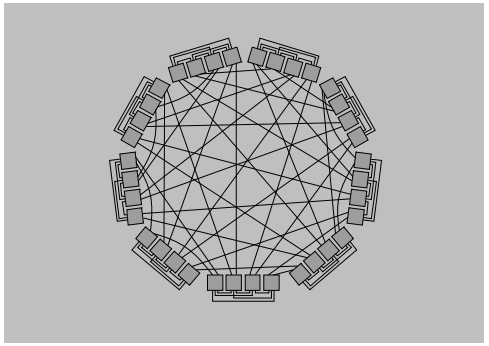
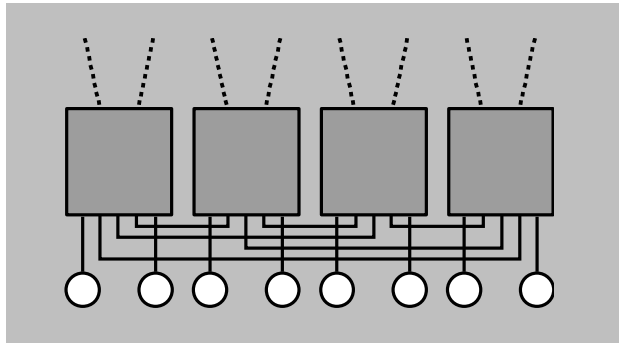
SAND2017-7773C



# Unveiling the Interplay Between Global Link Arrangements and Network Management Algorithms on Dragonfly Networks

Fulya Kaplan[†], Ozan Tuncer[†], **Vitus J. Leung**[‡], K. Scott Hemmert[‡], and Ayse K. Coskun[†]

[†] Boston University, Boston, MA [‡] Sandia National Laboratories, Albuquerque, NM

U.S. DEPARTMENT OF ENERGY | NNSA National Nuclear Security Administration

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and 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.

# Motivation

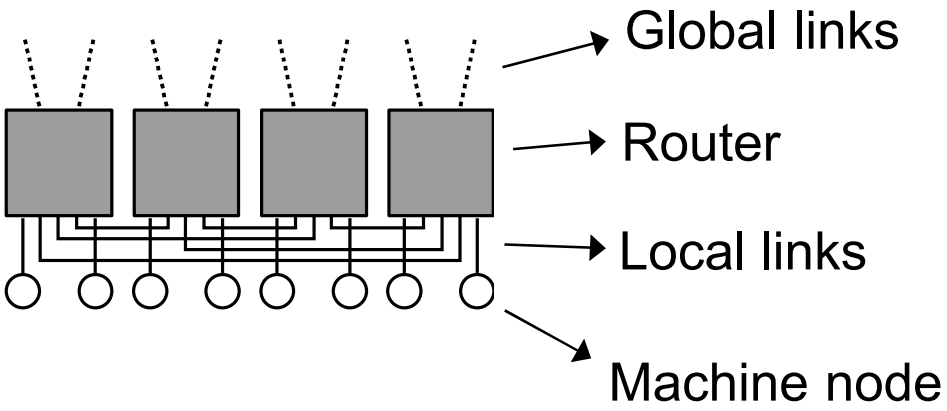
Sandia National Laboratories

- High performance computing (HPC) jobs run for hours, days
- Communication bottlenecks increase the job execution time by 50% to 1000% [*Prisacari et al., HPDC'14]*
- Various runtime techniques have been proposed to reduce communication time
  - Job placement
  - Message routing
  - Low-diameter network topologies
    - Cray BlackWidow *[Abts et al., SC'07]*
    - Flattened butterfly *[Kim et al., ISCA'07]*
    - Slim fly *[Besta et al., SC'14]*
    - Dragonfly *[Kim et al., ISCA'07]*
      - Real system implementations: Cray Cascade, IBM Percs

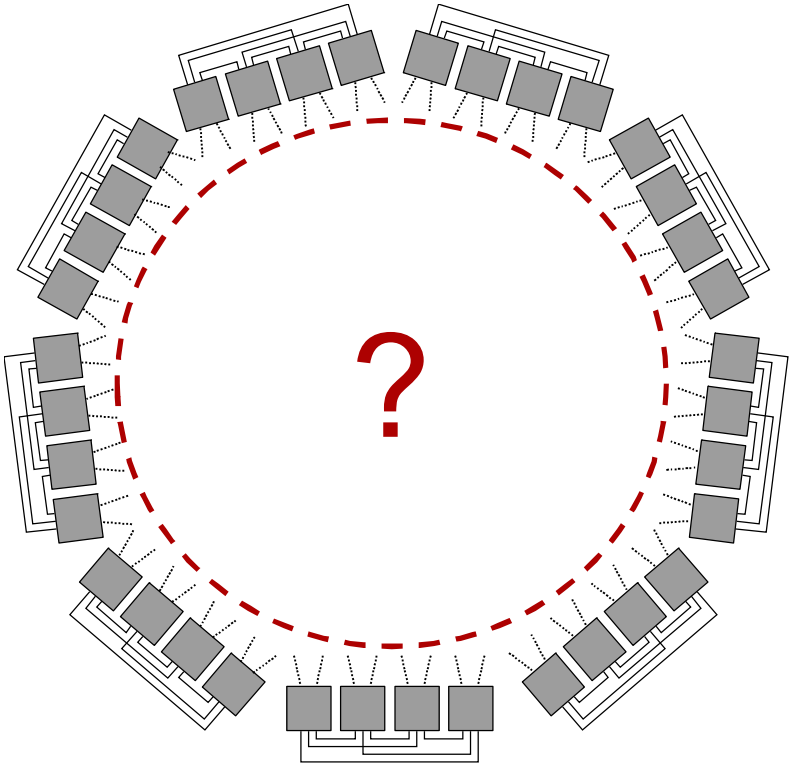
# Motivation – Dragonfly Topology

Dragonfly consists of groups of routers that are closely-connected

A dragonfly group



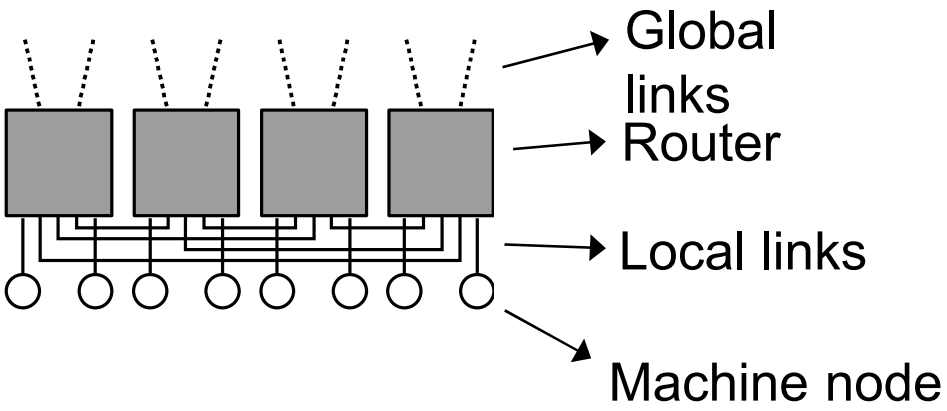

A dragonfly machine



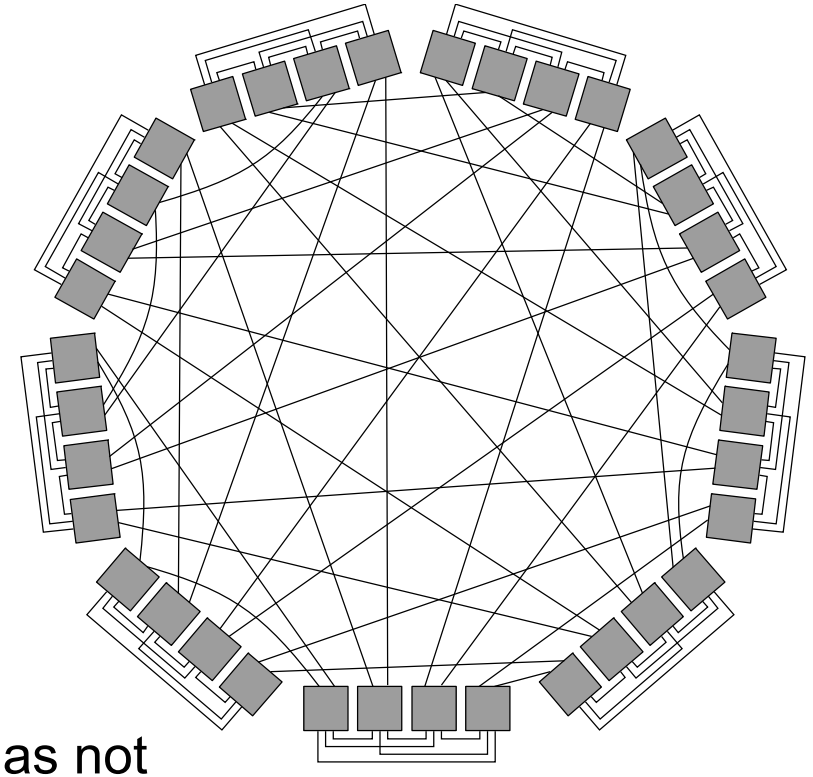
# Motivation – Dragonfly Topology

Dragonfly consists of groups of routers that are closely-connected

A dragonfly group




A dragonfly machine



- Network bisection bandwidth changes by up to 33% based on the global link arrangement
  *[Hastings et al., CLUSTER'15]*
- The impact of global link arrangement has not been studied extensively in tandem with runtime management techniques

# Contributions

Sandia National Laboratories

- We introduce a packet-level simulation framework that considers the coupling between:
  - Global link arrangements
  - Link bandwidths
  - Job placement
  - Application communication patterns
  - Routing algorithms

- We show that the “circulant-based” global link arrangement provides up to 15% lower communication overhead

- We demonstrate that the combined impact of job placement and routing highly depends on network settings and communication patterns (we show up to a 44% difference)

# Outline

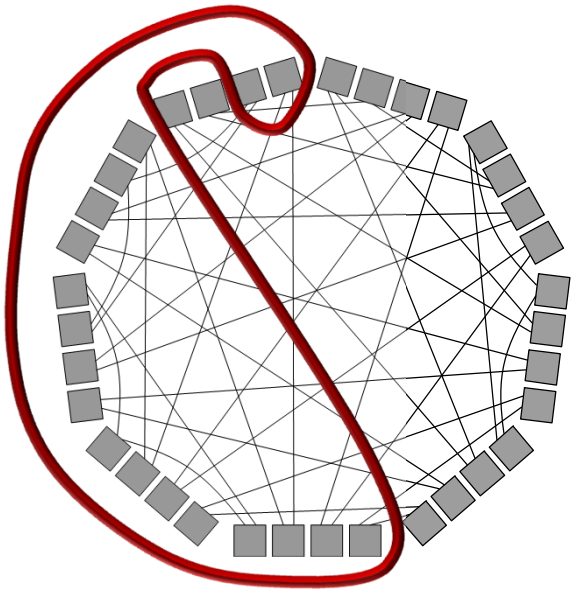
- Global link arrangements background
- Related work
- Simulation framework
- Experimental methodology
- Results
- Conclusion

# Global Link Arrangements

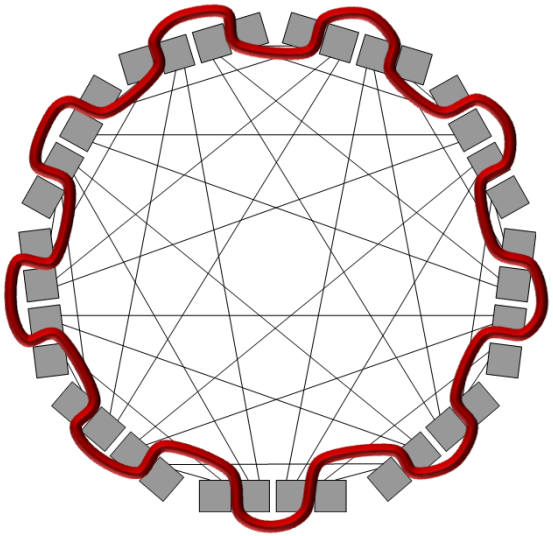
Sandia National Laboratories

Three arrangements proposed by Camarero et al. in 2014

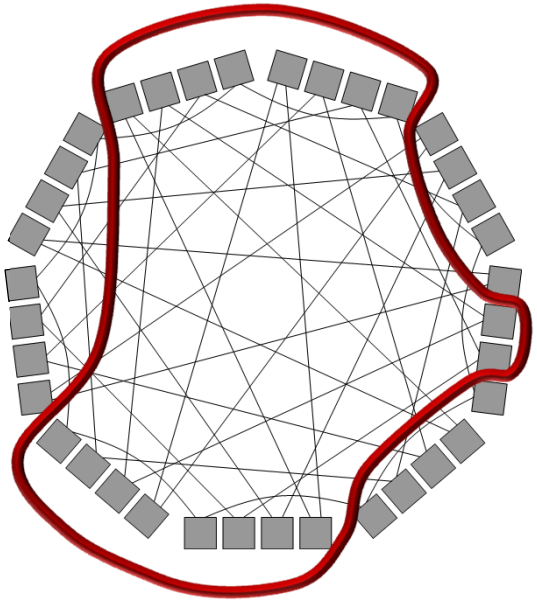
Absolute



Relative



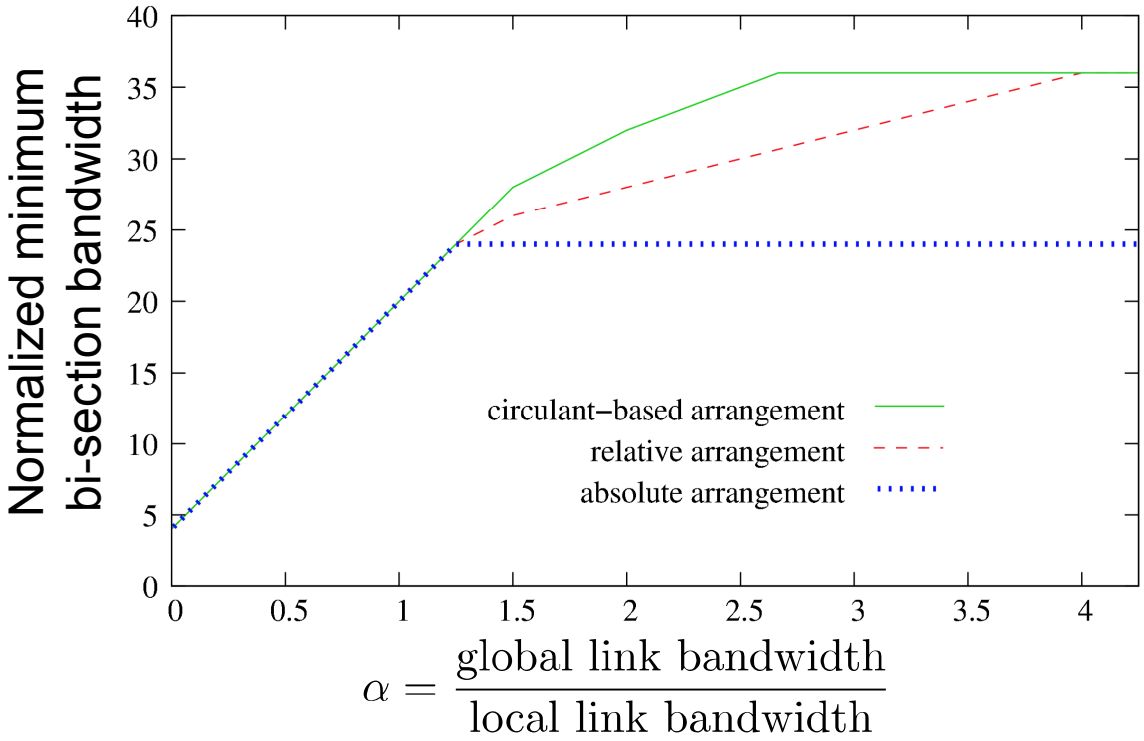
Circulant-based



The arrangements differ by their
theoretical minimum bisection bandwidth,
a commonly-used measure of the potential network bottlenecks

# Global Link Arrangements

Theoretical minimum bisection bandwidth *[Hastings et al., Cluster'15]*




What does the minimum bandwidth difference correspond to in terms of communication time, in which condition?

# Related Work on Reducing Communication Overhead on Dragonflies

- Focusing on a specific global link arrangement
  - Routing and job placement
    *[Bhatele et al., SC'11], [Chakavarty et al., HiPC'12], [Prisacari et al., HPDC'14]*
  - Global link bandwidth
    *[Bhatele et al., Tech. Report'15], [Groves et al., Cluster'16]*
- Routing and message passing for specific application communication patterns
  *[Fuentes et al., Cluster'15], [Prisacari et al., IPDPS'13]*
- Theoretical analysis of global link arrangements
  *[Hastings et al., Cluster'15], [Belka et al., HiPINEB'17]*

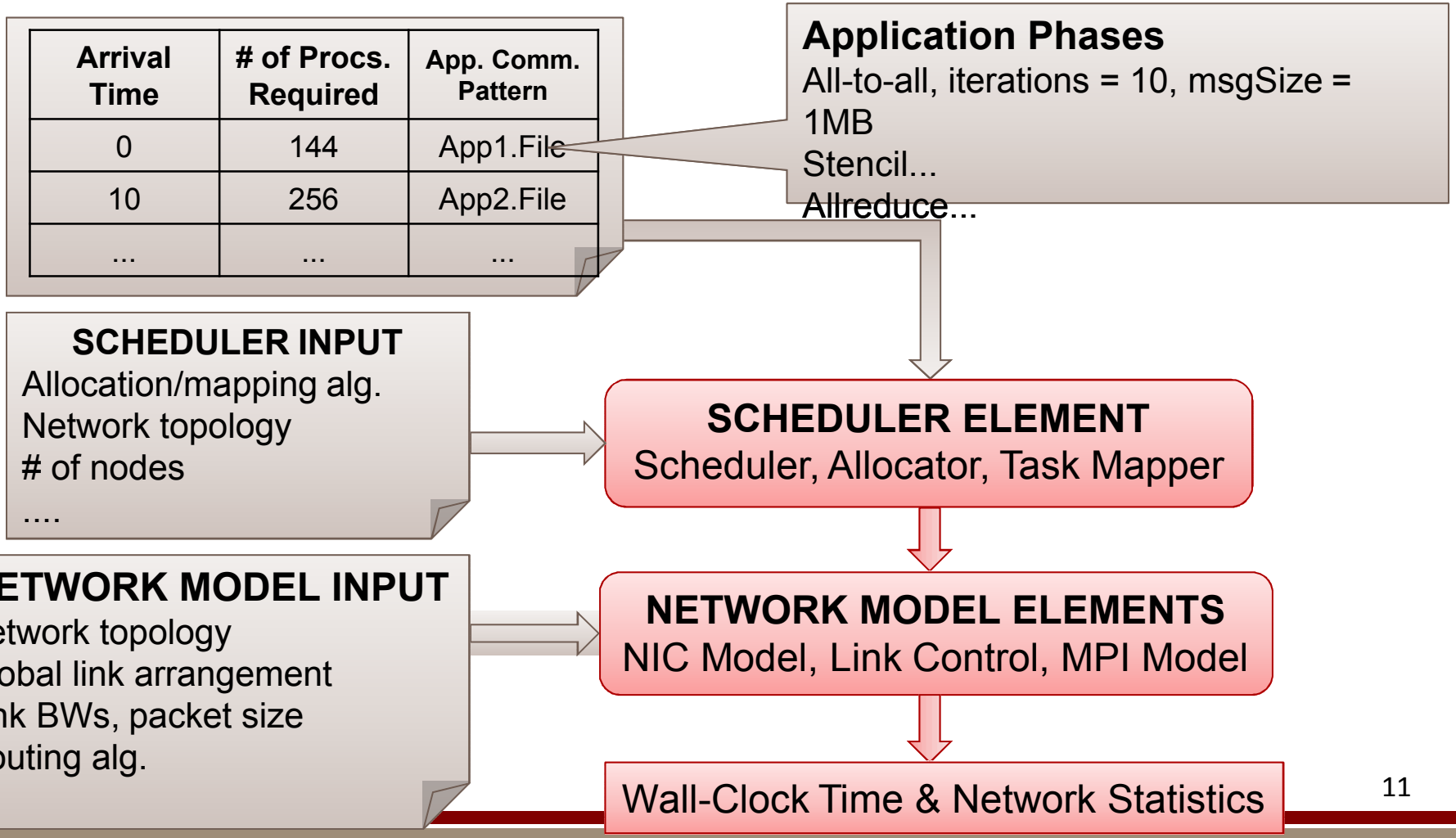
Existing works do not compare global link arrangements considering various communication patterns, routing and job placement algorithms, and link bandwidths

# Simulation Framework

- Packet-level simulation of HPC applications considering
  - Job placement
  - Application communication patterns
  - Routing algorithms
  - Network topology
  - Global/local link bandwidths
- Based on Structural Simulation Toolkit (SST)
  *[Rodrigues et al., SIMUTOOLS'12]*
  - Scheduler: Job placement
  - Ember: MPI messaging
  - Merlin: Router

# Simulation Framework

- Based on Structural Simulation Toolkit (SST)
- Available at sst-simulator.org



# Experimental Methodology

Sandia National Laboratories

- Target machines

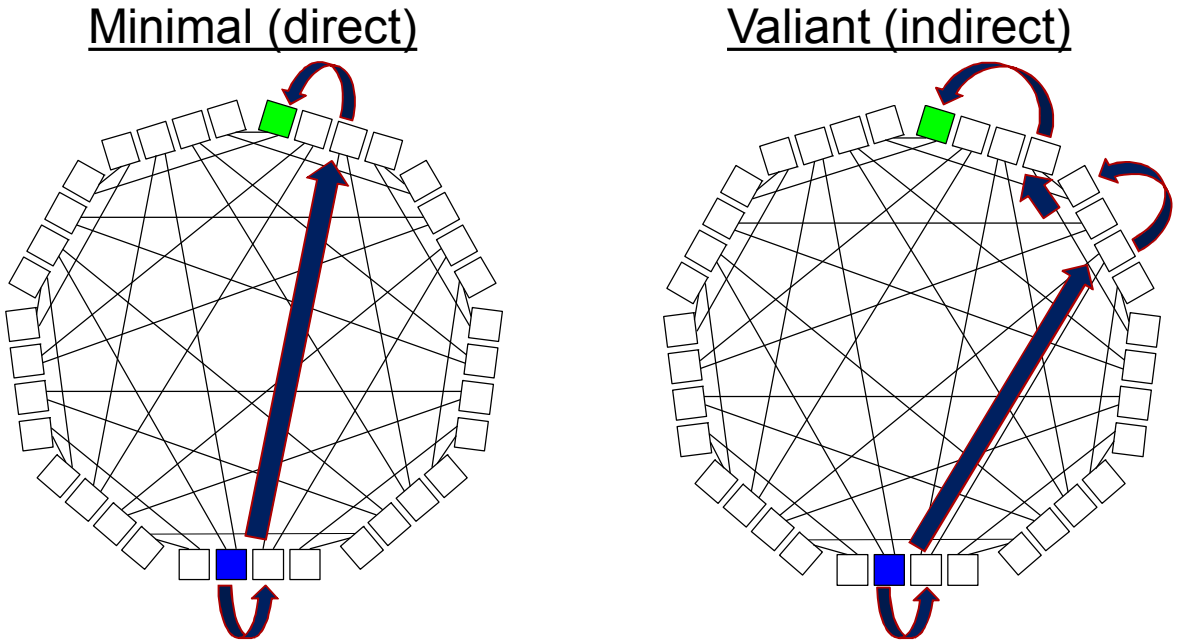
| Machine | # of groups | # of routers per group | # of nodes per router | # of cores per node | Total # of cores |
|---|---|---|---|---|---|
| small | 9 | 4 | 2 | 2 | 144 |
| medium | 17 | 8 | 2 | 2 | 544 |
| large | 17 | 8 | 4 | 2 | 2176 |

  - # of groups is determined by # of nodes and routers per group
  - All-to-all intra-group network
- Link bandwidths
  - Local: 1GB/s
  - Global: 0.5, 1, 1.5, …, 4GB/s

# Experimental Methodology

- Routing algorithms

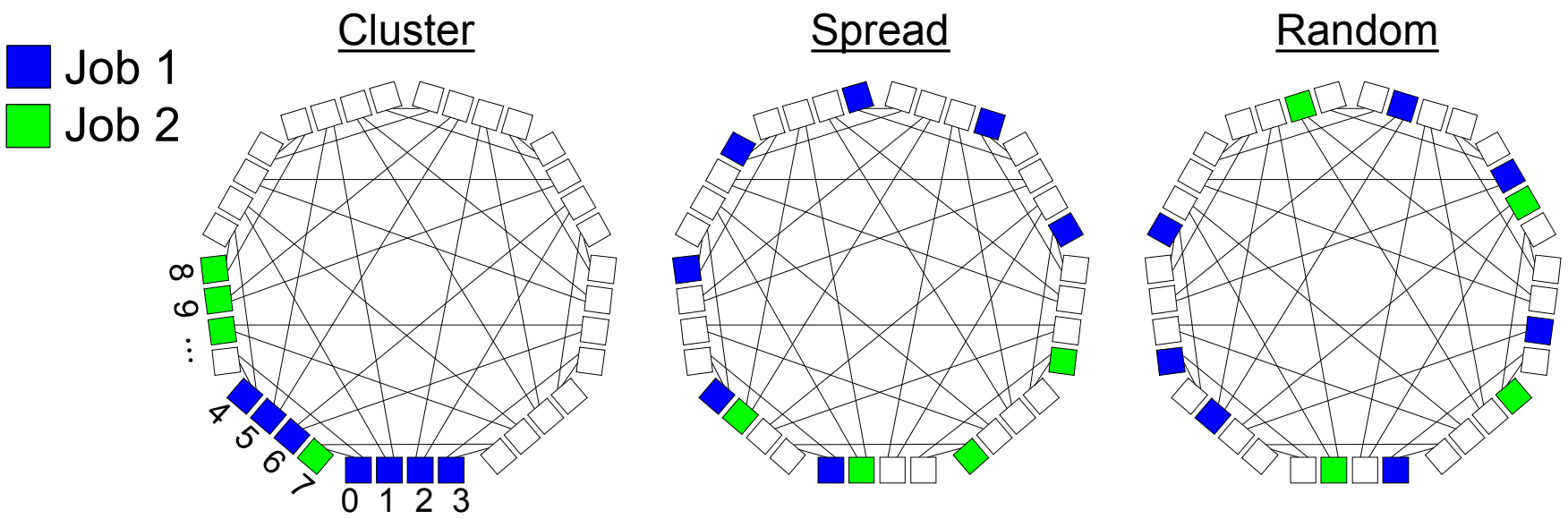
Source
Destination




- Workloads
  - All-to-all: maximum communication intensity (e.g., FFT)
  - 3D stencil: Common HPC communication pattern (e.g., CFD)
  - Bisection: stresses the bisection bandwidth
- 3 different message sizes (100, 1000, 4000KB)

# Experimental Methodology

- Job allocation algorithms

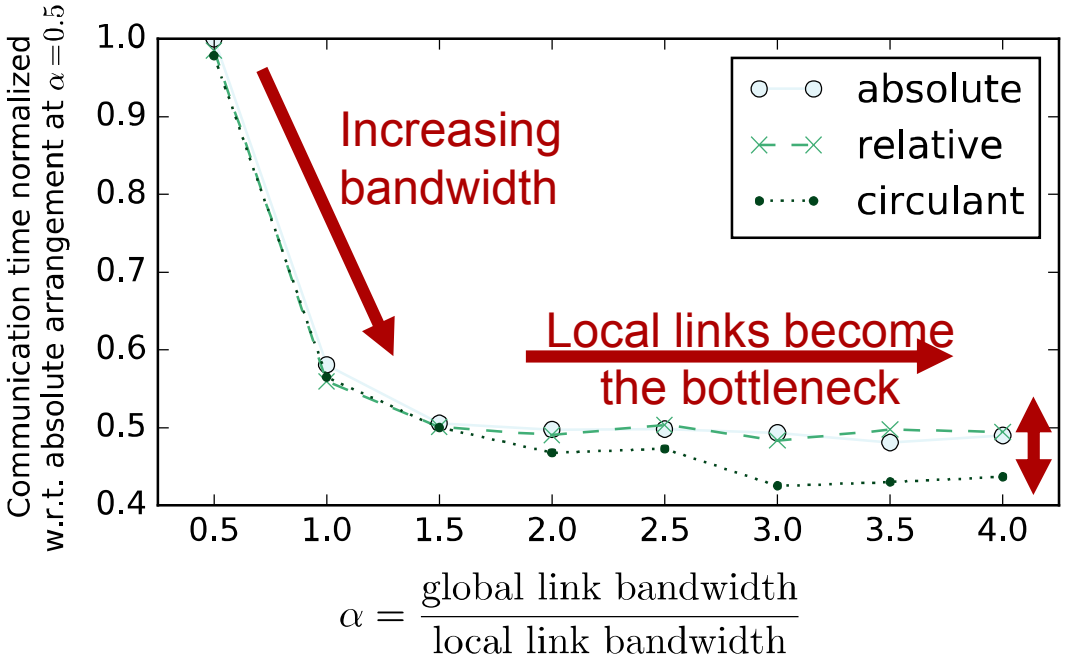



- Task-to-core mapping (15 times, randomized)
  - Random for all-to-all and bisection communication patterns
  - Recursive graph bisection for 3D stencil communication pattern *[Hoefler et al., ICS'11]*

# The Impact of Link Arrangement on Bisection Communication Pattern

Sandia National Laboratories

- Circulant-based arrangement has higher bisection bandwidth at high $\alpha$
  - Agreement with the theory using minimal routing *[Hastings et al., Cluster'15]*



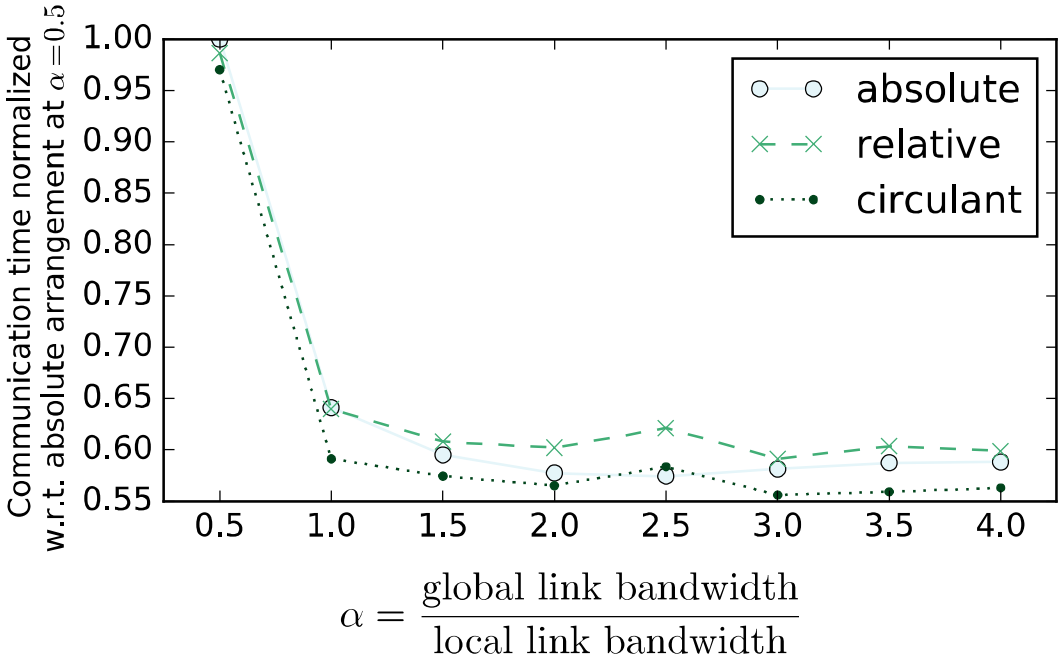

Minimal routing
Bisection pattern
Single job

Up to **15%** difference due to the change in bisection bandwidth

# The Impact of Link Arrangement on Bisection Communication Pattern

Sandia National Laboratories

- Circulant-based arrangement has higher bisection bandwidth at high $\alpha$
  - Agreement with the theory using minimal routing *[Hastings et al., Cluster'15]*
- The conclusion holds with Valiant (indirect) routing



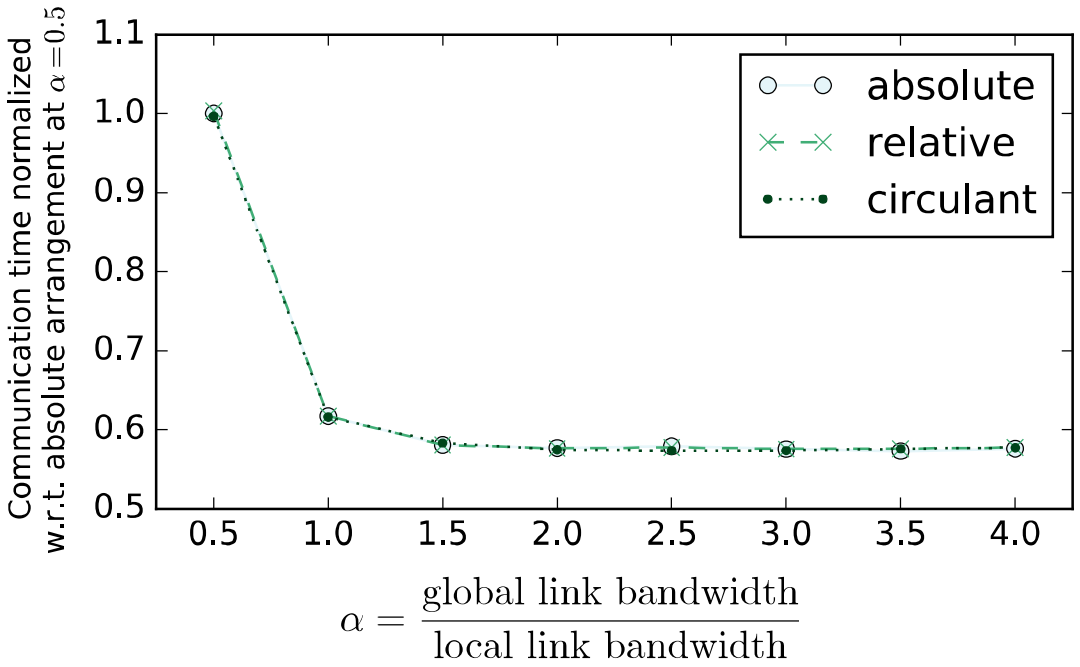

Valiant routing
Bisection pattern
Single job

**3-7%** difference

# The Impact of Link Arrangement on All-to-all Communication Pattern

Sandia National Laboratories

- Circulant-based arrangement reduces communication time by up to 3% for all-to-all communication patterns
- The same conclusion holds for valiant routing, stencil communication pattern, medium & large machines



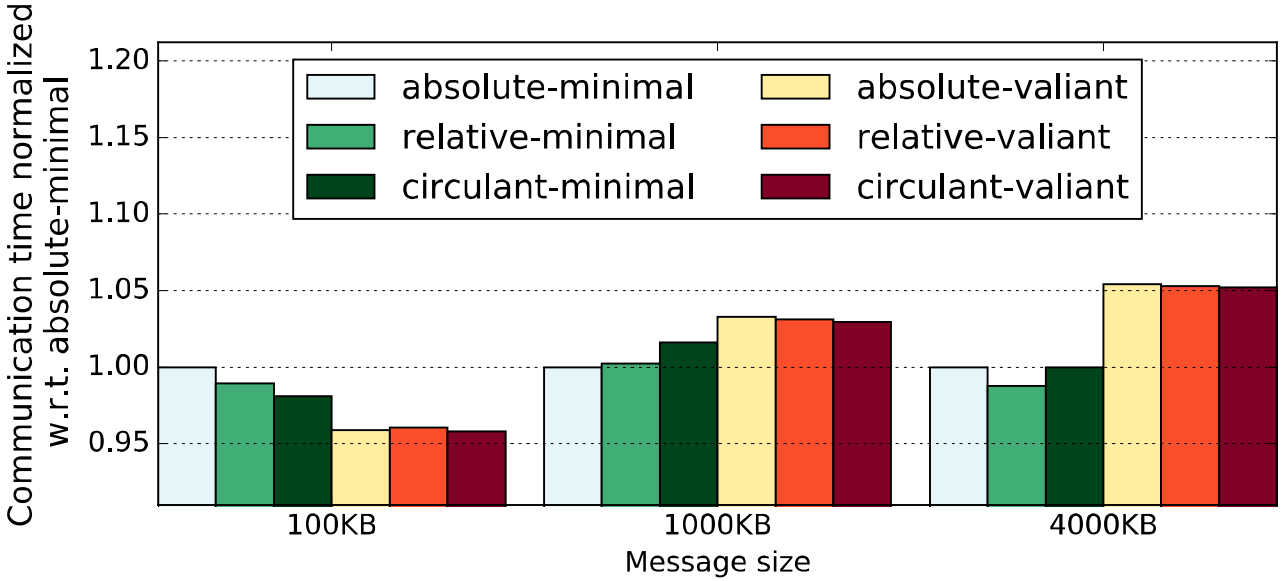

Minimal routing
All-to-all pattern
Single job

Up to **3%** difference

# The Impact of Routing Algorithm and Network Load on Communication Time

Sandia National Laboratories

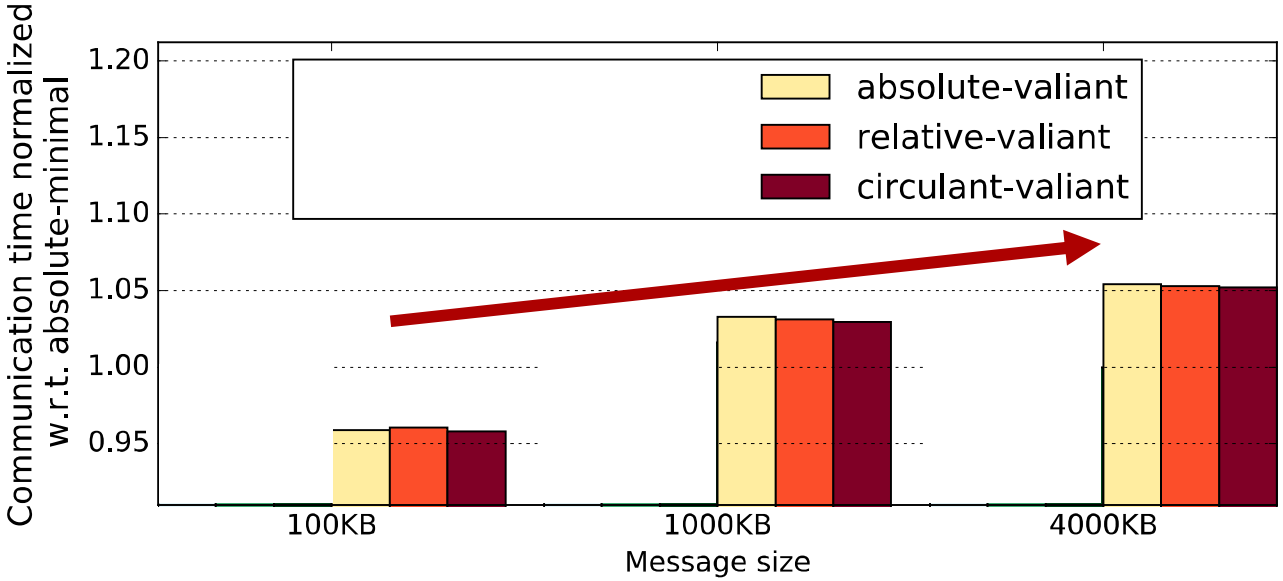
- Routing algorithm and network load level has minimal impact on the performance of global link arrangements
- Up to 6% difference in communication time between {link arrangement, routing algorithm} pairs in any network load level




All-to-all pattern
Single job
$\alpha = 0.5$

# The Impact of Routing Algorithm and Network Load on Communication Time

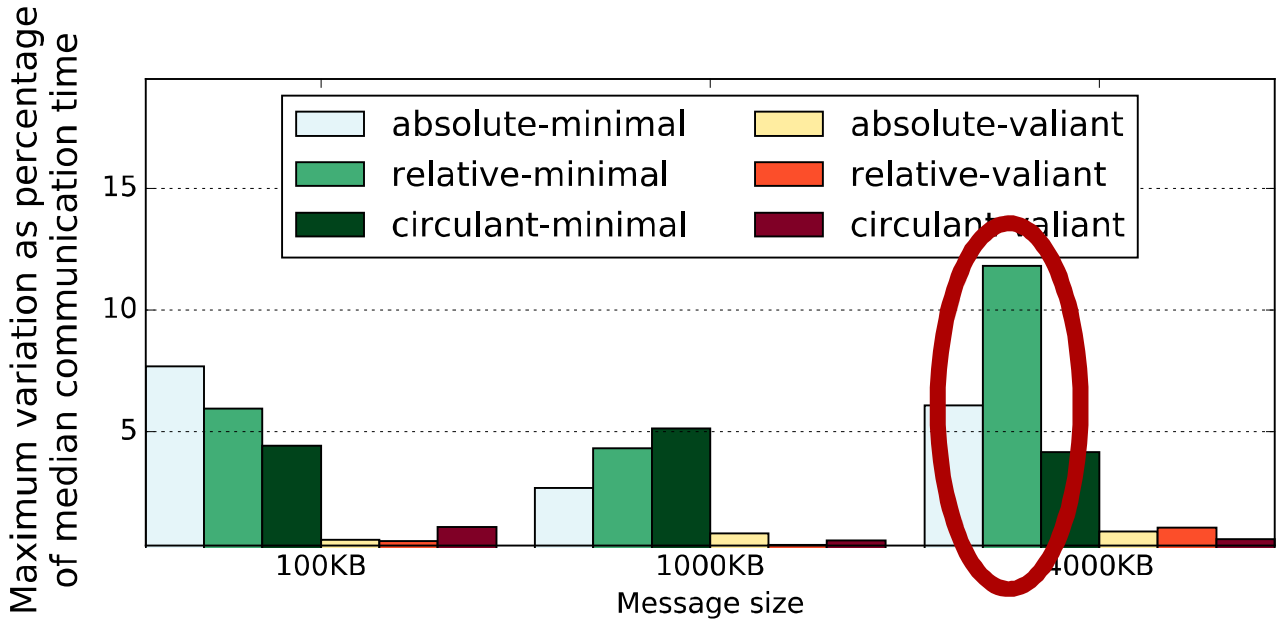
- Routing algorithm and network load level has minimal impact on the performance of global link arrangements
- Up to 6% difference in communication time between {link arrangement, routing algorithm} pairs in any network load level
- Valiant (indirect) routing clogs the network with increasing message size




All-to-all pattern
Single job
$\alpha = 0.5$

# Communication Time Variation due to Task-to-Core Mapping

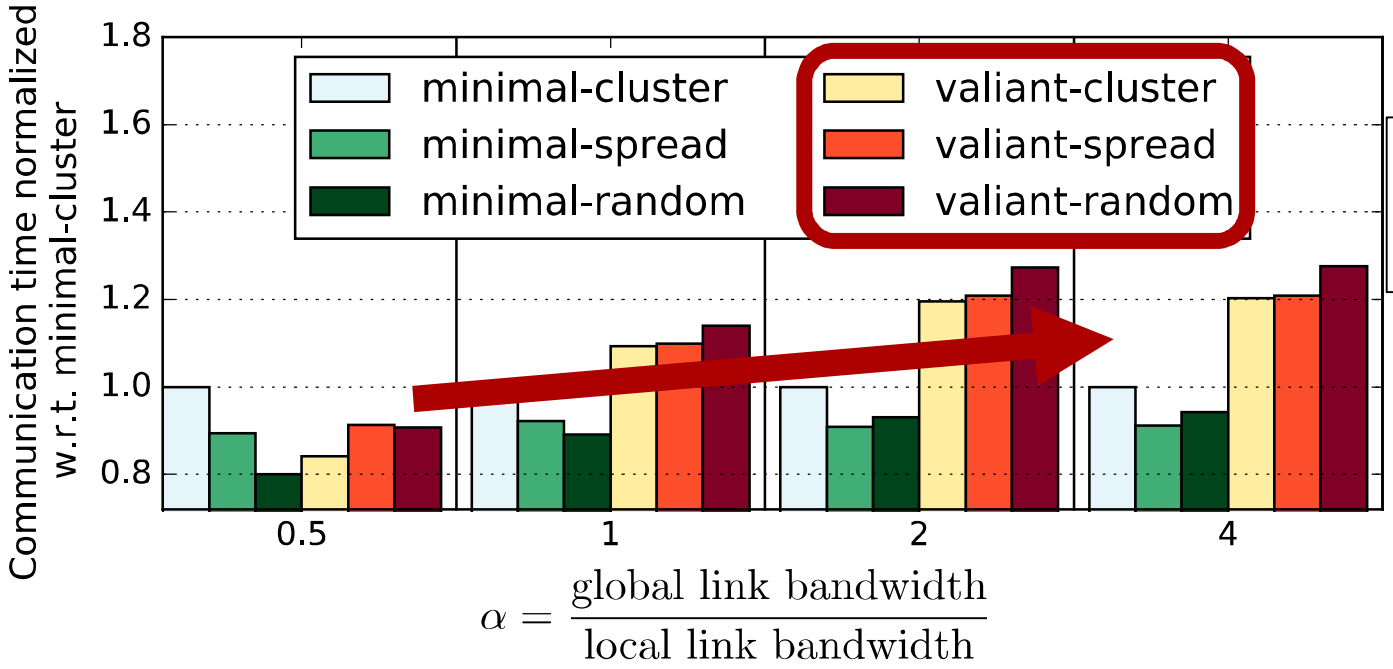
- Up to 11% communication time variation among random mappings
  - This is due to packet congestion in routers
- Valiant (indirect) routing reduces this variation by dispersing the traffic




Minimal routing
All-to-all pattern
Single job
$\alpha = 0.5$

# The Impact of  and Link Arrangements on Routing Algorithms

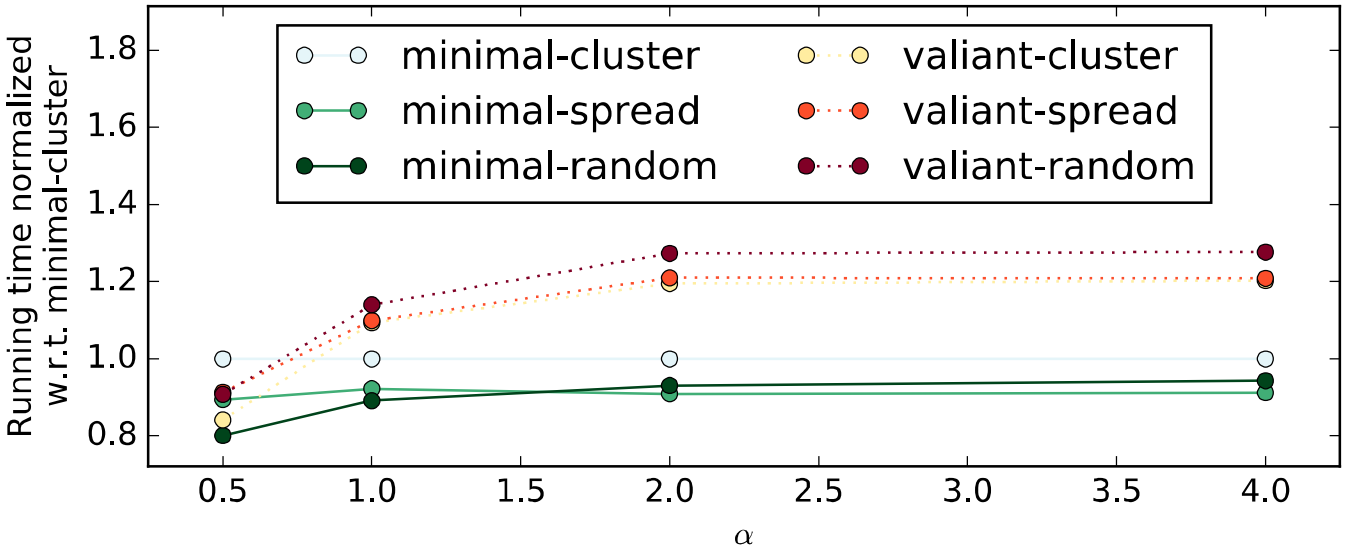
- When the global link is high, Valiant (indirect) routing creates bottlenecks in local links
  - The same effect is observed when the network load is high




3D stencil pattern
Two jobs, each utilizing 50% of the machine

# Multiple Job Analysis

- Running time of the stencil communication pattern using circulant arrangements with {routing, allocation} algorithm pairs
- Results are normalized with respect to the {minimal, cluster} pair at their corresponding α value




- This results in around 44% performance difference between best and worst pairs at α = 4.

# Conclusion

- We have studied the impact of dragonfly global link arrangements using various application communication patterns, routing algorithms, job placement algorithms, dragonfly parameters
- Main outcomes:
  - "Circulant-based" link arrangement decreases communication overhead compared to absolute arrangement by up to
    - 15% for bisection pattern
    - 3% for all-to-all and stencil patterns
  - The communication time increases by up to 11% due to inefficient task-to-core mapping
  - Valiant routing underperforms compared to minimal routing when the network load is high or when the global link bandwidth is high; it results in up to 44% longer running times in comparison to using minimal routing at $\alpha = 4$