SAND2022-1084R

HIHE01-36: Evaluate how various topologies perform in the context of link failures

Approved for public release SAND2022-

EXASCALE COMPUTING PROJECT

K. Scott Hemmert, PI (Sandia National Laboratories*)

Joseph P. Kenny (Sandia National Laboratories*)

*Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy?s National Nuclear Security Administration under contract DE-NA0003525.

U.S. DEPARTMENT OF ENERGY | Office of Science

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Explore the effects of failed links on workloads

- Study looks at the effect that failed links have on the throughput of HPC systems
  - What workloads are most effected?
  - How many links need to be down before throughput of the machine is noticeably affected?
    - Can help facilities determine how often system time needs to be taken to replace downed links

Topologies Studied

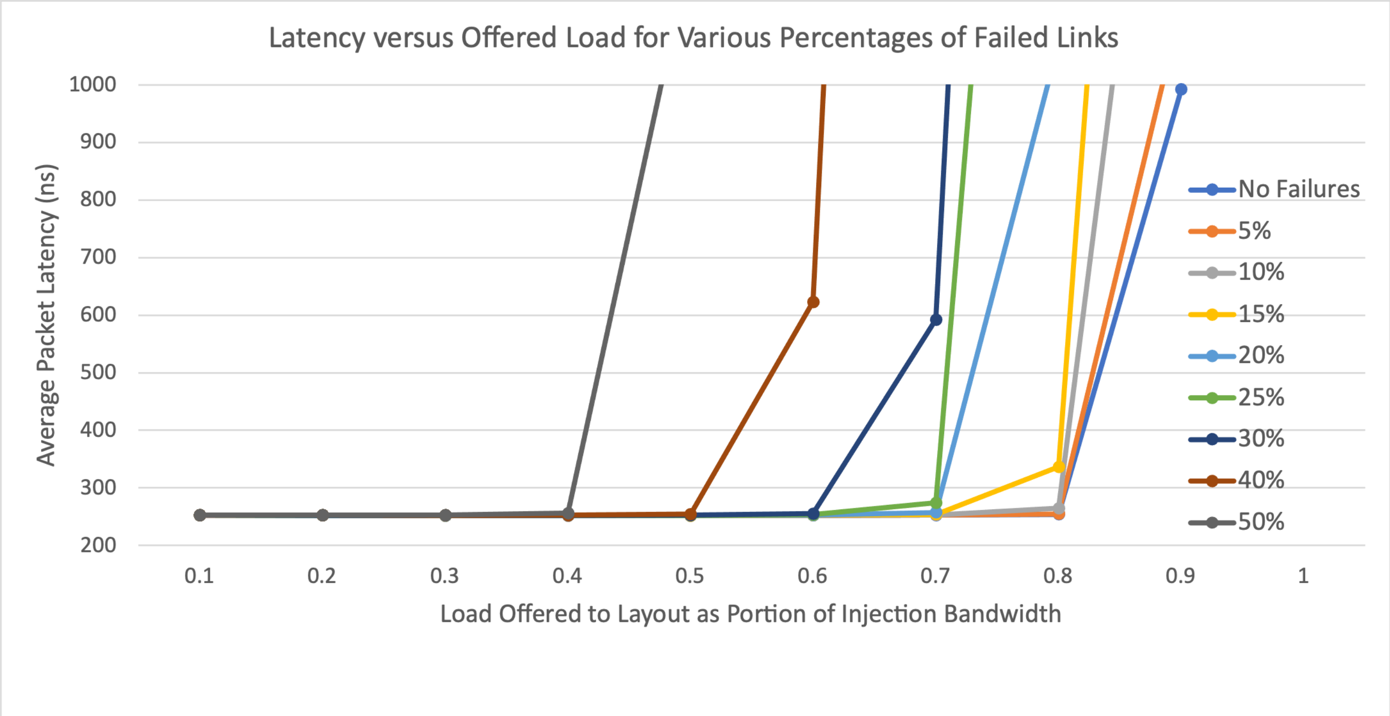
- Dragonfly
  - Only global failed links studied
    - Experience shows that the optical global links fail at a much higher rate that the electrical intragroup links
  - UGAL routing algorithm used
    - UGAL: traditional adaptive routing algorithm where each packet can select between a direct route or a valiant route (indirect route through an intermediate group) based on load on output ports of injection router
      - The valiant group is selected statistically based on the minimum number of links connecting the source and valiant groups and the valiant and destination groups. This better balances traffic to account for the location of failed links
      - This algorithm can route even when two groups are no longer connected due to failed links (only indirect routes available)
- HyperX
  - All non-host links subject to failure
    - Assuming all optical links with equal chance of failure
  - DOAL routing algorithm used
    - Packet will traverse the network in a specified dimension order, but can take one mis-route per dimension to avoid traffic

Dragonfly models successfully route in face of large numbers of failed links

Network Configuration:

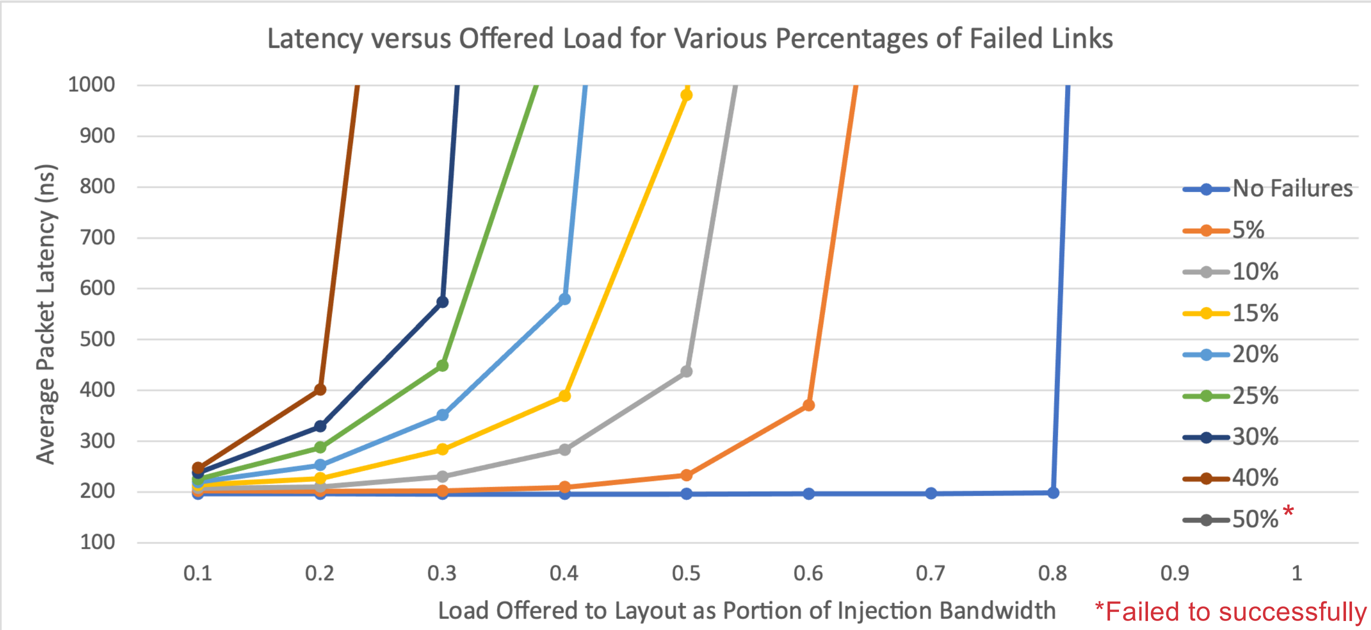
8704 Nodes
- 512 node groups
- 17 groups
- 32 links between each set of groups (maximum global bandwidth for group configuration)
- 4352 global links

544 - 64 port switches
- 400 Gbps link bandwidth
- 16 byte flit size
- 50% switch speedup (600 Gbps/port)



Test uses uniform random traffic. Results show a decrease in throughput based on number of failed links, but the fabric continues to route successfully.

HyperX routing runs into issues at higher link failure rates



*Failed to successfully route with 50% failed links.

Network Configuration:

8820 Nodes
- 21x21 hyperX
- 20 endpoints per switch

441 - 64 port switches
- 400 Gbps link bandwidth
- 16 byte flit size
- 50% switch speedup (600 Gbps/port)

Test uses uniform random traffic. Results show a decrease in throughput based on number of failed links. Routing failures occur at higher percentage of failed links because there are no 1 or 2 hop routes to align packet in a dimension. Routing constraints could be relaxed to fix this issue (not considered as part of this study).

Studies of Application Communication Patterns

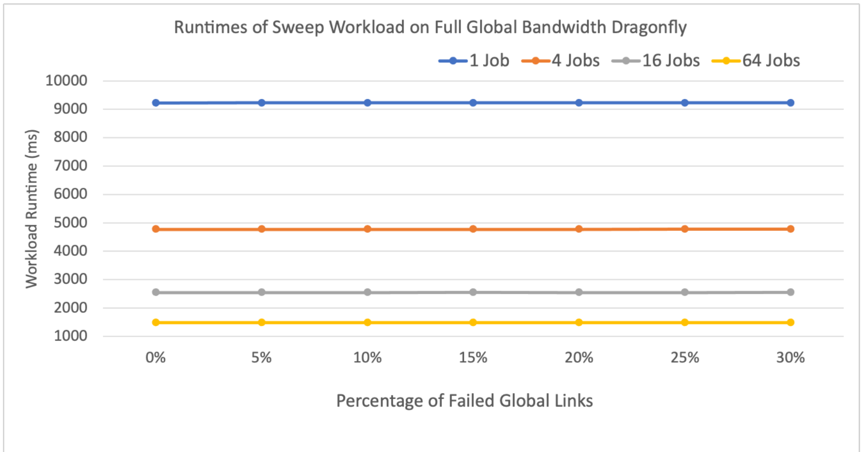
- Halo3D26
  - 27-point stencil operations, communicates with 26 neighbors
  - Uses weak scaling with a 30x30x30 problem size per rank
  - Jobs are set to the maximum size perfect cube that will fit into the available nodes. Available nodes per job is computed by dividing 32k by the number of jobs on the system.
- Sweep3D
  - Sweep communication pattern using a 2D processor decomposition, with the third dimension on-node
  - Uses weak scaling with nx=8, ny=8, nz=100, kba=10 per rank
  - Jobs are set to the maximum size perfect square that will fit into the available nodes. Available nodes per job is computed by dividing 32k by the number of jobs on the system
- LQCD
  - Motif that models the communication patterns of Lattice Quantum Chromo Dynamics
  - Jobs are based on a maximum size using parameters: nx=144, ny=144, nz=144, nt=192 at 32k MPI ranks
  - Smaller jobs are created by halving the size at each step
    - Job size halved by reducing one of nx, ny, nx, nt (in that order) by half at each size reduction
      - i.e. nx=72, ny=144, nz=144, nt=192, then nx=72, ny=72, nz=144, nt=192, etc.

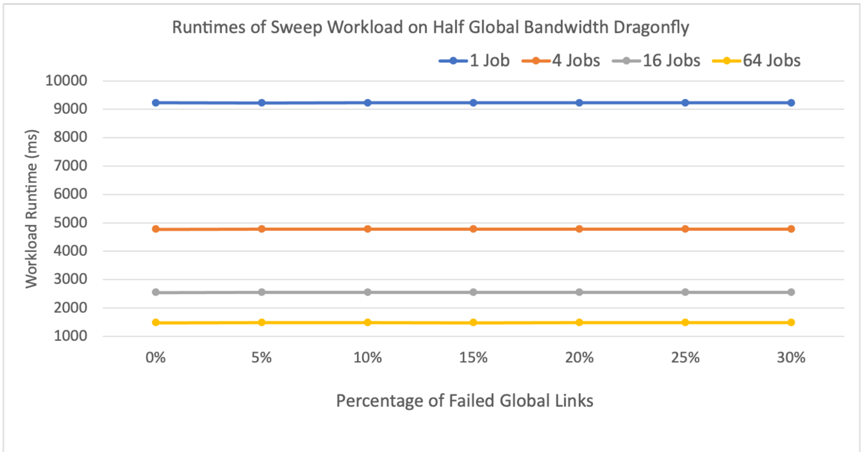
Architectures Studied

- Targeted closest “reasonable” configuration with more than 32k network endpoints
  - 4 NICs per node (8k or more nodes)
  - 4 MPI ranks per node (1 MPI rank per NIC)
  - Endpoints not needed to fit problems are allocated randomly across the network, then jobs are mapped linearly thereafter
- Dragonfly: 33,280 network endpoints
  - 512-node groups, 65 groups
  - Studied full-, half-, and quarter-global bandwidth configurations
    - 8, 4, 2 links between each group
- HyperX: Number of endpoints varies by bisection bandwidth
  - Full: 36,864 network endpoints
    - 16 endpoints per router, 16x16x9 hyperX with 2 links between each router in Z dimension, 1 link for the others
  - Half: 34,560 network endpoints
    - 20 endpoints per router, 12x12x12 hyperX with 1 link between each router in all dimensions
  - Quarter: 36,000 network endpoints
    - 36 endpoints per router, 10x10x10 hyperX with 1 link between each router in all dimensions

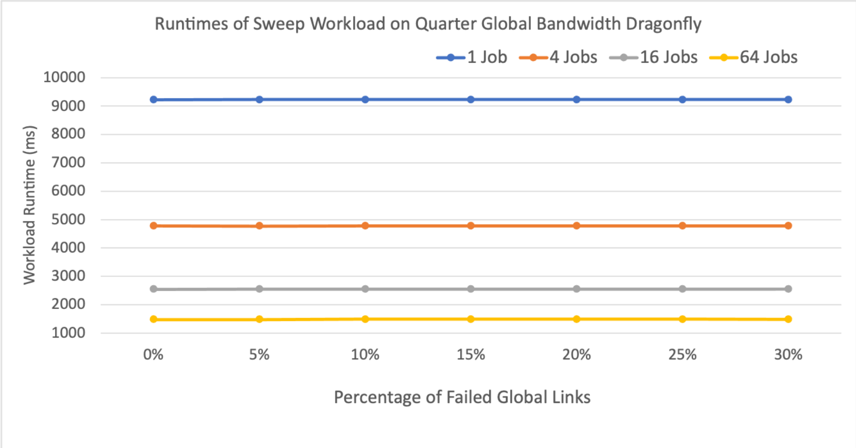
Dragonfly

Sweep3D on Dragonfly



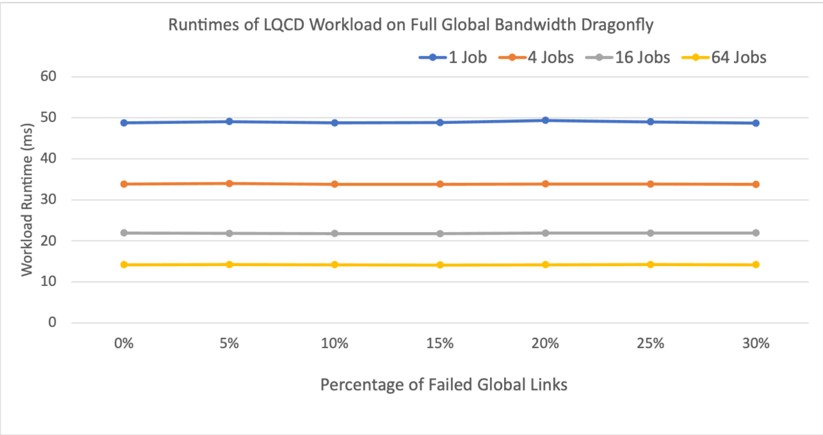


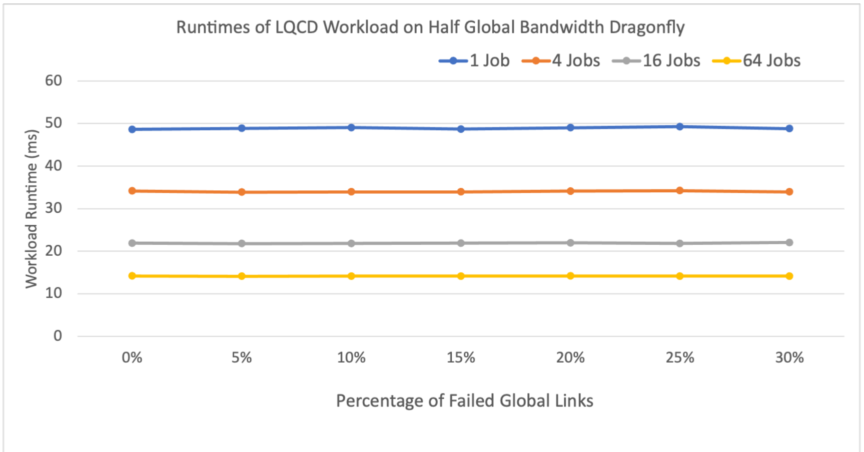
Sweep3D pattern showed no sensitivity to failed links in the cases studied.



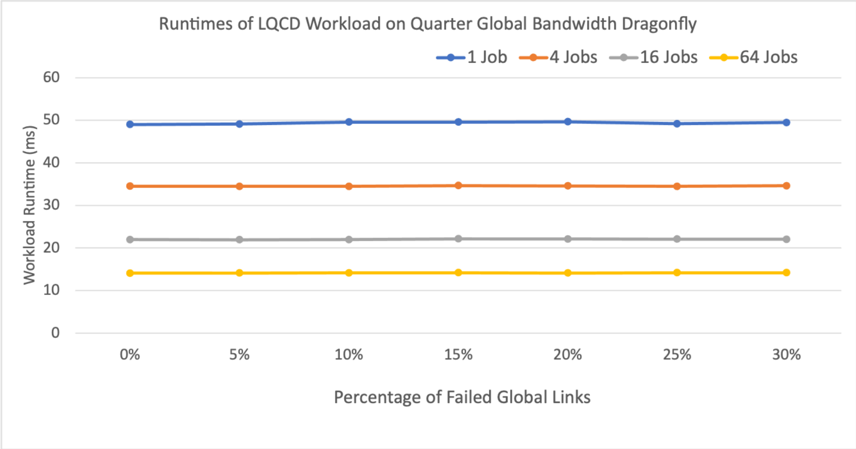
EXASCALE COMPUTING PROJECT

LQCD on Dragonfly



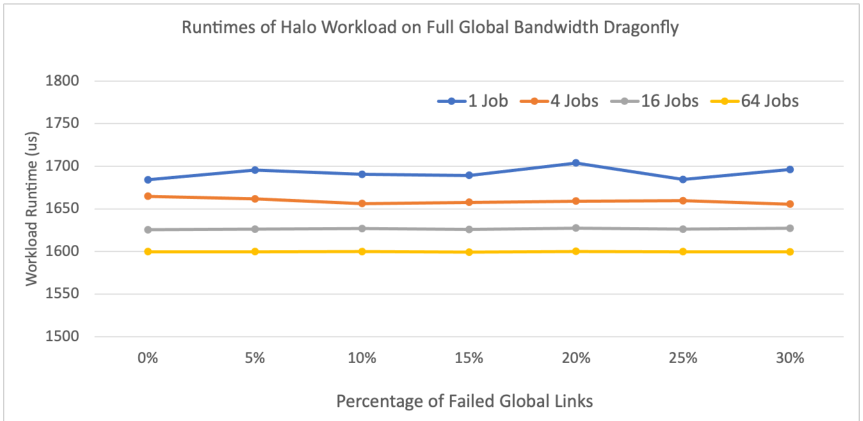


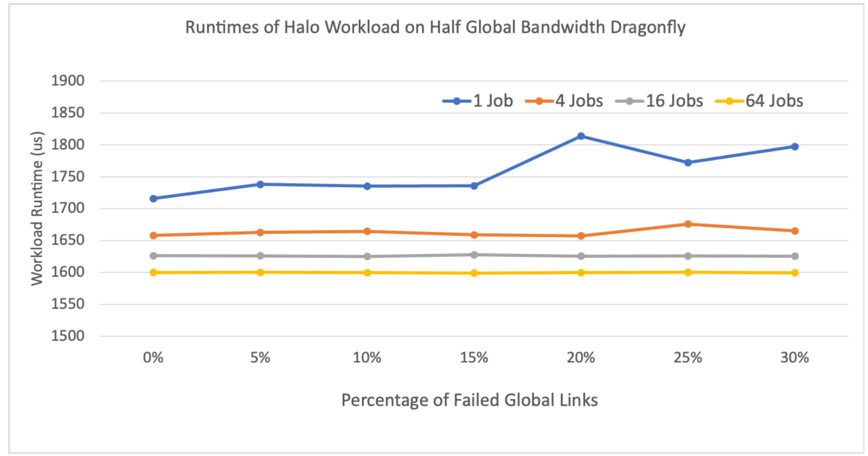
LQCD pattern showed no sensitivity to failed links in the cases studied.



EXASCALE COMPUTING PROJECT

Halo3D26 on Dragonfly – graph view

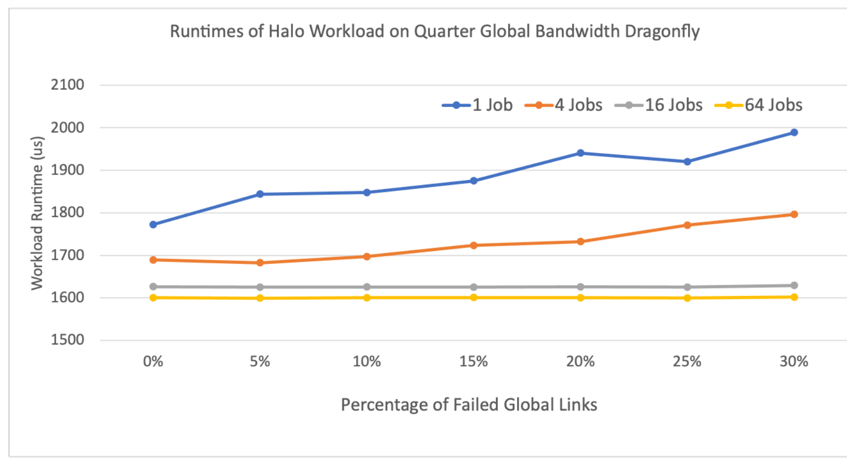




Large Halo3D jobs (1 and 4 jobs on the system) showed sensitivity to failed links

- Particularly in the case of reduced global bandwidth
  - Half – 1-2% throughput reduction at 5-15% failed links
  - Quarter – 4-6% throughput reduction at 5-15% failed links
- Typical DOE systems have had between ¼ and ½ global bandwidth

Systems that run large jobs with nearest neighbor communication patterns may need to be more aggressive at repairing failed links, depending on the system global bandwidth configuration



EXASCALE COMPUTING PROJECT

Halo3D26 on Dragonfly – table view

Performance Degradation with Full Global Bandwidth	5%	10%	15%	20%	25%	30%
1 Job	0.68%	0.38%	0.31%	1.18%	0.02%	0.72%
4 Jobs	-0.18%	-0.51%	-0.42%	-0.34%	-0.30%	-0.55%
16 Jobs	0.04%	0.09%	0.03%	0.12%	0.05%	0.11%
64 Jobs	0.00%	0.01%	-0.03%	0.03%	0.00%	0.00%

Performance Degradation with Half Global Bandwidth	5%	10%	15%	20%	25%	30%
1 Job	1.29%	1.13%	1.16%	5.70%	3.28%	4.75%
4 Jobs	0.29%	0.39%	0.06%	-0.04%	1.07%	0.43%
16 Jobs	-0.02%	-0.07%	0.09%	-0.05%	-0.04%	-0.05%
64 Jobs	0.01%	-0.02%	-0.06%	-0.01%	0.01%	-0.03%

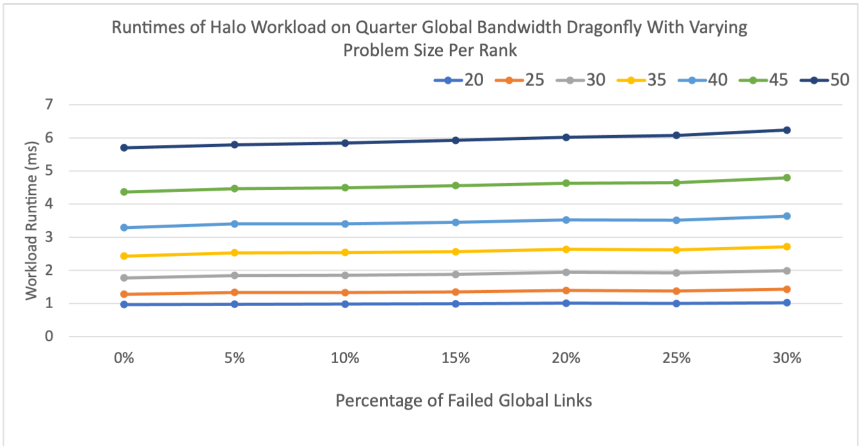
Large Halo3D jobs (1 and 4 jobs on the system) showed sensitivity to failed links

- Particularly in the case of reduced global bandwidth
  - Half – 1-2% throughput reduction at 5-15% failed links
  - Quarter – 4-6% throughput reduction at 5-15% failed links
- Typical DOE systems have had between ¼ and ½ global bandwidth

Systems that run large jobs with nearest neighbor communication patterns may need to be more aggressive at repairing failed links, depending on the system global bandwidth configuration

Performance Degradation with Quarter Global Bandwidth	5%	10%	15%	20%	25%	30%
1 Job	4.03%	4.24%	5.79%	9.48%	8.34%	12.21%
4 Jobs	-0.40%	0.45%	2.01%	2.54%	4.83%	6.33%
16 Jobs	-0.05%	-0.03%	-0.05%	-0.01%	-0.06%	0.18%
64 Jobs	-0.06%	0.01%	0.02%	0.00%	-0.04%	0.09%

Halo3D26 on Dragonfly: Sensitivity to Problem Size



Problem size affects sensitivity of Halo workload to failed links. Sensitivity peaks at mid-range problem sizes. This effect is likely due to the fact that compute time increases with the cube of the problem size and communication volume only increases with the square of problem size. Thus, communication becomes a smaller percentage of total iteration time as the problem size increases, which overcomes the degradation in absolute communication times at large problem sizes.

Performance Degradation with Simulated Compute Time

	5%	10%	15%	20%	25%	30%
20	1.19%	1.28%	2.22%	4.33%	3.31%	5.76%
25	4.14%	3.82%	5.08%	8.91%	7.47%	11.67%
30	4.03%	4.24%	5.79%	9.48%	8.34%	12.21%
35	4.17%	4.48%	5.57%	8.46%	7.68%	11.77%
40	3.39%	3.42%	4.80%	7.03%	6.78%	10.54%
45	2.25%	2.89%	4.39%	6.04%	6.34%	9.85%
50	1.58%	2.45%	3.95%	5.49%	6.60%	9.36%

Performance Degradation without Simulated Compute Time

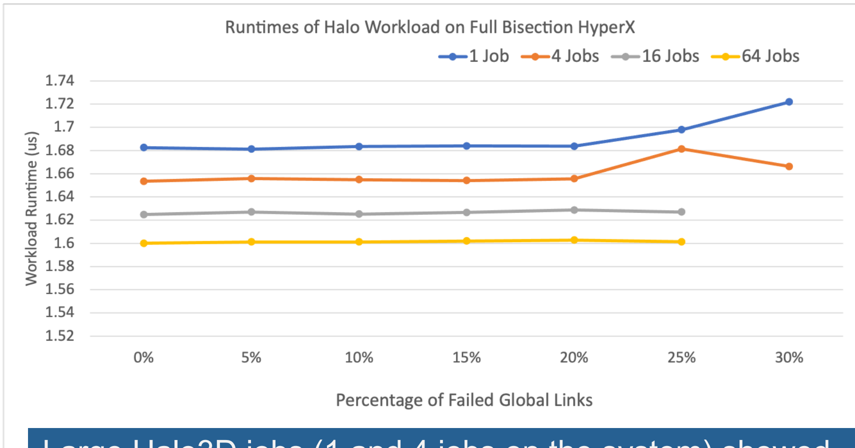
	5%	10%	15%	20%	25%	30%
20	1.68%	1.80%	3.12%	6.07%	4.64%	8.07%
25	6.25%	6.18%	8.01%	13.83%	12.12%	18.84%
30	8.53%	8.99%	12.27%	20.10%	17.68%	25.87%
35	10.76%	11.53%	14.34%	21.79%	19.79%	30.32%
40	10.20%	10.72%	15.70%	23.13%	21.25%	33.33%
45	8.12%	10.43%	15.86%	21.82%	22.91%	35.59%
50	6.57%	10.19%	16.44%	22.89%	27.50%	39.02%

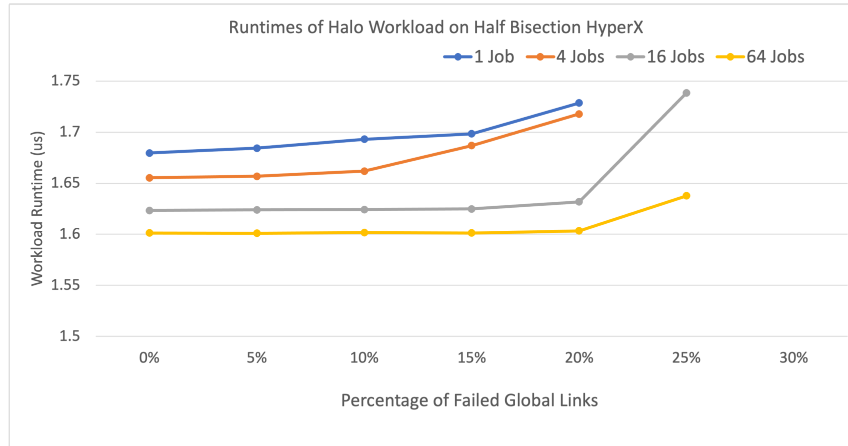
Tables show the performance degradation caused by failed links. Cases are shown for both complete iteration times, as well as the case where the compute time is not accounted for.

HyperX

*Sweep and LQCD patterns showed the same lack of sensitivity to failed links as under dragonfly; halo results are shown

Halo3D26 on HyperX – graph view

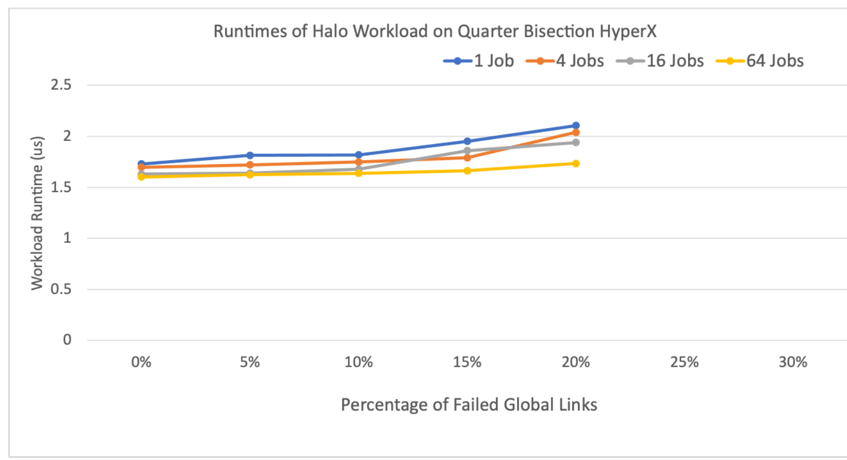




Large Halo3D jobs (1 and 4 jobs on the system) showed sensitivity to failed links

- Particularly in the case of reduced global bandwidth
  - Half – 1-4% throughput reduction at 5-15% failed links
  - Quarter – 5-15% throughput reduction at 5-15% failed links, even for smaller jobs
- No current or past DOE systems using hyperX, but would anticipate full to half bisection systems

Systems that run large jobs with nearest neighbor communication patterns may need to be more aggressive at repairing failed links, depending on the system bisection bandwidth configuration



*Missing points are due to fabric not being able to fully route with that many failed links

Halo3D26 on HyperX – table view

Performance Degradation with Full Global Bandwidth						
	5%	10%	15%	20%	25%	30%
1 Job	-0.08%	0.05%	0.08%	0.07%	0.91%	2.33%
4 Jobs	0.14%	0.08%	0.03%	0.13%	1.69%	0.77%
16 Jobs	0.13%	0.02%	0.12%	0.24%	0.14%	
64 Jobs	0.07%	0.07%	0.12%	0.17%	0.08%	

Performance Degradation with Half Global Bandwidth						
	5%	10%	15%	20%	25%	30%
1 Job	0.27%	0.79%	1.12%	2.91%		
4 Jobs	0.09%	0.39%	1.91%	3.78%		
16 Jobs	0.03%	0.05%	0.09%	0.52%	7.10%	
64 Jobs	-0.01%	0.03%	0.00%	0.13%	2.28%	

Large Halo3D jobs (1 and 4 jobs on the system) showed sensitivity to failed links

- Particularly in the case of reduced global bandwidth
  - Half – 1-4% throughput reduction at 5-15% failed links
  - Quarter – 5-15% throughput reduction at 5-15% failed links, even for smaller jobs
- No current or past DOE systems using hyperX, but would anticipate full to half bisection systems

Systems that run large jobs with nearest neighbor communication patterns may need to be more aggressive at repairing failed links, depending on the system bisection bandwidth configuration

Performance Degradation with Quarter Global Bandwidth						
	5%	10%	15%	20%	25%	30%
1 Job	4.91%	5.14%	12.86%	21.75%		
4 Jobs	1.34%	3.03%	5.49%	20.10%		
16 Jobs	0.46%	2.85%	14.03%	18.95%		
64 Jobs	1.41%	2.22%	3.84%	8.28%		

*Missing entries are due to fabric not being able to fully route with that many failed links

Conclusions

- Of the three communication patterns studied, only one (27-point stencil) showed sensitivity to failed links.
- HyperX topology more sensitive to failed links than dragonfly (NOTE: see caveats below)
  - For higher link failure rates, valid routes within the constraints of the routing algorithm could not be found
    - These results are likely somewhat pessimistic, as a routing table based approach should be able to do a better job of steering packets away from the troubled areas. The routing algorithm used in the study only looked at routing around failures in the current dimension.
- Results suggest that systems can withstand at least 5% link failure before the throughput of the machine is noticeably affected, even for the sensitive communication pattern.
  - Larger jobs more sensitive to link failures than smaller jobs. Systems with only smaller jobs may be able to get away with more failed links.
  - Lower global/bisection bandwidth will increase sensitivity to failed links
  - Still need to evaluate throughput in presence of many-to-one communication patterns and other traffic that causes network congestion even in the best case