Random Regular Graph and Generalized De Bruijn Graph with k-shortest Path Routing

Peyman Faizian Md Atiqul Mollah Xin Yuan
Department of Computer Science
Florida State University
Tallahassee, FL 32306
{*faizian, mollah, xyuan*}*@cs.fsu.edu*

Scott Pakin Michael Lang
Computer, Computational, and Statistical Sciences
Los Alamos National Laboratory
Los Alamos, NM
{*pakin, mlang*}*@lanl.gov*

***Abstract*—Random regular graph (RRG) has recently been proposed as an interconnect topology for future large scale data centers and HPC clusters. While various studies have been performed, this topology is still not well understood. RRG is a special case of *directed regular graph* (DRG) where each link is unidirectional and all nodes have the same number of incoming and outgoing links. In this work, we establish bounds for DRG on diameter, average k-shortest path length, and a load balancing property with k-shortest path routing, and use these bounds to evaluate RRG. The results indicate that RRG with k-shortest path routing is not ideal in terms of diameter and load balancing. We further consider the Generalized De Bruijn Graph (GDBG), a deterministic DRG, and prove that for most network configurations, GDBG is near optimal in terms of diameter, average k-shortest path length, and load balancing with a k-shortest path routing scheme. Finally, we explore the strengths and weaknesses of RRG for different traffic conditions by comparing RRG with GDBG.**

***Keywords*-network; topology; random regular graph; generalized De Bruijn graph; k-shortest path routing**

I. Introduction

The quest for low latency and high bandwidth interconnects has led to the idea of using random topologies for future extreme scale data centers and HPC clusters [1]–[4]. In a *random regular graph* (RRG), which is also known as the Jellyfish topology [2], all nodes have the same degree and links are bidirectional and connected randomly. RRG is a special *directed regular graph* (DRG) where links are unidirectional and the incoming and outgoing degrees of all nodes are the same. RRG offers low diameter and high bisection bandwidth. In addition, it also provides high path diversity such that using k-shortest path routing or its variants is sufficient to exploit the network capacity [2]. It has been shown that RRG achieves higher performance than similar cost fat-trees that are widely deployed in the current data centers and HPC clusters, which argues for using RRG as a design alternative for the future extreme-scale interconnects [2]–[4].

The RRG topology represents a class of high capacity topologies with sufficient diversity among short paths such that k-shortest path routing and its variants are sufficient to explore the network capacity. This is different from other recent interconnect proposals such as SlimFly [5] and Dragonfly [6] that rely on Valiant Load Balance (VLB) routing to exploit network capacity for some traffic patterns. For a network topology to be effective with k-shortest path routing, it must have the following properties.

- **Property 1** (large aggregate capacity): The topology must have high aggregate capacity. This property is historically measured by bisection bandwidth.
- **Property 2** (minimal resource usage with k-shortest path routing): The diameter and the average path length of the k-shortest paths between source-destination (SD) pairs must be small. By using shorter paths to communicate data, less network resources are needed on average to transmit each packet, which results in more efficient network resource utilization and higher aggregate throughput.
- **Property 3** (load balance with k-shortest path routing): For k-shortest path routing to fully exploit the network capacity, the paths must be evenly distributed over the network.

From the graph theory, the bisection bandwidth of RRG is close to the optimal with very high probability [7]. However, it is unclear how close RRG is to the optimal in terms of diameter, average k-shortest path length, and load balance with k-shortest path routing. Existing studies of RRG either compare its performance to other topologies such as fat-tree [2], [4] or performs a coarse grain upper bound performance estimation [3]. Although some studies have been performed, RRG is still not well understood. For example, it is unclear if there exists any type of traffic that can result in poor performance on RRG.

Since RRG is a special DRG, its performance is bounded by the bounds for DRG. In this work, we establish the bounds for DRG on diameter, average k-shortest path length, and a load balancing property with k-shortest path routing that is quantified by the maximum link load for all-to-all communication. These bounds are then used to evaluate RRG. The results show that for most values of k, RRG

performs well for average k-shortest path length, but not for diameter and load balancing.

We further study *Generalized de Bruijn Graph* (GDBG) [8], a deterministic DRG. We identify a form of k-shortest path routing, which we call hop-limited all path routing (ALLPATH(H)), for GDBG. We prove that with ALLPATH(H), for most network configurations, GDBG is near optimal in diameter, average k-shortest path length, and load balance. Hence, GDBG can achieve near optimal performance among all DRGs when the ALLPATH(H) routing is used. This indicates that GDBG is a highly effective topology with k-shortest path routing, and can be directly applied in many situations such as the logical topology for a distributed peer-to-peer network where k-shortest path routing is used. Moreover, GDBG can be used in simulation and modeling to evaluate the performance of other regular graph or directed regular graph topologies such as RRG to investigate their performance for some specific traffic patterns. We compare the performance of GDBG with that of RRG and identify the strengths and weaknesses of RRG. Our results refine the earlier conclusion in [3] that the performance of RRG is close to the optimal: we show that RRG is close to the optimal for diverse traffic where each switch communicates with many other switches; yet it is much less effective than GBDG (let alone the optimal) in dealing with concentrated traffic patterns such as the switch-level permutation or shift communication where each switch only communicates with one other switch. The contributions of this work include the following.

- We identify a near optimal topology for k-shortest path routing: the generalized De Bruijn graph. We prove that with hop-limited all path routing, a form of k-shortest path routing, generalized De Bruijn graph achieves near optimal performance for most network configurations.
- We show that RRG with k-shortest path routing is not ideal in diameter and load balancing, and identify the strengths and weaknesses of RRG with k-shortest path routing in dealing with different traffic patterns.

II. Background and related work

A. Random Regular Graph (RRG)

Random Regular Graph or the Jellyfish topology is proposed recently by Singla et al. [2] as a flexible and high-capacity topology for large scale interconnects. Unlike the deterministically constructed interconnections currently deployed in HPC and data centers, in an RRG, switches are interconnected randomly. In an interconnect with an RRG topology, switches have the same line rate and using the same number of ports to connect to other switches. The detailed method to build an RRG topology can be found in [2]. An interconnect with an RRG topology is characterized by three parameters: the number of switches (N), the radix of each switch(x), and the number of ports used by each switch to connect to other switches(r). In this paper, we focus on switch level topology with parameters N and r, and denote the topology as $RRG(N, r)$. RRGs have been shown to have good topological properties such as small diameter [9] and high bisection bandwidth [7].

B. Directed regular graph

Each link in an RRG topology can be represented as two unidirectional links, one in each direction. Hence, $RRG(N, r)$ belongs to the class of *directed r-regular graph* where links are unidirectional and each node has r incoming and r outgoing links. We will use $DRG(N, r)$ to denote an arbitrary directed r-regular graph with N nodes. Some special DRG topologies are considered in this paper: the Kautz graph [10] and the generalized De Bruijn graph (GDBG) [8], [11], [12]. The Kautz graph achieves the optimal diameter given a degree and a network size. GDBG is also a low-diameter and high-connectivity topology. While many results for these topologies have been obtained, our work focuses on k-shortest path routing properties of GDBG, which is a new contribution.

III. Bounds of DRG topological metrics

In order for a $DRG(N, r)$ that uses k-shortest path routing to achieve high performance, the topology must have the three properties listed in Section I. Since it is known that the bisection bandwidth for an RRG is likely to be close to the optimal [7], we will study Properties 2 and 3 and derive bounds on diameter, average length of k-shortest paths, and load balancing.

A. Diameter

Lemma 1 (diameter): The diameter of any $DRG(N, r)$ is at least

$$\lceil log_r(N(r-1)+1) \rceil - 1.$$

Proof: Consider any source node. From Moore's bound, we know that the node can reach at most $1 + r + r^2 + ...r^H$ nodes in H hops. Let D be the diameter of the $DRG(N, r)$. We have $1+r+r^2+...r^D \geq N$. Simplifying the inequation, we obtain $D \geq \lceil log_r(N(r-1)+1) \rceil - 1$. □

B. k-shortest path properties

There are different ways for k-shortest path routing to work. One is to fix k, and select k-shortest paths for each source-destination (SD) pair [2]. In this case, the average path length of k-shortest paths for all SD pairs is an important performance metric since it directly reflects the amount of resources used to send a packet. Note that the average shortest path length is a special case when $k = 1$. The second method is to fix a hop limit (H), and use all paths between an SD pair whose hop counts are no more than H (assuming that all SD pairs have at least one path that is H-hop or shorter) [4]. We will call this variant of k-shortest path routing *hop-limited all path routing* (ALLPATH). Later

in the paper, we will show that GDBG with a form of ALLPATH achieves near optimal performance for most network configurations. With ALLPATH, the number of short paths whose hop count are no more than H between each SD pair is an important parameter. Next, we will derive bounds on these two metrics for $DRG(N, r)$.

The average k-shortest path length is a direct measurement of the path quality for k-shortest path routing. In [13], it is shown that the lower bound on the average shortest path length of any r-regular network of size N is $D \geq \frac{\sum_{j=1}^{h-1} jr(r-1)^{j-1}+hR}{N-1}$ where $R = N-1-\sum_{j=1}^{h-1} r(r-1)^{j-1} \geq 0$ and h is the largest integer such that the inequality holds. Note that the result is for single path routing on an undirected r-regular graph. We extend this bound for k-shortest path routing on $DRG(N, r)$.

Lemma 2: For any $DRG(N, r)$, the average path length of all k-shortest paths between all SD pairs, $AKH(N, r, k)$, is

$$AKH(N, r, k) \geq \frac{\sum_{j=1}^{h-1} jr^j + hR}{k(N-1)}$$

Where

$$R = k(N-1) - \sum_{j=1}^{h-1} r^j \geq 0$$

and h is the largest integer such that the inequality holds.

Proof: Consider the paths from any source node, since the nodal degree is r, the number of 1-hop paths is at most r, the number of 2-hop paths is at most r^2, ..., the number of i-hop paths is at most r^i. Let h be the largest integer such that

$$R = k(N-1) - \sum_{j=1}^{h-1} r^j \geq 0,$$

Consider the $k(N-1)$ k-shortest paths from the source node to all other $N-1$ nodes. Among the $k(N-1)$ paths, at most r paths can be 1-hop paths, r^2 paths can be 2-hop paths, and r^i i-hop paths for all $i = 1, 2, ..., h-1$. The remaining $R = k(N-1) - \sum_{j=1}^{h-1} r^j$ paths have at least h-hops. Hence, the average path length for the $k(N-1)$ shortest paths from the source is at least

$$\frac{\sum_{j=1}^{h-1} jr^j + hR}{k(N-1)}$$

Since this applies to all source nodes,

$$AKH(N, r, k) \geq \frac{\sum_{j=1}^{h-1} jr^j + hR}{k(N-1)} \square$$

We will use $LK(N, r, k) = \frac{\sum_{j=1}^{h-1} jr^j + hR}{k(N-1)}$ to denote the lower bound of average k-shortest path length for any $DRG(N, r)$.

Given a fixed H, the following lemma gives the upper bound of the number of paths for each SD pair whose length is no more than than H in $DRG(N, r)$.

Lemma 3: For a given H, the upper bound of the smallest number of paths whose length is no more than H between each SD pair in any $DRG(N, r)$ is $\lfloor \frac{r+r^2+...r^H}{N-1} \rfloor$.

Proof: Consider any source, there are at most r 1-hop paths, at most r^2 2-hop paths, ..., at most r^i i-hop paths. Hence, the total number of paths whose length is no more than H is at most $r + r^2 + ... + r^H$. Since there are $N-1$ destinations from the source node, at least one SD pair will have no more than $\lfloor \frac{r+r^2+...r^H}{N-1} \rfloor$ paths of H or less hops. $\square$

C. Load balancing

Another property to ensure that a DRG topology performs well with k-shortest path routing and its variants is how evenly paths are distributed among all links in the topology. Ideally, for an SD pair that is uniform randomly selected among all SD pairs, all links in the network should have an equal probability to be used by its paths. This property is reflected in the maximum link load for the all-to-all communication that is calculated as follows. The load on each link is initiated to be 0. For each SD pair (s, d) that uses X paths to carry traffic, if a link l is used by one of its paths, the load on the link is increased by $\frac{1}{X}$. The maximum link load is the load on the link with the largest load after all SD pairs are considered.

Clearly, if each path for an SD pair is equally likely to be selected to carry traffic for the SD pair, the load of a link, as it is calculated, directly reflects the possibility that the link is selected to carry traffic. If the link load for the all-to-all communication is not evenly distributed among all links, some links will have much higher loads (based on the calculation) than others. These links will have a higher chance of being used to carry traffic; and this can lead to network hot-spots. On the other hand, if the traffic is evenly distributed across the network, the maximum link load should be very similar to the average link load: all links will have a similar probability to be selected to carry a traffic. The next theorem establishes the lower bound of the maximum link load for all-to-all communication.

Lemma 4: For any $DRG(N, r)$, the lower bound of the maximum link load for all-to-all communication is

$$ML(N, r) \geq \frac{LK(N, r, 1) \times (N-1)}{r}.$$

Proof: All-to-all communication has $N(N-1)$ SD pairs. Consider an arbitrary SD pair (s, d). Let $shortest(s, d)$ be the hop count of the shortest path for SD pair (s, d). Let there be X paths the SD pair. Since all of the paths have no less than $shortest(s, d)$ hops, the total load contributed by this SD pair to all links in the network is no less than $shortest(s, d) \times \frac{1}{X} \times X = shortest(s, d)$. As defined earlier, $LK(N, r, 1)$ is the lower bound of the average shortest path length for any $DRG(N, r)$. The total load contributed by all SD pairs in the all-to-all communication is no less than $N(N-1) \times LK(N, r, 1)$. The bound of the maximum link

load is achieved when this lower bound of contributed load is evenly distributed among all $N \times r$ links in the network. Hence the lower bound of the maximum link load for all-to-all communication is

$ML(N,r) \geq \frac{N \times (N-1) \times LK(N,r,1)}{N \times r} = \frac{LK(N,r,1) \times (N-1)}{r}$. □

IV. Generalized De Bruijn Graph and Its Properties

In this section, we will formally prove that for most network configurations (values of N and r), the Generalized De Bruijn Graph (GDBG) with a specific hop-limited all path routing is near optimal for diameter, average k-shortest path length, and load balancing, and is thus a near optimal topology for k-shortest path routing. We will first describe the construction of the Generalized De Bruijn Graph (GDBG), and then prove the properties of GDBG with the k-shortest path routing scheme.

A. GDBG construction and routing

An N-node r-regular GDBG, denoted as $GDBG(N,r)$, consists of N nodes, numbered from 0 to $N-1$. Each node has r incoming links and r outgoing links. The nodes are connected as follows. The r outgoing links of Node 0 is connected to Nodes 0, 1, ..., $r-1$; the r outgoing links of Node 1 is connected to Nodes r, $r+1$, ..., $r+r-1$; ...; the r outgoing links of Node i is connected to Nodes $(i*r) \ mod \ N$, $(i*r+1) \ mod \ N$, ..., $(i*r+r-1) \ mod \ N$, for all $i = 0, 1, ..., N-1$. Basically, the outgoing links from the nodes are arranged in a round-robin fashion, cycling through all N nodes and then repeat. From the construction, it can easily be seen that the outgoing degree of each node is r. Since the outgoing links cycle through N nodes in a round-robin fashion, the $N * r$ incoming links connect to the N nodes and each node has $\frac{Nr}{N} = r$ incoming links.

A node may sometimes connect to itself through a loop-back link. For example, Node 0 connects to Node 0. It is proved in [12] that the number of such loop-back links is usually small as summarized in the following Lemma.

Lemma 5: The number of loop-back links in $GDBG(N,r)$ is at most $2r$, when $r \geq 2$. □

Regardless of the network size, the number of potential loop-back occurrences in $GDBG(N,r)$ is at most $2r$ when $r \geq 2$. When $r \ll N$, the impact of such links is negligible. Such loop-back links may be re-wired to increase the network capacity. However, since the number is small, we will not consider re-wiring in this paper. Figure 1 visualizes an example $GDBG(6,2)$.

For $GDBG(N,r)$, let H be the smallest integer such that $r^H \geq N$, the hop-limited all path routing will use all paths that are no more than H hops to route traffic. We will use the notion ALLPATH(H) to denote the routing scheme for GDBG. Note that the value of H can be derived from N and r. For example, for $GDBG(N = 3000, r = 40)$, $H = 3$. As we will show in the following, when $r^H \geq N$, at least one such path exists between each SD pair. However, in most configurations, the number of short paths is sufficiently large to provide path diversity. For example, for $GDBG(3000, 40)$, $H = 3$, the topology guarantees that there are at least 21 paths that are no more than 3 hops between every SD pair.

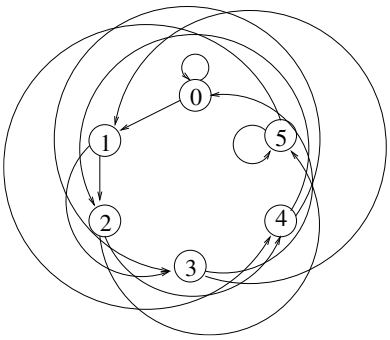


Figure 1. A GDBG(6,2) topology

B. Properties of the GDBG topology

We will show that GDBG with ALLPATH(H) is a near optimal topology in terms of average k-shortest path length (Lemma 8) and load balancing (Lemma 10 and Lemma 11). For the completeness of this paper, we also show the bound for the diameter (Lemma 7), which has been proven before [11], [12]. The proofs of these lemmas are based on Lemma 6 where a concept called *raw path* is used. In contrast to a simple path that does not contain loops, a raw path may be a simple path or may be a path that contains loops. A raw h-hop path from a source node contains h links starting from the source node, potentially including loop-back links: it can be realized by either a h-hop simple path when there is no loop or a shorter simple path when the raw path contains loops, which are removed in the shorter simple path.

Lemma 6: For any source node i in $GDBG(N,r)$, its raw h-hop paths reach r^h continuously numbered nodes (with wrap around due to the modular operation) $i \times r^h \ mod \ N$, ..., $(i \times r^h + r^h - 1) \ mod \ N$.

Proof: We prove by induction. Base case: when $h = 1$, from the construction of the topology, node i connects to nodes $i*r \ mod \ N$, $(i*r+1) \ mod \ N$, ..., $(i*r+r-1) \ mod \ N$.

Induction case: assuming that when $h = n$, node i in its raw n-hop paths reaches r^n continuously numbered nodes $i \times r^n \ mod \ N$, $(i \times r^n + 1) \ mod \ N$, ..., $(i \times r^n + r^n - 1) \ mod \ N$. Consider the case when $h = n + 1$. Extending from node $i \times r^n \ mod \ N$ one more hop can reach r nodes $((i \times r^n \ mod \ N) * r) \ mod \ N$ (which is node $i \times r^{n+1} \ mod \ N$), $(i \times r^{n+1} + 1) \ mod \ N$, ..., $(i \times r^{n+1} + r - 1) \ mod \ N$. Extending from node $(i \times r^n + 1) \ mod \ N$ one more hop can reach r nodes $((i \times r^n + 1) \ mod \ N) * r) \ mod \ N$ (which is node

$(i \times r^{n+1} + r)\ mod\ N$), $(i \times r^{n+1} + r + 1)\ mod\ N$, ..., $(i \times r^{n+1} + 2r - 1)\ mod\ N$. Hence, the r^n continuously numbered nodes $i \times r^n\ mod\ N$, ..., $(i \times r^n + r^n - 1)\ mod\ N$ from the raw n-hop paths are extended into r^{n+1} continuously numbered nodes starting from node $i \times r^{n+1}\ mod\ N$. The lemma is also true when $h = n + 1$. □

Built upon Lemma 6, we can derive the bound for the diameter of a GDBG topology.

Lemma 7 (GDBG Diameter): The diameter of $GDBG(N, r)$ is no more than D such that $r^D \geq N$.

Proof: See [11], [12].

Lemma 1 shows that the optimal diameter for a $DRN(N, r)$ is the smallest integer D_{opt} such that $1 + r + ... + r^{D_{opt}} \geq N$. In practice, only the Kautz graph is known to achieve the bound for some very specific values of N and r. Lemma 7 indicates that for a vast majority of values of N and r, the diameter of $GDBG(N, r)$ is optimal since in most cases when $1 + r + ... + r^{D_{opt}} \geq N$, $r^{D_{opt}} \geq N$. Consider a specific case when $r = 20$. Table I shows the theoretical optimal diameters and the diameters achieved by GDBG for 8420 configurations ($r = 20$, and $N = 2..8421$). It can be seen from the table that among the 8420 networks, 7974 GDBG networks(94.7%) (N=2-20, 22-400, 422-8000) achieve the optimal diameter. Only for less than 5.3% of the cases, the GDBG's diameter is one more than the optimal.

Lemma 8 (average k-shortest path length): In $GDBG(N, r)$, for k-shortest path routing where $\frac{r^{H-1}}{N} < k \leq \frac{r^H}{N}$, the average k-shortest path length is near optimal.

Proof: Since $\frac{r^{H-1}}{N} < k \leq \frac{r^H}{N}$, from Lemma 6, all of the k shortest simple paths (for all pairs of SD nodes) can be realized by paths whose length is no more than H. Consider all k-shortest paths from one source node. There are at least $r - 1$ 1-hop simple paths (r raw 1-hop paths with at most 1 loop-back path that is not counted in the simple paths). Similarly, there are at least $r^2 - r - 1$ 2-hop simple paths (r^2 raw 2-hop paths with at most $r+1$ being 1-hop or 0-hop paths). In general, there are at least $r^i - r^{i-1} - ... - r - 1$ i-hop simple paths for all $i = 1, 2, ..., H - 1$. The rest of the paths will all be H-hop paths since all of the k paths are no more than H hops. The number of such paths is $k(N - 1) - \sum_{i=1}^{H-1}(r^i - r^{i-1} - ... - r - 1)$. Let $RR = k(N - 1) - \sum_{i=1}^{H-1}(r^i - r^{i-1} - ... - r - 1)$, the average k-shortest path length is at most

Diameter	Number of Nodes (Optimal)	Number of Nodes (GDBG)
1	2 to 21	2 to 20
2	22 to 421	21 to 400
3	422 to 8421	400 to 8000

Table I
NUMBER OF NODES FOR A GIVEN DEGREE ($r = 20$) AND A GIVEN DIAMETER

$$
\begin{aligned}
&\frac{\sum_{i=1}^{H-1} j(r^j - r^{j-1} - ... - r - 1) + H \times RR}{k(N-1)} \\
&\leq \frac{\sum_{i=1}^{H-1} j \times r^j + H \times RR}{k(N-1)} \\
&= \frac{\sum_{i=1}^{H-1} j \times r^j + H \times (k(N-1) - \sum_{j=1}^{H-1} r^j)}{k(N-1)} + \frac{H \times \sum_{i=1}^{H-1}(r^{i-1} + ... + r + 1)}{k(N-1)} \\
&= LK(N, r, k) + \frac{H \times \sum_{i=1}^{H-1}(r^{i-1} + ... + r + 1)}{k(N-1)} \\
&= LK(N, r, k) + \frac{H \times \sum_{i=1}^{H-1} \frac{r^i - 1}{r-1}}{k(N-1)} \\
&\leq LK(N, r, k) + \frac{H \times (r^H - 1)}{(r-1)^2 \times k \times (N-1)}
\end{aligned}
$$

Since $\frac{r^H - 1}{N} < k \leq \frac{r^H}{N}$, the additional term $\frac{H \times (r^H - 1)}{(r-1)^2 \times k \times (N-1)}$ is roughly in between $\frac{H}{(r-1)^2}$ and $\frac{H}{r-1}$. In practical networks, H is usually a small number such as 2, 3, or 4; while r is a reasonably large number such as 20, 30, and 40. Hence, the average k-shortest path length in $GDBG(N, r)$ is very close to the optimal that can be achieved. □

The following lemma states that GDBG distributes its short paths evenly among all SD pairs.

Lemma 9: For a given H, there are at least $\lfloor \frac{r^H}{N} \rfloor$ paths between any SD pair in $GDBG(N, r)$ whose length is no more than H.

Proof: From Lemma 6, the raw H-hop paths from an arbitrary source node reach r^H continuously numbered nodes. Thus, each node will be the destination of a raw H-hop path $\lfloor \frac{r^H}{N} \rfloor$ times. Hence, For a given H, there are at least $\lfloor \frac{r^H}{N} \rfloor$ paths between any SD pair in $GDBG(N, r)$ whose length is no more than H. □

Lemma 3 states that the upper bound of the smallest number of paths whose length is no more than H between any SD pair in any $DRG(N, r)$ is $\lfloor \frac{r + r^2 + ... + r^H}{N-1} \rfloor$. Lemma 9 shows that $GDBG(N, r)$ guarantees to provide $\lfloor \frac{r^H}{N} \rfloor$ paths that are no more than H hops for each SD pair. When r is the degree in practical networks (e.g. r = 16, 24, 36), $\lfloor \frac{r^H}{N} \rfloor$ is very close to $\frac{r + r^2 + ... + r^H}{N-1}$ (asymptotically the same). Thus, GDBG distributes its short paths evenly among all SD pairs.

Lemma 10 (load balance): For $GDBG(N, r)$ with ALLPATH(H), each SD pair will at least have $\lfloor \frac{r^H}{N} \rfloor$ paths, and the maximum link load for all-to-all traffic is no more than

$$\frac{(H)r^H - \frac{r^H - 1}{r-1}}{r-1} \times \frac{1}{\lfloor \frac{r^H}{N} \rfloor}$$

Proof: By the definition of ALLPATH(H), H is the smallest integer such that $r^H \geq N$. Since all raw H-hop paths are used in ALLPATH(H), from Lemma 9, each SD pair will have at least $\lfloor \frac{r^H}{N} \rfloor$ paths. Clearly, all of these paths can be derived from raw paths that are H hops or less by removing loop-backs. Hence, the number of links used in the raw paths are strictly no less than those used in the actual simple path: we can bound the maximum link load by counting all potential load contributions from the links in raw paths that are no more than H-hops.

In $DRG(N,r)$, to realize all 1-hop raw paths, each link is used one time. To count the number of times each link is used in all 2-hop raw paths. Consider 2-hop raw paths from each node. There are r first hop-paths, each will be used r times since there are r second hop branches. There are r^2 second hop links. From the construction of $DRG(N,r)$, all links in each hop are evenly distributed among all links in the network: all first hop links in the 2-hop raw paths are even distributed among all links in the network; all second hop links in the 2-hop raw paths are evenly distributed among all links in the network. The example for $DRG(6,2)$ is shown in the Figure 2. There are a total of $N \times r$ first hop links evenly distributed among all $N \times r$ links, each being used r times; and there are a total of $N \times r^2$ second hop links evenly distributed among all $N \times r$ links. Hence, to realize all 2-hop raw paths, each link is used $\frac{N\times r}{N\times r} \times r + \frac{N\times r^2}{N\times r} = 2r$ times. Following the similar logic, to realize all 3-hop raw paths, each link is used $3 \times r^2$ times; and to realize all i-hop raw paths, each link is used $i \times r^{i-1}$ times. Hence, the total number of times each link is used in all of the raw paths that are no more than H-hop is

$$1 + 2r + 3r^2 + ... + (H)r^{H-1} = \frac{(H)r^H - \frac{r^H-1}{r-1}}{r-1}$$

Since the number of paths between every SD pair is at least $\lfloor \frac{r^H}{N} \rfloor$, each time a link is used, the load contribution to the link is at most $\frac{1}{\lfloor \frac{r^H}{N} \rfloor}$. Hence, the link load for all-to-all traffic for any link is at most:

$$\frac{(H)r^H - \frac{r^H-1}{r-1}}{r-1} \times \frac{1}{\lfloor \frac{r^H}{N} \rfloor} \square$$

Lemma 11: Let H be the smallest integer such that $r^H \geq N$, for GDBG with ALLPATH(H), the maximum link load for all-to-all communication is near optimal.

Proof: Since H is the smallest integer such that $r^H \geq N$, $LK(N,r,1) \approx H-1$ for most of the network configurations. From Lemma 4, the lower bound on the maximum link load for all-to-all traffic is

$$ML(N,r) \geq \frac{LK(N,r,1) \times (N-1)}{r} \approx \frac{(H-1)(N-1)}{r}$$

.

From Lemma 10, for $GDBG(N,r)$, the maximum link load is at most

$$\frac{(H)r^H - \frac{r^H-1}{r-1}}{r-1} \times \frac{1}{\lfloor \frac{r^H}{N} \rfloor} \approx \frac{H \times N}{r-1}$$

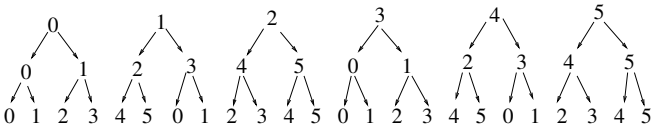


Figure 2. All 2-hop raw paths in $DRG(6,2)$: all first hop links are evenly distributed; and all second hop links are evenly distributed.

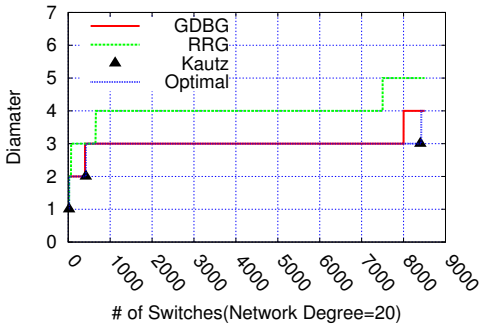


Figure 3. Diameters for different sized topologies with $r = 20$

Since $\frac{H\times N}{r-1}$ is very close to $\frac{(H-1)(N-1)}{r}$, the maximum link load for all-to-all communication on $GDBG(N,r)$ with ALLPATH(H) is near optimal. This lemma is confirmed in our numerical study (Figures 10 and 11). $\square$

V. Empirical comparison of topological properties

We empirically evaluate the topological properties for random regular graph (RRG), generalized De Bruijn graph (GDBG), Kautz graph, and the theoretical bounds (optimal). The empirical results confirm the theoretical findings in the previous section.

A. Diameter

Figure 3 shows the results for topologies with size ranging from 1 to 8500 and a fixed nodal degree ($r = 20$) while Figure 4 shows the results for a fixed network size ($N = 901$) but different nodal degrees. As can be seen from the figures, Kautz graphs are optimal, but only apply for a very limited number of network configurations. The diameter for RRG is mostly 1 or 2 more than the optimal. In most of the cases when the optimal diameter is 3 (Figure 3), the diameter of RRG is 4, which is 33.3% worse than the optimal. On the other hand, GDBG achieves optimal diameter for the majority of the network configurations. This reaffirms the conclusion in Table I that RRG is not ideal in terms of network diameter in most cases and GDBG is optimal in most cases.

B. Average k-shortest path length

Let us now examine the average k-shortest path length. We first show the average shortest path length ($k = 1$). Figure 5 shows the results for topologies with size ranging from 1 to 8500 and a nodal degree of 20 while Figure 6 shows the results for a fixed network size ($N = 901$) but different nodal degrees. As can be seen from the figures, depending on the network configuration, the average shortest path length for RRG can be up to more than 15% worse than the optimal. In all of the configurations, GDBG has very similar average shortest path length to the optimal.

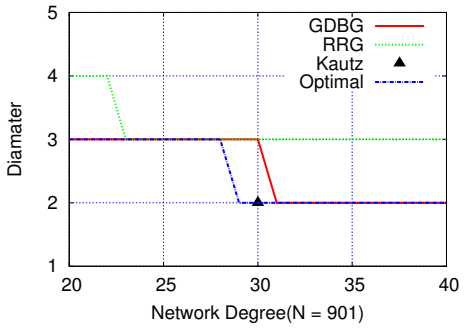


Figure 4. Diameters for the same sized topologies ($N = 901$) with different nodal degrees

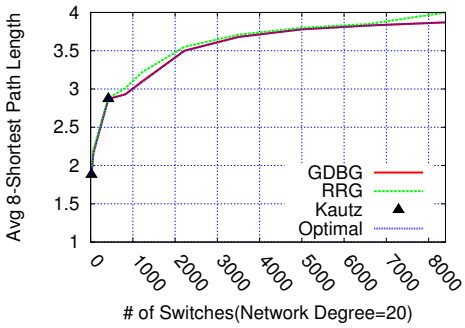


Figure 7. Average 8-shortest path lengths for different sized topologies with same nodal degree ($r = 20$)

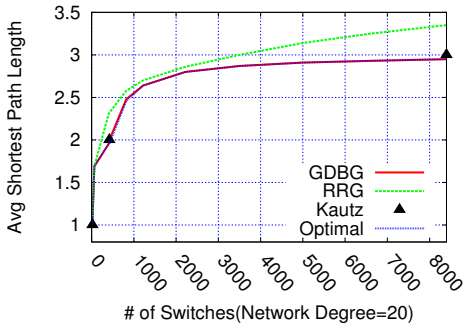


Figure 5. Average shortest path lengths for different sized topologies with same nodal degree ($r = 20$)

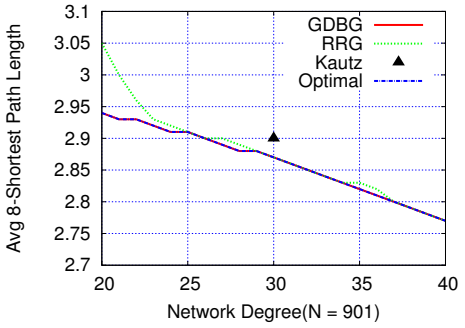


Figure 8. Average 8-shortest path length for the same sized topologies ($N = 901$) with different nodal degrees

Kautz graphs are always optimal but only work for a few configurations.

Figure 7 shows the average 8-shortest path lengths for topologies with size ranging from 1 to 8500 and a fixed nodal degree ($r = 20$) while Figure 8 shows the results for a fixed network size ($N = 901$) but different nodal degrees. RRG and Kautz graphs are slightly worse than the optimal (less than 3% in all cases) while GDBG virtually has an identical average 8-shortest path length as the optimal. When the number of paths used is not 1, the gap between RRG and the optimal is small. For Kautz graphs, even though it achieves optimal average shortest path length for single path routing, when multiple paths are used, the topology is no longer optimal.

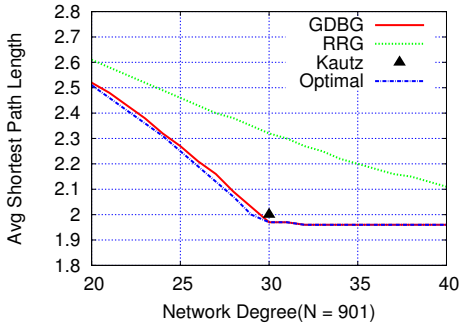


Figure 6. Average shortest path length for the same sized topologies ($N = 901$) with different nodal degrees

Figure 9 shows the average k-shortest path length for different topologies with $N = 901$, $r = 30$ and varying k. The trend in this figure is representative for other network configurations. As demonstrated in the figure, when $k = 1$, RRG is noticeably worse than the other topologies (GDBG, Kautz, optimal) in average shortest-path length. For higher values of k, the gap among different topologies is small. Depending on the value of k, for some configurations, all topologies have similar average k-shortest path length; for some other configurations, RRG has up to 5% longer average k-shortest path length than other topologies. GDBG consistently has near optimal average k-shortest path length for all network configurations.

C. Load balancing

In the comparison of the load balancing property, we assume ALLPATH(H) for GDBG. For other topologies, we first compute the average number (A) of paths for each SD pair for GDBG. All other topologies assume k-shortest path routing with $k = \lceil A \rceil$.

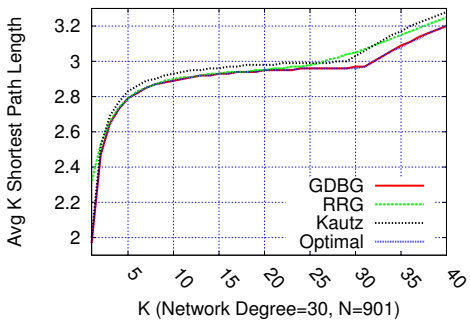


Figure 9. Average k-shortest path length ($N = 901$, $r = 30$)

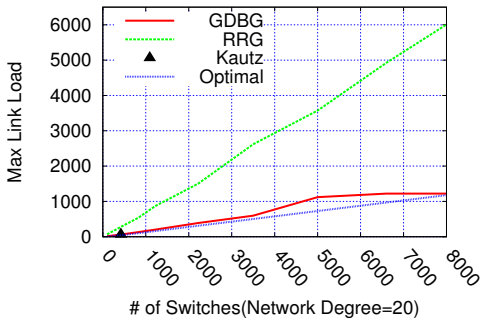


Figure 10. Maximum link load for all-to-all communication on different topologies with increasing network size($r = 20$)

Figure 10 shows the changes in the maximum link load while increasing the network size, whereas, Figure 11 changes the nodal degree while keeping the number of switches fixed. In both cases, there is a significant gap between maximum link loads in RRG and the optimal case. GDBG distributes the load much more evenly among all links in the network and therefore, reaches almost optimal maximum link load in all instances.

Figure 12 shows the normalized bisection bandwidth of RRG against GDBG for a fixed network size ($N = 901$) and different nodal degrees. The normalized bisection bandwidth

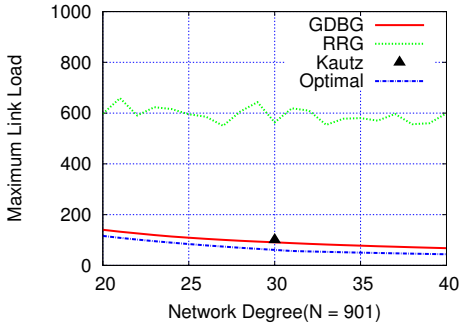


Figure 11. Maximum link load for all-to-all communication on different topologies with increasing network degree($N = 901$)

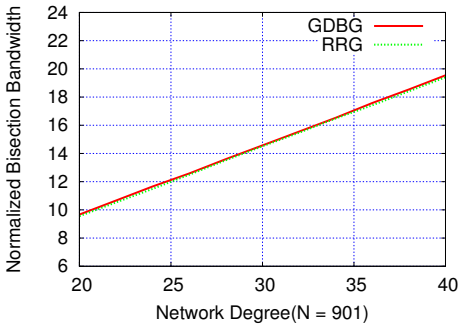


Figure 12. Bisection Bandwidth of RRG vs GDBG with increasing network degree($N = 901$)

is the total number of links in the minimum cut divided by $\frac{N}{2}$, the size of each partition. As can be seen in the figure, RRG and GDBG maintain almost identical bisection bandwidths. This is observed for all other network configurations.

D. GDBG versus related topologies

Here we summarize GDBG topological properties. The key properties are the following

- The diameter of $GDBG(N, r)$ is near optimal among all $DRG(N, r)$ (Lemma 7).
- For any k, the average k-shortest path length of $GDBG(N, r)$ is near optimal (Lemma 8).
- For any value H, $GDBG(N, r)$ evenly distributes its short paths whose length is no more than H among all SD pairs (Lemma 9).
- With ALLPATH(H), $GDBG(N, r)$ achieves near perfect load balancing for most network configurations (Lemma 11).

In comparison to the Kautz graph that achieves the optimal diameter for very specific values of N and r, $GDBG(N, r)$ can be constructed for any values of N and r and achieves optimal diameter for a vast majority of network configurations. In addition, GDBG is able to distribute short paths among all of its SD pairs evenly and achieves almost perfect load balancing with ALLPATH(H). Kautz graphs do not have these properties that are important to achieve high performance when k-shortest path routing or its variants are used. In comparison to RRG, GDBG is better in almost all important topological properties including diameter, average k-shortest path length, short path distribution, and load balancing.

VI. Comparison between RRG and GDBG

With GDBG, one can comparatively study RRG with a provably near optimal topology (GDBG) for arbitrary communication patterns. In the study, we assume that the switch level topology is either an RRG or a GDBG. One or more compute nodes can attach to each switch. For both topologies, the switches are numbered from 0 to N-1; and the compute nodes within each switch are numbered contiguously.

The ALLPATH(H= 3) routing is used for GDBG in all experiments. For RRG, we initially planned to use the same hop-limited all path routing with $H = 3$ in the evaluation. Interestingly, the experiments failed because RRG fails to provide connectivity among all SD pairs with the same $H(= 3)$ as the corresponding GDBG does: some SD pairs do not have any path that is no more than 3 hops. To use ALLPATH(3), RRG will require one or two more hops than GDBG just in order for the routing to work. Such a comparison would be unfair to RRG since traffic would take longer paths. This itself indicates that GDBG is a better topology. To address this issue and compare the two topologies in the experiments, we compute the average number (A) of paths for all SD pairs with the ALLPATH(3) routing on GDBG. We then compare GDBG using ALLPATH(3) with RRG using k-shortest path routing, where $k = \lceil A \rceil$. The total number of paths available for communication in RRG is slightly more than that in GDBG: the experiments slightly favor RRG. Note that k-shortest path routing was suggested for RRG [2].

Two traffic patterns from LANL-FSU Throughput Indices (LFTI) [14] are used in the evaluation: the random permutation pattern and the random shift pattern. In a random shift pattern, each node i communicates with node $(i + a)\ mod\ nprocs$, where $nprocs$ is the number of processing nodes and a is a random value from 1 to $nprocs$. The two patterns are applied in two different ways: at the compute node level with multiple compute nodes connecting to each switch, and at the switch level. When a communication pattern is applied at the compute node level, it is assumed that the links connecting compute nodes to switches have the same speed as the links between switches. For the permutation pattern, the traffic is diverse since different nodes in one switch may communicate to different nodes at different switches. For the shift pattern, the traffic is more concentrated: all nodes in each switch communicate to nodes in one or two other switches. When the traffic patterns are at the switch level, we assume that one compute node is attached to each switch and has infinite bandwidth to and from the switch. This allows for evaluating the potential throughput between switches. The traffic is concentrated with switch level traffic patterns since each switch only communicates with another switch.

For all traffic patterns, the aggregate throughput is computed using a linear programming formulation that maximizes the minimum concurrent flow, the same metric used in [3]. The linear programming formulation is solved using IBM CPLEX. Each point in the figure is the average of 20 to 200 random samples; the data collection stops when the 95% confidence interval is less than 1% of the average: with high confidence, the reported numbers are close to the theoretical average.

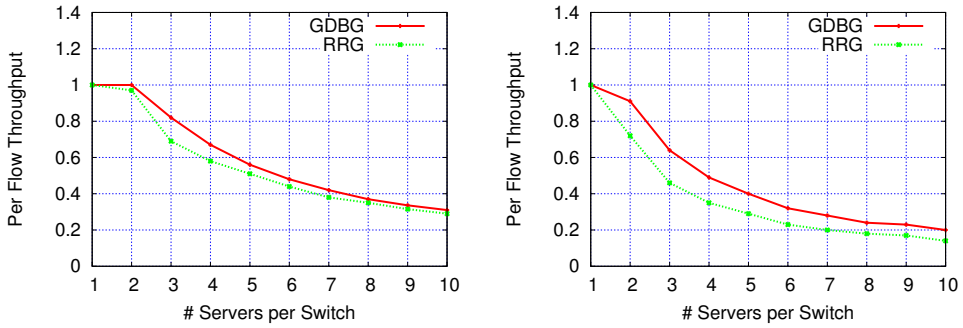
Figure 13 shows the results for compute-node level traffic pattern on RRG(150, 8) and GDBG(150, 8) with the number of compute nodes on each switch ranging from 1 to 10. The per flow throughput is normalized to the link speed: 1 means 100% of the link speed. As can be seen from the figure, when the number of nodes in each switch is 1, both RRG and GDBG can support the full speed of each compute node for both patterns. As the number of compute nodes per switch ($persw$) increases, both cannot support full bandwidth from the compute nodes. For both patterns, GDBG consistently performs noticeably better than RRG. For the permutation pattern, when the number of nodes per switch is between 3 and 10, GDBG is between 3% ($persw = 10$) and 20% ($persw = 3$) better; for shift patterns, GDBG is better by up to 43% (achieved when $persw = 10$). The figure also shows that when the number of nodes per switch is large (e.g. $persw = 10$), RRG has similar performance as GDBG for permutation patterns. In this case, the traffic pattern at the switch level is diverse as each switch communicates with many other switches. This confirms earlier results [3] that RRG can achieve near optimal performance when traffic is diverse. On the other hand, for the more concentrated shift patterns, RRG is less effective.

Figure 14 shows the results for the switch-level patterns where traffic is concentrated. In the experiments, the network size ranges from 100 to 1000 and the nodal degree is fixed at 16. In the figure, the per-flow throughput is normalized to the (between switch) link speed. Since there are multiple links between switches, the per-flow throughput is usually more than 1. For both patterns, GDBG performs significantly better than RRG, especially as the network grows in size. For example, on the 900-node system, GDBG achieves a per-flow throughput of 2.64 for the shift pattern, 2.64 times that of RRG. As the network size increases, the advantages of GDBG manifest.

These results refine the current understanding of RRG when Singla concluded that RRG is near optimal in general [3]. Our results indicate that RRG is near optimal for diverse traffic when each switch communicates with many other switches. However, for concentrated traffic where each switch only communicates with one other switch, the performance of RRG is far from GDBG, let alone the optimal.

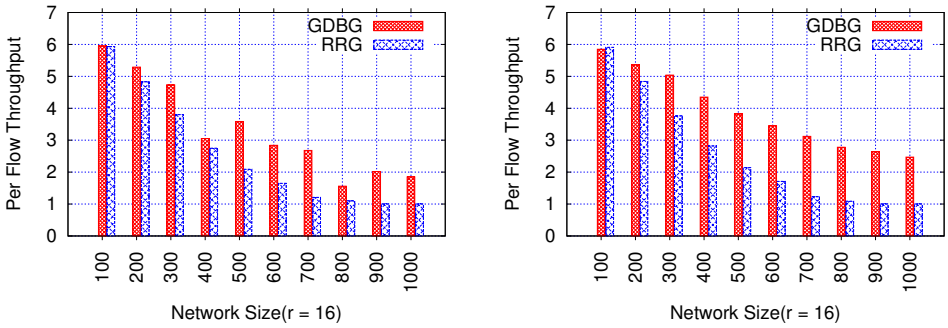
VII. Conclusion

We derive bounds for DRG on diameter, average k-shortest path length, and load balance with k-shortest path routing and show that RRG is not ideal for diameter and load balancing. We identify a near optimal topology for k-shortest path routing with the associated k-shortest path routing scheme: the generalized De Bruijn graph (GDBG) with the hop-limited all-path routing (ALLPATH(H)). Not only GDBG is an effective topology for k-shortest path routing, it can also serve as a near optimal benchmark to evaluate other topologies when k-shortest path routing is appropriate. We show that while random regular graph performs well for diverse traffic, it is not effective in dealing with concentrated



(a) Permutation Traffic Pattern (b) Shift Traffic Pattern

Figure 13. Average per flow throughput for compute node level traffic patterns $N = 150$, $r = 8$



(a) Permutation Traffic Pattern (b) Shift Traffic Pattern

Figure 14. Average per flow throughput for switch-level traffic patterns ($N = 100, .., 1000$ $r = 16$).

traffic patterns such as switch-level permutation and shift patterns.

REFERENCES

[1] M. Koibuchi, H. Matsutani, H. Amano, D. F. Hsu, and H. Casanova, "A case for random shortcut topologies for hpc interconnects," *SIGARCH Comput. Archit. News*, vol. 40, no. 3, pp. 177–188, Jun. 2012.

[2] A. Singla, C.-Y. Hong, L. Popa, and P. B. Godfrey, "Jellyfish: Networking data centers randomly," in *Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation*, ser. NSDI'12, 2012, pp. 17–17.

[3] A. Singla, P. B. Godfrey, and A. Kolla, "High throughput data center topology design," in *Proceedings of the 11th USENIX Conference on Networked Systems Design and Implementation*, ser. NSDI'14. Berkeley, CA, USA: USENIX Association, 2014, pp. 29–41.

[4] X. Yuan, S. Mahapatra, W. Nienaber, S. Pakin, and M. Lang, "A new routing scheme for jellyfish and its performance with hpc workloads," in *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, ser. SC '13, 2013, pp. 36:1–36:11.

[5] M. Besta and T. Hoefler, "Slim fly: A cost effective low-diameter network topology," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, ser. SC '14, 2014, pp. 348–359.

[6] J. Kim, W. Dally, S. Scott, and D. Abts, "Technology-driven, highly-scalable dragonfly topology," in *Computer Architecture, 2008. ISCA '08. 35th International Symposium on*, June 2008, pp. 77–88.

[7] B. Bollobás, "The isoperimetric number of random regular graphs," *Eur. J. Comb.*, vol. 9, no. 3, pp. 241–244, May 1988.

[8] N. G. de Bruijn, "A combinatorial problem," *Nederl. Akad. Wetensh. Proc. Ser.*, vol. A 49, pp. 758–764, 1946.

[9] "The diameter of random regular graphs," *Combinatorica*, vol. 2, no. 2, 1982.

[10] W. H. Kautz, "Bounds on directed (d, k) graphs," *Theory of Cellular Logic Networks and Machines, AFCRL-68-0668 Final Report*, pp. 20–28, 1968.

[11] M. Imase and M. Itoh, "Design to minimize a diameter on building block network," *IEEE Transactions on Computers*, vol. C-30, pp. 439–443, 1981.

[12] M. Imase and M. Itoh, "A design for directed graph with minimum diameter," *IEEE Transactions on Computers*, vol. C-33, pp. 782–784, 1983.

[13] V. G. Cerf, D. D. Cowan, R. C. Mullin, and R. G. Stanton, "A lower bound on the average shortest path length in regular graphs," *Networks*, vol. 4, no. 4, pp. 335–342, 1974.

[14] X. Yuan, S. Mahapatra, M. Lang, and S. Pakin, "Lfti: A new performance metric for assessing interconnect designs for extreme-scale hpc systems," in *Proceedings of the 2014 IEEE 28th International Parallel and Distributed Processing Symposium*, ser. IPDPS '14, 2014, pp. 273–282.