INFINIBAND NETWORK ANALYSIS AND MONITORING USING OPENSM

N. Dandapanthula[1], H. Subramoni[1], J. Vienne[1] , K. Kandalla[1] , S. Sur[1] , D. K. Panda[1] , and R. Brightwell[2]

Presented By Xavier Besseron[1]

Date: 08/30/2011

[1]Network-Based Computing Laboratory, The Ohio State University

[2]Sandia National Laboratories

Outline

- Introduction
  - InfiniBand
  - OpenSM
  - Problem Statement
- INAM – Scalable InfiniBand Network Analysis & monitoring tool
- Experimental Analysis
- Conclusions and Future work

InfiniBand

- An industry standard for low latency, high bandwidth System Area Networks

- 41.20% of the top 500 most powerful supercomputers in the world are based on the InfiniBand interconnects (JUNE 2011)
  - Pleiades – 111,104 cores - NASA
  - Road Runner – 122,400 cores - LANL
  - Red Sky – 42,440 cores - Sandia National Labs
  - Ranger – 62,976 cores - TACC

- Multiple Virtual Lanes (VL) supported by IB
  - Logical channel under the same physical link
  - Separate buffer and flow control
  - Service Differentiation

OpenSM

- InfiniBand Subnet Manager (IBA Specifications)
- Part of OFED software package
  - Open Fabrics Enterprise Distribution
  - Open source software for RDMA and kernel bypass applications
  - Needed by the HPC community for applications which need low latency and high efficiency and fast I/O
- Scans, Initiates and Monitors the InfiniBand Fabric
- Performance Counters and Subnet Management Attributes (Not supported at VL granularity)
- Subnet Manager (SM), Subnet Management Agent (SMA)
- At least one instance required per Subnet
- Usage of Virtual Lanes

Existing Monitoring Tools

- Nagios [Agent Based]
  - + Easily Integratable & Configurable
  - + Supports multiple interconnects
  - - No discovery process
  - - Involves more overhead
  - - No Layer 2, Switch Dependent

- Ganglia [Agent Based]
  - + Portable and Scalable
  - + Distributed Modules provide higher sampling rates
  - + Supports multiple interconnects
  - - Use of Daemons (gmond) involves more overhead
  - - Metric measurements in compiled code
  - - Adding custom metrics can be a bit complicated

Existing Monitoring Tools (Contd)

- Fabric IT [Agent Less]
  - + Good Sampling Rates
  - + Agent less
  - + Integrated into the Subnet Manager
  - - Proprietary by Mellanox, Specific for IB
  - - Does not show communication patterns or Link usage pertaining to a Job
  - - No long term data storage

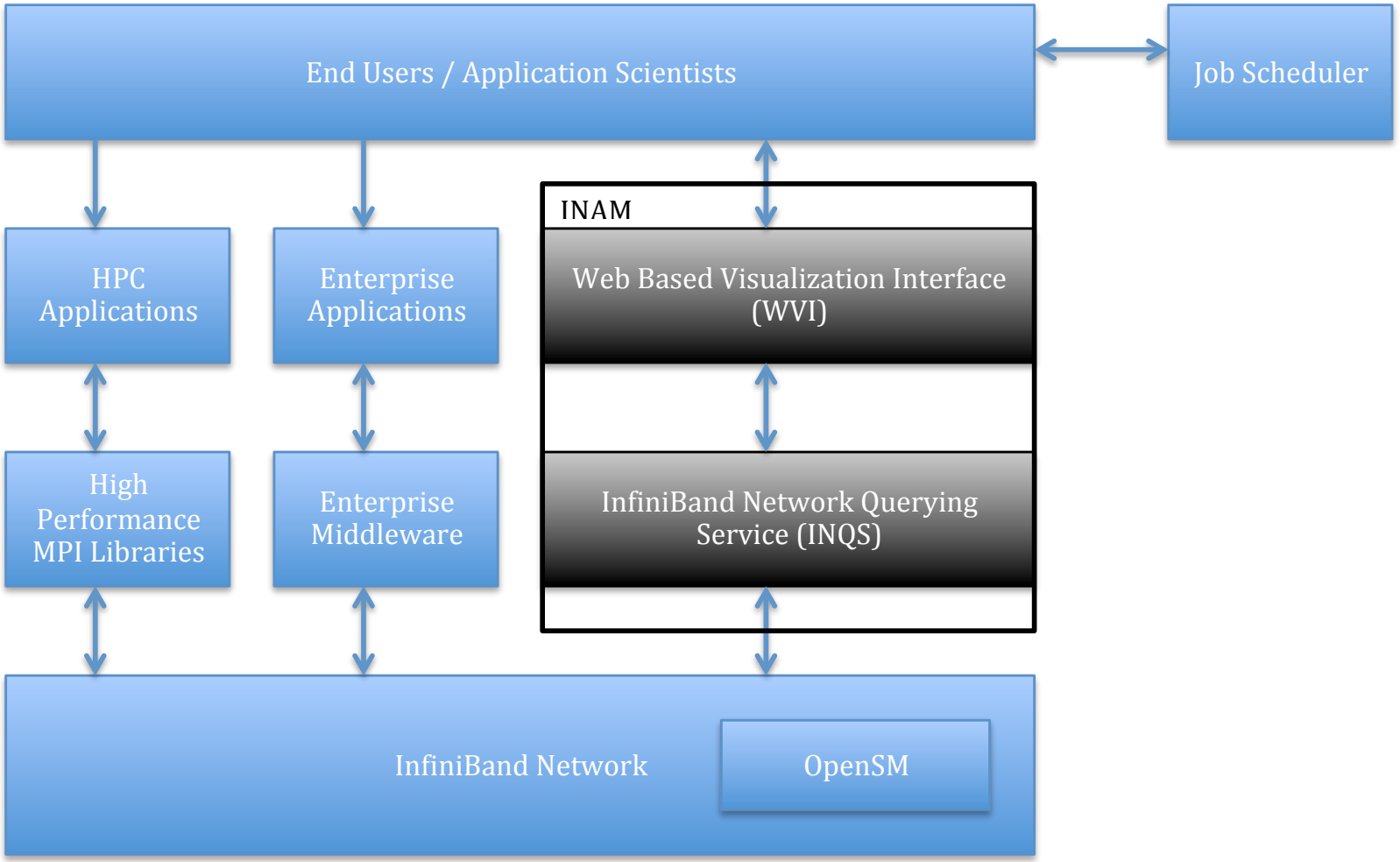
InfiniBand Network Analysis and Monitoring Tool

- Can an InfiniBand network monitoring tool be designed such that:
  - Shows the various performance counters and attributes
  - Is Agentless
  - Has low overhead
  - Depicts the communication matrix of target applications
  - Shows the link usage statistics

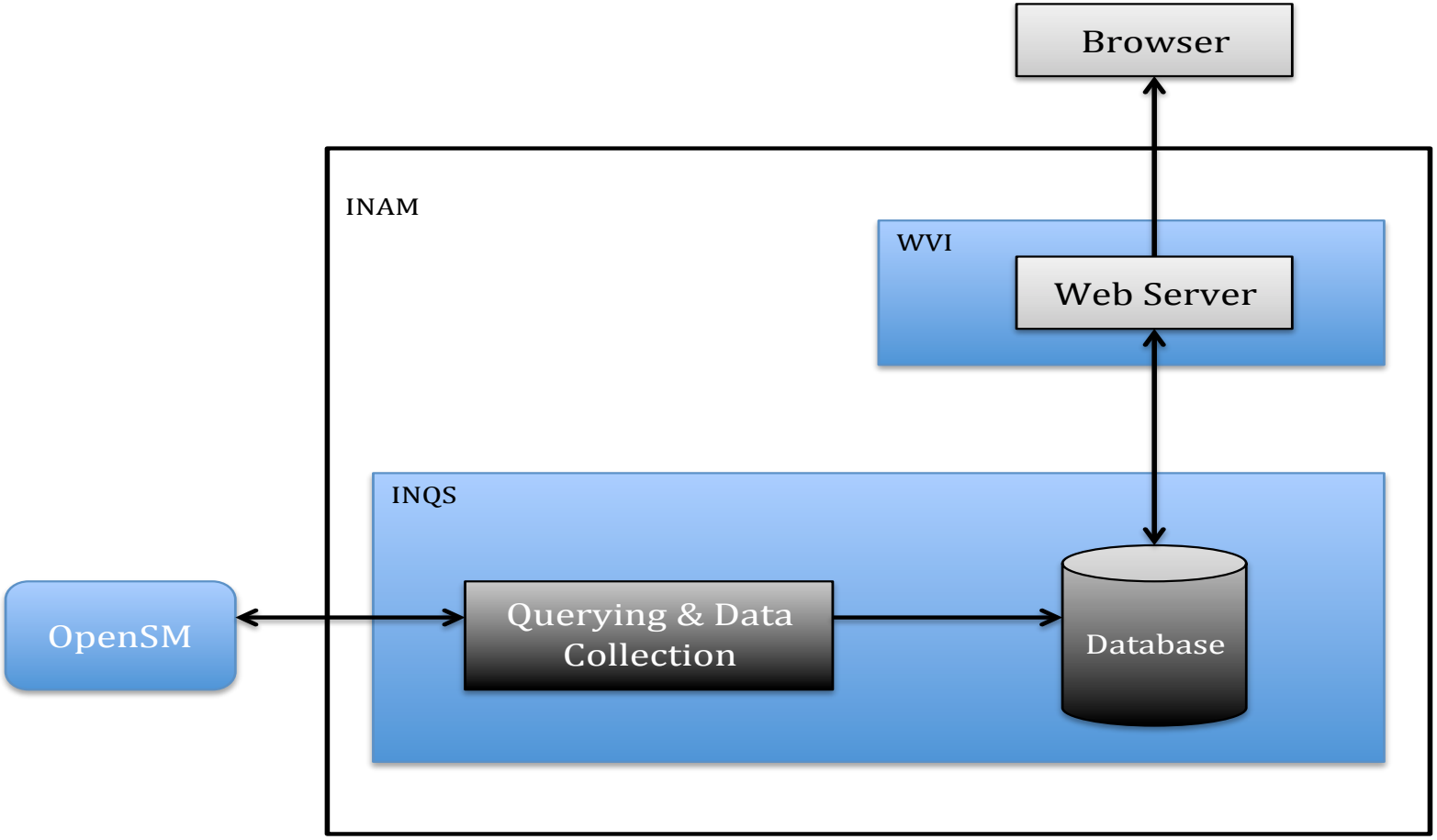
Outline

- Introduction
- INAM – Scalable InfiniBand Network Analysis & monitoring tool
  - Framework & Design
  - Network Monitoring
  - Link Utilization & Communication Pattern
- Experimental Analysis
- Conclusions and Future work

INAM-Framework



INAM-Framework (Contd)

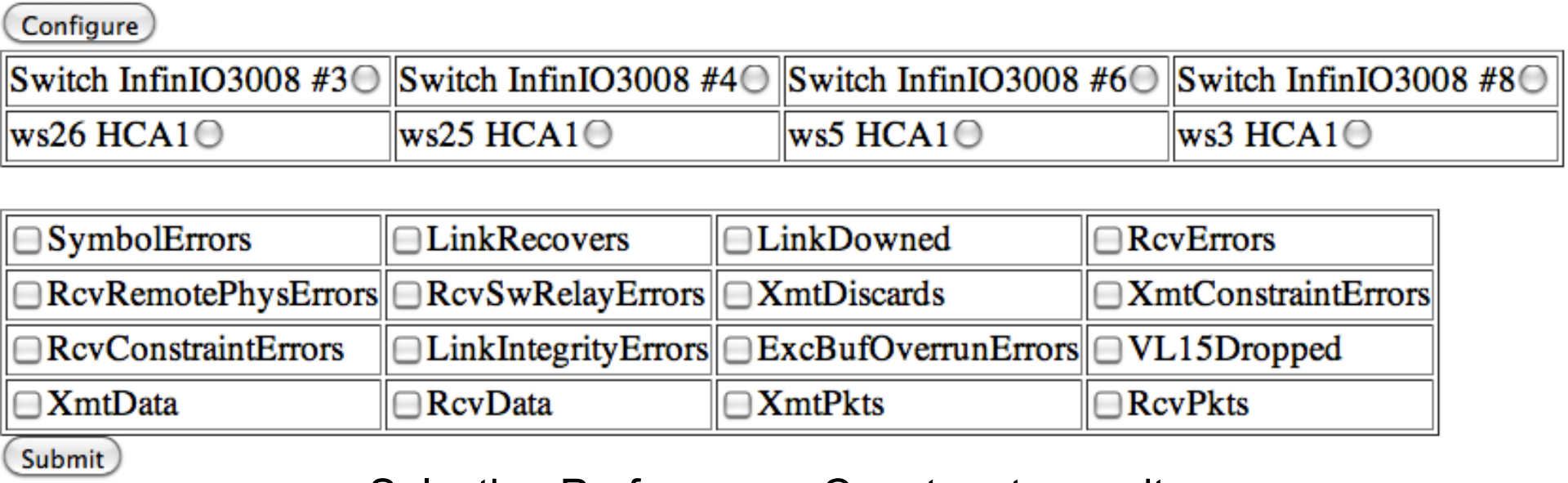


INAM – Network Monitoring

- Network Monitoring
  - Query the SMAs on the host nodes to obtain the performance counters and Subnet Management attributes and SM info
  - Temporary Database.
    - Real time monitoring with visualization.
  - Permanent Database
    - Keeps track of events in the subnet
    - Stores them for the time period mentioned by the user
    - Query this database to obtain the behavior of network traffic over a period of time
  - Modify rate of data collection (Sampling rate) as per user input
  - Modify rate of display as per user input

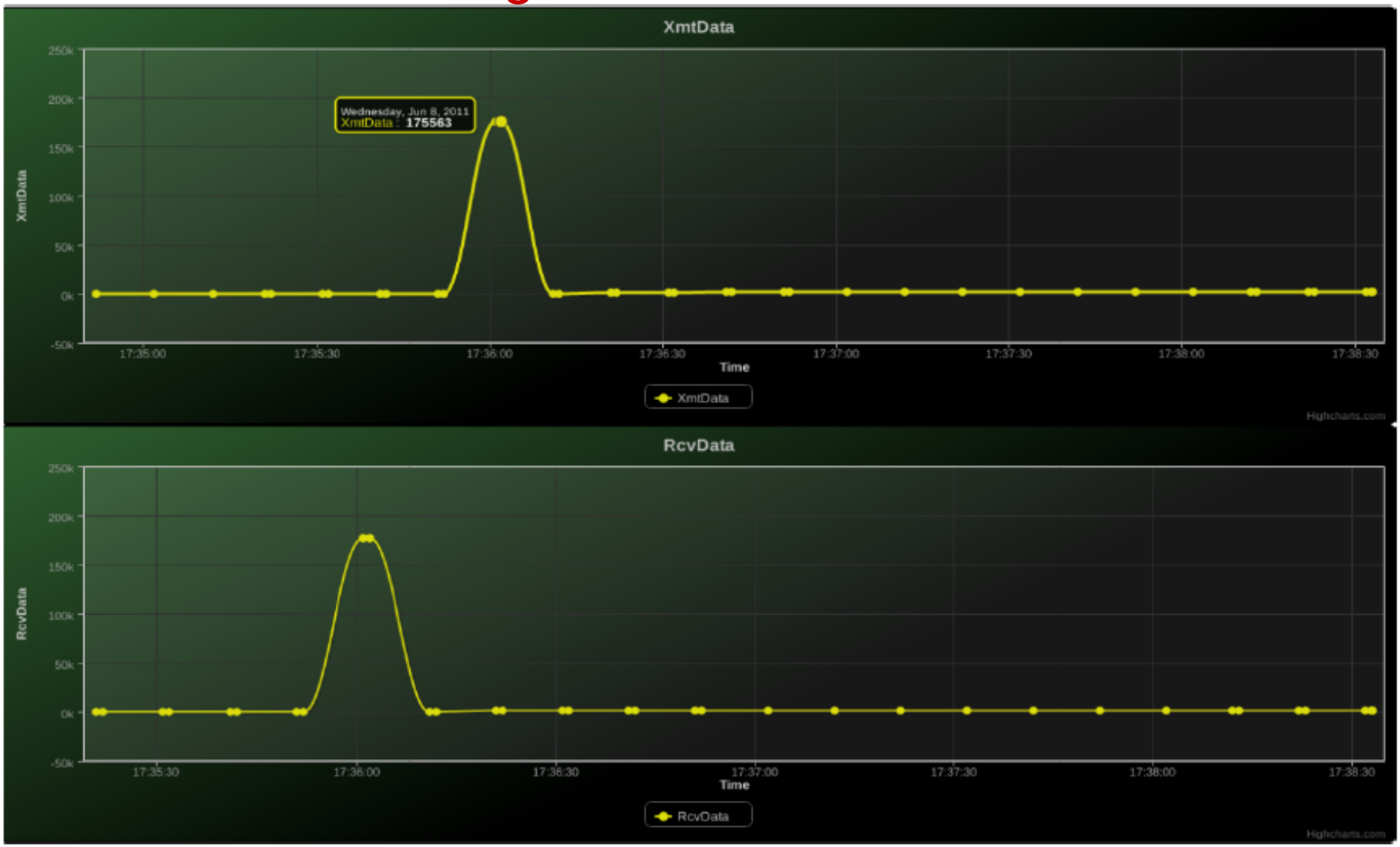
INAM – Network Monitoring

- Monitors the following in real time
  - Performance Counters
  - Subnet Management Attributes
  - Subnet Manager information in real time.



Selecting Performance Counters to monitor

Monitoring Performance Counters



Comparing Transmitted and Received Data on a Port

INAM – Network Monitoring

Configure

Switch InfinIO3008 #3 ◉	Switch InfinIO3008 #4 ○	Switch InfinIO3008 #6 ○	Switch InfinIO3008 #8 ○
ws26 HCA1 ○	ws25 HCA1 ○	ws5 HCA1 ○	ws3 HCA1 ○

Submit

LinkWidthEnabled	1X or 4X
LinkWidthSupported	1X or 4X
LinkWidthActive	4X
LinkSpeedSupported	2.5 Gbps
LinkState	Active
PhysLinkState	LinkUp
LinkDownDefState	Polling
LinkSpeedActive	5 Gbps
LinkSpeedEnabled	2.5 Gbps

Link Attributes

LID	2
GUID	0x2c902002135dd
Activity Count	787060
Priority	0
Status	3

Subnet Manager Information

VLCap	VL0-7
VLHighLimit	0
VLArbHighCap	8
VLArbLowCap	8
VLStallCount	7
OperVLs	VL0-3

VL Attributes

Monitoring Subnet Management Attributes

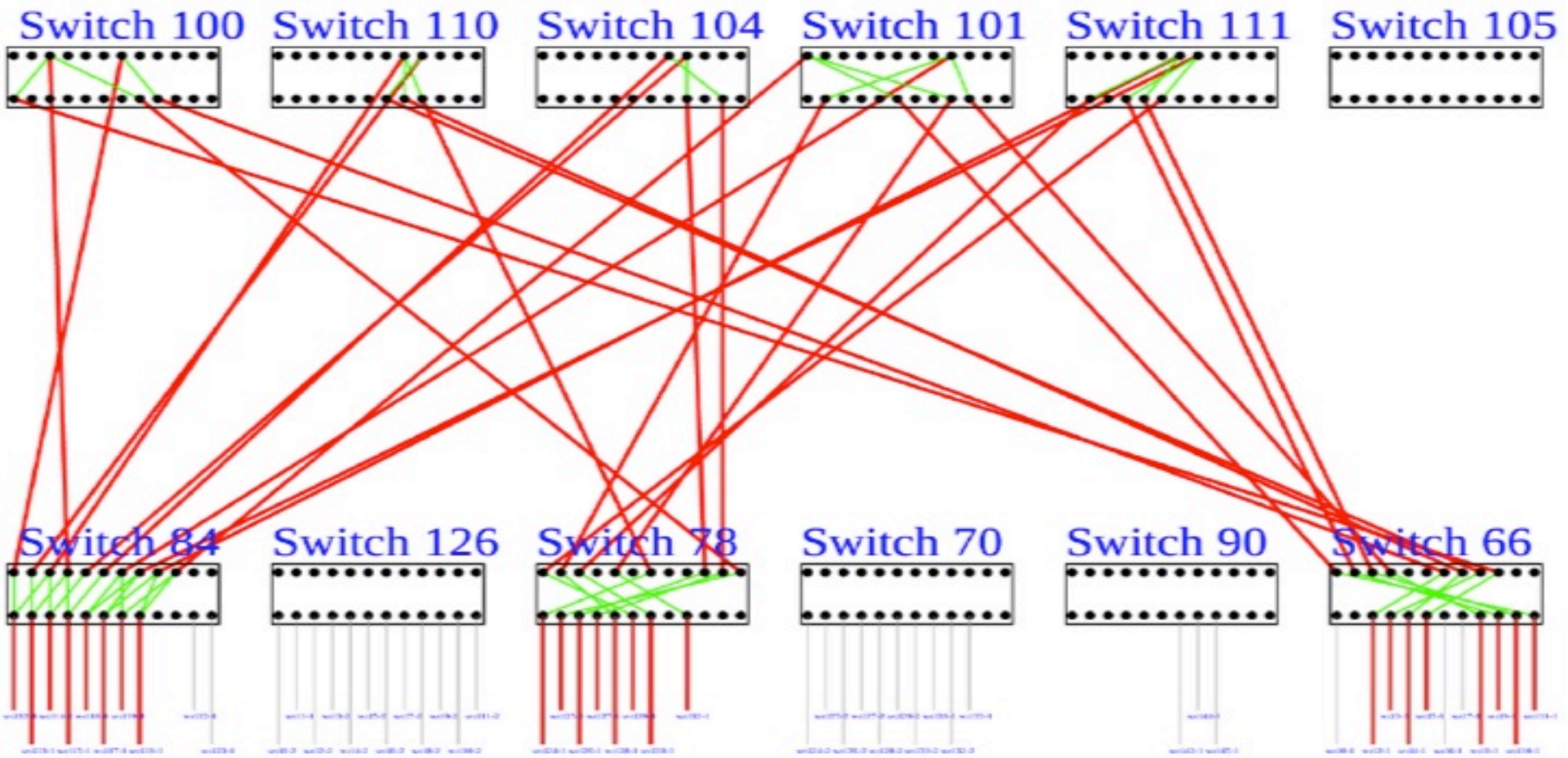
INAM – Link Utilization

- Link Utilization
  - Attributes Used

    XmtWait attribute
    - The number of units of time a packet waits to be transmitted from a port
    - Used for determining Link overutilization

    Received Packets, Sent Packets, Link Speed
    - Used for determining data exchange
  - Based on the host file provided by the user, obtain all possible paths between every source & destination pairs
  - Color variation of the links dependent on the amount of data transferred
  - Keep track of how many times each link is traversed and the amount of data flowing through it

INAM – Link Utilization



Screenshot Showing Link Utilization

Outline

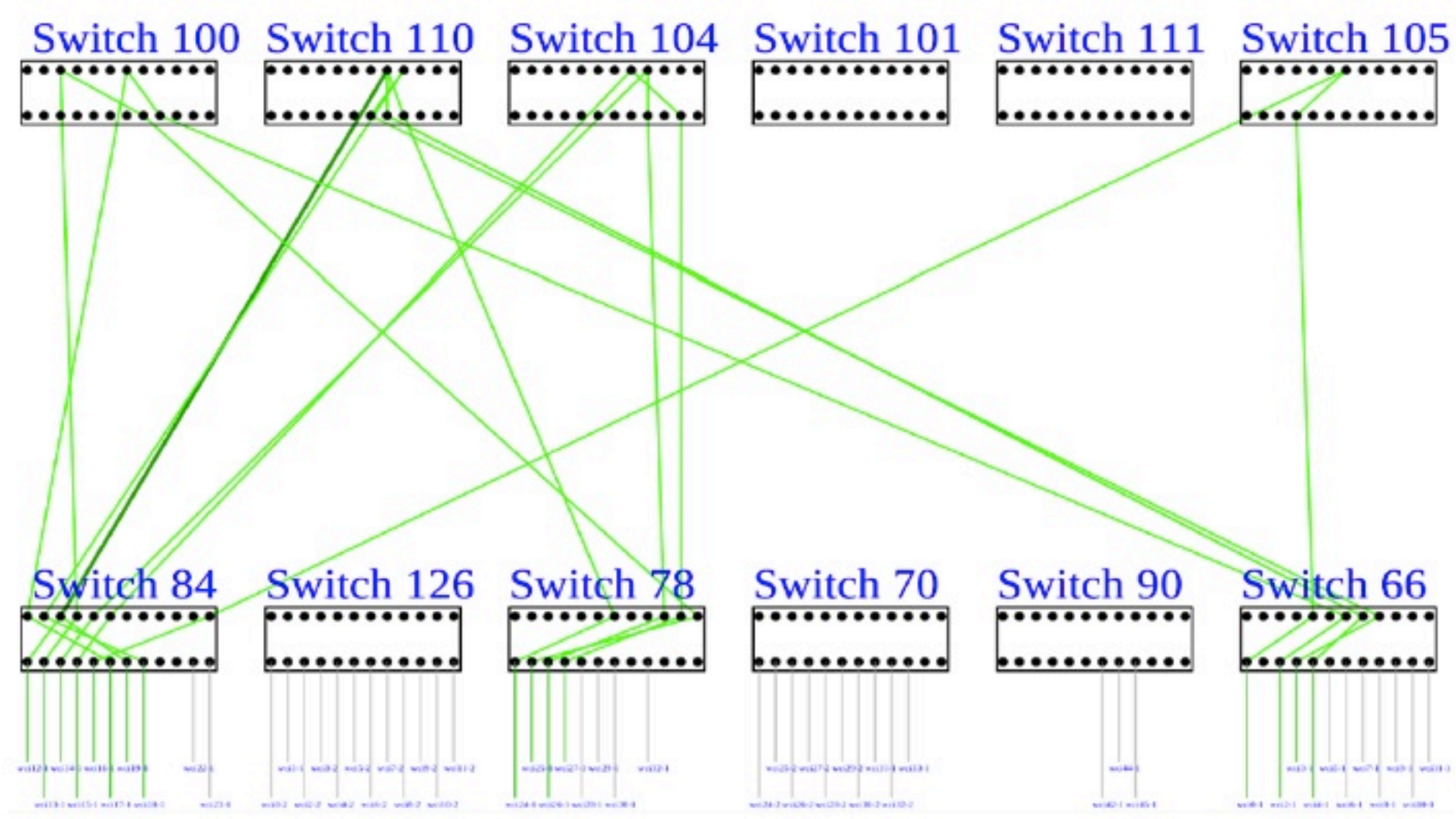
- Introduction
- INAM – Scalable InfiniBand Network Analysis & monitoring tool
- Experimental Analysis
- Conclusions and Future work

Experimental Analysis

- Experimental Setup
  - 6 Leaf Switches, 6 Spine Switches with 24 ports per switch
  - 35 nodes

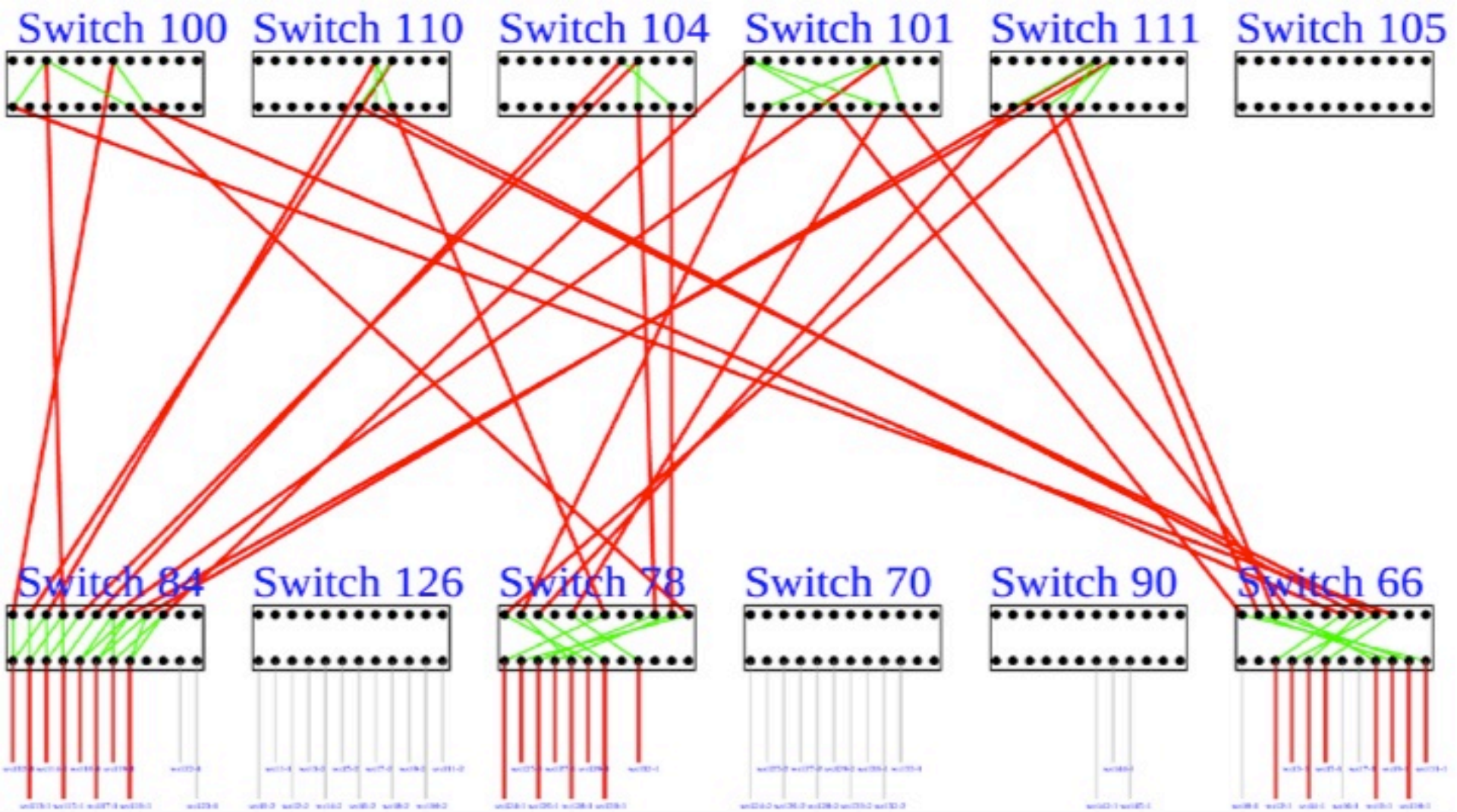
- Experiments
  - Communication pattern analysis for 16 processes and 64 processes
  - Communication pattern analysis for MPI_Bcast Operation with 16KB and 1 MB and with 6 processes. One process on each leaf switch
  - Communication Pattern for LU benchmark from Spec MPI Suite

Network Traffic Pattern for 16 processes



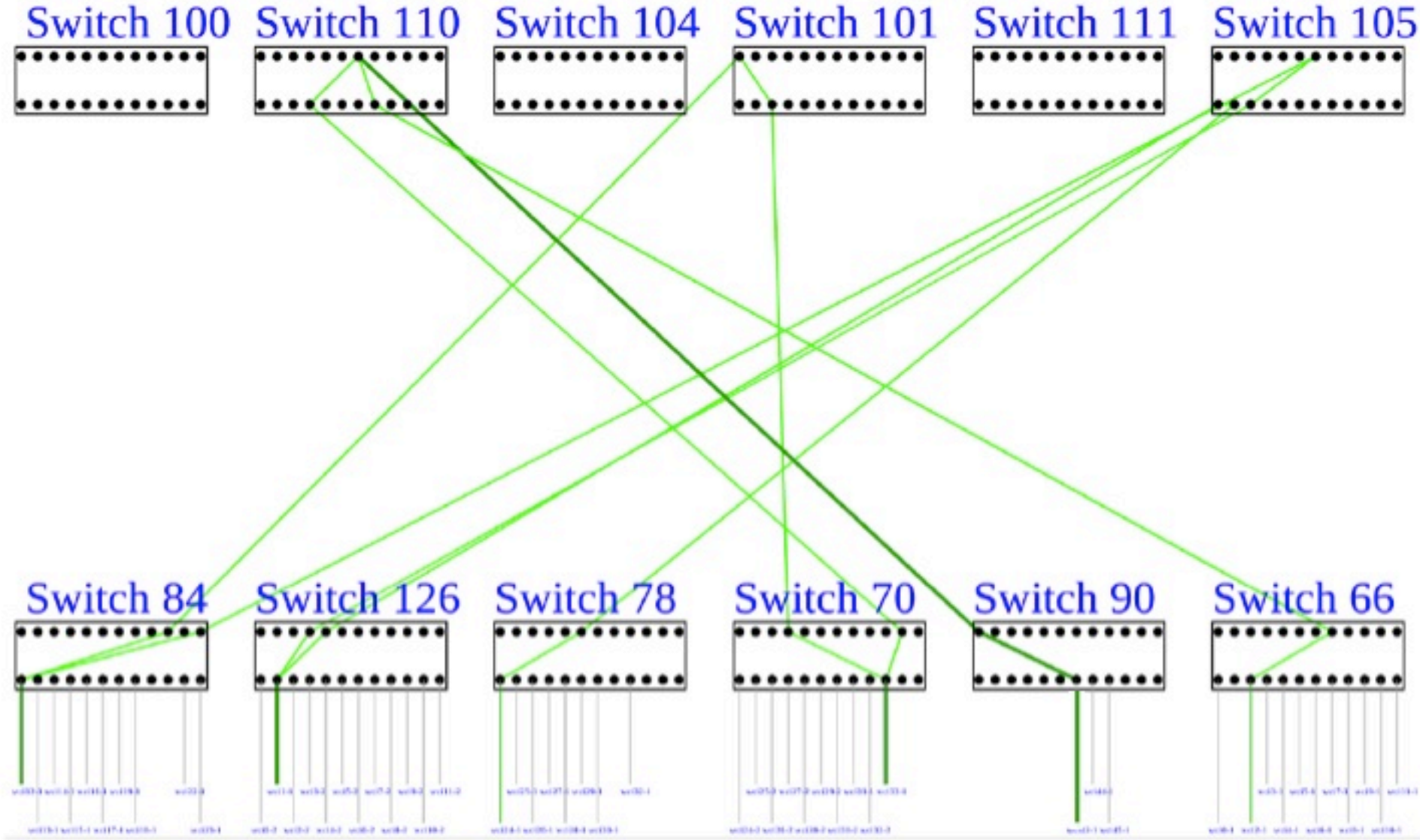
P2P communication with 8 processes on switch 84 and 4 processes each on switch 78 and 66

Network Traffic Pattern for 64 processes



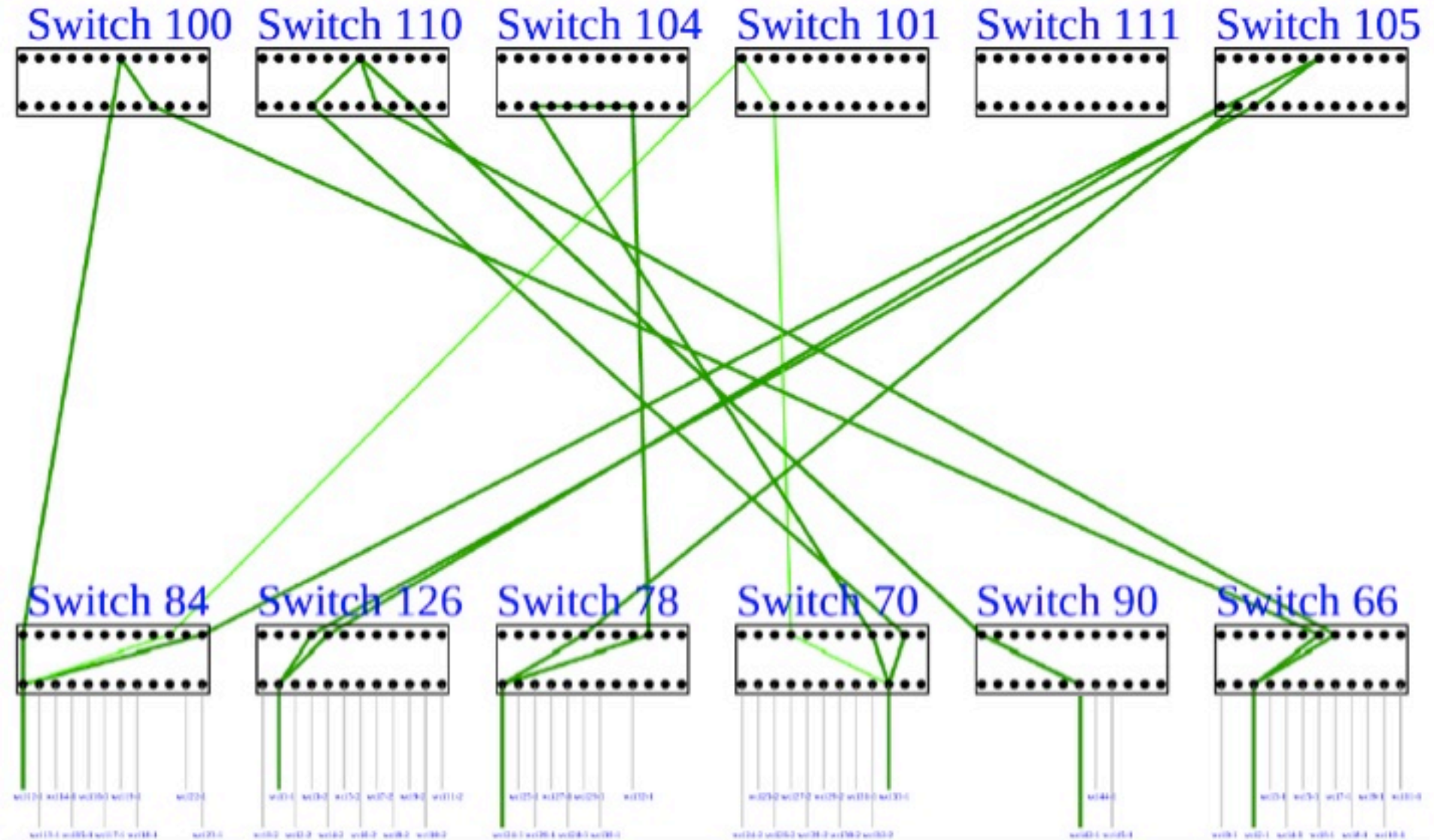
P2P communication with 32 processes on switch 84 and 16 processes each on switch 78 and 66

Link Utilization of Binomial Algorithm



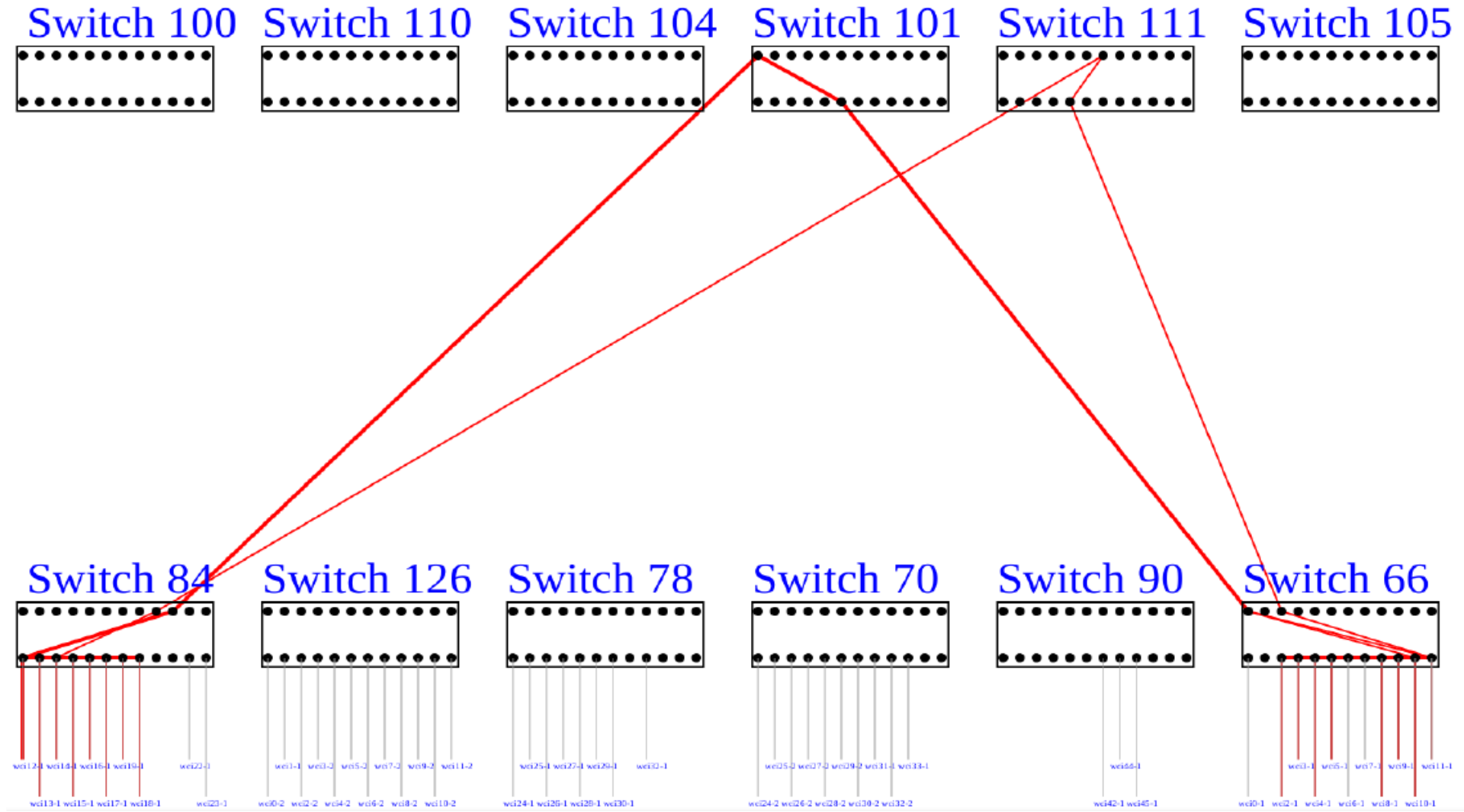
MPI_BCAST – 16 KB – 6 nodes – 1 node / switch

Link Utilization of Scatter Allgather Algorithm



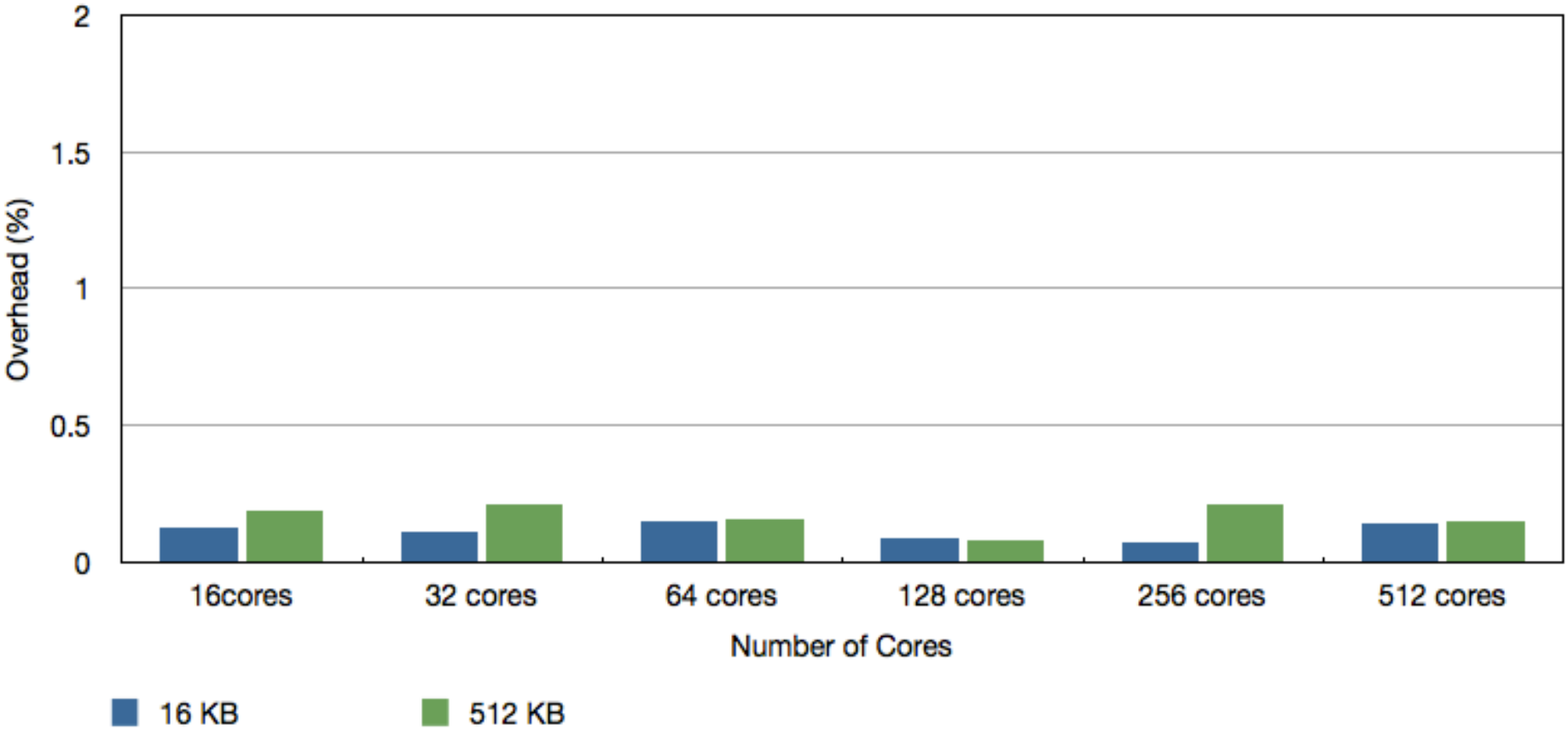
MPI_BCAST – 1 MB – 6 nodes – 1 node / switch

Communication Pattern of LU Benchmark



128 processes – 8 nodes / switch – 8 processes / node

INAM - Overhead



IMB alltoall – 8 cores / node
Overhead less then 0.5 % as we increase the system size

Outline

- Introduction
- INAM – Scalable InfiniBand Network Analysis & monitoring tool
- Experimental Analysis
- Conclusions and Future work

Conclusion & Future Work

- Conclusion
  - INAM - a scalable network monitoring and visualization tool for InfiniBand networks
  - Low Overhead
  - Agent less
  - Link Utilization
  - Communication Pattern

- Future Work
  - Time line graphical pattern display which shows the entire cluster's traffic at every instant.
  - Scalability Studies
  - On line analysis of the Communication patterns on a cluster
  - Incorporate support for counters per virtual lane