Dragonfly Interconnect Topology

Emily Hastings — Anda Xu

Image from http://markbw.com/id68.html

Introduction

A ***supercomputer*** is a computer that performs at or near the currently highest operational rate for computers.

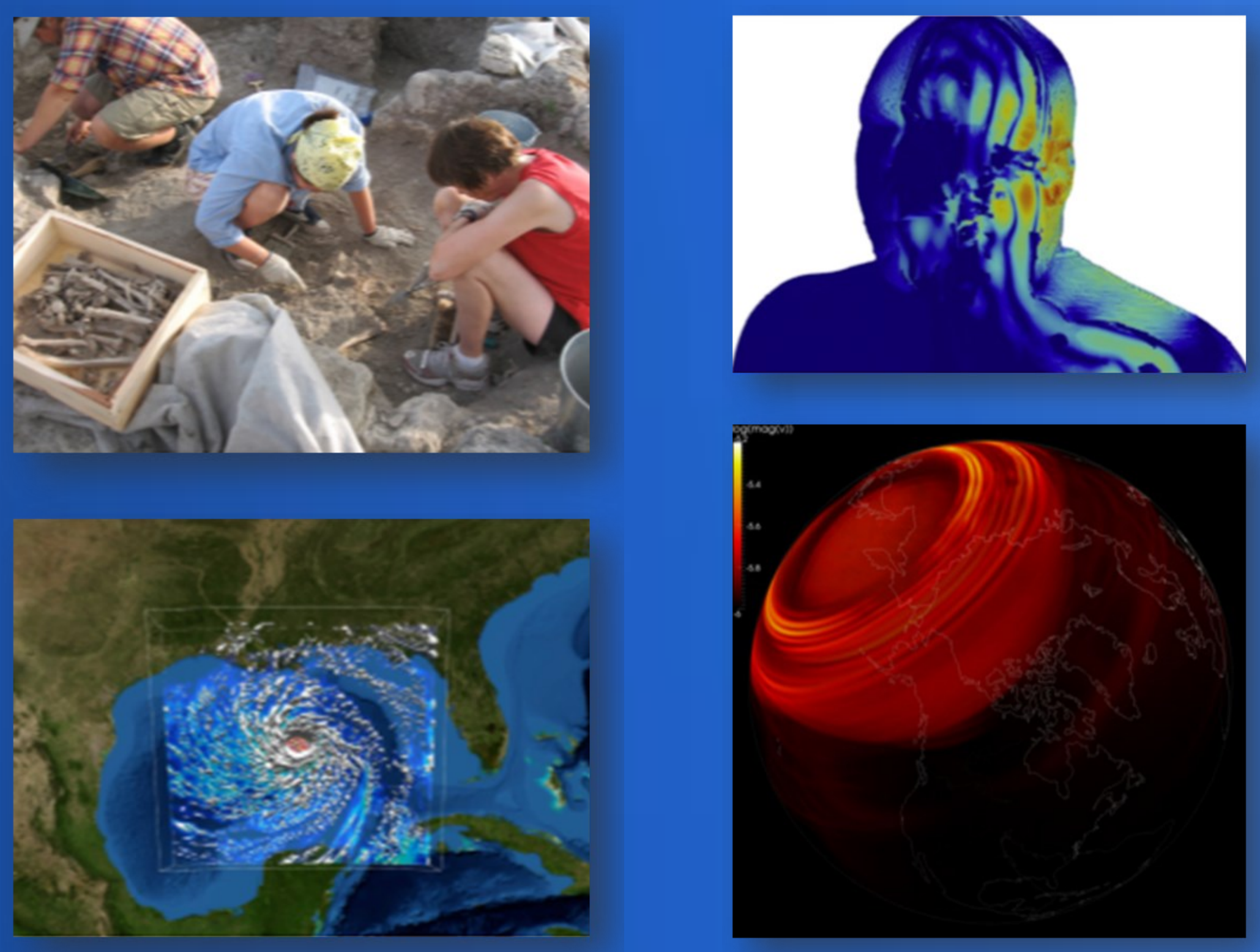
Image from http://www.webestigate.com/2010/11/13/u-s-building-next-powerful-supercomputers/ibm_supercomputer/

Where do we apply them?

- Preserving data at archeological sites
- Modeling the human body
- Weather prediction
  - Hurricanes
  - Earthquakes
  - Disaster relief

Why we care about High Performance Computing:

Because of remarkable advances in computer technology, scientists now have a new problem solving tool, a supercomputer. Supercomputers supported by high powered graphics workstations are extremely valuable resources for a wide range of scientific investigations. High performance computing is becoming increasingly more important to many scientific and engineering disciplines, so it is important for educators to prepare future generations to be ready for that growing demand.



Images from http://www.tacc.utexas.edu/documents/13601/498275/DigitizingPast.pdf, http://www.tacc.utexas.edu/documents/13601/498275/WirelessWorries.pdf , http://www.tacc.utexas.edu/documents/13601/498275/CenteroftheStorm.pdf , http://www.tacc.utexas.edu/documents/13601/498275/EmergencyComputingJapan.pdf

Our Research:

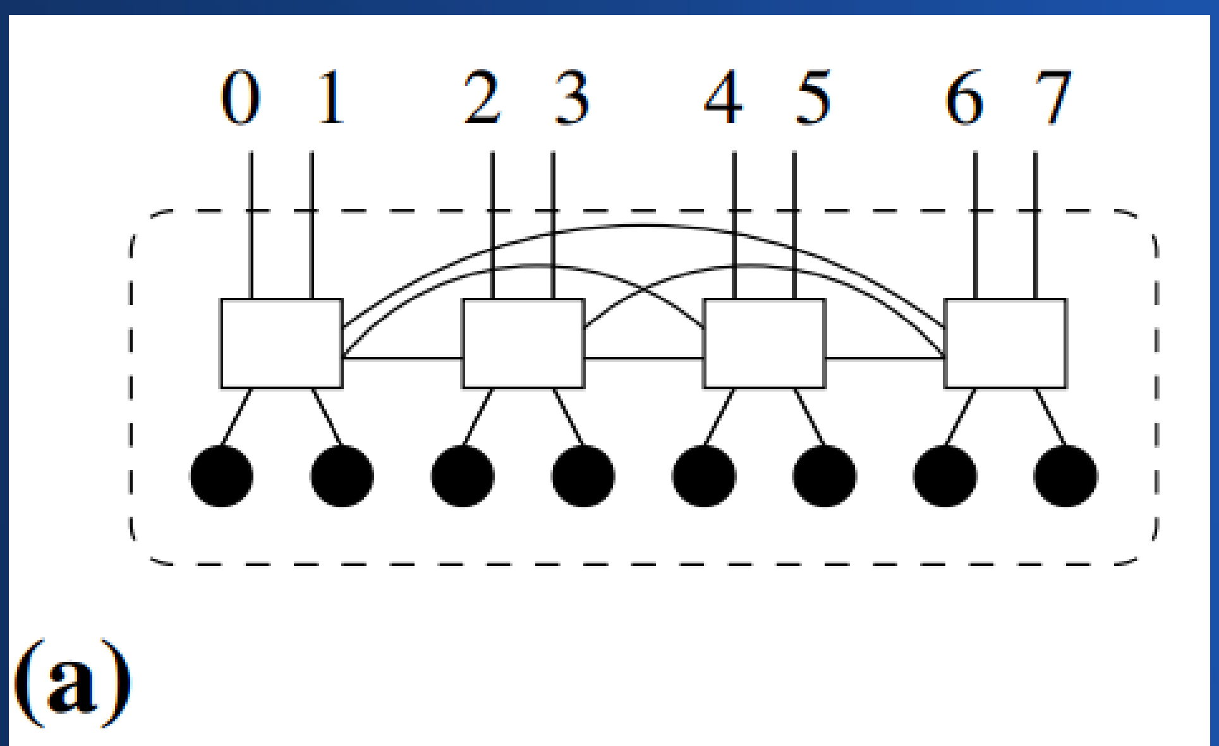
Over the course of last summer, our team researched the recently-developed Dragonfly topology for supercomputers, ran tests on the performance of the various cabling methods, and investigated task mapping on this new machine. We would like to thank our faculty mentor Prof. David Bunde and support from the Richter program and contract 899808 from Sandia National Laboratories. 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.

What Is Dragonfly?

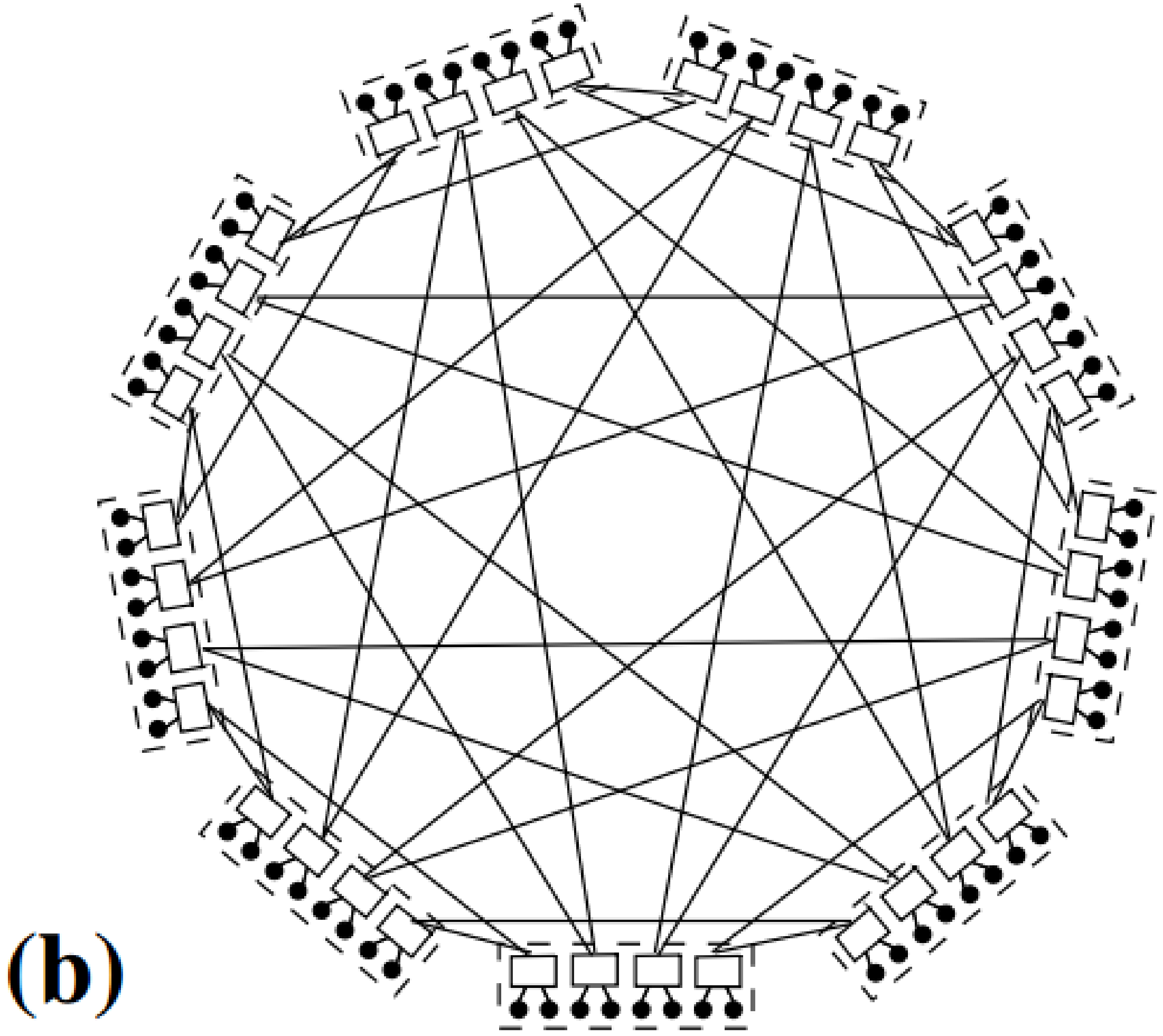
Dragonfly is an interconnect topology for supercomputers introduced in 2008 by scientists at Stanford University and Cray, Inc.

In a Dragonfly machine, nodes (processors) are connected to routers, and routers are connected in groups of equal size. Within each group, all routers are connected to each other, and as such can act as a single "virtual router" that is much larger than its components (a). These virtual routers are then connected to each other, with one edge between each pair of groups (b).

With this configuration, direct communication between any pair of nodes can occur in no more than three hops (locally to router connected to destination group, globally across inter-group edge, and locally within destination group).



(a)



(b)

Advantages Over Other Topologies:

- Reduced latency (time to send a message) and hop counts
- Scalable
- Design minimizes necessary global channels, reducing cost significantly

Dragonfly Interconnect Topology

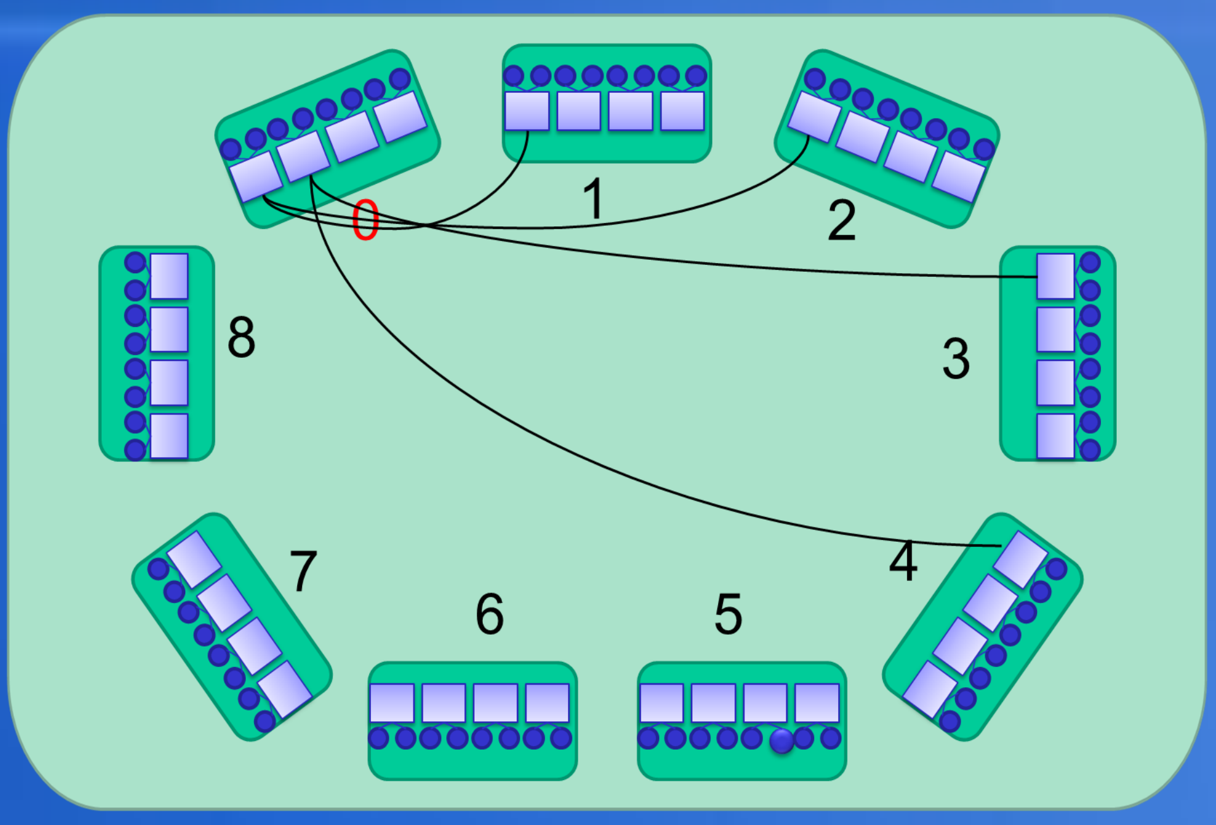
Emily Hastings — Anda Xu

Image from http://markbw.com/id68.html

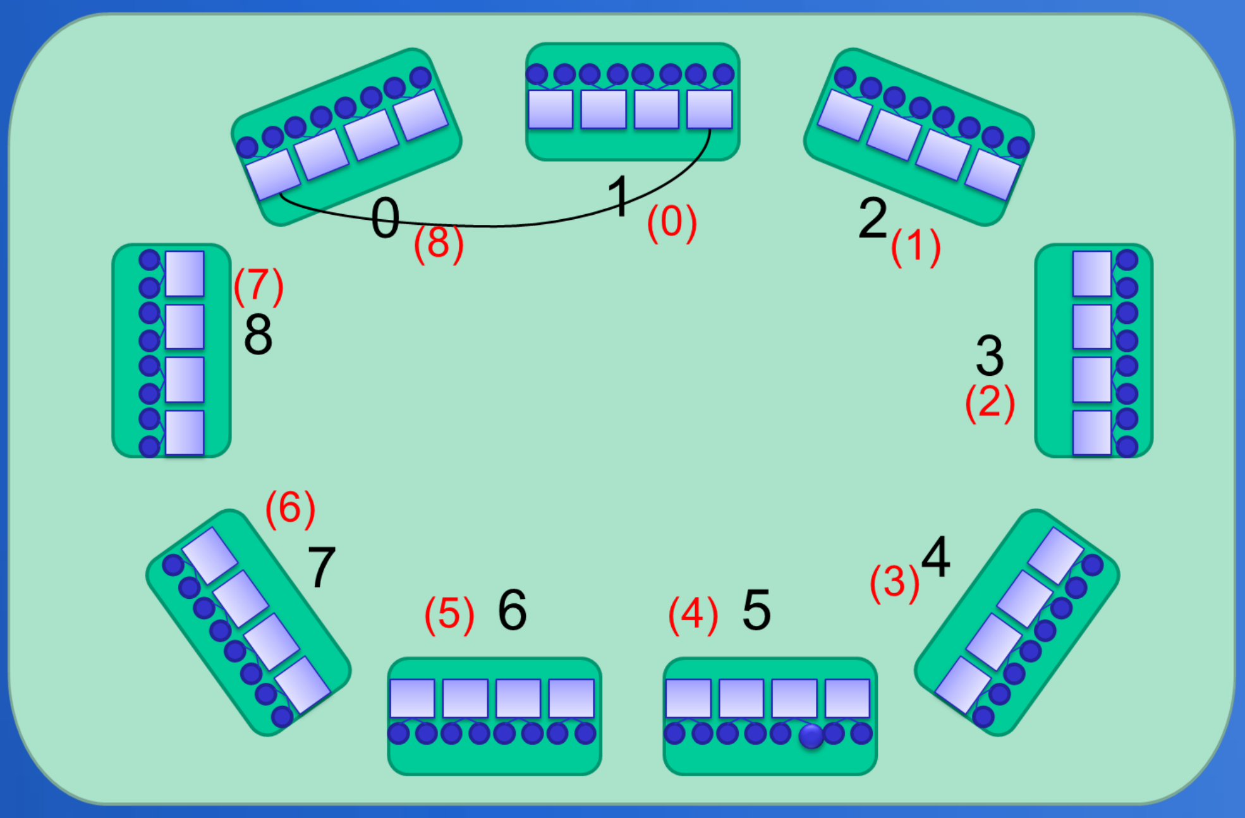
Cabling

The ***absolute cabling scheme*** conceptually connects port i of each group's virtual switch to group i. Once we ignore the unnecessary connection that this would give each group to itself, this scheme connects port i of group j to group i if i < j and to group i + 1 otherwise. The absolute cabling scheme is depicted in Figure (a)

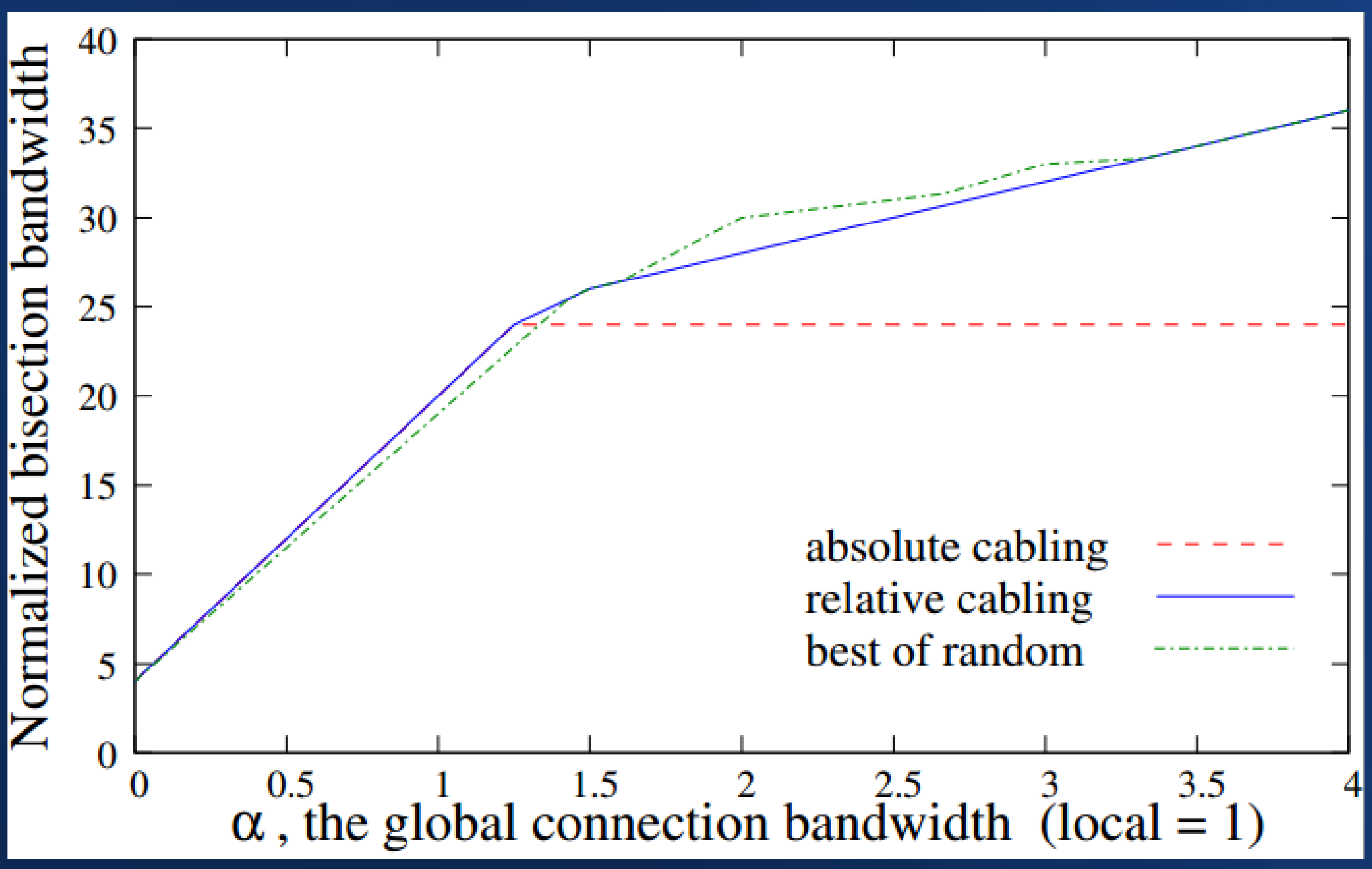
In the ***relative cabling scheme***, each group uses its port 0 to connect to the "next" group, its port 1 to two groups down, etc. More formally, port i of group j (i in {0, . . . , g − 2} and j in {0, . . . , g }, with g = ah+1 being the number of groups) connects to group (i+1+j) mod g. This is the absolute cabling scheme with each group using its own numbering for the groups so it takes the role of group 0. The relative cabling scheme is depicted in Figure (b).



(a)

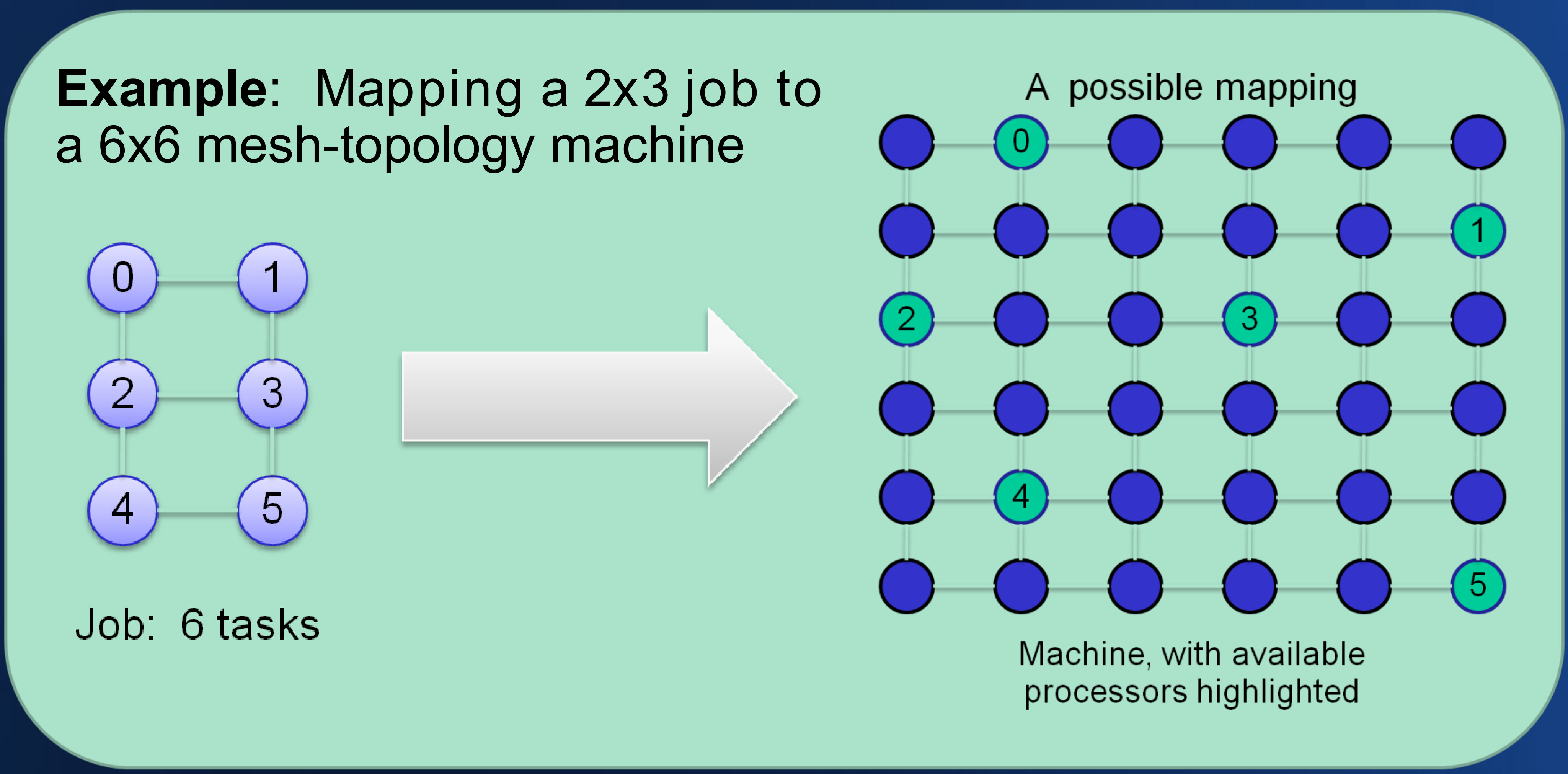


(b)



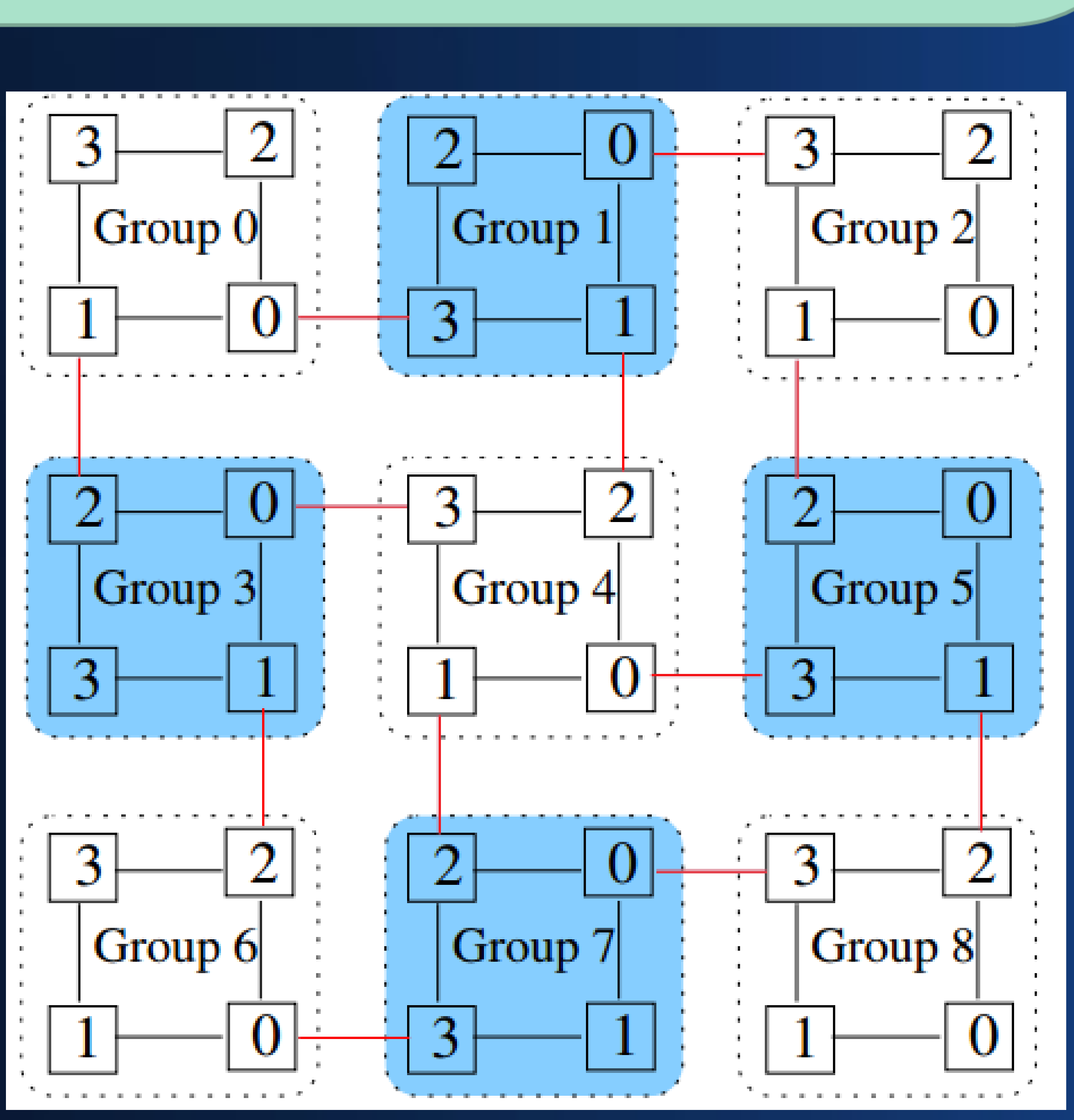
To compare these cabling schemes, we used ***bisection bandwidth***, the minimum bandwidth between 2 equal-sized parts of the system. This common quality measure tries to capture the worst-case communication bottleneck in a large computation. Overall, the relative cabling scheme seemed to perform better than the absolute.

Task Mapping



Example: Mapping a 2x3 job to a 6x6 mesh-topology machine

6x6 job, divided into 9 2x2 sub-jobs that correspond with Dragonfly groups. Tasks are numbered according to the Dragonfly router they have been mapped to, and inter-subgraph connections served by direct global connections on the machine are shown in red.



Task mapping is the process of assigning the tasks of a given job to available processors in the machine. The edges in the job graph represent tasks that need to communicate with each other, so in order to task-map efficiently, these nodes need to be close together on the machine. On Dragonfly, all nodes are close together, but mappings can be optimized so that many of these communication paths need only 1 hop on the machine.

Part of our research dealt with trying to find a general pattern for such mappings, by dividing jobs into sub-jobs that were the same size as the Dragonfly groups and assigning each router a task. This way, all edges within a sub-job are served by local connections, and many of the inter-group edges use direct global connections. For a relative-cabled Dragonfly as shown above, a 6x6 job can be mapped optimally using a checkerboard pattern of two router arrangements. We found that for, larger machines and jobs, this pattern can be generalized to a single repeated arrangement.

In the future, we hope to investigate applying this generalization to different kinds of jobs, including three-dimensional meshes.