Exceptional service in the national interest

Sandia National Laboratories

Center for Computing Research

Comparing global link arrangements for Dragonfly networks

Hastings, Rincon-Cruz, Spehlmann, Meyers, Xu, and Bunde (Knox College) and Vitus Leung (Discrete Math & Opt)

New Challenges in Scheduling Theory 2016

U.S. DEPARTMENT OF ENERGY

NNSA National Nuclear Security Administration

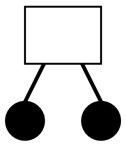
Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Dragonfly

- Hierarchical architecture to exploit high-radix switches and optical links

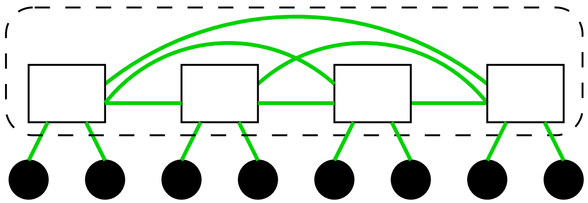
Dragonfly

- Hierarchical architecture to exploit high-radix switches and optical links
  - Nodes attached to switches



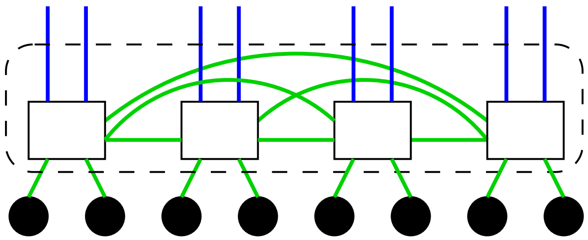
Dragonfly

- Hierarchical architecture to exploit high-radix switches and optical links
  - Nodes attached to switches
  - Switches form groups
  - Group members connected w/ local edge (electrical)



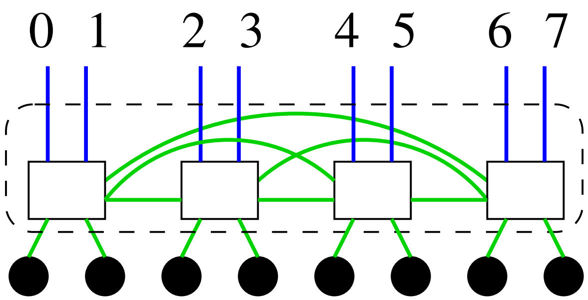
Dragonfly

- Hierarchical architecture to exploit high-radix switches and optical links
  - Nodes attached to switches
  - Switches form groups
  - Group members connected w/ local edge (electrical)
  - Each pair of groups connected w/ global edge (optical)



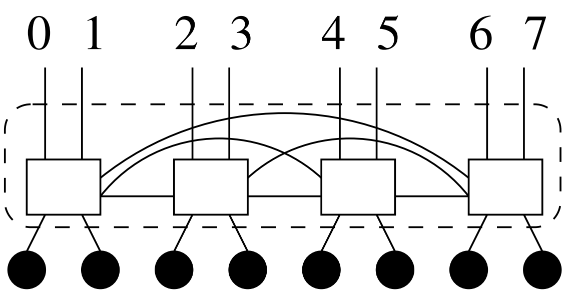
Dragonfly

- Hierarchical architecture to exploit high-radix switches and optical links
  - Nodes attached to switches
  - Switches form groups
  - Group members connected w/ local edge (electrical)
  - Each pair of groups connected w/ global edge (optical)



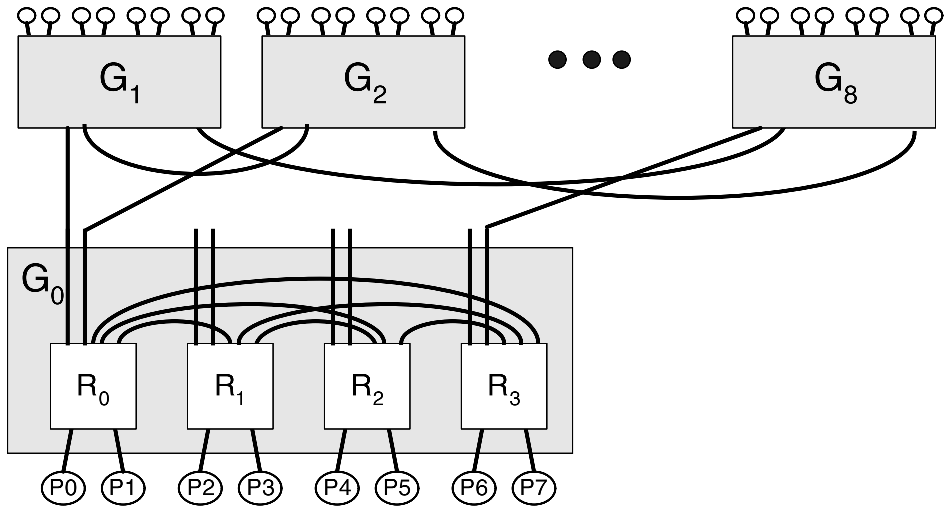
Dragonfly parameters

- p = number of nodes connected to a switch
- a = number of switches in a group
- h = number of optical links on a switch



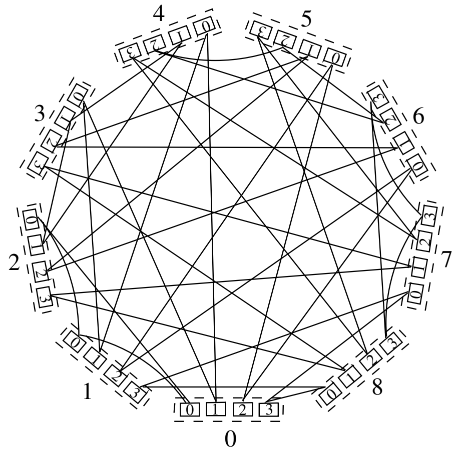
- Number of groups $g = ah+1$

Which port connects to which group?

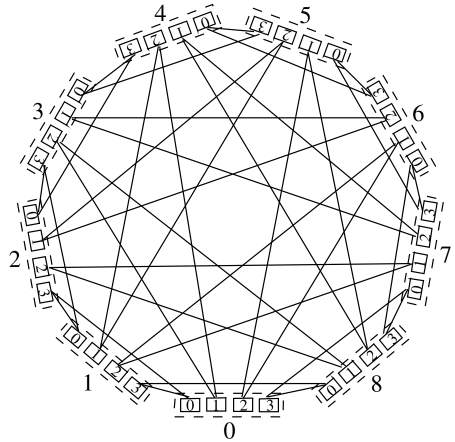


From original Dragonfly paper: Kim et al., ISCA 2008

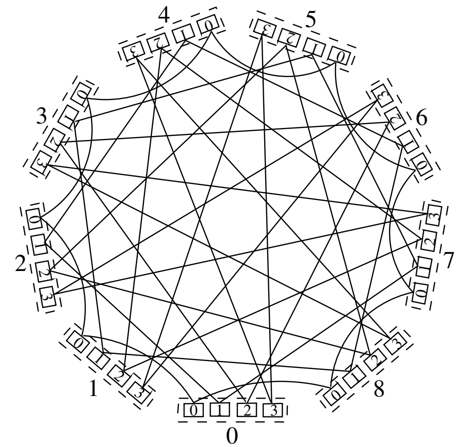
Three distinct global link arrangements



Absolute arrangement



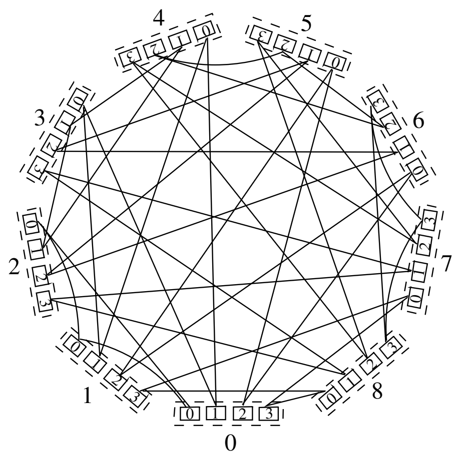
Relative arrangement



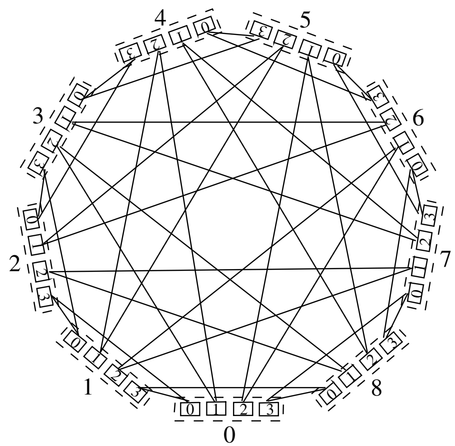
Circulant-based arrangement

Arrangements defined in Camarero et al. ACM Trans. Architec. Code Optim., 2014.

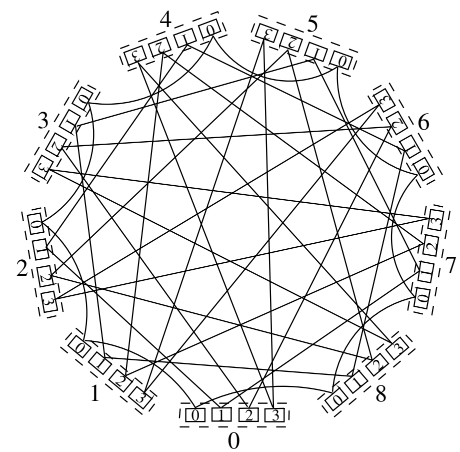
Three distinct global link arrangements



Absolute arrangement



Relative arrangement



Circulant-based arrangement

Arrangements defined in Camarero et al. ACM Trans. Architec. Code Optim., 2014.

Note:

IBM implementation (PERCS) uses absolute

Researchers who draw entire system in their papers use relative

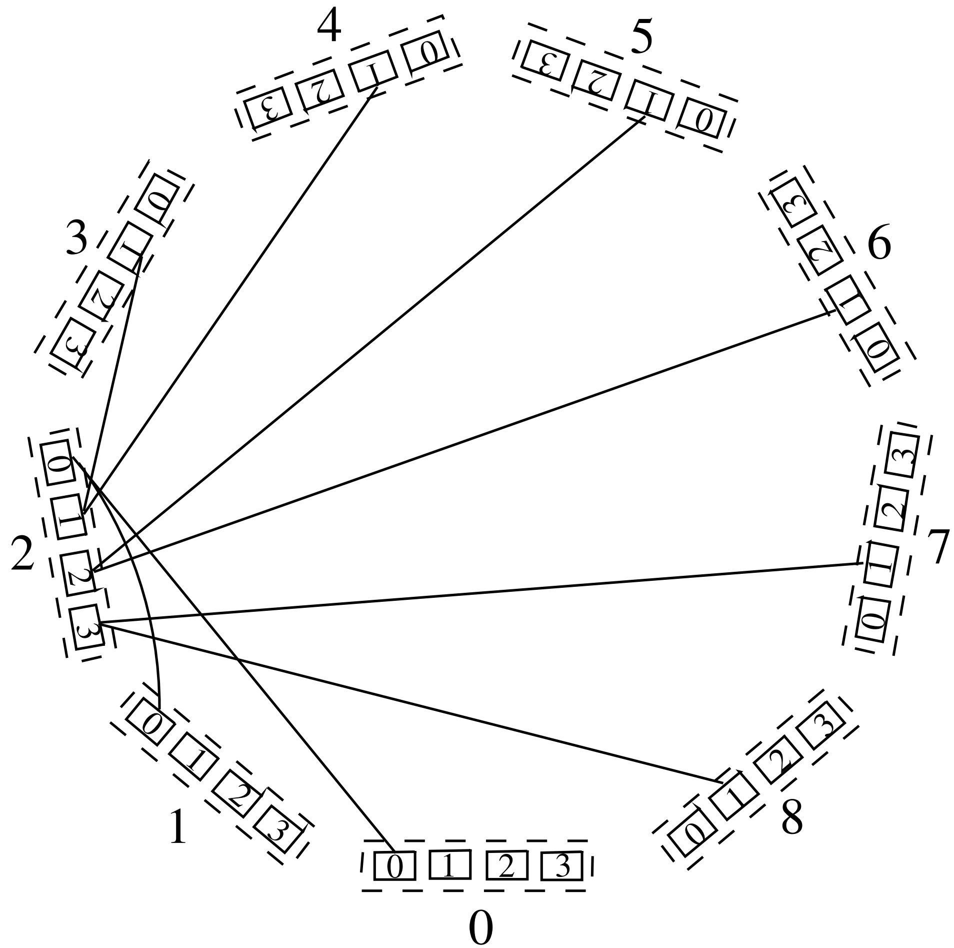
Absolute arrangement

(aka Consecutive arrangement)

Port *k* connects to group *k*
(except skip own group)

Equivalently, port *k* of group *i* connects to

group *k*	if $k < i$
group *k*+1	if $k \geq i$

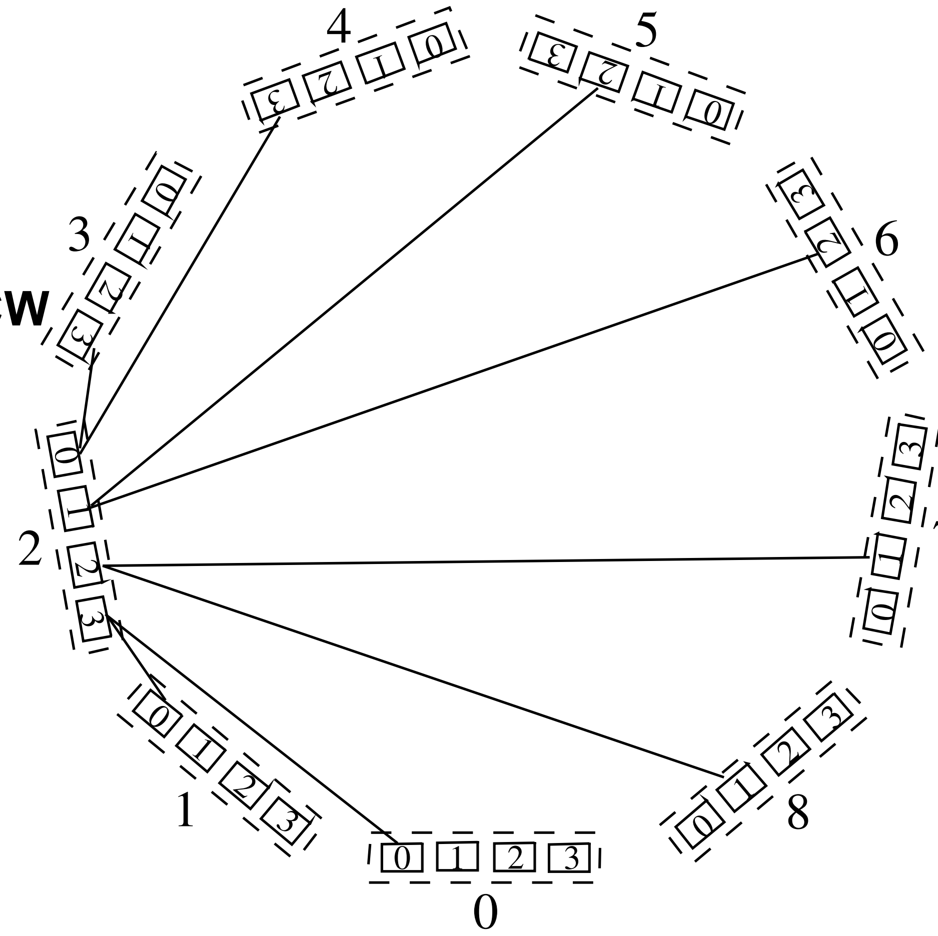


Relative arrangement

(aka Palmtree arrangement)

Port *k* connects (*k*+1)st group CW

Equivalently, port *k* of group *i* connects to group (*i*+*k*+1) mod *g*



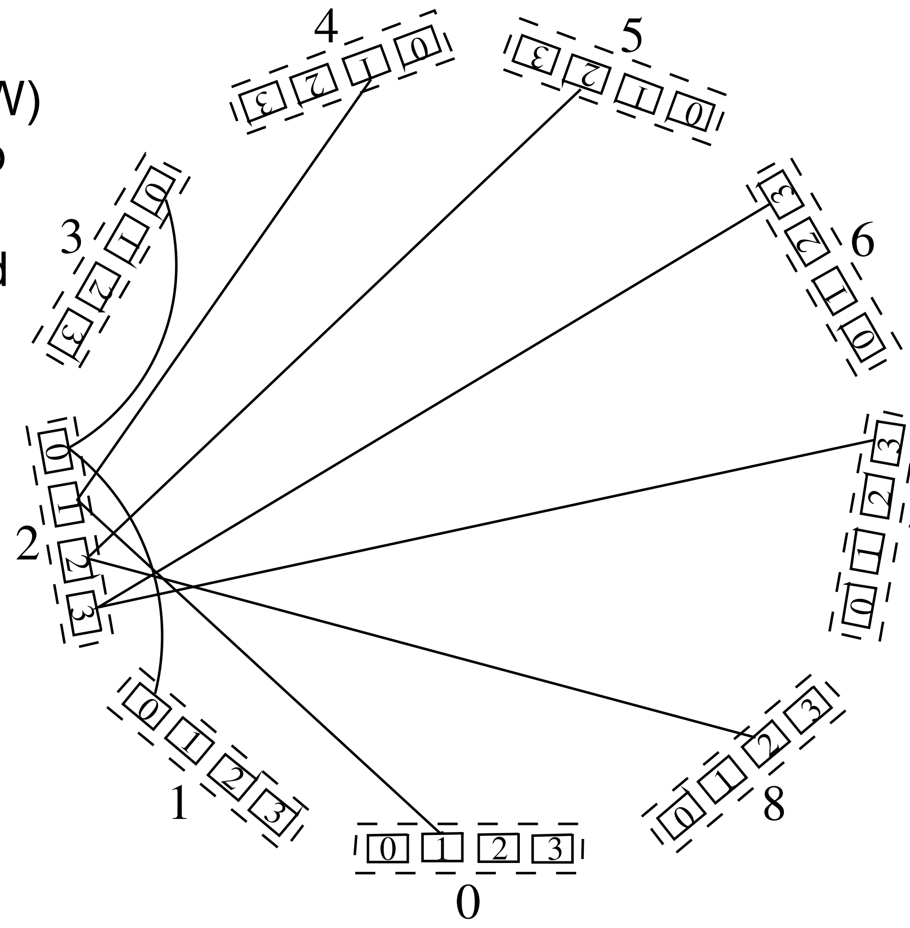
Circulant-based arrangement

Port 0 connects to next group (CW)
Port 1 connects to previous group
Port 2 connects to group 2 ahead
Port 3 connects to group 2 behind
...

Equivalently, port *k* of group *i* connects to group

$(i+k/2+1) \bmod g$ if k is even

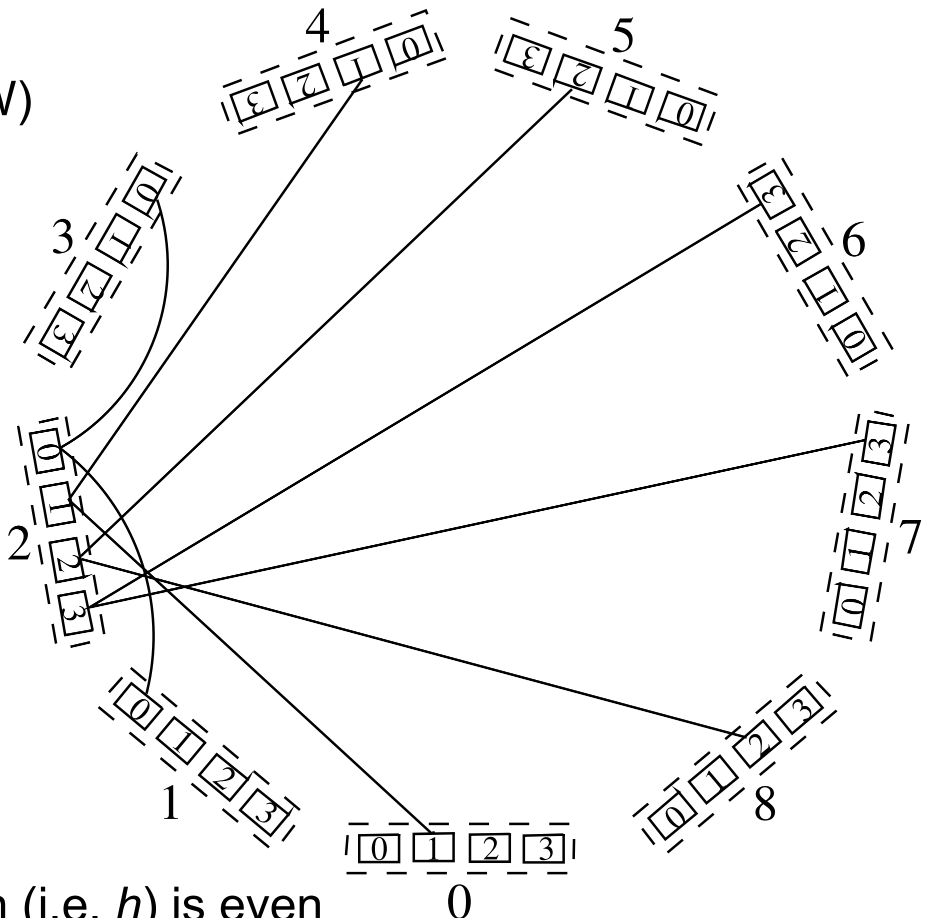
$(i-k/2-1) \bmod g$ if k is odd



Circulant-based arrangement

Port 0 connects to next group (CW)
Port 1 connects to previous group
Port 2 connects to group 2 ahead
Port 3 connects to group 2 behind
...

Equivalently, port *k* of group *i* connects to group
$(i+k/2+1) \bmod g$ if *k* is even
$(i-k/2-1) \bmod g$ if *k* is odd



Notes:
Assumes # global links/switch (i.e. *h*) is even
Always connects switches at same position in their groups

Our contribution

- Comparing absolute, relative, and circulant-based arrangements
  - Bisection bandwidth
  - "Ease of use" with task mapping
    - Criteria for good mapping adapted from Prisacari et al., IPDPS 2013
    - Communication in phases such that
      - Messages distributed evenly on links
      - All paths in a phase have same length

Bisection bandwidth

- Minimum bandwidth between two equal-sized parts of the system
  - Bandwidth for a particular bisection is the (weighted) number of edges crossing from one part to the other
  - Minimize this over all bisections

- Tries to measure worst-case communication bottleneck in a large computation

Initial exploration

- Branch and bound on small Dragonfly system (NP-hard …)

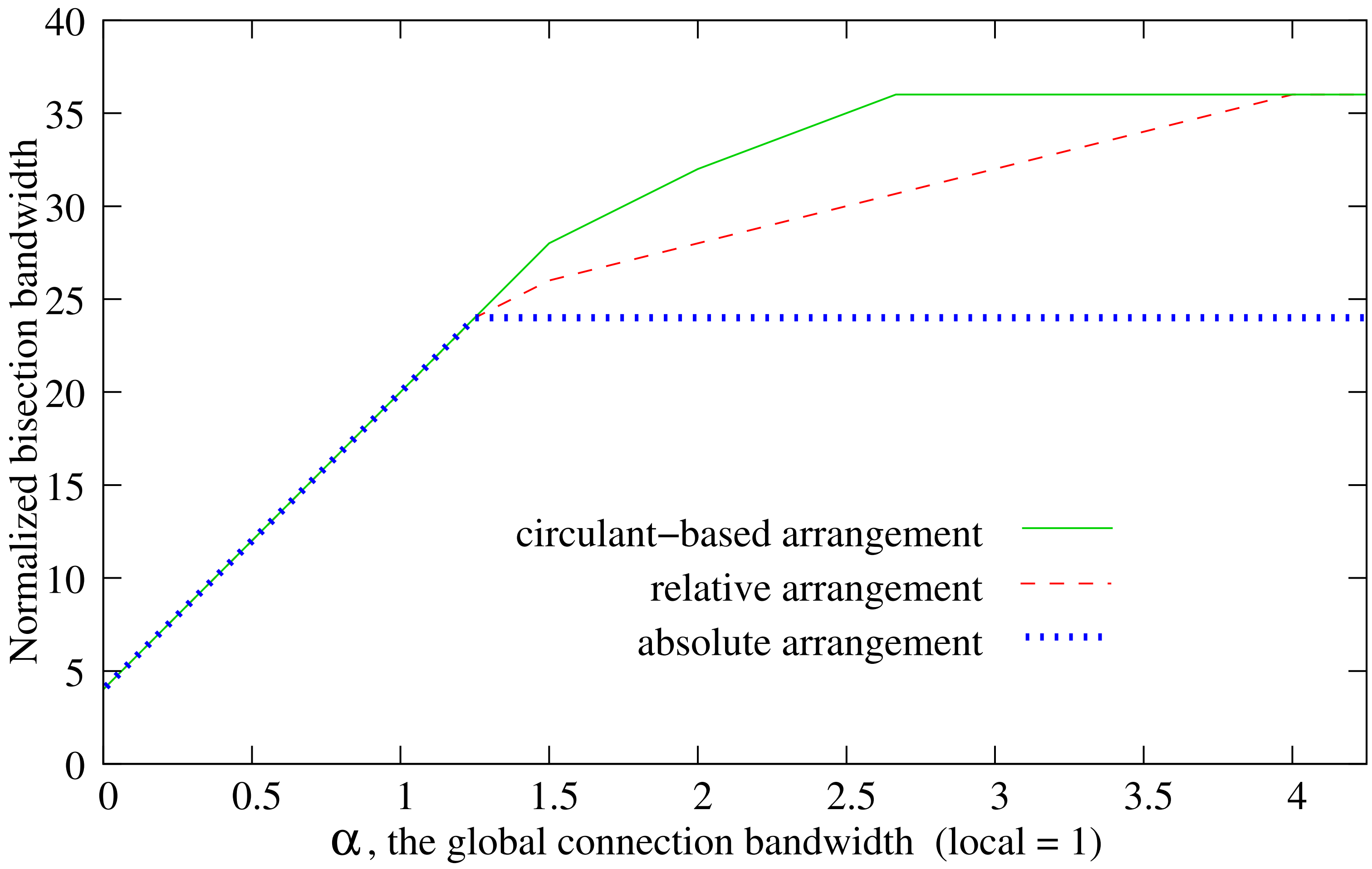
  (p,4,2): 4 switches per group

  2 global links per switch

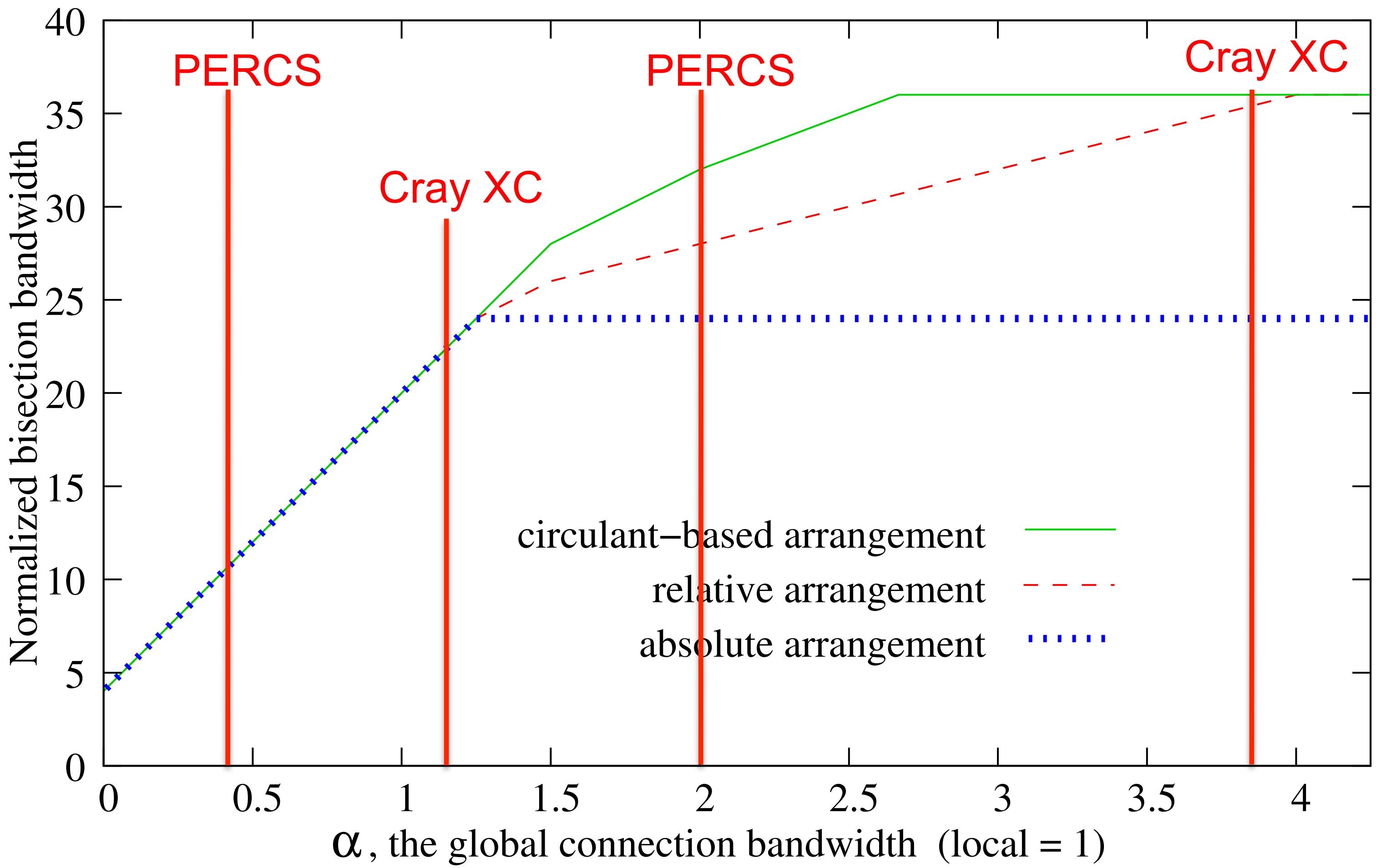
  Has 36 switches

- Treat types of edges separately
  - local edges have bandwidth 1
  - global edges have bandwidth α

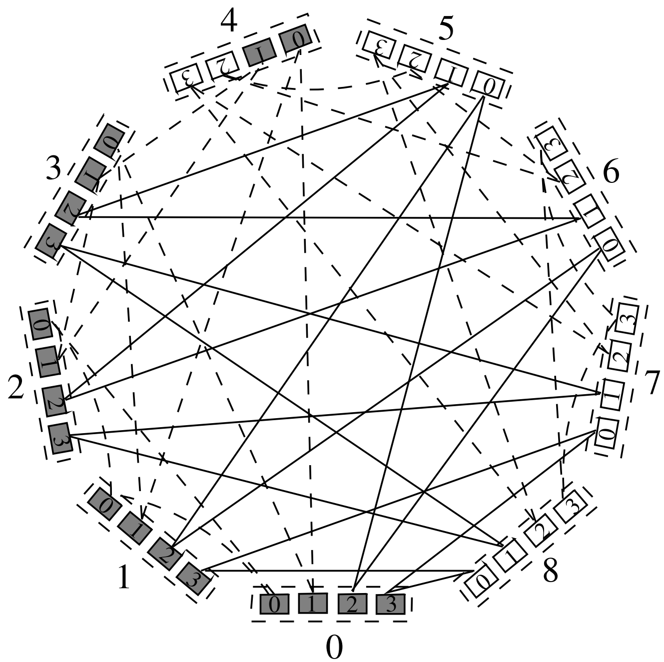
Bisection bandwidth as function of α



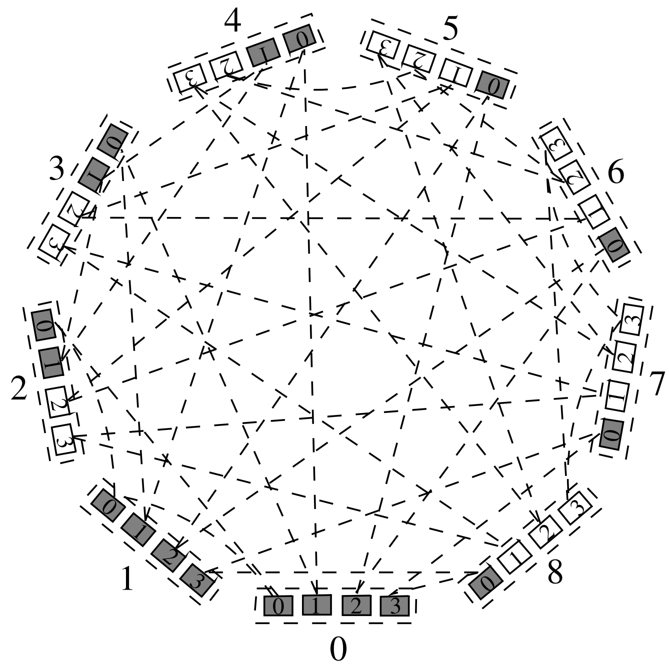
Bisection bandwidth as function of α



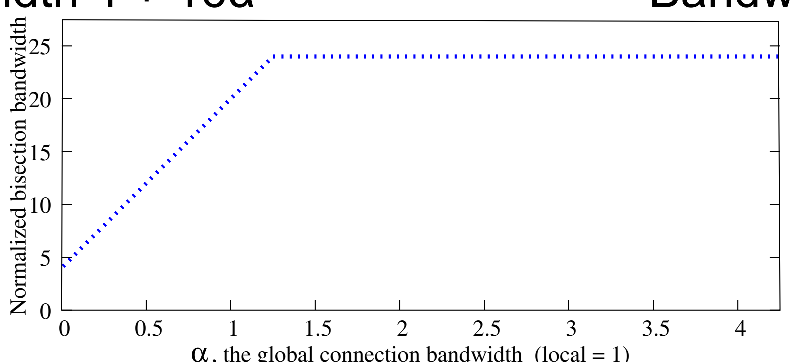
Min-bandwidth cuts for absolute arrangement



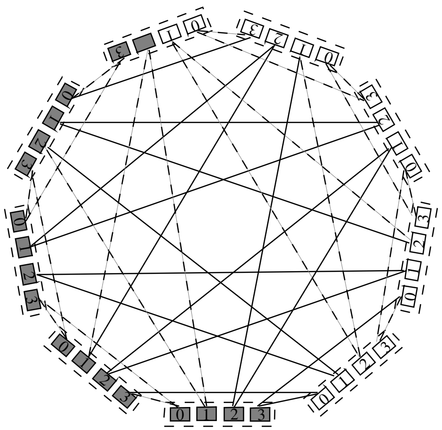
Bandwidth 4 + 16α



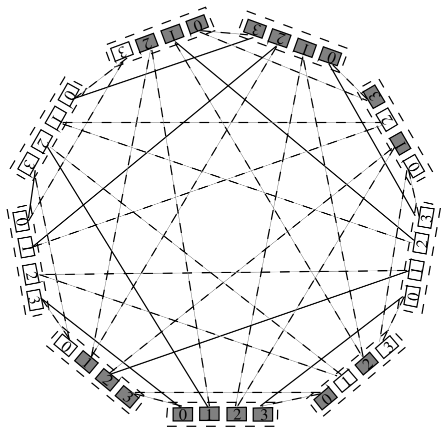
Bandwidth 24



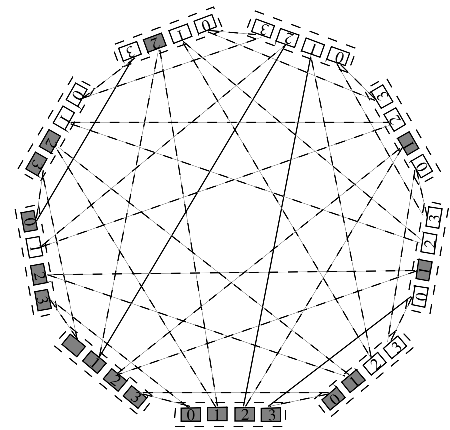
Min-bandwidth cuts for relative arrangement



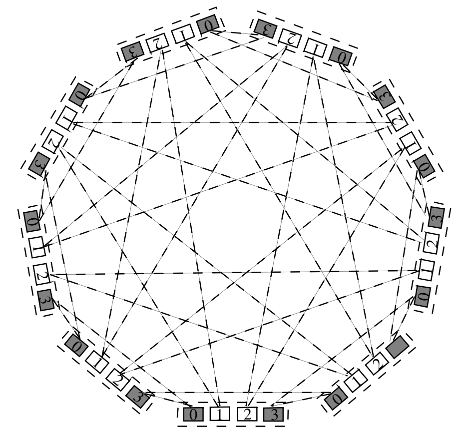
bandwidth 4 + 16α



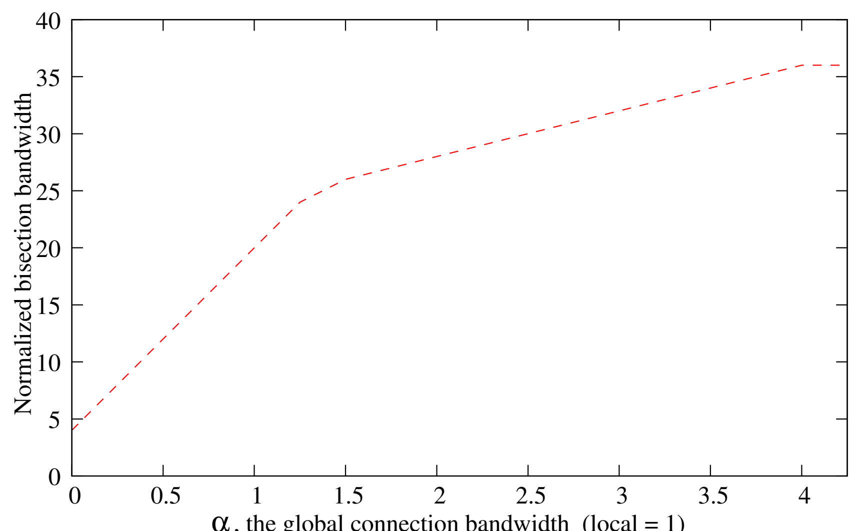
bandwidth 14 + 8α



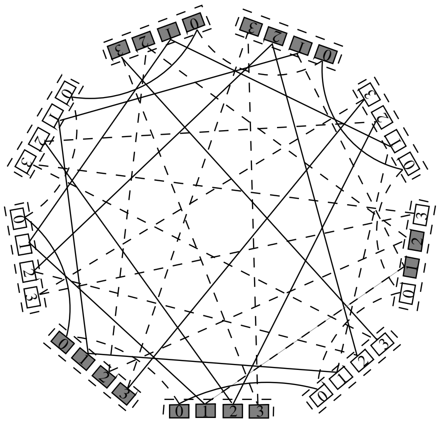
bandwidth 20 + 4α



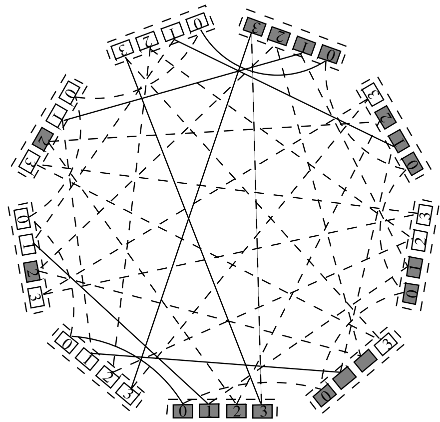
bandwidth 36



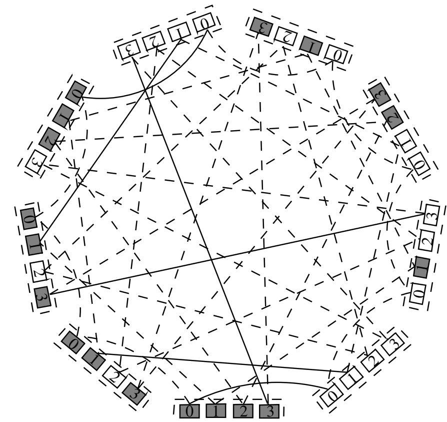
Min-bandwidth cuts for circulant-based arrangement



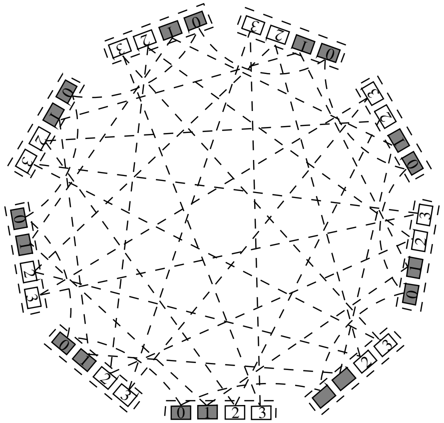
bandwidth 4 + 16α



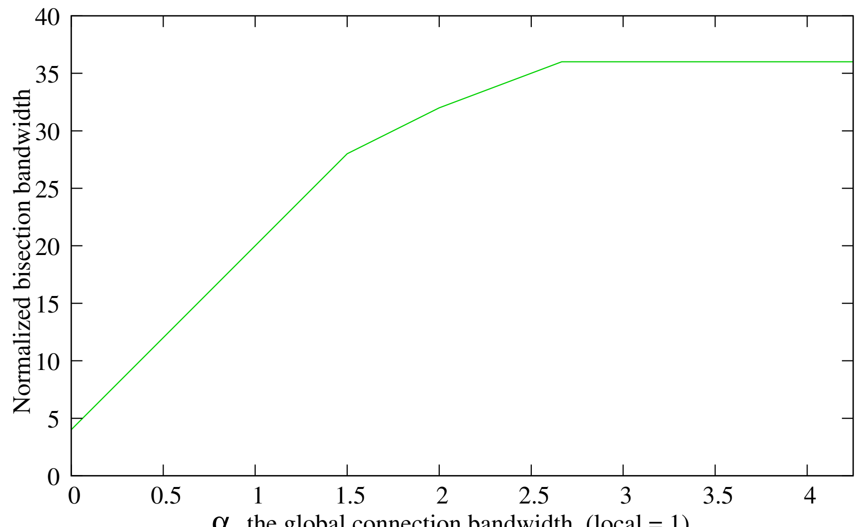
bandwidth 16 + 8α



bandwidth 20 + 6α



bandwidth 36



Observations from (*p*,4,2)

- In terms of bisection bandwidth:

  Absolute ≤ Relative ≤ Circulant-based

- For all three arrangements, maximum bisection bandwidth is bounded

Larger networks

- Focus on large α
  - Determine when bisection bandwidth is ultimately limited by local edges
- Globally Connected Component (GCC): Switches that form connected component in graph without local edges

GCCs in Circulant-based arrangements

Recall: Every edge connects two switches at same position in their respective groups

GCCs in Circulant-based arrangements

Recall: Every edge connects two switches at same position in their respective groups

There are at least *a* GCCs (*a* = #switches/group)

GCCs in Circulant-based arrangements

Recall: Every edge connects two switches at same position in their respective groups

There are at least a GCCs (a = #switches/group)

If a is even and α is sufficiently large, the bisection bandwidth is $(a/2)^2 g$ (g = #groups)

GCCs in Circulant-based arrangements

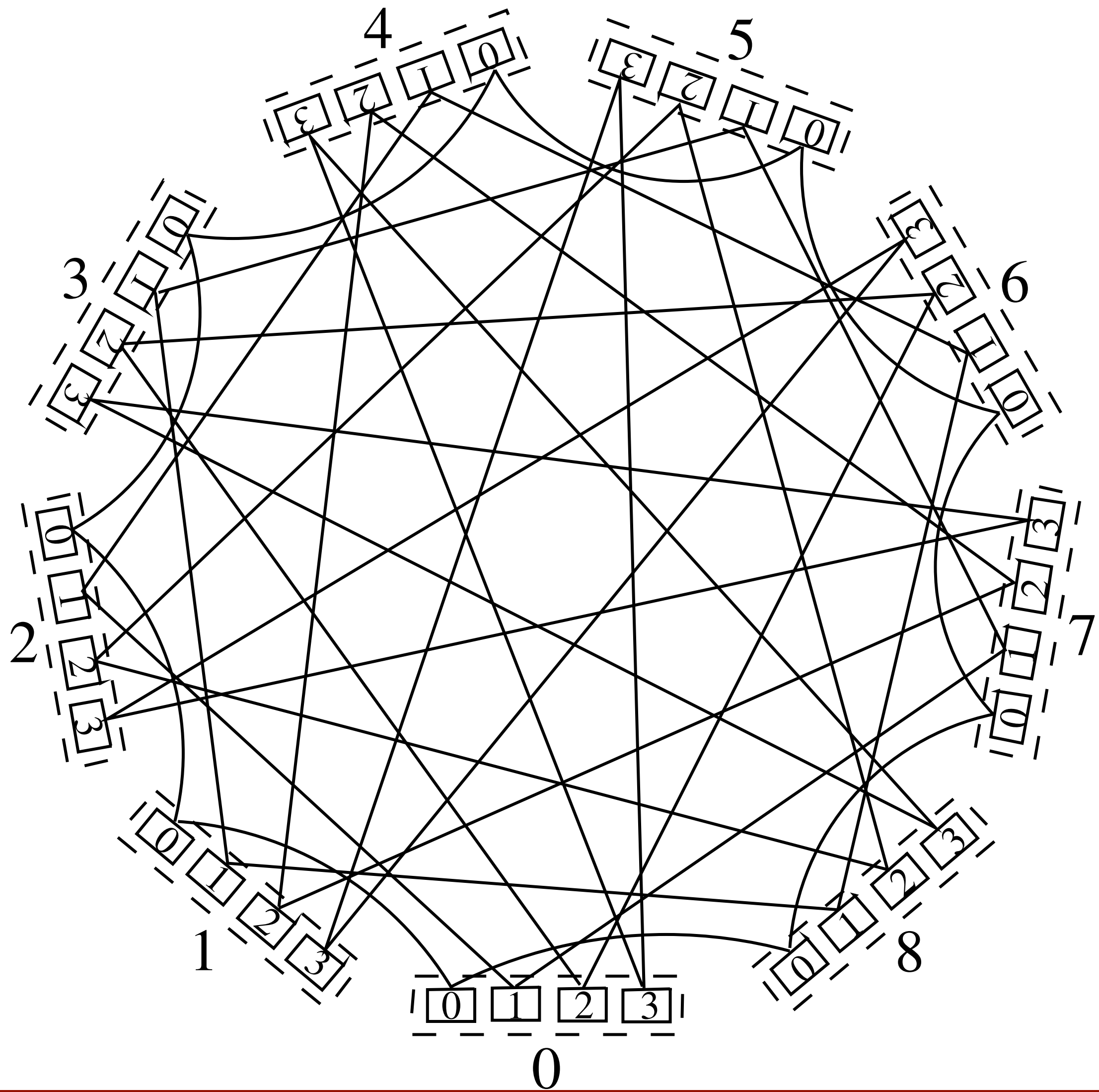
Recall: Every edge connects two switches at same position in their respective groups

There are at least *a* GCCs (*a* = #switches/group)

If *a* is even and α is sufficiently large, the bisection bandwidth is $(a/2)^2 g$ (*g* = #groups)

Structure of GCCs potentially more complicated than that, single switch number can be split into multiple GCCs if *g* is multiple of distance traversed by switch's links

Circulant-based arrangement



GCCs in Relative arrangements

Recall: Port k connects to $(k+1)^{st}$ group CW

GCCs in Relative arrangements

Recall: Port k connects to $(k+1)^{st}$ group CW

Switch 0 connects to switch (a-1) in next group

GCCs in Relative arrangements

Recall: Port k connects to $(k+1)^{st}$ group CW

Switch 0 connects to switch $(a$-1) in next ~~group~~

h groups

GCCs in Relative arrangements

Recall: Port k connects to $(k+1)^{st}$ group CW

Switch 0 connects to switch (a-1) in next ~~group~~

h groups

All 0^{th} and $(a-1)^{st}$ switches form a GCC

GCCs in Relative arrangements

Recall: Port k connects to $(k+1)^{st}$ group CW

Switch 0 connects to switch $(a-1)$ in next ~~group~~

h groups

All 0^{th} and $(a-1)^{st}$ switches form a GCC

Generalizes:

$a/2$ GCCs of size $2g$ (plus 1 of size g if a is odd)

Bisection bandwidth in Relative arrangement

When α is sufficiently large, bisection bandwidth is

$(a/2)^2 g$ if a is a multiple of 4

θ(α) otherwise

GCCs in Absolute arrangements

Recall: Port k connects to group k (skip own group)

Gives

- $a(a-1)/2$ GCCs of size $2h$
- a GCCs of size $h+1$

GCCs in Absolute arrangements

Recall: Port k connects to group k (skip own group)

Gives

- $a(a-1)/2$ GCCs of size $2h$
- a GCCs of size $h+1$

If a is a multiple of 4, bisection bandwidth is $\leq (a/2)^2 g$.

(Also 3 other times, including when $h \leq a/2$)

GCCs in Absolute arrangements

Recall: Port k connects to group k (skip own group)

Gives

$a(a\text{-}1)/2$ GCCs of size $2h$

a GCCs of size h+1

If a is a multiple of 4, bisection bandwidth is $\leq (a/2)^2 g$.

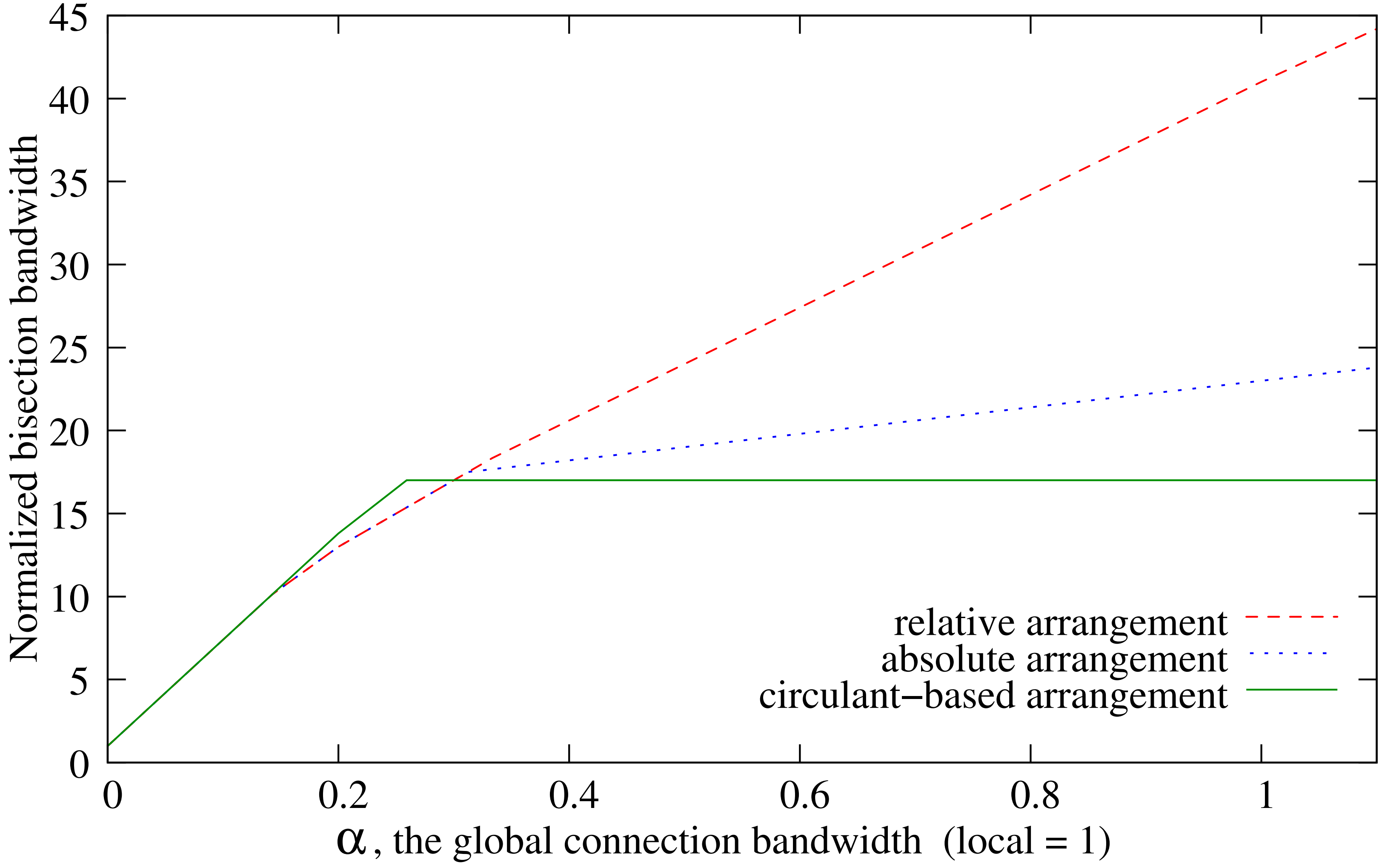
(Also 3 other times, including when $h \leq a/2$)

Otherwise, $\theta(\alpha)$

When bisection bandwidth is bounded

- Circulant: a is even (& other times)
- Relative: a is a multiple of 4
- Absolute: a is a multiple of 4 (& 3 other times, including when $h \leq a/2$)

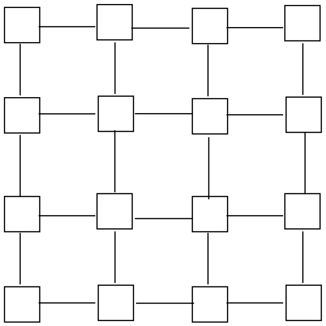
Normalize bisection bandwidth for (*p*, 2, 8)



Task mapping

- Assignment of tasks to compute nodes to minimize contention

- Our assumptions:
  - Stencil jobs



  - Tasks blocked to fit on entire switch

Criteria for good mapping

Adapted from Prisacari et al., IPDPS 2013:

Communication in phases such that:

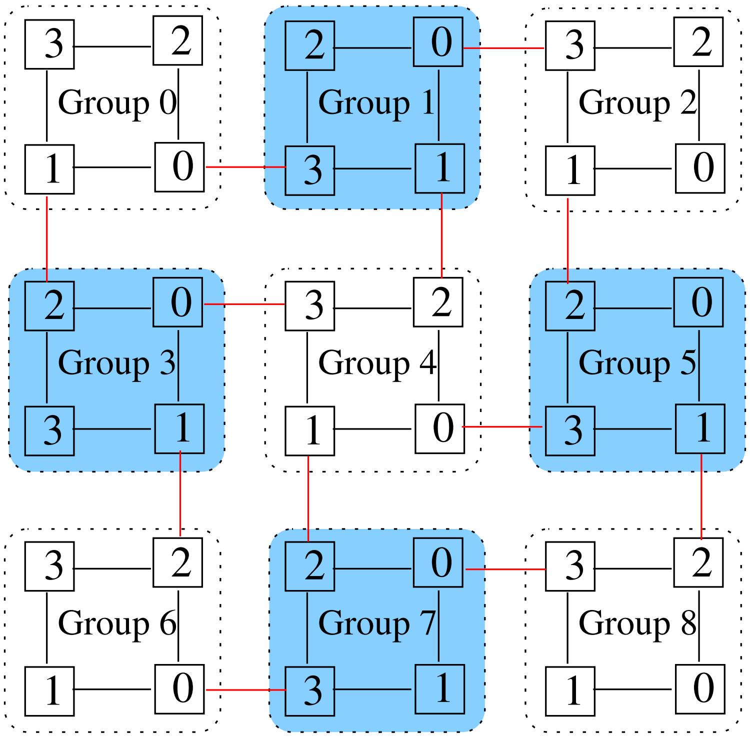
1. Messages distributed evenly on links
2. All paths in a phase have same length

Criteria for good mapping

Adapted from Prisacari et al., IPDPS 2013:

Communication in phases such that:

1. Messages distributed evenly on links
2. All paths in a phase have same length



Mapping of 6 × 6 job onto (p,4,2) Dragonfly with relative global link arrangement

Criteria for good mapping

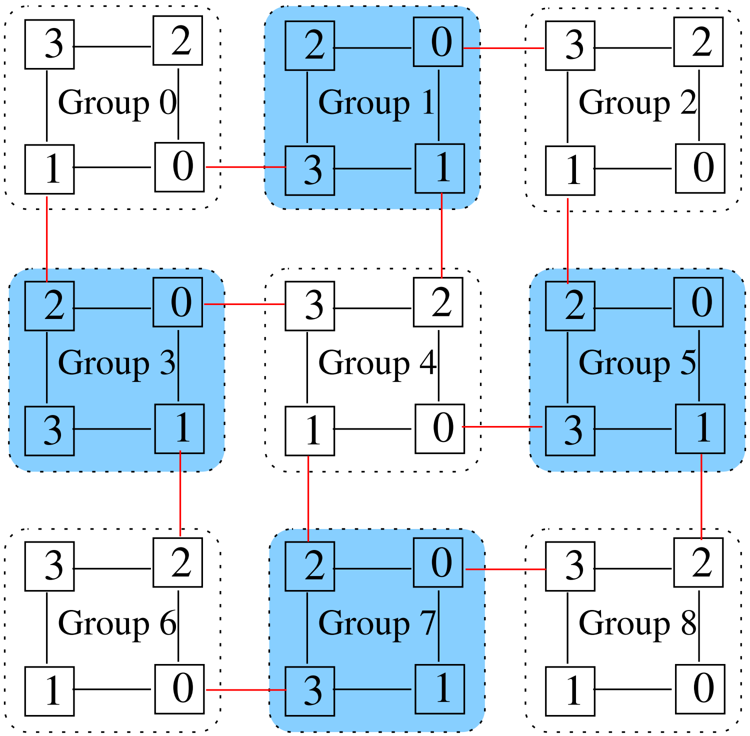
Adapted from Prisacari et al., IPDPS 2013:

Communication in phases such that:

1. Messages distributed evenly on links
2. All paths in a phase have same length

Phases for this mapping:

- Neighbors w/ local links
- Neighbors directly connected by global link
- Neighbors with multi-hop path



Mapping of 6 × 6 job onto (*p*,4,2) Dragonfly with relative global link arrangement

Criteria for good mapping

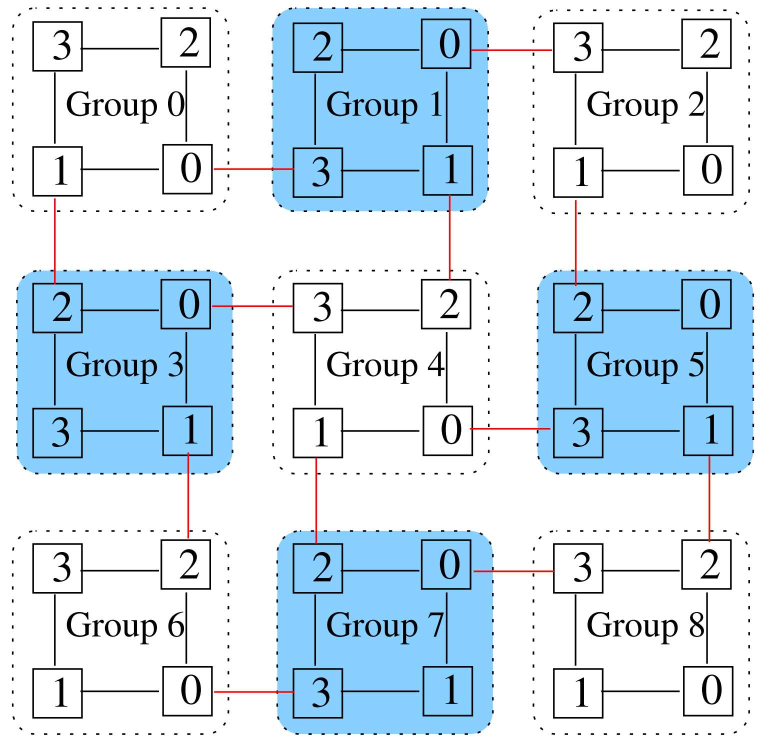
Adapted from Prisacari et al., IPDPS 2013:

Communication in phases such that:

1. Messages distributed evenly on links
2. All paths in a phase have same length

Phases for this mapping:

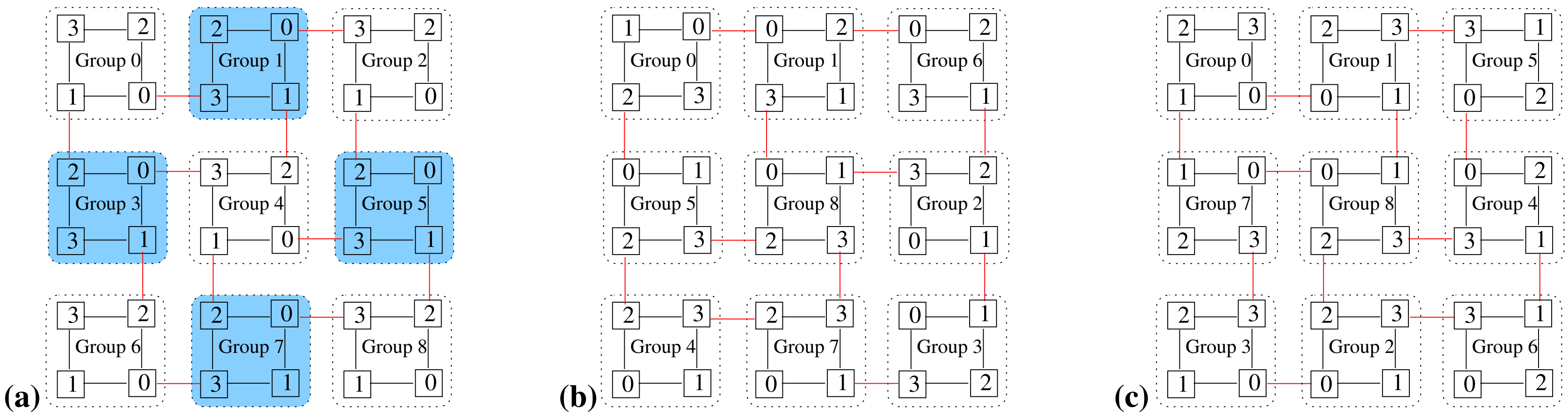
- Neighbors w/ local links
- Neighbors directly connected by global link
- Neighbors with multi-hop path



Mapping of 6 × 6 job onto (*p*,4,2) Dragonfly with relative global link arrangement

Nothing this regular seems to exist for absolute or circulant-based arrangements

Three mappings for a 6 x 6 stencil job



(a) Relative

(b) Absolute

(c) Circulant-based

Conclusions

- On original (p, 4, 2) graph, for bisection bandwidth:

  Absolute ≤ Relative ≤ Circulant-based

- On large graphs, Circulant-based is most often bounded, then Absolute, then Relative

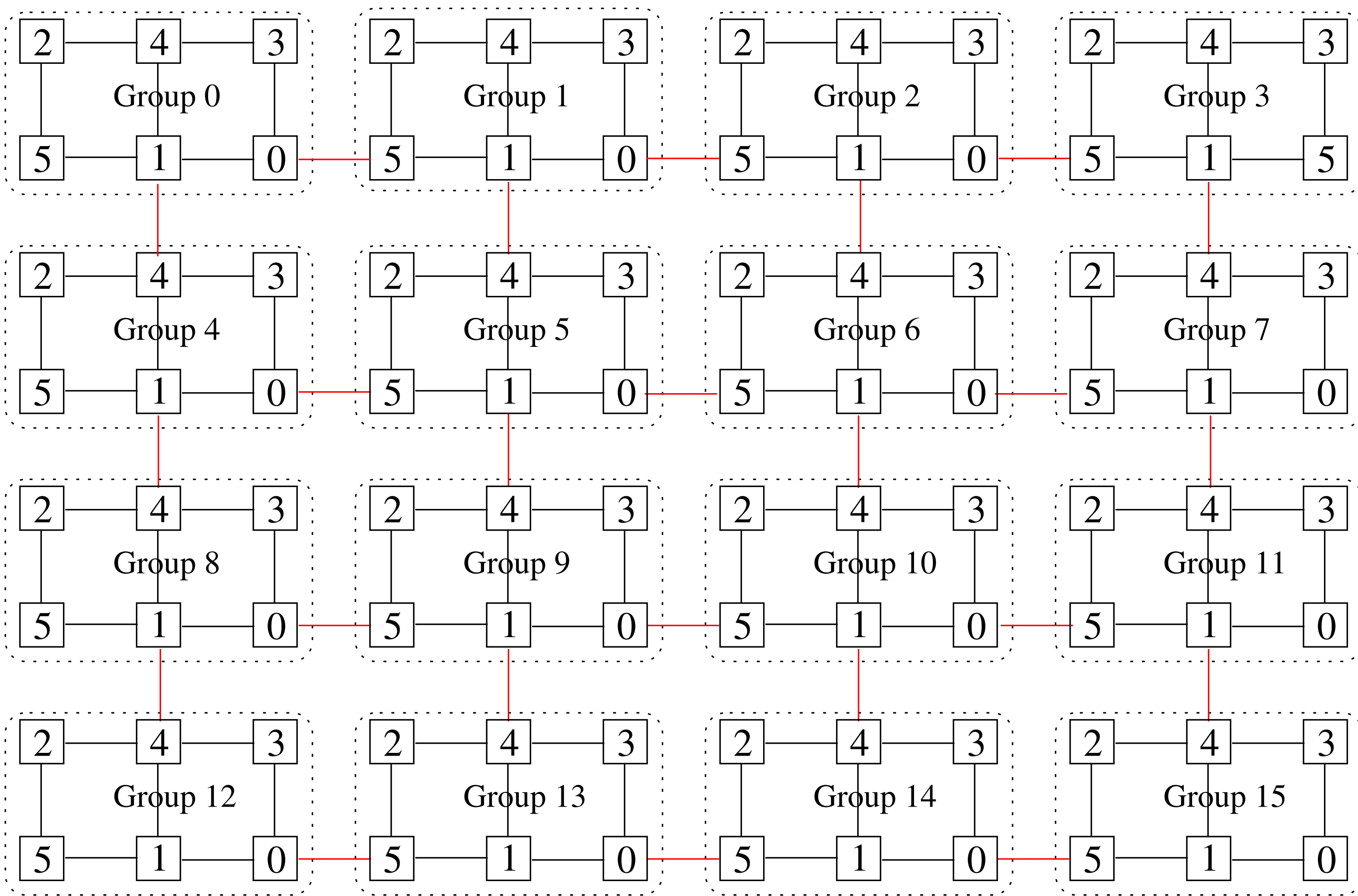
- On (p, 2, 8) graph, at large α:

  Circulant-based ≤ Absolute ≤ Relative

  and Absolute and Relative unbounded

- For mapping stencils, Relative gives much more regular mappings

Mapping for a 12 x 8 stencil job on 16 groups of (p, 6, 3)-Dragonfly with rel.



Future work

- Bisection bandwidth at smaller values of α
- Other global link arrangements
- Generalize task mapping and evaluation by simulation
- Communication scheduling recommended by Prisacari et al. may be difficult to implement
- Early Sandia Trinity applications measurements
  - Communications stalls surprisingly high
  - Thermal problems in turbo mode, 25°F swings

Thanks!

- vjleung@sandia.gov