Performance Evaluation of I/O Traffic and Placement of I/O Nodes on a High Performance Network

Salvador Coll[*†], Fabrizio Petrini[*], Eitan Frachtenberg[*] and Adolfy Hoisie[*]

[*]CCS-3 Modeling, Algorithms, and Informatics

Los Alamos National Laboratory

[†]Digital Systems Design and Parallel Architectures Groups

Technical University of Valencia - SPAIN

scoll@lanl.gov

Outline

- Introduction
- Quadrics network design overview
- Experimental framework
- Experimental results
- Conclusions

Introduction

- Common trend in large-scale clusters: high performance data networks
- I/O can be limited by the interconnect performance

Introduction

- Common trend in large-scale clusters: high performance data networks
- I/O can be limited by the interconnect performance
- Open problems:
  - influence of the I/O servers placement
  - effect of using dedicated or shared I/O servers
  - potential interference of background I/O traffic with computation

Los Alamos
NATIONAL LABORATORY

Introduction

- Some of the most powerful systems in the world use the Quadrics interconnection network:
- The Terascale Computing System (TCS) at the Pittsburgh Supercomputing Center – the second most powerful computer in the world

Los Alamos NATIONAL LABORATORY

Introduction

- Some of the most powerful systems in the world use the Quadrics interconnection network:
- The Terascale Computing System (TCS) at the Pittsburgh Supercomputing Center – the second most powerful computer in the world
- ASCI Q machine, currently under development at Los Alamos National Laboratory (30 TeraOps, expected to be delivered by the end of 2002)

Los Alamos NATIONAL LABORATORY

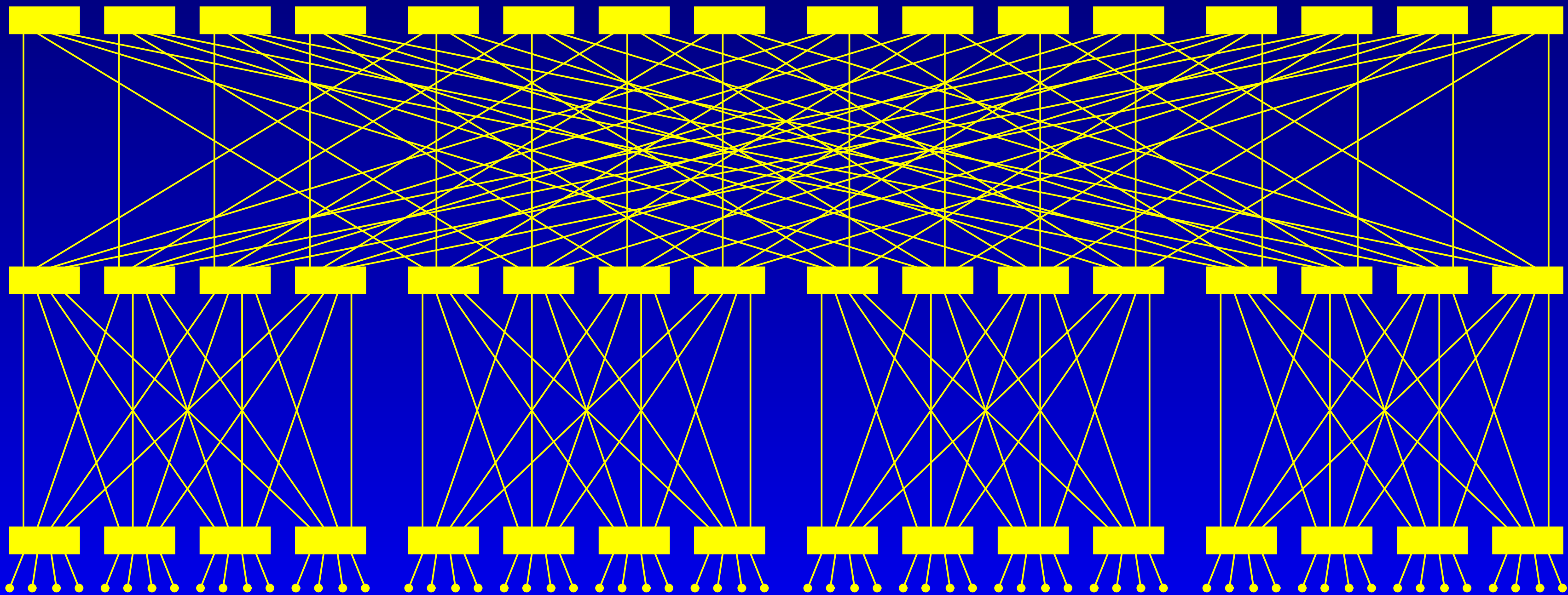
Introduction

- Some of the most powerful systems in the world use the Quadrics interconnection network:

- The Terascale Computing System (TCS) at the Pittsburgh Supercomputing Center – the second most powerful computer in the world
- ASCI Q machine, currently under development at Los Alamos National Laboratory (30 TeraOps, expected to be delivered by the end of 2002)
- Objective: experimental evaluation of a Quadrics-based cluster under I/O traffic

Quadrics Network Design Overview

- Fat-tree
- Based on 4x4 switches
- Wormhole switching
- 2 virtual channels per physical link
- Adaptive routing



Los Alamos
NATIONAL LABORATORY

Quadrics Network Design Overview

- Fat-tree
- Based on 4x4 switches
- Wormhole switching
- 2 virtual channels per physical link
- Adaptive routing

Some of the most important aspects of this network are:

- the integration of the local memory into a distributed virtual shared memory,
- the support for zero-copy remote DMA transactions and
- the hardware support for collective communication.

Los Alamos NATIONAL LABORATORY

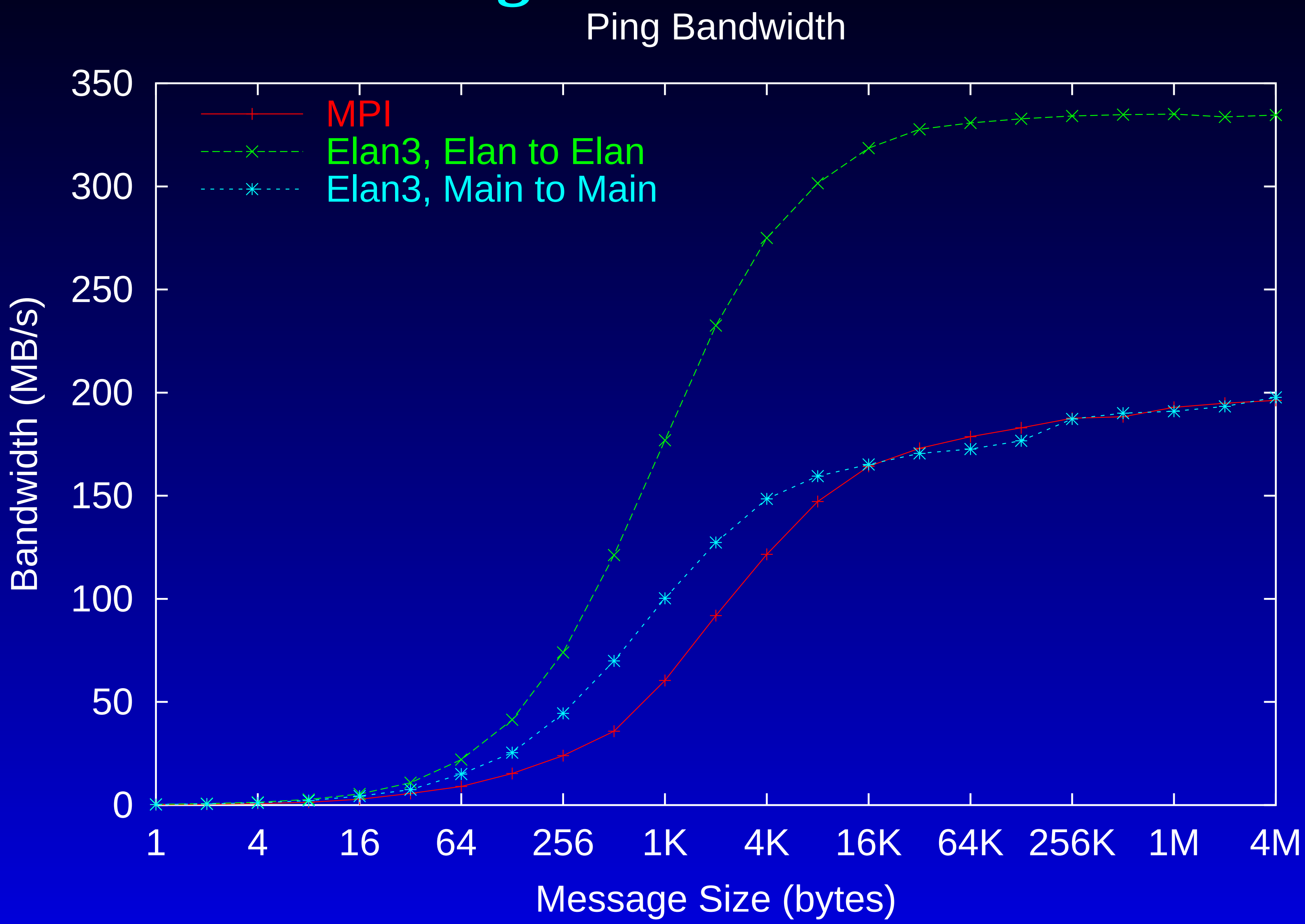
Experimental Framework

- The experimental results are obtained on a 64-node cluster of Compaq AlphaServer ES40s running Tru64 Unix.
- Each Alpahserver is attached to a quaternary fat-tree of dimension three through a 64 bit, 33 MHz PCI bus using the Elan3 card.
- In order to expose the real network performance, we place the communication buffers in Elan memory.

Experimental Results

- We present:
  - unidirectional and bidirectional ping results, as a reference, and
  - single hot-spot
  - multiple hot-spots
  - combined traffic: I/O plus uniform traffic

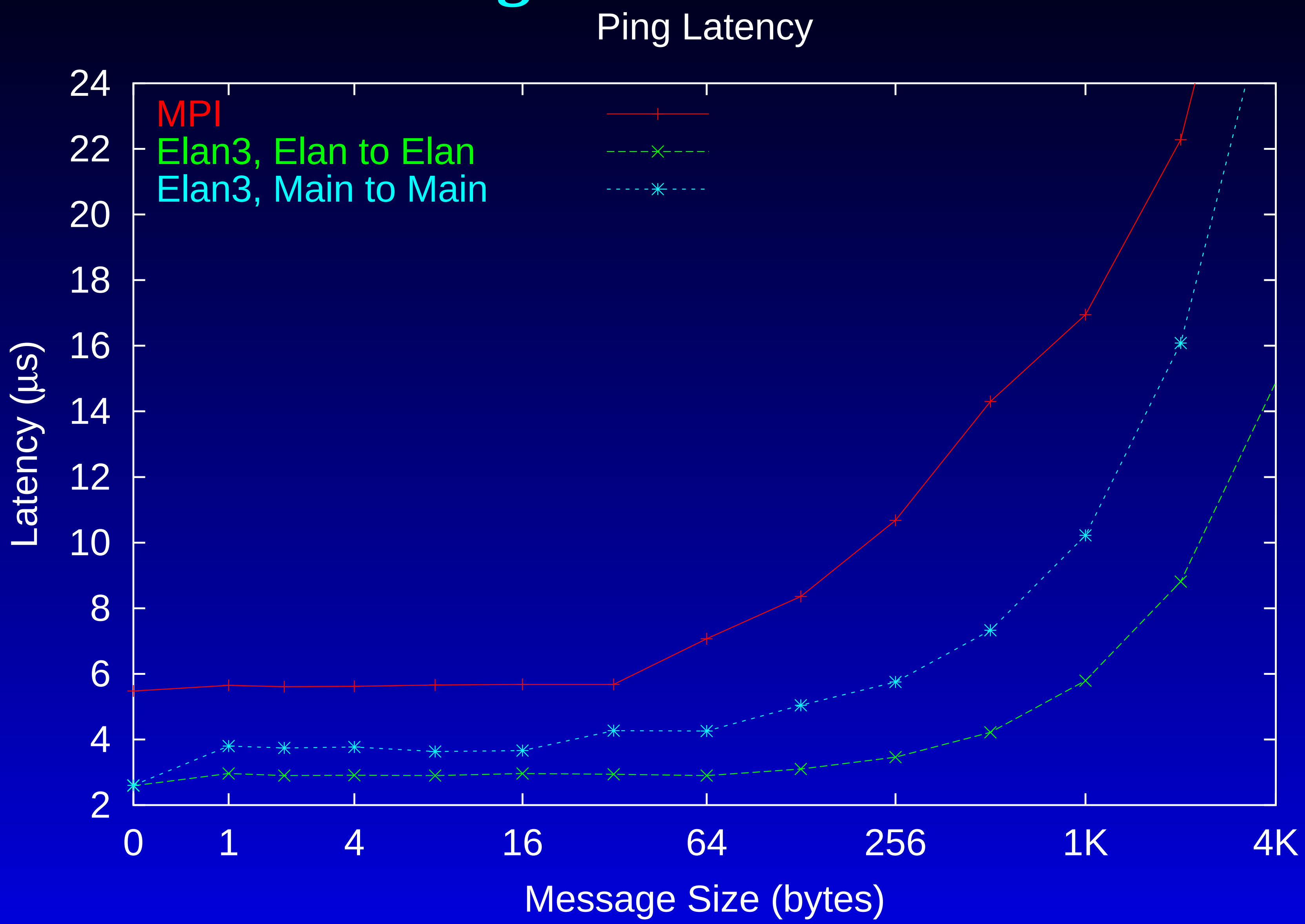
Unidirectional Ping



- Peak data bandwidth (Elan to Elan) of **335 MB/s** $\simeq$ 396 MB/s
- Main to main memory asymptotic bandwidth of 200 MB/s

Los Alamos NATIONAL LABORATORY

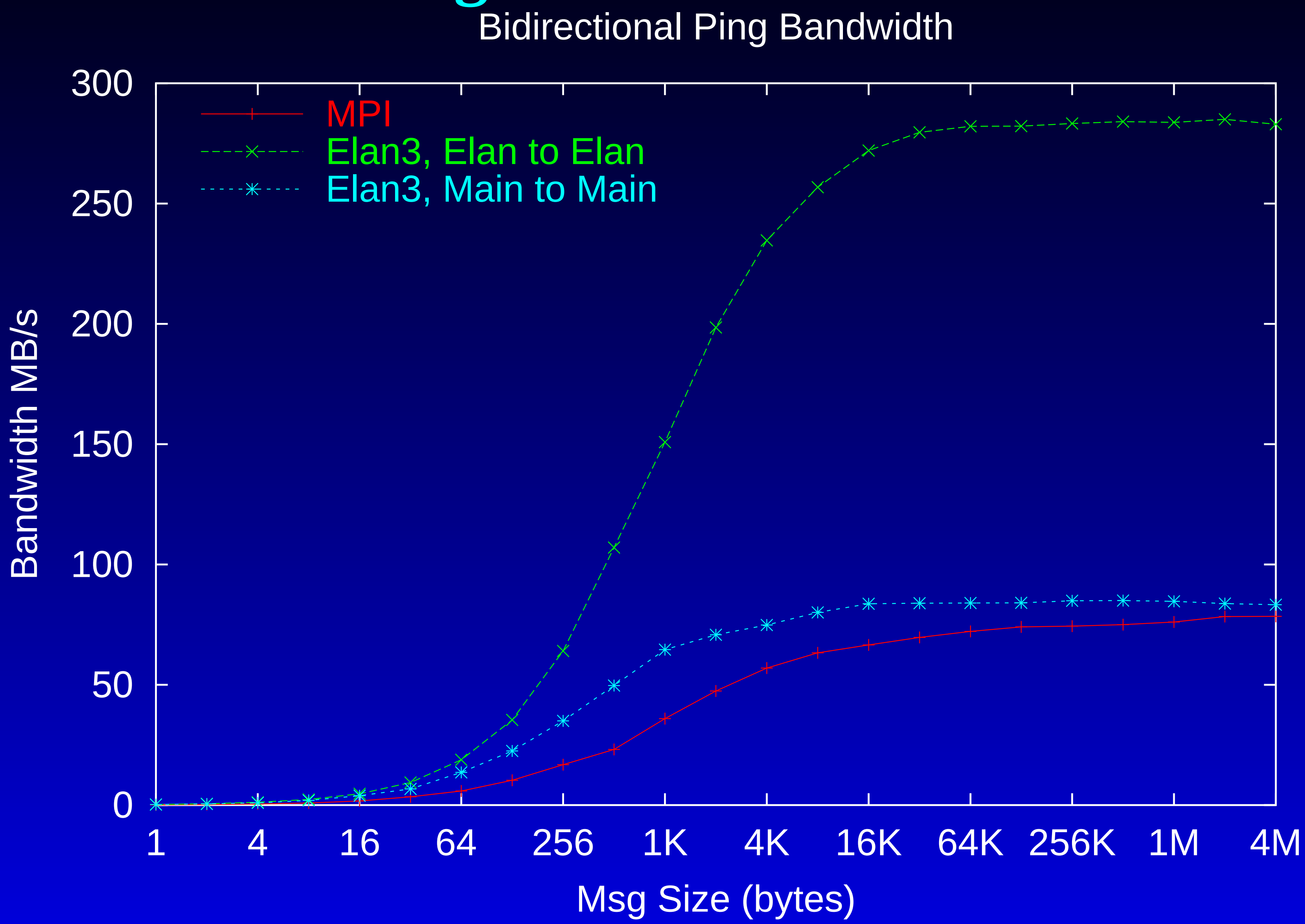
Unidirectional Ping



- Latency of **2.4 μs** up to 64-byte messages (Elan to Elan memory)
- Higher MPI latency due to message tag matching

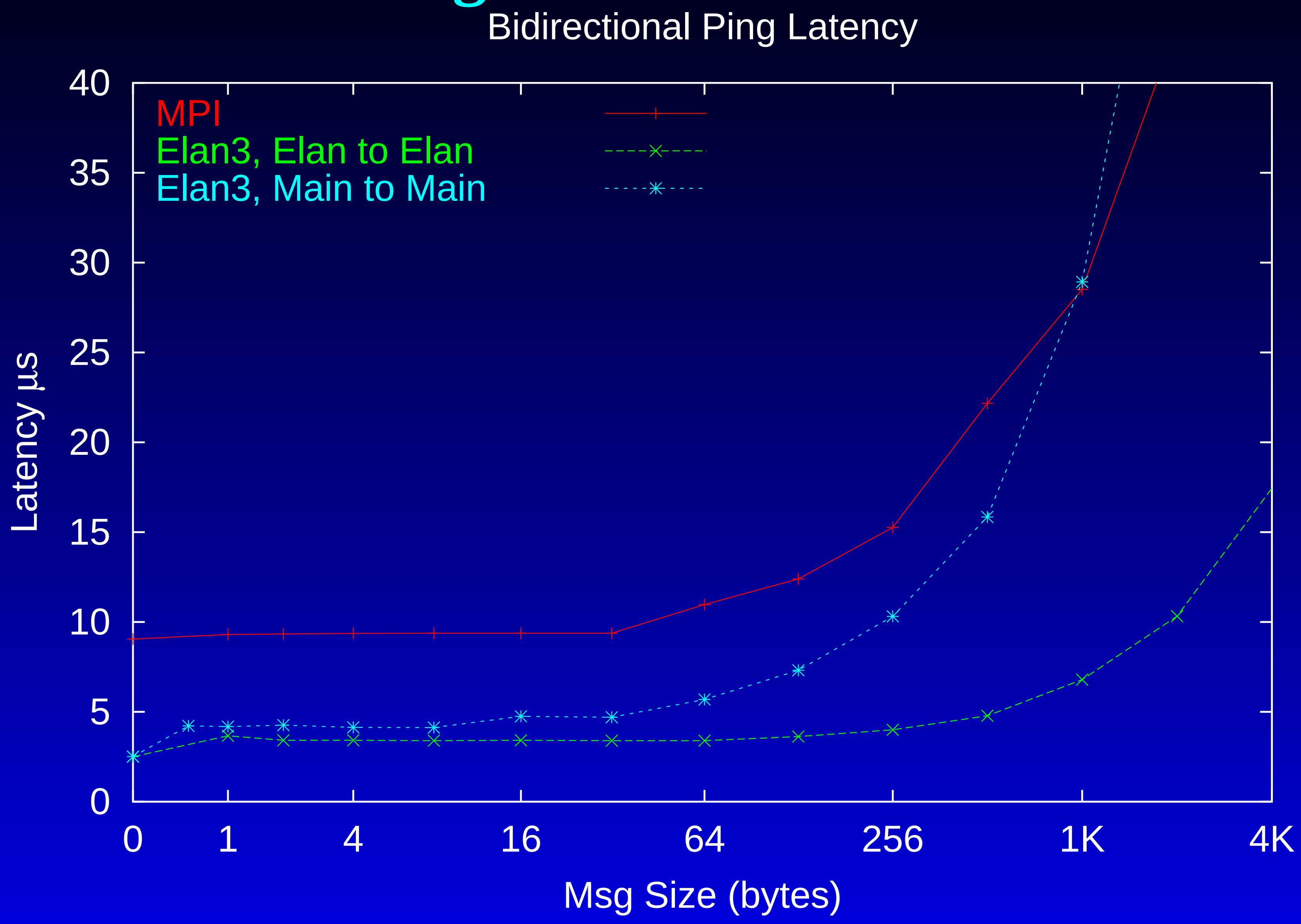
Los Alamos
NATIONAL LABORATORY

Bidirectional Ping



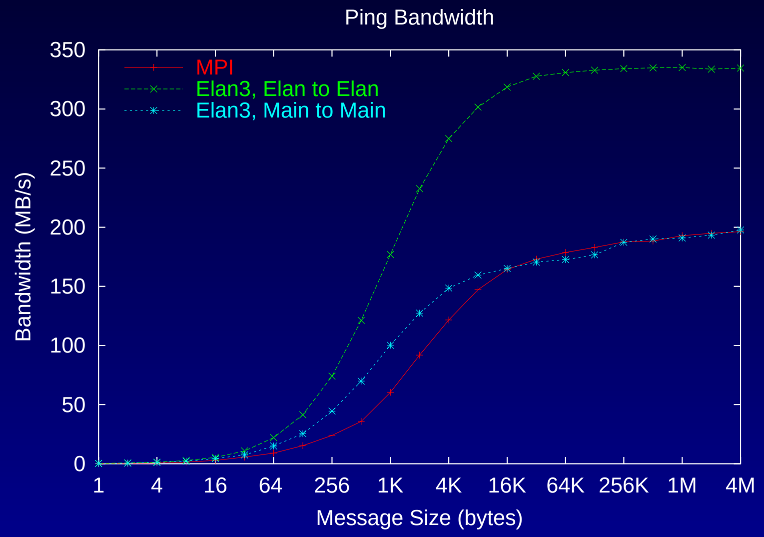
- Peak data bandwidth (Elan to Elan memory) of **280 MB/s**
- Main to main memory asymptotic bandwidth of 80 MB/s

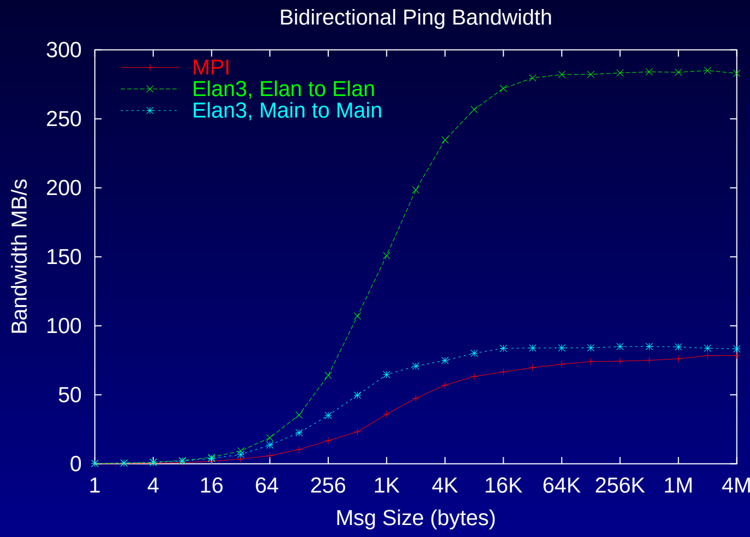
Bidirectional Ping



- Latency of **4 μs** up to 64-byte messages (Elan to Elan memory)

Ping Summary



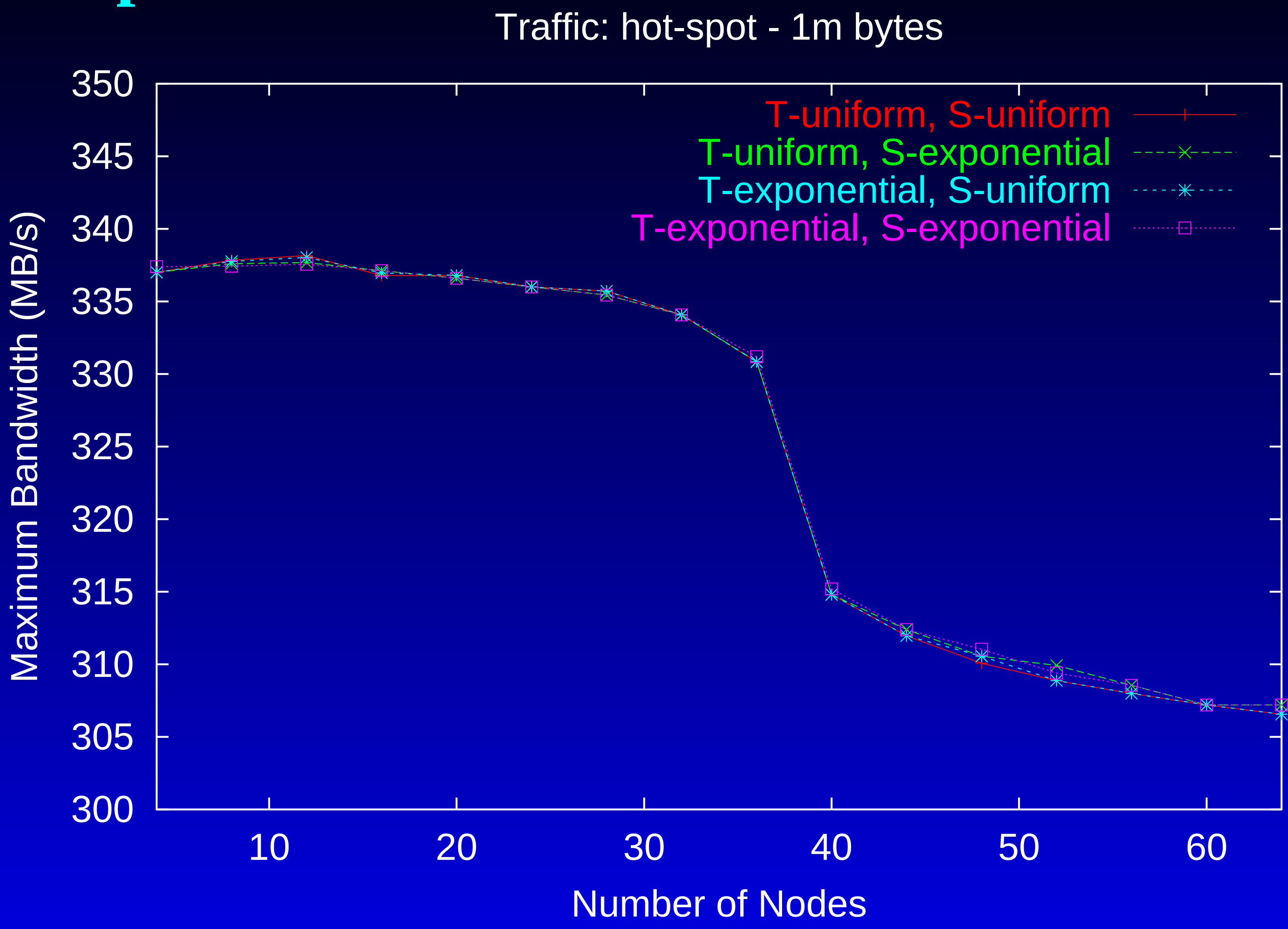


	Unidirectional	Bidirectional
Elan Memory	335 MB/s	280 MB/s
Main Memory	200 MB/s	80 MB/s

Hot-spot

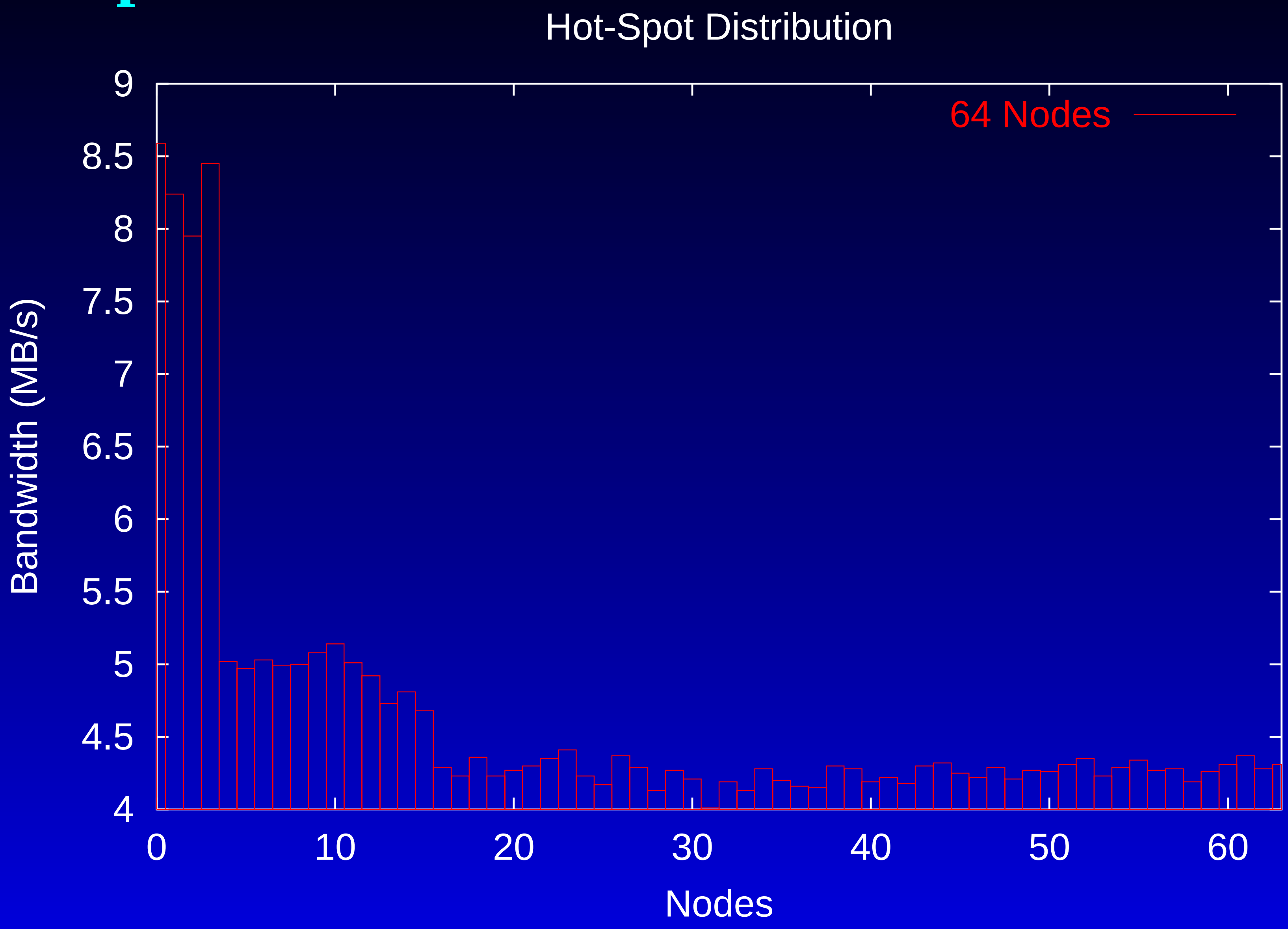
Objective: analyze the behavior of a single I/O node

Hot-spot



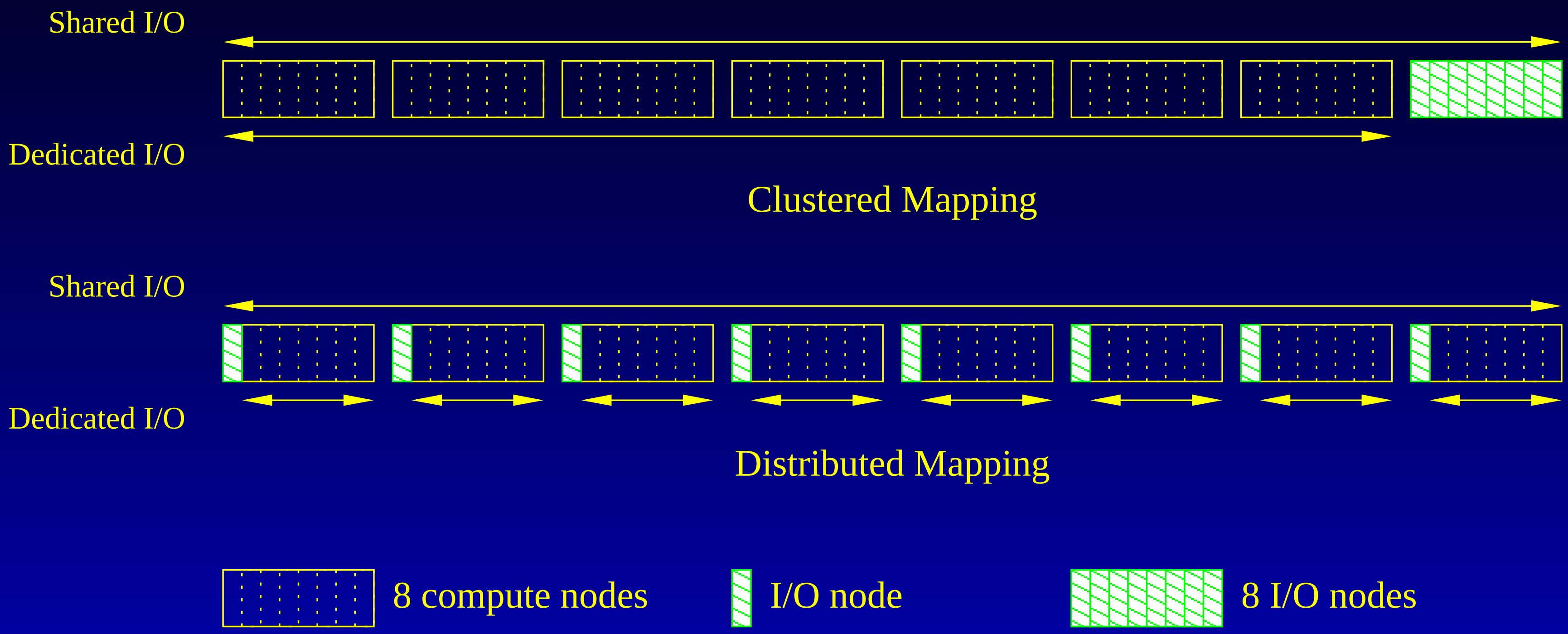
- Peak data bandwidth > **335 MB/s** up to 32 nodes

Hot-spot



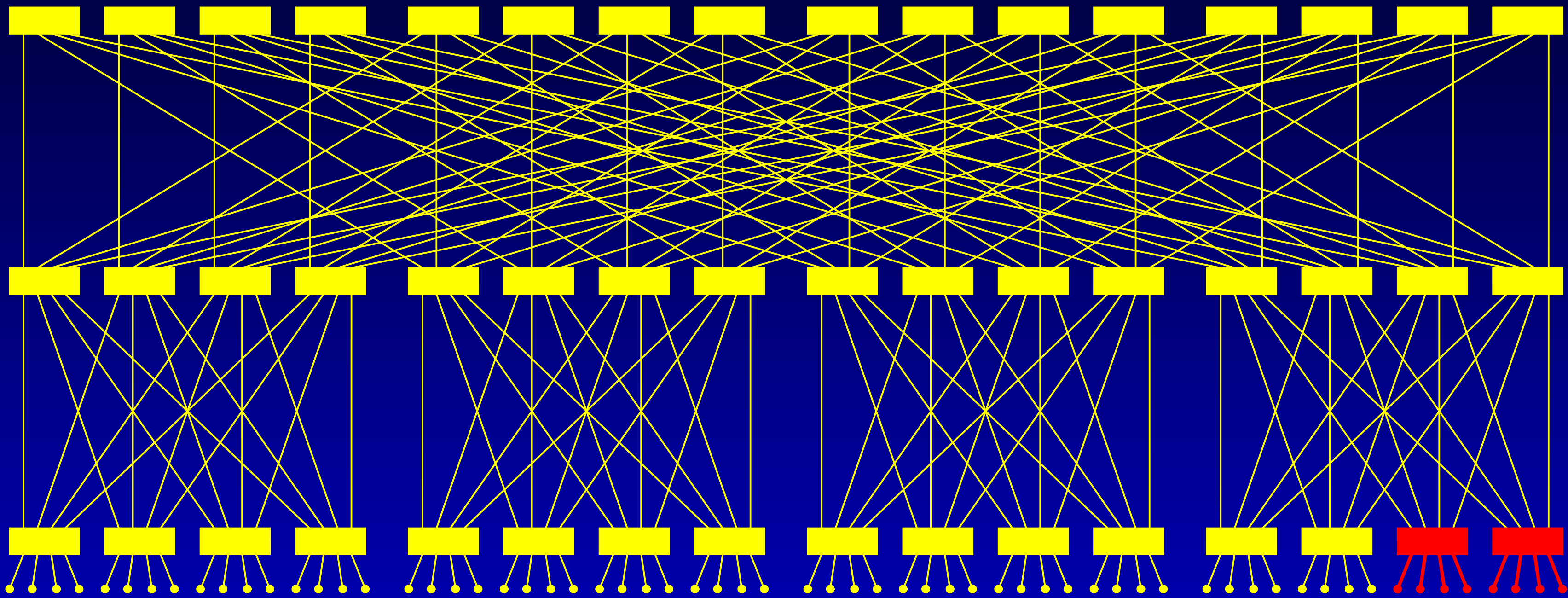
- Bandwith delivered to each node unevenly distributed

Multiple Hot-spots



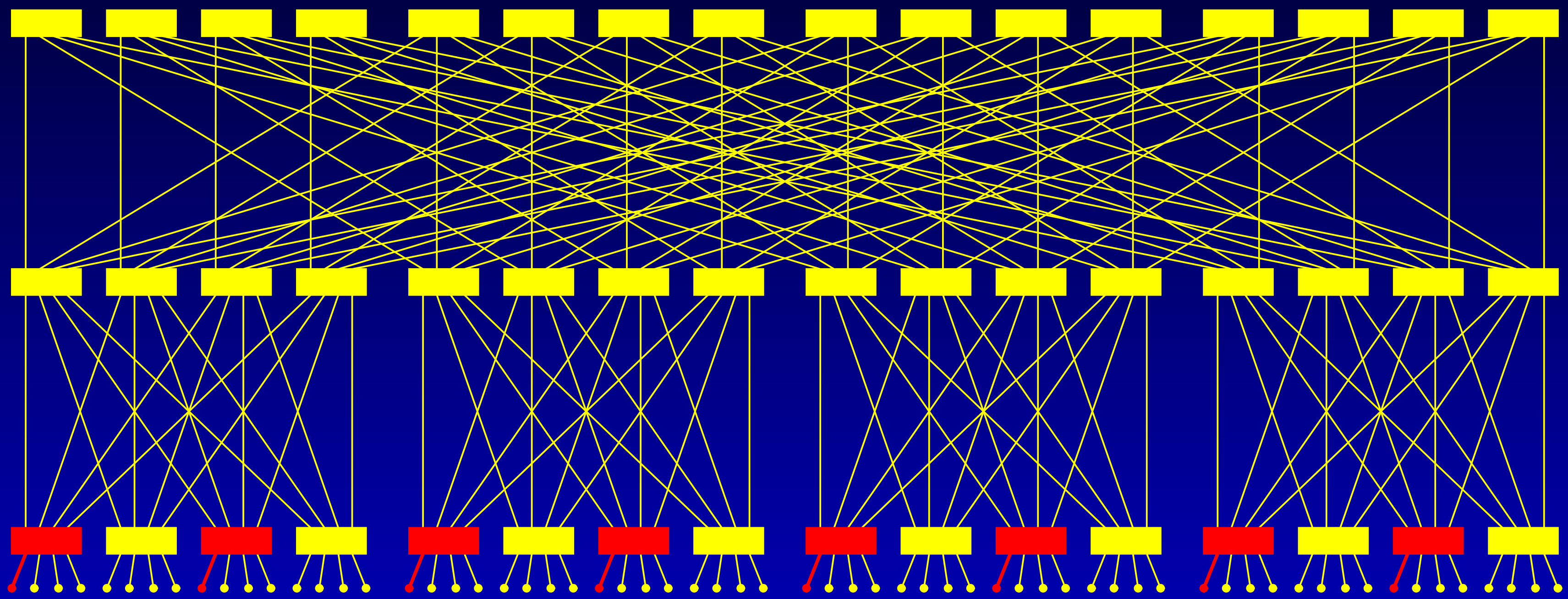
Multiple Hot-spots

Clustered I/O mapping



Multiple Hot-spots

Distributed I/O mapping

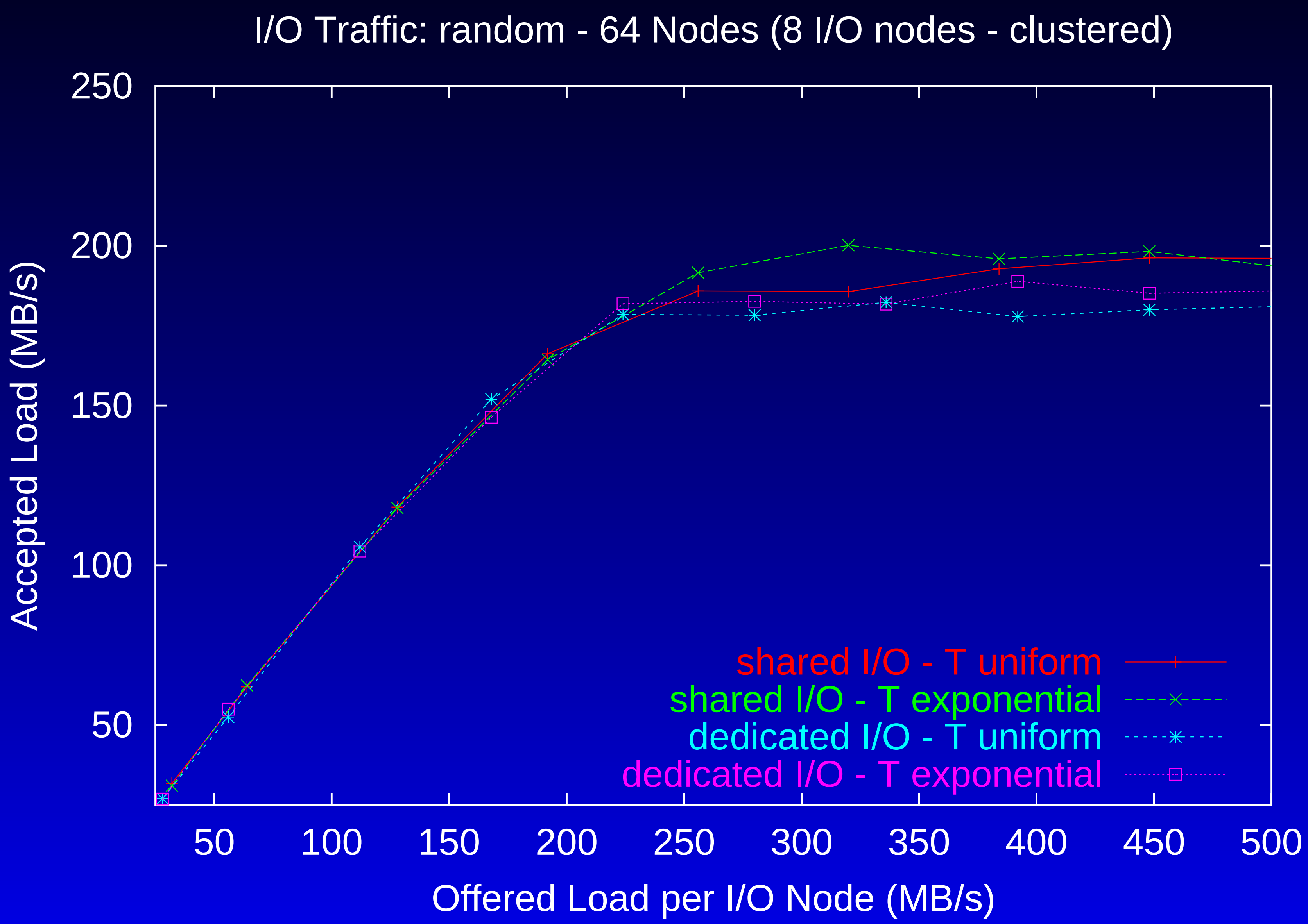


Multiple Hot-spots

Objectives:

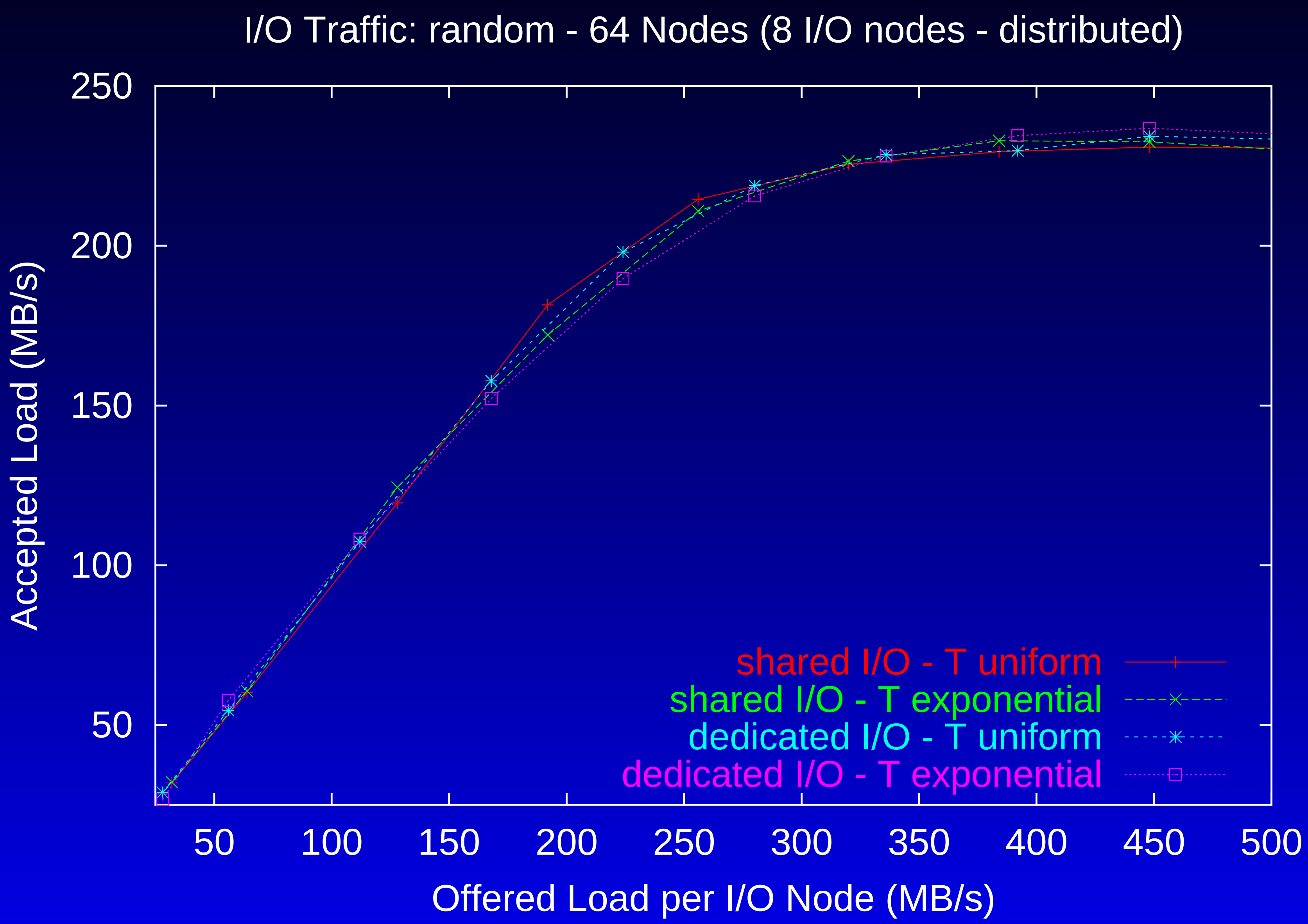
- behavior of multiple I/O nodes
- influence of the I/O node (hot-node) mapping: clustered and distributed
- effects of the application mapping: shared I/O and dedicated I/O
- influence of the traffic pattern: random and deterministic
- effect of the I/O read/write ratio

Multiple Hot-spots



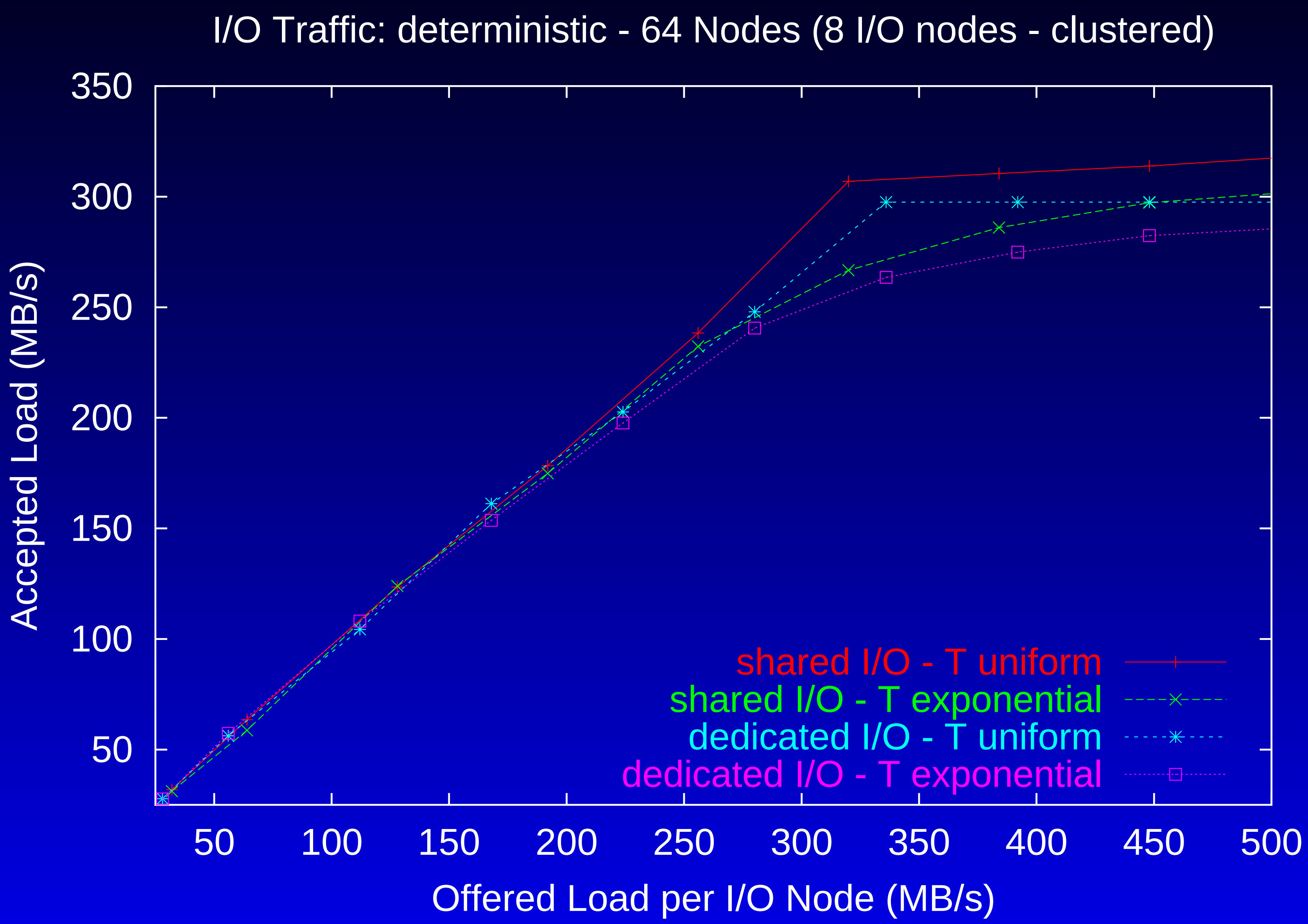
- Asymptotic bandwidth delivered by each I/O node of 196 MB/s

Multiple Hot-spots



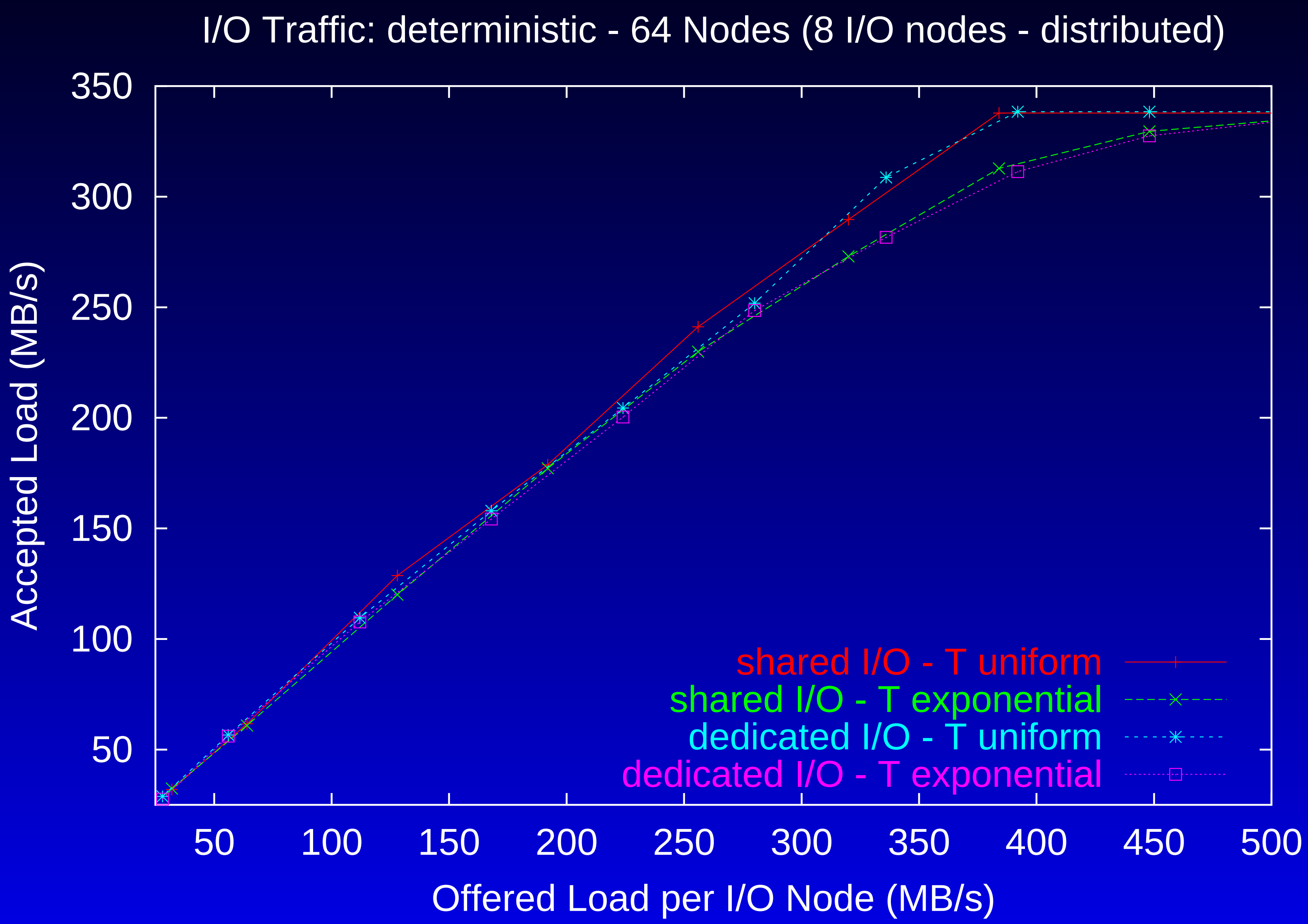
- Asymptotic bandwidth of 234 MB/s

Multiple Hot-spots



- Asymptotic bandwidth of 320 MB/s

Multiple Hot-spots



- Asymptotic bandwidth of 338 MB/s

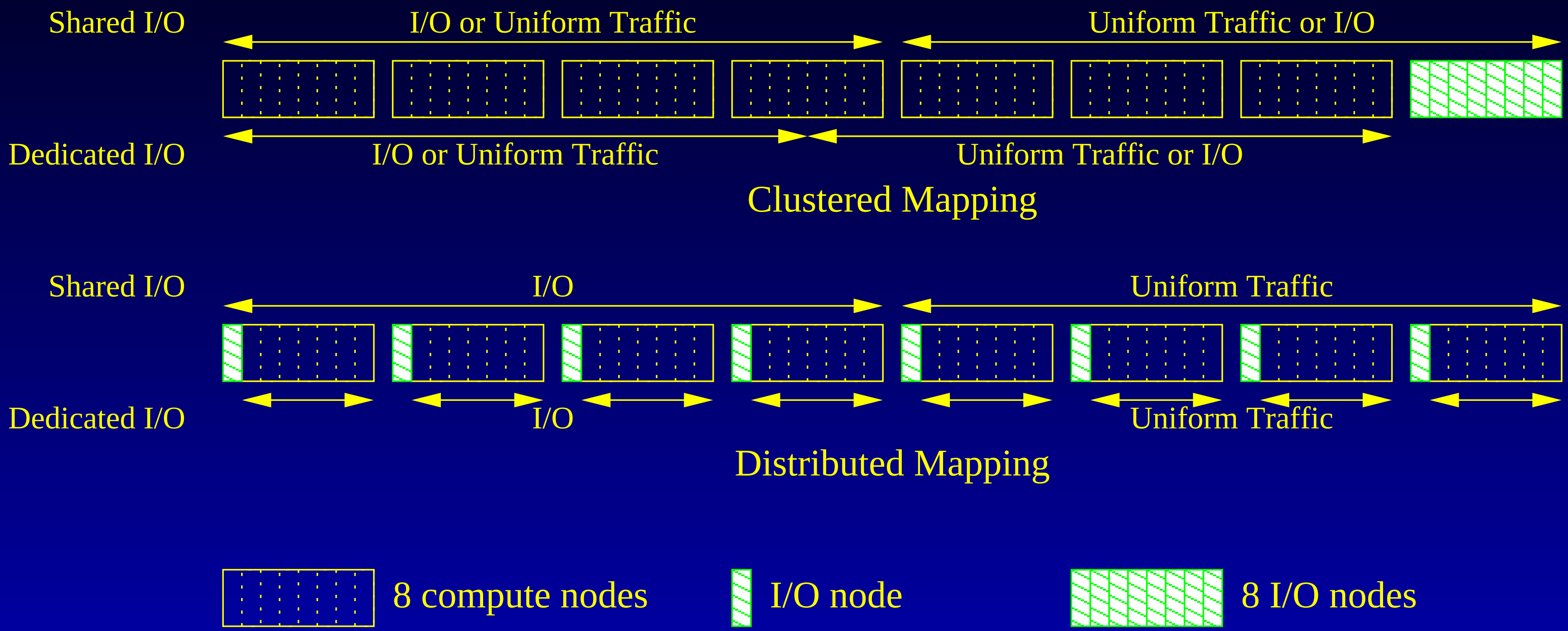
Multiple Hot-spots Summary

	Clustered I/O	Distributed I/O
Random Traffic	196 MB/s	234 MB/s
Deterministic Traffic	320 MB/s	338 MB/s

- Better results obtained with:
  - distributed I/O
  - deterministic traffic
- No significant effect of the application mapping
- Insensitive to read/write ratio
- Insensitive to time and message size distributions

Los Alamos NATIONAL LABORATORY

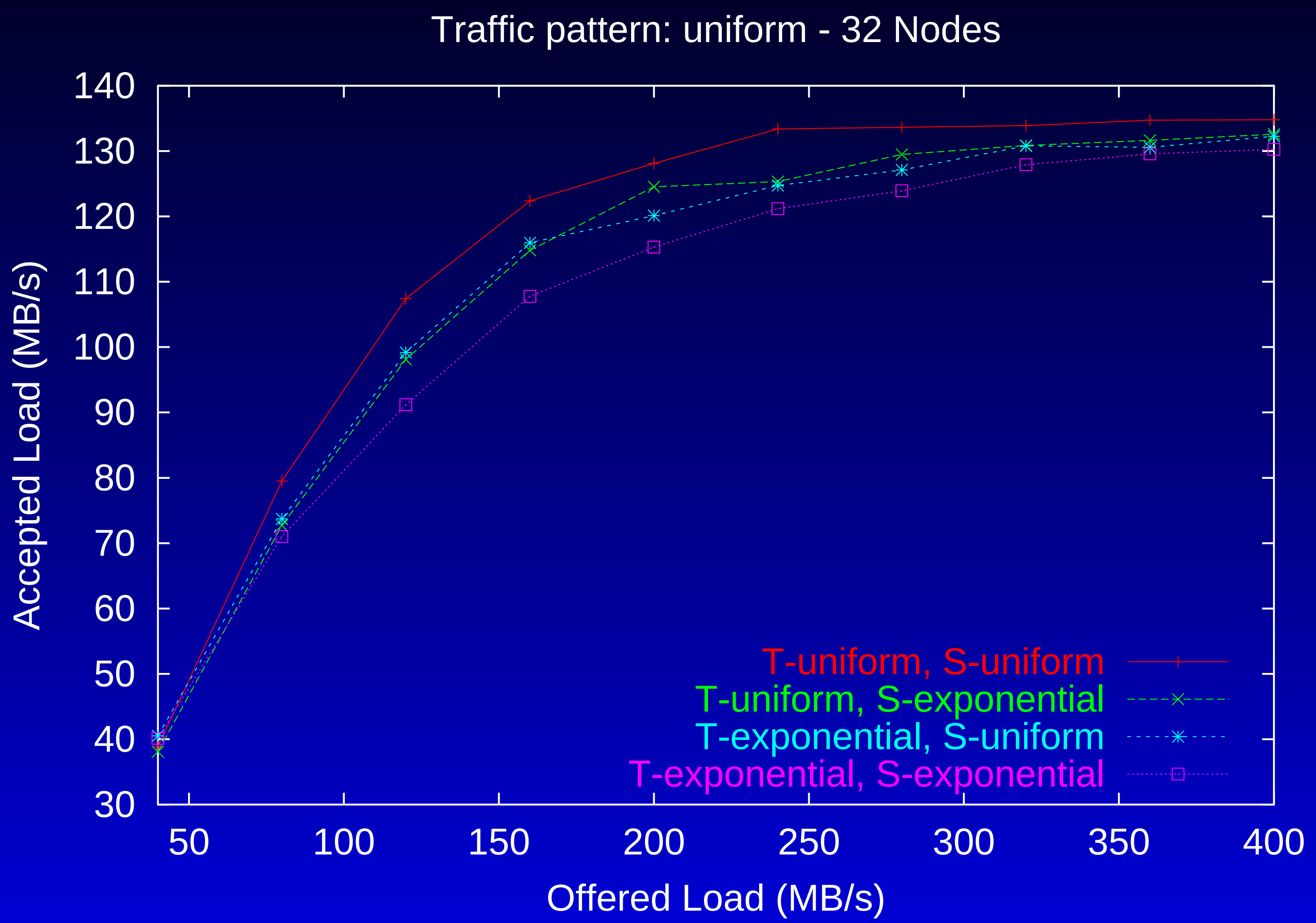
Combined Traffic



Objective:

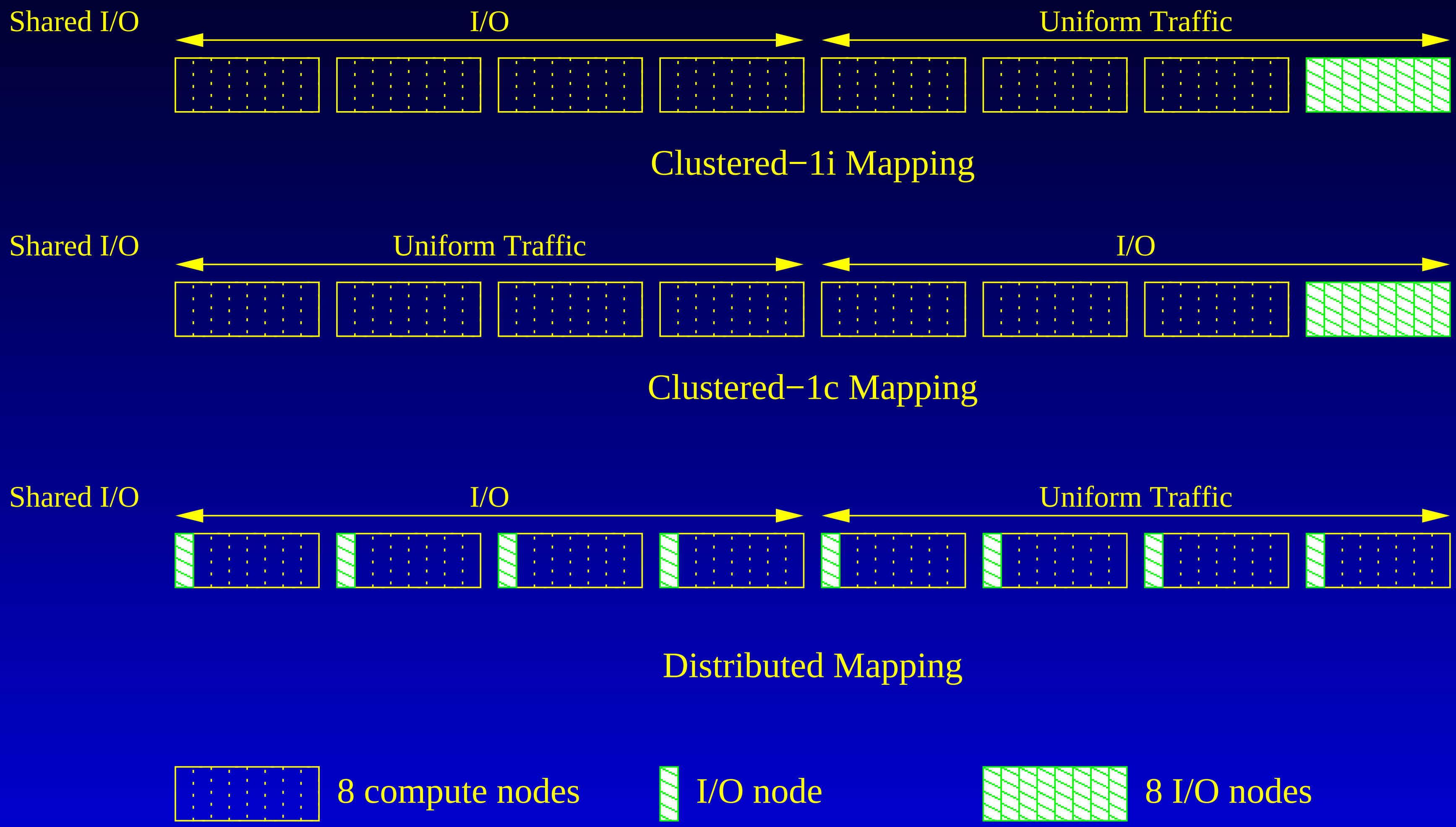
- interference of the I/O on a parallel job

Combined Traffic



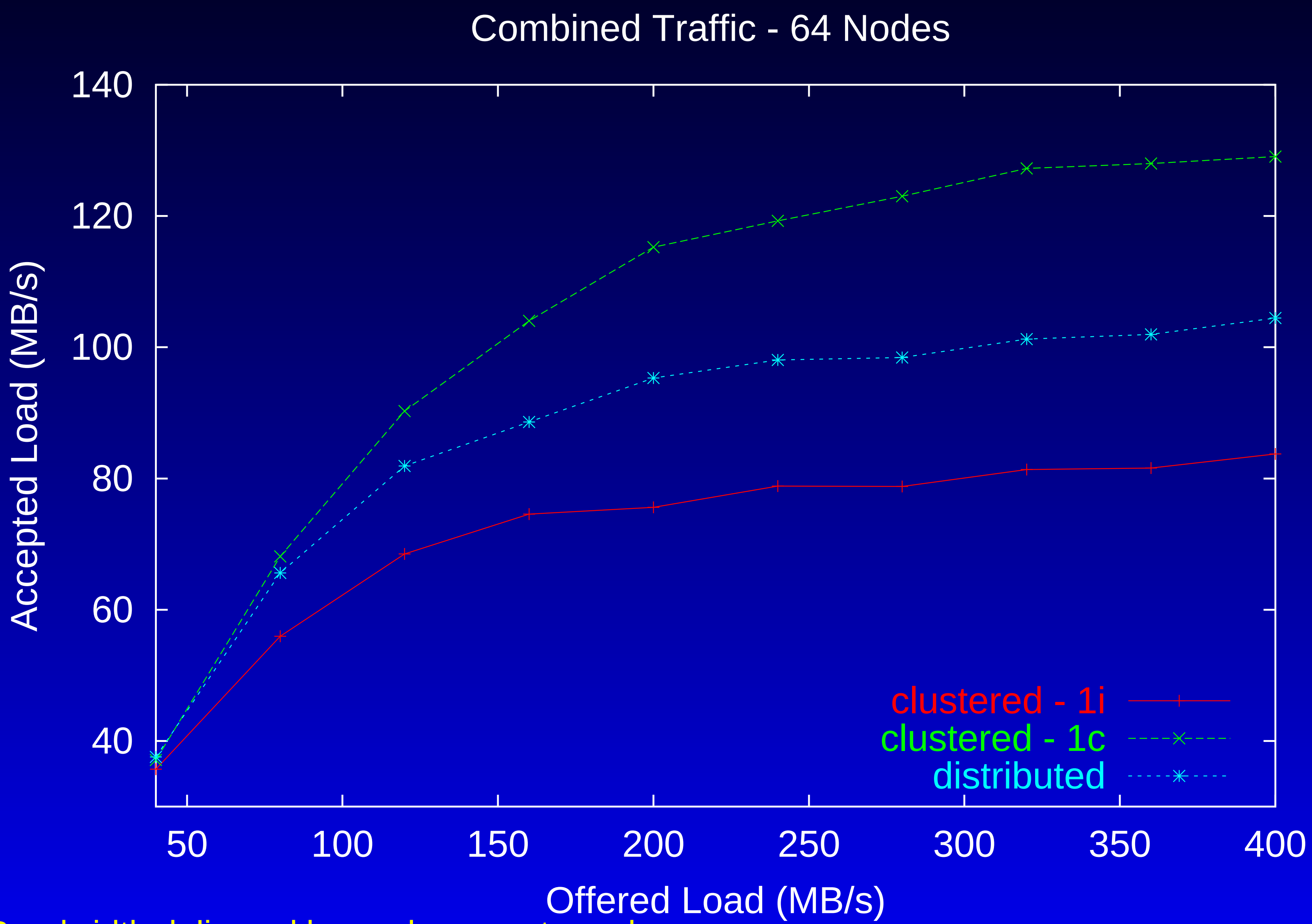
Uniform traffic with no background I/O. Results for 32 nodes.

Combined Traffic with Shared I/O



Combined Traffic with Shared I/O

I/O load = 0.1

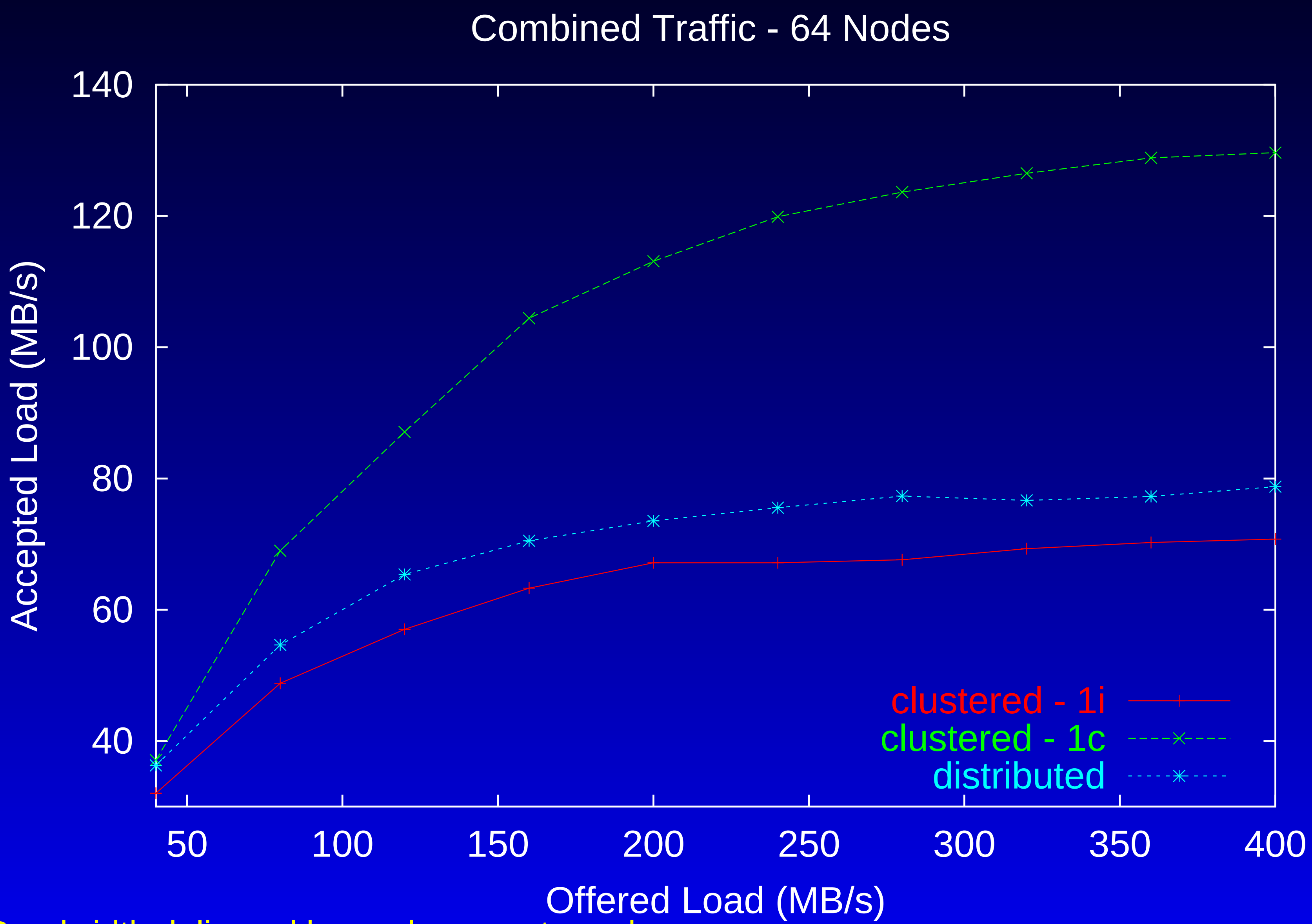


Bandwidth delivered by each compute node.

Los Alamos NATIONAL LABORATORY

Combined Traffic with Shared I/O

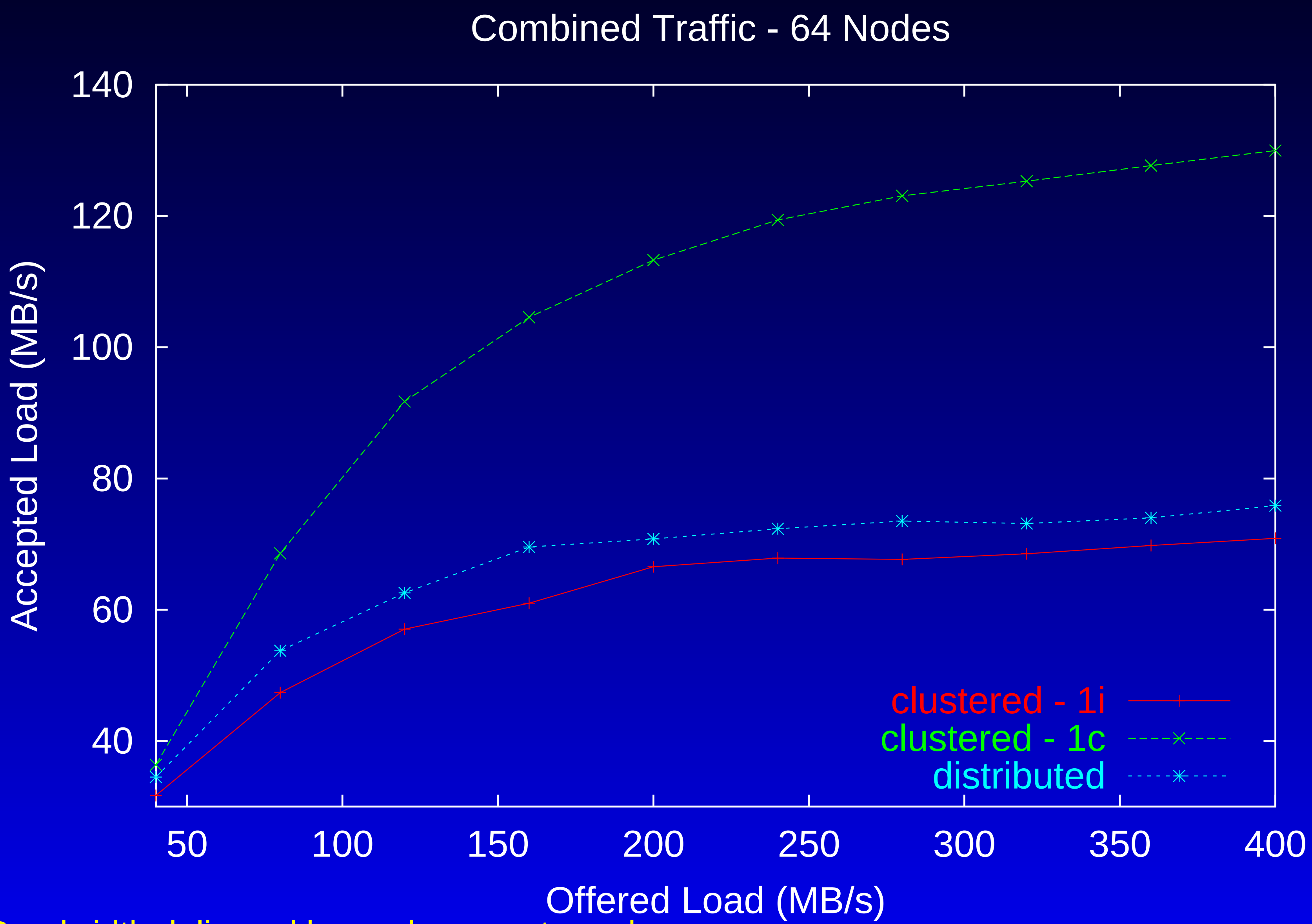
I/O load = 0.3



Bandwidth delivered by each compute node.

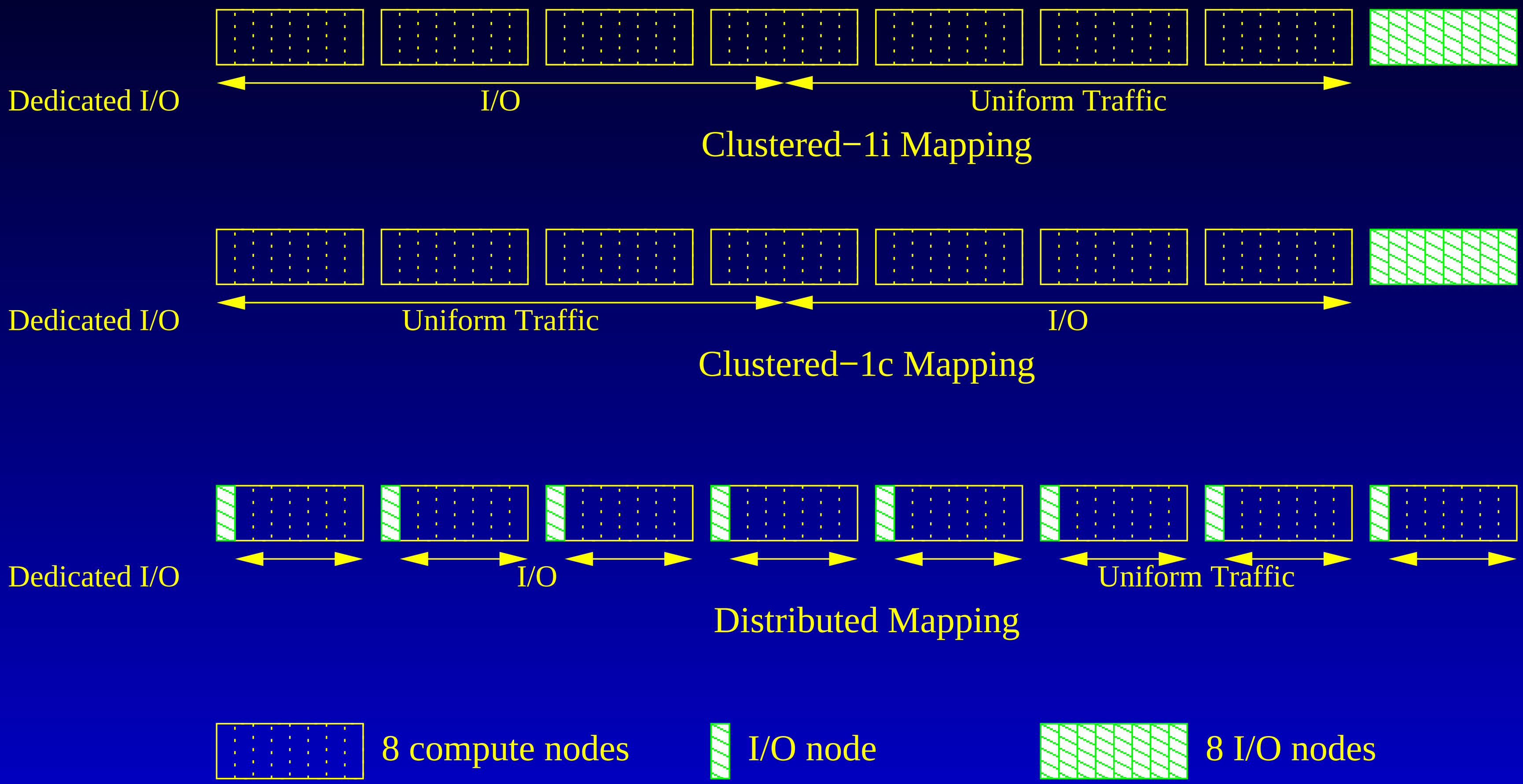
Combined Traffic with Shared I/O

I/O load = 0.5



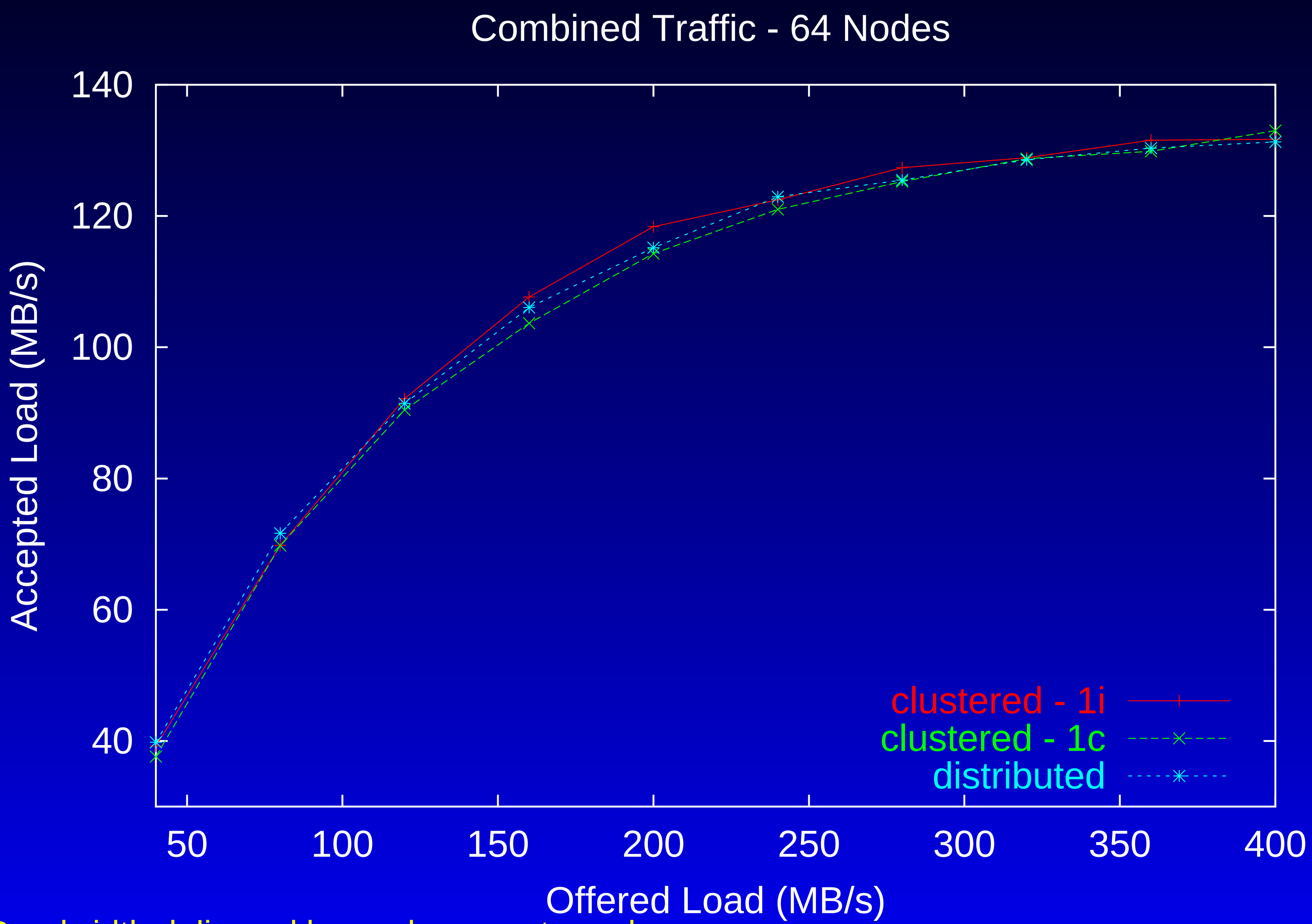
Bandwidth delivered by each compute node.

Combined Traffic with Dedicated I/O



Combined Traffic with Dedicated I/O

I/O load = 0.1

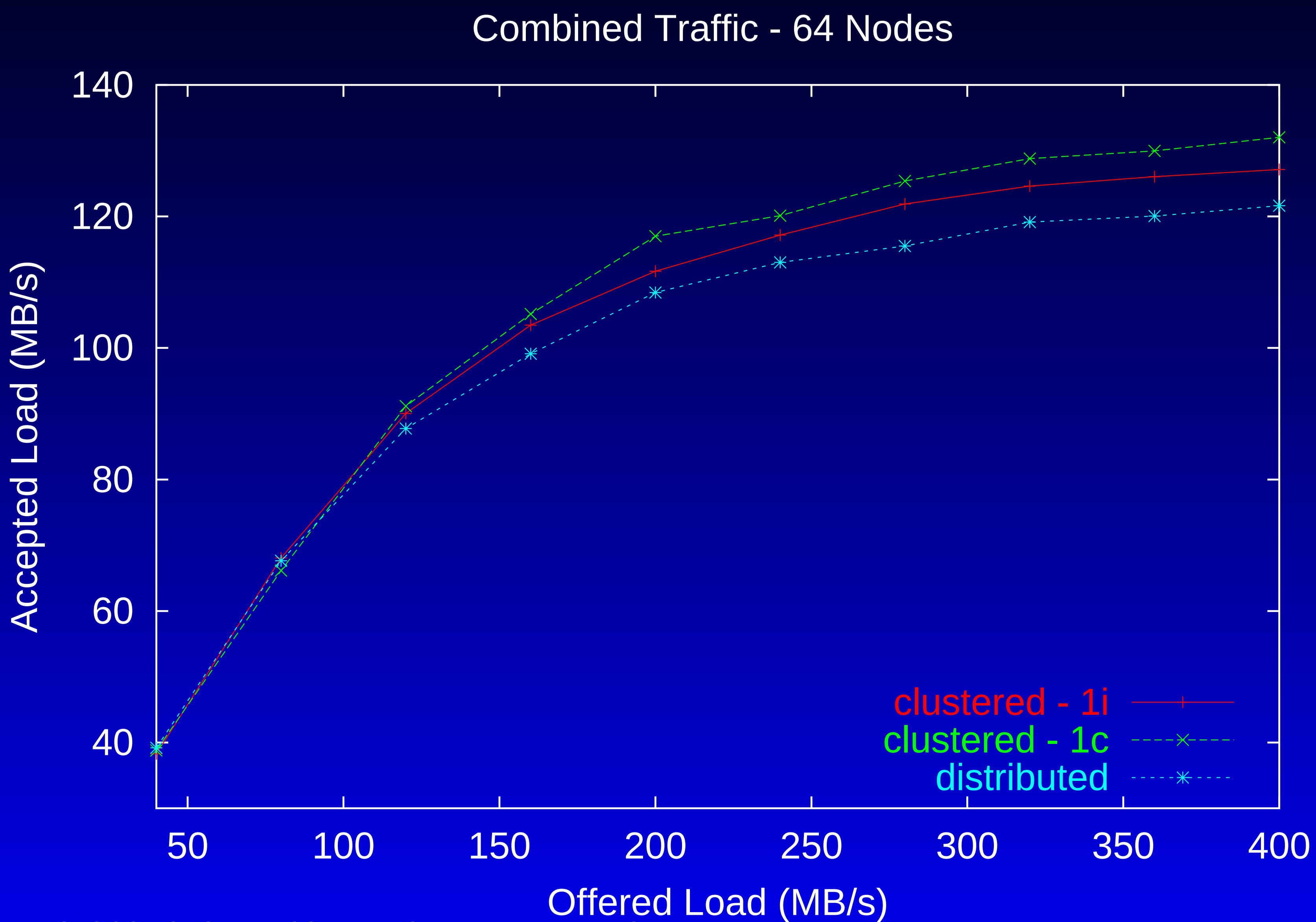


Bandwidth delivered by each compute node.

Los Alamos NATIONAL LABORATORY

Combined Traffic with Dedicated I/O

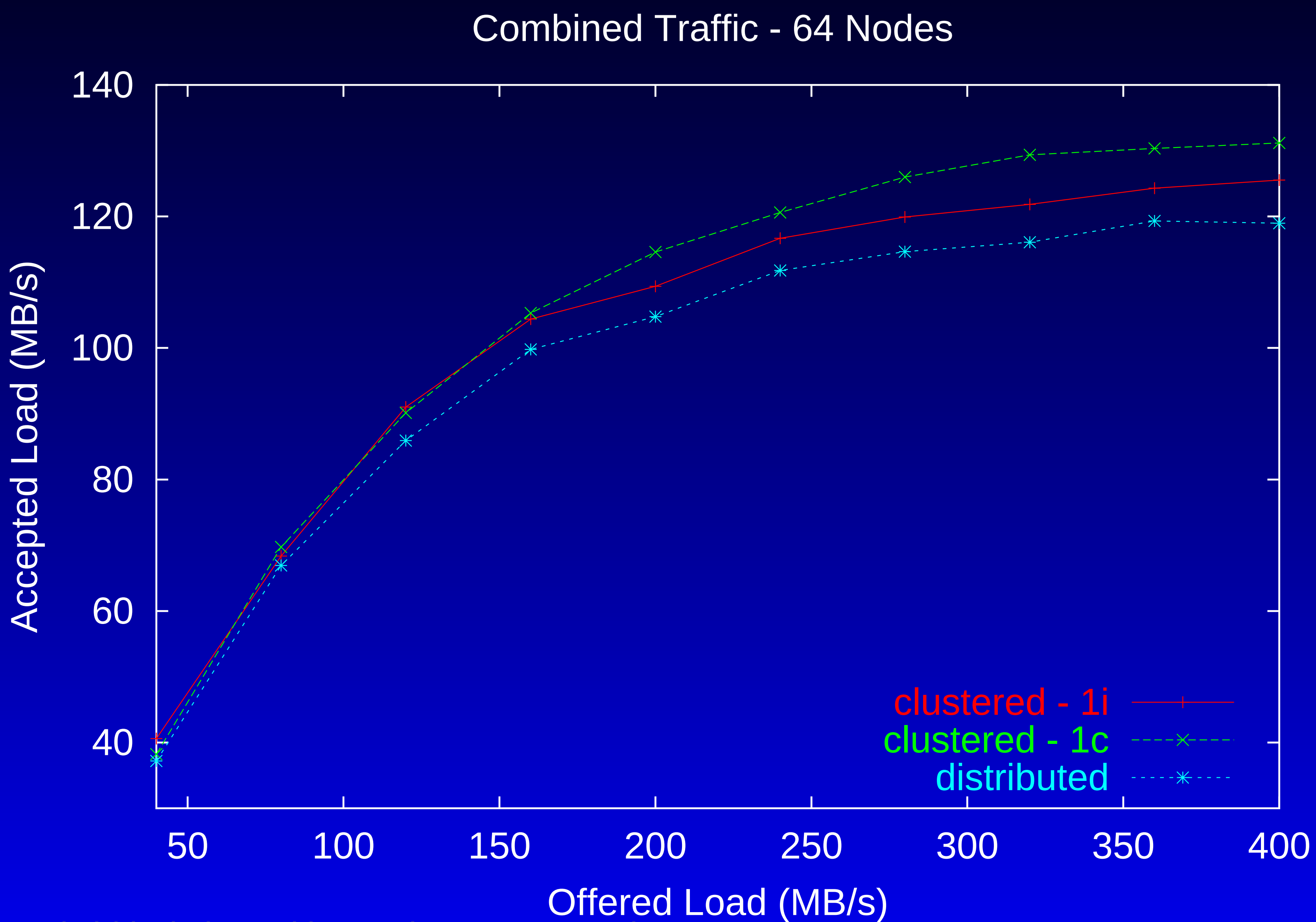
I/O load = 0.3



Bandwidth delivered by each compute node.

Combined Traffic with Dedicated I/O

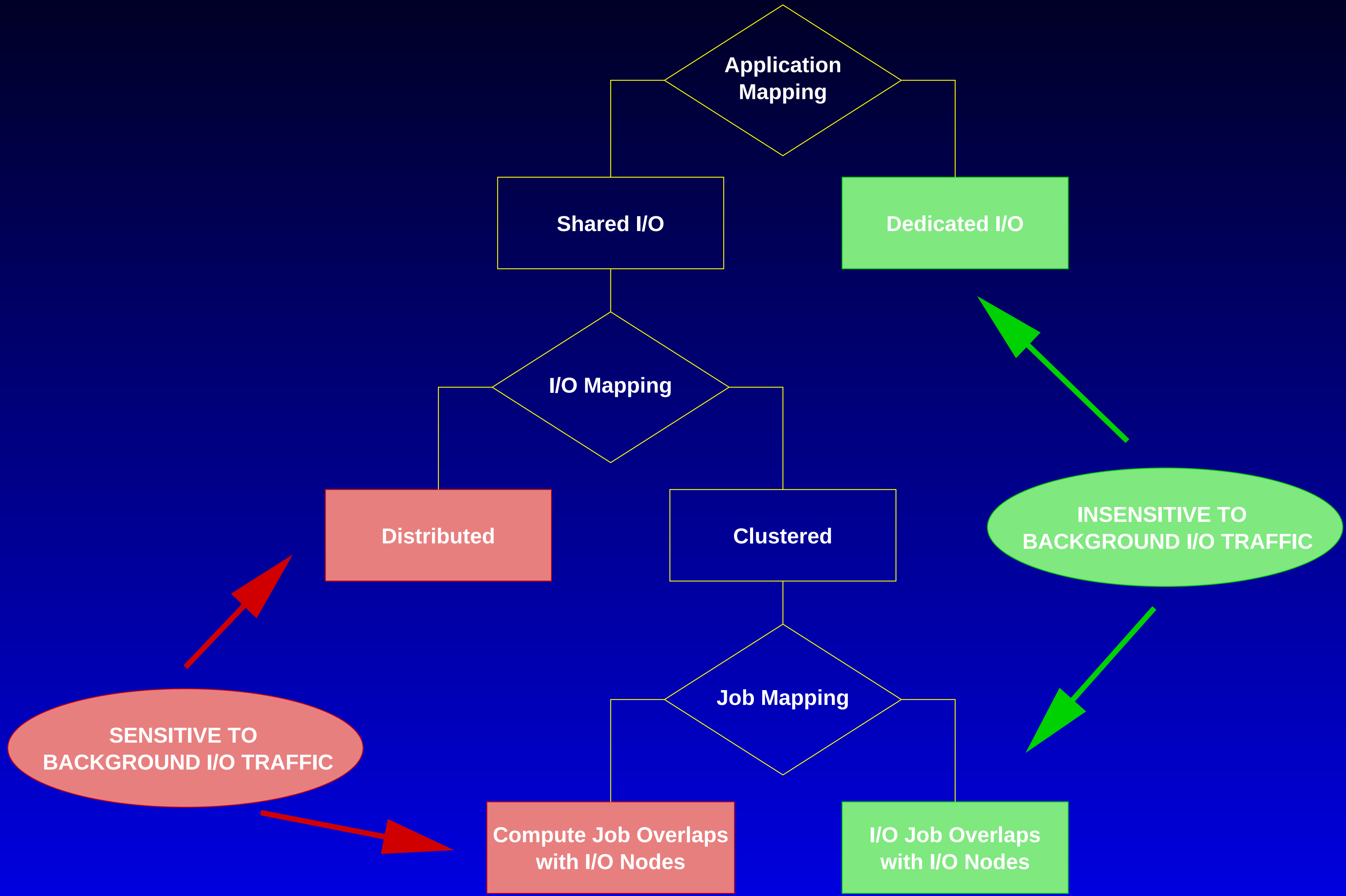
I/O load = 0.5



Bandwidth delivered by each compute node.

Los Alamos NATIONAL LABORATORY

Combined Traffic Summary



Los Alamos NATIONAL LABORATORY

Conclusions

- A single hot-node (I/O server) can handle, without performance degradation, traffic generated by up to 32 nodes.

Conclusions

- A single hot-node (I/O server) can handle, without performance degradation, traffic generated by up to 32 nodes.

- With multiple I/O servers it is more efficient to distribute them rather than cluster them, with a bandwidth increase of up to 20%.
- The performance is insensitive to both the fraction of I/O reads and writes and to the mapping of the parallel job.

Los Alamos
NATIONAL LABORATORY

Conclusions

- A single hot-node (I/O server) can handle, without performance degradation, traffic generated by up to 32 nodes.
- With multiple I/O servers it is more efficient to distribute them rather than cluster them, with a bandwidth increase of up to 20%.
- The performance is insensitive to both the fraction of I/O reads and writes and to the mapping of the parallel job.
- Multiple jobs can be run in parallel without interference, as long as these jobs are not mapped on the I/O nodes.
- The I/O job can interfere with the compute job when the latter is mapped on the I/O nodes.

Los Alamos NATIONAL LABORATORY

Additional Information

http://www.c3.lanl.gov/~fabrizio/quadrics.html

Los Alamos NATIONAL LABORATORY

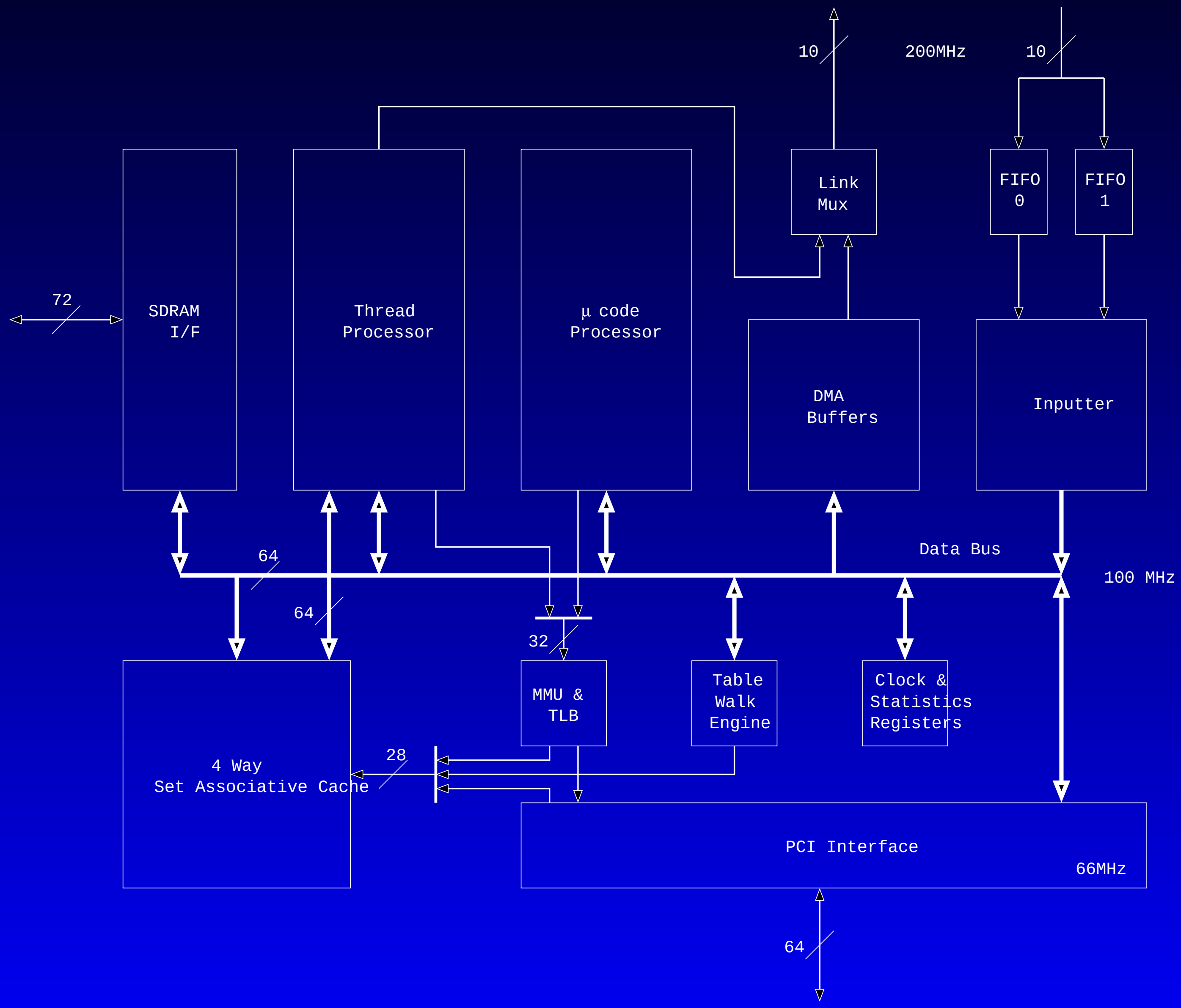
APPENDIX

Quadrics Network Design Overview

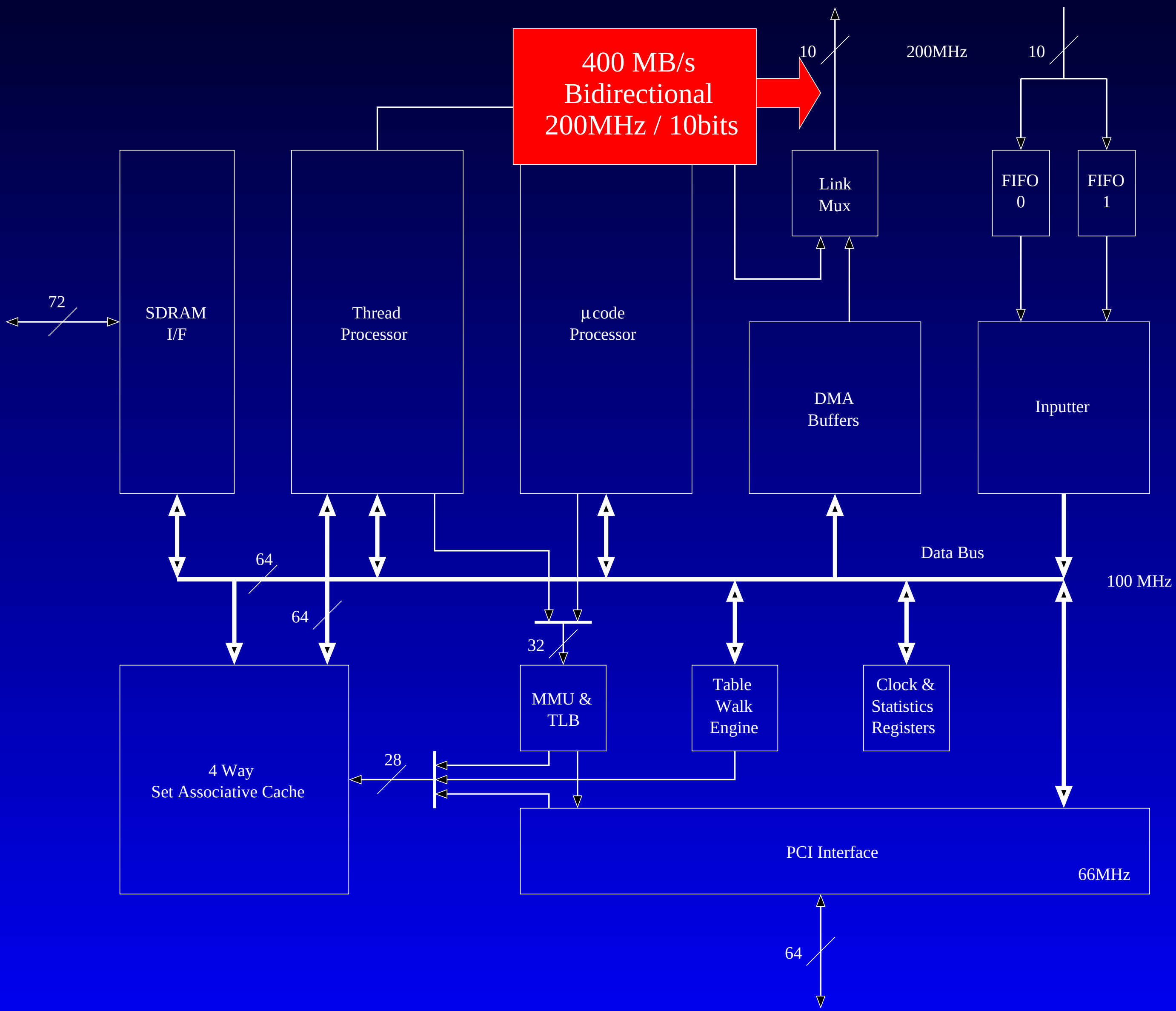
- QsNET provides an abstraction of distributed virtual shared memory
- Each process can map a portion of its address space into the global memory
- These address spaces constitutes the virtual shared memory
- This shared memory is fully integrated with the native operating system
- Based on two building blocks:
  - a network interface card called Elan
  - a crossbar switch called Elite

Los Alamos NATIONAL LABORATORY

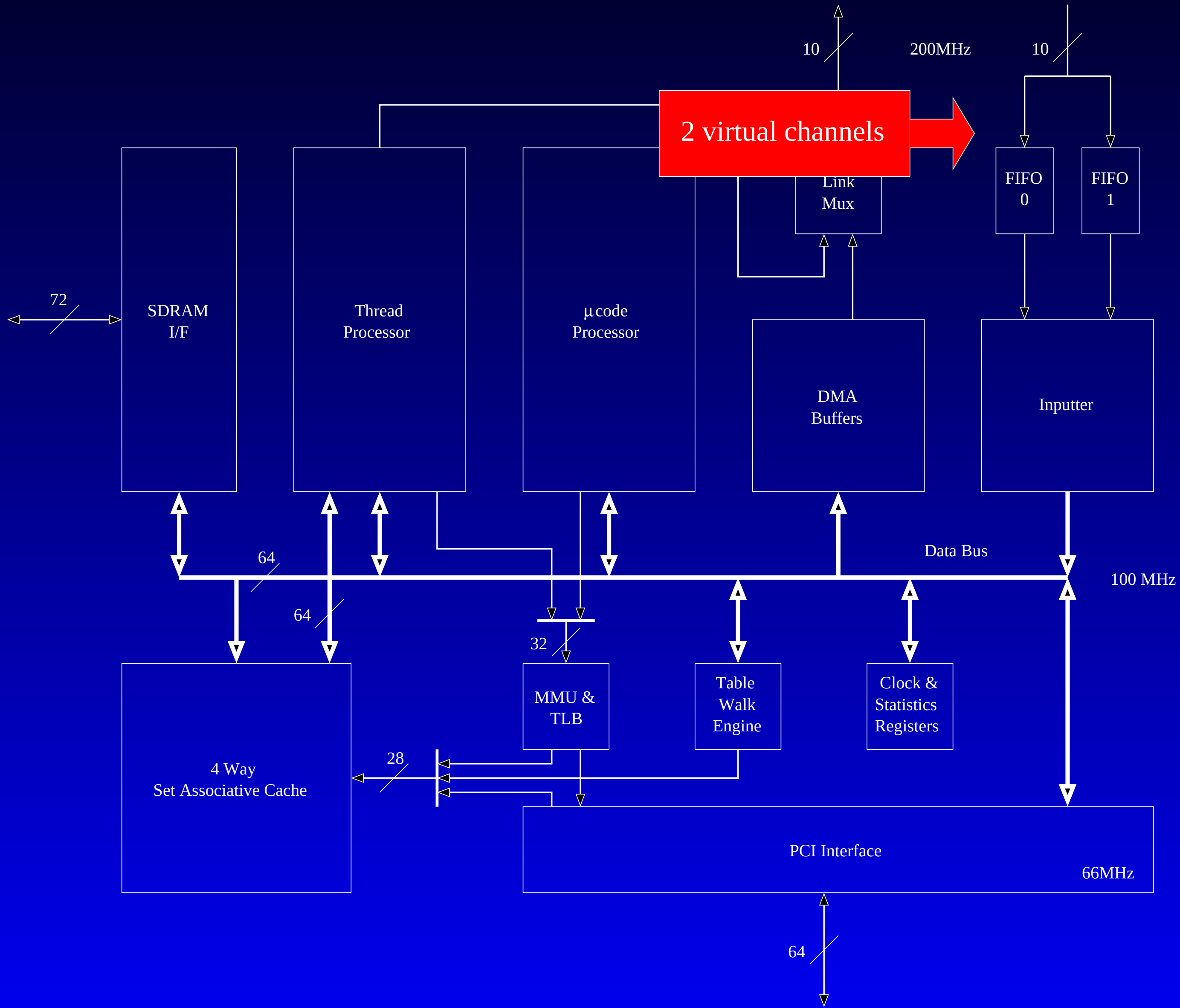
Elan


10
200MHz
10
Link Mux
FIFO 0
FIFO 1
72
SDRAM I/F
Thread Processor
µ code Processor
DMA Buffers
Inputter
Data Bus
64
100 MHz
64
32
MMU & TLB
Table Walk Engine
Clock & Statistics Registers
4 Way Set Associative Cache
28
PCI Interface
66MHz
64

Elan

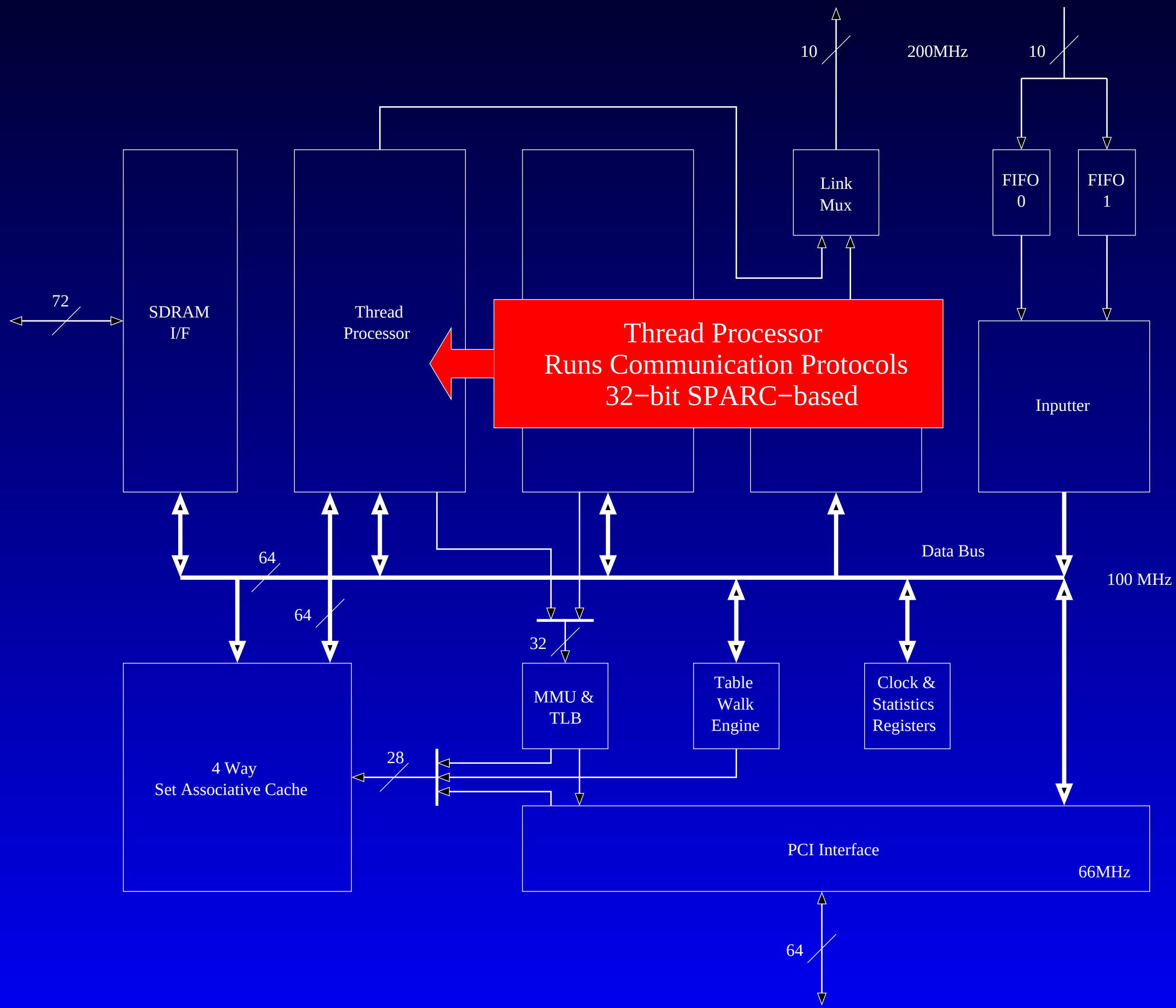

400 MB/s
Bidirectional
200MHz / 10bits
10
200MHz
10
Link
Mux
FIFO
0
FIFO
1
72
SDRAM
I/F
Thread
Processor
µcode
Processor
DMA
Buffers
Inputter
Data Bus
100 MHz
64
64
32
MMU &
TLB
Table
Walk
Engine
Clock &
Statistics
Registers
4 Way
Set Associative Cache
28
PCI Interface
66MHz
64

Elan

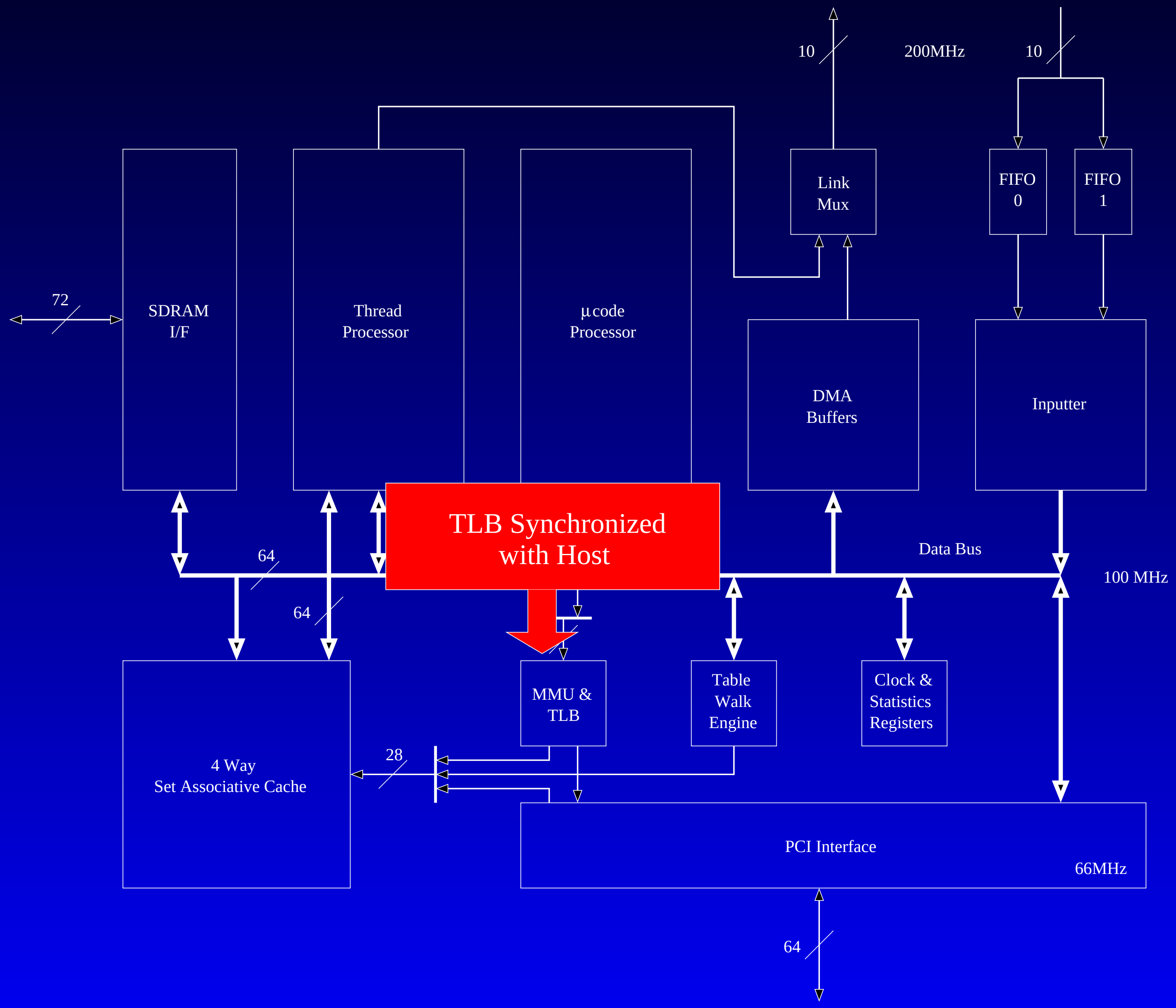


10
200MHz
10
2 virtual channels
Link
Mux
FIFO
0
FIFO
1
72
SDRAM
I/F
Thread
Processor
µcode
Processor
DMA
Buffers
Inputter
Data Bus
64
100 MHz
64
32
MMU &
TLB
Table
Walk
Engine
Clock &
Statistics
Registers
4 Way
Set Associative Cache
28
PCI Interface
66MHz
64

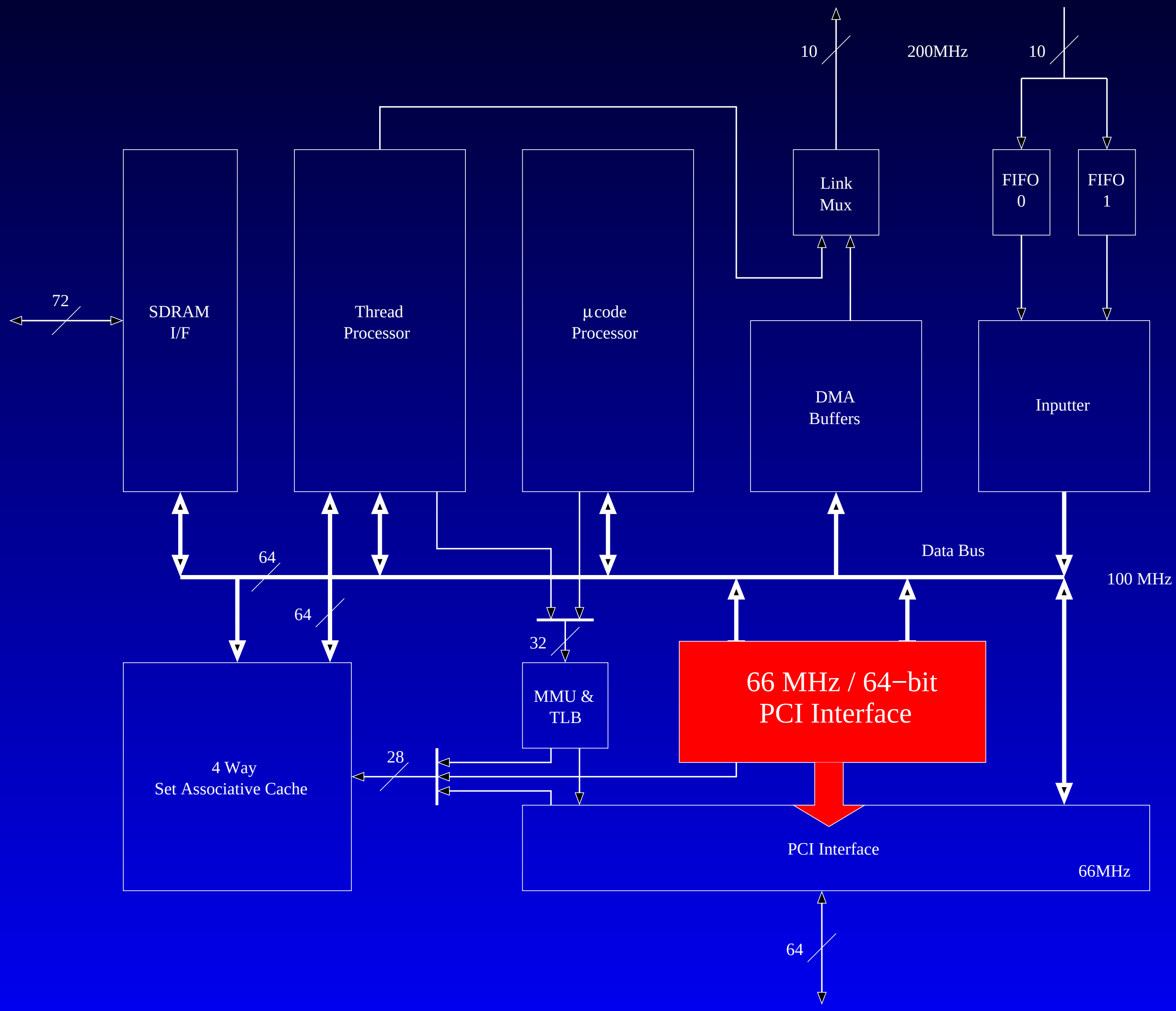
Elan


10
200MHz
10
Link Mux
FIFO 0
FIFO 1
72
SDRAM I/F
Thread Processor
Thread Processor
Runs Communication Protocols
32−bit SPARC−based
Inputter
Data Bus
100 MHz
64
64
32
MMU & TLB
Table Walk Engine
Clock & Statistics Registers
28
4 Way Set Associative Cache
PCI Interface
66MHz
64

Elan


10
200MHz
10
Link
Mux
FIFO
0
FIFO
1
72
SDRAM
I/F
Thread
Processor
µcode
Processor
DMA
Buffers
Inputter
TLB Synchronized
with Host
64
Data Bus
100 MHz
64
MMU &
TLB
Table
Walk
Engine
Clock &
Statistics
Registers
4 Way
Set Associative Cache
28
PCI Interface
66MHz
64

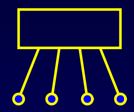
Elan


10
200MHz
10
Link
Mux
FIFO
0
FIFO
1
72
SDRAM
I/F
Thread
Processor
µcode
Processor
DMA
Buffers
Inputter
Data Bus
100 MHz
64
64
32
MMU &
TLB
66 MHz / 64−bit
PCI Interface
4 Way
Set Associative Cache
28
PCI Interface
66MHz
64

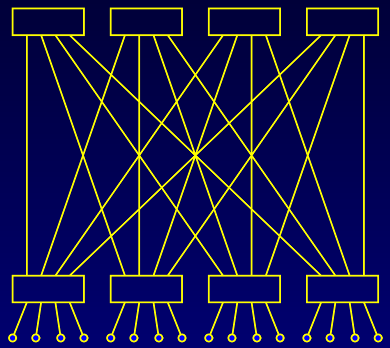
Elite

- 8 bidirectional links with 2 virtual channels in each direction
- An internal 16x8 full crossbar switch
- 400 MB/s on each link direction
- Packet error detection and recovery, with routing and data transactions CRC protected
- 2 priority levels plus an aging mechanism
- Adaptive routing
- Hardware support for broadcast

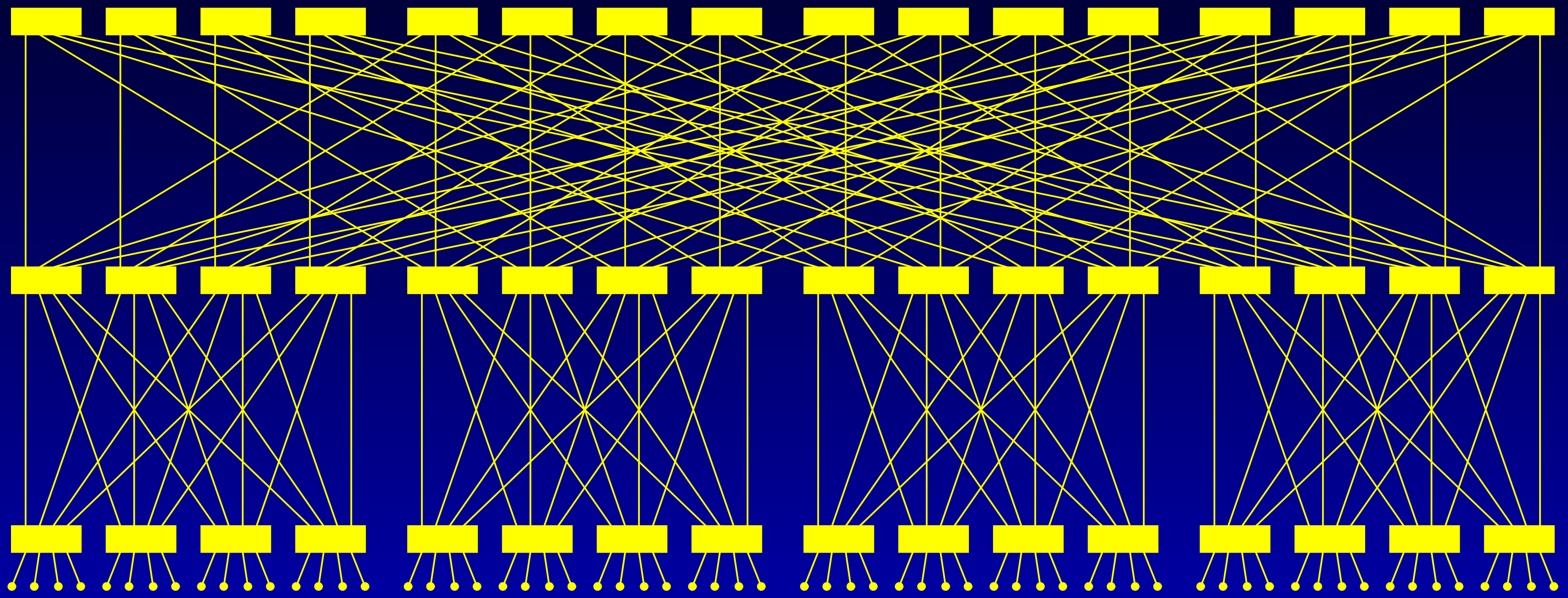
Network Topology: Quaternary Fat-Tree



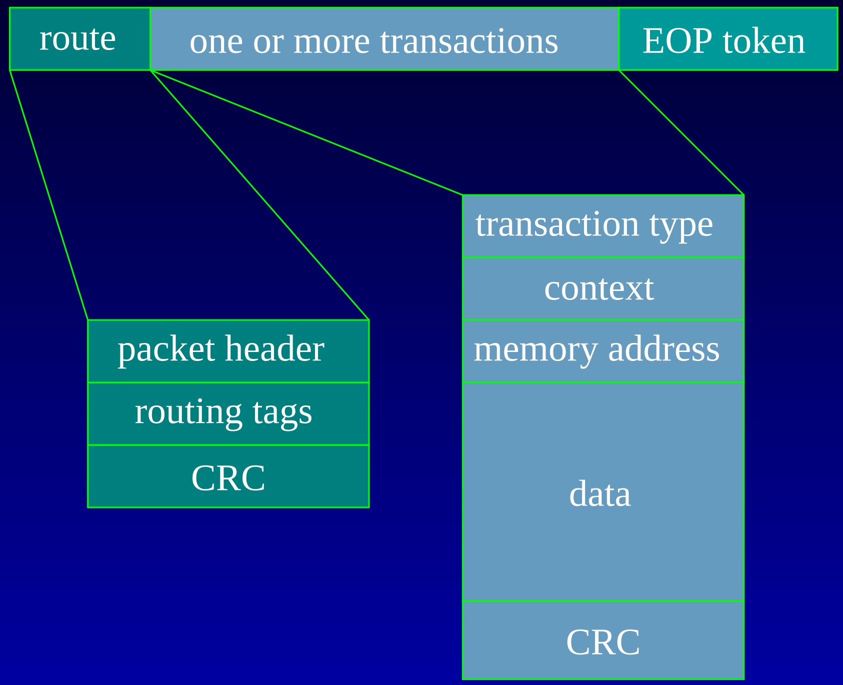
Network Topology: Quaternary Fat-Tree



Network Topology: Quaternary Fat-Tree



Packet Format



- 320 bytes data payload (5 transactions with 64 bytes each)
- 74-80 bytes overhead

Los Alamos
NATIONAL LABORATORY

Programming Libraries

- Elan3lib
  - event notification
  - memory mapping and allocation
  - remote DMA
- Elanlib and Tports
  - collective communication
  - tagged message passing
- MPI, shmem