Overview of the BlueGene/L Torus Network

Philip Heidelberger

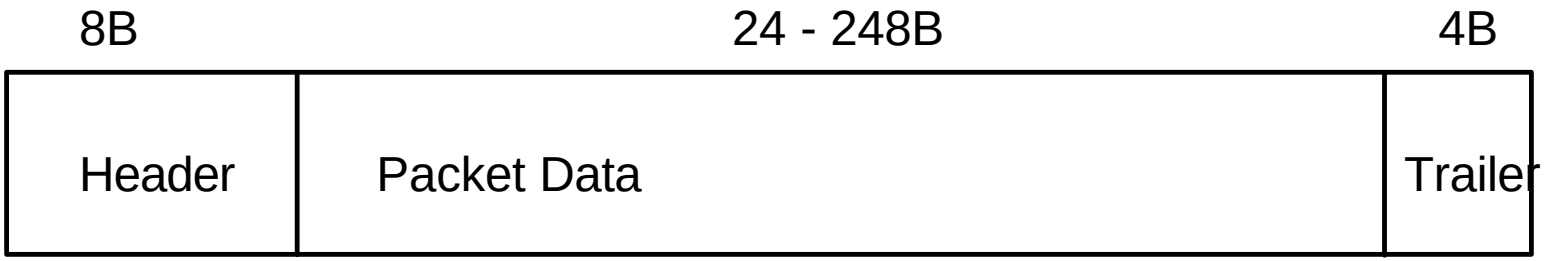
Outline

- Hardware overview
  - Basic structure and capabilities
  - Reliability features
  - Dynamic routing and arbitration details
- Simulators
  - Near cycle accurate simulator
    - Used during design phase to study performance tradeoffs
    - Sample performance studies
  - Multi-node VHDL testbench
    - Using to verify the actual hardware design

Torus Network Overview

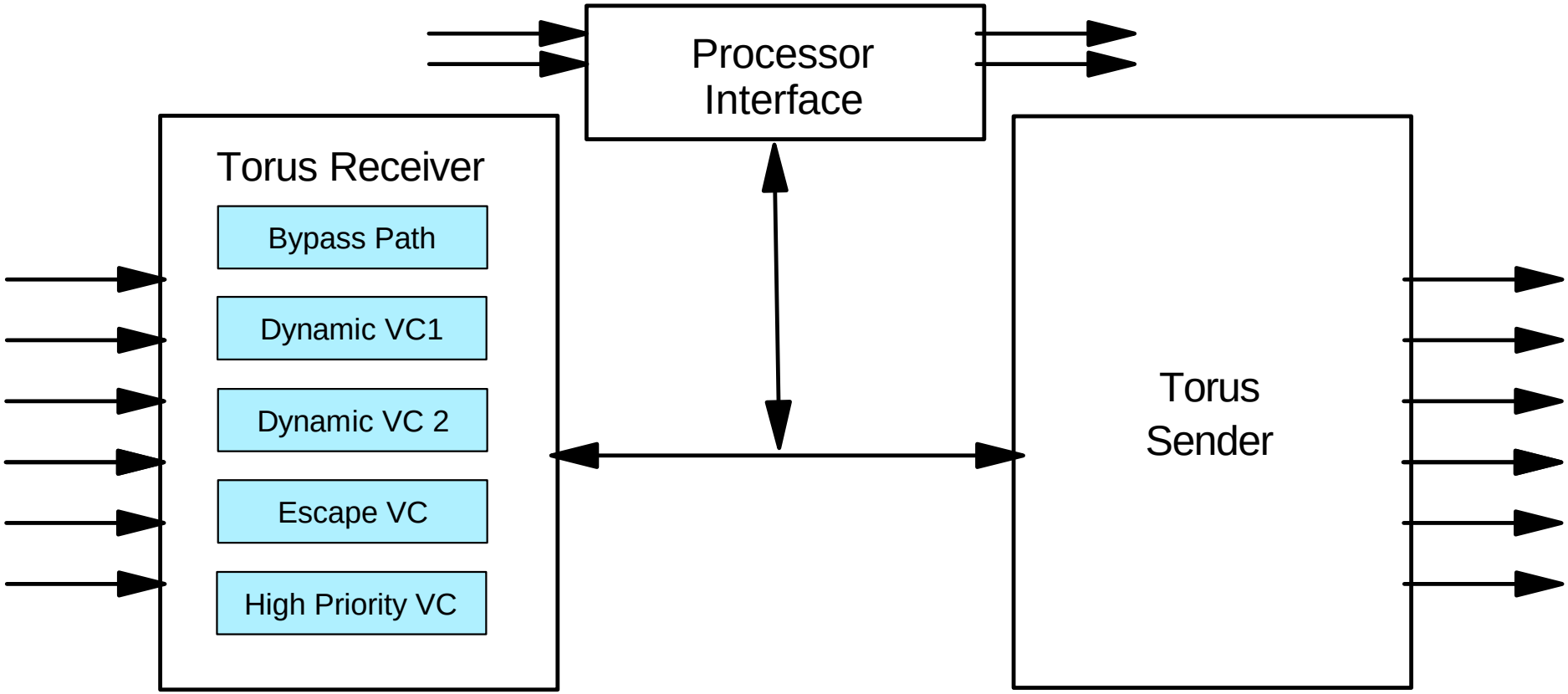
- 3-D torus with dynamic routing
- Point-to-point and broadcast down row messages
- 32 to 256 byte packets (multiples of 32) + 4 byte trailer
- Packet CRC + retransmission protocol for reliable delivery
  - 8 byte acks and token/acks
- Target bandwidth:
  - 1.4 Gbs/link (1 bit wide): 1 byte/4 processor cycles
  - 175 MBs/link
  - 1.05 GBs/node
  - Flexibility to clock 2x faster/slower, depending on link error rates
- Per-hop latency
  - = capture + torus logic + drive + wire/cable
  - torus logic: 12 cycles*5.7 ns/cycle = 69 ns
  - capture + drive = 12 ns
  - wire/cable: 1 to 135 ns 8 ns (average)
  - total (average) 89 ns/hop

Packet Format



- Header
  - Routing information (destination, vc, size, dynamic, broadcast)
  - Sequence number and CRC
- Packet Data
  - Expect 8B MPI software overhead/packet
  - MPI Payload: up to 240B/packet
- Trailer: CRC
- Payload Efficiency
  - = 240 / [256 (packet) + 4 (trailer) + 8 (token/ack) + 2-4 (idles)]
  - = 88%

Torus Router Structure



- Virtual Cut Through Architecture
  - Dynamic routing with 2 Virtual Channels (VCs) to improve throughput
  - Token flow control for VCs to prevent overflows
  - Escape VC used for deadlock prevention and deterministic routing
  - High priority VC for inter-node OS messaging
  - 1 KB/buffer
- Multiple simultaneous transfers from receiver to sender/processor
- Multiple injection and reception fifos
- Complex arbitration policies to avoid contention
- Error checks and retransmission of corrupted packet

Dynamic Routing

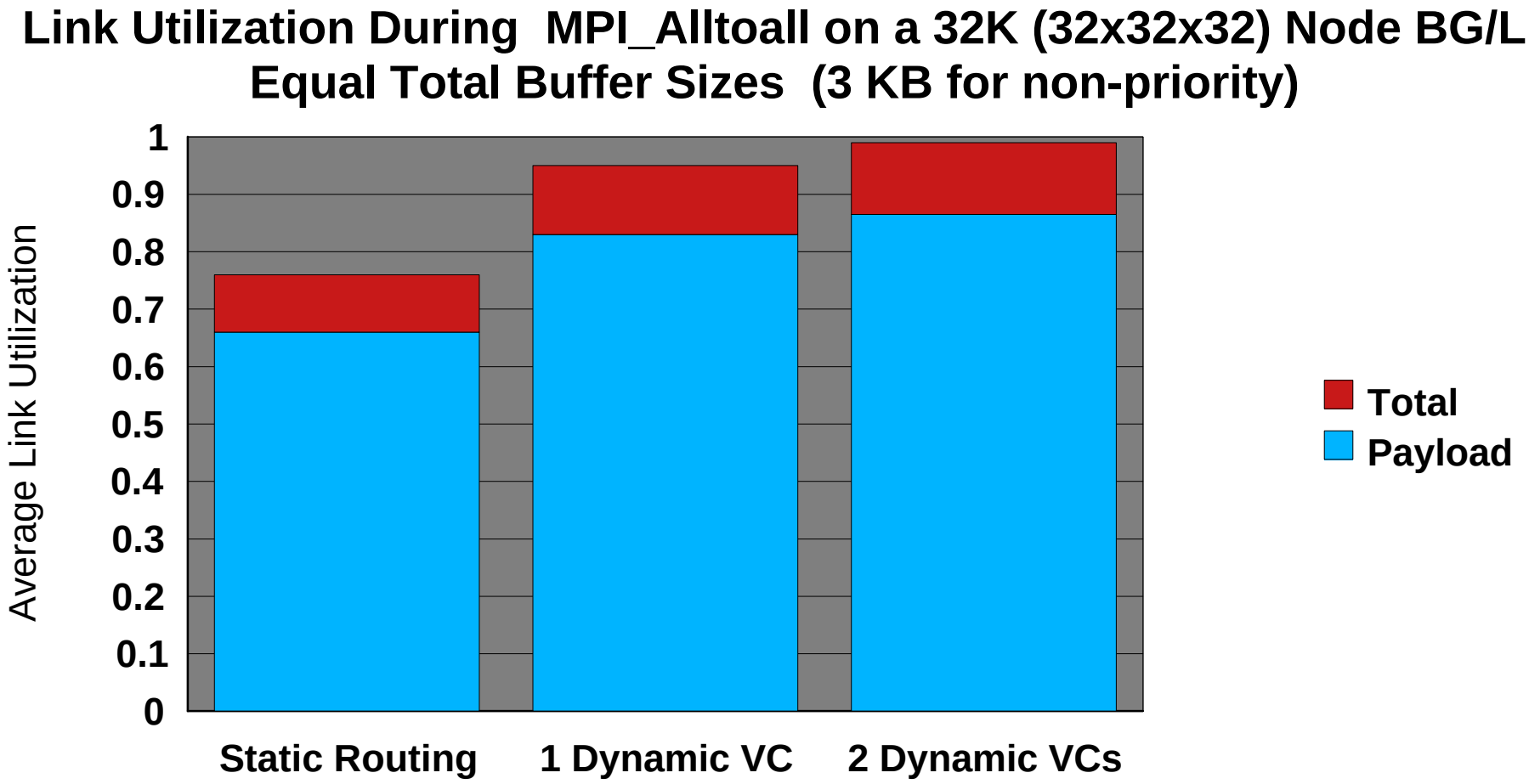
- No routing tables
- Packet may be dynamically or statically routed
- "hint bits" in header determine directions
  - eg, 011000 x- and y+ moves are allowed
  - set by hardware upon injection, or software
  - modified as packet flows through network
    - no direction reversal
    - set to 0 when packet reaches destination
- Arbitration
  - Join the shortest queue approximation to select direction & vc
  - Serve the longest queue approximation to select winner
- Fault Tolerance
  - Torus node can still operate even if other parts of node cannot
  - Dead node or link can be avoided by having sw set hint bits
  - Full connectivity with up to 3, non-colinear, faulty nodes in partition
  - Link CRC and (different) packet CRC

Near Cycle Accurate Simulator

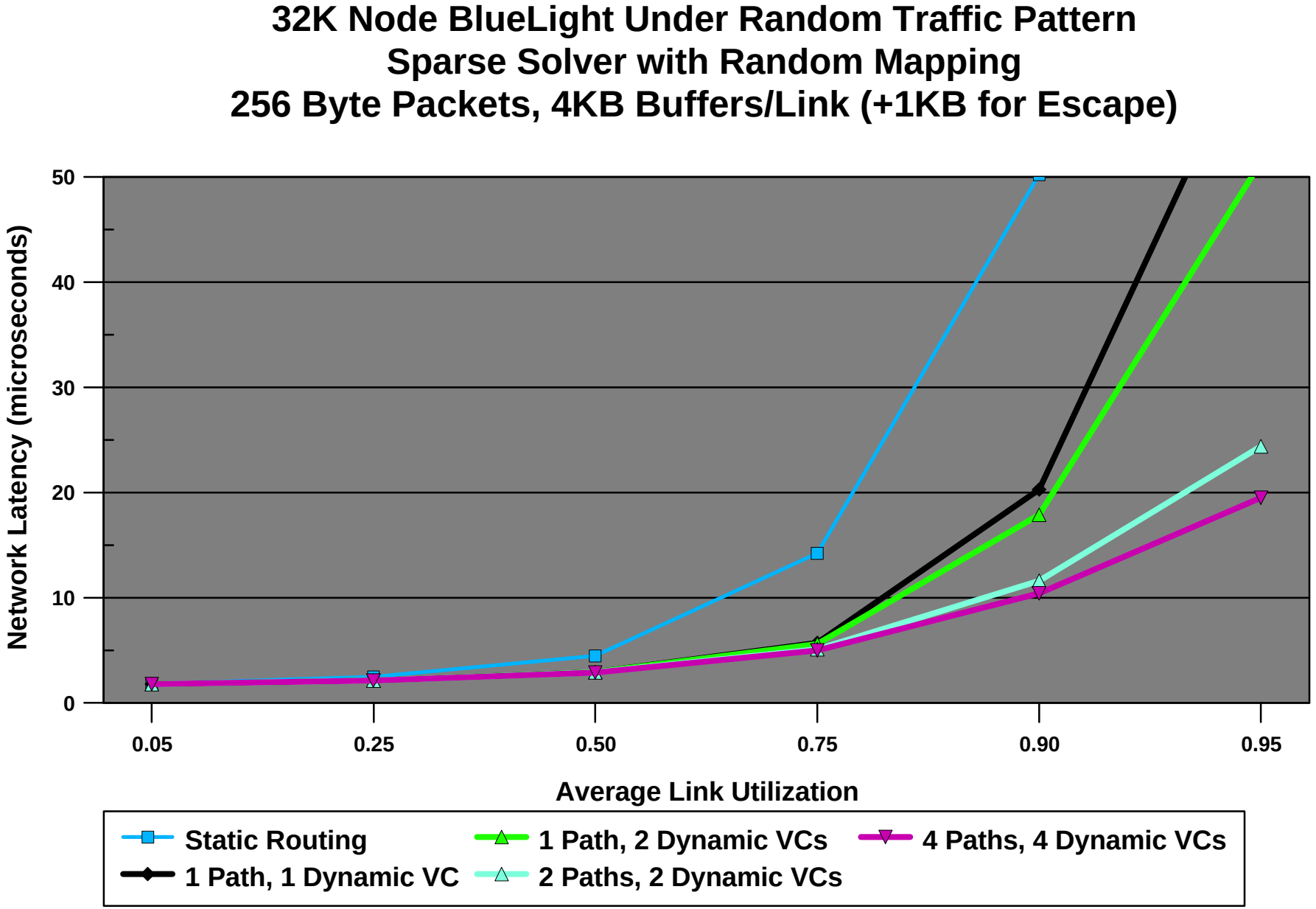
- Extensively used in design of BG/L
  - number and size of virtual channels
  - arbitration policies
- Runs in parallel on an SMP
  - Excellent speedup
  - ~ 0.5 BG/L microseconds / second (large torus, heavy traffic)
- Driven by
  - pseudo-codes
  - UTE traces
- MPI messaging protocol model
- Delivered to LLNL, Cal Tech., SDSC

Sample Performance Study: MPI_Alltoall

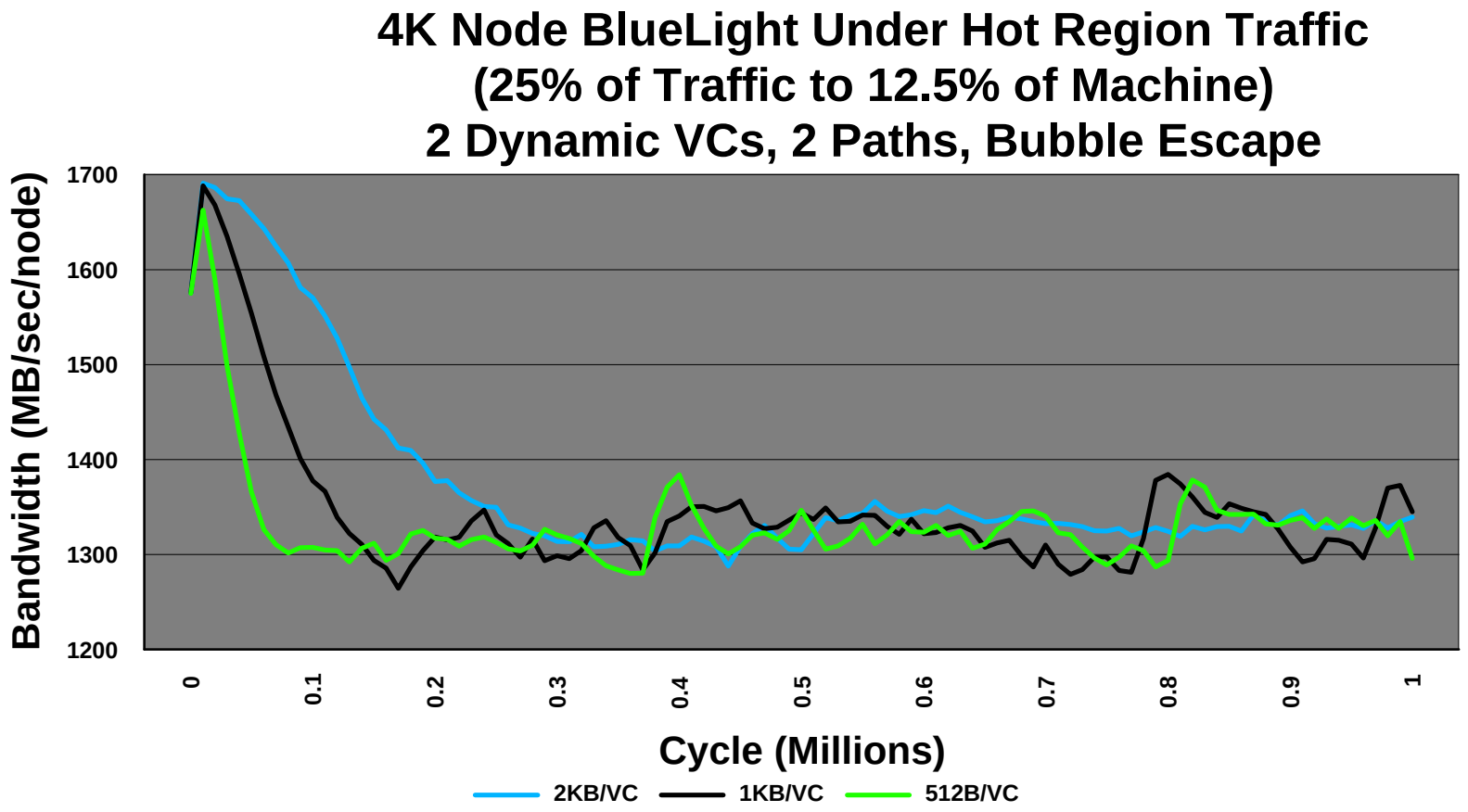
- Each node sends a different message to every other node
- Important MPI call used extensively in applications
- Stresses the network



Average Network Latency

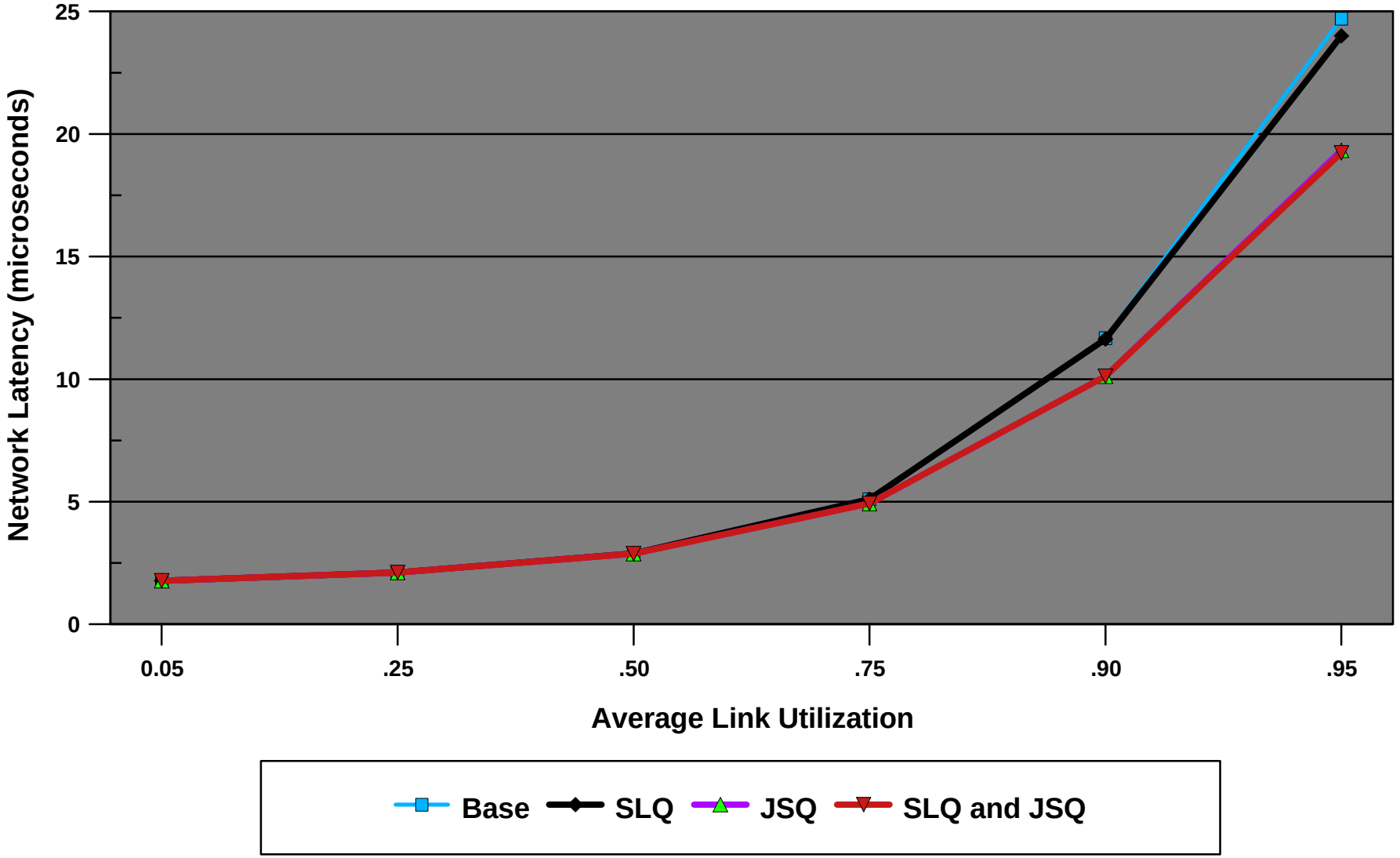


Network Performance Under Extreme Non-Uniform Load

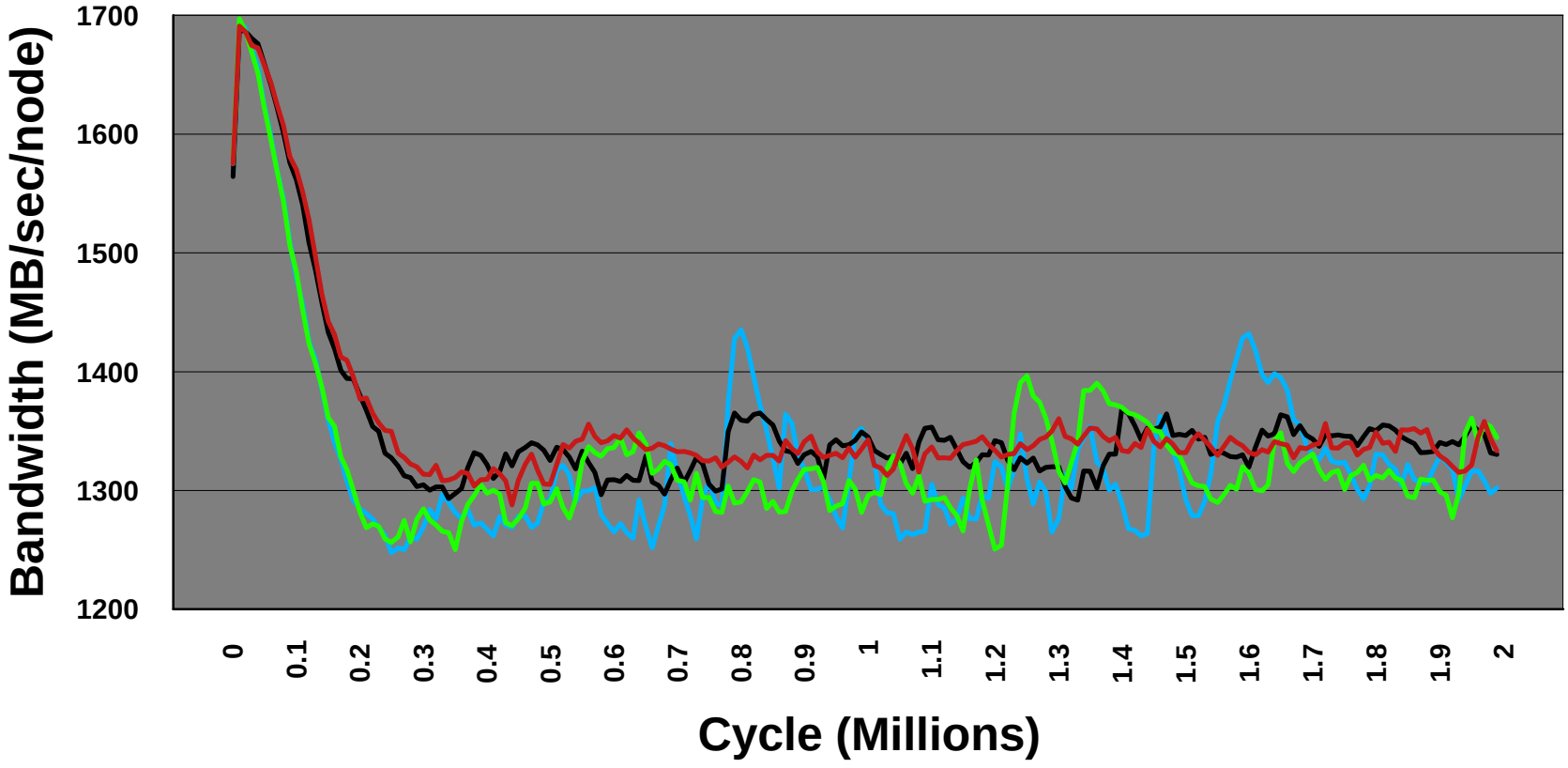


- Sends only
- Throughput into hot region stabilizes near peak bandwidth
- Larger buffers slow decline in throughput, but stabilize near same level

32K Node BlueLight Under Random Traffic Pattern
256 Byte Packets, 4KB Buffers/Link (+1KB for Escape)



4K Node BlueLight Under Hot Region Traffic
(25% of Traffic to 12.5% of Machine)
2 Dynamic 2KB VCs, 2 Paths, 2KB Bubble Escape



3D - FFT Communications Study

- Two communication phases
- Planar Phase:
  - send/receive a message to/from every node in the same x,y plane
- Row Phase:
  - send/receive a message to/from every node in the same z row
- Data set size: 1024^3 (doubles) on 16x16x16 BlueLight
- 8 KB messages for Planar phase
- 128 KB messages for Row phase
- Key performance metric is sustained link bandwidth
- Planar:
  - 86% payload out of max physical
- Row:
  - 88% payload out of max physical

- Can maintain close to full link utilization