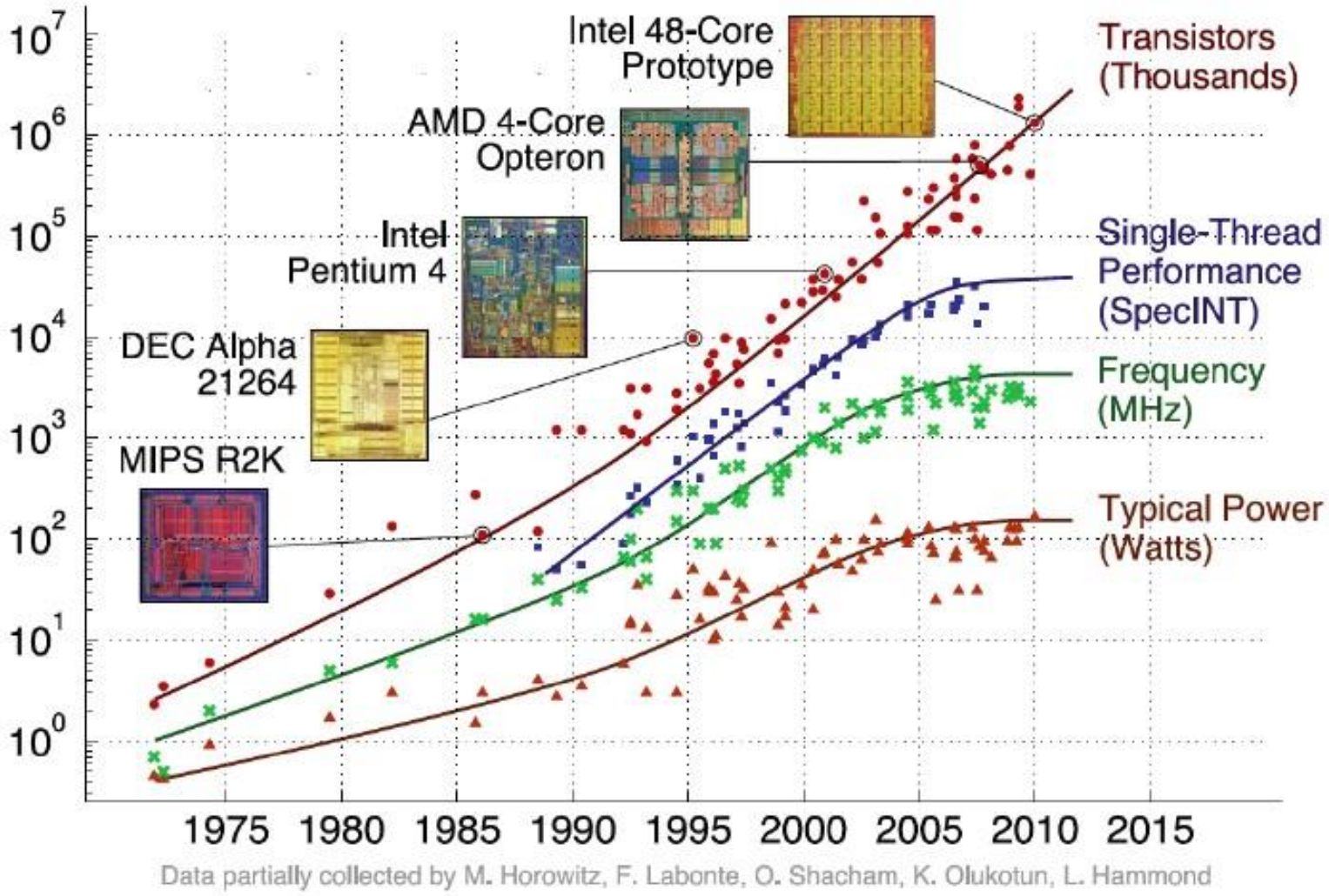
NoC and Off-Chip Networks

LNMIIT
The LNM Institute of Information Technology

- **NoC**
  - Sensitive to cost:
    - area and power
  - Wires are relatively cheap
  - Latency is critical
  - Traffic is known a-priori
  - Design time specialization
  - Custom NoCs are possible
- **Off-Chip Networks**
  - Cost is in the links
  - Latency is tolerable
  - Traffic/applications unknown
  - Changes at runtime
  - Adherence to networking standards

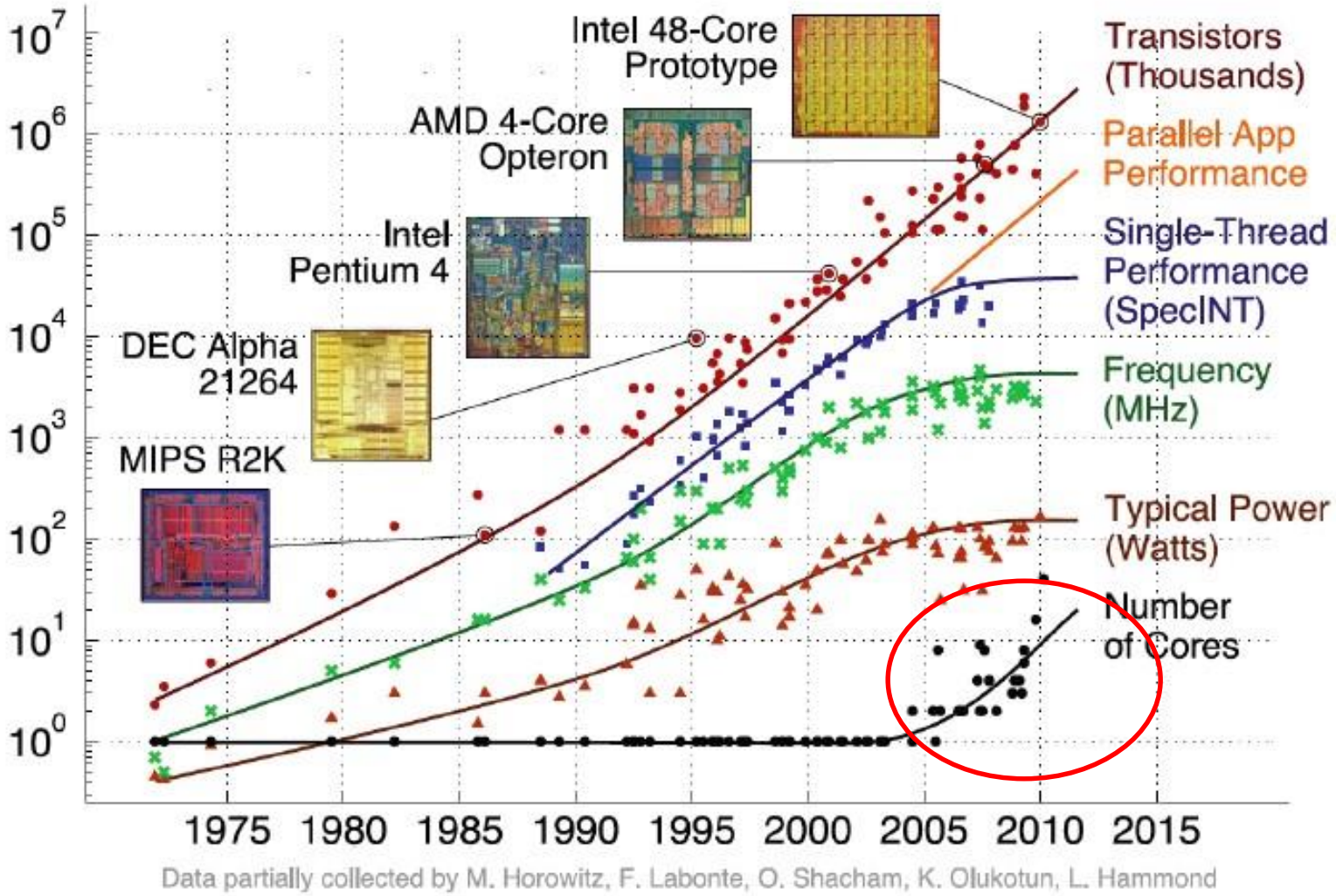
Why Study NoC?

LNMIIT
The LNM Institute of Information Technology



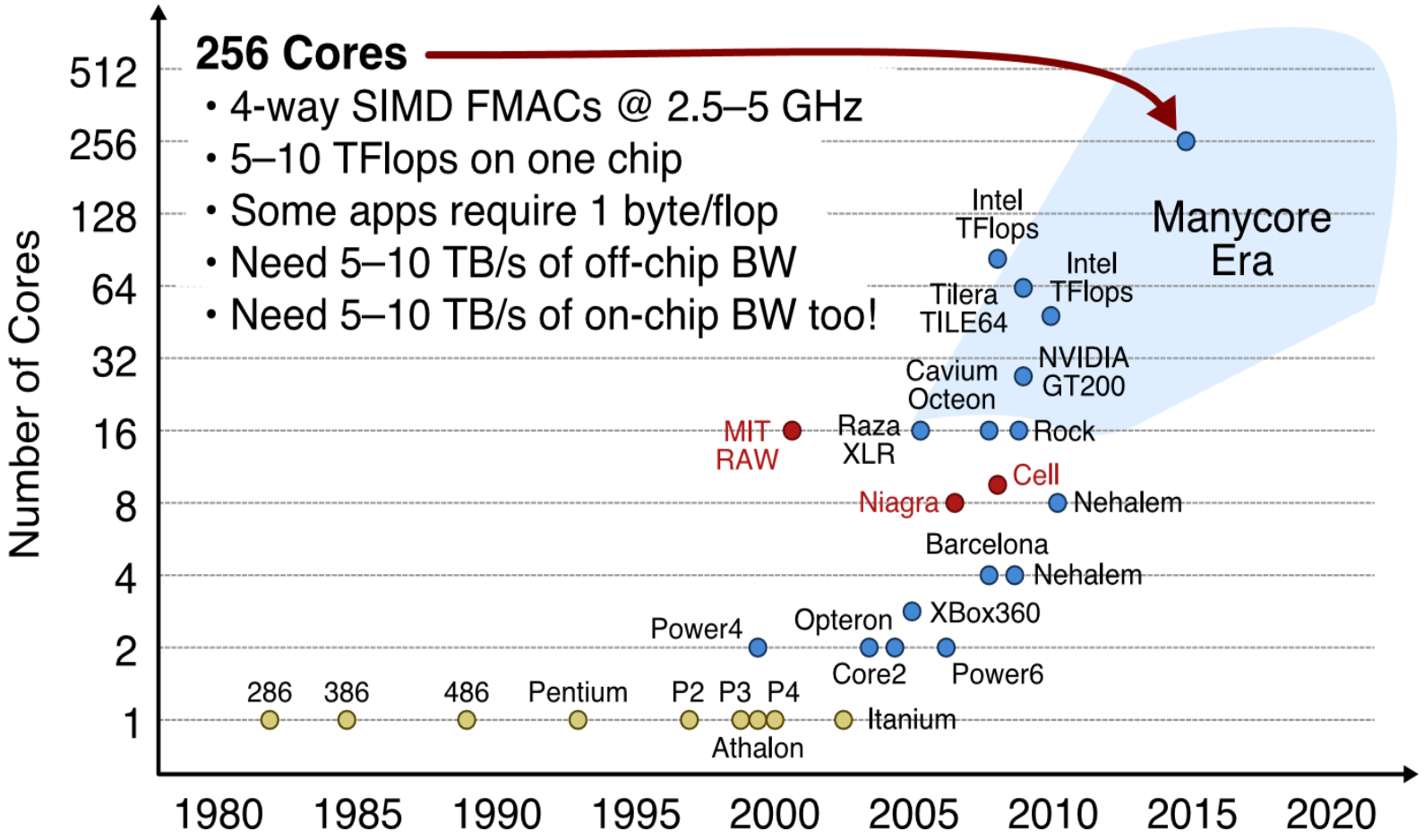
Why Study NoC?

LNMIIT
The LNM Institute of Information Technology

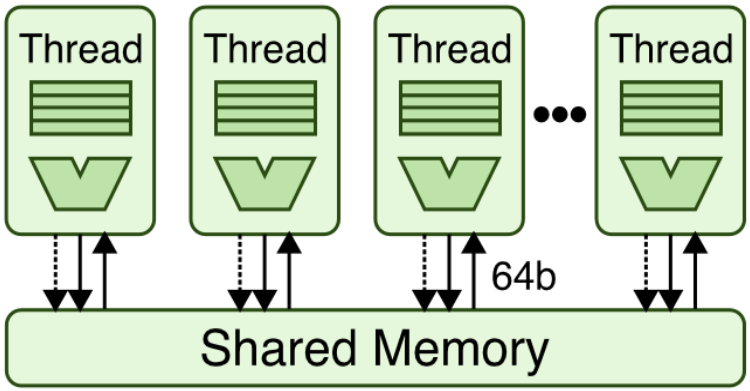


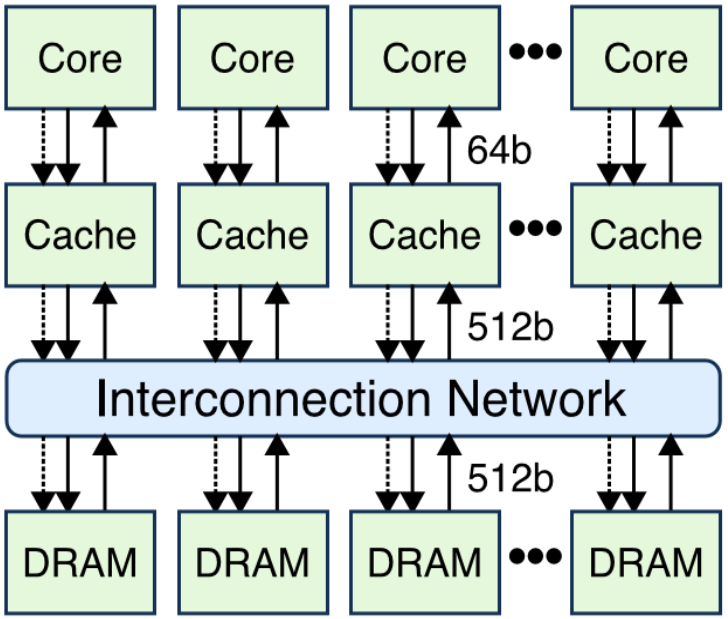
Examples of Multicore and Manycore Processors

LNMIIT
The LNM Institute of Information Technology



Irregular Threaded Application Running on Processor-to-Memory Network





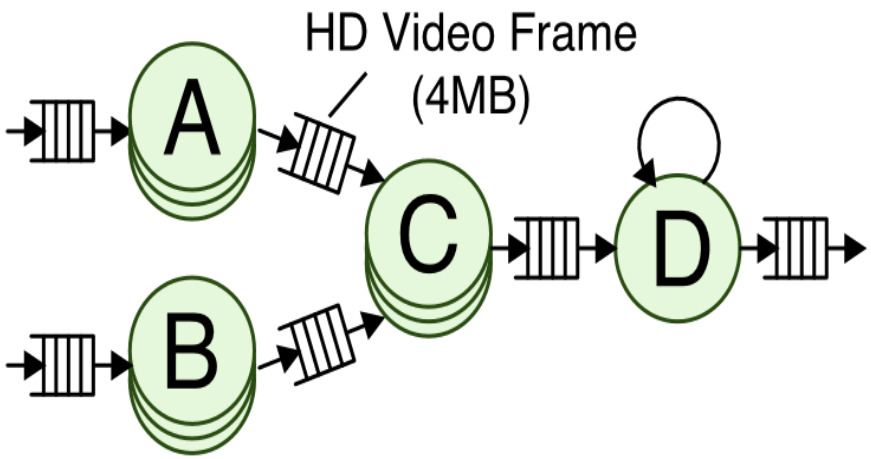
Application Requirements

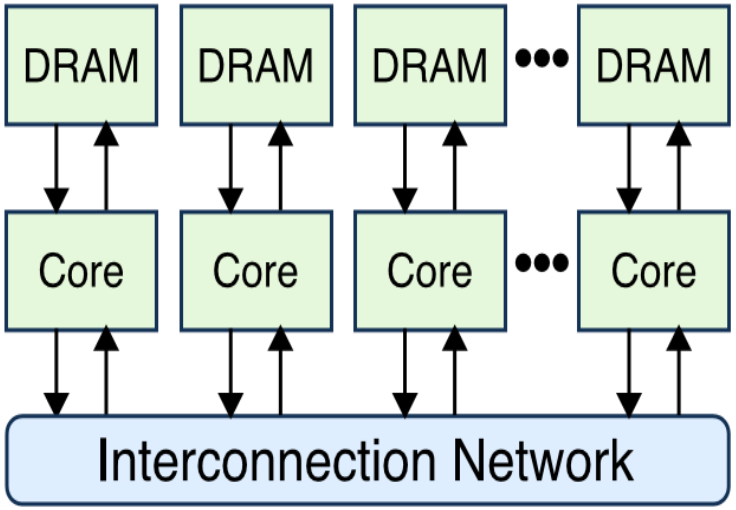
Message Size	512b
Average Bandwidth	400 MB/s
Peak Bandwidth	8 GB/s
Latency	Minimum
Traffic Pattern	Arbitrary

What network design meets these requirements within the technology constraints and with the least a ea, power, and lowest latency?

Streaming Application Running on Processor-to-Processor Network

LNMIIT
The LNM Institute of Information Technology





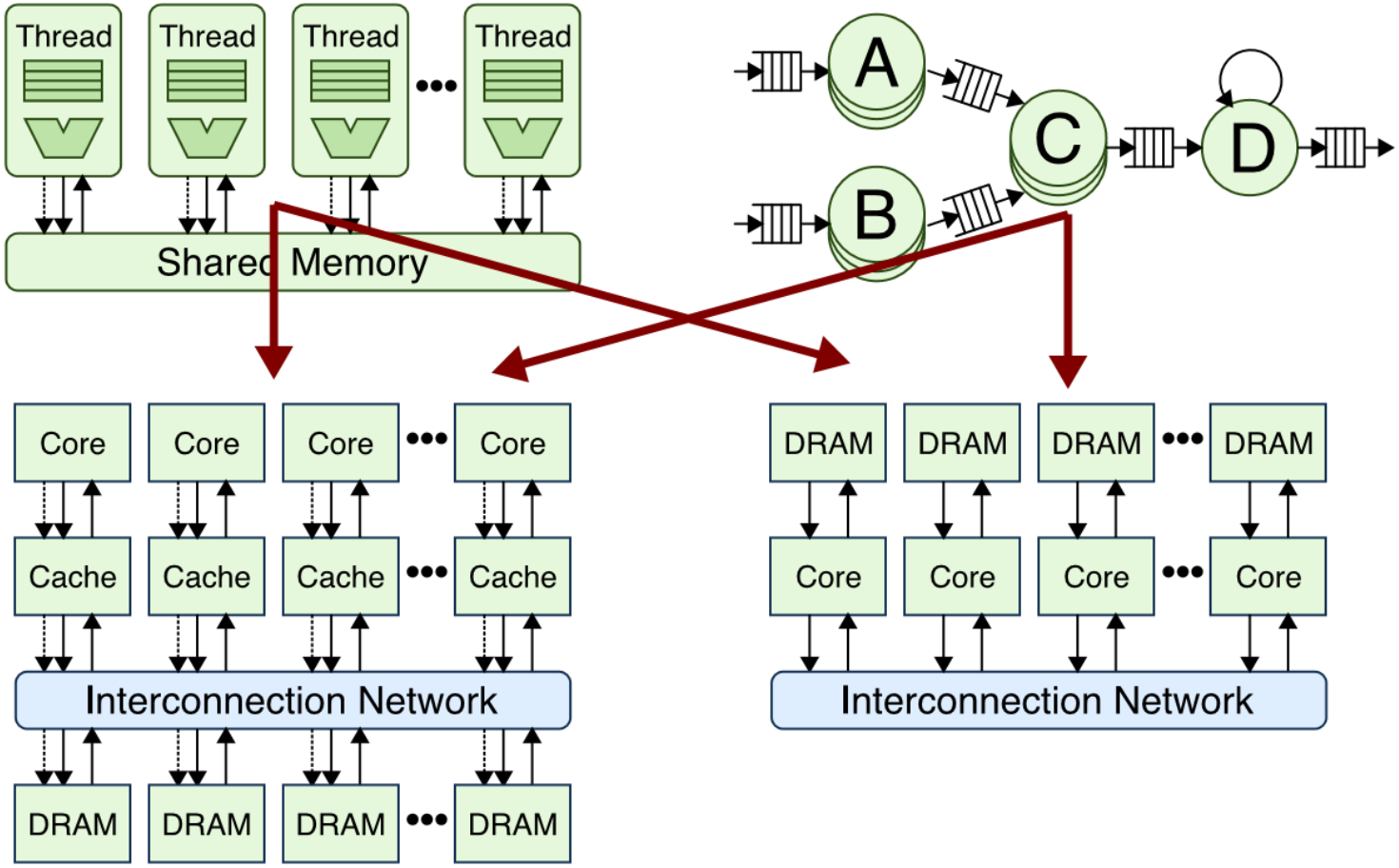
Application Requirements

Message Size	4MB
Average Bandwidth	120 MB/s
Peak Bandwidth	120 MB/s
Latency	Tolerant
Traffic Pattern	Streaming

What network design meets these requirements within the technology constraints and with the least area, power, and maximum bandwidth?

Goal: Flexible Networks Capable of Running Many Different Kinds of Applications

LNMIIT
The LNM Institute of Information Technology

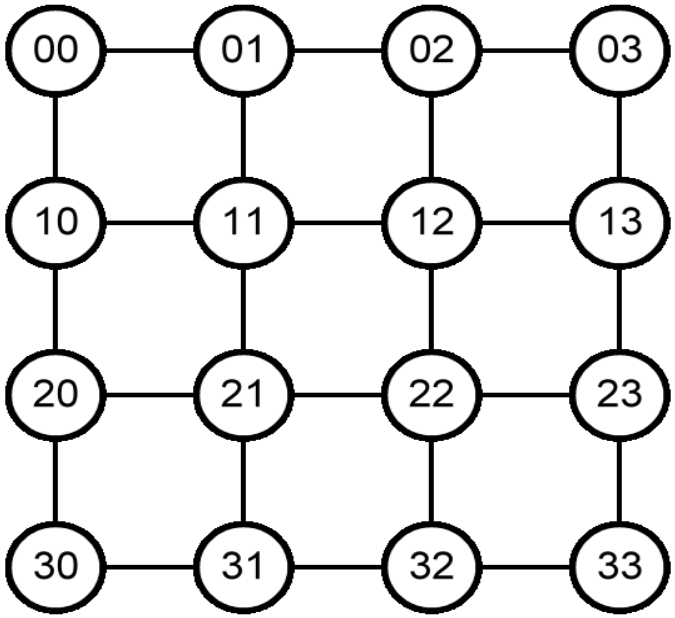


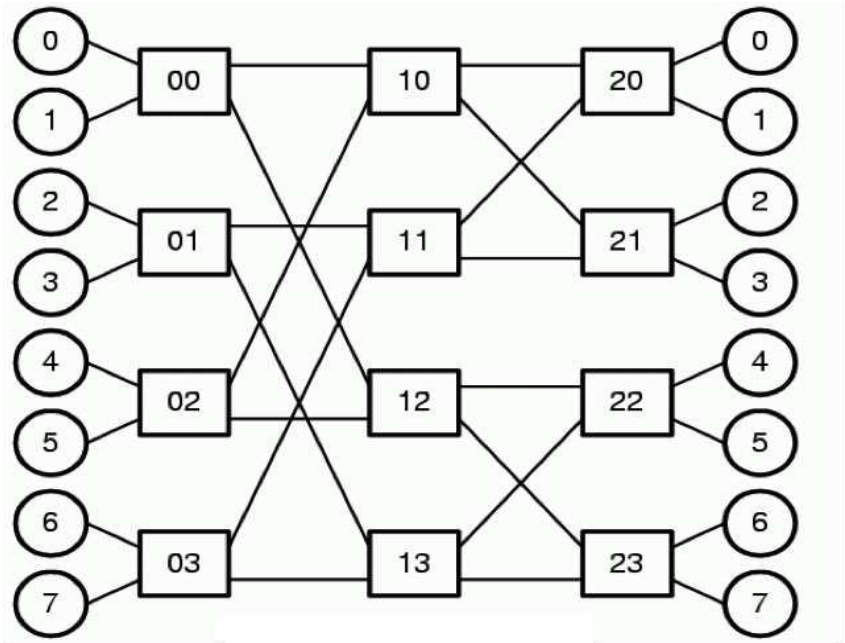
Interconnection Network Basics

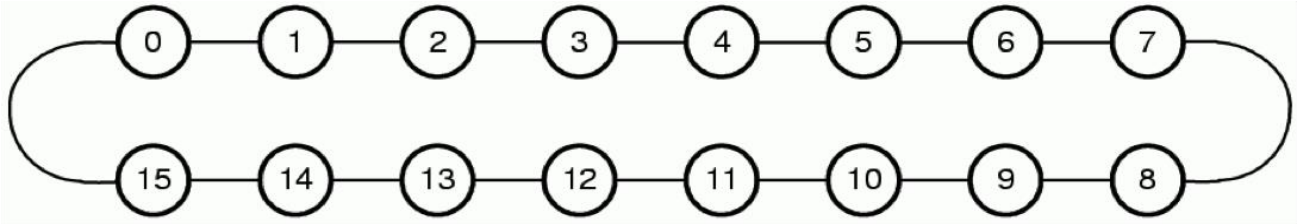
LNMIIT
The LNM Institute of Information Technology

- Topology
  - Arrangement of Nodes and Channels
- Routing
  - Determining Path Between Terminals
- Flow Control
  - Managing Allocation of Resources
- Router Microarchitecture

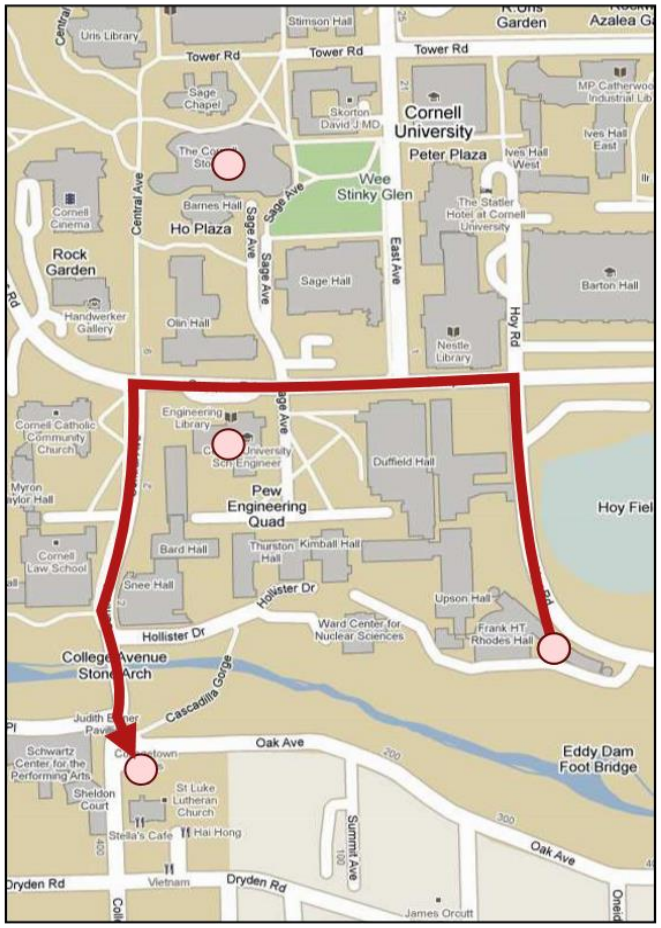
Topology

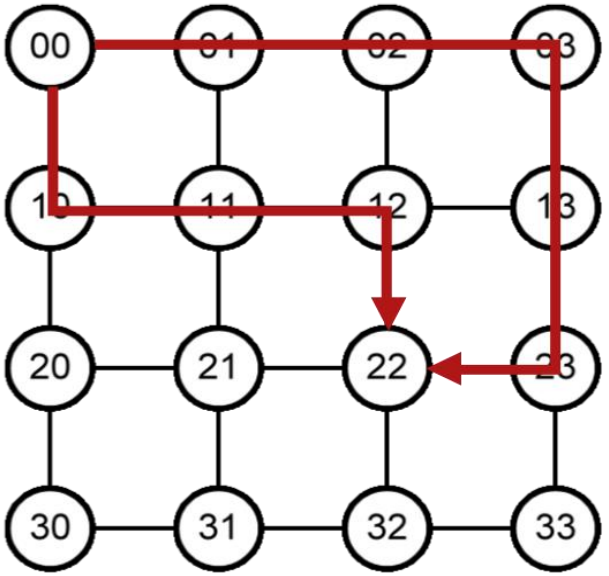

00
01
02
03
10
11
12
13
20
21
22
23
30
31
32
33


0
1
2
3
4
5
6
7
00
01
02
03
10
11
12
13
20
21
22
23
0
1
2
3
4
5
6
7


0
1
2
3
4
5
6
7
15
14
13
12
11
10
9
8

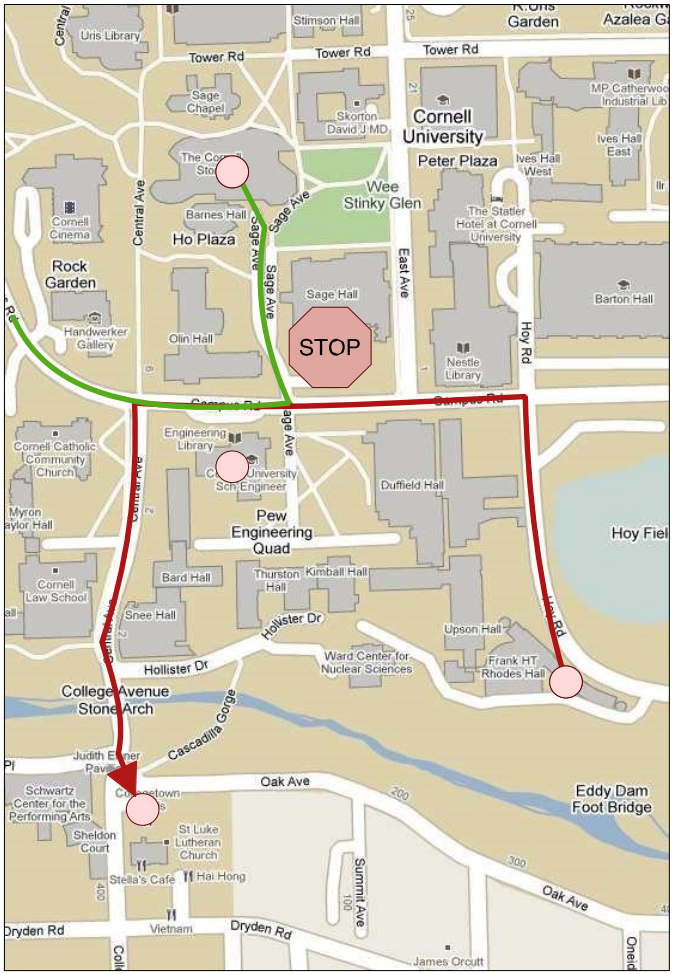
Routing

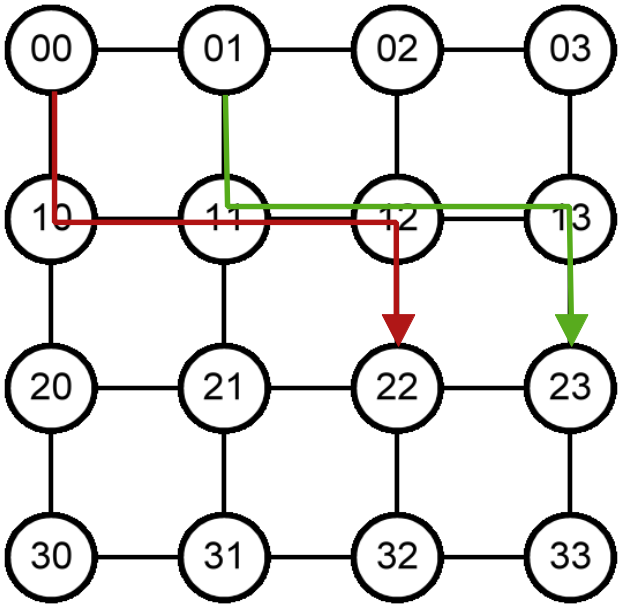


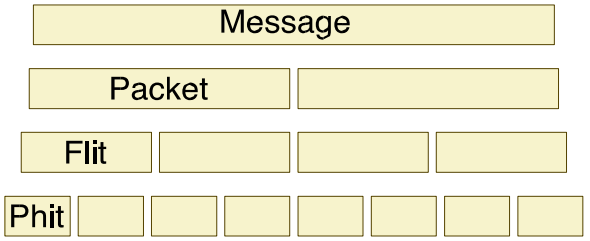


Minmal Routing vs. Non-Minimal Routing
Oblivious vs. Adaptive Routing
Deterministics vs. Randomized Routing

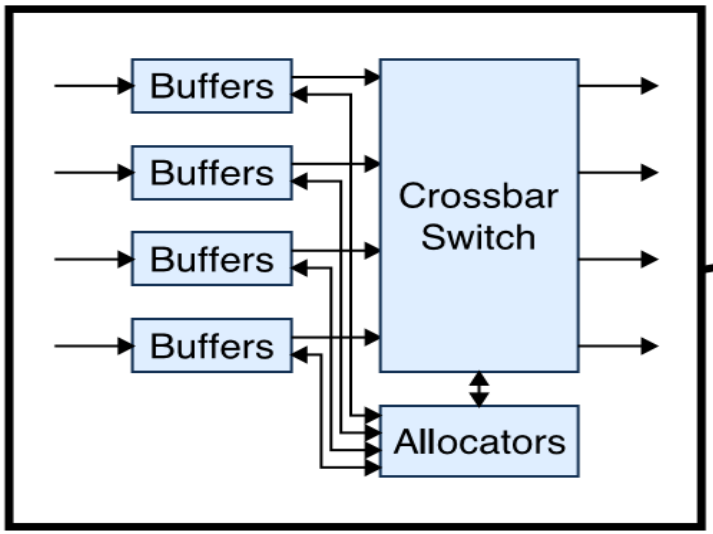
Flow Control

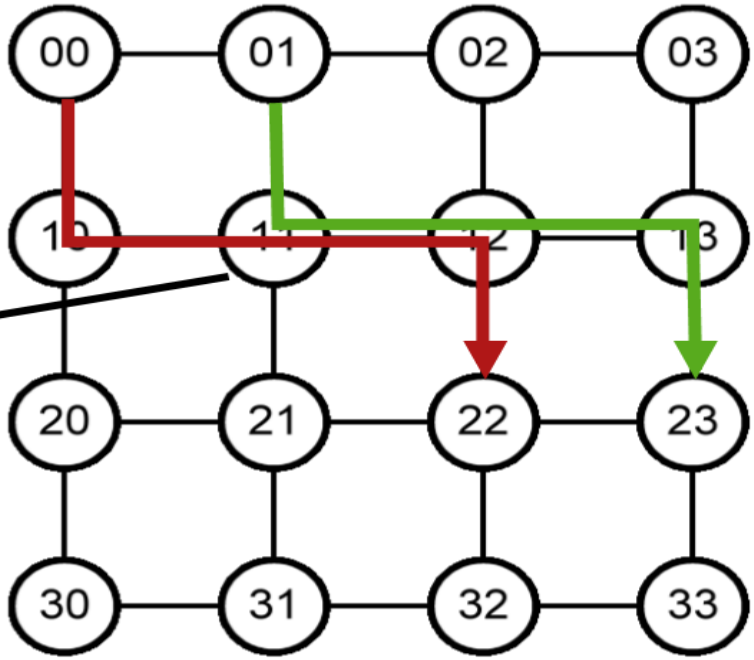


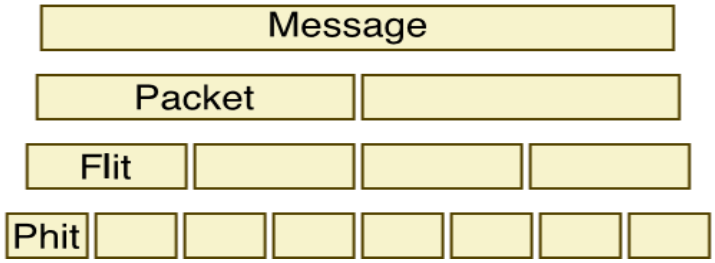




Router Architecture

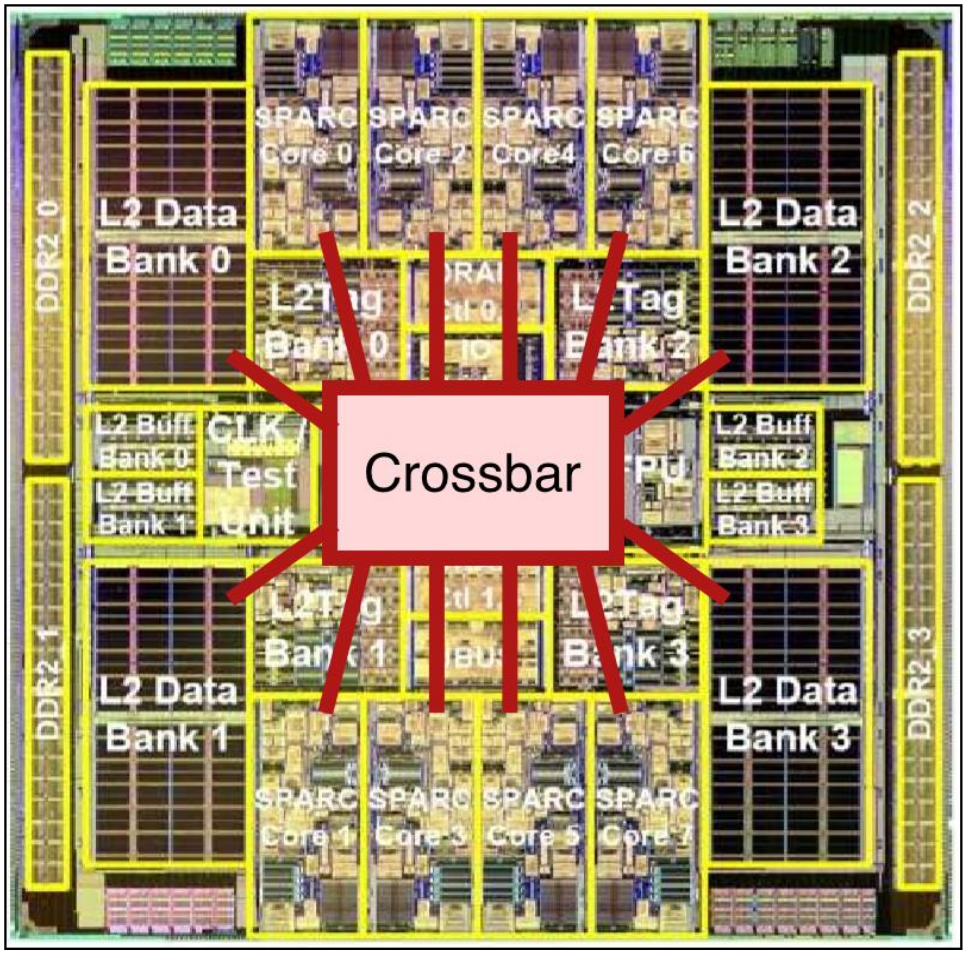






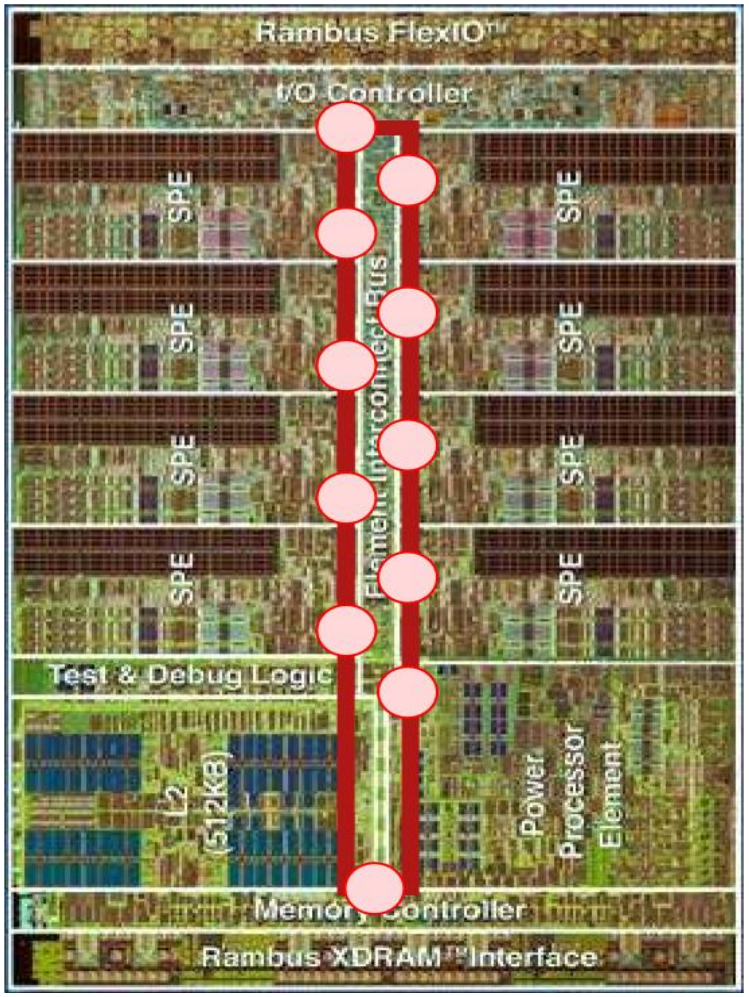
Example Architectures

Sun Niagara Processor



- 8 multithreaded processors
- Single-stage crossbar connecting 8 cores to 4 L2 cache banks
- "200 GB/s" total bisection BW

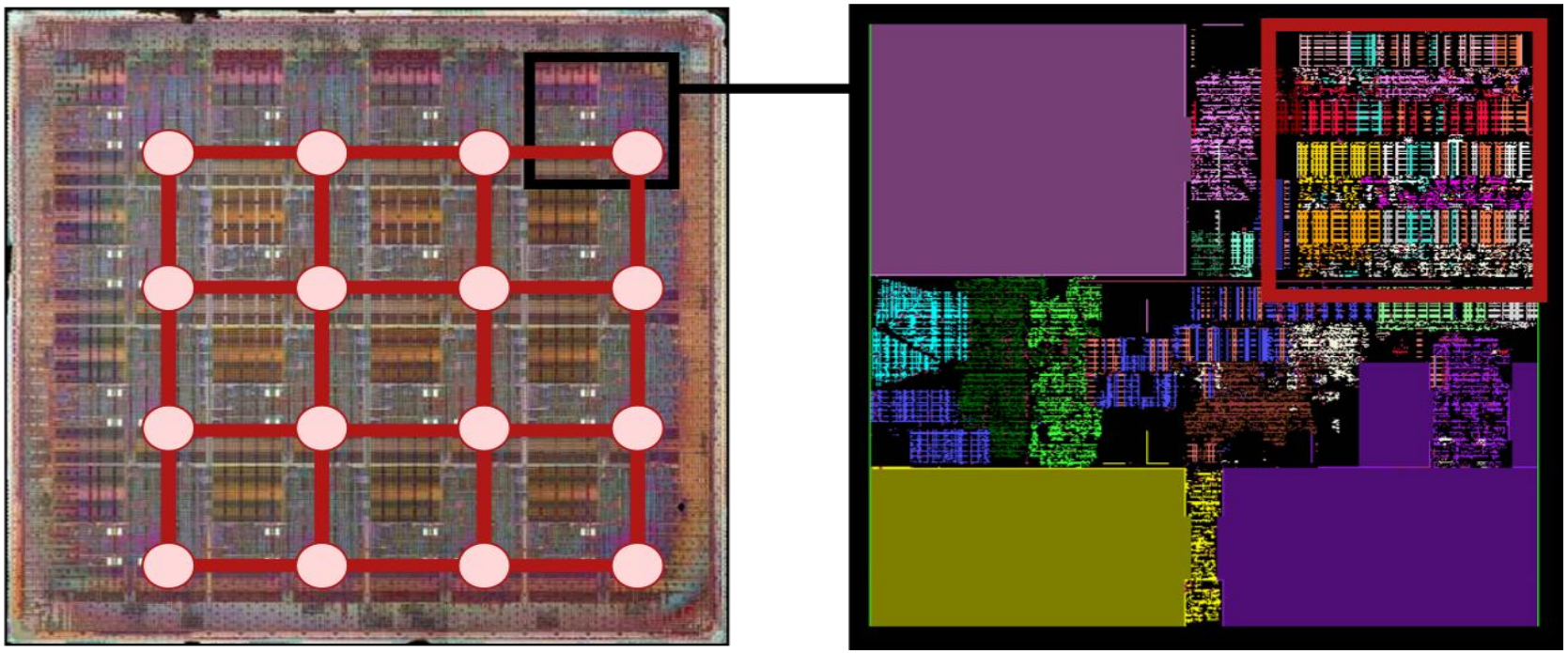
IBM Cell Processor



- 1 general-purpose processor
- 8 processors specialized for data-parallelism
- 4 uni-directional rings
- Each ring is 128b wide at 1.6 GHz
- Network Bisection BW = 25.6 GB/s
- Total Bisection = 102.4 GB/s

MIT Raw Processor

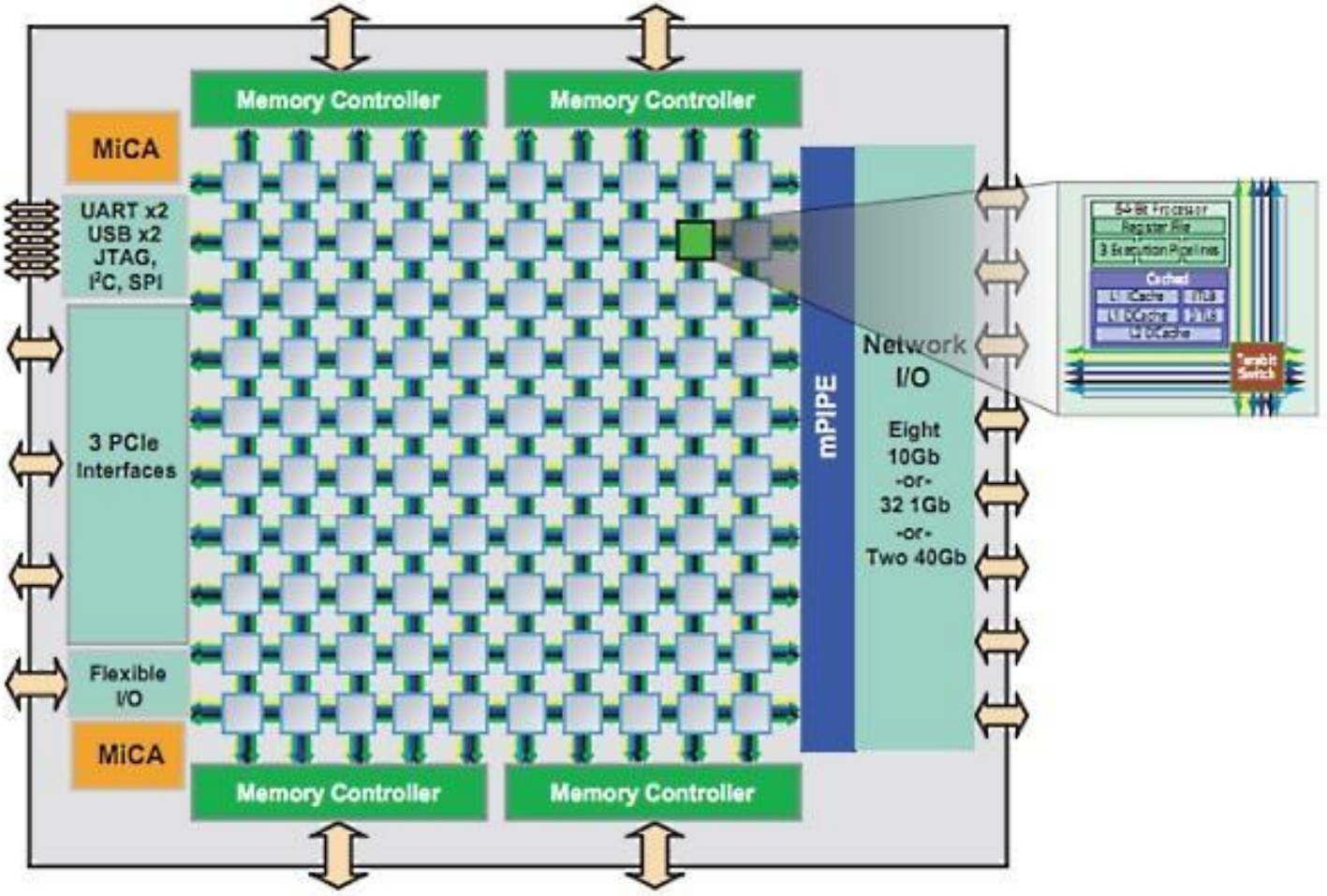
LNMIIT
The LNM Institute of Information Technology



- 16 simple RISC cores
- Two dynamically routed mesh networks (32b/channel)
- Two statically routed mesh networks for message passing (32b/channel)
- Bisection bandwidth per network is 8*32b at 400 MHz 12.8 = 12.8 GB/s
- Total bisection bandwidth is 51.2 GB/s
- Network consumes 20-30% of total chip power

Tilera Tile100

LNMIIT
The LNM Institute of Information Technology



Any Question?