Comparing global link arrangements for Dragonfly networks

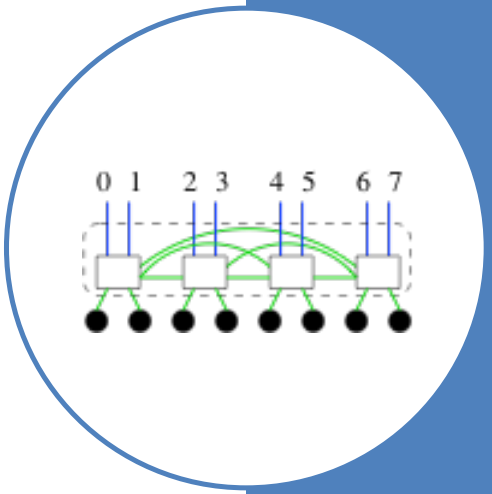
Mercy Swinney

Agenda

- Choice global link arrangement can change the achievable network performance.
  - Absolute
  - Relative
  - Circulant-based
- Global link arrangement's impact on regularity of task mapping

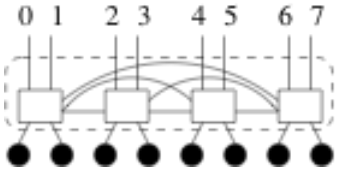
Dragonfly

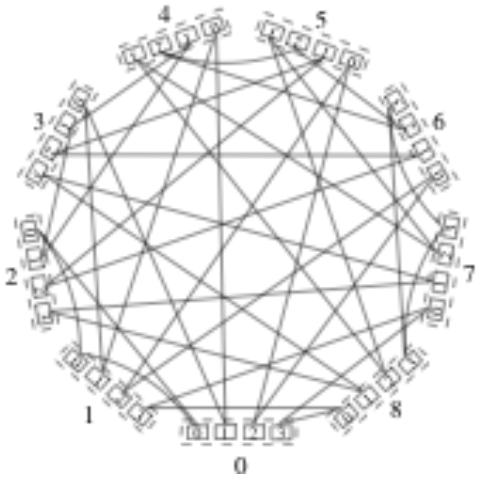
- Hierarchical architecture to exploit high-radix switches and optical links
  - Nodes attached to switches
  - Switches form groups
  - Group members connected w/ local edge
  - Each pair of groups connected w/ global edge



Dragonfly parameters

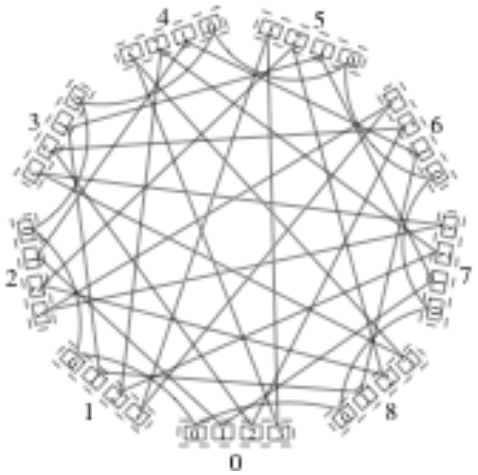
- p = number of nodes connected to a switch
- a = number of switches in a group
- h = number of global links on a switch
- Number of groups g = ah+1





Absolute arrangement

Relative arrangement



Circulant-based arrangement

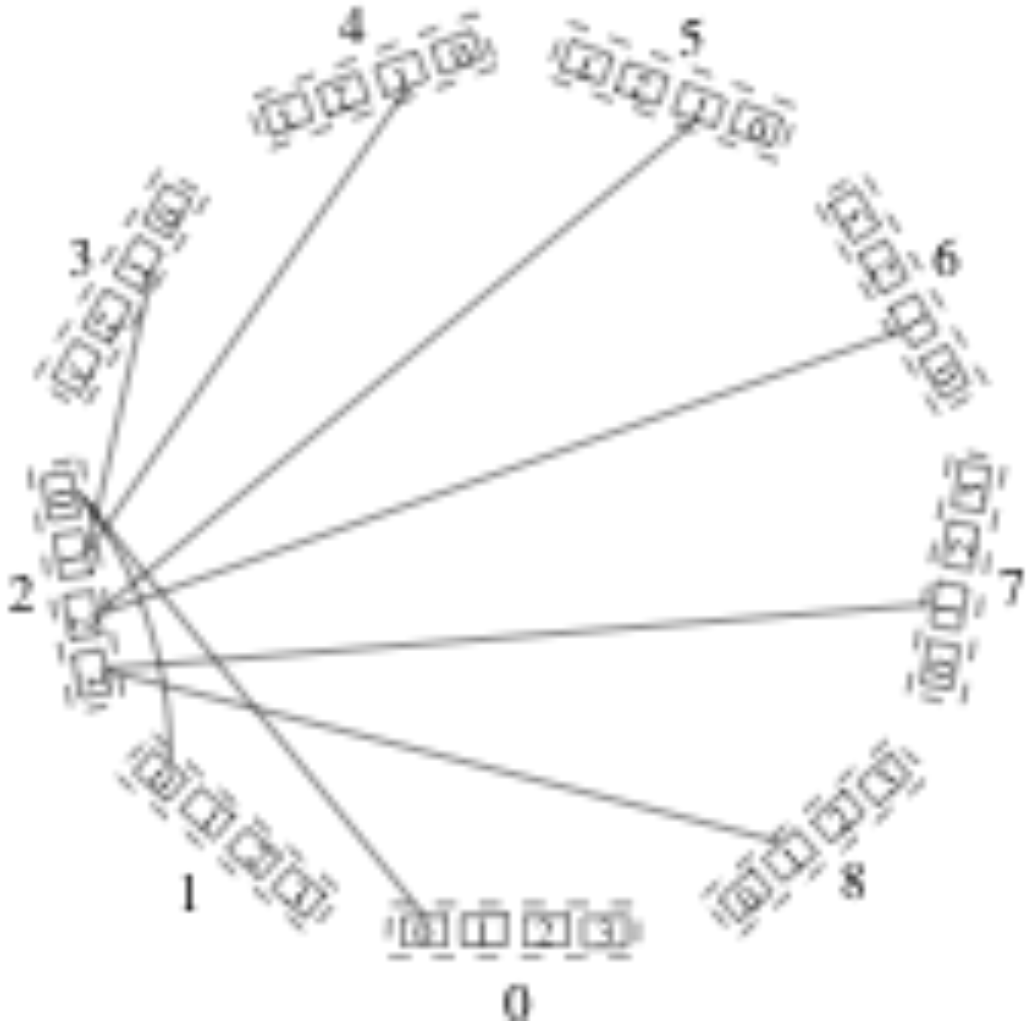
Three distinct global link arrangements

Absolute 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 ≥ i



Relative arrangement

Port k connects $(k+1)^{st}$ group

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

Circulant-based arrangement

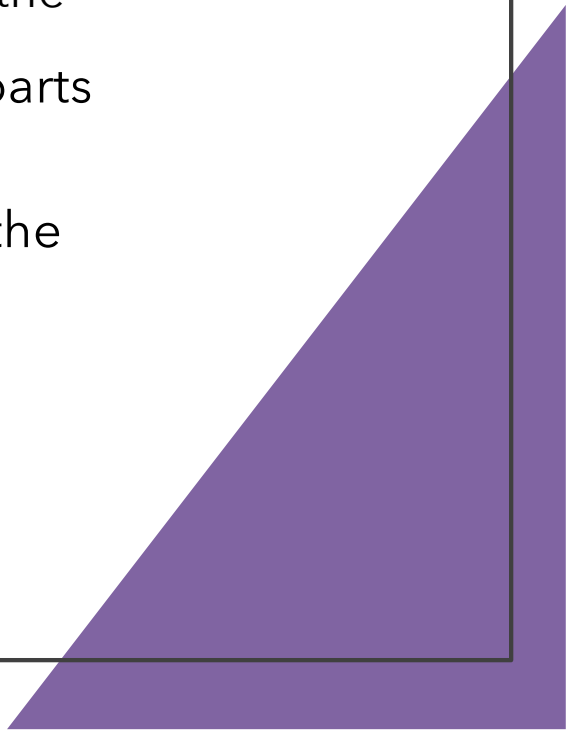
Port 0 connects to next group
Port 1 connects to previous group
Port 2 connects to group 2 ahead
Port 3 connects to group 2 behind
...

Equivalently, port k of group i connects to group

(i+k/2+1) mod g if k is even
(i-k/2-1) mod g if k is odd

Bisection bandwidth

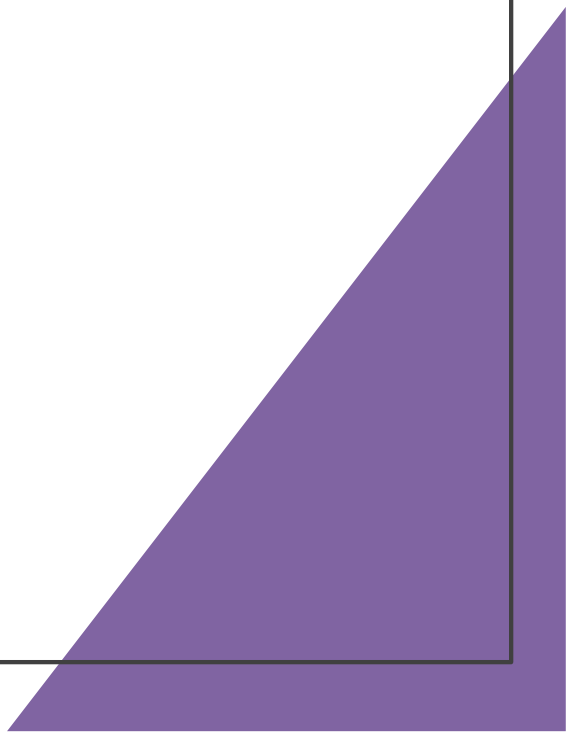
- Bandwidth caries linearly with α, with the rate depending on the number of global links crossing the cut
- Minimum bandwidth between two equal-sized parts of the system
  - Bandwidth for a particular bisection is the number of edges crossing from one part to the other
  - Minimize this over all bisections
- Tries to measure worst-case communication bottleneck in a large computation



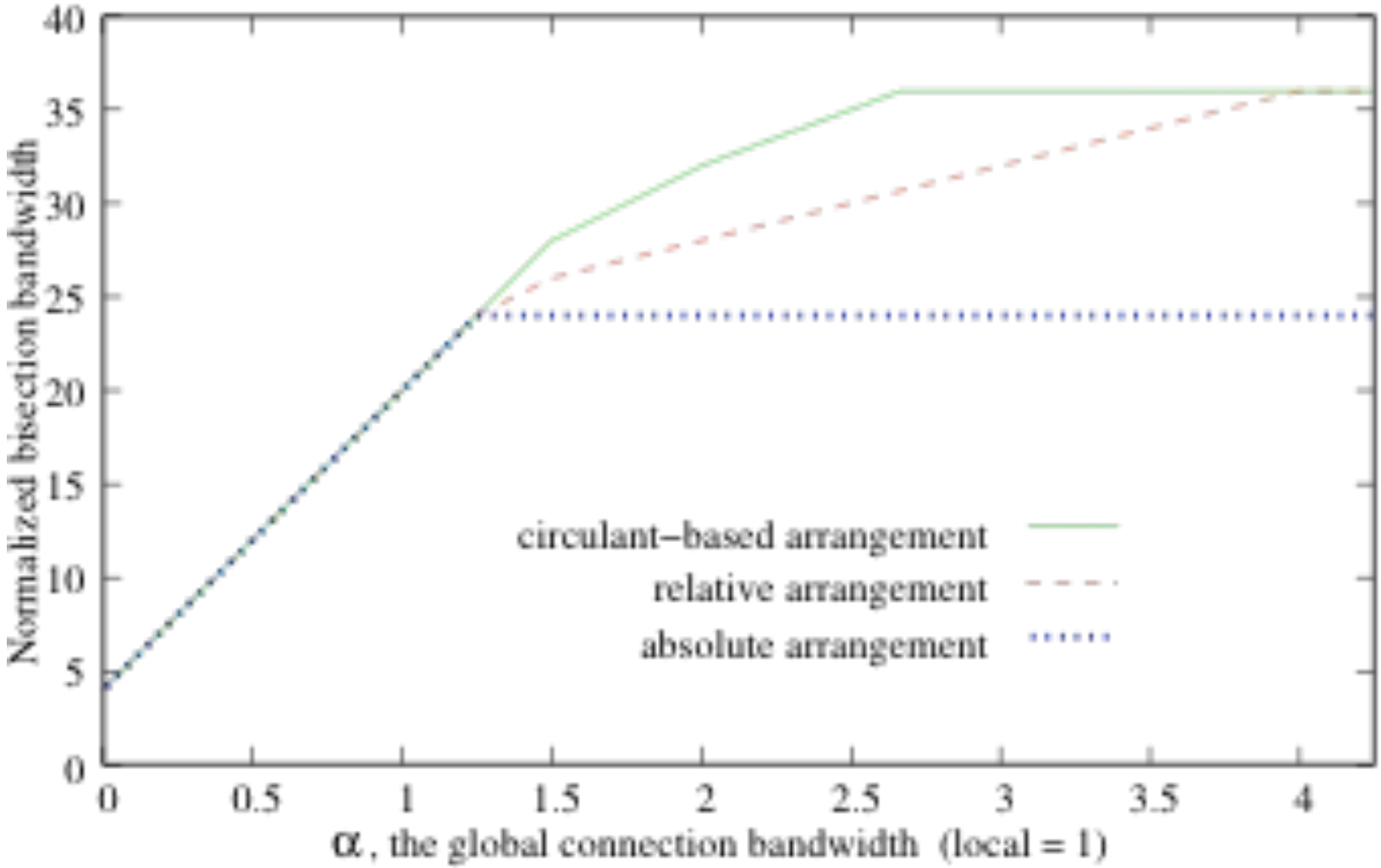
Initial exploration

- Small Dragonfly system
  (p,4,2): 4 switches per group
  2 global links per switch
  Has 36 switches

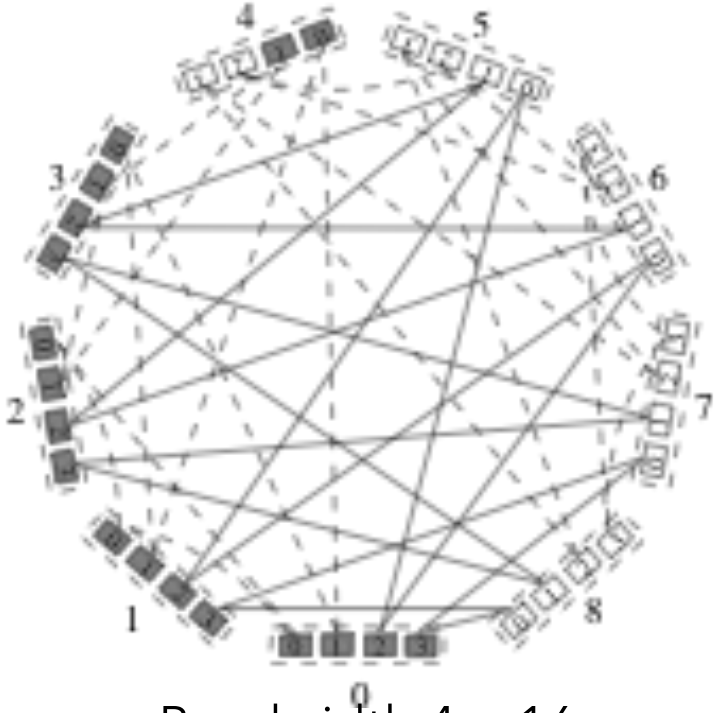
- Treat types of edges separately
  - local edges have bandwidth 1
  - global edges have bandwidth α



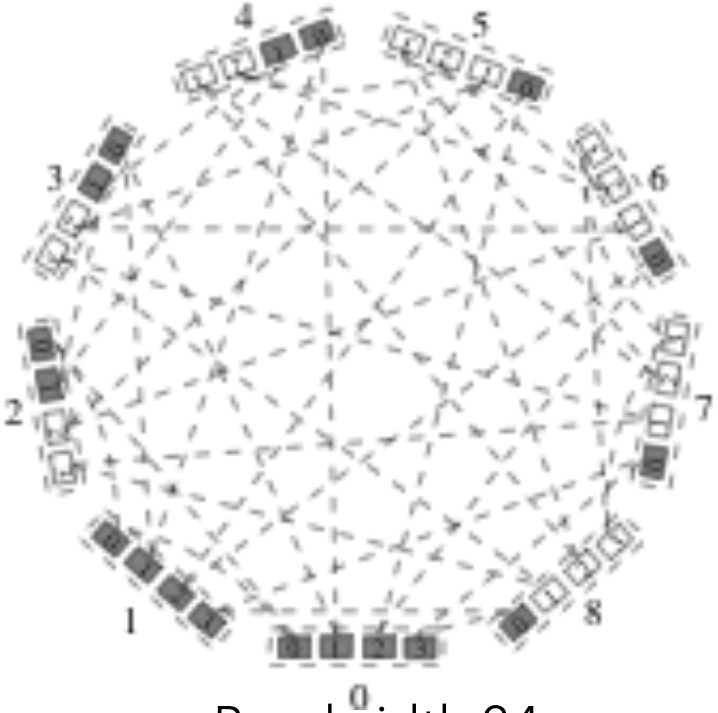
Bisection bandwidth as function of α



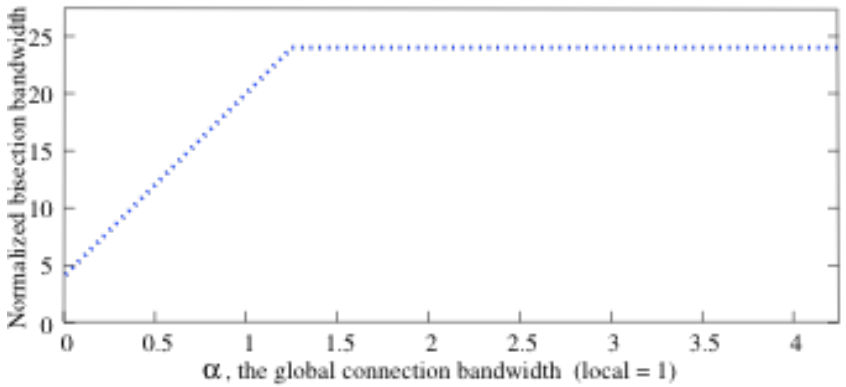
Min-bandwidth cuts for absolute arrangement



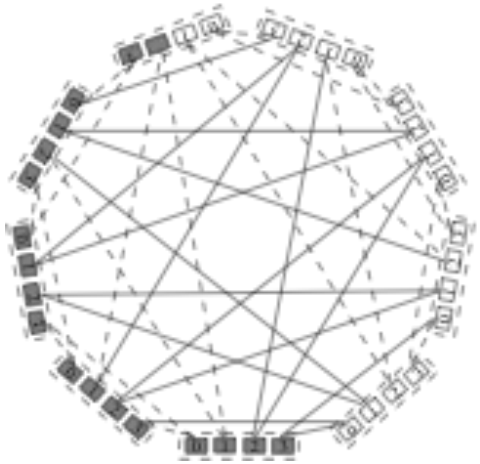
Bandwidth 4 + 16α



Bandwidth 24



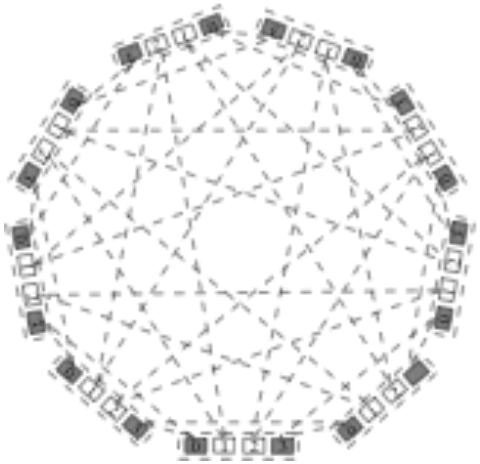
Min-bandwidth cuts for relative arrangement



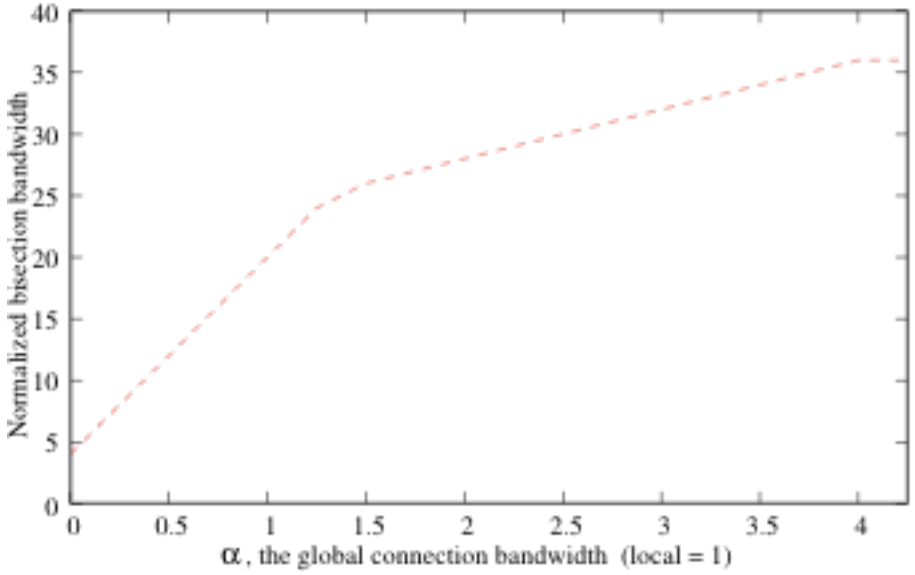
bandwidth 4 + 16α

bandwidth 14 + 8α

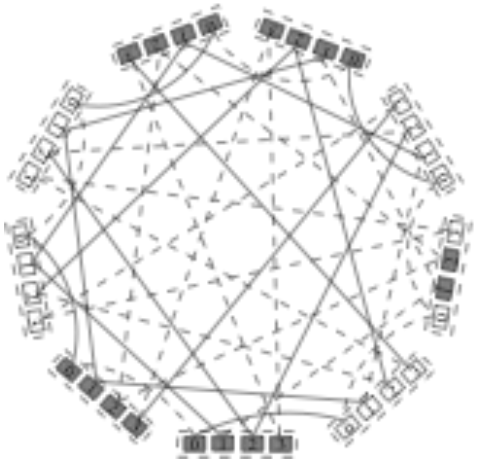
bandwidth 20 + 4α



bandwidth 36

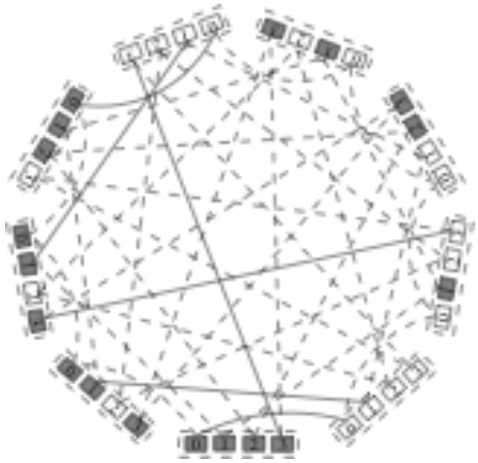


Min-bandwidth cuts for circulant-based arrangement

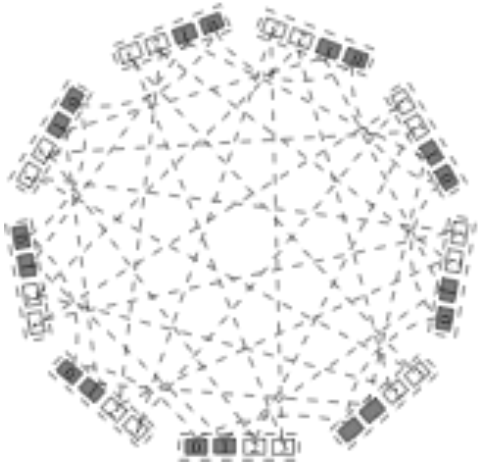


bandwidth 4 + 16α

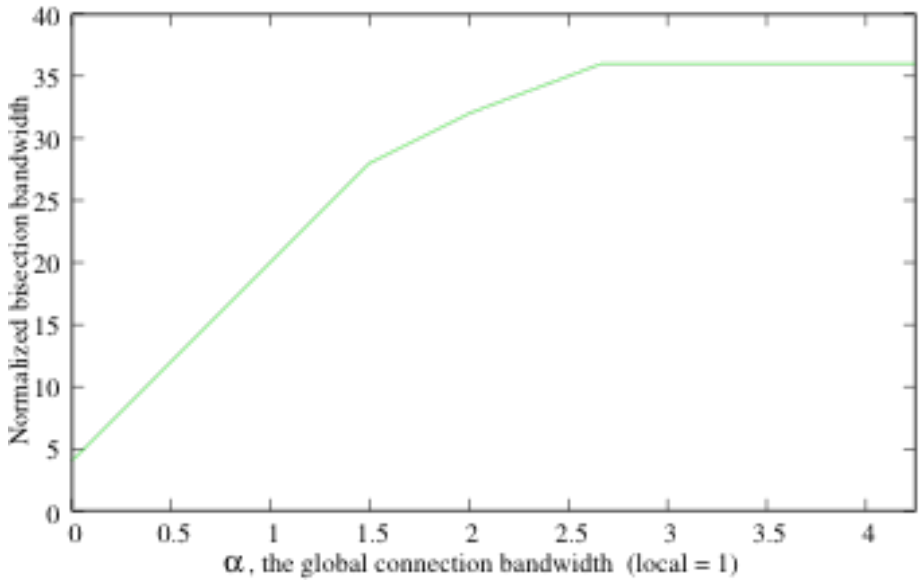
bandwidth 16 + 8α



bandwidth 20 + 6α



bandwidth 36

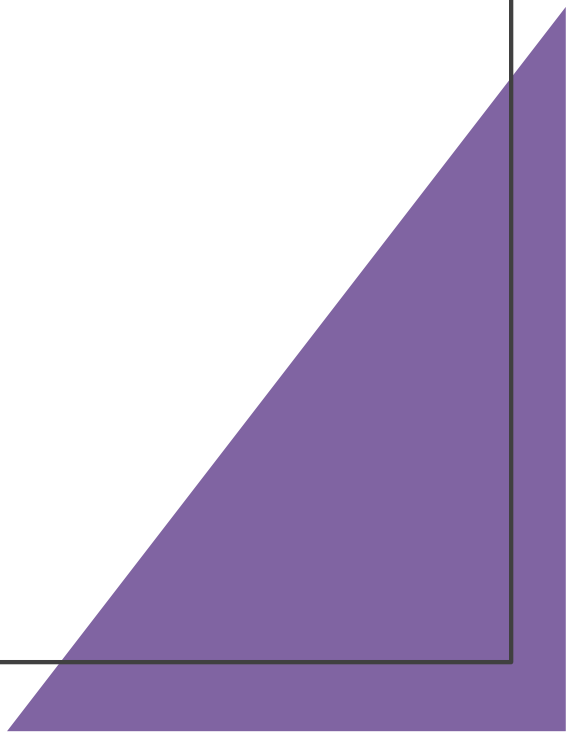


Observations from (p,4,2)

- In terms of bisection bandwidth:

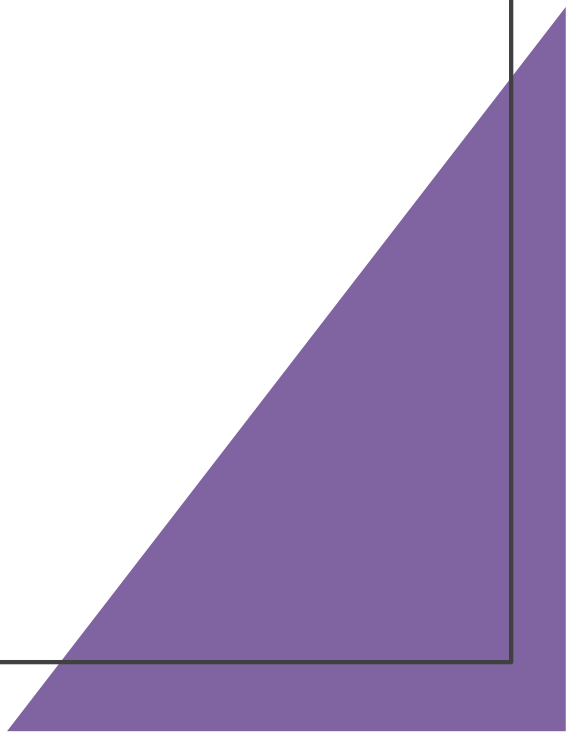
  Absolute ≤ Relative ≤ Circulant-based

- For all three arrangements, maximum bisection bandwidth is bounded



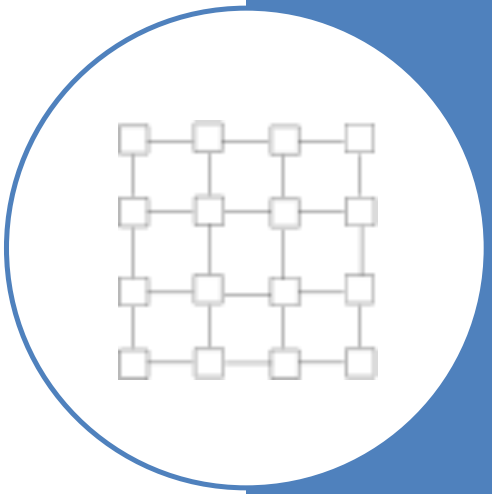
When bisection bandwidth is bounded

- Circulant: a is even
  - $(a/2)^2 g$
- Relative: a is a multiple of 4
  - $(a/2)^2 g$
- Absolute: a is a multiple of 4
  - $\theta(1)$



Task mapping

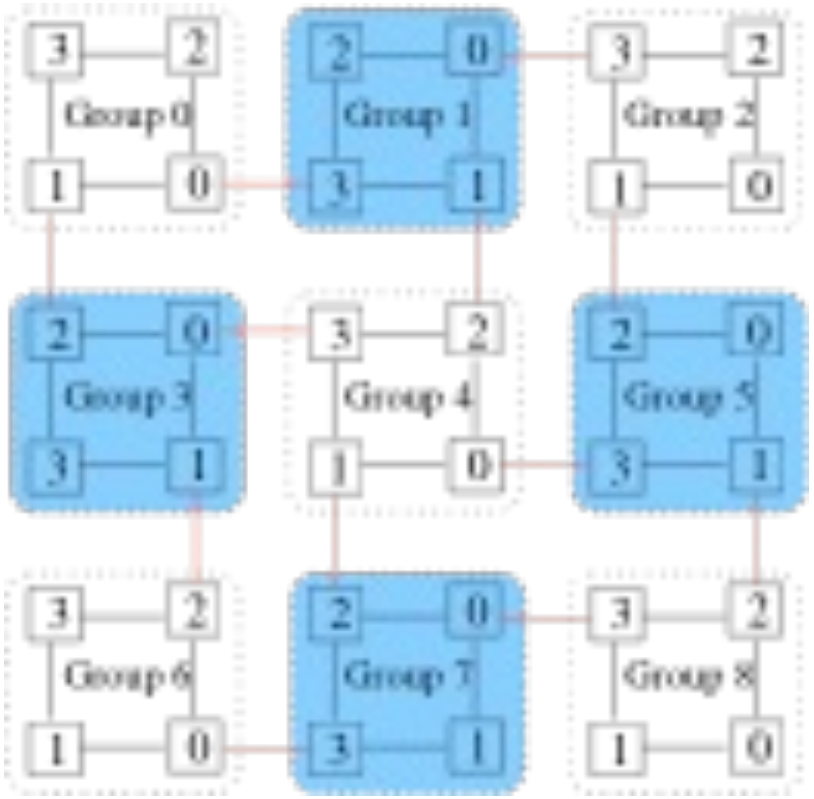
- Hotspots can still occur
- Assignment of a job's tasks to the processing elements assigned to that job
- Quality task mapping improves application performance by improving bandwidth utilization
- Good mappings are possible under each, the relative arrangement allows mappings with a regularity that will simplify code and improve generalizability



Criteria for good mapping

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

Final Thoughts

- On small graph, for bisection bandwidth:

  Absolute ≤ Relative ≤ Circulant-based

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

- For mapping stencils, Relative gives much more regular mappings

Thank You