The Journal of Supercomputing (2023) 79:8729–8753
https://doi.org/10.1007/s11227-022-05005-1

Performance evaluation of multi-exaflops machines using Equality network topology

Chi-Hsiu Liang[1] · Chun-Ho Cheng[1] · Hong-Lin Wu[1] · Chao-Chin Li[1] · Po-Lin Huang[1] · Chi-Chuan Hwang[1]

Accepted: 4 December 2022 / Published online: 26 December 2022

Abstract

In modern computing architectures, graph theory is the soul of the play due to the rising core counts. It is indispensable to keep finding a better way to connect the cores. A novel chordal-ring interconnect topology system, Equality, is revisited in this paper to compare with a few previous works. This paper details the procedures for constructing the Equality interconnects, its special routing procedures, the strategies for selecting a configuration, and evaluating its performance using the open-source cycle-accurate BookSim package. Four scenarios representing small- to large-scale computing facilities are presented to assess the network performance. This work shows that in 16,384-endpoint systems, the Equality network turns out to be the most efficient system. The results also show the steady scalability of Equality networks extending to 48–320K, and a million endpoints. Equality networks are adjustable to fit with commodity hardware and resilient under ten common traffic models. It is suggested that Equality network topology can be used in constructing efficient multi-exaflops supercomputers and data centers.

Keywords Network topology · Routing protocols · Discrete-event simulation

✉ Chi-Hsiu Liang
alvyn.liang@gmail.com

✉ Chi-Chuan Hwang
chchwang@mail.ncku.edu.tw

Chun-Ho Cheng
chunho.cheng@gmail.com

Hong-Lin Wu
winching.hi@gmail.com

Chao-Chin Li
kin_2000@hotmail.com

Po-Lin Huang
50ta2012@gmail.com

[1] Department of Engineering Sciences, National Cheng Kung University, Tainan 701, Taiwan

Mathematics Subject Classification 05C12 · 68M10 · 68Q85 · 68R10 · 90C35 · 94C15

1 Introduction

High-performance computing (HPC) is a type of computing that uses high-end computing components to cooperatively address large-scale tasks that cannot be solved easily by ordinary computers. The computing components are connected by HPC networks to achieve better efficiency.

An HPC network differs from other networks in that it often seeks to synchronize communication and computation so that the communication does not interrupt the computation too much to increase efficiency. An HPC network also tends to use homogeneous computing hardware, such as the same model of switches (with an equal number of ports), CPUs, and accelerators across the entire implementation. Homogeneous products in a system ensure lower prices for each component due to mass production and more straightforward restoration by prompt replacement when some parts go wrong.

Hwang et al. have shown the potential of Equality network compared with a few popular HPC network topologies [1–4] such as 2-tier fat-tree, 3-tier fat-tree, 3D torus, and 5D torus. In this work, we further analyze the performance of Equality networks in different scales to compare with Slim Fly, Dragonfly, and two popular network topologies, Fat-tree and Tori. We also extend the focus on applying the Equality networks to enable machines capable of reaching multi-exaflops based on current hardware craftsmanship.

The main contributions of the current work that are different from previous works include the following:

- The development and implementation of the systematic routing tables for Equality networks,
- The modified routing algorithm bottleneck-UGAL to refrain from over-subscribed paths,
- The introduction and explanation of a new measure called bisection ratio in addition to bisection bandwidth,
- The analysis of the resulting network properties (diameter, average distance, latency, and throughput) of various scales of the Equality networks and the comparison to other existing publications,
- The strategies of finding a suitable configuration for a future HPC system utilizing Equality network topology, and
- The largest cycle-accurate simulation ever calculated by BookSim (a 1 M-endpoint system).

2 Network architecture

Different network topologies often are designed to fit specific workloads when designed beforehand. To justify the quality of a network and whether it is suitable for the target application workloads, one can inspect the performance measures of a network and additionally perform simulations on the network. The standard network measures used in this paper include the network diameter d and average distance a. The standard communication measures are the message latency and the network's overall throughput under different traffic patterns and injection intensities.

A well-balanced topology should have a reasonable network diameter and also accompanied by tailored routing algorithms to reduce the latency and increase the throughput. Nevertheless, for any application, a given network has an effective diameter, d_{eff}, if all communication patterns of the specific application use no more than d_{eff} hops in the network regardless of the actual network diameter. The same idea goes for the average hop count.

Under low injection rates, when a packet never contends with other packets for resources, the latency is called 'zero-load latency,' l_0, which is the sum of the latency bits when no queue or blocking is involved. Under high injection rates, when the network is saturated under the specific load, traffic pattern, and routing algorithm, if the latency goes high, the throughput approaches the saturation throughput, t_{sat}.

Predominantly, the latency can be reduced if the network diameter d is reduced; however, the network latency and throughput still depend on the traffic patterns of real applications. Several traffic patterns [5] are devised based on communication patterns triggered by real-world applications. For instance, the transpose traffic mode is encountered in corner-turn [6] and matrix transpose applications. Choosing a network that outperforms others on most traffic patterns contributes to a better system.

2.1 Switching delays

The port-to-port latency of the contemporary high-end switches ranges from tens of nanoseconds to a few μs. The switching delay depends on many parameters, including the cable length, cable material, optical-electrical conversion, buffer size, switching logic cycle frequency, switching memory access time, routing table size, routing calculation complexity, hop count, etc. The hardware performance can be calculated by summing up all the instruction cycles contributing from each logic element, multiplied by the time in nanoseconds per cycle. The hardware issues, such as memory access time and logic frequency, are out of the scope of the current paper.

If the latency per hop is fixed when the switching hardware is selected, the only way to reduce the overall latency of the system is to fine-tune the topology. In reality, the messages can be waiting in channel buffers in packet-based routing architectures or for the channel to be freed in wormhole routing architectures. In the events of network congestion, the hopping distance may have a lesser impact on the network latency; therefore, the design of the topology and routing algorithm should not only focus on the diameter of the network.

2.2 Network selection

There are many topology inherited shortcomings in the existing topologies. The torus networks bear the routing difficulty of long hopping distances and over-subscribed paths. Although fat-tree networks perform well on all permutation traffic patterns, the cost for large fat-tree networks is considerably higher, and the number of layers determines the zero-load latency (l_0) of a fat-tree network.

Most topologies have the total router number being the product of integers, meaning the number of routers cannot be changed easily if the budget is modified. Dragonfly networks suffer from low global bandwidth, which becomes the bottleneck of global traffic. The irregularity of Slim Fly network links predestined the routing intricacy when the network diameter is more than two. Slim Fly MMS requires very high radix routers according to its mathematical expression when it scales up. According to its expression, with 80-port routers, the largest configuration it can offer is $k' = 53$, $\delta = -1$ MMS, which can hold about 2450 routers and 63,700 endpoints. Once a model of router (in this case: 80-port) is selected for Slim Fly MMS, it almost means the number of endpoints is fixed. Other closest solutions in this range are:

- $\delta = 0$, holding 2048 routers and 49,152 endpoints.
- $\delta = 1$, holding 2402 routers and 52,272 endpoints.

and one can see these solutions differ significantly from the first solution.

Daryin and Korzh [7] have been looking for low-diameter topologies that have structures with the optimal performance for the Russian supercomputer manufactured by T-Platforms. They chose to go with the hybrid of Slim Fly and Flattened Butterfly for comparison with Dragonfly, tori, hypercube, and Flattened Butterfly. In the end, they chose Flattened Butterfly for their system #22 in November 2014 Top500 list [8] with *R*max 1.8 petaflops. This has shown that from the perspective of a system designer, the window to finding an appropriate system size can be very narrow. The comparison of our results with this petascale system is shown in Fig. 5 (Sect. 5).

Hwang et al. have compared the performance of Equality network against 2-tier fat-tree [1], 3-tier fat-tree [2], and 3D torus [3]. They also show that Equality can be used to design many-core computer chips [4].

3 Methods

In the following sections, we recapitulate the construction of an Equality network (Sect. 3.1). Section 3.2 addresses how we describe Equality networks and some related prior arts. Further, we detail the optimization procedures (Sect. 3.3) and the routing table construction procedures (Sect. 3.4). Section 3.5 briefs three routing algorithms utilized in this work, and Sect. 3.6 recites the routine for cycle-accurate simulation.

The concentration p depends on what application will run on the network, while the system is being designed. However, finding a proper p under saturated traffic is essential. The balance of network radix K and concentration p, together with the system's networking cost and bisection bandwidth, is discussed in Sect. 4.

To get the estimated performance of our designs, we utilize the open-source BookSim [9] for cycle-accurate simulation of our networks. The package we use is downloaded from its GitHub site (https://github.com/booksim) and later modified locally to include our home-brewed routing procedures and algorithms. We implemented our topology, routing algorithms, and a few extra traffic models into BookSim. The results in various system scales are discussed in Sect. 5.

Symbols and notations used in the current paper are listed in Table 1.

3.1 Connection rules

Since our previous works are short conference papers, we would like to have this chance to address more about the initial conception of Equality network topology.

Upon constructing a network, one of the intuitive approaches is to link all the nodes into a ring. A ring topology has a network radix of 2. At the dawn of the current study, we looked at Hamiltonian cycles and sought to find a way to reduce the network's diameter. Since we only focus on any Hamiltonian networks that have equal radix on all router nodes, upon adding one link on a router, the same link has to be applied on all routers. To keep the routing identical for all routers, we tried many strategies for adding connections on all nodes.

To make the interconnects, every member of the routers makes links through the following rules. An Equality network has N routers, where N is an even number. The routers are sequentially numbered from zero to $N-1$, i.e., $r_0, r_1, \dots, r_{N-1}$. The routers are connected to form a ring and later to other ring members, just like chordal ring topologies.

A set of positive integers, C, starting from 2 to $N-3$, excluding any even numbers greater than $N/2$, are used as the candidates C for making the physical links.[1] From C, a subset of integers, S, composed of S_A (a collection of odd positive integers) and S_B (a collection of even positive integers) is selected for the

[1] It is worth to note that all even numbers $S_j < N/2$ make equal links as the corresponding even numbers $N - S_j$; therefore, all even numbers greater than $N/2$ are excluded from the set.

Table 1 Symbols and notations used in the paper

General network terms	
Notation	Explanation
N2048K38P8	Notation for configuring a group of Equality networks. This example represents a network configuration consisting of 2048 switches, each with a network radix of 38, and is attached to 8 endpoints. This notation can also be written as n2048k38p8 (allowing both uppercase and lowercase). Each network in the group can be specified in more detailed notation
N14K6[−1,1,3,9](4)	A detailed specification of a network
Variable	Explanation
N	The total number of nodes/routers in the network
K	The number of inter-node links per node/router (network radix)
p	The number of endpoints per router
P	Router radix, i.e., $P = K + p$
r_i	ith node/router $(r_0, r_1, \dots, r_{N-1})$
d_{ij}	Shortest distance between r_i and r_j. Note that $d_{ij} = dji$
a	Average distance of the network, i.e., $\overline{d_{ij}}$
d	Diameter of the network, i.e., $\max(d_{ij})$
L	Number of layers in fat-tree networks
n	The number of endpoints in the network ($n = N \cdot p$ for Equality networks; $n = 2(P/2)^L$ for full fat-tree networks)
ϕ	Channel bandwidth
Φ	Total network bandwidth, $\Phi = \frac{N\phi(K+2p)}{2}$
B	Bisection bandwidth
B_r	Topology bisection ratio (counting only the inter-router links), $B_r = \frac{2B}{\phi NK}$
b_r	Network bisection ratio (including the links to servers)
Equality network terms	
$\mathbf{O}$	A set of positive odd integers for linking candidates
$\mathbf{E}$	A set of positive even integers for linking candidates
$\mathbf{C}$	A set of positive integers for linking candidates.$\mathbf{C} \equiv \{\mathbf{O} \cup \mathbf{E}\}$
$\mathbf{S_A}$	The array for selected odd alternative links.$\mathbf{S_A} \in \mathbf{O}$
$\mathbf{S_B}$	The array for selected even links. $\mathbf{S_B} \in \mathbf{E}$
$\mathbf{S}$	Array of selected positive integers from the collection $\mathbf{C}$ for a given interconnect. $\mathbf{S} \equiv \{\mathbf{S_A} \cup \mathbf{S_B}\}$
S_j	jth member of the array $\mathbf{S}$
Routing terms	
$\mathbf{P}(r_i, r_j)$	All paths from r_i to r_j
$\vec{\mathbf{P}}(r_i, r_j)$	Shortest paths from r_i to r_j
$\vec{\mathbf{P}}'(r_i, r_j)$	Sub-shortest paths from r_i to r_j
$\hat{\mathbf{R}}$	Shortest path routing table
L_{0j}	List of routers for sub-shortest paths from r_0 to r_j

Table 1 (continued)

Routing terms	
$\widehat{\mathbf{R}'}$	The combination of the shortest path routing table and sub-shortest alternatives, where $\widehat{\mathbf{R}'}$ includes $\widehat{\mathbf{R}}$ and $L_{0j}\forall r_j \neq r_0$
Performance terms	
l_0	Zero-load latency
t_{sat}	Saturation throughput

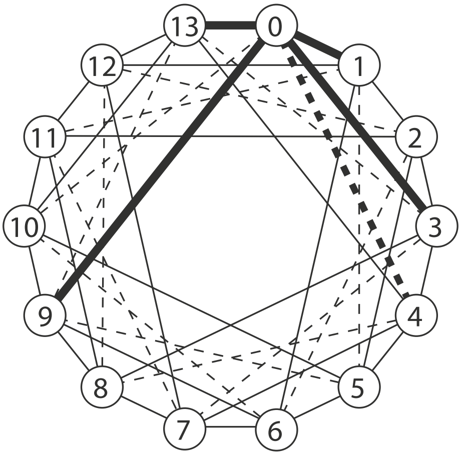

Fig. 1 A simple Equality network named N14K6[−1,1,3,9](4). The bold lines mark the links initiate from r_0. The solid lines mark odd links, and the dashed lines mark even links

interconnections. The connections are made for every router r_i with $r_{(i+S_j)\text{mod}N}$ if i is even, or with $r_{(i-S_j)\text{mod}N}$ if i is odd, for every number S_j.

3.2 Syntax

The general notation of Equality involves an '*n*' denoting the total number of routers, followed by a number, and a 'k' followed by another number indicating the network radix. The notation of '*p*' can be omitted if the number of attached endpoints is not yet specified. Both uppercase and lowercase are allowed in the notation of '*n*,' '*k*,' or '*p*' as long as they are in a sequence to describe the constraints of the target network. The notation is for the designers to have a rough idea of what configuration the network has followed rather than the full specification.

A pair of square brackets and a pair of parenthesis enclosing comma-separated numbers are for detailed specification of an Equality network, where S_{A} is listed in square brackets, and S_{B} is listed in the parentheses. For instance, an Equality network named N14K6[−1,1,3,9](4) is presented in Fig. 1. As described above, the '*N*' and '*K*' mark the number of router nodes and network radix constraints, respectively.

Hence, the number 14 means there are 14 routers in the network, and the number 6 indicates that the network radix of the routers is six. This specification is more detailed in the hops but does not say how many endpoints are attached.

Remark 1 For the general configuration of a network, one can also express in the short-hand notation of n14k6p3 to represent all Equality networks with 14 switches, network radix 6, and 3 endpoints per switch.

An Equality network has an equal number of inter-router connections in all switches. The number of inter-router connections, K, can be evaluated by the following equation:

$$K = \text{len}(S_\text{A}) + \begin{cases} \text{len}\,(S_\text{B}) \cdot 2 - 1 \text{ if } N/2 \in S_\text{B} \\ \text{len}\,(S_\text{B}) \cdot 2 \text{ if } N/2 \notin S_\text{B} \end{cases} \tag{1}$$

A direct explanation of Eq. 1 is that each odd number adds one inter-router connection for each router. Each even number adds two inter-router connections for each router, except the diameter link of the ring, which is the same as the odd numbers, adds one inter-router connection for each router. The diameter link of the ring can be either an odd or an even link.

The breakthrough of Equality in chordal ring topologies [10–14] is in the mixing of multiple even and odd links in a ring of even nodes while keeping the alternative nature of odd links. In addition, systematic routing rules are provided for derived networks.

For instance, in 2016, Faraha et al. discussed [13] about degree six 3-modified chordal ring networks, where the total number of nodes N must be divisible by 3, and every three nodes are grouped into a class. Zabłudowski et al. discussed [15] about modified network double ring structures. The only publication we found to show a likelihood of the Equality networks is [12] (modified chordal ring CHR5_a(20; 3,7), CHR5_c(16; 3,5,7) and CHR5_d(16; 3,5,8)); however, it is not based on the same construction rules and applies only to radix-5 networks.

3.3 Network optimization

Equality topology offers a plethora of networks to be assessed and made to practice. One needs to set a goal to find the candidates.

3.3.1 Assignment of $\mathbf{S_A}$ and $\mathbf{S_B}$

Upon the decision of the network radix, one follows Eq. 1 to confine the lengths of S_A and S_B to achieve a fixed K. For instance, he or she would like to construct a network of 1840 routers with network radix 17. If four even numbers are selected as $\mathbf{S_B}$, assuming hop 920 is not selected (i.e., N/2), they contribute eight connections to each router. If S_B has 920, they consume seven connections to each router. Let us say

920 is not in S_B; then, additional nine numbers can be added to S_A to get an Equality network of radix 17.

3.3.2 Optimization

We optimize Equality networks with a genetic algorithm. Ideally, an initial random seed is given for the generation of S_A and S_B from C, with the constraint of a predefined radix K. We then select the population size and other simulation parameters, such as the maximum number of generations, and mutation rate. The goal of the optimization is to minimize the product of the average distance a and network diameter d. At the end of the optimization, a series of best results from generations of evolution are reported. If the search space being explored is large, the optimized results are not necessarily the global minimum; however, the results are usually low enough for application. If a sufficient amount of evolution is conducted, one can get results close to the global minimum.

For large systems, the empirical approach can involve only optimizing S_A while keeping S_B fixed. For instance, some of our systems are derived by handpicking S_B (and possibly a part of S_A) and optimize the remaining S_A; otherwise, the phase space would be too large to explore. The designer has to adjust S_B many times in length and composition and optimize S_A to reach a better set of answers.

3.4 Routing table

Routing is an essential part of communication. In the current work, the routing procedures include a universal routing table and three routing algorithms.

Remark 2 An even node and an odd node see an entirely identical network structure only in the reverse direction in an Equality network.

Starting from r_0 in an Equality network, one can define a set of paths $P(r_0, r_j)$ to any other node r_j in the network. From $P(r_0, r_j)$, one can also find the shortest paths $\overrightarrow{P}(r_0, r_j)$ from r_0 to r_j and save the paths as the first routing table $\widehat{R}$ of the network.

From Remark 2, one can then derive the shortest paths $\overrightarrow{P}(r_i, r_j)$ between r_i and r_j by simple conversion. Since the network is symmetric, the shortest paths (in fact, any paths) between nodes depend on their respective relative difference modulus to total node number N in their IDs as described in Eq. 2.

$$P(r_i, r_j) \equiv \begin{cases} P(r_0, r_{j-i \bmod N}) \iff i\%2 = 0 \\ P(r_0, r_{i-j \bmod N}) \iff i\%2 \neq 0 \end{cases} \tag{2}$$

Apart from the shortest paths, $\overrightarrow{P}(r_0, r_j)$, we are also interested in the paths that are slightly longer, which are the sub-shortest paths $\overrightarrow{P}'(r_0, r_j)$, from r_0 to r_j.

For each target node r_j, the sub-shortest paths $\vec{P}'(r_0, r_j)$ can be evaluated by checking all neighbors r_k not in the shortest paths $\vec{P}(r_0, r_j)$, finding all the distances d_{0k} and pick the nodes r_k for all $d_{0k} \geq d_{0j}$ and collecting the list of nodes r_k for those where d_{0k} is minimal. The collected list L_{0j} of nodes r_k can then be incorporated with the shortest path routing table $\widehat{R}$ to form another routing table $\widehat{R}'$.

$\widehat{R}'$ is thus containing the shortest paths and alternative targets for sub-shortest paths.

The definition of $\widehat{R}'$ involves the knowledge of total router number N and the interconnection configuration without knowing the number of endpoints p for each router. Therefore, the complexity of the routing table of an Equality network is O(N) instead of O((Np)^2) in regards to an irregular network.

3.5 Routing algorithms

The routing algorithms including the *adaptive minimal* (abbreviated amin), *non-minimal UGAL* (global adaptive routing using local information, abbreviated ugal), and *bottleneck-UGAL* (abbreviated bgal) are provided in this work to assess the performance of the Equality networks under consideration.

3.5.1 amin

Adaptive minimal routing algorithm routes the packets through the shortest paths, where there may be one or more shortest paths.

From the universal routing table $\widehat{R}'$ the alternative intermediate router IDs mentioned above as L_{0j} are used for *ugal* and *bgal* routing algorithms.

3.5.2 ugal

ugal routing algorithm originally defined in [16] is implemented and simulated for comparison. In the current work, the packet is routed so that all the provided shortest and sub-shortest paths are considered and weighted based on the products of the paths' queue length and hop distance.

3.5.3 bgal

The bottleneck-UGAL routing is designed based on ugal algorithm; however, only in the bottleneck of all pair-wise relationships are allowed to utilize the sub-shortest paths. It differs from ugal in the quenching of the available sub-shortest paths when the number of the shortest paths is higher than a threshold value 2 (fixed in this work, but this value is adjustable). This means if the number of shortest paths exceeds the threshold value, only the shortest paths are included in the routing paths.

3.6 Cycle-accurate simulation conditions

3.6.1 Traffic models

Ten traffic models, including uniform random, asymmetric, random permutation, neighbor, tornado, bit rotation [5], bit complement, bit reverse, bit complement, and bit shuffle, are implemented in BookSim to evaluate the Equality networks. In the ten traffic models, only bit rotation is implemented locally, which performs a $d_i = s_{i+1} \mod b$ relation to set a target endpoint for each source.

3.6.2 Deadlock freedom

Deadlocks happen when several turning points wait for buffers of cyclic dependency. Deadlock-freedom is achieved with what was described in [17], which guarantees the deadlock-freedom by either limiting the routing to ensure cycle-freedom in the channel dependency graph [18] or utilizing virtual channels (VCs) to break such cycles into different sets of buffers [19].

The routing strategy used herein is similar to that introduced in [20, 21]. If we consider a packet sent from router r_i to r_j, we use the number of virtual channels equal to the diameter for minimal routing. If r_i and r_j are directly connected (i.e., one hop), then the packet is routed using VC0. If the path between r_i and r_j consists of two hops, then VC0 is used for the first hop, and VC1 is used for the second hop, respectively. Consider a network of diameter two; for example, only one turn can be taken on the path, and therefore, the maximum number of required VCs is two [22]. That way, one flit only depends on the virtual channel one above its current virtual channel to make progress. Thus, no cyclic dependency can happen.

For adaptive routing, on the other hand, the number of virtual channels is equal to the maximum hop of paths in $\widehat{R'}$. For $d = 2$ systems, the maximum hop in $\widehat{R'}$ is 4, and for $d \geq 3$ systems, the maximum hop in $\widehat{R'}$ is $d + 1$. BookSim gives an error when the number of VCs is insufficient. To generalize the algorithm above, we use a VCk ($0 \leq k < n$) on a hop k for an n-hop path between r_i and r_j.

The general parameters in all simulations are the same as described in [17], which are:

```
credit_delay   = 2; routing_delay  = 0; vc_alloc_delay = 1; sw_alloc_delay = 1;
st_final_delay = 1; input_speedup  = 1; output_speedup = 1; internal_speedup = 2.0;
```

Single flit packet is used to avoid flow control issues as described in [17] and [22], where the virtual channel buffer size is set to 64 flit entries, and the number of virtual channels is set as described above depending on the network diameter.

We then collected latency and throughput benchmarks under various traffic patterns and injection rates until the values converged.

4 Cost and balance

4.1 The balance of *K, p* and *a*

The most intuitive interpretation of K/p is the ratio of ports used for inter-router connections to the number of ports used for endpoints on each router. It is easy to see that if K/p is lower, the interconnect's price would be lower for a fixed number of endpoints.

In general, if budget is of principle consideration, Equality networks can be restricted in the range of $p \cdot (d+1) \geq P > p \cdot (a+1)$; alternatively, $pd \geq K > pa$. The K/p value can be relaxed if performance is of principal concern. The design can be fine-tuned depending on what applications will be run on the final system. Equality is probably the only topology that does not need to sacrifice any port to adjust this ratio. If the number of endpoints is reduced (reduced p), the empty ports can be used for larger K values.

On fat-tree systems, the value of K/pd is always one (and a is very close to d). For instance, a 3-tier fat-tree network utilizes two times the number of links to the endpoints in the number of connections between routers, making four times inter-router ports (each inter-router link consumes two ports) to the number of ports on average on each router. Coincidently, the diameter (maximum inter-router hop count) of a 3 L fat-tree network is four. The same idea applies to 2 L and 4 L fat-trees.

The behavior of the power model acts similar to the networking cost as the number of SerDes for the inter-router links to the SerDes for the endpoint links has the same ratio as K/p.

We calculated the total networking cost of all Equality systems with a model similar to what is described in [17], where each of the cabinets is assumed to be 1 m x 1 m without aisle in the cluster.[2] Each of the endpoints and routers is assumed to be 1U in size. The overhead cable pathways are 1 m above the cabinet. The endpoints and the routers are allocated sequentially on the cabinet until the cabinet is filled to 42U standard cabinet size. All routers are situated on the top of the cabinet. Depending on the remaining cabinet space, endpoints can be allocated in the same or adjacent cabinet to the router.

Manhattan distance is calculated for each cable to include the distance from each module to the overhead cable pathway, the space on the aisle, and an additional 0.5 m horizontal distance from the side of the cabinet to the port. Therefore, from this model, a cable to the adjacent module is 104.45 cm, i.e., 2×0.5 m + 44.45 mm. Cables longer than 8 m are fiber, otherwise copper. The results are summarized in Sect. 5.2.

[2] The cost of copper cables: $f(x) = 0.4079x + 0.5771$[\$/Gb/s], the cost of fiber cables: $f(x) = 0.0919 + 7.2745$[\$/Gb/s], and the cost of router: $f(k) = 350.4k - 892.3$[\$] are consistent to [17] for comparison.

4.2 Cost per node

From the result of the cost model presented by Besta et al. (Fig. 11(c) of [17]) and Kim et al. (Fig. 19 of [23]), it is evident that all topologies follow almost steady cost/endpoint ratio. While the copper cable reduces the networking cost in small systems, the effect is insignificant in larger systems. In practice, this ratio will still depend on the cable length; i.e., larger computing nodes consume larger space, leading to a higher proportion of inter-router connection cost. For any two non-torus networks with computing nodes of a given form factor (for instance, 0.5U or 2U per endpoint), regardless of the topology used, the copper cable prices of the two networks should be equal if the number of routers and servers are the same. The analysis is reported in Sect. 5.3.

4.3 Bisection ratio

By definition, the bisection bandwidth, B, is evaluated by cutting a minimal number of cables to separate the system into two parts.

Instead of discussing the bisection bandwidth, we introduce two variables named network bisection ratio (b_r) and topology bisection ratio (B_r) to take the networking cost into account. To clarify, the network bisection ratio is the bisection bandwidth divided by total network bandwidth (including all links to endpoints), i.e., $b_r = B/\Phi$, whereas the topology bisection ratio, B_r, is the bisection bandwidth divided by total inter-router bandwidth (excluding all links to endpoints). The total network bandwidth, Φ, for a direct network is the total number of cables in the system multiplied by the channel bandwidth ϕ, i.e., $\Phi = N\phi(\frac{K}{2} + p)$.

For a fully connected 12-port 10 Gb/s Ethernet switch, the bisection bandwidth is 60 Gb/s, which is half the total cable number multiplied by the channel bandwidth. It is easy to see that the bisection ratio for this fully connected switch is half, i.e., $b_r = 1/2$. For a two-tier fat-tree, this value becomes 1/4, as the number of cables in the network doubles, whereas for a three-tier fat-tree, the value of b_r is 1/6. The lower the bisection ratio, the higher the cost of networking hardware because a higher fraction of networking cost is contributed to the inter-router links.

For symmetric chordal ring networks like the Equality networks, the bisection bandwidth is equal to the minimum number of links that are cut if one splits the network into two semicircles. During the generation of new networks, the topology bisection ratio B_r is evaluated by counting the minimal number of links being cut when splitting the network into two equivalent halves.[3]

The relation of B_r and network bisection ratio b_r is shown in Eq. 3.

[3] The number of links to the endpoints, p, is a variable and therefore is not considered in the topology generation stage.

Table 2 Latency (lat@0.9) of uniform traffic under injection rate 0.9 flit/cycle using adaptive_min routing for networks demonstrated for a set of selected Equality networks. The detailed selected sets of links, S_A and S_B, are not included in this table for brevity. All throughput values of Equality networks in this table are to 0.9 flit/cycle. 3FT44 is the half-span fat-tree system simulated in [17] using 44-port switches. For the three fat-tree systems (3FT36, 3FT48, 3FT80), we assume the simulated results from BookSim also apply to full-span fat-tree systems. Equality networks with ID marked as boldface in the table are systems mentioned in the main text

ID	*N*	*K*	*p*	*n*	*a*	*K*/*p*	*d*	lat@0.9	MB%
E361	2048	28	8	16,384	2.717	3.50	3	23.45	9.66
E362	18,000	30	6	108,000	3.45	5.00	5	24	0.08
E363	16,000	29	7	112,000	3.51	4.14	5	29	0.09
E364	20,000	30	6	120,000	3.611	5.00	5	26.1	0.09
E365	20,000	29	7	140,000	3.648	4.14	5	32.7	0.11
E366	30,000	30	6	180,000	3.71	5.00	5	26.59	0.14
E367	30,000	29	7	210,000	3.74	4.14	5	33.3	0.16
E368	40,000	30	6	240,000	3.81	5.00	5	27.55	0.18
E369	200	24	12	2,400	1.879	2.00	2	31.01	34.7
***E481**	4800	38	10	48,000	2.805	3.8	4	23.27	0.24
E482	9000	40	8	72,000	2.975	5.00	4	21.55	0.36
E483	16,000	39	9	144,000	3.17	4.33	4	24.95	0.72
E484	20,000	39	9	180,000	3.262	4.33	4	26.2	0.91
E485	32,768	40	8	262,144	3.444	5.00	4	25.4	1.34
E486	36,000	39	9	324,000	3.486	4.33	4	29.62	1.64
E487	250	32	16	4,000	1.871	2.00	2	31.73	24.4
E801	2400	60	20	48,000	2.293	3.00	3	24.8	1.13
E802	16,384	64	16	262,144	2.822	4.00	4	23.8	0.1
E803	15,000	62	18	270,000	2.819	3.44	4	28.1	0.1
E804	20,000	62	18	360,000	2.888	3.44	4	28.47	0.14
E805	36,000	64	16	576,000	3.096	4.00	4	27.41	0.22
E806	64,000	64	16	1,024,000	3.224	4.00	4	29.03	0.4
***E807**	20,000	64	16	320,000	2.857	4.00	4	23.79	0.12
E808	1400	60	20	28,000	1.957	3.00	2	17.68	38.9
▷SF14	722	29	15	10,830	1.963	1.93	2	≫50	85.7
***‡E441**	900	32	12	10,800	2.346	2.67	3	28.67	2.83
***‡E442**	1000	33	11	11,000	2.38	3	3	25.14	2.87
***‡E443**	800	31	13	10,400	2.3	3	3	38.7	2.7
‡ ▷3FT44	1210	–	–	10,648	3.953⋄	4	4	36.3	–
▷DF14	1386	20	7	9702	2.896	2.86	3	≫50	18.2
◁T3D	10,648	6	1	10,648	16.5	6.0	33	≫50	–
◁T5D1	7776	10	1	7776	7.5	10.0	15	≫50	–
◁T5D2	16,807	10	1	16,807	8.5	10.0	15	≫50	–
▷DF08	264	11	4	1,056	2.783	2.75	3	≫50	21.6
3FT36	1620	–	–	11,664	3.943⋄	4†	4	35.2	–
3FT48	2880	–	–	27,648	3.957⋄	4†	4	36.4	–
3FT80	8000	–	–	128,000	3.975⋄	4†	4	37.9	–

Table 2 (continued)

ID	N	K	p	n	a	K/p	d	lat@0.9	MB%
SF16	53,138	255	19	1,009,622	1.995	13.42	2	N/A	81.7

* Systems where latency or throughput values are reported in this work.

† The effective K/p values for fat-tree systems are provided for reference purposes. For fat-tree systems, the number of wires for inter-router connections is $n(L-1)$, L being the number of layers. The total port number for inter-router connections is $2n(L-1)$

‡ Not shown in Fig. 2a and b to prevent overlapping with Slim Fly networks.

⋄ The a value here is calculated based on full-span fat-tree.

◁ T3D is a 22-ary 3-cube, T5D1 is a 6-ary 5-cube, and T5D2 is a 7-ary 5-cube. The number of virtual channels used is 4, and the virtual channel buffer size is 8.

▷ Citation for SF14, DF14 and 3FT44: [17]. Citation for DF08: [23]. Citation for SF16: [27]

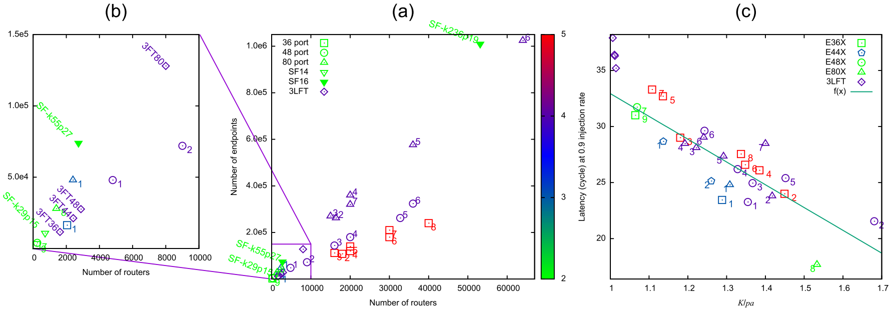

Fig. 2 **a** Various sizes of systems under consideration. The palette on the right-hand side denotes the network diameter of the system. **b** Zoom in the range where the router number is less than 10,000. **c** Latency against K/pa under injection rate 0.9 flit/cycle. Fitting function $f(x) = ax + b$, where $a = -20.3211$ (asymptotic standard error $\pm 2.736(13.46\%)$) and $b = 53.255$ (asymptotic standard error $\pm 3.577(6.717\%)$) fits against all Equality networks in Table 2. The Slim Fly results are shown as inverted triangles, whereas the fat-tree results are shown as diamonds. For all systems listed in the above sub-figures, the numbers next to the symbol represent the identification of the system in their respective router radix category. For instance, a triangle with six next to it represents the sixth system using 80-port routers; therefore, the system ID for this system is E806, as listed in Table 2

$$b_r = \frac{B}{\Phi} = \frac{\phi B_r \frac{NK}{2}}{\frac{N\phi(K+2p)}{2}} = \frac{B_r K}{K+2p} = B_r \frac{\frac{K}{p}}{\frac{K}{p}+2} \tag{3}$$

The data are collected and discussed in Sect. 5.4.

5 Results

We have selected a few publications as targets for comparison purposes. To make a sensible comparison, most networks utilize switches with the port number available in the market.

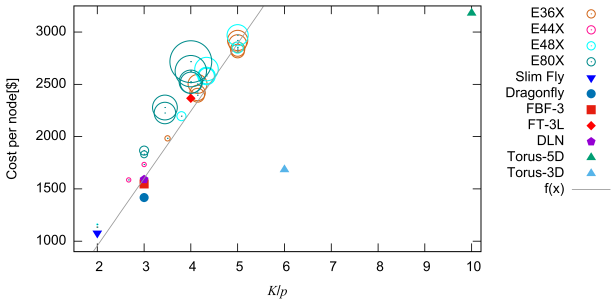

Fig. 3 Line $f(x) = 644.5x - 333.1$ is drawn to fit cost-per-node against K/p for networks with Slim Fly, Dragonfly, FBF-3, FT-3 L and DLN. The tori follow another trend as only copper cables are used. Both trends are considered to depend on K/p. The cost-per-node values of these topologies are only a rough estimation for systems under 40K endpoints, whereas the Equality systems are plotted in circles of different sizes depending on the number of endpoints. The largest circle is the E806 systems with 1 M endpoints. While other topologies use different symbols of the point size 1 for clarity, the Equality system circle point sizes are adjusted by $\sqrt{n/40000}$ points

5.1 Targeted systems

Equality can be used to achieve reasonably good ratios of the Moore bounds. We include a column in Table 2 to address the ratio of the network size against the Moore bound. We obtained networks reaching ratios 39% of diameter 2 Moore bound, 9.66% of diameter 3 Moore bound, 1.64% of diameter 4 Moore bound, and 0.18% of diameter 5 Moore bound.

We simulated networks of scales from small to 1,024,000 endpoints in this work with various router counts and radices. The listed systems, named according to the radix of the routers (i.e., E361 represents the first system using 36-port routers), are selected networks for each configuration. The listed networks are included in Appendix 1.

We have chosen carefully by optimizing the configuration as discussed in Sect. 3.3 and hand-picked one network with better performance in all traffic modes for each configuration. Some of the networks in the table are designed for exascale systems. The values of the torus, 3-tier fat-tree, Dragonfly, and Slim Fly networks are there for reference.

A Slim Fly network denoted by SF14 is compared with three diameter-3 Equality networks (E441, E442, and E443) in the table to show that with a bit of relaxation in router number, Equality allows better throughput using identical hardware specification. Slim Fly in this comparison, the latency under injection rate 0.9 flit/cycle is over 50 cycles, whereas all three Equality systems are under 50 cycles.

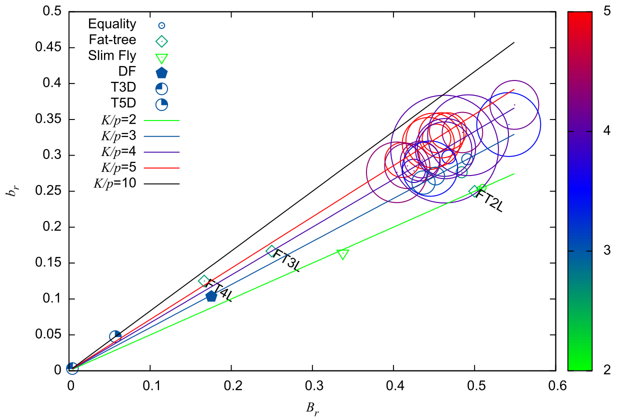

Fig. 4 The relation between b_r and B_r for all Equality systems listed in Table 2. The color of the circle follows the palette of *K*/*p*. Higher *K*/*p* gives a higher b_r for a given B_r value. The circle sizes of Equality systems have been adjusted by $\sqrt{n/3000}$. The FT-2 L diamond falls on the line of $K/p = 2$ as the Slim Fly $d = 2$ systems. The FT-3 L has effective $K/p = 4$, $B_r = 0.25$ and $b_r = 1/6$. The FT-4 L has effective $K/p = 6$, $B_r = 1/6$ and $b_r = 0.125$. The Slim Fly point takes the value $b_r = B/\Phi$, where $\phi = 10$ [Gb/s], $B = 60736$ [Gb/s](approx. from Fig. 5(c) of [17]) and $\Phi = N\phi(K/2 + p) = 370300$ [Gb/s] ($K = 34$, $p = 18$, and $N = 1058$). The Dragonfly point takes the values $B = 41666$ [Gb/s](approx.) and $\Phi = N\phi(K/2 + p) = 402480$ [Gb/s], where $N = 2064$, $K = 23$, $p = 8$. The 3D-torus point takes the values $B = 12654$ (approx.), where $N = 15625$ (i.e. 25^3), $K = 6$, and $p = 1$. The 5D-torus point takes the values $B = 48006$ (approx.), where $N = 16807$ (i.e., 7^5), $K = 10$, and $p = 1$

Figure 2a and b shows the comparison of scales (number of routers and endpoints) of the listed systems in the table. It can be seen that the maximum sizes 3-tier fat-trees can achieve are far smaller than Equality.

5.2 The balance of *K*, *p* and *a*

Figure 2(c) shows that the latency at the package injection rate 0.9 flit/cycle is negatively correlated with *K*/*pa* with amin routing algorithm. The value *K*/*pa* stands for the switch's balance of upward and downward flow ratios. The higher the *a* value, the higher the frequency a message has to consult the switches. By adjusting the weight of *K*/*p*, one can counterbalance the weight of *a*.

5.3 Cost per node

Figure 3 shows that with *K*/*p* being the variable, the 'cost per endpoint' values for all topologies (here shows Slim Fly, Dragonfly, FBF-3, FT-3 L, and DLN) follow the same trend if fiber cables are used for longer links. On the other hand, if only copper cables are used, the cost behaves like the tori. Equality systems are shown in circles,

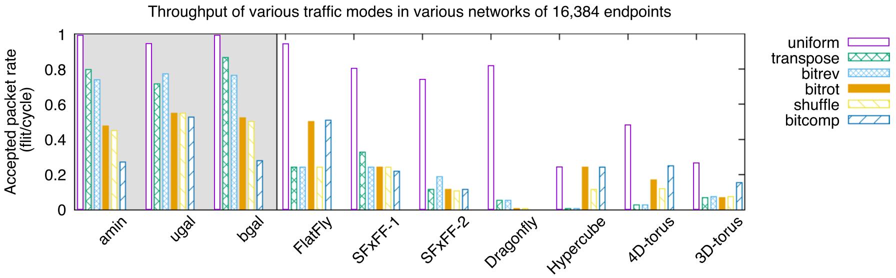

Fig. 5 Throughput comparison chart for networks of 16,384 endpoints. Columns in the gray box contain three routing algorithms amin, ugal, bgal running on the target Equality E361 system with the configuration of n2048k28p8 using 36-port switches. The data from the right panel contain the best results of the respective networks from [7]

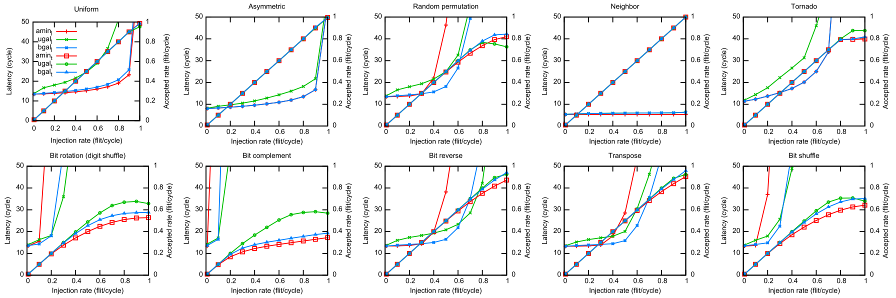

Fig. 6 Comparison of throughput and latency of the Equality E481 system with the configuration of n4800k38p10 using 48-port switches in 10 traffic modes. In total, there are 48,000 endpoints in this system. The transpose traffic uses 16,384 endpoints for calculation, whereas the other bit permutation traffics uses 32,768 endpoints for calculation

which follow the same trend as the fitted line, only that the system sizes are much larger. A higher number of endpoints will slightly increase the average networking cost per node, but not as significant as the ratio of *K*/*p*. The cluster model can be adjusted to include hot/cold aisle and different rack sizes, but it is not in the scope of the current study.

5.4 Bisection ratio

Figure 4 shows the distribution of the Equality networks compared with 3D-torus, 5D-torus, Slim Fly, Dragonfly, and fat-tree networks. The location of the network in this graph depends on the *K*/*p* and b_r values of the respective network. Since b_r/B_r is a function of *K*/*p*, the Slim Fly network sits close to the line of $K/p = 2$, where two of the Equality networks (E369 and E487) fall on that line (near FT2L). The

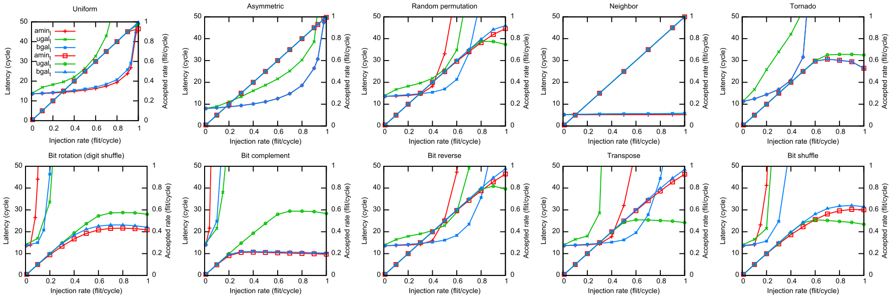

Fig. 7 Comparison of throughput and latency of the Equality E807 system with the configuration of n20000k64p16 using 80-port switches in 10 traffic modes. In total, there are 320,000 endpoints in this system. All bit permutation traffics uses 262,144 endpoints for calculation

distribution of $\mathbf{S_A}$ and $\mathbf{S_B}$ defines B_r in Equality networks. The fat-tree networks have good b_r in two tiers, whereas it degrades as the number of layers increases. The b_r values of tori also explain why uniform traffic is the nightmare of torus networks.

The introduction of bisection ratios b_r and B_r gives a new viewpoint to look at the bisection bandwidth of a network. The question becomes, "*With the constraint of the networking budget, what percentage of the budget is contributed into the bisection bandwidth?*," instead of "*How much is the bisection bandwidth of the network?*." The point is not to get a large bisection ratio, as one needs a proper balance to communicate with the nearby and remote routers. Our experiments found that B_r around 0.4 ~0.5 and b_r around 0.3 are suitable for most traffic models, where global and local traffic ratios are at a balance.

5.5 Individual scenarios

The 16, 384-endpoint scenario is prepared for the comparison to [7]. Figure 5 shows the throughput performance of the E361 network, with the best results shown in [7]. For the Equality network, three routing algorithms: amin, ugal, and bgal are shown inside the grayed box on the left-hand side. The other seven blocks: "FlatFly, SFxFF-1, SFxFF-2, Dragonfly, Hypercube, 4D-torus, and 3D-torus" are the best values directly taken from the paper. It is apparent that almost in all traffic modes, the E361 network using three routing algorithms, with the same number of switches of the same radix, performs better than all networks presented in [7].

It shows that the *ugal* algorithm performs akin to FlatFly in the bit complement traffic model, where all other topologies fail. Although fat-tree performs well in permutation traffic, we do not include it in this comparison because the maximum size 3-layer fat-tree can achieve with 36-port routers has only 11,664 endpoints.

The 48, 000-endpoint scenario is prepared for the situation where the InfiniBand LID limit is of major concern. Detailed simulations on ten traffic models are demonstrated in graphs containing both latency and throughput results.

All the injection processes are simulated in various injection frequencies to reflect the spectrum of latency values under 50 cycles. We focus on latencies lower than 50 cycles as they reflect the usable range of the system that is reliable under the given injection process. Figure 6 presents all ten traffic processes.

The 1E scenario features systems for future applications with multi-exaflops peak performance depending on the computing power of each endpoint. Figure 7 shows the ten traffic models. The E807 network tested here has the configuration of n20000K64p16. The 48, 000-endpoint and 1E networks with close *K*/*pa* values are here to show the scalability of Equality networks. With a significant increase in the system size, the performance is still consistent across the two networks.

The 10^6-endpoint system is a single-point (simulating at 0.9 flit/cycle injection rate) simulation to achieve a million endpoints. We only show the results of throughput and latency of amin routing algorithm under an injection rate of 0.9 flit/cycle for the E806 system with the configuration of n64000k64p16 in Table 2 having 1,024,000 endpoints. The simulation time of the BookSim package running this million-endpoints system took a maximum of 272 GB memory running on single-core in bitcomp traffic. As the only large-memory resource on our site is an IBM box, the calculation took around 20 days to complete. At this point, the BookSim package fails with an integer error, presumably due to the integer type range limitation of the _cur_pid variable, but all simulations are reasonably converged before the job ends.

All of the above systems show the flexibility and performance of the Equality networks.

6 Conclusion and future work

This paper shows that the low memory footprint routing logic of Equality networks is natively born for routing algorithms using minimal adaptive and sub-shortest paths. With the topology setting and routing logic implemented in BookSim, we demonstrate the simulation results from small to large scales. We also perform the first million-node cycle-accurate calculations based on BookSim package.

Many have advocated [17, 24–26] the use of low-diameter networks for the realization of enormous network size with high radix routers. Table 2 shows that excellent performance persists in Equality networks with reasonably low diameters and ordinary router radix under high injection rates. Conversely, extremely low-diameter topologies usually involve networks that are not flexible or need very high-radix routers. For the 1 M-endpoint network, the current work provides a solution to build a network of 64,000 routers (80-port) (shown in Table 2) for the replacement of the network solution of 53,138 routers, each with 264 ports reported in [27]. Moreover, the resulting network performance is plausible. We have shown that Equality networks have similar or better performance compared to many other topologies, including Tori, Dragonfly, Slim Fly, Flat Fly, and Fat-Tree.

Equality does not do permutation (e.g., bit complement and bit shuffle) quite as well as fat-tree. Still, it generally trades the zero-load latency with the throughput while keeping moderate- to high-level usabilities.

Equality networks can be used with low- or high-radix routers [28] available in the market. The results of this paper also show excellent performance of Equality networks under uniform random traffic even with an adaptive minimal routing algorithm, which suggests good performance using commodity hardware for general-purpose clusters. We have also performed routing a mini-sized Equality network on 12 HPE 5945 48SFP28 8QSFP28 switches for real applications. The routing of this small cluster is accomplished with multi-protocol label switching (MPLS), where all shortest paths and some sub-shortest paths are assigned between every pair of routers. On InfiniBand, we expect the network will work well with Nue routing introduced by Domke et al. [29].

The adjustability and outstanding performance of Equality networks give the industry a new network topology option for HPC systems. The system designers will have more flexibility in picking commodity hardware. It is also an opportunity for data centers to reorganize for better efficiency.

In future, we plan to run simulations using ROSS-CODES and TraceR as described in [30] to reflect the effects of the real application traffic, especially for other subtle effects while dealing with many jobs on the cluster [31].

To summarize, the current work reiterates the construction of networks based on a novel class of network topology, allowing routing with simple logic to achieve strong scalability from small to large systems. The work shows the performance of networks by comparing the cycle-accurate BookSim benchmarks against many previous works. The results show significant benefits of utilizing this new class of network topology for future high-performance computing applications.

Appendix 1 Configuration of the listed Equality networks

Table 3 lists the configurations of the Equality networks simulated in this work.

Table 3 Configurations of listed Equality networks

ID	Configuration
E361	n2048k28p8 ahops:[−1, 1, 101, 115, 191, 321, 387, 447, 481, 519, 697, 843, 925, 989, 1125, 1165, 1391, 1513, 1879, 1895] bhops:(200,410,614,824)
E362	n18000k30p6 ahops:[−1, 1, 55, 623, 851, 1729, 2401, 2519, 3535, 4239, 4517, 4703, 6241, 6959, 9851, 10,069, 10,077, 13,341, 15,155, 15,529, 15,853, 17,485] bhops: (1700,3500,5500,7300)
E363	n16000k29p7 ahops:[−1, 1, 815, 2301, 2523, 2807, 2967, 6819, 6985, 7603, 8275, 10,301, 11,043, 11,131, 11,893, 13,259, 14,189, 14,389, 14,875, 14,999, 15,553] bhops:(1500,3100,4900,6500)
E364	n20000k30p6 ahops:[−1, 1, 1329, 2155, 3889, 3937, 4265, 4843, 6133, 6235, 6463, 7497, 8015, 10,989, 11,015, 12,155, 14,573, 15,051, 16,235, 18,015, 18,293, 18,397] bhops:(3900,7900,2100,6100)
E365	n20000k29p7 ahops:[−1, 1, 1419, 1771, 2057, 2155, 2967, 5213, 5905, 6235, 6755, 8015, 8031, 12,155, 12,239, 12,693, 16,235, 16,441, 18,015, 18,547, 18,605] bhops:(3900,7900,2100,6100)
E366	n30000k30p6 ahops:[−1, 1, 1501, 2761, 2883, 3585, 5017, 6545, 6625, 8897, 9619, 11,095, 15,795, 15,849, 16,323, 16,501, 21,937, 22,325, 22,549, 25,811, 26,549, 28,215] bhops:(2900,5900,9100,12,100)
E367	n30000k29p7 ahops:[−1, 1, 1551, 3351, 3453, 5019, 6533, 11,769, 12,523, 12,923, 14,713, 14,927, 16,551, 16,657, 18,475, 19,047, 19,463, 20,825, 21,213, 22,283, 23,309] bhops:(3100,6100,8900,11,900)
E368	n40000k30p6 ahops:[−1, 1, 1825, 1893, 1901, 2719, 4953, 6429, 6747, 7825, 8693, 8811, 10,507, 12,729, 13,591, 13,743, 16,755, 21,901, 25,167, 28,359, 28,749, 29,395] bhops:(3900,7900,12,100,16,100)
E369	N200k24p12 ahops:[−1,1,11,13,19,35,39,59,97,107,109,115,117,137,155,157,187,193,195] bhops:(34,66,100)
E481	n4800k38p10 ahops:[−1, 1, 179, 245, 289, 389, 537, 541, 703, 731, 771, 1201, 1317, 1621, 1671, 2129, 2155, 2539, 2645, 2725, 3027, 3131, 3601, 3919, 4023, 4071, 4245, 4421, 4555, 4781] bhops:(490,970,1430,1910)
E482	n9000k40p8 ahops:[−1, 1, 451, 551, 929, 1175, 1259, 1351, 1691, 1919, 2201, 2251, 2477, 3151, 3865, 4049, 4083, 4787, 4951, 5131, 5649, 5851, 5861, 6751, 7185, 7281, 7651, 7951, 8237, 8481, 8549, 8637] bhops:(904,1796,2704,3596)
E483	n16000k39p9 ahops:[−1, 1, 43, 129, 751, 1889, 2301, 2417, 4001, 4073, 5477, 5833, 7025, 7693, 8751, 9347, 9413, 9903, 10,041, 10,301, 10,419, 10,749, 10,851, 12,001, 12,113, 12,229, 12,631, 13,465, 13,493, 14,027, 14,659] bhops:(1500,3100,4900,6500)
E484	n20000k39p9 ahops:[−1, 1, 405, 573, 1035, 2663, 3025, 3059, 3845, 4519, 5045, 7055, 7715, 9611, 9633, 10,863, 11,035, 13,025, 13,143, 14,827, 15,045, 15,157, 15,369, 16,525, 16,983, 17,055, 18,647, 18,723, 19,449, 19,733, 19,961] bhops:(1900,8100,3900,6100)
E485	n32768k40p8 ahops:[−1, 1, 1535, 1669, 2703, 3795, 4405, 4685, 4949, 5769, 7755, 9727, 9829, 10,435, 11,087, 12,877, 13,269, 15,947, 18,075, 19,287, 22,715, 22,791, 25,653, 27,149, 27,245, 27,325, 29,007, 32,681] bhops:(2600,13,784,5100,11,284,7100,9284)
E486	n36000k39p9 ahops:[−1, 1, 821, 893, 1473, 1751, 2701, 3719, 3861, 4219, 5401, 5989, 6677, 6745, 6947, 7493, 9901, 11,169, 11,429, 12,639, 14,503, 16,063, 19,751, 21,737, 23,401, 26,123, 26,783, 29,013, 32,253, 33,035, 34,393] bhops:(3500,7300,10,700,14,500)
E487	n250k32p16 ahops:[−1, 1, 9, 17, 21, 35, 37, 57, 65, 75, 83, 89, 109, 115, 125, 133, 151, 155, 163, 169, 199, 221, 241, 243] bhops:(24, 46, 78, 102)
E801	n2400k60p20 ahops:[−1, 1, 13, 31, 55, 165, 171, 217, 267, 287, 305, 327, 361, 405, 523, 605, 705, 709, 815, 975, 1041, 1073, 1255, 1293, 1365, 1505, 1577, 1605, 1663, 1719, 1737, 1805, 1869, 1905, 2015, 2167, 2239, 2291, 2381, 2389] bhops:(110, 230, 350, 470, 550, 650, 730, 850, 970, 1090)

Table 3 (continued)

ID	Configuration
E802	n16384k64p16 ahops:[−1, 1, 425, 721, 1165, 1435, 1501, 1659, 1869, 2095, 2581, 2635, 2751, 2809, 2961, 3245, 3667, 3767, 4635, 5727, 5867, 6617, 6881, 6885, 7153, 7165, 8617, 9065, 9373, 9627, 9833, 10,173, 10,629, 10,649, 10,741, 11,437, 11,873, 12,201, 12,297, 12,511, 13,227, 13,259, 14,545, 14,723, 15,933, 16,163, 16,275, 16,323] bhops:(900,1900,2800,3700,7292,6292,5392,4492)
E803	n15000k62p18 ahops:[−1, 1, 73, 411, 547, 779, 843, 1025, 1657, 1697, 1899, 1935, 2491, 2589, 2611, 2703, 2755, 3575, 4239, 4845, 5133, 5685, 5703, 6323, 6997, 7541, 8525, 8619, 8939, 9281, 9435, 9553, 10,255, 10,477, 10,907, 11,013, 11,075, 11,237, 11,641, 12,345, 13,073, 13,185, 13,531, 14,529, 14,667, 14,899] bhops:(800,1500,2300,3100,4400,5200,6000,6700)
E804	n20000k62p18 (Br 0.542) ahops:[−1, 1, 505, 783, 871, 1515, 2045, 2525, 3535, 3581, 3659, 4545, 5995, 6471, 6751, 6807, 7153, 7685, 8589, 8739, 8855, 9049, 9167, 9743, 10,359, 10,383, 10,505, 11,227, 11,473, 11,515, 11,673, 12,525, 13,355, 13,535, 14,349, 14,545, 15,365, 15,407, 15,433, 16,037, 17,103, 17,829] bhops:(900,9100,1900,8100,2900,7100,3900, 6100,4500,5500)
E805	n36000k64p16 ahops:[−1, 1, 13, 157, 335, 1057, 2349, 3489, 4075, 4177, 4387, 4491, 4567, 5747, 5913, 9745, 10,733, 11,063, 12,531, 13,391, 15,885, 17,999, 18,001, 18,013, 18,157, 18,223, 18,335, 19,057, 20,349, 20,583, 21,373, 21,489, 22,327, 22,567, 23,913, 25,599, 27,209, 27,695, 29,011, 29,461, 31,721, 33,269, 34,617, 35,829] bhops:(2222, 4098, 6328, 7336, 8954, 9046, 10,664, 11,672, 13,902, 15,778)
E806	n64000k64p16 ahops:[−1, 1, 445, 725, 1751, 2415, 2957, 5301, 5931, 7161, 9169, 11,601, 11,843, 13,007, 13,187, 13,499, 15,115, 16,001, 16,745, 18,003, 22,965, 23,031, 24,103, 26,701, 27,687, 28,455, 30,251, 30,651, 31,215, 31,795, 33,751, 37,301, 38,681, 39,319, 41,633, 45,683, 45,907, 48,001, 50,949, 51,417, 55,859, 56,573, 57,879, 58,701, 58,927, 59,455, 59,745, 62,251] bhops:(3500,7100, 10,600, 14,100, 17,900, 21,400, 24,900, 28,500)
E807	n20000k62p18 ahops:[−1, 1, 505, 783, 871, 1515, 2045, 2525, 3535, 3581, 3659, 4545, 5995, 6471, 6751, 6807, 7153, 7685, 8589, 8739, 8855, 9049, 9167, 9743, 10,359, 10,383, 10,505, 11,227, 11,473, 11,515, 11,673, 12,525, 13,355, 13,535, 14,349, 14,545, 15,365, 15,407, 15,433, 16,037, 17,103, 17,829] bhops:(900,9100,1900,8100,2900,7100,3900,6100,4500,5500)
E808	1400K60p20 ahops:[−1, 1, 13, 3, 33, 47, 49, 67, 81, 89, 101, 121, 131, 147, 183, 225, 259, 321, 377, 387, 409, 459, 503, 511, 519, 525, 595, 605, 617, 717, 755, 835, 859, 915, 919, 1003, 1157, 1183, 1185, 1313, 1355] bhops:(70, 140, 210, 280, 350, 420, 490, 560, 630, 700)
E441	N900k32p12 ahops:[−1,1,23,25,55,121,135,165,177,333,457,475,495,543,549,557,585,615,717, 727] bhops:(70,130,194,256,320,360)
E442	N1000k33p11 ahops:[−1, 1, 27, 39, 45, 105, 215, 327, 365, 401, 455, 491, 523, 545, 547, 605, 653, 701, 715, 771, 801, 813, 865, 875, 955] bhops:(70, 180, 320, 430)
E443	N800k31p13 ahops:[−1, 1, 27, 39, 45, 105, 215, 327, 365, 401, 455, 491, 523, 545, 547, 605, 653, 701, 715, 771, 801, 813, 865, 875, 955] bhops:(70, 180, 320, 430)

Acknowledgments The authors acknowledge incisive feedback from all the reviewers. This material is partially supported by the Air Force Office of Scientific Research under award number <FA9550-16-1-0499>. Morphing Supercomputer INC partially supports this material by backing the research budget.

Author contributions Chi-Hsiu Liang did the major invention of the network topology, 90% simulation coding, and manuscript writing. Chun-Ho Cheng and Hong-Lin Wu did the simulation of all the networks. Chao-Chin Li maintained the IT hardware and provided hardware support in our group. Po-Lin Huang did 10% of the simulation coding and its extending applications. Chi-Chuan Hwang, the PI, was the initiator of the project and provided most of the ideas committing to the brainstorming in this project.

Funding Ministry of Science and Technology, Taiwan, MOST 102-2911-I-006-301. Air Force Office of Scientific Research, United States, FA9550-16-1-0499. Morphing Supercomputer INC.

Data availability Not applicable.

Declarations

Conflict of interest Not applicable.

Ethical approval Not applicable.

References

1. Yang CY, Liang CH, Wu HL, Cheng CH, Li CC, Chen CM, Huang PL, Hwang CC (2019) Exceeding the performance of two-tier fat-tree: equality network topology. Future of Information and Communication Conference, 14. Springer, Cham, pp 1187–1199
2. Liang CH, Cheng CH, Wu HL, Li CC, Chen CM, Huang PL, Huang SL, Hwang CC. (2018) Beyond the performance of three-tier fat-tree: equality topology with low diameter. In 2018 international symposium on computer, consumer and control (IS3C) (pp. 22-29). IEEE
3. Wu HL, Cheng CH, Liang CH, Li CC, Chen CM, Huang PL, Huang SL, Hwang CC. (2020) Beyond the performance of 3D-torus: equality topology with low radix. In 2020 international symposium on computer, consumer and control (IS3C) (pp. 319-322). IEEE
4. Cheng CH, Wu HL, Liang CH, Li CC, Chen CM, Huang PL, Huang SL, Hwang CC. (2020) Equality NoC: a novel NoC topology for high performance and energy efficiency. In 2020 international symposium on computer, consumer and control (IS3C) (pp. 83-86). IEEE
5. Dally WJ, Towles BP (2004) Princ Pract Interconnect Netw. Elsevier
6. Lutomirski A, Tegmark M, Sanchez NJ, Stein LC, Urry WL, Zaldarriaga M (2011) Solving the corner-turning problem for large interferometers. Mon Not R Astron Soc 410(3):2075–80
7. Daryin A, Korzh A (2015) Early evaluation of direct large-scale InfiniBand networks with adaptive routing. Supercomput Front Innov 1(3):56–69
8. Dongarra JJ, Meuer HW, Strohmaier E (1997) TOP500 supercomputer sites. Supercomputer 15(13):89–111
9. Jiang N, Balfour J, Becker DU, Towles B, Dally WJ, Michelogiannakis G, Kim J. (2013) A detailed and flexible cycle-accurate network-on-chip simulator. In performance analysis of systems and software (ISPASS), 2013 IEEE international symposium on (pp. 86-96). IEEE
10. Beivide R, Martínez C, Izu C, Gutierrez J, Gregorio JÁ, Miguel-Alonso J (2003) Chordal topologies for interconnection networks. International symposium on high performance computing, 20. Springer, Berlin, pp 385–392
11. Parhami B (2008) Periodically regular chordal rings are preferable to double-ring networks. J Interconnect Netw 9:99–126
12. Dubalski B, Bujnowski S, Ledzinski D, Zabludowski A, Kiedrowski P. (2012) Analysis of modified fifth degree chordal rings. In New Frontiers in Graph Theory. InTech
13. Faraha RN, Chienb SLE, Othmanca M (2016) Graph theoretical properties of degree six 3-modifiled chordal ring networks. J Eng Appl Sci 11(9):1987–1999

14. Parhami B. (1995) Periodically regular chordal ring networks for massively parallel architectures. In frontiers of massively parallel computation, Proceedings. Frontiers' 95., fifth symposium on the (1995) pp. 315-322, IEEE
15. Zabłudowski Ł, Dubalski B, Kiedrowski P, Ledziński D, Marciniak T (2012) Modified NDR structures. Image Process Commun 17(3):29–45
16. Singh A. (2005) Load-balanced routing in interconnection networks (Doctoral dissertation, Stanford University)
17. Besta M, Hoefler T. (2014) Slim Fly: A cost effective low-diameter network topology. In High Performance Computing, Networking, Storage and Analysis, SC14: International Conference (pp. 348-359). IEEE
18. Duato J (1995) A necessary and sufficient condition for deadlock-free adaptive routing in wormhole networks. IEEE Trans Parallel Distrib Syst 6(10):1055–67
19. Dally WJ, Seitz CL. (1988) Deadlock-free message routing in multiprocessor interconnection networks. California Institute of Technology. (Unpublished)
20. Duato J, Yalamanchili S, Ni LM (2003) Interconnection networks: an engineering approach. Morgan Kaufmann
21. Gopal IS. (1994) Interconnection networks for High-performance parallel computers. chapter Prevention of Store-and-forward Deadlock in computer networks., p. 338-344. IEEE computer society press, Los Alamitos, CA
22. Kim J, Balfour J, Dally W. (2007) Flattened butterfly topology for on-chip networks. In microarchitecture. MICRO 2007. 40th annual IEEE/ACM international symposium on 2007 (pp. 172-182). IEEE
23. Kim J, Dally WJ, Scott S, Abts D (2008) Technology-driven, highly-scalable dragonfly topology. ACM SIGARCH Comput Archit News. IEEE Comput Soc 36(3):77–88
24. Kathareios G, Minkenberg C, Prisacari B, Rodriguez G, Hoefler T. (2015) Cost-effective diameter-two topologies: analysis and evaluation. In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (p. 36). ACM
25. Mubarak M, Carothers CD, Ross R, Carns P. (2012) Modeling a million-node dragonfly network using massively parallel discrete-event simulation. In high performance computing, networking, storage and analysis (SCC), 2012 SC Companion (pp. 366-376). IEEE
26. Kim J, Dally WJ, Abts D. (2007) Flattened butterfly: a cost-efficient topology for high-radix networks. In ACM SIGARCH Computer Architecture News (pp. 126-137). ACM
27. Wolfe N, Carothers CD, Mubarak M, Ross R, Carns P. (2016) Modeling a million-node slim fly network using parallel discrete-event simulation. In Proceedings of the 2016 annual ACM Conference on SIGSIM Principles of Advanced Discrete Simulation (pp. 189-199). ACM
28. Alistarh D, Ballani H, Costa P, Funnell A, Benjamin J, Watts P, Thomsen B. (2015) A high-radix, low-latency optical switch for data centers. In ACM SIGCOMM computer communication review (pp. 367-368). ACM
29. Domke J, Hoefler T, Matsuoka S. (2016) Routing on the dependency graph: a new approach to deadlock-free high-performance routing. In proceedings of the 25th ACM international symposium on high-performance parallel and distributed computing (pp. 3-14)
30. Jain N, Bhatele A, White S, Gamblin T, Kale LV (2016) Evaluating HPC networks via simulation of parallel workloads. Framework 20(21):22
31. Yang X, Jenkins J, Mubarak M, Ross RB, Lan Z. (2016) Watch out for the bully!: job interference study on dragonfly network. In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, (p. 64). IEEE Press