Task Mapping for Emerging Network Topologies

(CNS-1423413)

David P. Bunde dbunde@knox.edu

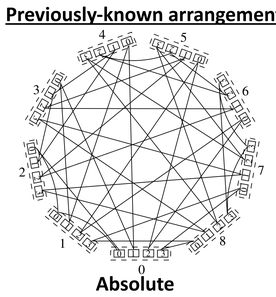
Different link arrangements for Dragonfly networks

Dragonfly networks (Kim et al., ISCA 2007) attach compute nodes to switches organized in a 2 level hierarchy. Within a *group*, all switches are connected using local (copper) links. Between each pair of groups there is exactly one global (optical) link. A (p, a, h)-Dragonfly is defined by

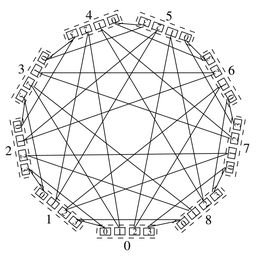
p the number of nodes connected to a switch,
a the number of switches in a group, and
h the number of global links on a switch.

We looked at the *global link arrangement*, the specific member of each group connected to each other group. Surprisingly, this significantly affected the *bisection bandwidth* (the weight of edges crossing the min-bisecting cut); local links have weight 1, global links have weight α.

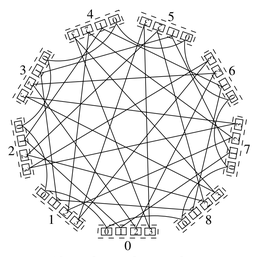
Previously-known arrangements:



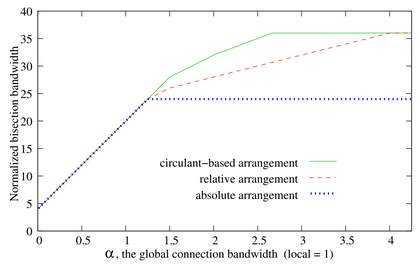
Absolute
i^{th} link from a group goes to group i (skipping own group)



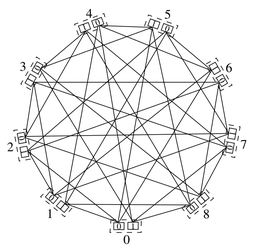
Relative
i^{th} link from a group goes i groups ahead



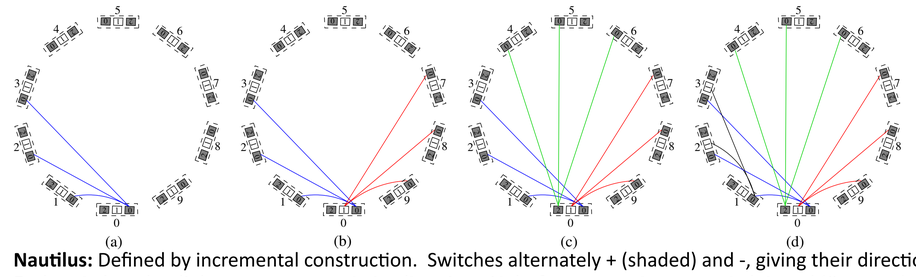
Circulant-based
Similar to relative, but links go alternately ahead and back



New arrangements:



Helix: Each switch send half links forward and back, plus advances the switch number



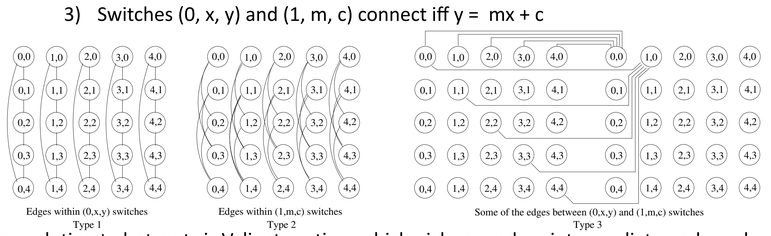
Nautilus: Defined by incremental construction. Switches alternately + (shaded) and -, giving their direction. Each switch makes links to next groups in its direction. Links from group j are to switch j mod (# groups).

Main finding: Arrangement matters. New arrangements slightly better, but in configuration-dependent way

Improved Valiant Routing for Slim Fly networks

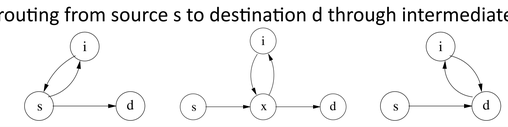
Slim Fly networks (Besta and Hoefler, SC 2014) are based on MMS graphs (McKay et al., *J. Combinatorial Theory Series B*, 1998), an attempt to maximize the number of vertices for graphs of diameter 2.

- Choose a prime q and find the primitive element of the field of elements mod q
- Populate X and X' with powers of it (in a specific way depending on q)
- Create two q × q grids of switches with coordinates (0, x, y) and (1, m, c). Edges are specified as follows:
  1) Switches (0, x, y) and (0, x, y') connect iff y – y' is in X
  2) Switches (1, m, c) and (1, m, c') connect iff c – c' is in X'
  3) Switches (0, x, y) and (1, m, c) connect iff y = mx + c

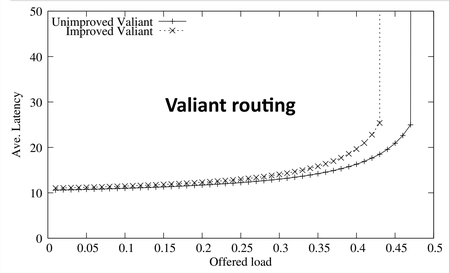


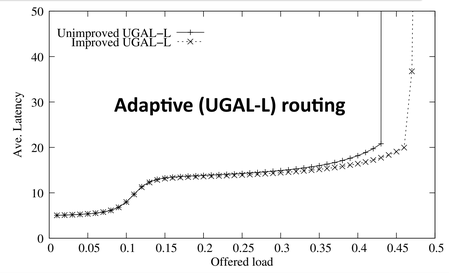
A common solution to hotspots is Valiant routing, which picks a random intermediate node and routes to it before routing to the destination. Adaptive strategies do this only used when congestion is detected.

On Slim Fly networks, routing from source s to destination d through intermediate i can give wasteful cycles:



Finding: Eliminating choices that result in cycles improves network performance for adaptive routing.

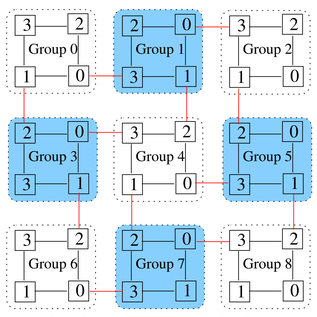




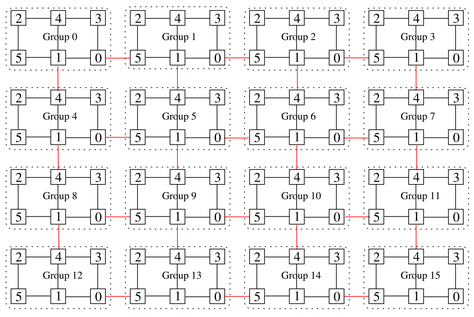
Task mapping for Dragonfly networks

Task mapping is the assignment of tasks (e.g. MPI ranks) to compute nodes. Placing communicating tasks near each other has been previously shown to improve performance on mesh systems.

We looked at ways to map stencil jobs onto Dragonfly systems so that Dragonfly groups process contiguous submeshes and a node on the boundary of each submesh was adjacent to the group processing neighboring tasks. (In the figure, tasks are laid out by their position in the job and labeled by the mapping. Groups are delimited by dashed lines and inter-group links are shown in red.)



6 × 6 job on a (p, 4, 2)-Dragonfly



12 × 8 job on a (p, 6, 3)-Dragonfly

Finding: Evaluation in progress, but seems to improve performance w/ minimal routing and interference to other jobs w/ Valiant routing

Undergraduate Collaborators

Previous years: Michael Anderson, Logan Ayers, Madison Belka, Harry Carpenter, Myra Doubet, Deyu Han, Emily Hastings, Hai Le, Sofia Meyers, Rosemary Momoh, David Rincon-Cruz, Kim Santos, Marc Spehlmann, Zhaofeng Wang

New this summer: Shogo Akiyama, Alex Brooks, Ink Chinavinijkul, Ryland Curtsinger, Jacob Newcomb, Rishav Sharma, Lingzhi Xi

Main publications

[1] E. Hastings, D. Rincon-Cruz, M. Spehlmann, S. Meyers, A. Xu, D. Bunde, V. Leung, "Comparing global link arrangements for Dragonfly networks," in *Proc. IEEE Cluster*, pp. 361-370, 2015.

[2] M. Belka, M. Doubet, S. Meyers, R. Momoh, D. Rincon-Cruz, D. Bunde, "New link arrangements for Dragonfly networks," in *Proc. 3rd IEEE International Workshop on High-Performance Interconnection Networks in the Exascale and Big-Data Era (HiPINEB)*, 2017.

[3] D. Han, Z. Wang, and D. Bunde, "Improving Valiant Routing for Slim Fly Networks," to appear in *Proc. Tenth International Workshop on Parallel Programming Models and Systems Software for High-End Computing (P2S2)*, 2017.

This work was partially supported by NSF CNS-1423413, the Paul K. and Evalyn Elizabeth Richter Memorial Funds, and Sandia Laboratories contract 899808. 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.

KNOX COLLEGE