Comparison study of scalable and cost-effective interconnection networks for HPC

Pablo Fuentes, Enrique Vallejo, Carmen Martínez, Marina García, Ramón Beivide
Department of Electronics and Computing
University of Cantabria
pablo.fuentes@alumnos.unican.es, {enrique.vallejo, carmen.martinez, marina.garcia, ramon.beivide}@unican.es

***Abstract*—This work attempts to compare size and cost of two network topologies proposed for large-radix routers: concentrated torus and dragonflies. We study and compare the scalability, cost and fault tolerance of each network. On average, we found that a concentrated torus can be a cost-efficient option for middle-range networks.**

***Index Terms*—concentrated torus; network topology; dragonfly;**

I. Introduction

In recent years multiple research fields have increased their computing requirements. Bioengineering (protein folding), Defense (cryptanalysis), or Meteorology (weather forecasting) are examples of this trend. There exists a general effort for pursuing the Exaflop (10^{18} floating point operations per second) [1], [2], [3]. The importance of the interconnection network increases with the system size. Since it can significantly condition performance and power consumption, new topologies are under study. The *Dragonfly* is an example of this search for a scalable and high performance network [4].

Despite this need for highly scalable networks, cost-efficiency is a restrictive factor. In this context, the design of many supercomputers is strongly constrained by both cost and power consumption. The use of *commodity* technology lowers the deployment costs of the systems. The Gordon Supercomputer [5], from San Diego Supercomputer Center, currently reaches 48^{th} position in Top500 list using only commodities, both in processor and network and a concentrated Torus topology. Such topology has been studied before in [6].

Network cost strongly depends on routers and longest wires, which must be implemented on optic fiber. However, these dependencies can vary relying on the amount of wires needed. Concentrated tori may happen to be a possibility to increase network performance and size while keeping costs below a reasonable limit. This is due to the lack of optic cables for the interconnection.

In a concentrated torus with D dimensions every router is attached to p compute nodes and to other $2D$ routers. Network connectivity is improved by assigning t physical wires to each topological link between routers (where t is called *"trunking factor"*), thus increasing the amount of traffic the network can cope with.

Bisection BandWidth (BBW) is a metric typically employed with regards to network dimensioning. BBW is defined as the aggregate bandwidth of a minimum cut which divides the network in two equal halves. Assuming a uniform traffic pattern, BBW should be enough to carry the traffic generated by the nodes in one side with destination to the other partition. If BBW is lower than such aggregated traffic, the network is *oversubscribed*. Oversubscription lowers network costs, but makes the network a performance bottleneck unless locality is heavily exploited. In this work we will consider networks *without* oversubscription. In such case, the concentrated torus parameters (D, t and the number of routers per dimension) restrict the maximum number of compute nodes attached to routers and thus the system scalability. In the present paper we compare network scalability, cost and connectivity for both concentrated torus and dragonflies.

II. Network scalability

We focus our work on symmetric torus, where the number of routers in any network dimension is the same, denoted by r. The routers on each dimension are connected as a ring, and in that case the following relation holds to prevent oversubscription

$$\frac{p}{2}\frac{r}{2} \leq 2t \Rightarrow pr \leq 8t.$$

Therefore, the maximum number of nodes N will be upper bounded by

$$N = pr^D = \frac{(pr)^D}{p^{D-1}} \leq 8t\left(\frac{8t}{p}\right)^{D-1}.$$

The largest configuration is obtained with the minimum p (no concentration) and maximum t (maximum trunking), which is obvious since wider links increase BBW. However, a sensible restriction is to assume that $p \geq t$ to avoid links overdimension with respect to computing nodes. In that case, the maximum size is obtained when the following expressions fulfill, where k denotes the router degree.

$$N = 8^D t \text{ and } t = p = \left\lfloor \tfrac{k}{2D+1} \right\rfloor$$

In comparison, a balanced dragonfly (as described in [7]) reaches about $\frac{k^3}{16}$ nodes. Figure 1 presents the maximum achievable network size with different router degree, using both balanced Dragonflies and concentrated tori with 1 to 4 dimensions D. In this chart we restrict to $D < 5$ since larger number of dimensions make layout very complex; in fact, machines using $D \geq 5$ (such as BlueGene/Q or the K-Computer) use asymmetric designs to be able to wrap multiple

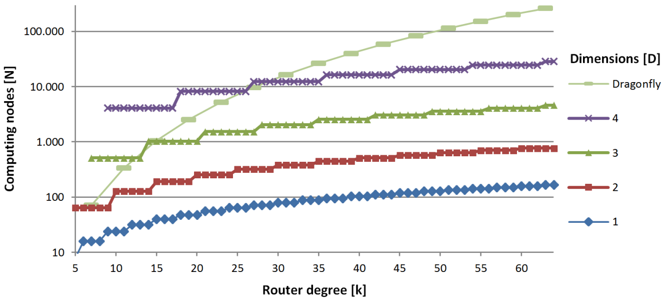

Fig. 1. Maximum scalability for a router size

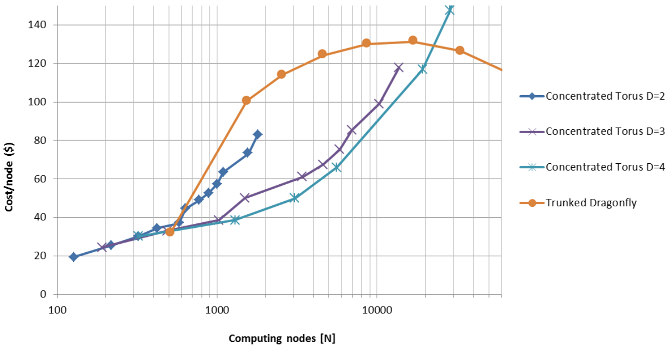

Fig. 2. Cost comparison using router size 64

dimensions within a single rack. Router degree is ranged below 64 ports, which is achievable with current or near future network fabrics for commodities. Figure 1 shows that 4D concentrated tori can reach a higher number of computing nodes without oversubscription when the router degree is below 27 ports.

III. Cost comparison

Network cost distinguishes between two components. Using the component prices from [8] we estimate that each router has an estimated cost of 390$. The links connecting the nodes to the routers will be 1m length (4,53$). All links in a concentrated torus can use electric wires. The length of links between routers depends on network size considering a folded torus layout and the cabinet distribution. Cost of Dragonfly is calculated similarly to [7], using trunking between groups of 512 nodes. Note that in large sizes the cost diminishes because there are unused global ports due to the rounding in the trunking factor calculation. Dragonfly networks differ to tori because they have some very long optical links (with a price of approximately 220$ per cable). Figure 2 displays the network cost per node built from 64-port routers. Note that for a size bellow 20.000 computation nodes it is cheaper to build a 3D or 4D concentrated torus using large-radix switches than a dragonfly.

IV. Router Connectivity

The probability of failures grows with the size of the computing system which makes fault tolerance an aspect of great importance in network design. As a first approach for evaluating a topology fault tolerance properties it is usual to take into account the node connectivity. Hence, in this section we consider the router connectivity in both concentrated torus and dragonfly topologies. Therefore, to compare both networks what we look is for the number of links which have to fail in order to completely disconnect a router.

In the case of concentrated D-dimensional torus, any router is connected by $2tD$ links to the $2D$ neighbour routers. Therefore, these number of links is exactly the connectivity of any router in the network, which corresponds to a proportion of $\frac{2D}{2D+1}$ to the total number of links. On the other hand, a similar calculation but for the dense dragonfly yields to a ratio of approximately $\frac{3}{4}$ of the links that have to be removed in order to disconnect completely one router. As a consequence, the toroidal topology provides for a greater router connectivity in any case. In particular, the ratio for concentrated 3D torus is $\frac{6}{7}$ while the one for the dragonfly is $\frac{3}{4}$ of the links.

V. Conclusions and Future Lines

In this paper we have explored a network topology for high-performance computing which we call concentrated torus. We have compared its size scalability and costs against dragonfly, which is a relatively novel topology expected to be both highly scalable and cost-efficient. Moreover, a first approximation to fault tolerance capacity of the toroidal topology has been made by calculating the router connectivity. As we have seen, the figures of merit that we have considered suggest that concentrated tori could be a good rival in medium-sized networks. Our future goals are to obtain performance charts via network simulation, using synthetic traffic loads and traces.

Acknowledgment

This work has been supported by the Spanish Ministry of Science under contracts TIN2010-21291-C02-02 and CONSOLIDER Project CSD2007-00050, and by the European HiPEAC Network of Excellence.

References

[1] K. Bergman, S. Borkar, D. Campbell, W. Carlson, W. Dally, M. Denneau, P. Franzon, W. Harrod, K. Hill, J. Hiller *et al.*, “Exascale computing study: Technology challenges in achieving exascale systems,” 2008.

[2] E. C. Joseph, S. Conway, C. Ingle, G. Cattaneo, C. Meunier, and N. Martinez, “A strategic agenda for european leadership in supercomputing: HPC 2020,” 2010.

[3] P. Dongarra, “Beckman et al.the international exascale software roadmap. volume 25, number 1, 2011,” *International Journal of High Performance Computer Applications*, pp. 77–83.

[4] J. Kim, W. Dally, S. Scott, and D. Abts, “Cost-efficient dragonfly topology for large-scale systems,” *Micro, IEEE*, vol. 29, no. 1, pp. 33–40, 2009.

[5] M. L. Norman and A. Snavely, “Accelerating data-intensive science with Gordon and Dash,” in *Proceedings of the 2010 TeraGrid Conference*, ser. TG ’10. New York, NY, USA: ACM, 2010, pp. 14:1–14:7. [Online]. Available: http://doi.acm.org/10.1145/1838574.1838588

[6] J. Navaridas and J. Miguel-Alonso, “Indirect cube: A power-efficient topology for compute clusters,” *Optical Switching and Networking*, vol. 8, no. 3, pp. 162 – 170, 2011.

[7] J. Kim, W. J. Dally, S. Scott, and D. Abts, “Technology-driven, highly-scalable dragonfly topology,” in *ISCA ’08*. Washington, DC, USA: IEEE Computer Society, 2008, pp. 77–88.

[8] J. Kim, W. J. Dally, and D. Abts, “Flattened butterfly: a cost-efficient topology for high-radix networks,” in *ISCA ’07*. New York, NY, USA: ACM, 2007, pp. 126–137.