Avoiding hot-spots on two-level direct networks

Abhinav Bhatele, Nikhil Jain, William D. Gropp, Laxmikant V. Kale

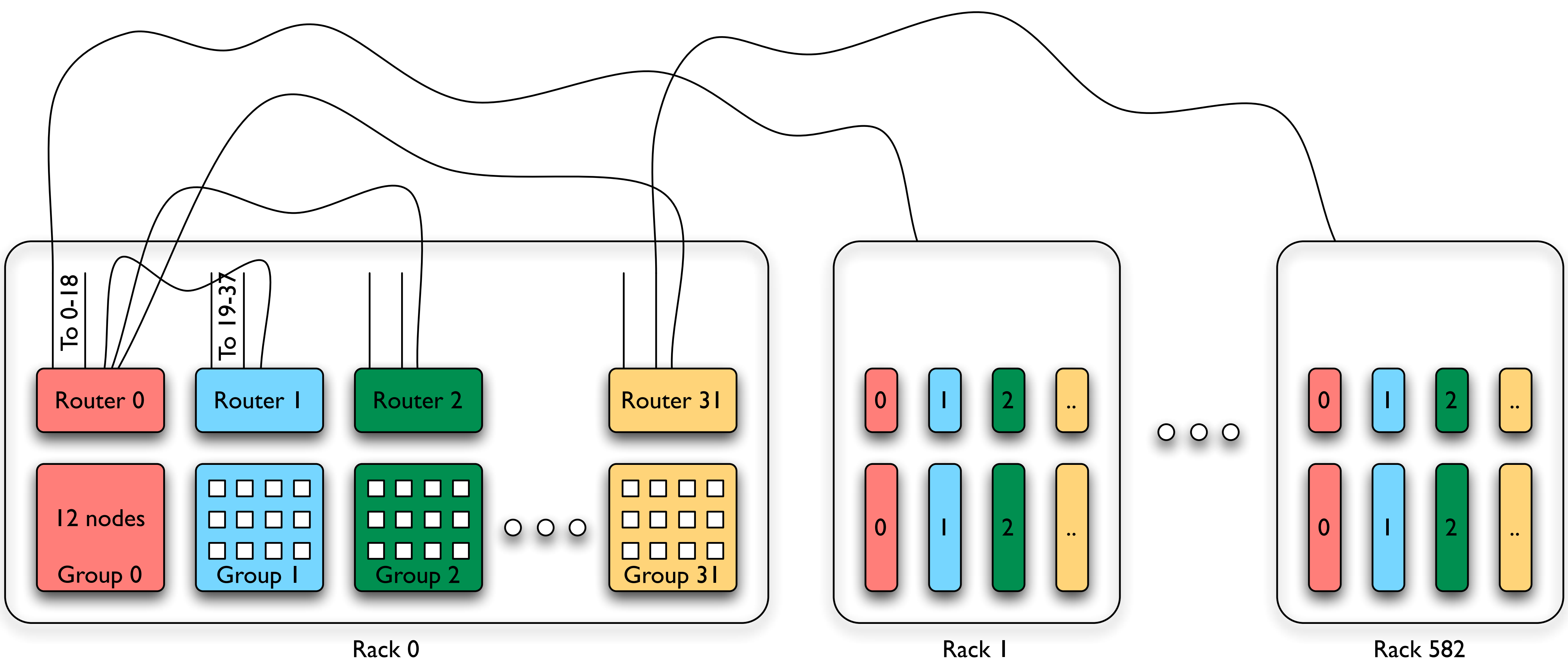
Supercomputing '11 ◆ November 17, 2011

LLNL-PRES-511461

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344

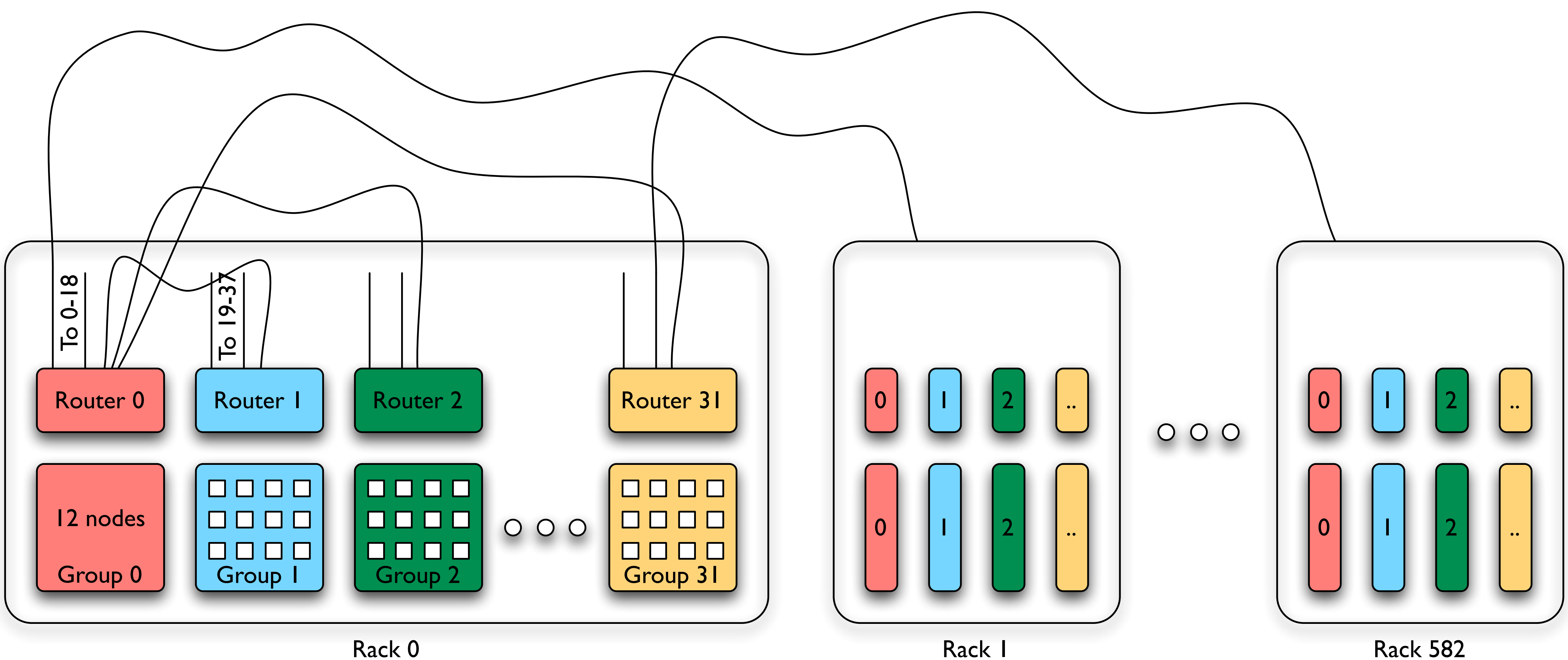
Lawrence Livermore National Laboratory, P. O. Box 808, Livermore, CA 94551

Interconnects for exascale



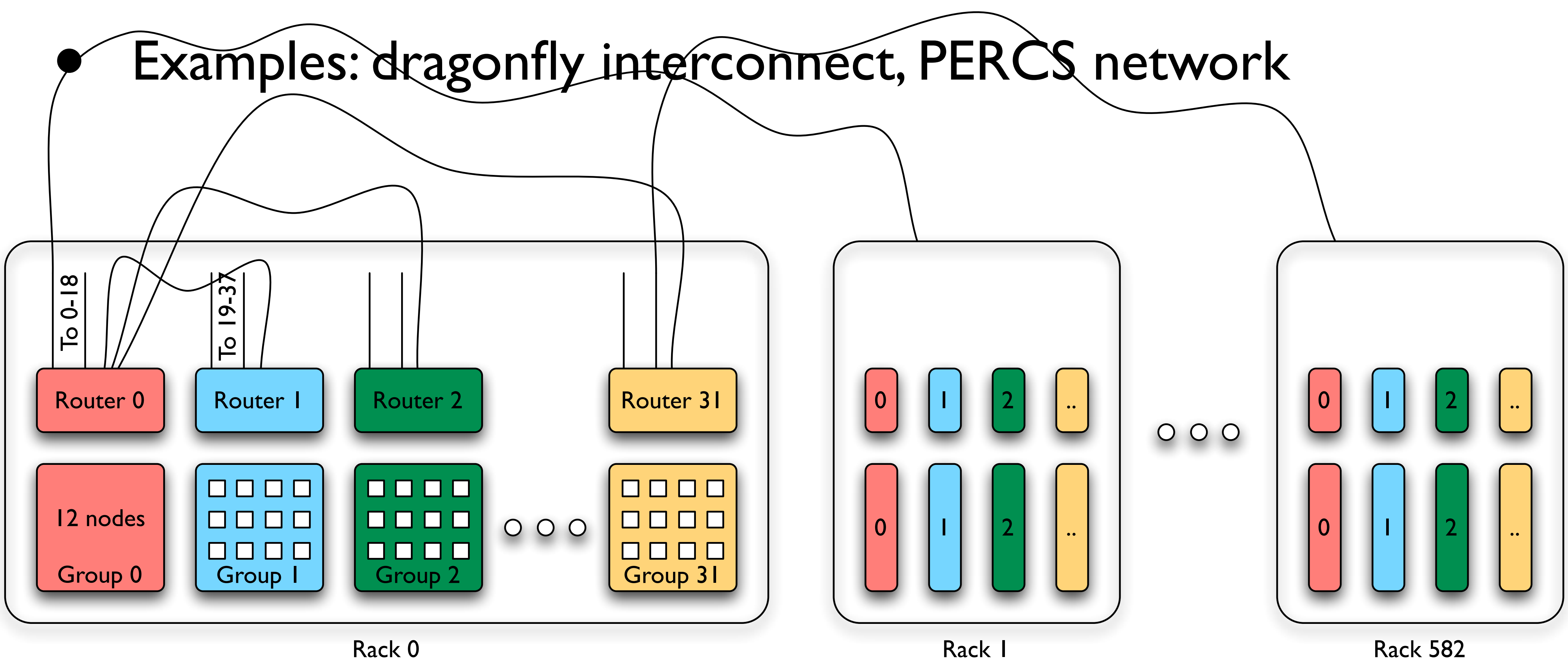
Interconnects for exascale

- Multi-level direct (all-to-all connection) networks
  - Higher bandwidth links at lower levels
  - Low diameter: few hops on the average



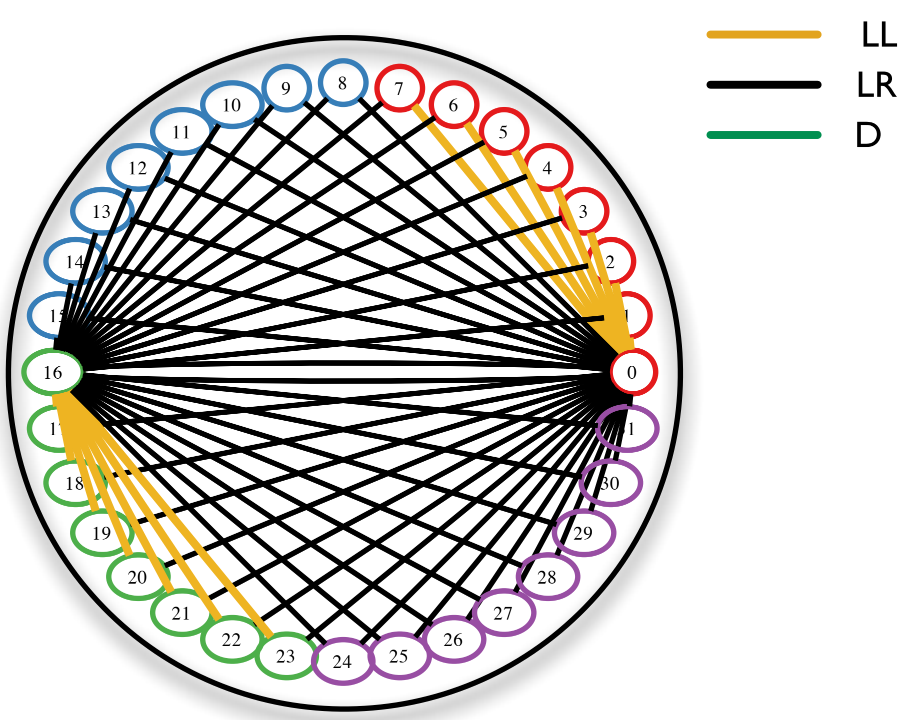
Interconnects for exascale

- Multi-level direct (all-to-all connection) networks
  - Higher bandwidth links at lower levels
  - Low diameter: few hops on the average
- Examples: dragonfly interconnect, PERCS network



IBM's PERCS network

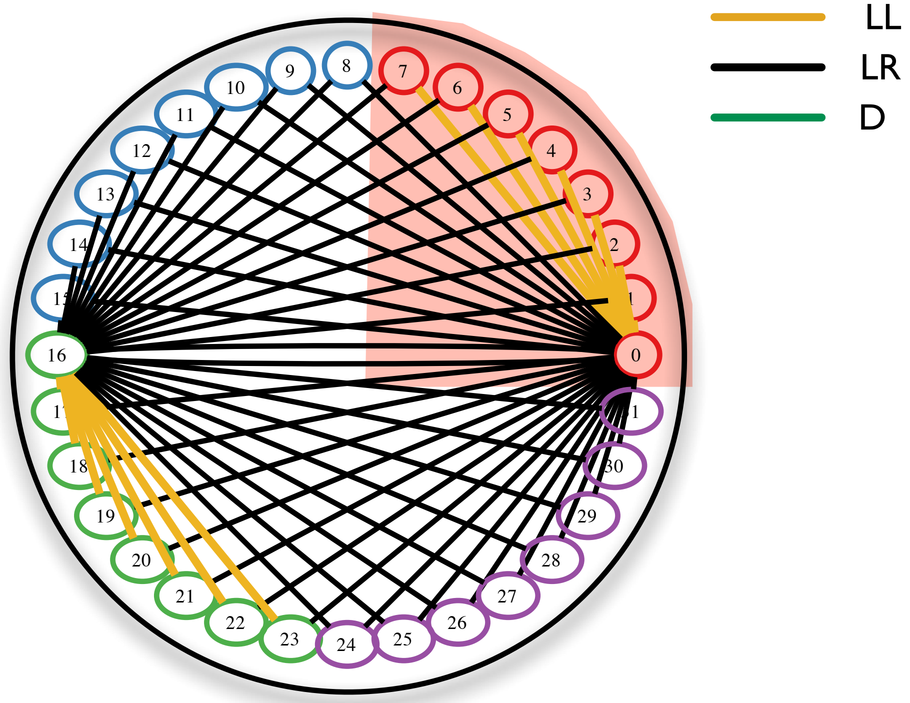
- Each QCM has four 8-core POWER7 chips, 8 such QCMs form a drawer, 4 drawers form a supernode
- Two-level network with 512 supernodes
- Three types of links: LL (24 GB/s), LR (5 GB/s), D (10 GB/s)



One supernode in the PERCS topology

IBM's PERCS network

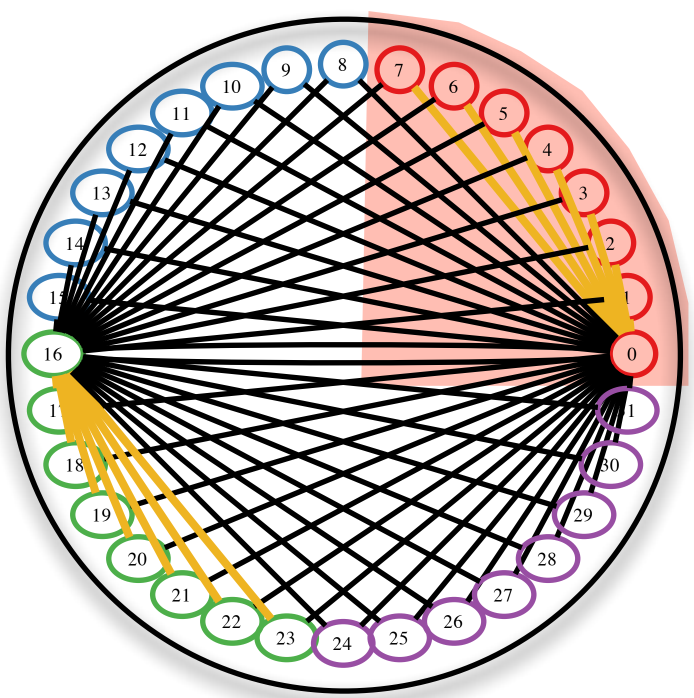
- Each QCM has four 8-core POWER7 chips, 8 such QCMs form a drawer, 4 drawers form a supernode
- Two-level network with 512 supernodes
- Three types of links: LL (24 GB/s), LR (5 GB/s), D (10 GB/s)



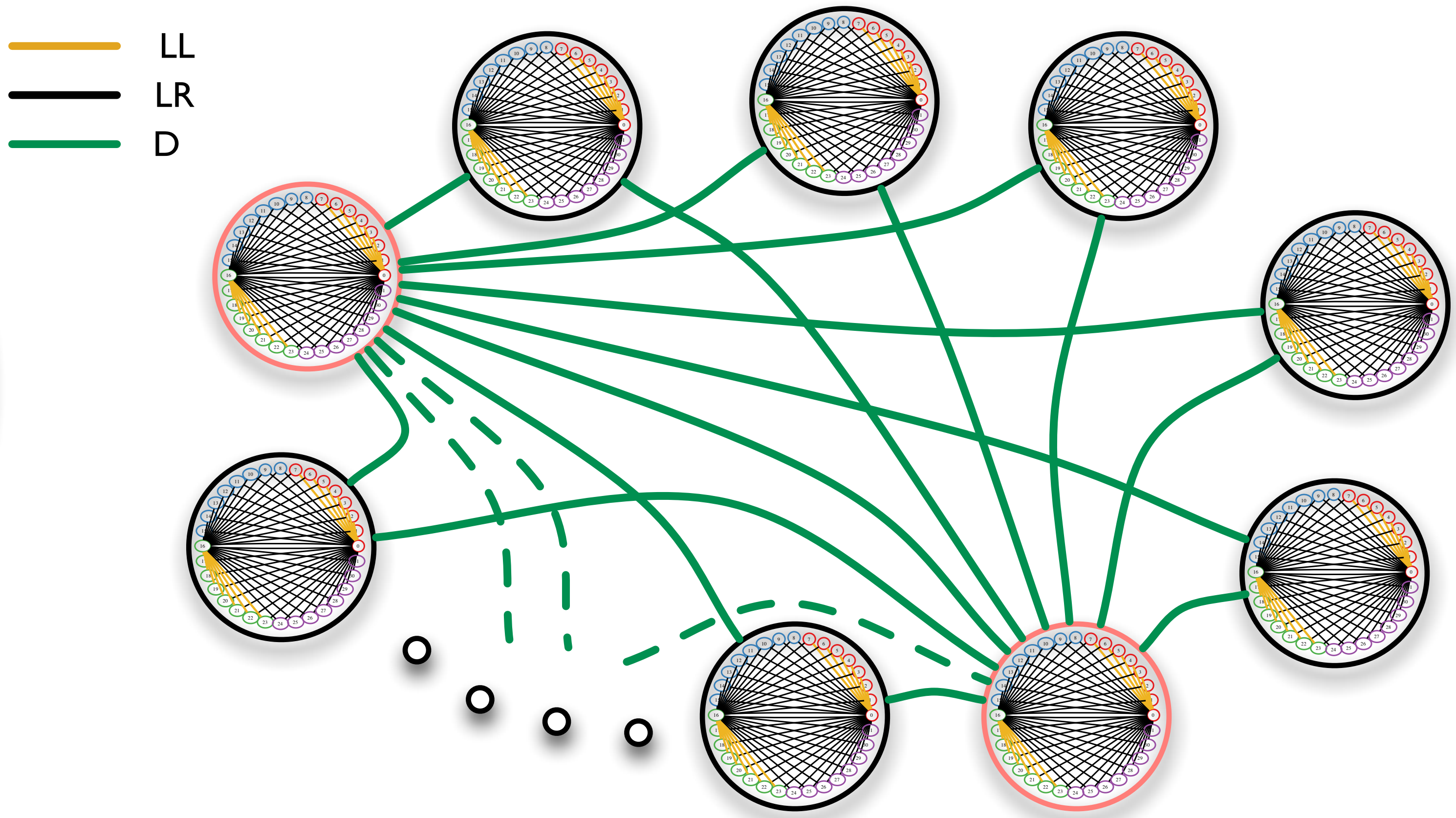
One supernode in the PERCS topology

IBM's PERCS network

- Each QCM has four 8-core POWER7 chips, 8 such QCMs form a drawer, 4 drawers form a supernode
- Two-level network with 512 supernodes
- Three types of links: LL (24 GB/s), LR (5 GB/s), D (10 GB/s)

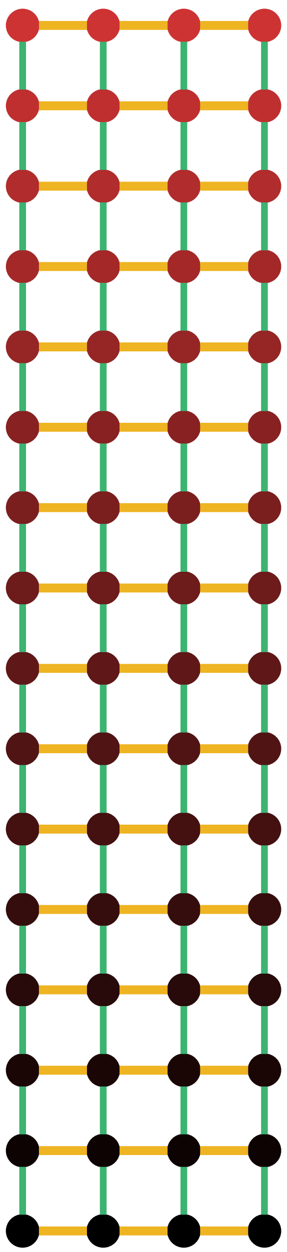


One supernode in the PERCS topology



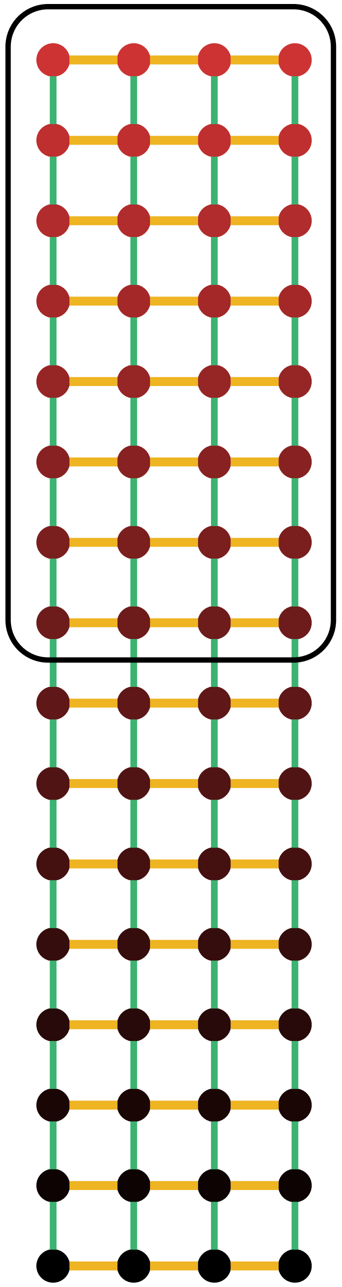
The 'D' link bottleneck

- Lets say we want to run a 2D Stencil on 2 supernodes (64 QCMs)
- Application communication graph: 16 x 4
- MPI-rank ordered mapping leads to significant contention on the single D link
- This is true for running any application with O(1) communicating partners per MPI process



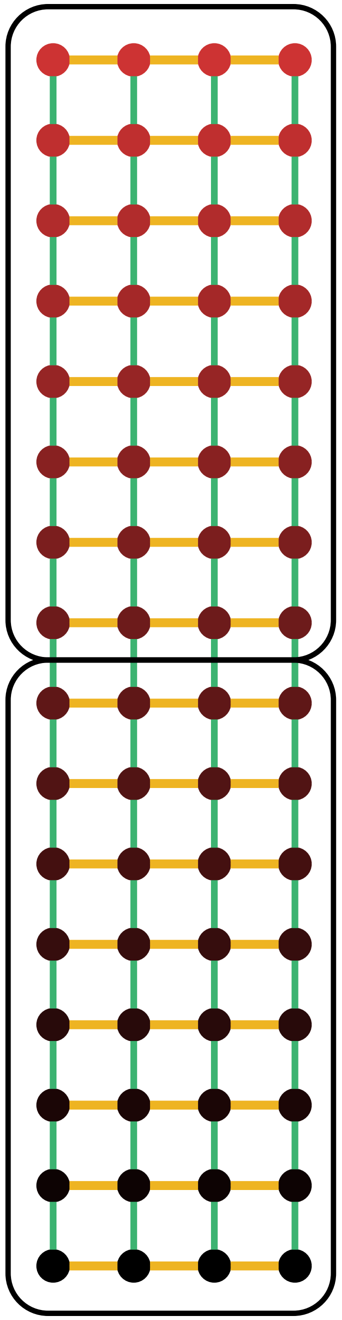
The ‘D’ link bottleneck

- Lets say we want to run a 2D Stencil on 2 supernodes (64 QCMs)
- Application communication graph: 16 x 4
- MPI-rank ordered mapping leads to significant contention on the single D link
- This is true for running any application with O(1) communicating partners per MPI process



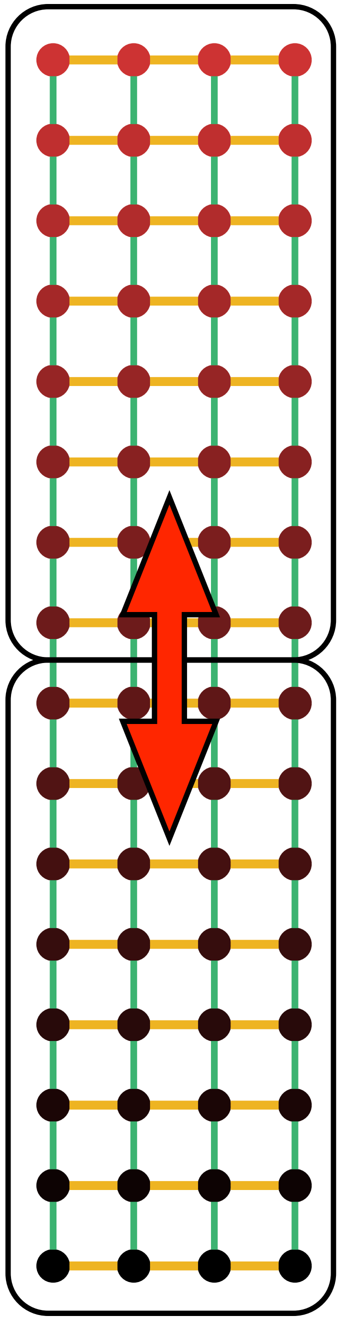
The ‘D’ link bottleneck

- Lets say we want to run a 2D Stencil on 2 supernodes (64 QCMs)
- Application communication graph: 16 x 4
- MPI-rank ordered mapping leads to significant contention on the single D link
- This is true for running any application with O(1) communicating partners per MPI process



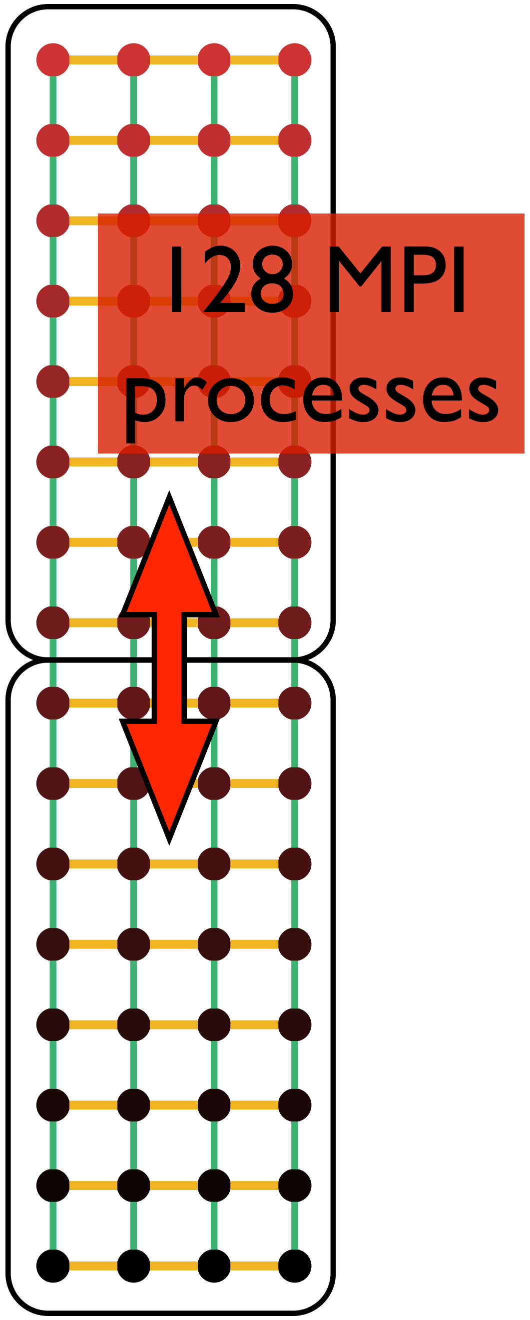
The ‘D’ link bottleneck

- Lets say we want to run a 2D Stencil on 2 supernodes (64 QCMs)
- Application communication graph: 16 x 4
- MPI-rank ordered mapping leads to significant contention on the single D link
- This is true for running any application with O(1) communicating partners per MPI process

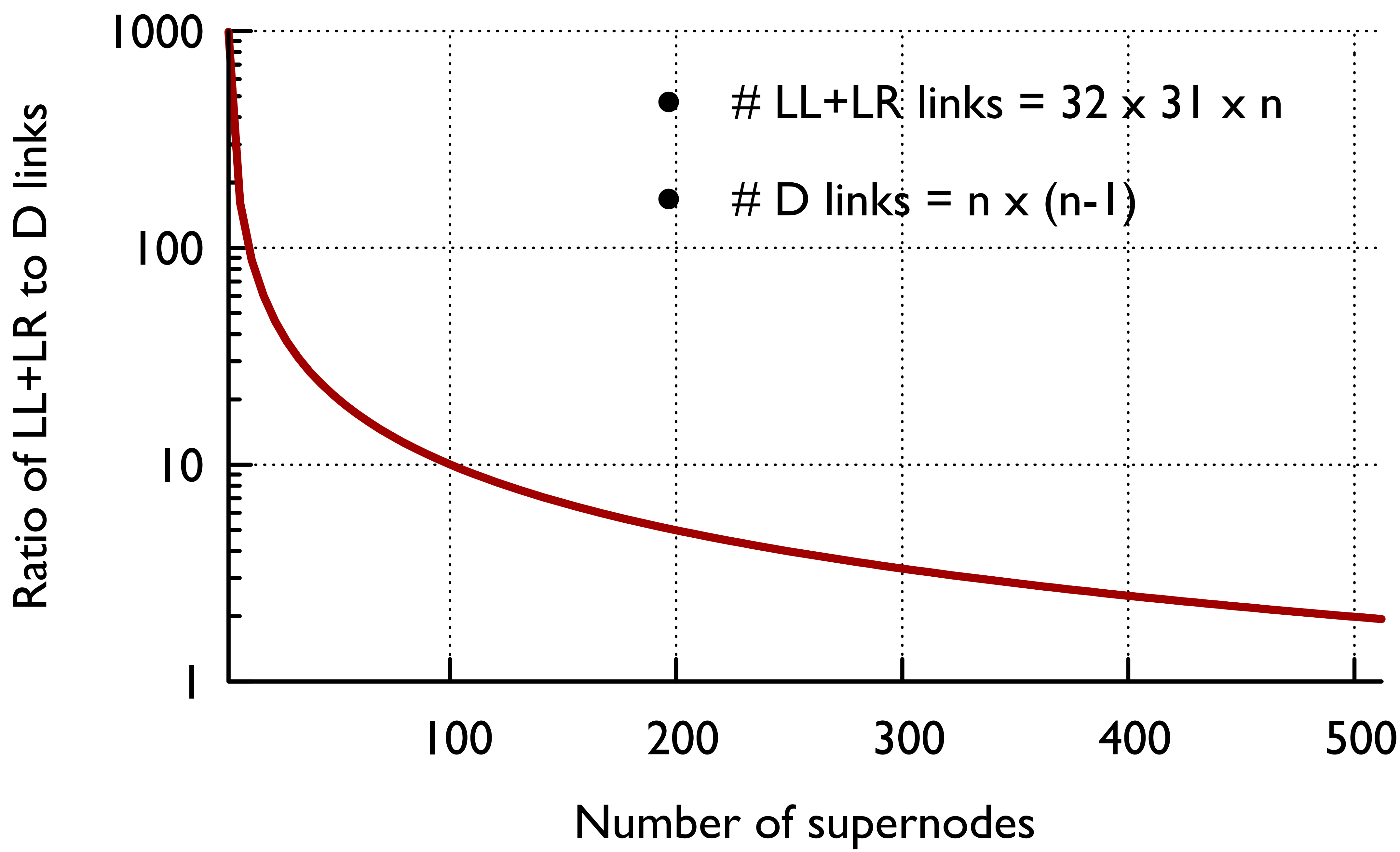


The ‘D’ link bottleneck

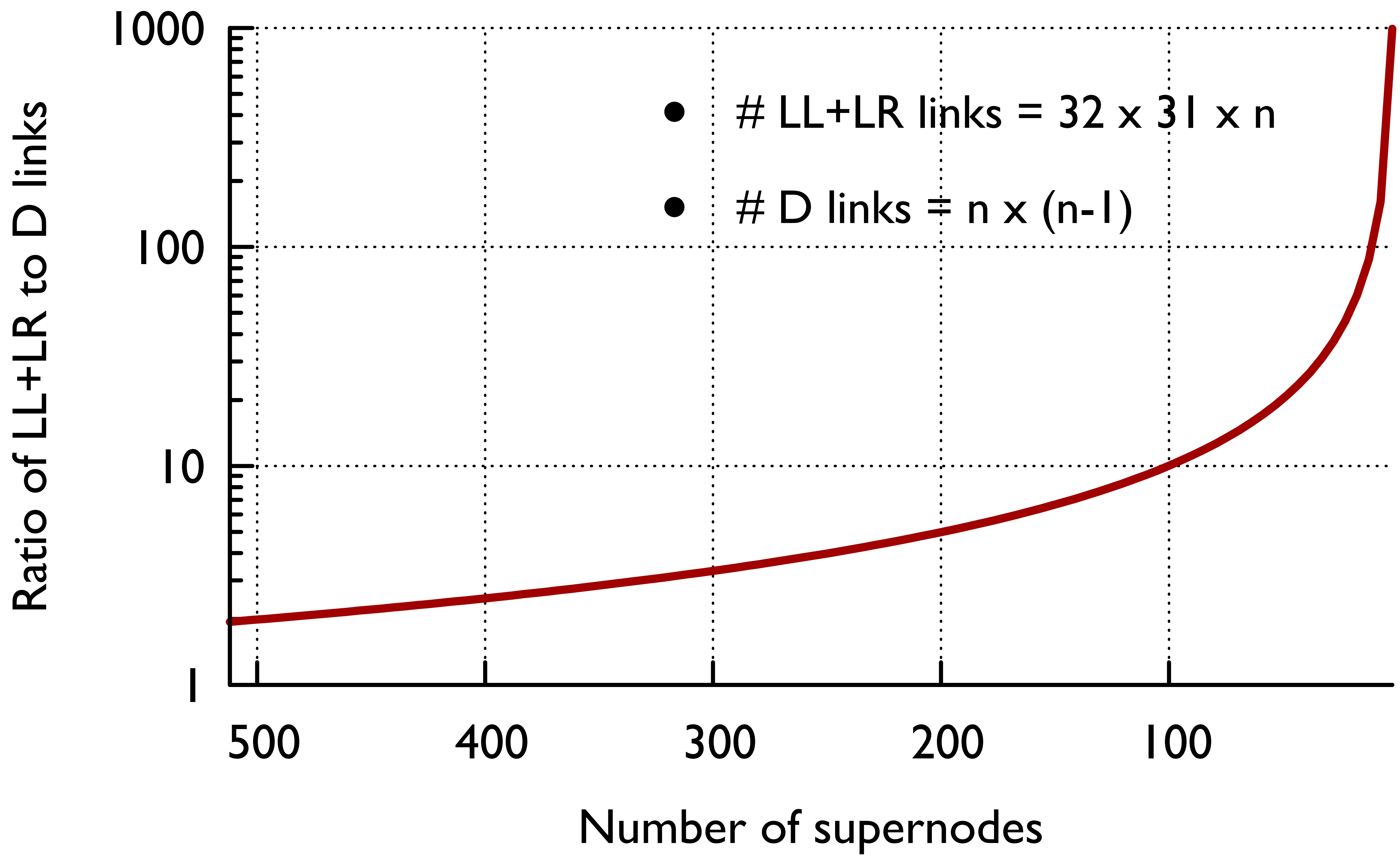
- Lets say we want to run a 2D Stencil on 2 supernodes (64 QCMs)
- Application communication graph: 16 x 4
- MPI-rank ordered mapping leads to significant contention on the single D link
- This is true for running any application with O(1) communicating partners per MPI process



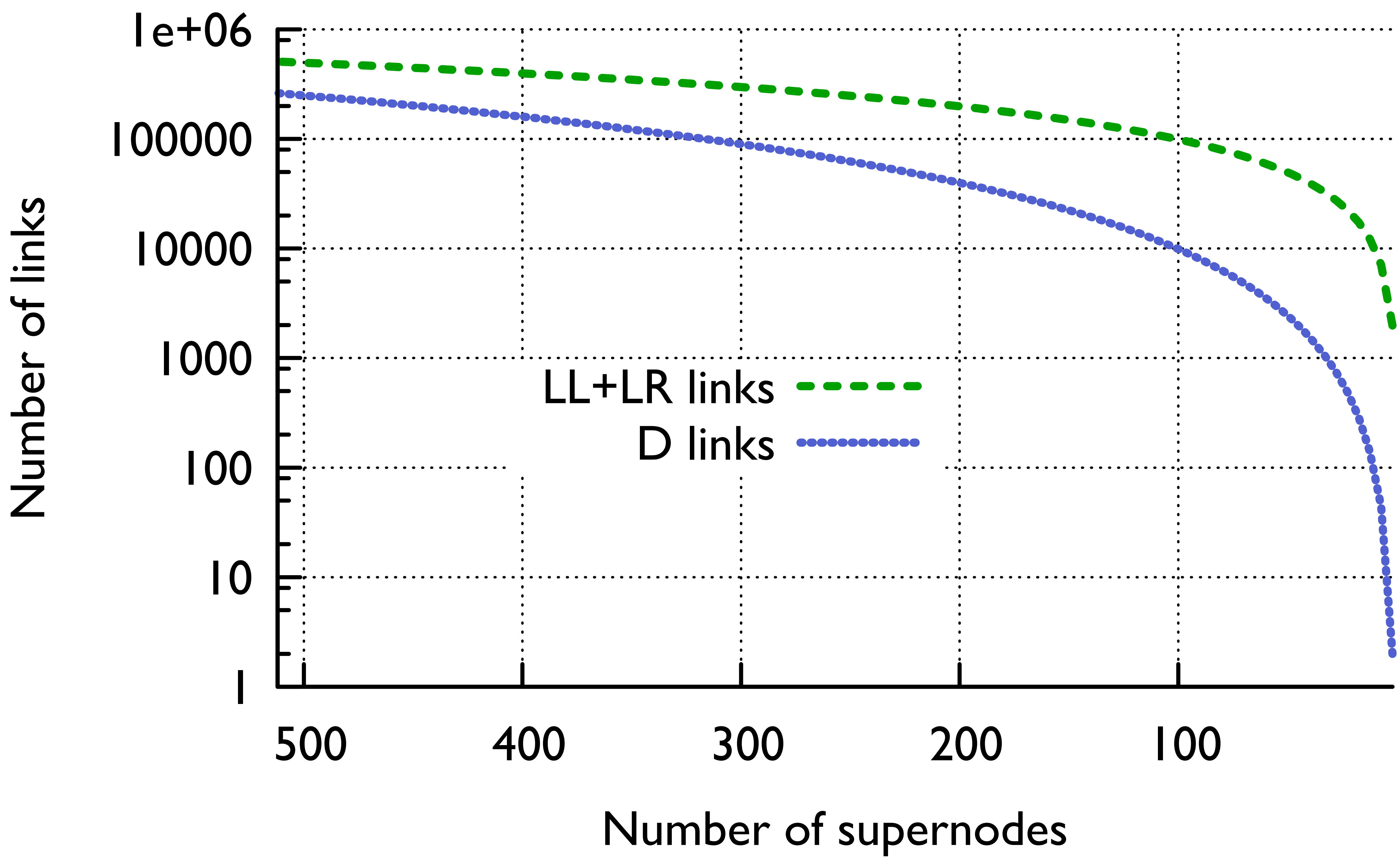
The ‘D’ link bottleneck



The 'D' link bottleneck



The ‘D’ link bottleneck



Software stack/runtime choices

- Job scheduler:
  - Granularity of allocation: QCM (node), drawer, supernode
  - Contiguous allocation, random allocation or careful topology-aware allocation
- Routing:
  - Direct versus random indirect
- Mapping
  - Is it important for optimal performance?

Simulation study – BigSim

Simulation study – BigSim

- Use BigSim for emulation and simulation of a future machine - detailed packet-level network simulation
  - Compute time prediction, trace collection, simulation

Simulation study – BigSim

- Use BigSim for emulation and simulation of a future machine - detailed packet-level network simulation
  - Compute time prediction, trace collection, simulation
- Three different benchmarks
  - 2-dimensional five-point stencil
  - 4-dimensional nine-point stencil
  - n-targets multicast pattern

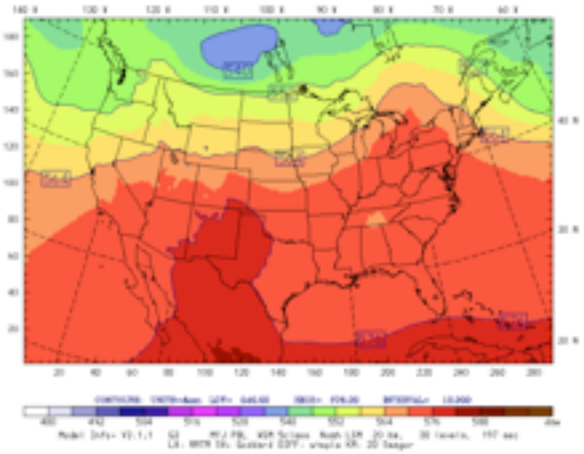
Simulation study – BigSim

- Use BigSim for emulation and simulation of a future machine - detailed packet-level network simulation
  - Compute time prediction, trace collection, simulation
- Three different benchmarks
  - 2-dimensional five-point stencil
  - 4-dimensional nine-point stencil
  - n-targets multicast pattern
- Two job allocation sizes
  - 64 supernodes
  - 300 supernodes

Simulation study – BigSim

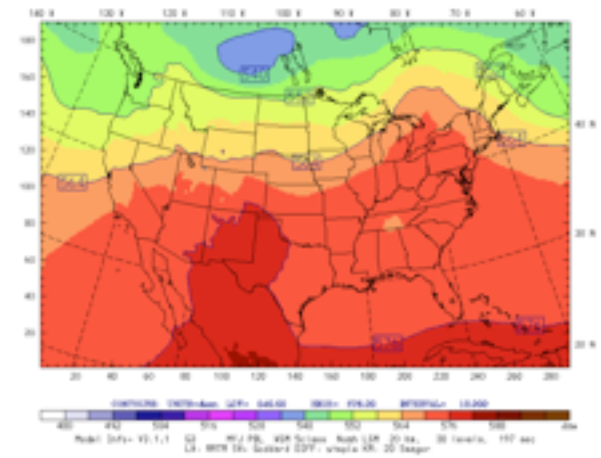
- Use BigSim for emulation and simulation of a future machine - detailed packet-level network simulation
  - Compute time prediction, trace collection, simulation
- Three different benchmarks
  - 2-dimensional five-point stencil
  - 4-dimensional nine-point stencil
  - n-targets multicast pattern
- Two job allocation sizes
  - 64 supernodes
  - 300 supernodes

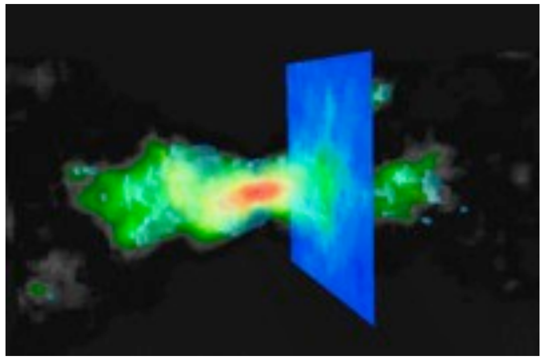
WRF



Simulation study – BigSim

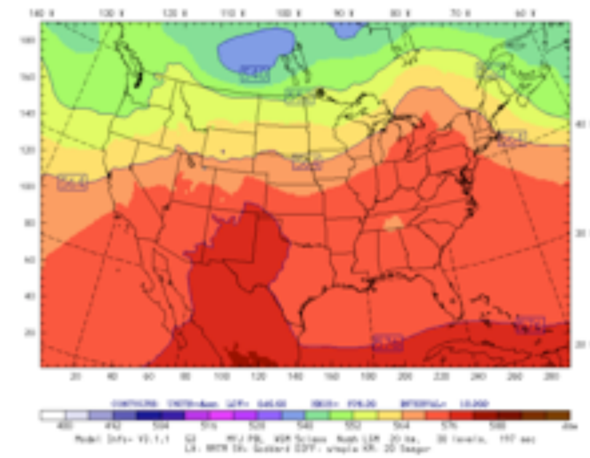
- Use BigSim for emulation and simulation of a future machine - detailed packet-level network simulation
  - Compute time prediction, trace collection, simulation
- Three different benchmarks
  - 2-dimensional five-point stencil WRF
  - 4-dimensional nine-point stencil MILC
  - n-targets multicast pattern
- Two job allocation sizes
  - 64 supernodes
  - 300 supernodes

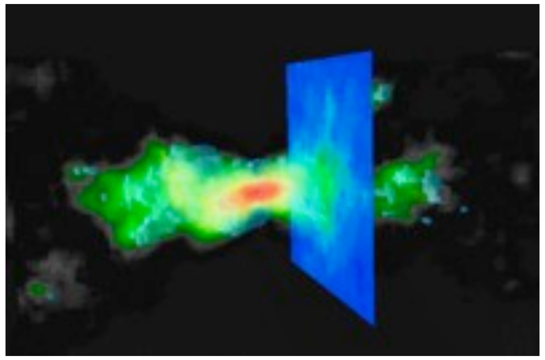


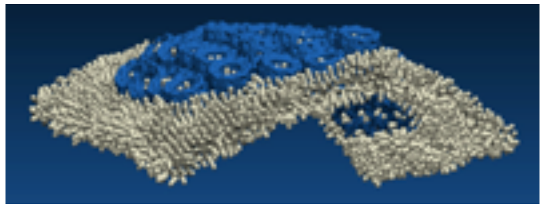


Simulation study – BigSim

- Use BigSim for emulation and simulation of a future machine - detailed packet-level network simulation
  - Compute time prediction, trace collection, simulation
- Three different benchmarks
  - 2-dimensional five-point stencil WRF
  - 4-dimensional nine-point stencil MILC
  - n-targets multicast pattern NAMD
- Two job allocation sizes
  - 64 supernodes
  - 300 supernodes





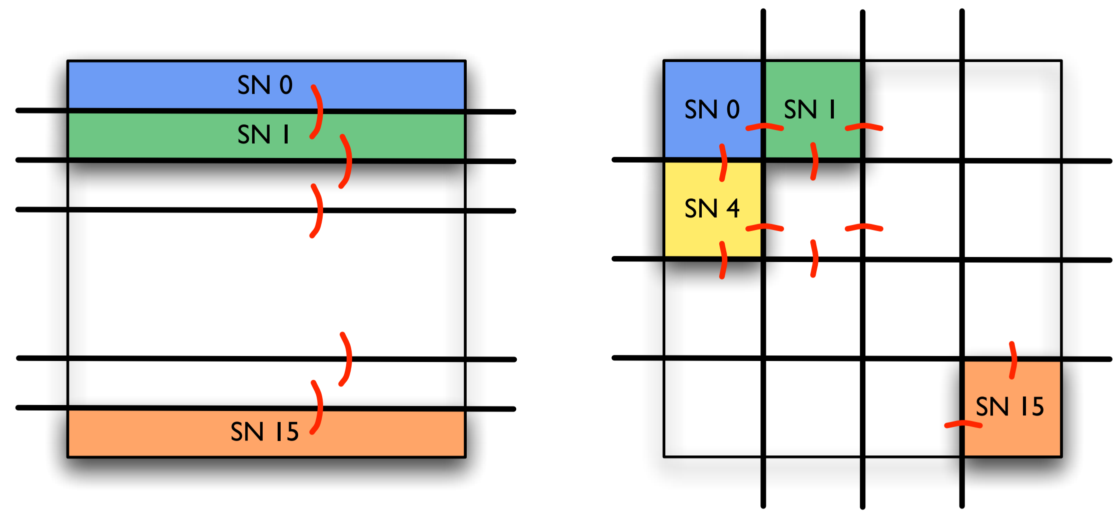


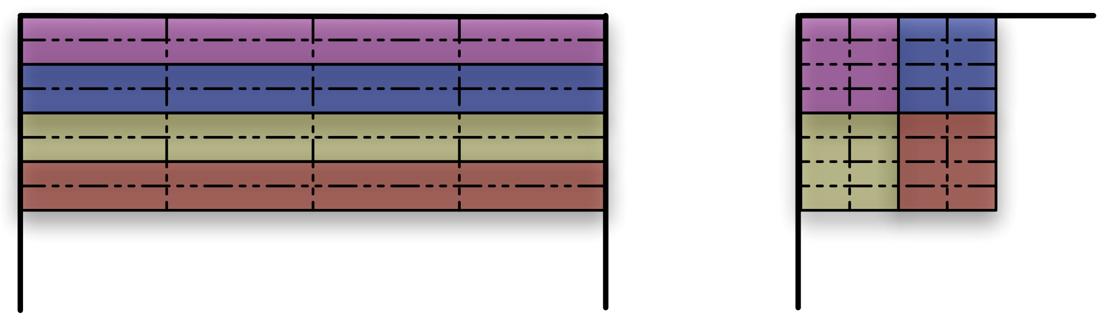
Prediction Methodology

- Compute time prediction: run on a 3.8 GHz Power7 processor to get timings for sequential computation
- Emulation: obtain traces by running on 512-1360 cores of a 1.9 GHz Power5 cluster
- Simulations on one node of a SGI Altix 1000 shared memory machine

Mapping and Routing

- Default mapping: MPI rank-ordered mapping
- Blocking: at the level of nodes, then drawers and supernodes
- Random mapping on nodes
- Routing choices:
  - Indirect routing w/ default mapping
  - Indirect routing w/ random drawers mapping





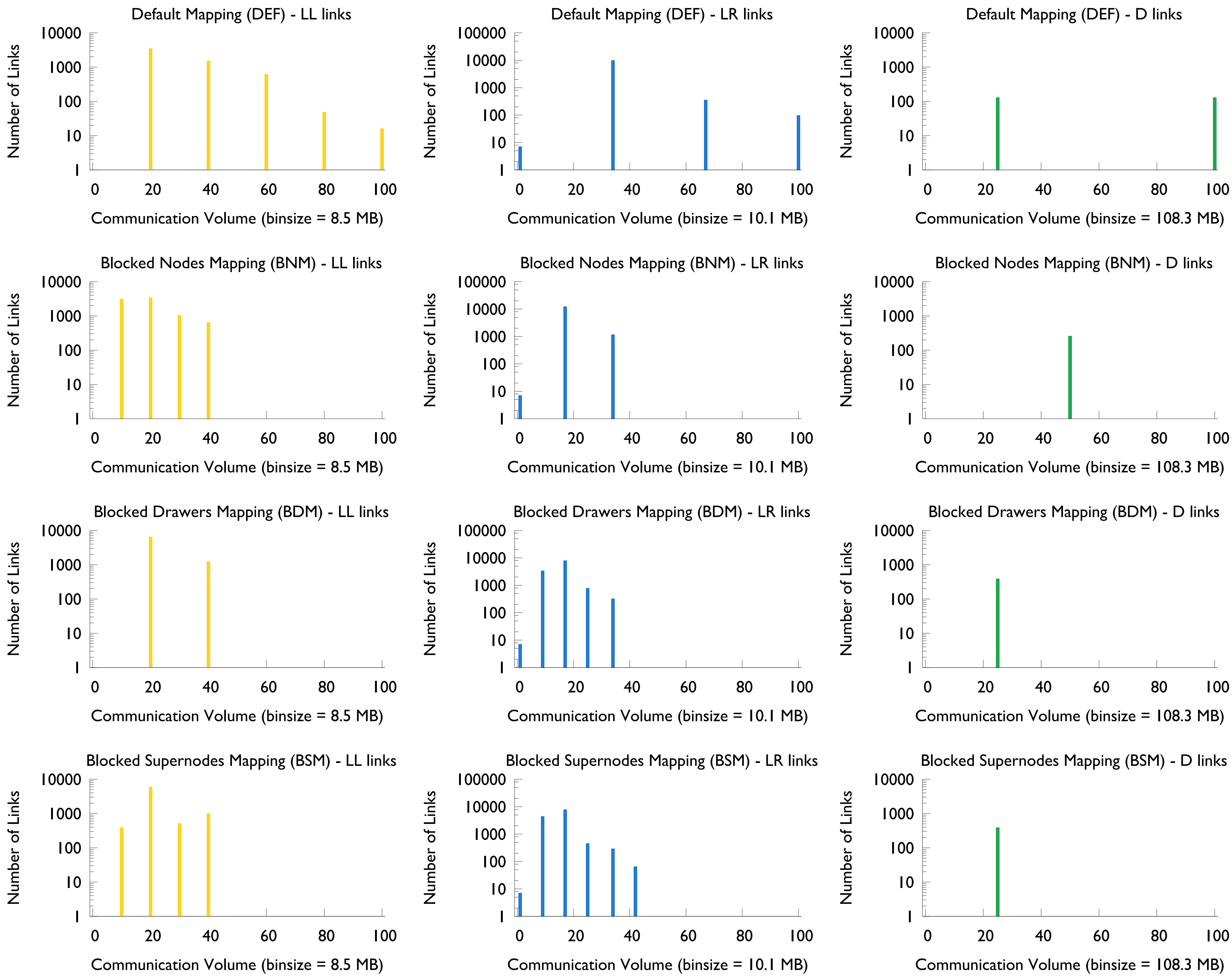
4-dimensional stencil

- Representative of MILC, a Lattice QCD code
- Each MPI task has 64 x 64 x 64 x 64 elements
- Size of messages exchanged = 2 MB

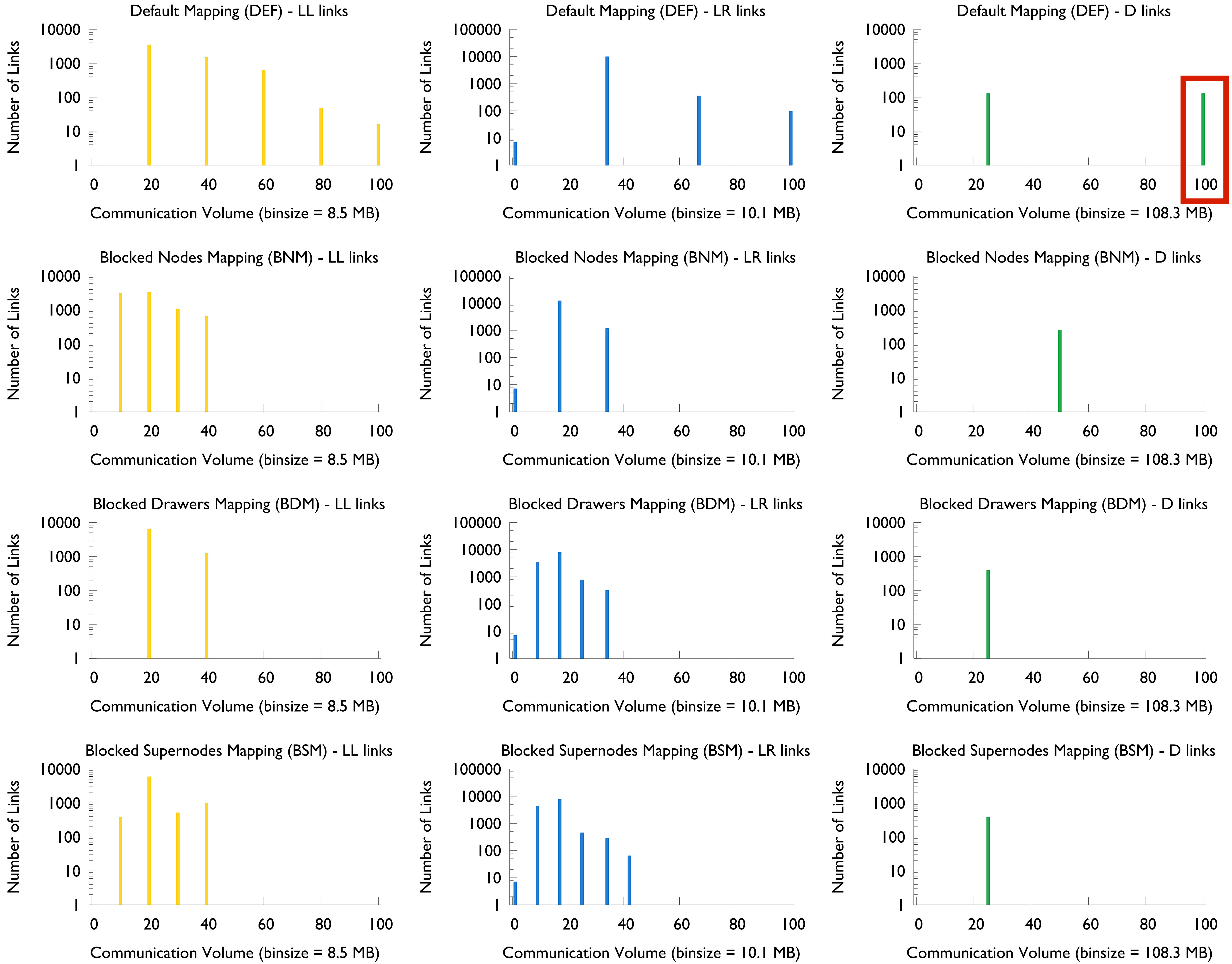
Experiments

- Direct routing:
  - Default MPI rank-ordered mapping (DEF)
  - Blocking MPI tasks at the level of nodes (BNM)
  - Blocking at the level of drawers (BDM)
  - Blocking at the level of supernodes (BSM)
  - Random mapping at the level of nodes (RNM)
  - Random mapping at the level of drawers (RDM)
- Indirect routing
  - Default MPI rank-ordered mapping (DFI)
  - Random mapping at the level of drawers (RDI)

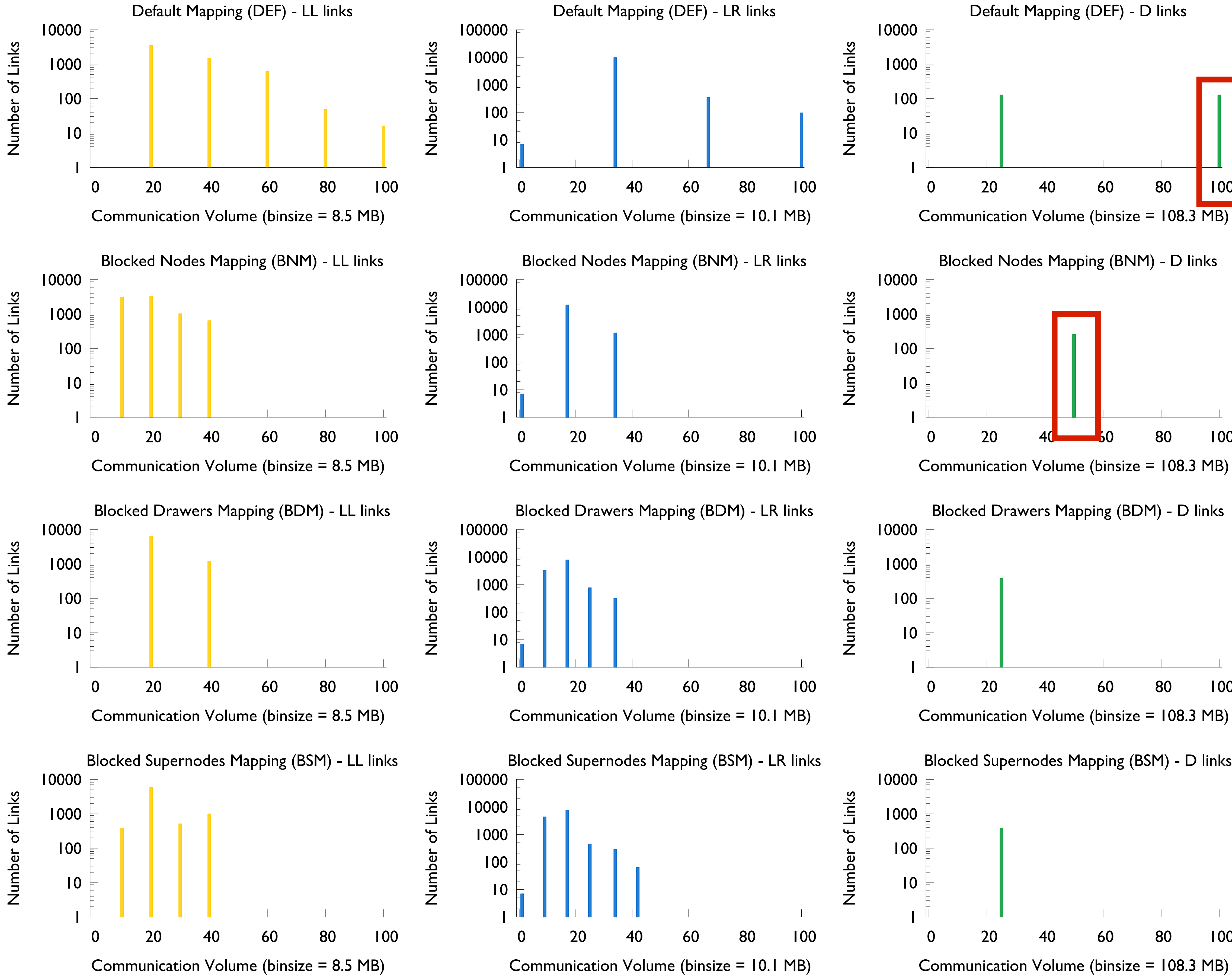
4D Stencil on 64 supernodes



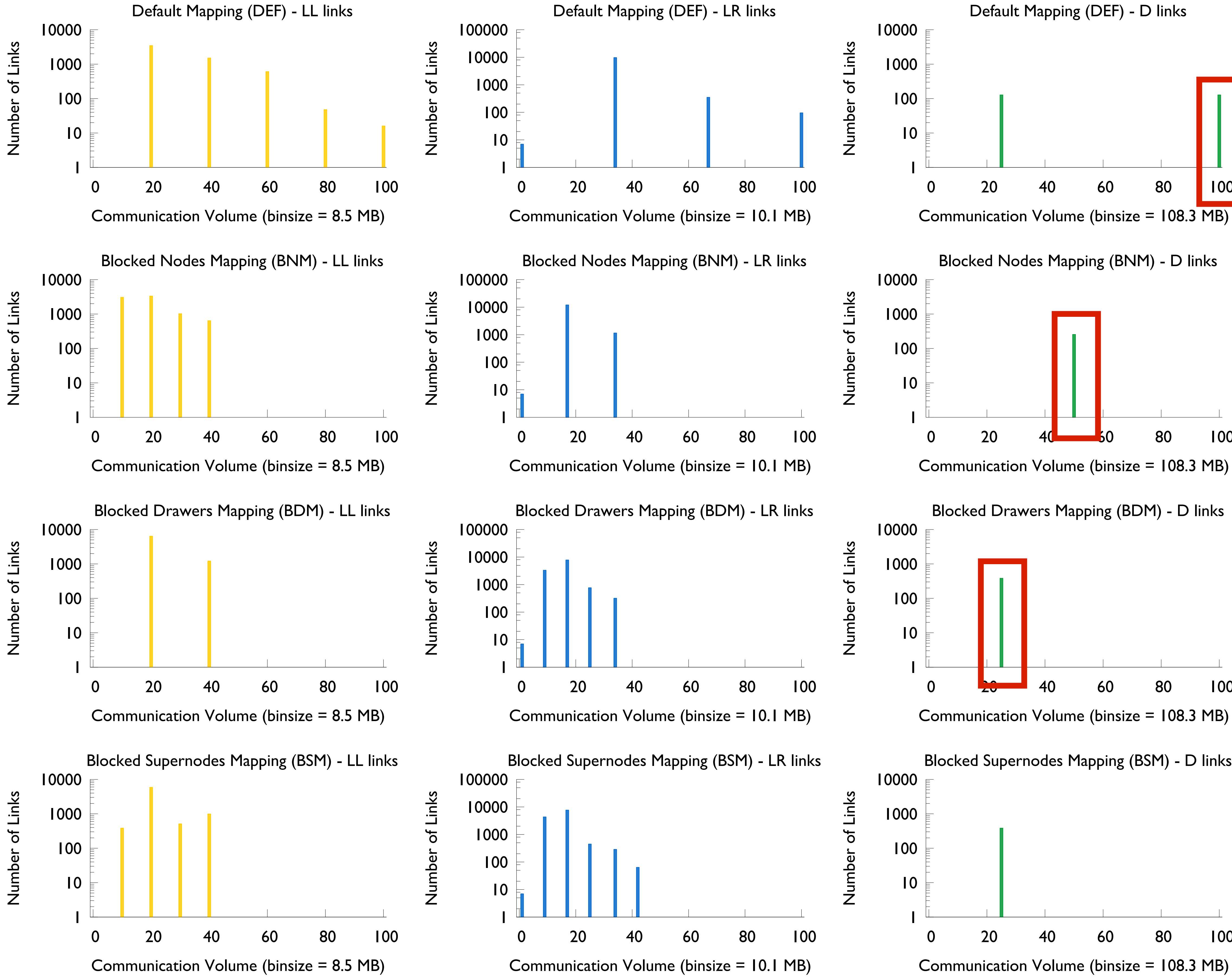
4D Stencil on 64 supernodes



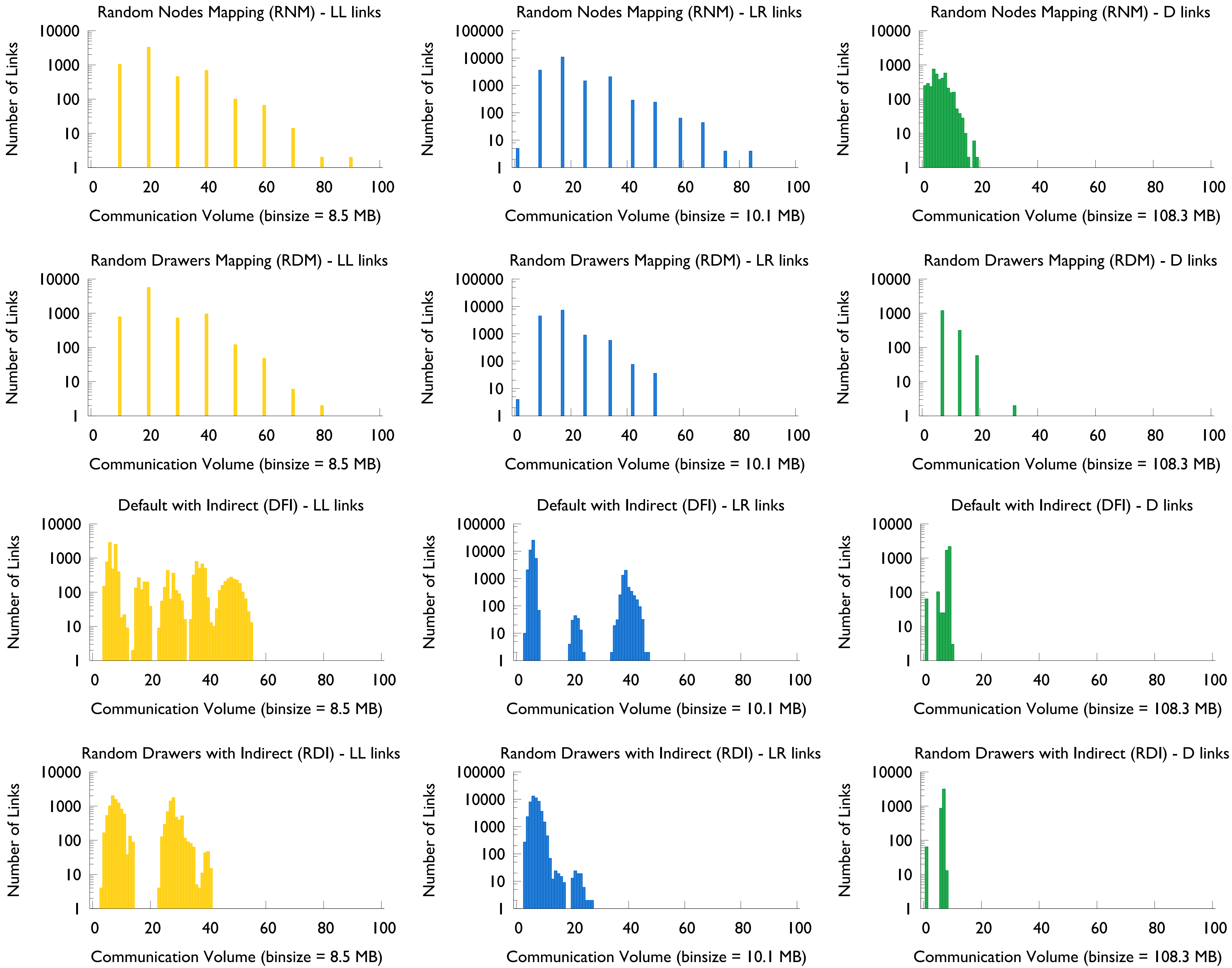
4D Stencil on 64 supernodes



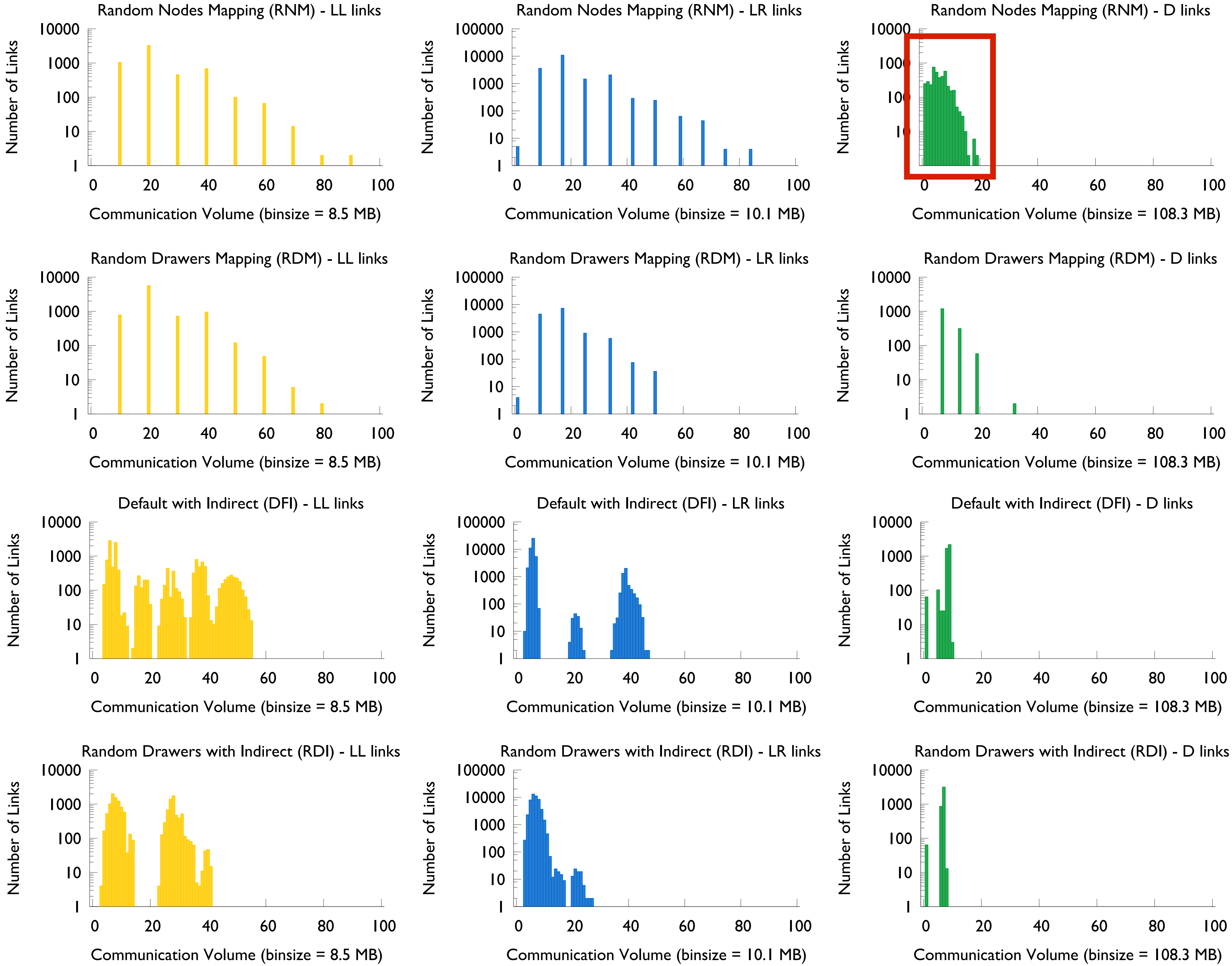
4D Stencil on 64 supernodes



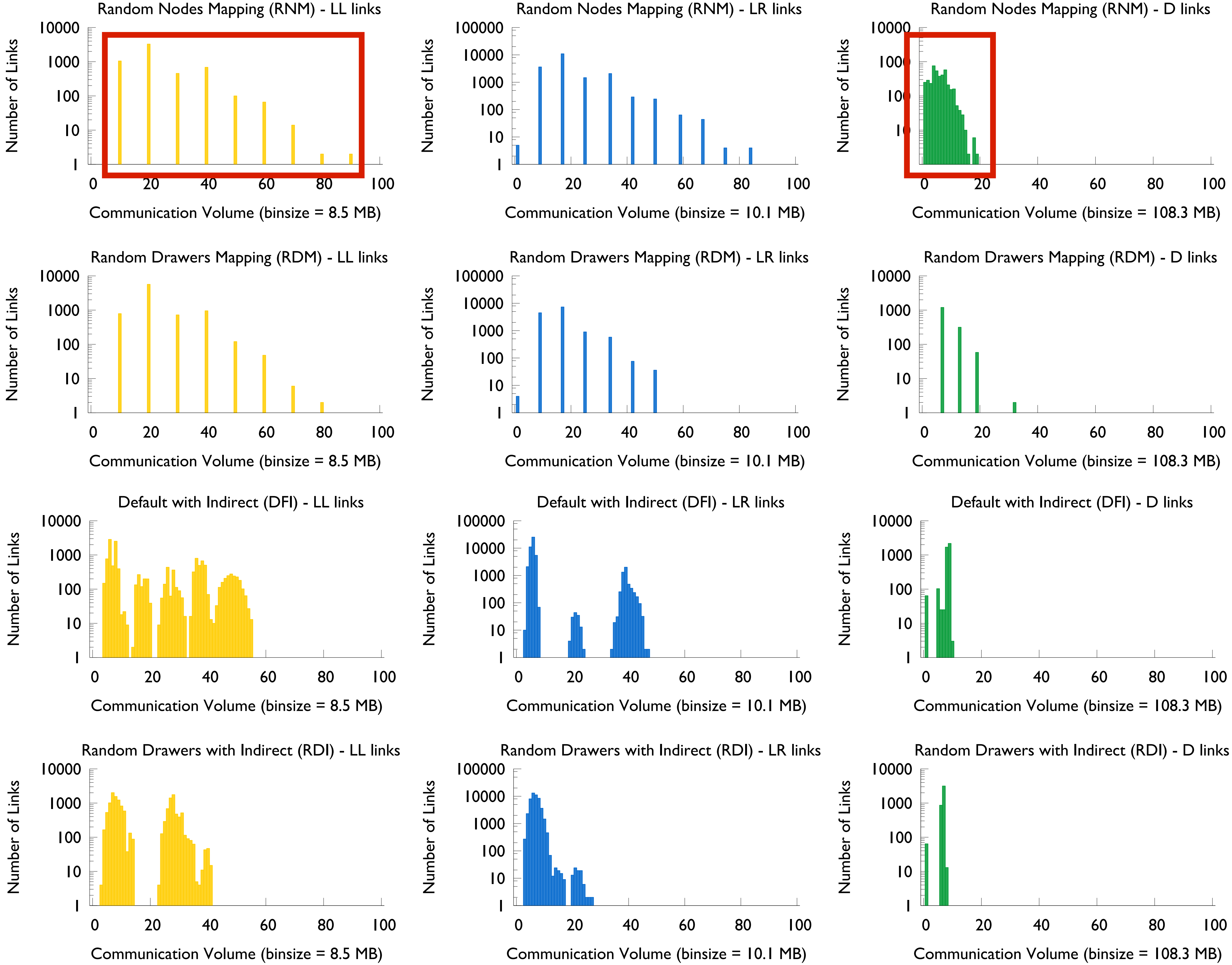
4D Stencil on 64 supernodes



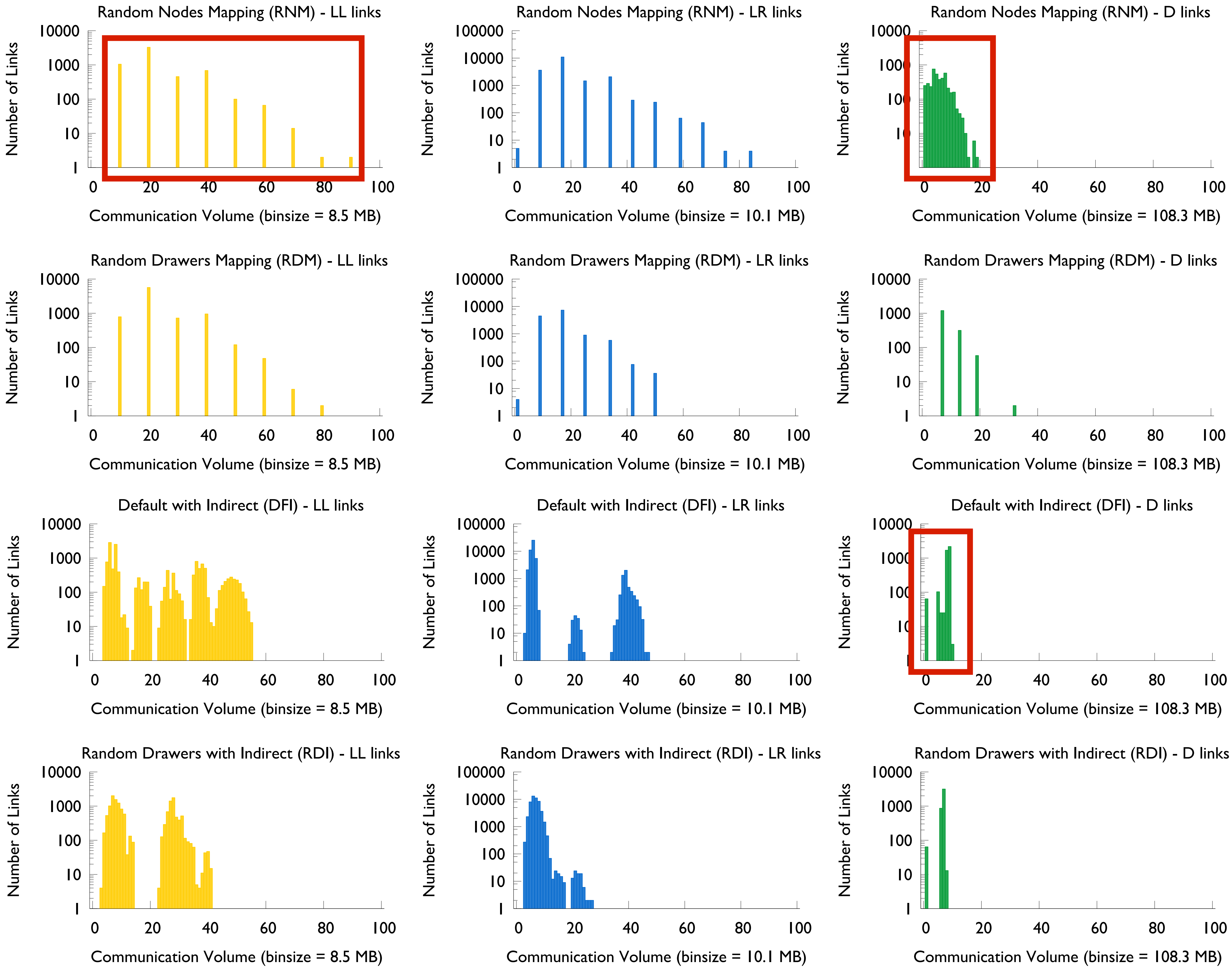
4D Stencil on 64 supernodes



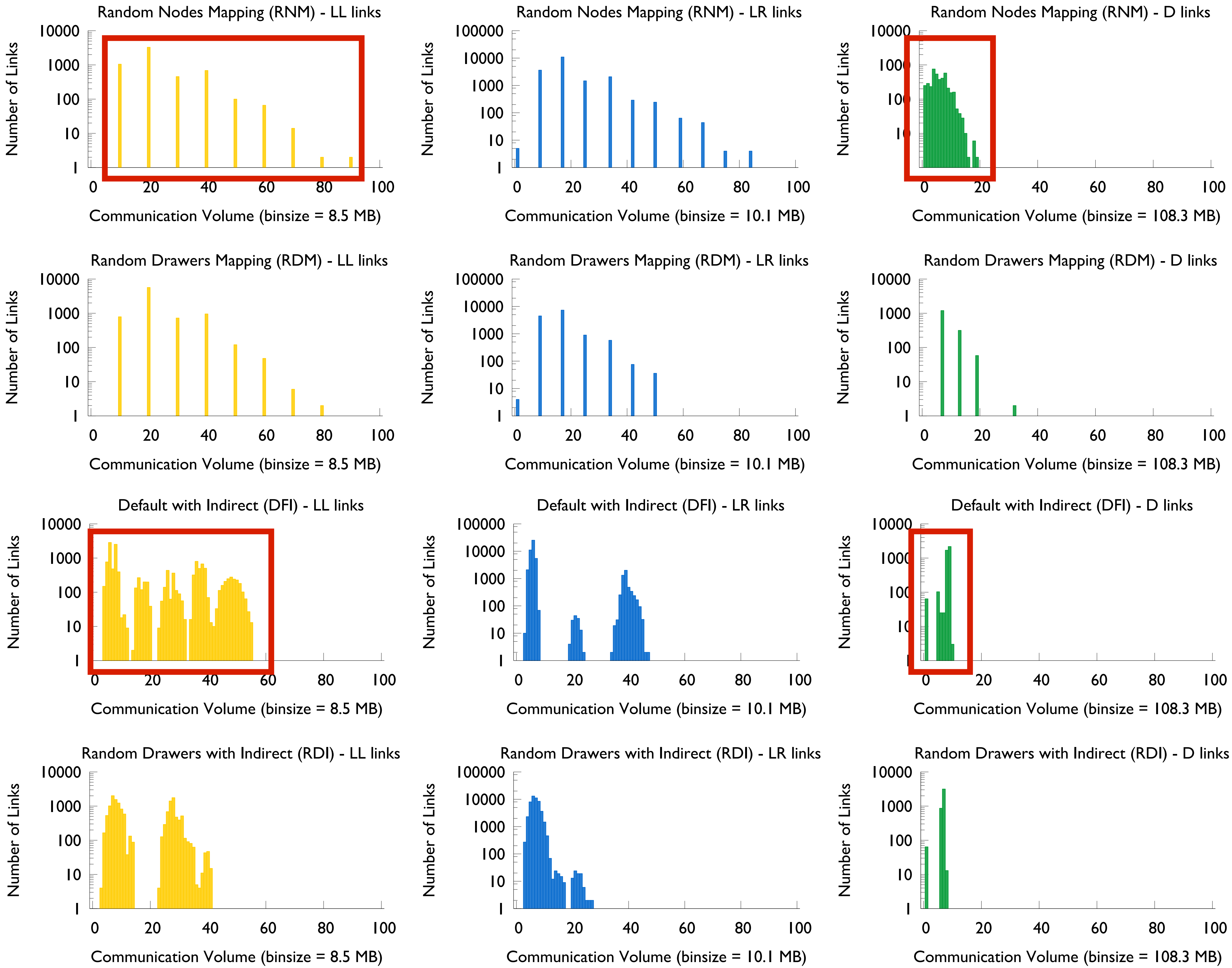
4D Stencil on 64 supernodes



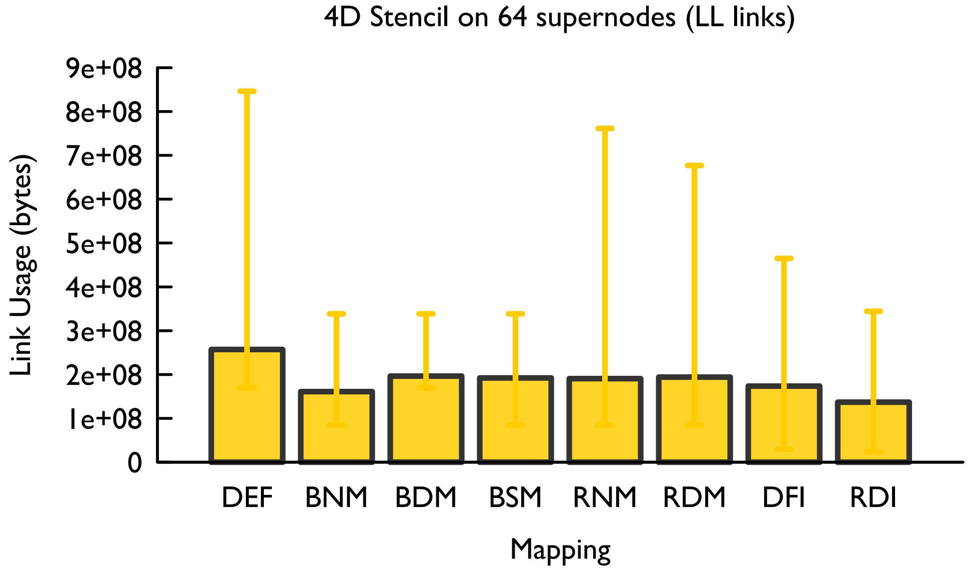
4D Stencil on 64 supernodes

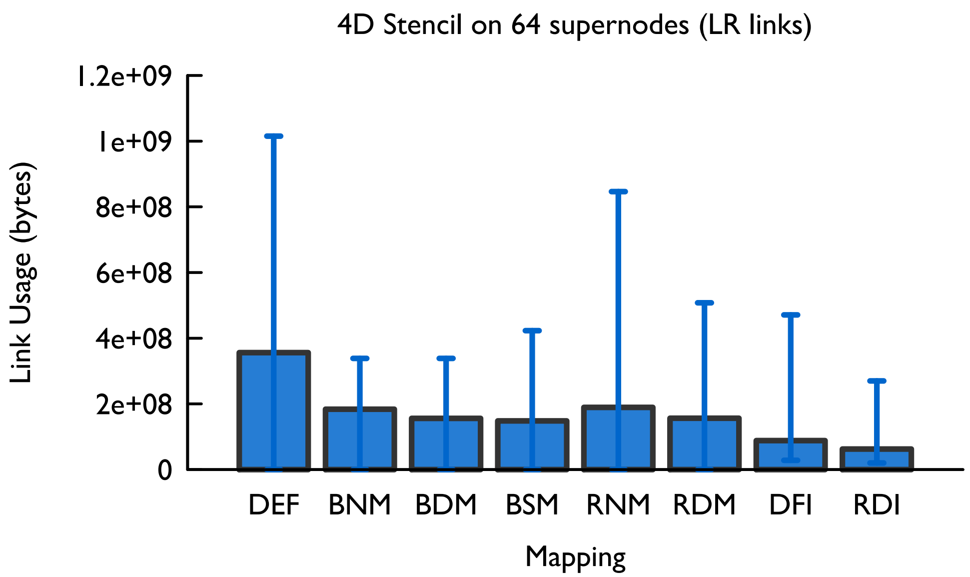


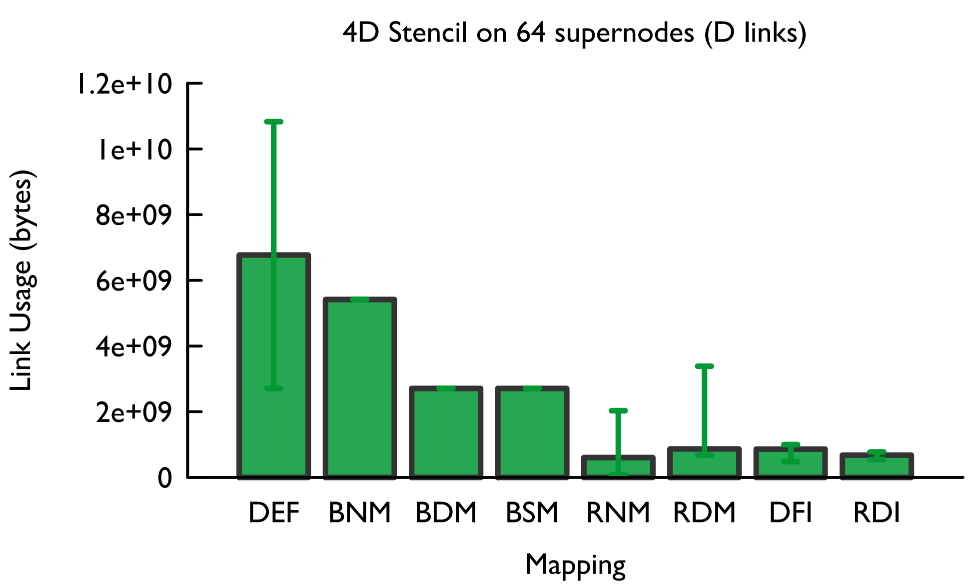
4D Stencil on 64 supernodes



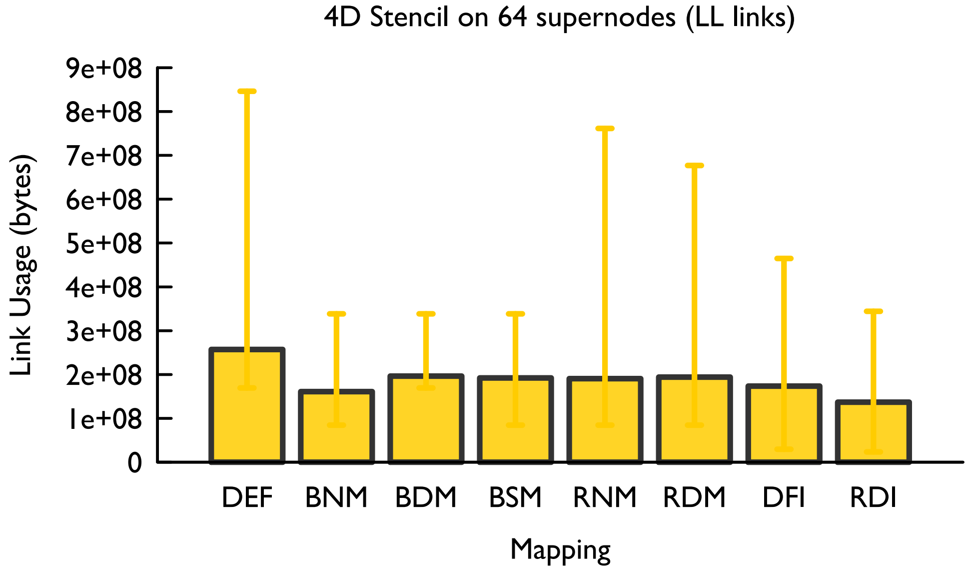
4D Stencil on 64 supernodes

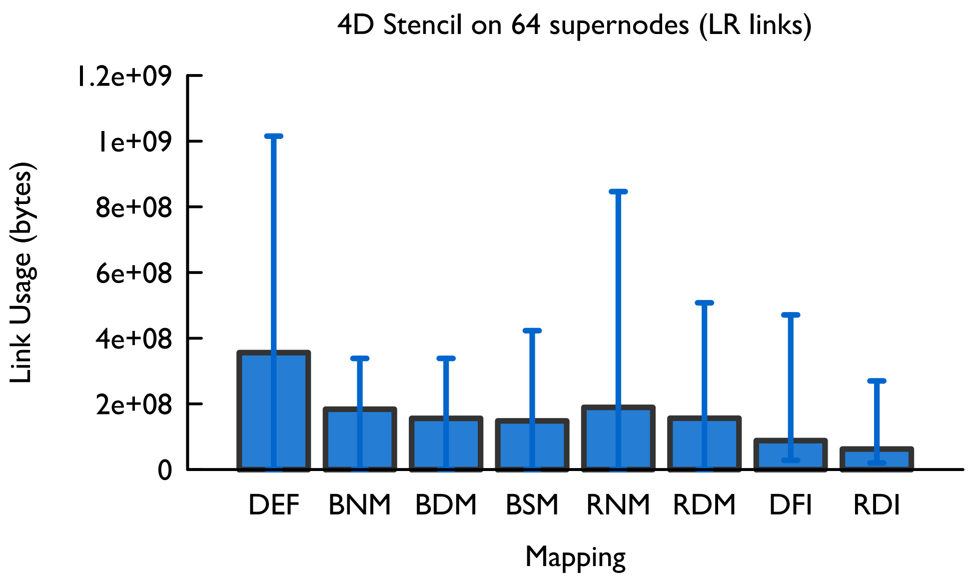


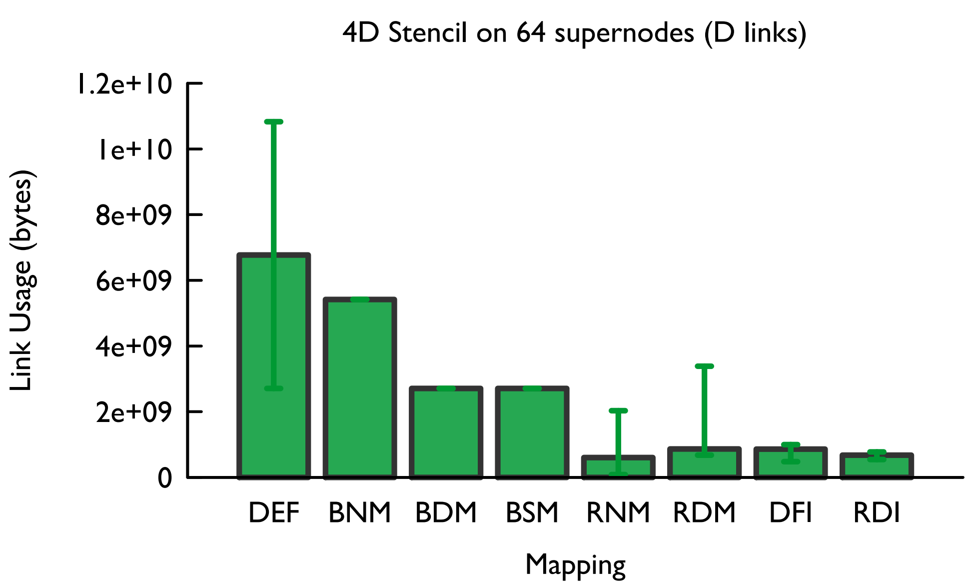


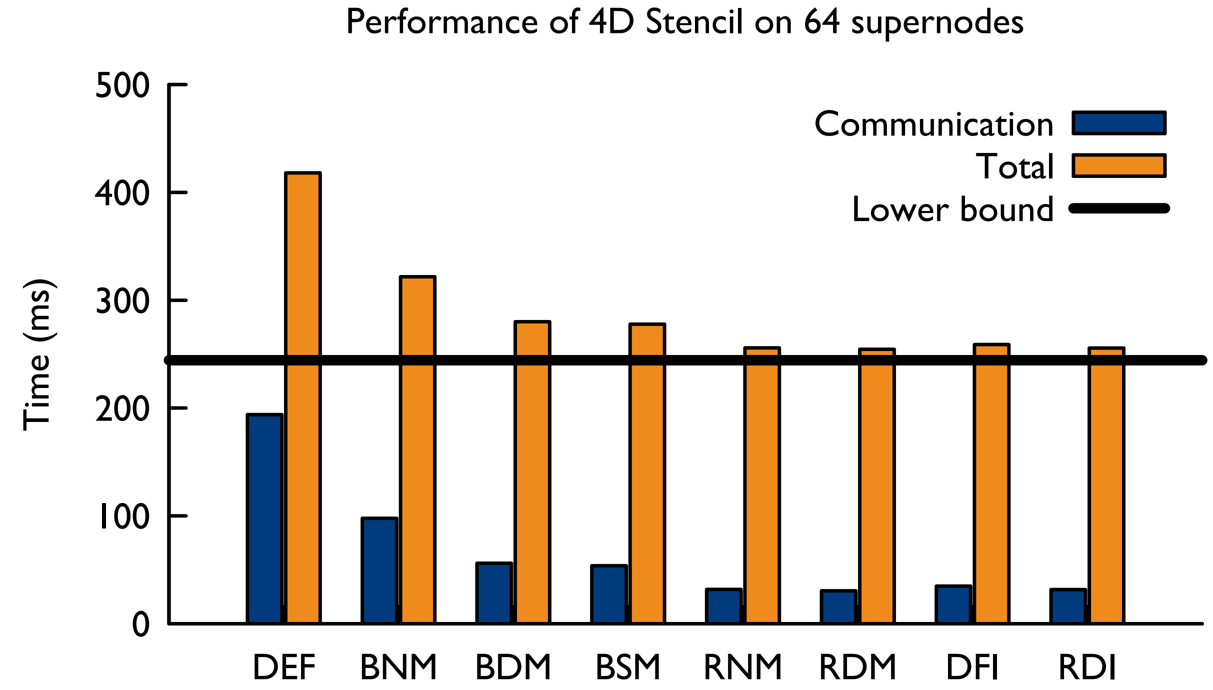


4D Stencil on 64 supernodes

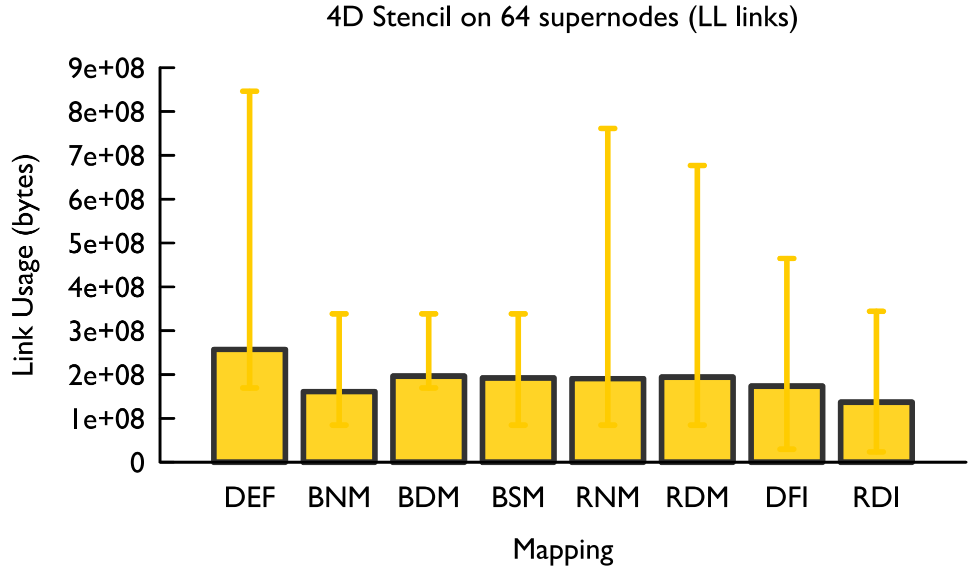


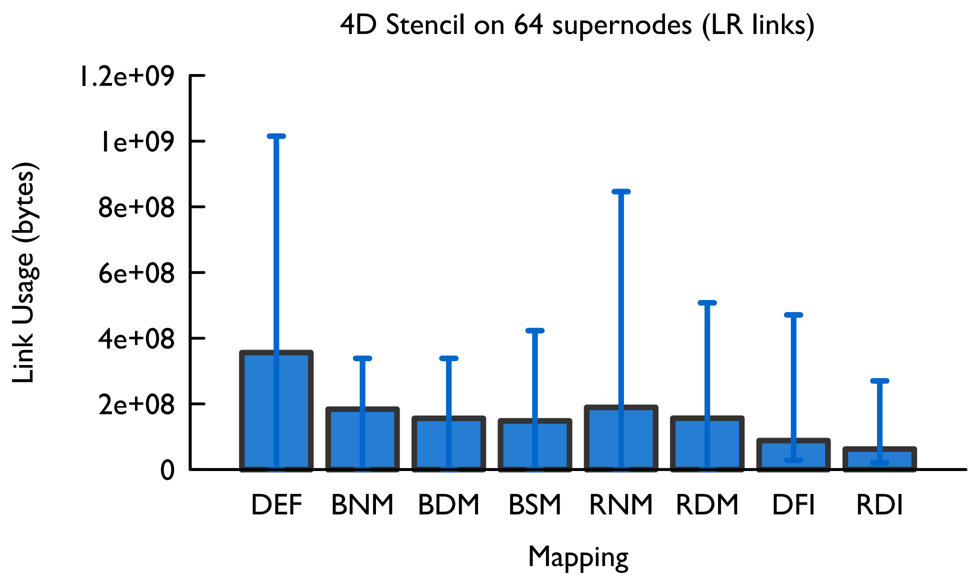


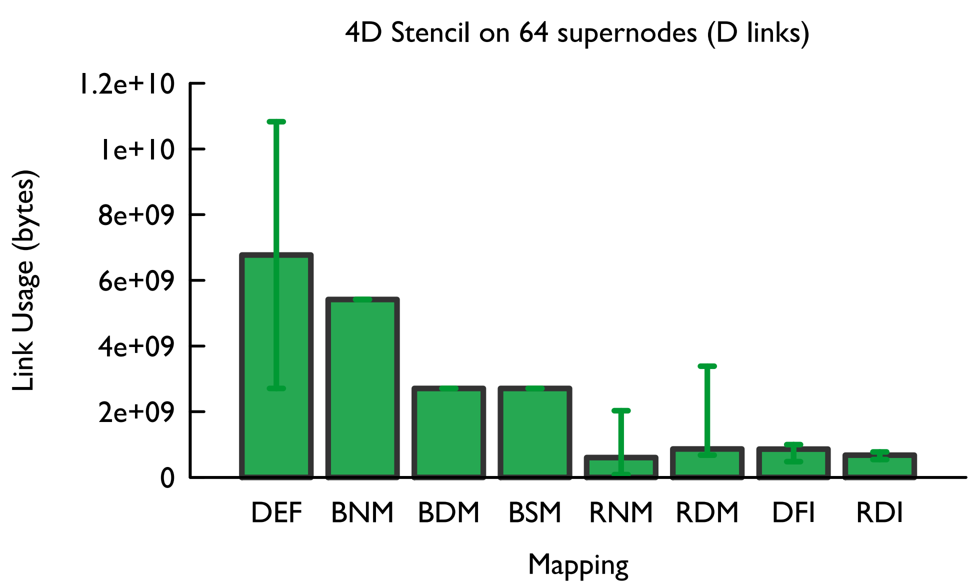


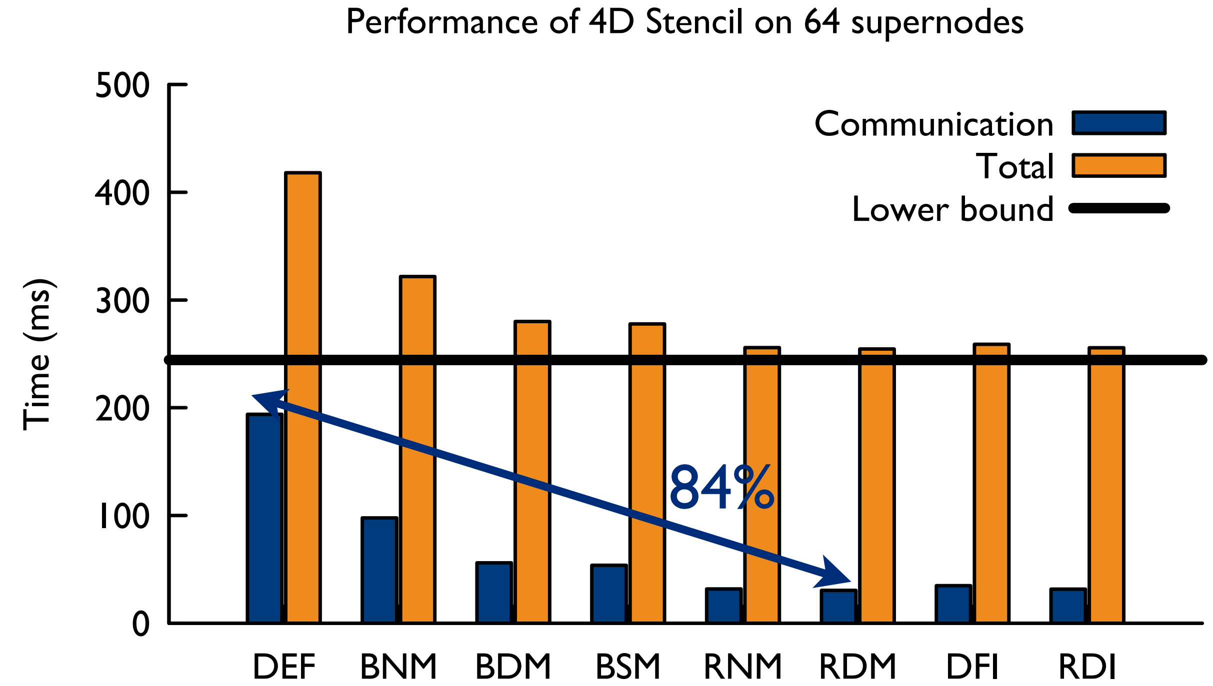


4D Stencil on 64 supernodes

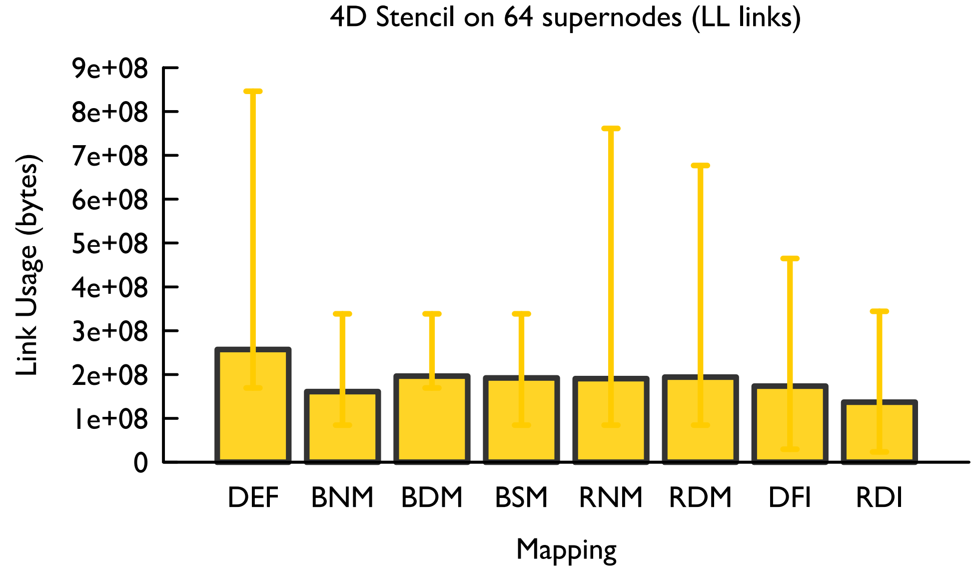


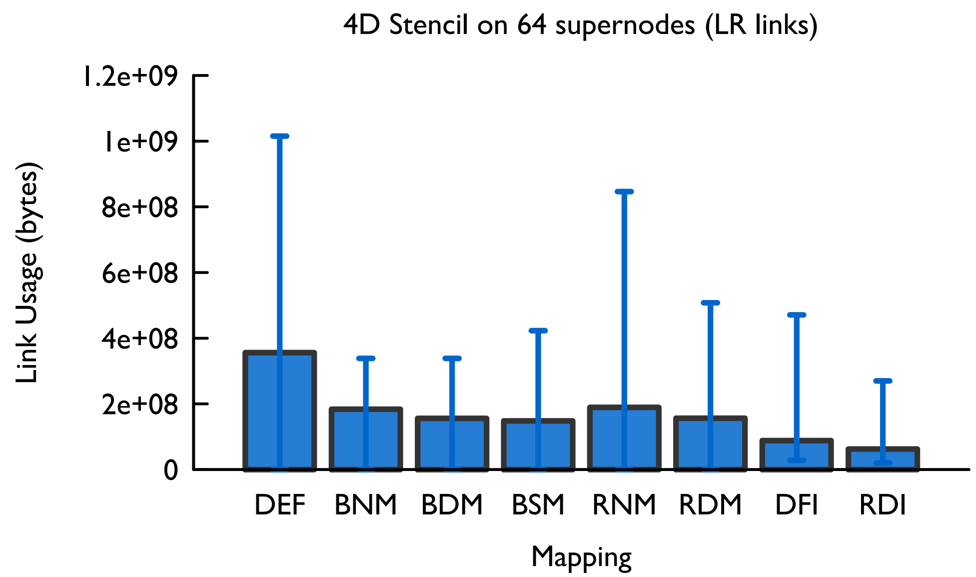


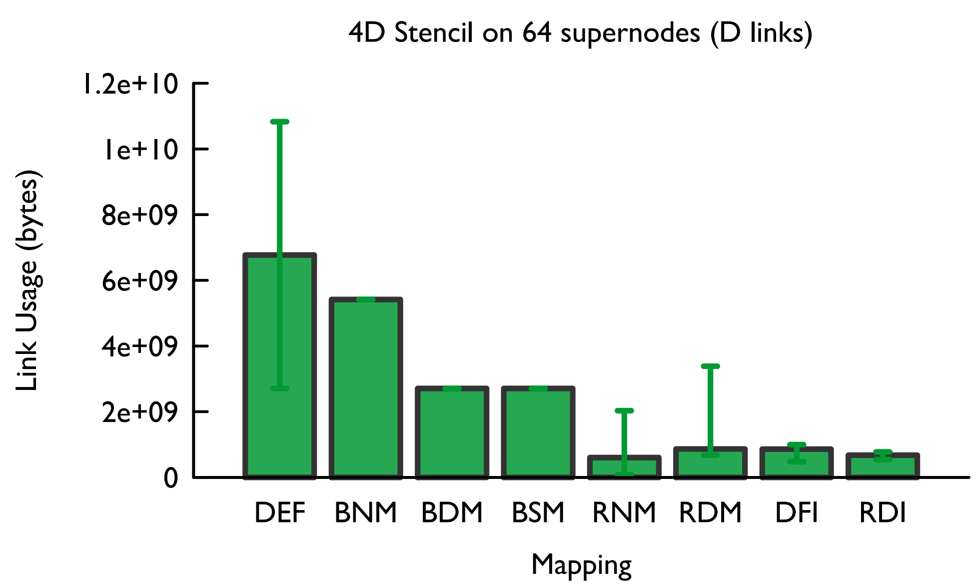


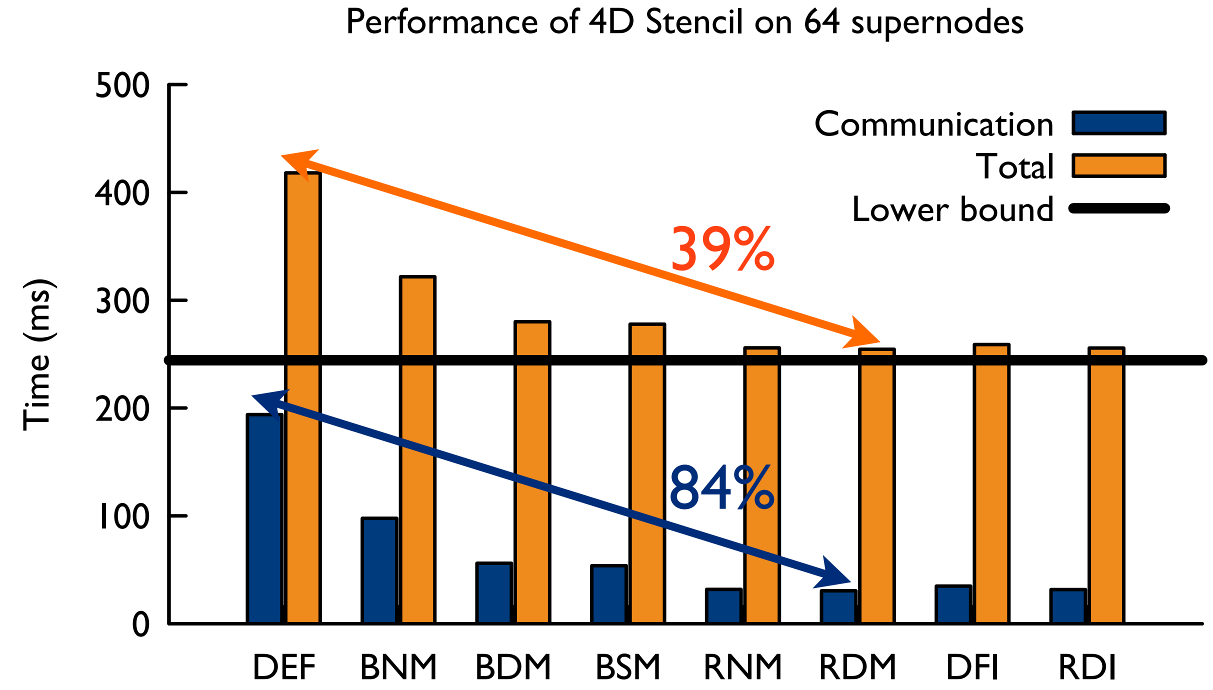


4D Stencil on 64 supernodes

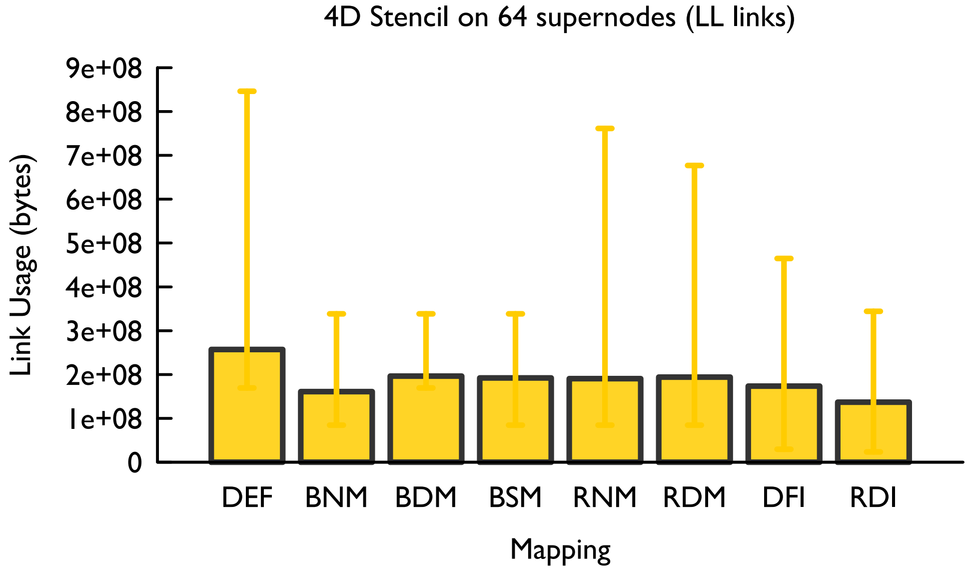


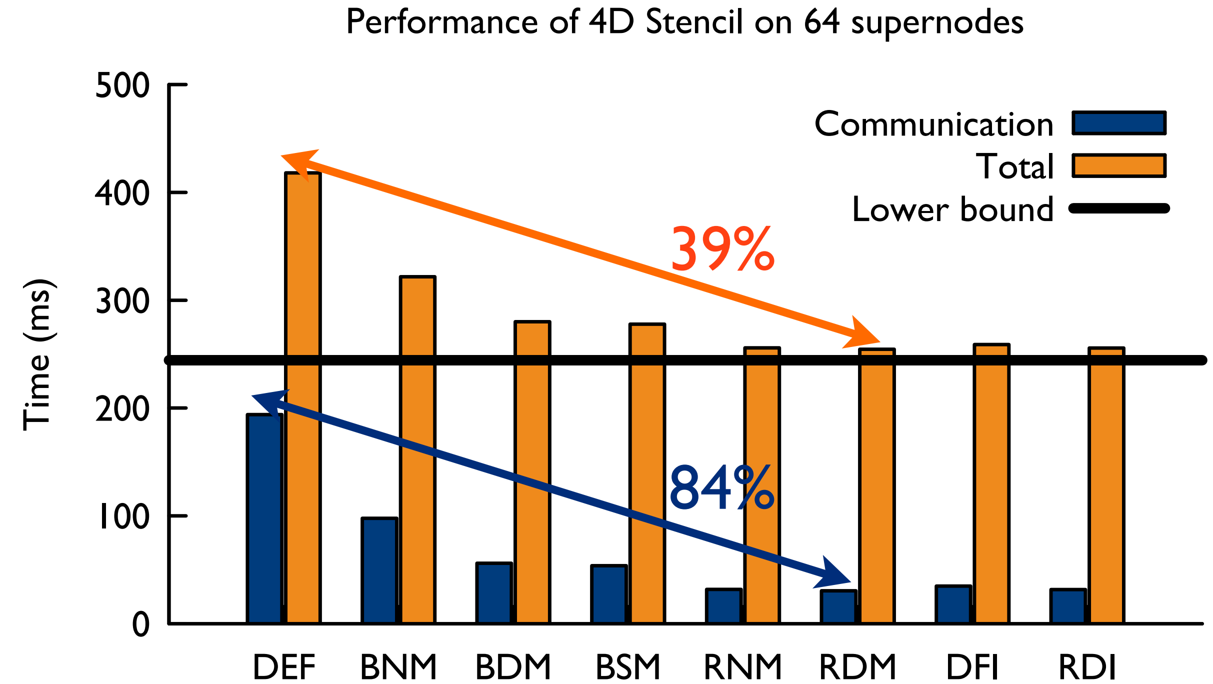


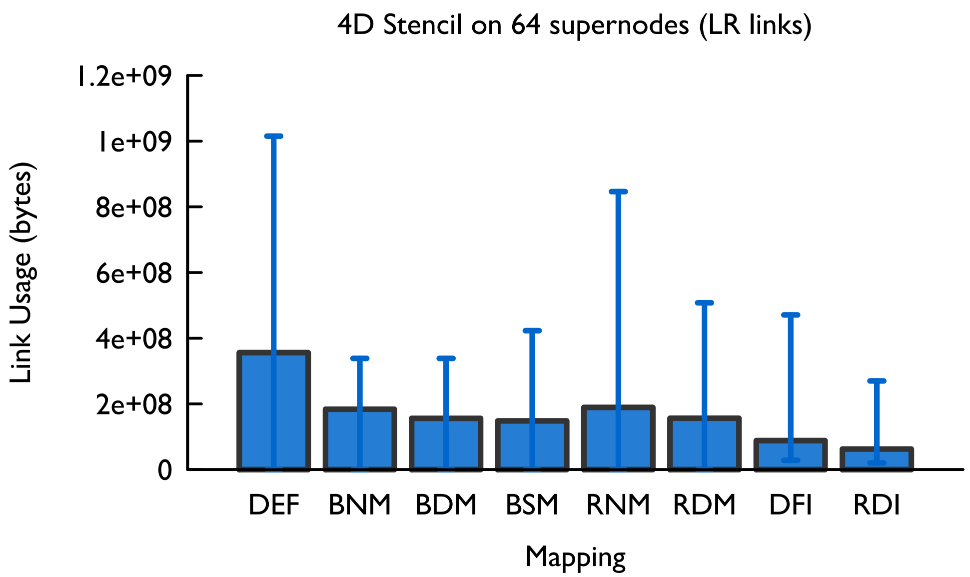


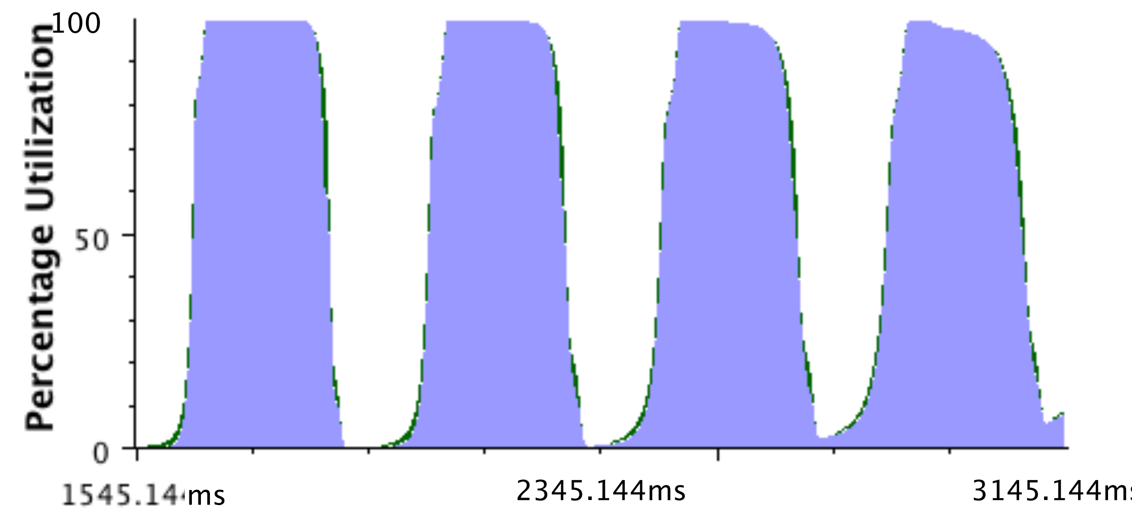


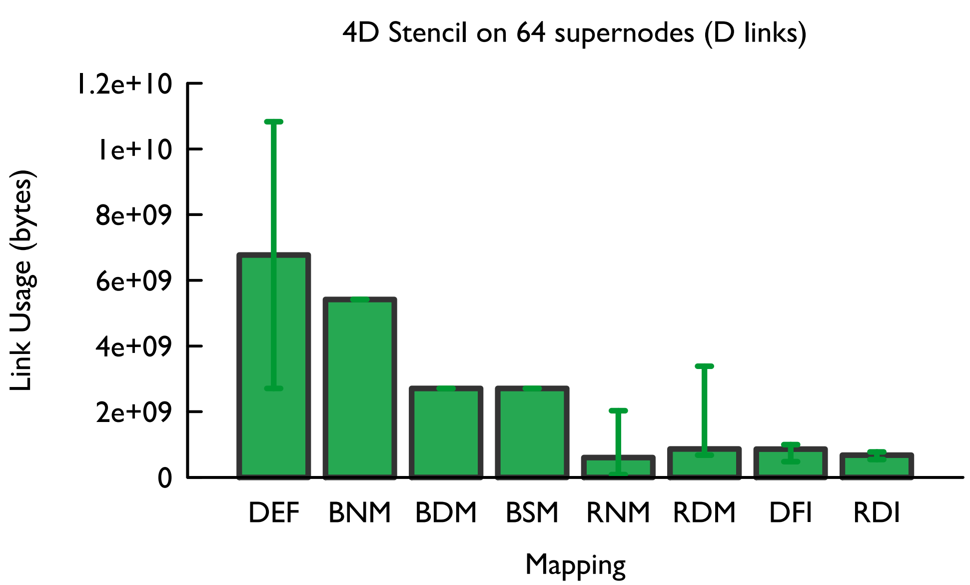
4D Stencil on 64 supernodes

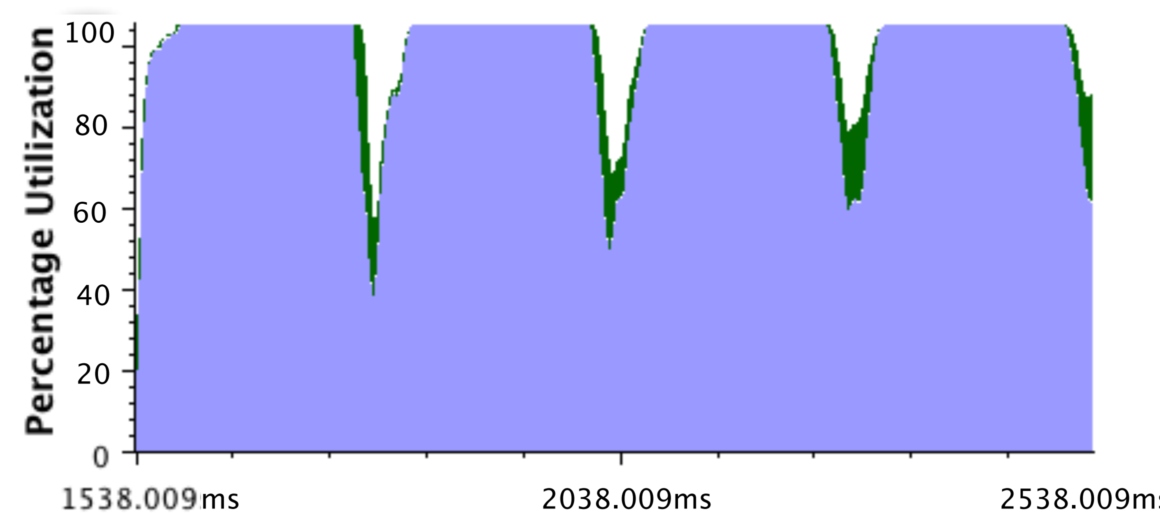








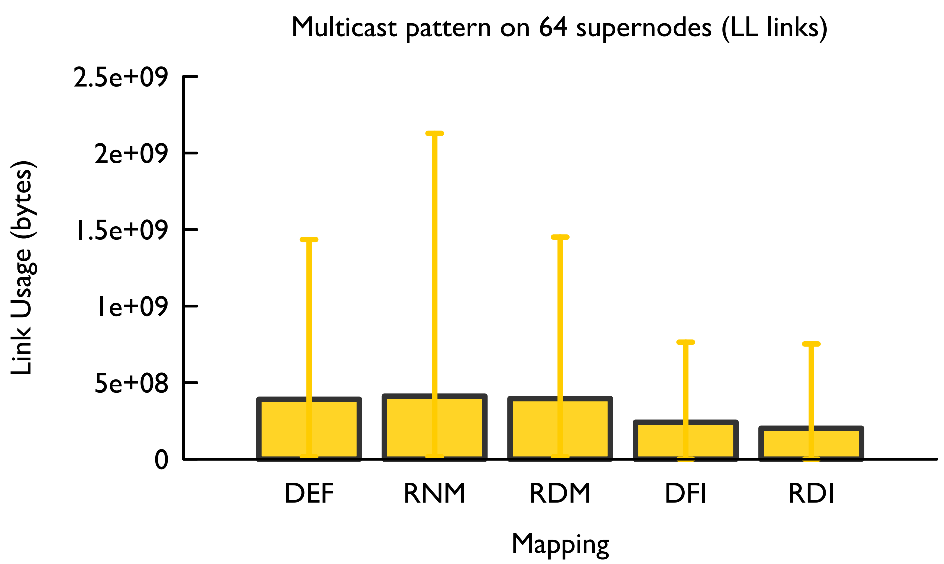


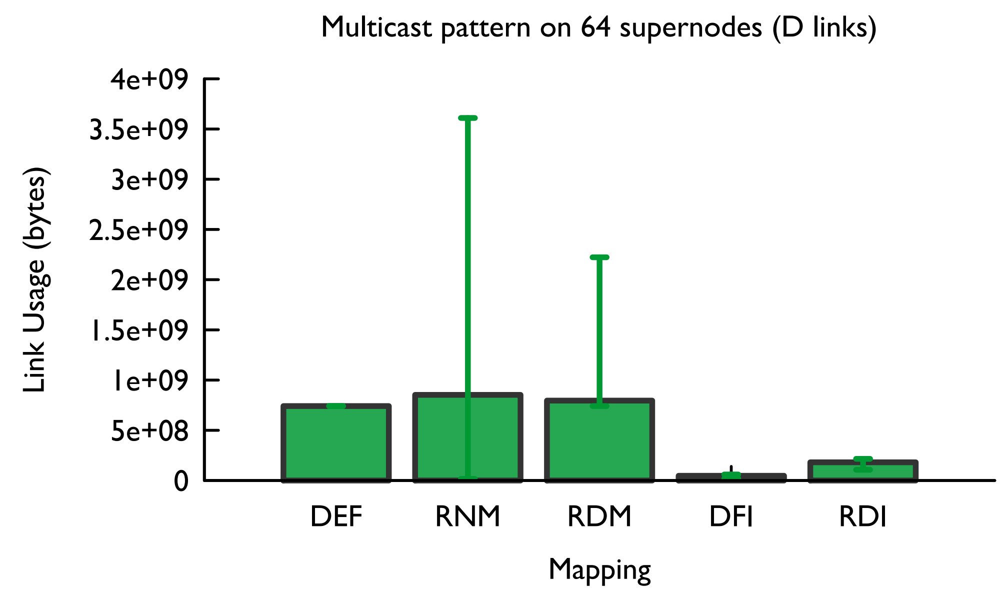


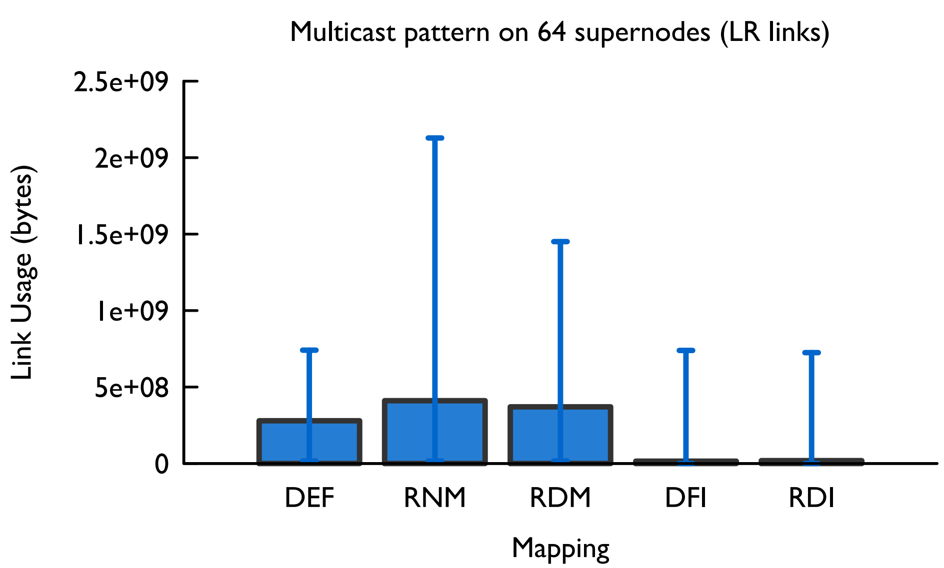
n-targets multicast pattern

- Similar to the communication pattern in NAMD
- Each MPI task sends messages to 14 others
- Message size = 1 MB

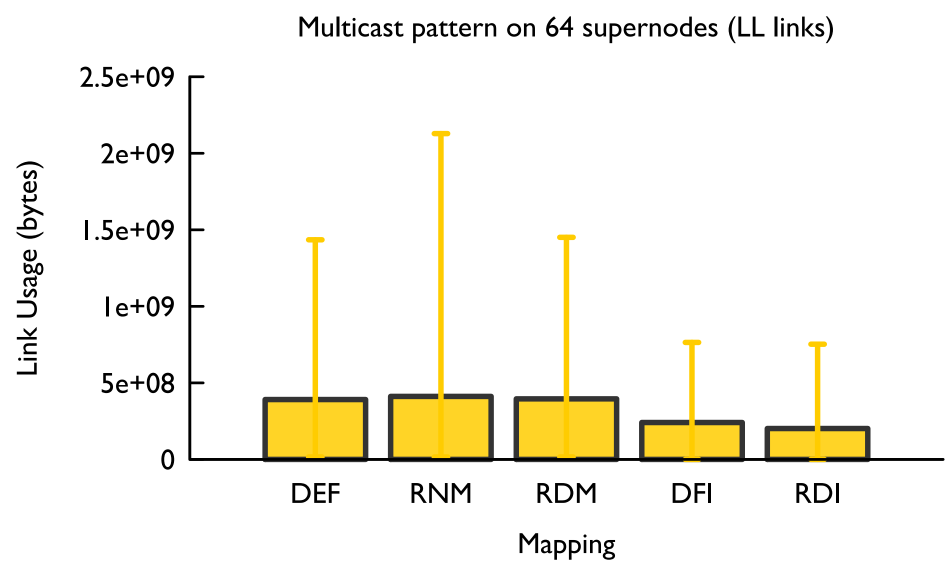
n-targets multicast pattern

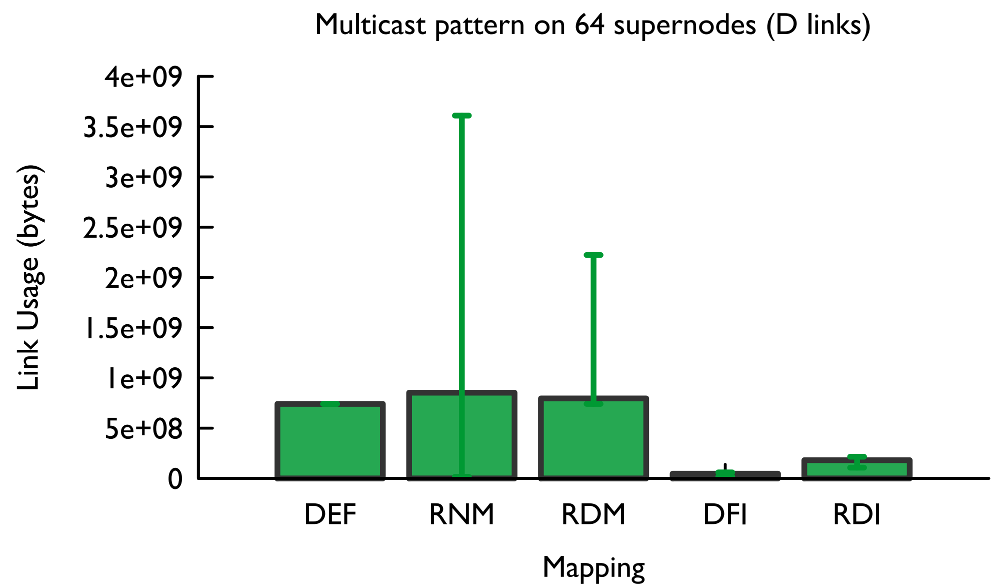


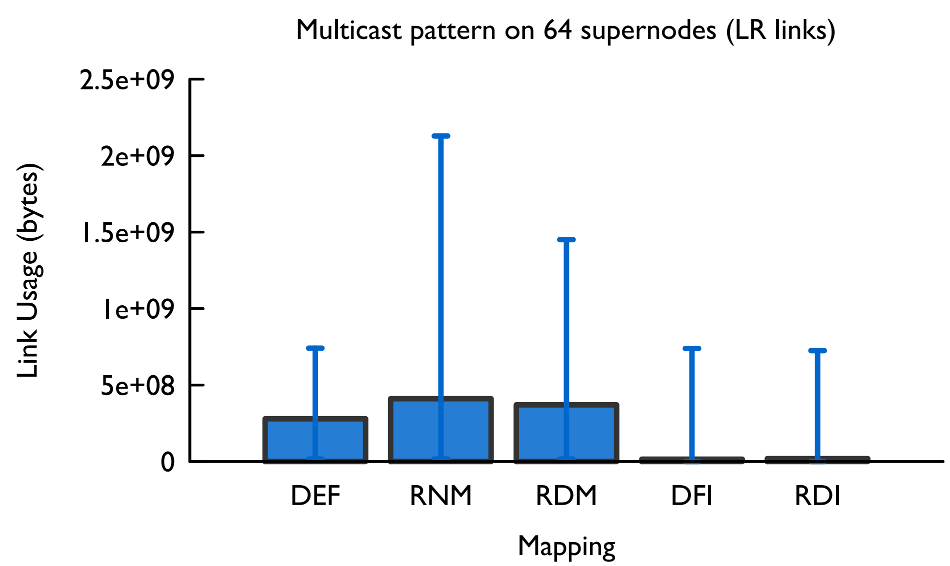


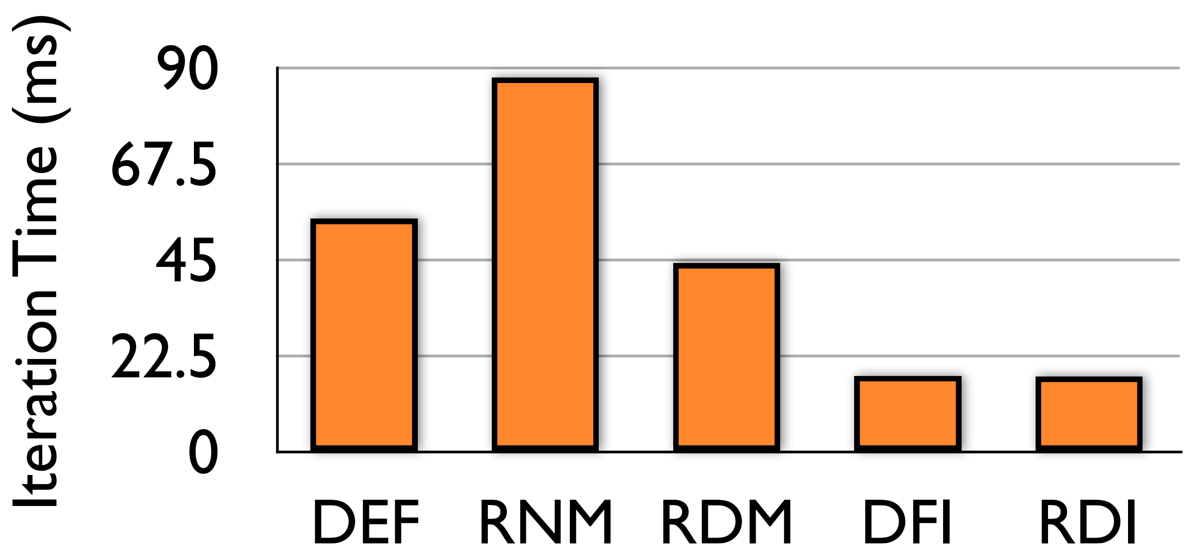


n-targets multicast pattern

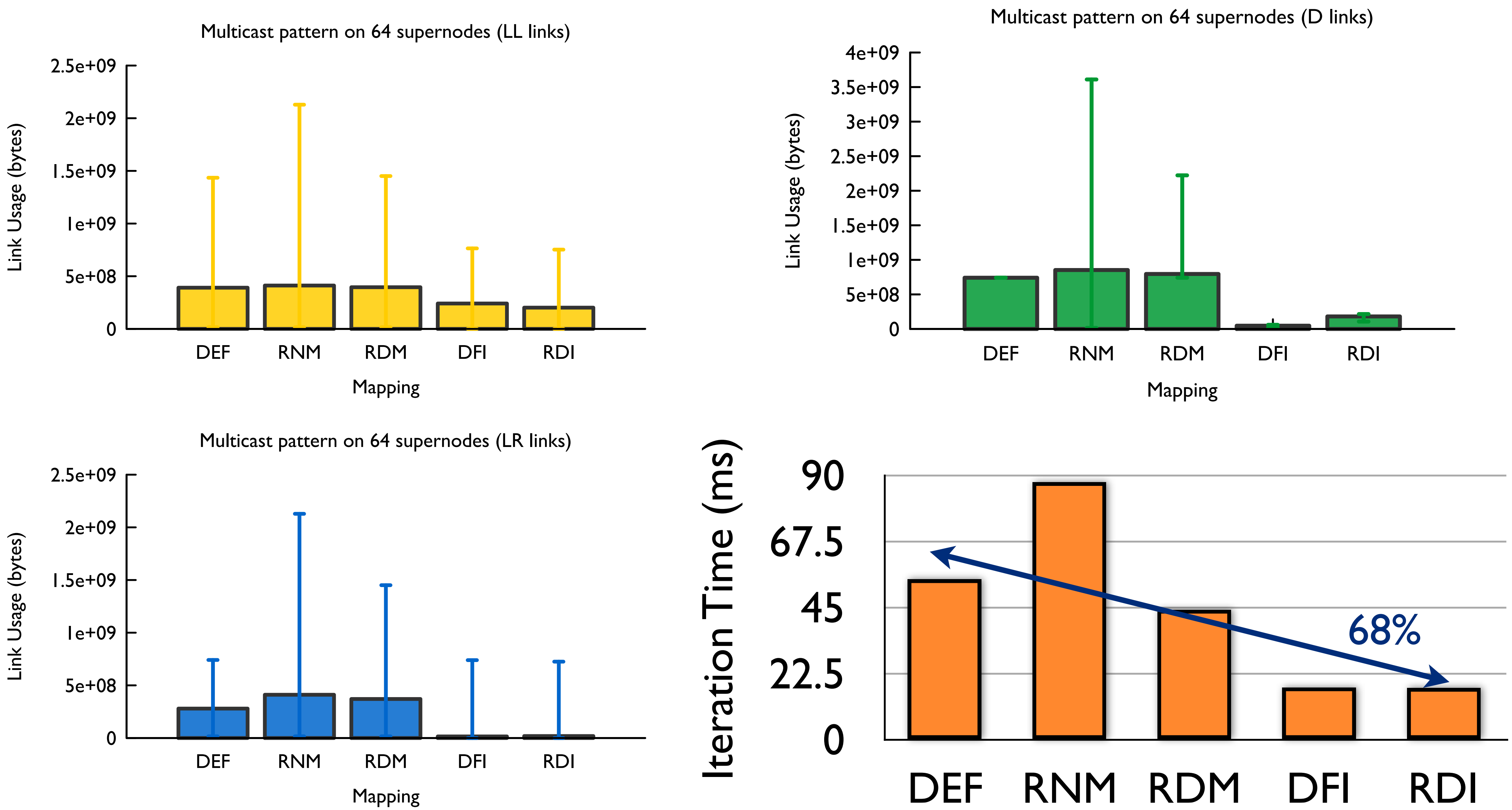




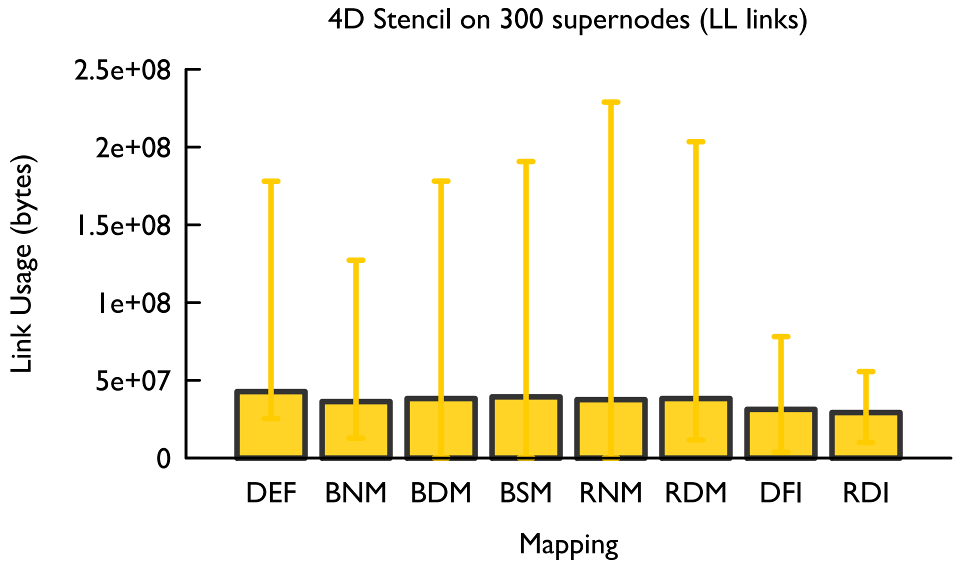


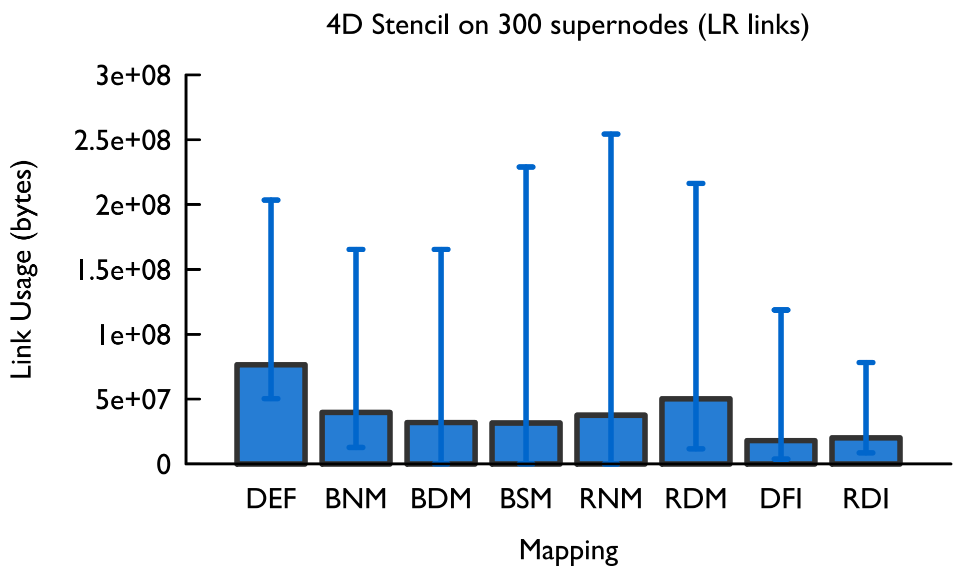


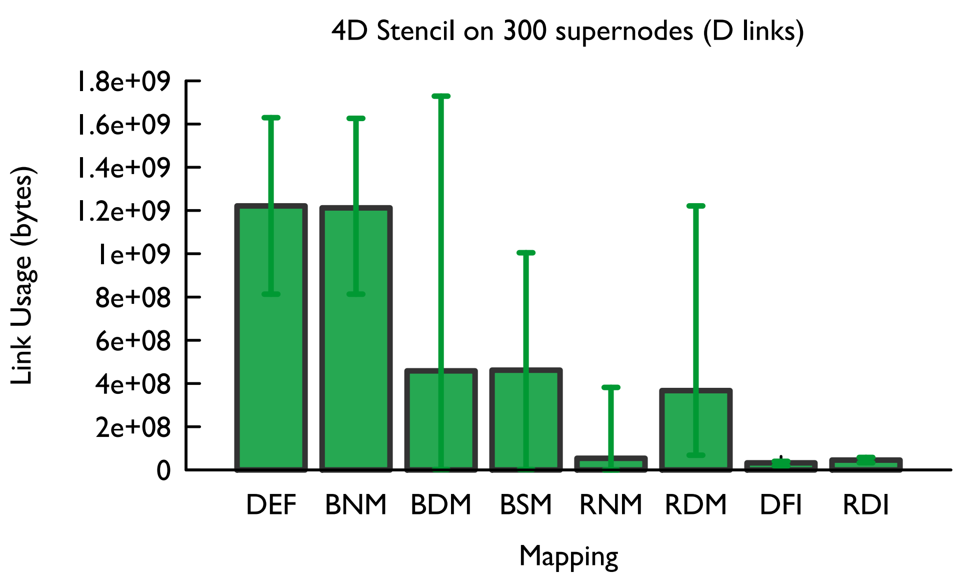
n-targets multicast pattern



4D Stencil on 300 supernodes



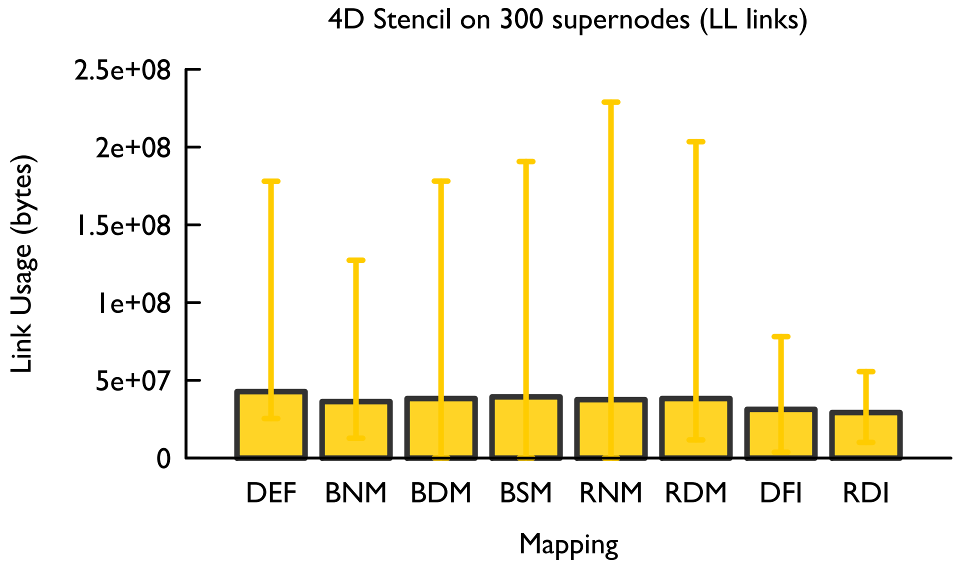


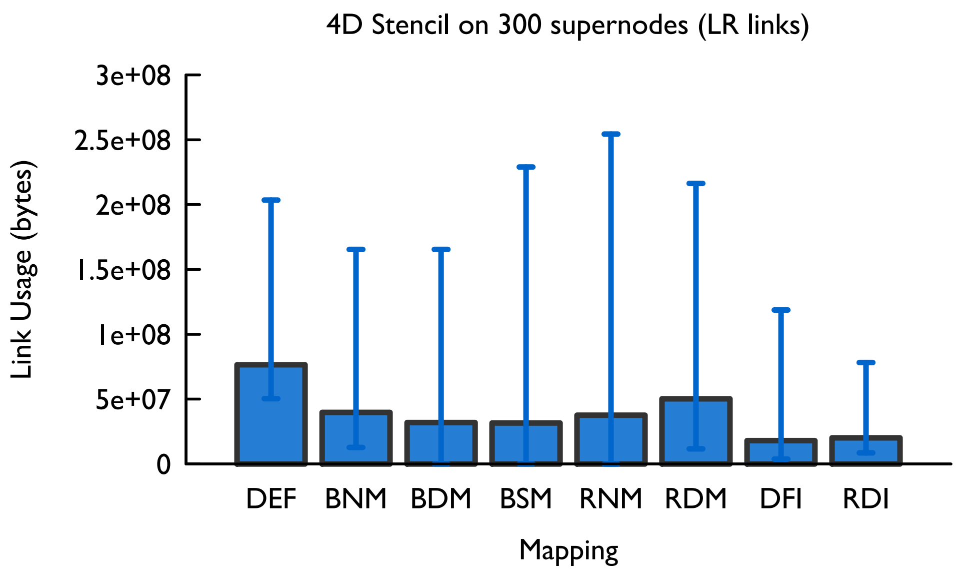


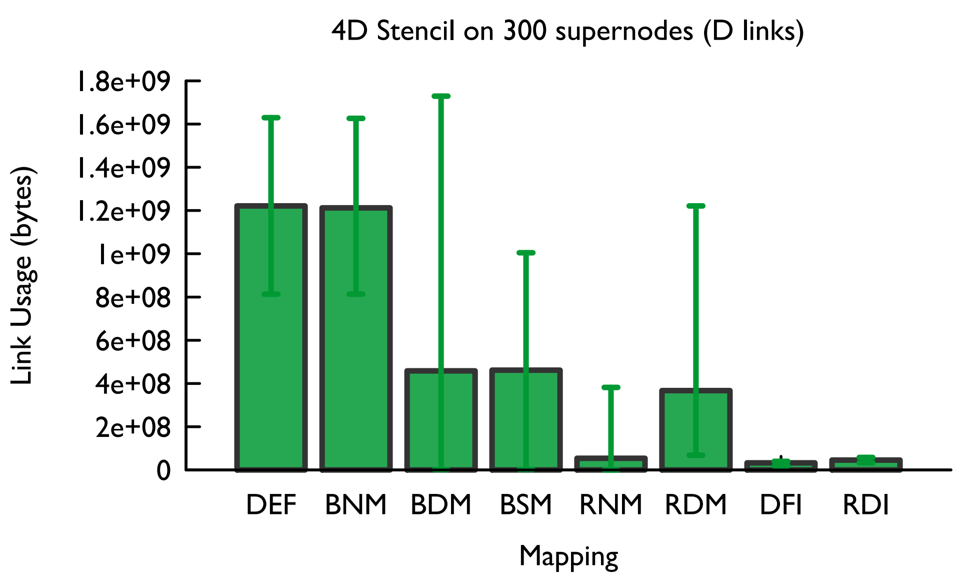
- Largest detailed network simulation so far = 307,200 MPI tasks
- Non-power-of-2 leads to more complex mapping
- Message size = 1 MB

Text

4D Stencil on 300 supernodes

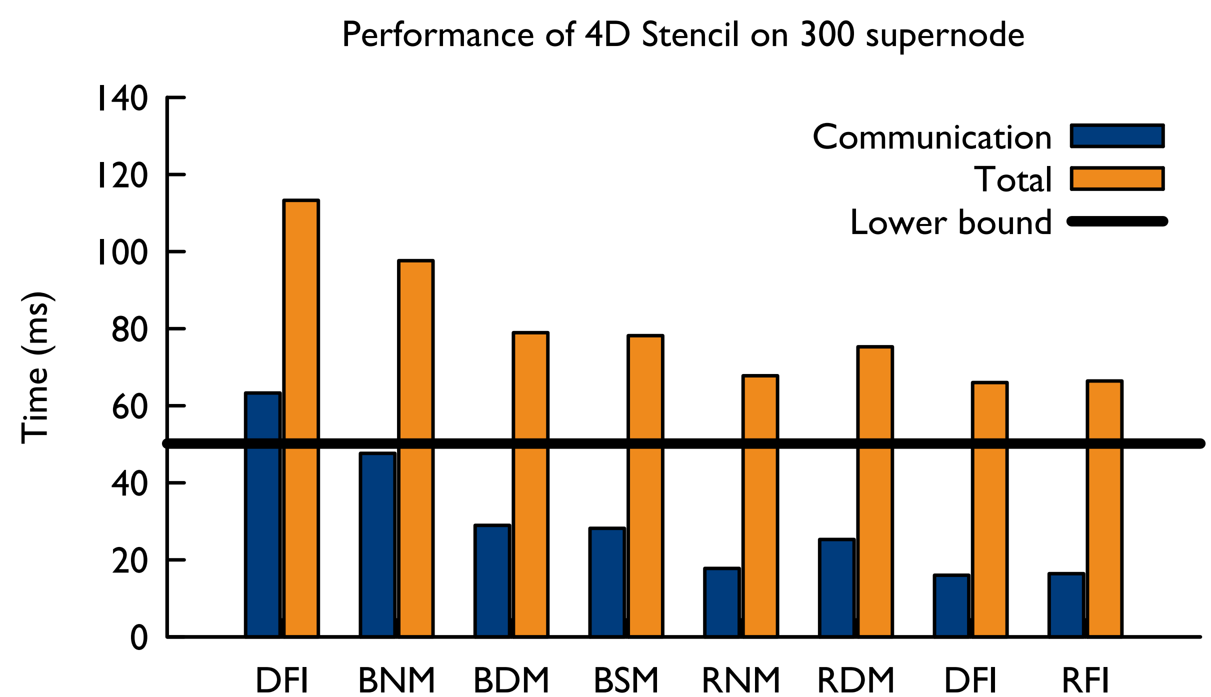




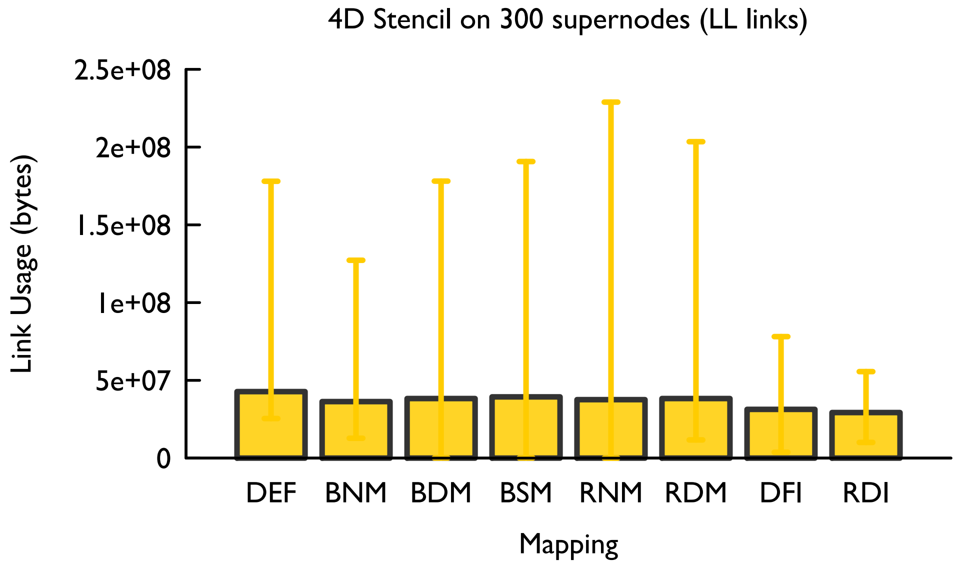


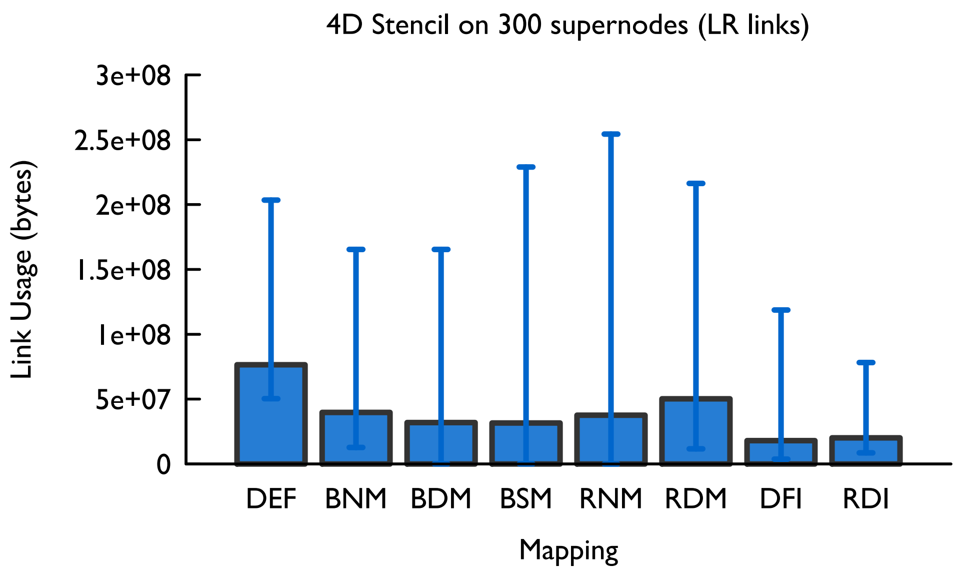
- Largest detailed network simulation so far = 307,200 MPI tasks
- Non-power-of-2 leads to more complex mapping
- Message size = 1 MB

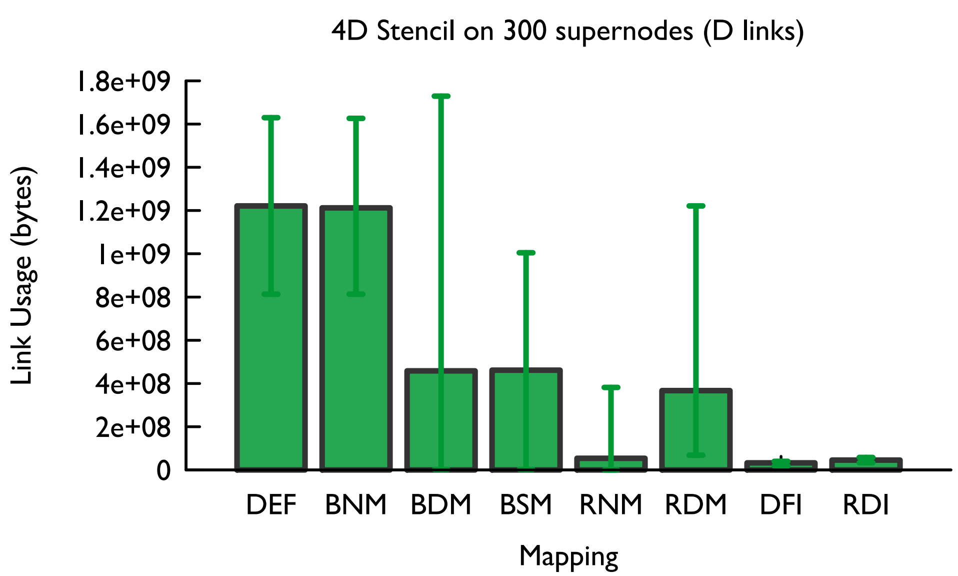
Text



4D Stencil on 300 supernodes

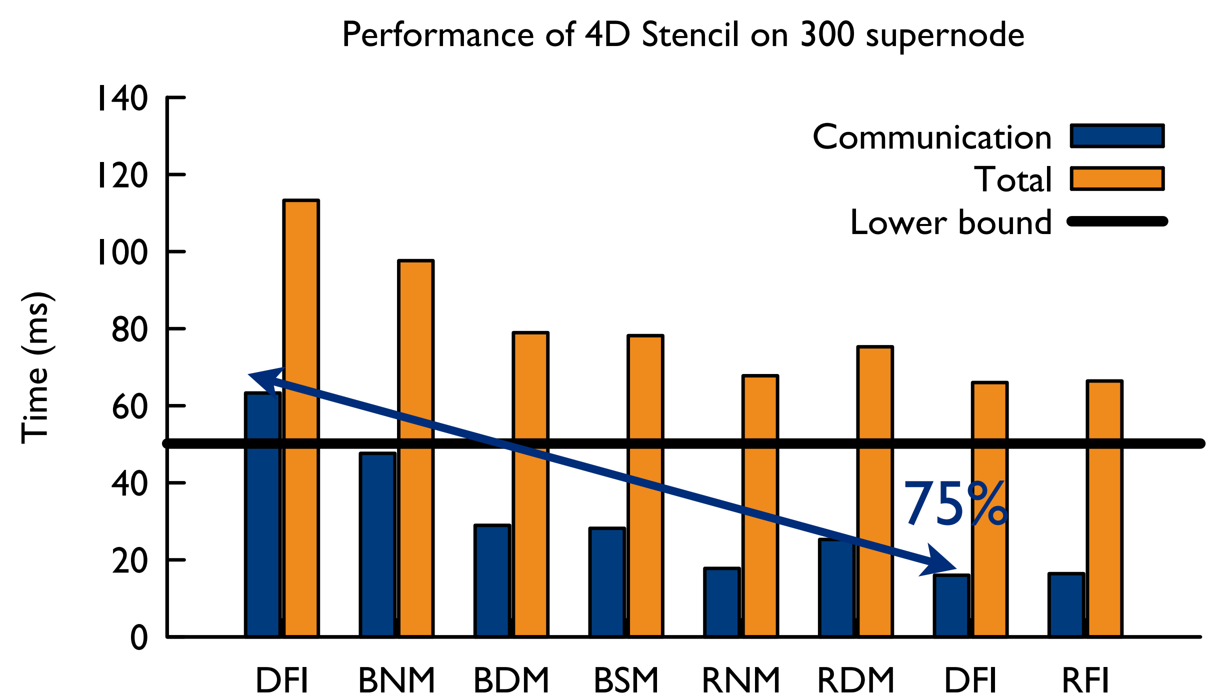




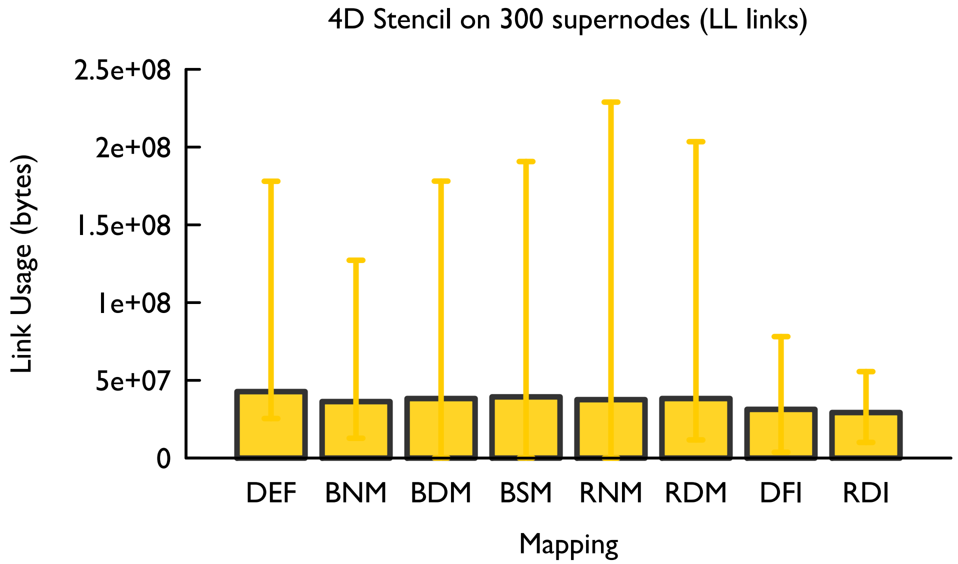


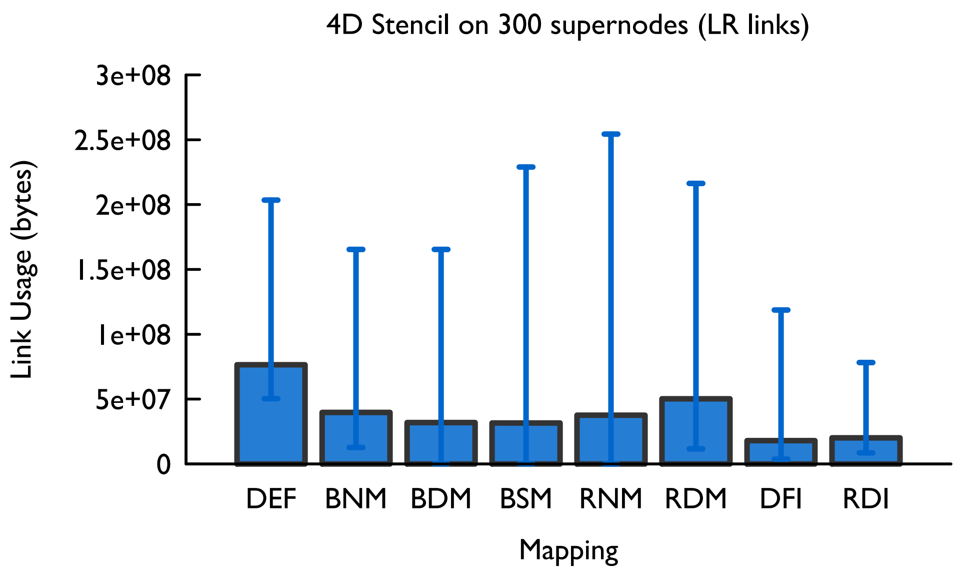
- Largest detailed network simulation so far = 307,200 MPI tasks
- Non-power-of-2 leads to more complex mapping
- Message size = 1 MB

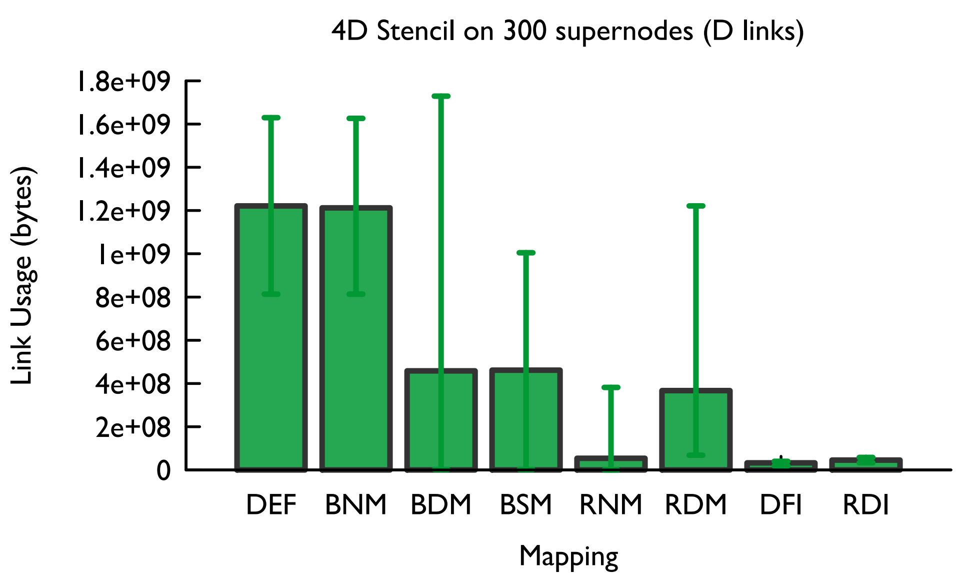
Text



4D Stencil on 300 supernodes

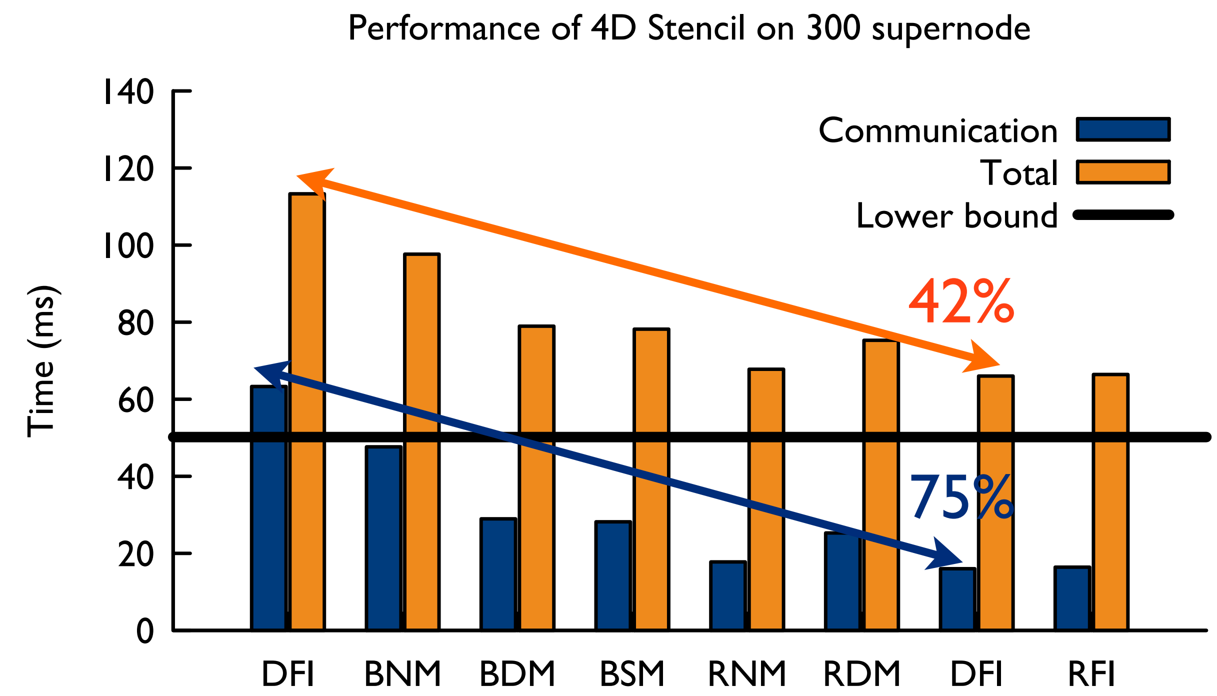






- Largest detailed network simulation so far = 307,200 MPI tasks
- Non-power-of-2 leads to more complex mapping
- Message size = 1 MB

Text



Summary

- Default MPI rank-ordered mapping on multi-level direct networks can lead to hot-spots
- Packet-level simulation to assist machine architects and application developers in making routing and mapping choices
- Conclusions:
  - With direct routing, random mapping at node granularity is best
  - With indirect routing, default mapping is good enough
- Utility of simulation-based analysis to analyze algorithms and design choices for future machines

Questions?

More information at: http://charm.cs.illinois.edu/research/topology

Thanks to Ryan Mokos (PPL) for implementing the PERCS network model.
Funding support: NSF grant OCI 07-25070, DOE grant DE-SC0001845

Supercomputing '11 ◆ November 17, 2011

LLNL-PRES-511461

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344