PolarStar：拓展直径为3的网络的边界

Kartik Lakhotia
kartik.lakhotia@intel.com 英特尔 圣克拉拉，加利福尼亚州，美国

Laura Monroe
lmonroe@lanl.gov 洛斯阿拉莫斯国家实验室 洛斯阿拉莫斯，新墨西哥州，美国

Kelly Isham
kisham@colgate.edu
科尔盖特大学
美国纽约州汉密尔顿

Maciej Besta
mbesta@inf.ethz.ch 苏黎世联邦理工学院 苏黎世，瑞士

Nils Blach
nils.blach@inf.ethz.ch 苏黎世联邦理工学院 苏黎世，瑞士

Torsten Hoefler
torsten.hoefler@inf.ethz.ch
苏黎世联邦理工学院
苏黎世，瑞士

Fabrizio Petrini
fabrizio.petrini@intel.com
英特尔
圣克拉拉 , CA, USA

arXiv:2302.07217v2 [cs.NI] 6 Aug 2024

摘要

我们提出了 PolarStar，这是一类新颖的直径为 3 的网络拓扑，由低直径因子图的星积构造而成。

对于几乎所有基数，PolarStar 都能提供已知规模最大的直径为 3 的网络拓扑，从而构成目前已知最佳的可扩展直径为 3 的网络。与当前最先进的直径为 3 的网络相比，PolarStar 的规模几何平均值分别比 Bundlefly 增加 1.3×、比 Dragonfly 增加 1.9×、比 3-D HyperX 增加 6.7×。PolarStar 还具备许多其他理想特性，包括模块化布局、高二分带宽、对链路故障的高韧性，以及针对每种基数的大量可行配置。

我们通过模拟合成和真实世界的流量模式进行了详细评估，结果表明，PolarStar 的性能与当前直径为 3 的网络相当或更优。

CCS 概念

• **网络** → **网络架构**；**网络性能评估**；**数据中心网络**；• **计算数学** → **图论**；**极值图论**。

关键词

网络；高性能网络；网络拓扑；网络评估；极值图论

1 引言

1.1 研究动机

目前规模最大的超级计算机包含数万至数十万个处理节点。例如，Frontier 是 TOP500 榜单中性能最强的系统 [49]，，配备 9，408 个中央处理器和 37，632 个图形处理器 [13]。排名第四的超级计算机富岳拥有 158，976 处理节点 [20]。未来的超级计算机规模将会更大，因为系统规模决定了峰值计算性能。

直径同样至关重要，因为它会影响通信延迟以及每台交换机的注入带宽。低直径网络，尤其是直径为 2 [33] 和 3 [36]，的网络，备受关注；它们能够提供低延迟且经济高效的高带宽通信基础设施。

我们在此探讨这个问题：使用给定基数的交换机构建时，最大的直径为 3 的网络拓扑能有多大?

当前的技术进步使这个问题尤为迫切。高岸线密度的高基数光学输入输出模块的出现，提升了人们对可扩展低直径网络 [17, 34, 45, 61]的兴趣。将这些模块与计算节点共同封装在同一芯片上，可大幅提升每节点带宽。每个数据包只占用少数链路的带宽，从而限制尾延迟带来的负面影响，并提升系统性能 [19]。

因此，低直径网络对于高效利用共封装芯片上的带宽至关重要。由于每个路由器都与一个计算节点集成在一起，共封装系统的可扩展性便由定义网络拓扑的图的阶来衡量。

1.2 相关工作

鉴于低直径拓扑的重要性以及交换机基数所受的技术限制，在给定直径 D和路由器度数 d的条件下，设计阶（节点数）尽可能大的网络是合理的。这就是度-直径问题 [46]，，其阶 G的上界由 Moore 界 [28]给出。具有良好*Moore* 界效率 （接近 Moore 界）的网络不仅可扩展性强，而且成本效益高、能效高，因为它们能够以基数相对较低的交换机和较少的线缆构建给定规模的系统。遗憾的是，目前已知的 $D > 2$ 和 $d > 2$ 最大图的规模远小于 Moore 界。

一些直径为2的网络，例如 PolarFly [37] 和 Slimfly [7] ，已接近 Moore 界。然而，其规模受限于直径为2时较小的 Moore 界（对于基数 d，为 $d^2 + 1$ ）。由于在可行的基数下，直径为2的网络只能容纳数千个节点，因此并不适用于大规模数据中心和高性能计算系统。

另一方面，直径为3的网络具有足够高的 Moore 界，能够满足可扩展性要求（$d^3 - d^2 + d + 1$ 基数为 d）。Dragonfly [36] 和 HyperX [2] 是大型系统中部署的热门直径为 3 的拓扑 [13, 39]，，但如图1所示，这些拓扑的 Moore 界效率较低，因而

本研究部分基于美国国家情报总监办公室（ODNI）情报高级研究计划局（IARPA）通过高级图形智能逻辑计算环境（AGILE）研究计划提供的支持，并依据美国陆军研究办公室（ARO）合同编号 W911NF22C0081 开展。本文所述观点和结论仅代表作者本人，不应被视为必然代表 ODNI、IARPA、ARO 或美国政府明示或暗示的官方政策或认可。本研究还获得了美国能源部的支持，资金通过洛斯阿拉莫斯国家实验室提供；该实验室由 Triad National Security, LLC 运营，服务于美国能源部国家核安全管理局（合同编号 89233218CNA000001）。此外，本研究还获得了美国能源部洛斯阿拉莫斯国家实验室 2023年 实验室指导研发计划的支持，项目编号为 20230692ER。美国政府保留一项不可撤销、非独占、免版税的许可，可对本作品的已发表版本进行出版、翻译、复制、使用或处置，并授权他人进行同样操作。本文的洛斯阿拉莫斯国家实验室编号为 LA-UR-22-30347，版本 2。

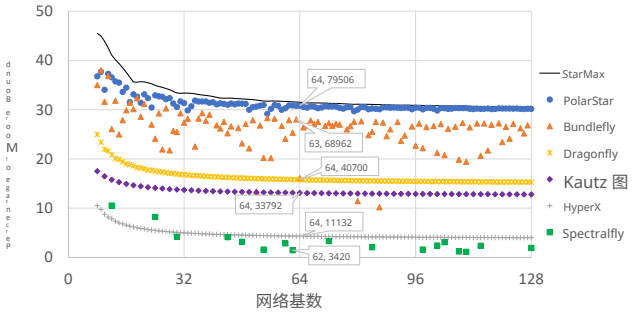


图1：直接直径为3的拓扑相对于 Moore 界的可扩展性。数据标签显示了每种拓扑在基数为 ≤ 64 时可支持的最大节点数及对应基数。*StarMax* 表示采用 $P-$ 和 $R-$ 星乘积（即最先进的 Bundlefly 和 PolarStar 网络中使用的数学构造）时理论上可实现的最大图的上界。对于 Spectralfly，由于它不是固定直径的拓扑，我们仅比较直径为 ≤ 3 且在给定基数下规模最大（若存在）的设计点。对于 Kautz 网络，我们将每条链路视为双向链路。

推高了网络成本和功耗。Kautz 网络 [44] 用于 SiCortex 的网络架构 [59],，是面向有向图、具有接近 Moore 界可扩展性的有向拓扑。然而，几乎所有已部署的片外网络系统 [9, 24, 31, 51, 60] 和互连网络 [2, 35, 36, 42] 都采用双向链路。将 Kautz 网络中的每条链路视为双向链路，会使节点的度翻倍，从而导致 Moore 界更高、效率更低。对于直径为3的情况，双向 Kautz 网络的渐近 Moore 界效率为 $< 13\%$，比 PolarStar 2.4×低。

最近，Lei 等人提出了 Bundlefly [40],，这是一种基于两个图的星积构造的模块化直径为3的网络。其模块间链路可以捆绑到多芯光纤中，以降低布线成本。然而，如图1所示，其 Moore 界效率并不稳定。

构造最大的直径为3的图是数学中的一个开放问题。已知的最佳直径为3的图与 Moore 界之间存在很大差距。组合学维基排行榜 [46] 列出了一些截至2013年已知的最大直径为3的图，基数最高达20。[6], 中的构造（最先进的 Bundlefly [40], 也采用了该构造）在基数为 $18-20$ 时超越了排行榜。据我们所知，在我们的构造提出之前，[6] 中的构造也为大多数基数 > 20 提供了已知的最大图。

本文构造了一类新的直径为3的图，几乎在所有基数下都优于此前的最佳结果 [6, 46]，其规模的几何平均值比 Bundlefly 高出 1.3× 个数量级 [40]。

1.3 贡献

我们提出了一类名为 PolarStar 的新型网络拓扑，它利用一种称为星积的数学构造，将 PolarFly [37] 扩展到大型直径为3的网络。

- 对于几乎所有基数，PolarStar 都提供了目前已知规模最大的直径为3直连网络，其规模相较于 Bundlefly、Dragonfly 和 HyperX 分别提升 1.3×，1.9× 和 6.7× 的几何平均值。其基础图是目前已知规模最大的直径为*3*的图中的一部分。
- PolarStar 在利用目前已知的、可构造低直径星积的因子图性质时，对于直径为 3 的星积实现了近乎最优的可扩展性。进一步优化这些直径受限的星积，可能难以带来显著收益。
- PolarStar延续了 *PolarFly* 的多项网络优势，包括便于将链路捆绑到多芯光纤中的模块化布局，以及较大的二分割。
- PolarStar 拥有广阔的设计空间，并且对于 $[8, 128]$ 中的每个基数，都存在多种配置。
- 我们为 PolarStar 设计了一种路由机制，利用其理论性质，并表明在合成结构和现实世界结构上，其性能与现有直径为 *3* 的网络相当或更优。

2 背景

2.1 网络模型

在直连网络中，每个交换机都与一个计算端点直接相连或集成在一起。因此，直连网络的拓扑可以建模为无向图 $G(V, E)$，其中 $V(G)$ 是交换节点（或顶点）的集合，$|V(G)|$ 是 G的阶，E 是链路（或边）的集合。每个节点都通过 d 条链路连接到其他节点，其中 d 是网络基数，也称度。任意一对节点之间最短路径的最大长度称为直径 D。本文考虑直径为 3 的网络。

2.2 Moore 界

Moore 界 [28] 是图在度为 d 且直径为 D 时所能拥有的节点数 N 的上界。该上界为

$$N \leq 1 + d \cdot \sum_{i=0}^{D-1} (d-1)^i.$$

对于直径为3的情况，Moore 界为 $N \leq d^3 - d^2 + d + 1$.

能够达到 Moore 界且具有 $D \geq 2$ 和 $d \geq 2$ 的图只有环、霍夫曼-辛格尔顿图、彼得森图[5, 15, 28] 以及一个假想的直径为2、度为57的图 [14]。这些图不适用于大规模网络设计：低直径的度为2的环规模很小，而其他图各自只有一个设计点。很少有图能接近 Moore 界。[46]最新发布的度-直径问题排行榜显示，相关度数下已知的最佳直径为3的图仅达到 Moore 界的 $25-30\%$。本文提出的 PolarStar构造规模大于已知的最佳图，相关讨论见 [6]和 [40]。

2.3 网络属性

本节根据若干理想的网络属性，将 PolarStar 与 Slimfly [7], PolarFly [37],Dragonfly [36], HyperX [2], MegaFly [23, 57], SpectralFly [63],Bundlefly [40] 和胖树进行比较。比较结果见表1。

直连性：直连网络中的每台交换机都连接到一个或多个端点，而间接网络中则有一些交换节点未连接到任何端点。直连性是一项理想特性；例如，如果采用共封装模块，Fat-tree 和 MegaFly 等间接网络就需要制造两种类型的芯片，从而增加成本。此外，从理论上讲，这些拓扑中的纯交换机芯片所需的端口数量，是带端点的共封装芯片的两倍。

拓扑	直接可扩展性	稳定设计空间	$D \leq 3$ 可捆绑性
胖树	✖		✖
PolarFly	✖		
Slimfly	✖		
三维 HyperX			
Dragonfly			
Bundlefly			
Megafly	✖		
Spectralfly			
PolarStar			

表1：网络属性评估：电池等级表示网络在相应属性“▬”方面的表现：“▭”表示非常好，“✖”表示一般，“✖”表示不好。 D 是网络直径

可扩展性：拓扑所能达到的绝对规模取决于其直径对应的 Moore 界以及 Moore 界效率。PolarFly [37]等直径为2的网络接近 Moore 界，但由于直径为2时 Moore 界较小，其规模受到限制。三级胖树的扩展规模与直径为2的网络相近，但可以通过增加层级来提高可扩展性。

直径为3的网络具有较大的 Moore 界，使用常见的交换机基数即可覆盖数十万个节点。然而，已知的直径为3的网络（如 HyperX 和 SpectralFly）Moore 界效率较低。PolarStar 在直径为3的网络拓扑中具有最高的 Moore 界效率。

低直径：直径较小的网络可实现低延迟的远程访问，并能维持较高的单交换机输入带宽。直径为 ≤ 3 的网络更受青睐，因为它能以较低成本兼顾性能和可扩展性。

稳定的设计空间：理想的拓扑应能为所有交换机基数提供多种配置，并实现稳定扩展（Moore 界效率始终如一，不出现波动）。这样一来，系统便可根据不同规模（由客户需求决定）进行设计，同时只需从有限的交换机基数中选择（受重新设计成本、可用性等因素限制）。Slimfly 和 PolarFly [7] 的可行基数和配置都很少。Bundlefly 的 [40] Moore 界效率会随着基数显著波动（图1）。PolarStar 为每种基数都提供多种配置，且 Moore 界效率更加稳定。尽管 SpectralFly 可采用多种基数构建，但如图1所示，只有极少数基数能实现直径为3。

可捆绑性：Bundlefly [40] 提出了适合捆绑的网络这一概念，即采用模块化拓扑，在相邻模块（节点的逻辑组）之间设置多条链路，并将这些链路封装在多芯光纤 [3]中。这降低了布线成本和复杂度。按照这一描述，Bundlefly 和 PolarStar 都适合捆绑。规模最大的 Dragonfly 和 Megafly 构造则不太适合捆绑——每对节点组之间只有一条链路。可以在相邻组之间使用多条链路，使这些拓扑具备可捆绑性 [13]。然而，这样做会按比例降低其可扩展性和 Moore 界效率。此外，能够容纳多个逻辑组的大型机架也有助于捆绑机架间线缆。

3 方法

图的星积是一种非交换乘积，由 Bermond、Delorme 和 Farhi 于 [6] 提出。在特定条件下，该乘积的直径可以小于各因子的直径之和。Lei 等人将星积构造用于 Bundlefly [40]。

本文同样使用星积，但为我们的因子图设计了不同于[6] Bundlefly 原有因子图的新性质。这些新性质使我们能够构造规模更大的自定义因子图，并构建 PolarStar 星积拓扑，其规模超过 Bundlefly 以及此前已知的所有直径为3的网络。

我们使用的因子图不仅具有较大的阶，还具备许多其他特性，这些特性可以延续到星积中，对网络构建很有用。本文将对这些特性以及 PolarStar 的网络特征进行分析。

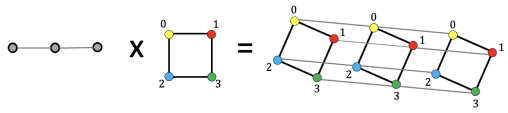
4 星积

4.1 星积的直观理解

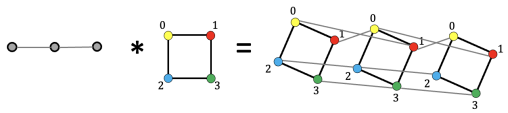
星积是笛卡尔积 $G \times H$的推广。在笛卡尔积中， H的副本称为超节点，用于替代结构图 G的顶点。当任意两个副本中的对应顶点 H_1 和 H_2在 G中通过一条边相连时，这两个副本中的对应顶点之间也会通过边相连。图 2a对此进行了说明。

星积的边同样由相邻超节点中顶点之间的双射决定。不过在这种情况下，要求边必须连接对应顶点的条件有所放宽。不同相邻超节点对之间的双射甚至不必相同。仔细选择双射连接，可能会使所得积的直径最多比结构图的直径大 1。我们稍后将讨论如何通过选择双射来实现这一受约束的直径。

直观来说，星积保留了结构图 G 的大尺度形态，并将超节点 H 的副本嵌入到 G的顶点位置。随后，按照所需的方式，通过双射连接这些超节点。图 2 对比了在相同因子图上构造的笛卡尔积和一个星积示例，并展示了它们的大尺度结构图、超节点嵌入及连接方式。图 5 展示了我们的星积构造。



(a) 笛卡尔积 $L_3 \times C_4$。相邻副本 C_4 中的对应节点由一条边相连。该双射就是恒等映射。

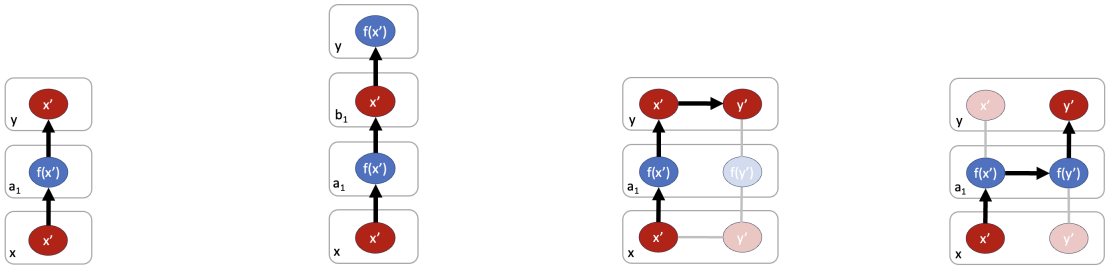


(b) 一个星积示例 $L_3 * C_4$。在这个特定的星积中，相邻超节点都通过同一个双射 $f = (01)(2)(3)$相连。

图 2：笛卡尔积 $L_3 \times C_4$ 与一个星积示例 $L_3 * C_4$的对比。结构图 L_3 是一个具有三个顶点的路径图，超节点 C_4 是一个具有四个顶点的环图。

4.2 正式定义

以下正式的定义1摘自 [6]。在本文中，我们将 G称为结构图，将 G' 称为超节点。



(a) $y' = x'$：一条长度为2 、穿过结构图的 x'-交替路径。

(b) $y' = f(x')$：一条 x'-交替路径，长度为3，穿过结构图，使用长度为-2 x'-a_1 和 y之间的交替路径。

(c) $(x', y') \in E(G')$： x'- 和 y'-长度为 2 的交替路径经过结构图，其中通过超节点 y （或 x）的一跳。

(d) $(f(x')\ f(y')) \in E(G')$： x'- 和 y'-交替路径通过结构图，长度为 2，并经过超节点 a_1跳。

图3：该图说明了定理 1：在具有性质 R的直径为2的结构图上，展示了如何由2跳和3跳交替路径构造路径。星积路径可能包括一个1跳超节点绕行，如图3c和3d所示，这是性质 $\mathbf{R}^*$ 在 G'上所允许的。 ing

定义1。 *[6]*设 $G = (X, E)$和 $G\prime = (X\prime, E\prime)$是两个图。对 G 的边作任意定向，并令 U 为所得弧的集合。对于 U 中的每条弧(x, y)，令 $f_{(x,y)}\colon X\prime \to X\prime$为一个双射。星积 $G_* = G * G\prime$定义如下：

(1) G_*的顶点集是笛卡尔积 $X \times X\prime$。*(2)* 当且仅当以下任一条件成立时，顶点$(x, x\prime)$与顶点$(y, y\prime)$相连：*(a)* $x = y$和$(x\prime, y\prime) \in E\prime$，或*(b)* $(x, y) \in U$和 $y\prime = f_{(x,y)}(x\prime)$。

在此定义中，条件 $2a$将 G 的节点替换为 $G\prime$ 的副本。条件 $2b$将 $G\prime$ 的副本彼此连接。$2b$ 中的双射 $f_{(x,y)}$规定了如何跨越弧（x, y），将 G 的顶点与 $G\prime$ 的顶点进行双射连接。

4.3 阶、度与直径

利用以下事实，可以由两个较小的图构造出一个直径受限的大图。

(1) 顶点数为 $|V(G_*)| = |V(G)||V(G\prime)|$。(2) 若 G 和 $G\prime$ 的最大度分别为 d 和 $d\prime$，则 G_* 的最大度为 $d_* \le d + d\prime$。(3) 若 G 和 $G\prime$ 的直径分别为 D 和 $D\prime$，则 G_* 的直径为 $D_* \le D + D\prime$。

为相邻超节点之间的连接选择合适的双射，可以将直径 D_* 降至至多 $D + 1$。本文将对结构图的所有弧使用同一个双射 f，并构造直径受此限制的星积。第5.2节将介绍这样一个双射。

5 低直径星积

第 4.3 节给出了星积的度和直径上界。在第 5.1 节中，我们定义了因子图上的一些性质，利用这些性质可以构造出直径至多比结构图大 1 的星积。

Bermond 等人描述了图 G和 $G\prime$上的性质 P 和 P*，可构造出直径相较于 G [6]最小增加或完全不增加的大型星积。我们提出的新性质 R 和 $\mathbf{R}^*$ 在理念上类似，但并不相同。

具体而言，我们的 $\mathbf{R}^*$ 通过允许任意直径放宽了 P*，而 P* 则将 $G\prime$的直径限制为 ≤ 2。我们的 R 通过以下方式强化了 P：要求所有顶点对之间都存在一条长度为直径 D的路径，因此也存在一条长度为 $D + 1$的路径。相比之下，P 只要求在彼此之间存在长度为 D的路径的顶点之间，也存在一条长度为 $D + 1$ 的路径。

弱化 P* 使我们能够设计出 IQ，这是一种具有性质 $\mathbf{R}^*$ 的 $G\prime$ 超节点图，其直径可以为 > 2，且阶大于此前已知的具有性质 P*的超节点。利用 IQ，并注意到 Erdős-Rényi 极性图具有性质 R，我们构造出了一个更大的星积，其直径为 3。

因此，利用这些新的 R 性质，我们改进了 [6], 中给出的两个最大的已知星积图；对于范围为 $[8, 128]$的基数，其 1.06× 和 1.21× 几何平均值提升分别为.....。事实上，对于几乎所有基数，这个新的直径为 3 的图族都大于此前设计的任何直径为 *3* 的图。

更重要的是，对于本领域而言，PolarStar 的规模远超最先进的 Bundlefly [40], ——后者是基于 [6] 的 P星积。Bundlefly 既没有采用我们更大的超节点，也没有采用最优的结构图。因此，对于 $[8, 128]$中的网络基数，PolarStar 的规模比 Bundlefly 大出 1.3× 几何平均值。

5.1 有用的因子图性质

5.1.1 结构图性质 R_0。 这一性质及其相关命题适用于结构图 G，并突显了其路径多样性。我们稍后将看到，这种多样性使星积中任意顶点对之间都能通过 $D + 1$ 跳实现可达。

性质 R。 直径为 D 的图 G 具有性质 R，当且仅当任意顶点对 $x, y \in V(G)$ 都可由长度为 D 的路径连接。

请注意，在性质 R 的定义中，自环（若图中存在 G）可以作为长度为D 的路径的一部分。

5.1.2 超节点性质 R^* 和 R_1。 接下来的性质 $\mathbf{R}^*$ 和 $\mathbf{R}_1$适用于超节点 $G\prime$。这些性质讨论的是超节点内部的路径。由于超节点之间的路径是通过将双射 f应用于超节点顶点来定义的，我们可以利用这些性质来定义一条穿过星积的短路径。

性质 $\mathbf{R}^*$。 图 $G\prime$ 满足性质 R^*，如果存在一个对合f，使得对任意 $x\prime, y\prime \in V(G\prime)$，以下至少有一项成立：

(a) $y' = x'$

(b) $y' = f(x')$

(c) $(x', y') \in E(G')$

(d) $(f(x'), f(y')) \in E(G')$

命题1。 一个度为 $d_\prime$ 且具有性质 R^* 的图 $G_\prime$ 至多有 $2d_\prime + 2$ 个顶点。

证明。固定一个顶点 y 属于 $G_\prime$。我们将计算还可以有多少个其他顶点。令 x 为 $G_\prime$中的一个顶点。根据性质 R*，以下至少有一项成立：

(1) $y = x$，因此 x恰好有一种选择。(2) $y = f(x)$：由于 f 是一个双射，因此 x恰好有一种选择，即 $f^{-1}(y)$。(3) $(x, y) \in E(G_\prime)$，因此 x有 deg(y) 种选择。(4) $(f(x)\ f(y)) \in E(G_\prime)$：由于 f是一个函数， $f(y)$ 已固定。因此 x有 deg($f(y)$) 种选择。

总而言之，假设所有条件互不相交，我们至多有 $2 + \deg(y) + \deg(f(y))$ 个顶点。由于 $\deg(y) \le d_\prime$ 且 $\deg(f(y)) \le d_\prime$， $G_\prime$ 至多可以有 $2 + 2d_\prime$ 个顶点。□

定义*2*。设 $G * G'$ 是一个星积，其中 G' 具有性质 R^*。$G * G'$ 中的 x'-交替路径是超节点之间形如 $((a_0, x'), (a_1, f(x')), (a_2, x') \ldots)$ 的路径，其中 $(a_0, a_1, a_2, \ldots)$ 是 G 中的一条路径，并且路径元素的第二项在 x' 和 $f(x')$ 之间交替。

根据星积的定义，当 $G_\prime$ 具有性质 R* 时，交替路径是超节点之间唯一存在的路径类型。这是因为 f 是一个对合，即 $f^2(x_\prime) = x_\prime$。

性质 R* 的子性质 (c) 和 (d) 允许从任意一条交替路径跳转到另一条。这一性质十分强大，也是定理1的关键，它确立了这些星积的低直径。

引理1。 设 $G * G_\prime$ 是一个星积，其中 $G_\prime$ 具有性质 R^*。对于所有 $x_\prime \in G_\prime$，对于 G中的每条路径，在 $G * G_\prime$ 中都有一条交替路径。所有交替路径都具有这种形式。

证明。由双射星积的定义以及对于所有 $x_\prime$， $f^2(x_\prime) = x_\prime$ 这一事实可得。□

下面的性质 R1 是特殊情形 $i = 1$下 [6] 中的性质 P_i，这是直径为 3 的星积所需要的。

性质 R1。 *[6]* 若存在一个双射 f f^2 G $E\ G$ ∪，使其与$_\prime$的一个自同构相 $f\ E\ G$ $V\ G$ 对应，那么图 $G_\prime$ 具有性质 R_1； 此时边集($_\prime$) (($_\prime$)) 恰好是($_\prime$)上的完全图的全部边。

5.2 低直径星积

这里给出星积 $G * G_\prime$ 的直径至多为 $D + 1$ 的判据。当 G的直径为 D，且满足以下任一条件时

- 结构图 G 具有性质 R，且超节点$G_\prime$ 具有性质 R* （定理1），或者
- 超节点 $G_\prime$ 具有性质 R1 （定理 2, [6]）。

我们定义星积中从 $(x,\ x_\prime)$ 到 $(y,\ y_\prime)$ 的一条路径。这里的策略是找出长度为 D 的 $x_\prime$ 和 $y_\prime$ 交替路径，起点为超节点 x （或 x 的邻居），终点为 y。如有必要，我们从 $x_\prime$ 交替路径跳一步到 $y_\prime$ 交替路径，使用性质 R* 的子性质 (c) 和 (d)。

图3阐释了定理1，展示了本文关注的直径为 3 的问题。该定理、其证明和图示有助于深入理解第 9.2 节讨论的路由策略。

定理1。 设 G和 $G_\prime$ 分别为满足性质 R 和 R^* 的图。对于 G 的每条边 (x, y)， 定义 $f_{(x,y)}(x_\prime) = f(x_\prime)$ （根据性质 R^*） 。若 G 的直径为 D，则星积 $G_* = G * G_\prime$ 是一个直径至多为 $D + 1$ 的图。

证明。设 $(x, x_\prime)$ 和 $(y, y_\prime)$ 是 $G * G_\prime$ 中的节点。根据 G 的性质 R，存在长度为 D $x_\prime$– 和 $y_\prime$– 的交替路径，分别连接超节点 x 与 y，以及超节点 a 与 y，且这些路径位于 $G * G_\prime$ 中，其中 a 是 G 中 x 的邻居。

我们考虑性质 R*中所述的关于 $x_\prime$ 和 $y_\prime$ 的穷尽情形集。为简化表述，我们假设 D 为偶数。对于 D 为奇数的情况，反转情形 (a) 和 (b) 的推理，同样的证明仍然成立。

(a) 如果 $y_\prime = x_\prime$，我们沿着长度为D $x_\prime$的交替路径前进。(b) 如果 $y_\prime = f(x_\prime)$，我们跳 1 次到达 $(a, f(x_\prime))$（x 在 G中的一个邻居），然后沿着从 $(a, f(x_\prime))$ 到 $(y, f(x_\prime))$、长度为D $x_\prime$的交替路径前进，得到一条长度为 $D + 1$ 的路径。

在其余情况下，我们从一条长度为 D跳、 $x_\prime$交替的路径开始，并在某个超节点中跳 1 次，转到一条平行的长度为 D跳、 $y_\prime$交替的路径，从而得到一条长度为 D+1的路径。超节点的选择取决于 $(x_\prime y_\prime)$ 或 $(f(x_\prime)\ f(y_\prime))$ 中哪一个属于 $E(G_\prime)$。

(c) 如果 $(x'\ y') \in E(G')$，那么在 x'-交替路径上的任意节点 (b, x')，我们都可以从 (b, x') 跳 1 跳到 (b, y')。我们会在偶数跳时到达这样的 (b, x')。(d) 如果 $(f(x')\ f(y')) \in E(G')$，那么在 x'-交替路径上的任意节点 $(b, f(x'))$，我们都可以从 $(b, f(x'))$ 跳 1 跳到 $(b, f(y'))$。我们会在奇数跳时到达这样的 $(b, f(x'))$。

□

定理1证明中的每种情形都在图3中作了说明，适用于 $D = 2$。图 3c 说明了情形 (c)：超节点内跳是在沿路径前进 2 跳后，于超节点 y处进行的。我们也可以在超节点 x 处进行这次跳跃，此时沿路径前进了 0 跳。同样，图 3d 说明了情形 (d)：超节点内跳是在沿路径前进 1 跳后，于超节点 a_1处进行的。

下面的定理 [6] 是针对性质 P_i提出的。我们针对特殊情形重述该定理，其中 $i = 1$， 从而得到直径 $D + 1$。证明与 [6]中的证明相同。

定理 2. *[6]* 设 G是一个直径为 $D \ge 2$的图，并设 $G_\prime$ 具有性质 R_1。对于 G的任意定向中的每条弧 (x, y)， 定义 $f_{(x,y)}(x_\prime) = f(x_\prime)$ （由性质 R_1给出） 。那么 $G * G_\prime$ 的直径至多为 $D + 1$。

6 选择优良因子图

我们希望选择尽可能大的可行结构图和超节点，因为星积的阶等于各因子图的阶之积。我们还希望因子图具有对网络构建有用的性质（包括 R 性质等）。

- 对于结构图，我们采用 *Erdős-Rényi* (ER_q) 极性图族，该图族由 Erdős 和 Rényi [21] 以及 Brown [11]提出。这个图族具有性质 R，且规模

大于任何其他已知的直径为 2 的图族，因此是优秀的结构图候选。ER_q 具有许多其他网络优势，星积也继承了这些优势 [18, 37, 38]。

• 对于超节点，我们构造了一个具有性质 R* 的新图，称为 *Inductive-Quad*（IQ）。IQ 大于所有其他已知能够与ER_q生成直径为 3 的星积、且具有相关性质的超节点，因此，它是比任何现有构造都更好的 G' 候选方案。IQ 达到了 R* 图的上界；因此，没有具有性质 R*、P* 或 P_1的超节点能够大于 IQ。

6.1 结构图 G的选择

6.1.1 ER_q 及其网络属性。Erdős-Rényi（ER）或 Brown 族极性图 [11, 21] 基于有限射影几何，其中邻接关系由正交性定义。这些图因其诸多优势 [37, 38]，被用于 PolarFly [37]，、SymSig [10] 和 demi-PN [12]网络。

这些理想的网络属性也适用于星积。因此，ER 图族是结构图的绝佳候选，因为根据定理 3，这些图确实具有性质 R。其优点包括：

• ER 图在几乎所有度数下都比其他已知的直径为 2 的图更大，如图 4 所示。度数为 $q+1$ 的 ER 图的阶为 q^2+q+1，因此渐近地达到直径为 2 的 Moore 界（q^2+2q+2）。

• ER 具有大量可行的度数：对于任意度数 $q+1$，都存在一个 ER 图，其中 q 是素数幂。
• ER 具有模块化布局。
• ER 具有较高的二分带宽。

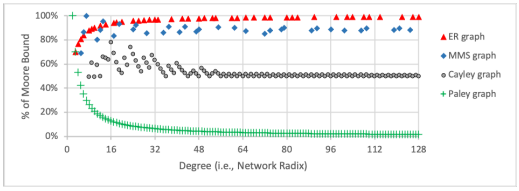


图 4：一些已知直径为2的图族与 Moore 界的比较： ER **图、McKay-Miller-Širáň 图 [48], 最优的凯莱图 [1], 以及 Paley 图。可以看出，增大结构图的规模只会略微增加星积的规模。**

6.1.2 的构造 ER_q。ER_q的顶点是向量（x，y，z），其中 x，y，$z \in \mathbb{F}_q$ 是阶为 q 的有限域。若顶点 v 和 w 的点积 $v \cdot w = 0$，则它们相邻，其中加法和乘法均按 $\mathbb{F}_q$ 中的规则进行。由于邻接关系是根据两个向量的正交性定义的，任意两个向量的倍数仍保持相同的邻接关系。因此，我们转向射影空间，只考虑每个向量的左归一化形式（即每个向量最左侧的非零项为 1）。ER图以这些左归一化向量为顶点，并在所有正交向量对之间连边。请注意，使用 $\mathbb{F}_q$ 上的算术来计算正交性。有关 $\mathbb{F}_q$ 上算术的详情，请参见 [47]；有关 ER图构造的信息，请参见 [37]。

ER_q 是一个直径为2的图。直观而言，可以从欧几里得空间中的垂直关系来理解这一点。每一对不同的向量 v_0 和 v_1 都与一个共同的 $w = v_0 \times v_1$ 正交，即它们的叉积为 v_0和 v_1的叉积。从 v_0到 v_1的两跳路径为（v_0，w，v_1）。有限几何中的情况也有类似的直观解释。

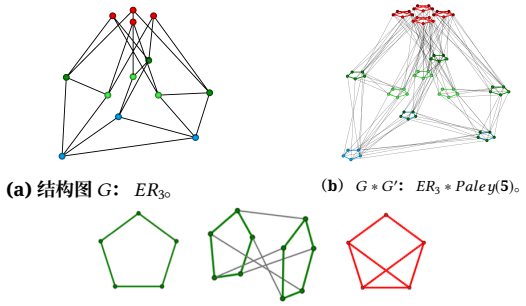
由于使用 $\mathbb{F}_q$ 的算术，ER_q 中的某些顶点是自正交的。如果允许将自环作为因子图边，则所有顶点都满足性质 R 的条件。

定理 3. 对于所有素数幂 q，ERq 都具有性质 R。

证明。ER图中任意一对顶点（x，y）都通过叉积顶点 $w = x \times y$由一条 2–跳路径相连。对于某些顶点对{x，y}，w可能与 x或 y相同；在这种情况下，这条 2–跳路径包含一个自环。□

先为因子图自环添加边，再将其加入星积中，如图5c所示，并删除乘积中剩余的所有自环。

6.2 超节点 G'的选择



(a) 结构图 G：ER_3。

(b) $G * G'$：$ER_3 * Paley(5)$。

(c) 超节点 G'：**Paley(5)。我们展示一个非自环超节点（绿色）以及两个非自环超节点（绿色）之间的连接。环超节点（绿色）。** ER_3 **恰好有带自环的顶点。在这种情况下，星积会向这些超节点添加边，如红色所示。**

图 5：星积 $ER_3 * Paley(5)$ **的构造。**

对于超节点G'，我们构造了一个新的图族，称为 Inductive-Quad（IQ）。该图族在存在时具有性质 R^*，且达到了超节点阶的上界。我们还提到了 Paley 图 [22, 50]，将其作为 G'候选方案。它的规模仅略小一些，并具有对称性，这在网络设计中很有用。对于几乎所有基数，Inductive-Quad 和 Paley 图都能产生最大的星积：IQ超节点几乎总是产生最大的星积，而 Paley 则在其他图中产生最大的星积。我们在图 5 中展示了 PolarStar 的星积构造。

超节点	阶	允许的 d'	对称的	R^*	R_1
Inductive-Quad	$2d'+2$	0或3（模4	N	Y	N
Paley [22, 50]	$2d'+1$	偶数，$2d'+1$ 一个素数幂	Y	N	Y
BDF [6]	$2d'$	全部	N	Y	N
凯莱 [48]	$2d'+\delta$，$\delta \in \{0, \pm 1\}$	$2d'+\delta$素数幂	Y	N	Y
完全	$d'+1$	all	Y	Y	Y

表 2：度为 d' **的超节点的参数比较** **s.**

其他超节点拓扑也可能值得关注，因此我们在此一并提及。例如，完全图提供了局部性良好的密集连接区域，而 Cayley 图具有高度对称性 [48]。BDF 图专为 [6] 中的大型星积而设计。我们在表 2 中列出超节点选项，并在图 7 中展示相关基数下 IQ和 Paley 的阶。

6.2.1 归纳四元图的构造。 我们在此展示一种新的度为-$d_\prime$ R* 图的构造，该图在阶方面达到上界 $2d_\prime + 2$ R* 图。

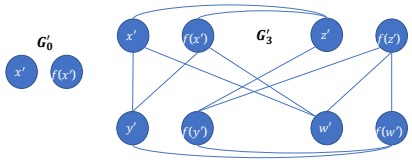
我们从图 6a 所示的两个基础 R* Inductive-Quad 图开始：一个度为 $d_\prime = 0$ 且有 2 个顶点的 IQ_0，以及一个度为 $d_\prime = 3$ 且有 8 个顶点的 IQ_3。两者都有 $2d_\prime + 2$ 个顶点。

接下来，我们通过归纳法，从一个具有性质 R*、度为 $d_\prime$ 且有 $2d_\prime + 2$ 个顶点的图 $IQ_{d'}$ 出发，构造一个具有性质 R*、度为 $d_\prime + 4$ 的图。如图 6b 所示，$V(IQ_{d'})$ 可以划分为两个互不相交的集合，每个集合包含 $d+1$ 个顶点：A 和 $f(A)$。为构造 $IQ_{d'+4}$，我们加入 IQ_3的 8 个顶点：

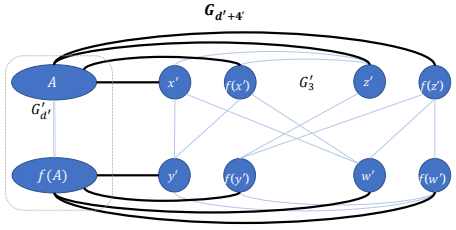
$$\{x', f(x'), y', f(y'), z', f(z'), w', f(w')\}$$

以及 IQ_3 条边。接下来，我们加入以下边：

- between $\{x', f(x'), z', f(z')\}$ and all vertices in A, and
- between $\{y', f(y'), w', f(w')\}$ and all vertices in $f(A)$.



(a) 度为 $d' = 0$ 和 $d' = 3$的基础归纳四元图。



(b) 从度为 d' 和 IQ_3的归纳四元图构造度为 $d' + 4$ 的归纳四元图。

图 6：归纳构造具有嵌入双射 f的归纳四边形拓扑，该拓扑满足性质 R*。

命题 2。 $IQ_{d'}$ 有 $2d_\prime + 2$ 个顶点，度为 $d_\prime = 4n$ 或 $d_\prime = 4n + 3$，并且具有性质 R^*。

证明。阶和度由构造得出。通过归纳法可得性质 R*，注意到经检查，IQ_0 和 IQ_3都具有性质 R*，而归纳构造随后保证了性质 R*。□

推论 1。 对于每个整数 $n \geq 0$，都存在一个度为 $d_\prime = 4n$ 或 $d_\prime = 4n + 3$ 、满足 R^* 且有 $2d_\prime + 2$ 个顶点的图 $G_{d'}$。

推论 1 表明，IQ 图比 [6], 中讨论的图更大，因为那些图的超节点较小 $G_\prime$。

满足 R^* 且有 $2d_\prime + 2$ 个顶点的图 $G_{d'}$ 也只在 $d_\prime = 4n$ 或 $d_\prime = 4n + 3$（其中 n 为某个非负整数）时存在，但我们在此不予证明。

7 POLARSTAR 的设计空间

我们评估了 PolarStar 可实现的网络规模，并将其与现有的直径为3的拓扑进行比较。

7.1 理论分析

回顾一下，星积 $G * G_\prime$ 的度为 $\deg(G) + \deg(G_\prime)$，阶为 $|V(G)| \cdot |V(G_\prime)|$。我们的结构图为 ER_q，其度为 $d = q + 1$，阶为 $q^2 + q + 1$，其中 q 为素数幂。使用一个度为 $d_\prime$的 Inductive-Quad 超节点，我们得到一个度为 $d_* = d + d_\prime$、阶为的 PolarStarG_*

$$|V(G_*)| = (q^2 + q + 1)(2d' + 2) = (q^2 + q + 1)(2d_* - 2q).$$

阶最大化时为

$$\arg\max_q V(G_*) = \frac{(d_* - 1) + \sqrt{(d_* - 1)(d_* - 2)}}{3} \approx \frac{2d_*}{3}. \quad (1)$$

代入 q的这个值，我们可以得到给定度 d_* 时 PolarStar 的最大阶为

$$\max_{\text{Inductive-Quad}} |V(G_*)| \approx \frac{8d_*^3 + 12d_*^2 + 18d_*}{27}. \quad (2)$$

同样，maxPaley $|V(G_*)| \approx \frac{8d_*^3}{27}$。因此，PolarStar 渐近地达到了直径为3的图的827倍 Moore 界。

实际上，因子图的度分布受到以下约束：(a) 在度为$q + 1$ 的 ER 图中，q 必须是素数幂；(b) 如表 2 所示，Inductive-Quad 图和 Paley 图仅存在于部分整数基数中。因此，我们评估 Inductive-Quad 和 Paley 超节点的 d 与 $d_\prime = d_* - d$的所有可行组合，以找出最大的配置。

7.2 实际可扩展性

图1比较了 PolarStar 与其他直径为3的直接拓扑的可扩展性，衡量指标为它们的 Moore 界效率。显然，PolarStar 的可扩展性超越了所有已知的直径为3的拓扑。与 HyperX [2] 和 Dragonfly [36], 相比，它们在阶上的几何平均值分别提升了 6.7× 和 1.9×。

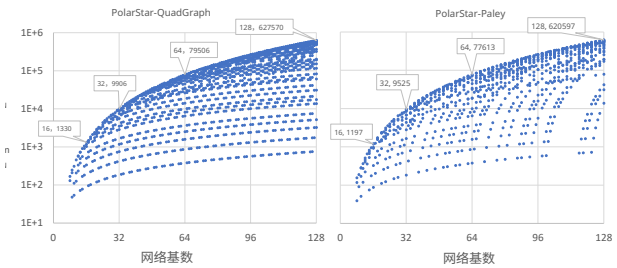


图 7：PolarStar 中基数与阶的可行组合。

与最先进的 Bundlefly [40],不同，PolarStar 为每种网络基数都提供了一种构造方式，并且相对于 Moore 界具有更一致的扩展能力。总体而言，它的几何平均规模比 Bundlefly 大 1.3×，这得益于其星积中采用了可扩展性更强的结构图和超节点，以及更多可用于优化扩展的度分布选择。

对于具有目前已知能够产生直径为3的乘积的因子图性质的星乘积网络，PolarStar 也接近最优规模。这是因为 (a) Erdős-Rényi 结构图渐近地达到直径为2的 Moore 界，并且

(b) 对于满足性质 R* 的图，Inductive-Quad 超节点拓扑达到了最优阶。

对于所有基数，度为 $k \in [8, 128]$时阶数最大的 PolarStar 均由 Inductive-Quad 超节点构造，除 $k = 23\ 50\ 56\ 80$ 外；在这种情况下，Paley 超节点能构造出更大的拓扑。PolarStar 还为每种基数提供了丰富的网络阶数选择，如图 7 所示。通过改变 (a) 结构图与超节点图之间的度分布，以及 (b) 超节点图的选择，可以实现如此多样的可行设计。

8 布局

物理部署倾向于采用模块化拓扑，由较小且相同的子图组成，这些子图可通过刀片或机架实现。此外，如果相邻模块之间有多条链路，可将它们捆绑到多芯光纤（MCF）中，以减少线缆数量 [3, 40]。我们将说明图 8 所示 PolarStar 的层次化模块结构，并指出 PolarStar 中捆绑链路的机会。

考虑一个度为 d_* 、具有度为 $q+1 \approx \frac{2d_*}{3}$ 的 ER_q结构图（方程（1））的最大阶 PolarStar，以及一个度为 $d_\prime = d_* -$ 的归纳式 Quad 超节点（$q+1$）。在此 PolarStar 中，超节点 是由 $2d_* - 2q$ 个节点构成的最小构建模块（图 5），并复制 q^2+q+1 次（每个 ER_q节点对应一个）。每对相邻超节点之间有 $2(d_* - q)$条链路。如果这些链路可以捆绑到一个 MCF 中，我们就能得到 $q(q+1)^2$ 条模块间 MCF（与 ER_q [37] 中的非自环边相同）。因此，这种捆绑可将全局线缆数量减少 $\approx \frac{2d_*}{3}$倍。

模块化层级中的下一级是超节点簇。如 [37]所示， ER_q 图可划分为 $q+1$ 个簇，每对簇之间约有 q 条链路。在 PolarStar 中，这对应于 $q+1$ 个超节点簇，如图 8c 所示，每对超节点簇之间约有 q 组链路。

9 评估：合成模式

9.1 拓扑

我们将 PolarStar 与 Bundlefly [40],Megafly [23, 57] 和 Spectralfly [63] 这几种最先进的直径为 3 的网络、实践中常用的直径为 3 的网络 3-D Hy- perX [2] 和 Dragonfly [36] ，以及应用最广泛的网络三级 Fat-tree [43] 进行比较，并按照惯例使用 Booksim 模拟器 [30]。

环面、超立方体或 Flattened Butterfly 等网络的性能已被证明低于这些基准 [7, 36]。我们还探索了灵活的低直径拓扑 Galaxyfly 系列 [41]。直径为 3 的 Galaxyfly 与 Dragonfly 同构，后者也纳入了比较。

网络	参数	# 路由器	网络基数	# 端点
PolarStar 与 Inductive-Quad (PS-IQ)	d=12, d'=3, p=5	1,064	15	5,320
PolarStar 与 Paley (PS-Pal)	d=9, d'=6, p=5	993	15	4,965
Bundlefly (BF)	d=11, d'=4, p=5	882	15	4,410
三维 HyperX（HX）	9 × 9 × 8，段落=8	648	23	5,184
Dragonfly（DF）	锚点=12, 标题=6, 段落=6	876	17	5,256
Spectralfly（SF）	ρ=23, q=13, p=8	1,092	24	8,736
Megafly（MF）	ρ=8, a=16, p=8	1,040	16	4,160
三级胖树（FT）	n=3, p=18	972	36	5,832

表 3：拓扑的模拟配置。每个路由器的端点数 器 = p。

所用拓扑的配置如表3所示。对于直径为3的直连网络， p（即每个路由器对应的端点数）为端口总数的 1/4（网络基数的 1/3）。对于胖树和 Dragonfly，我们使用 BookSim 内置的构造方式和路由 [30]。当路由器基数为 $2p$ 且有 3 层时，Booksim 的胖树拓扑每层有 $p^2 = 324$ 个路由器，顶层路由器的基数为一半（18）。此配置支持 $p^3 = 5,\ 832$ 个端点。在 Megafly 和胖树中，分别有一半和三分之一的路由器在一半的端口上连接端点。

9.2 PolarStar 中的最短路径计算

我们利用结构图 G和超节点 $G_\prime$ 的属性，计算 PolarStar 中的每条最短路径（minpath）。与基于表的路由相比，这显著降低了路由器的存储需求；基于表的路由会存储网络中每个路由器沿最短路径通往目的地的输出端口。

PolarStar 中的结构图是一个直径为 $D = 2$ 且具有性质 R 的 *Erdős-Rényi* 极性图。我们描述满足性质 R* 的 Inductive-Quad 超节点中的最短路径；Paley 超节点的计算方法类似。我们的最短路径计算适用于增量路由，因此适合基于目的地的路由。

如果源端和目的地相邻，最短路径就是它们之间的边。接下来，我们将说明，当源端和目的地位于相邻超节点中时，它们之间存在一条两跳最短路径。从 $(a_1,\ f(x_\prime))$ 到 $(y,\ y_\prime)$ 的路由路径在图3中以图形方式展示。为清晰起见并与图3保持一致，我们假设 $(a_1,\ f(x_\prime))$ 是源端，而 $(y,\ y_\prime)$ 是目的地。在这种情况下，我们通过逐一评估性质 R* 的所有情形来计算最短路径：

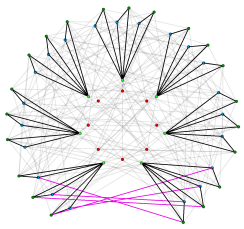
(a) $y' = x' \rightarrow (a_1, f(x'))$ 和 (y, x') 相邻。(b) $y' = f(x') \rightarrow$ 我们不会直接从源端跳到超节点 y。根据性质 R，结构图 G 中存在一条两跳路径 (a_1, b_1, y)。我们在 G_* 中取对应的路径，即 $((a_1, f(x')), (b_1, x'), (y, f(x')))$，见图 3b。(c) $(y'\ x') \in E(G') \rightarrow$ 我们使用路径 $((a_1, f(x')), (y, x'), (y, y'))$，如图 3c 所示。(d) $(f(y')\ f(x')) \in E(G') \rightarrow$ 我们使用路径 $((a_1, f(x')), (a_1, f(y')), (y, y'))$，如图 3d 所示。

最后，我们考虑源端和目的地既不相邻、也不位于相邻超节点中的情况。图3也展示了这种情况，其中源端为（$x,\ x_\prime$），目的地为（$y,\ y_\prime$）。根据性质 R，G中存在一条两跳路径（$x,\ a_1,\ y$）。在 G_*中，我们总是从（$x,\ x_\prime$）跳转到（$a_1,\ f(x_\prime)$）。由于（$a_1,\ f(x_\prime)$）位于与（$y,\ y_\prime$）相邻的一个超节点中，因此接下来我们按上述方式沿两跳路径前往（$y,\ y_\prime$），从而得到一条长度至多为 3 的整体路径。

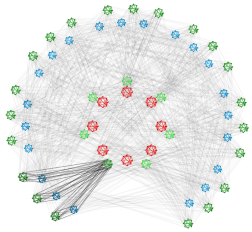
9.3 路由

我们使用以下广为人知的路由方案：

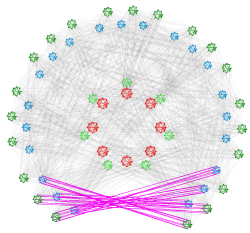
• 最短路由（*MIN*）：源端与目的地之间的每个数据包都沿最短路径进行路由。在 MF 中，我们利用同一超节点（组）内路由器之间的路径多样性。使用每对路由器之间的单条最短路径时，我们发现 SF 和 BF 的性能不佳，因此会在大型路由表中存储前往每个目的地的所有最短路径。HX 也使用所有最短路径



(a) 图 ER_7 [37]的模块化布局。每组三角形（共3 个）若共享一个公共节点，则构成一个非四次簇。红色节点构成一个四次簇。洋红色边连接两个簇。



(b) 具有 ER_7 结构图的 **PolarStar**$_{11}$ 布局图。每个 ER_7 节点都会变成一个 IQ_3 超节点在 **PolarStar** 中。高亮显示的链路及其相邻超节点构成一个超级节点簇。



(c) PolarStar 中的每对超级节点簇都由多条洋红色链路束连接。每条链路束对应图 8a 中所示的一个簇间链路 ER_7 。

图 8：PolarStar 的分层模块化布局，源自 PolarFly 网络中用于 ER 结构图的布局 [37]。相邻超节点通过一组链路连接，相邻超节点簇则通过多组这样的链路束连接。

但它通过对齐每个维度中的坐标来计算这些路径，因此不需要大型路由表 [2]。对于 PolarStar，最短路径将按照第 9.2 节所述进行解析计算。它只在结构图中存储最短路径，与 SF 和 BF 相比所需内存显著更少。

- 负载均衡自适应路由（*UGAL*）：Valiant 路由通过让每个数据包经由随机选择的中间路由器进行错误路由来平衡负载。在我们的实现中，我们随机采样 4 个中间路由器用于错误路由。我们利用本地缓冲区占用率预测 Valiant 路径和最短路径的路径延迟，并选择延迟最小的路径。请注意，SF、BF 和 HX 中的 UGAL 路由使用存储在路由表中的多条最短路径。

9.4 仿真设置

我们使用周期精确的 BookSim 模拟器 [30]分析网络性能。仿真参数（延迟、带宽）均按单条链路的数值归一化。我们采用基于信用的流量控制来产生反压，并限制注入以控制拥塞。数据包大小为 4 个流控单元，输入排队路由器的每个端口配有 128 个流控单元的缓冲区和 4 个虚拟通道（VC）。Dragonfly 和 Megafly 在 MIN 路由中分别使用 2 个和 1 个虚拟通道，在 UGAL 路由中分别使用 3 个和 2 个虚拟通道。仿真开始前会进行预热阶段，使网络在测量前达到稳态。

我们分析代表关键应用的合成流量模式下的网络性能。这类模式广泛用于评估网络拓扑 [2, 7, 36, 37, 40]。

(1) 均匀 随机流量——每个数据包的目的地均匀随机选择（代表图处理和不规则算法，例如稀疏线性代数 [8, 16]）。（2）随机排列 流量——源路由器到目的路由器的固定排列映射 τ 以均匀随机方式选定。路由器 R_s 上的端点仅向路由器 τ 上的对应端点传输（R_s）。这种模式还模拟了同时包含计算和路由器的共封装节点的排列流量。排列流量常见于快速傅里叶变换、物理模拟和集合通信 [7]。随机排列模式表示这些应用在进程 ID 随机映射到节点时产生的流量。

(3) 位洗牌 流量——通过将源地址的位向左移动 1 位来获得目的地址（$d_i = s_{(i-1) \bmod b}$）。这种模式常见于快速傅里叶变换和排序算法 [4]。（4）位反转 流量——通过反转源地址的位顺序来获得目的地址（$d_i = s_{b-i-1}$）。这种模式出现在库利-图基快速傅里叶变换、二分查找和动态树数据结构中 [25, 55, 62]。

位洗牌流量和位反转流量使用 2^b 个端点，其中 2^b 是不大于端点总数的最大 2 的幂。在分层拓扑（PolarStar、Bundlefly、Dragonfly、Megafly、胖树）中，每个路由器以及超节点/组内的端点 ID 均连续。在这类拓扑中，位洗牌流量下，超节点中的几乎所有端点都只与另外两个超节点通信。

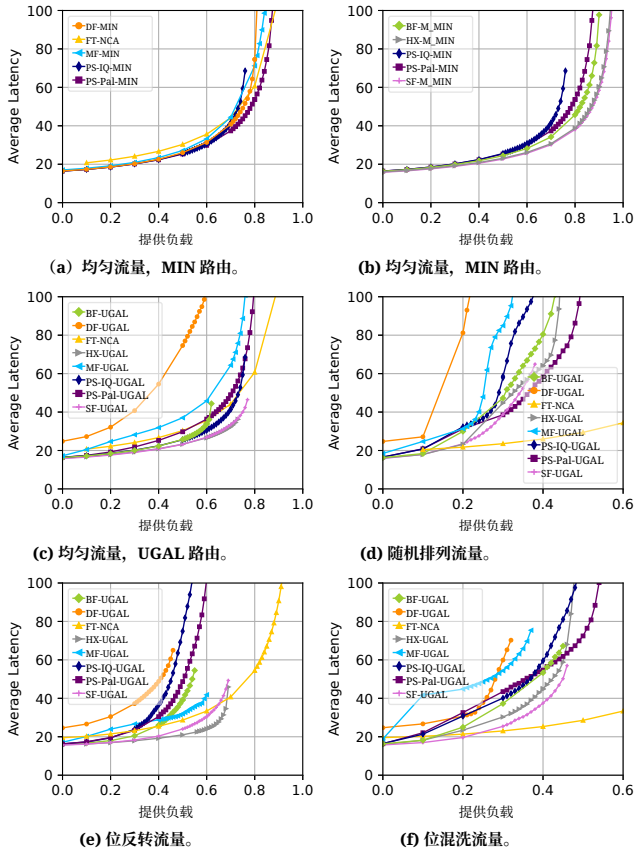
9.5 结果

图 9 比较了 PolarStar 与基线拓扑在不同路由方案和流量模式下的性能。标签遵循 <拓扑>-<路由> 的格式。负载按峰值注入带宽归一化。

总体而言，PS-Pal 和 PS-IQ 在大多数流量模式下表现出色。采用 MIN 路由时，PS-Pal 和 PS-IQ 在均匀流量下可维持超过满注入带宽（负载）的 75%。SF 和 BF 使用所有可用的最短路径，性能略高，但与 PS-* 相比，其路由表所需的存储空间显著更多。

采用自适应 UGAL 路由时，PS-Pal 和 PS-IQ 在各种流量模式下可维持满负载的 0.4 至 0.6。它们的性能显著优于 DF，与 BF（同样采用星积）相当，尽管 BF 为 UGAL 存储了多条最短路径。

MF 和 DF 的位洗牌性能较差，因为它们在每对超节点之间只有一条链路 [23], ，而 BF 和 PS-* 在超节点之间有多条链路。因此，这种模式凸显了星积拓扑相较于 DF 和 MF 的优势。相比之下，在负载分布更均衡的位逆序模式下，MF 的表现更好。对于大多数流量模式，HX 和 SF 能维持最高的注入率。它们具有高度对称性和较高的链路密度。然而，它们缺乏扩展效率（图1），且其几何平均规模分别比 PolarStar 6.7× 和 12.8× 低。



(a) 均匀流量，MIN 路由。 (b) 均匀流量，MIN 路由。

(c) 均匀流量，UGAL 路由。 (d) 随机排列流量。

(e) 位反转流量。 (f) 位混洗流量。

图9：多种拓扑的性能比较。延迟测量至模拟仍能保持稳定的最高注入率为止。超过此值后，网络将达到饱和，延迟会随模拟时间增加。在 SF 和 BF 中，每个路由表都存储了到每个目的地的所有最短路径。与其他拓扑不同，当仅使用一条最短路径时，它们的性能较差。

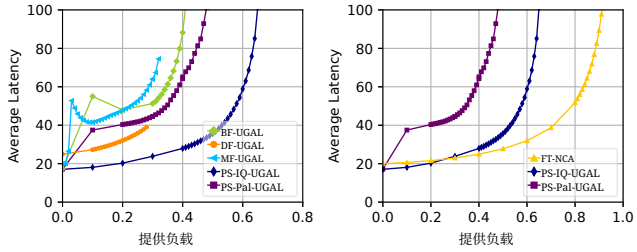


图10：对抗性流量下各拓扑的性能。

9.6 对抗性流量

对于分层拓扑 PS-*、BF、DF 和 MF，我们采用一种对抗性模式：让一个超节点中的所有端点都向另一个超节点中的端点发送数据，从而导致全局（超节点间）链路发生拥塞。对于每个源—目的地对，我们强制采用可能的最长最短路径（PS-*、BF、DF 和 MF 中为 3 跳，FT 中为 4 跳）；在 PS-* 和 BF 中，还采用最多的全局跳数（PS-* 中为 3，BF 中为 2）。这种流量模式曾用于 BF、DF 和 MF 的最坏情况分析 [23, 36, 40]。这可能并非 PS-* 的最坏情况模式。真正的最坏情况可能取决于路由算法和最小二分带宽；对于一般图，计算最小二分带宽是 NP难问题。

图10展示了网络在这种流量模式下的性能。DF 和 MF 在每对超节点之间仅有一条链路时，会在最低带宽下达到饱和。BF 和 PS-* 的表现更好，因为每对超节点之间都有多条链路。由于该配置中全局链路所占比例相对较高，PS-IQ 的表现优于 BF 和 PS-Pal。

10 评估：真实世界通信模式

10.1 模拟设置

为了比较不同拓扑在真实世界通信模式上的性能，我们使用SST模拟器 [54] 以及表3中列出的网络配置，并利用SST的Merlin库 [27]构建拓扑。

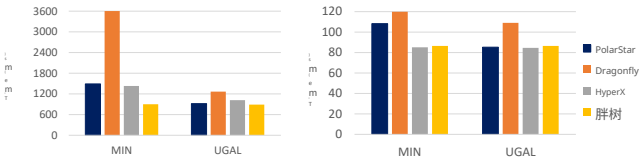
我们将PolarStar（PS-IQ）与Merlin中内置的Dragonfly、HyperX和胖树模块进行比较。对于这些拓扑，我们评估MIN和自适应UGAL路由机制。对于HyperX，Merlin提供DOAL路由，该路由在每个维度中最多自适应路由一次。在模拟中，我们将路由器缓冲区大小设置为每个端口64 KB，将路由器和链路延迟分别设置为20 ns，并将链路带宽设置为4 GBps。

我们使用从Ember通信中获得的两种通信模式 SST 的模式库 [26]：

- *Allreduce* – 科学计算、线性求解器和分布式机器学习中的一项重要集体通信操作 [52, 53, 56]。
- *Sweep3D* – 这是一种沿二维处理器网格对角线移动的波前通信模式。它对网络延迟要求较高，常见于粒子模拟和迭代求解器中 [29]。

Allreduce 的消息大小为 64KB。对于这两种通信模式，我们测量了 10 次迭代的时间。进程 ID 按线性方式映射到端点。

10.2 结果



(a) Allreduce 集体通信 (b) Sweep3D 通信模式

图11：Allreduce 和 Sweep3D 工作负载的性能。

Allreduce 在胖树上的执行速度最快，在 MIN 路由和自适应 UGAL 路由下的性能相近。另一方面，在 PolarStar、Dragonfly 和 HyperX 上，UGAL 的性能明显优于 MIN。PolarStar 在自适应路由下的执行时间与胖树相当，优于 HyperX。与采用 MIN 和 UGAL 路由的 Dragonfly 相比，其加速比分别为 2.4× 和 1.4×。对于 Sweep3D，采用 MIN 路由时，PolarStar 的速度略快于 Dragonfly。采用 UGAL 路由时，PolarStar 的性能与 HyperX 和胖树相近，且比 Dragonfly 好 1.28×。其相对于 MIN 路由的性能提升小于 Allreduce 上的提升。

总体而言，PolarStar 的性能优于 Dragonfly，与胖树和 HyperX 相当。不过，它的

Moore 界效率显著高于 HyperX（图1），并且可以用更小的交换机基数实现相近的规模（表3）。

11 结构分析

11.1 二分分析

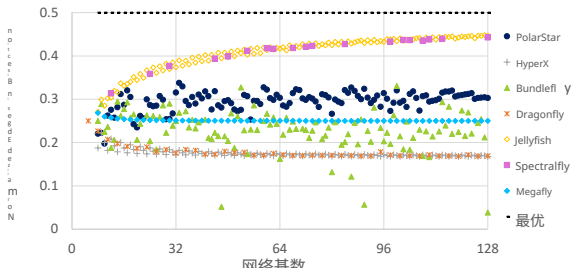


图12：METIS [32]估算的穿过最小二分割的链路比例。对于每种基数，PolarStar、Bundlefly 和 Spectralfly 均采用其最大可行（直径为3）构造。Jellyfish 的基数和规模与 PolarStar 相同。胖树和 Megafly 的二分带宽以连接了端点的路由器所关联的网络链路数进行归一化。

图12展示了网络基数范围为 [8, 128] 时不同拓扑的最小二分带宽。对于 PolarStar、Spectralfly、Megafly、Bundlefly 和 Dragonfly，使用 METIS [32] 估算最小二分带宽。我们还与采用随机图拓扑 [58]的 Jellyfish 进行了比较。在直连网络中，由于顶点之间采用随机连接，Jellyfish 的二分带宽链路占比最高，但其直径大于3。Spectralfly 使用能够优化扩展性质的拉马努金图。因此，它具有较大的二分带宽（与 Jellyfish 相当），但可行的直径为 3 的构造非常少。在其他直径为3的拓扑中，PolarStar 穿过二分割的链路比例最高，所有基数下的平均值为29。6%。相比之下，Bundlefly、Dragonfly、HyperX 甚至间接网络 Megafly 在二分割中的链路占比分别仅为22。9%、17 8%、17。4%和25。5%。二分割的改善可归因于：（a）结构图 [37], 拓扑 *ER* 具有较大的二分带宽；（b）由于超节点基数的选择较多，超节点与结构图之间的基数分布良好。

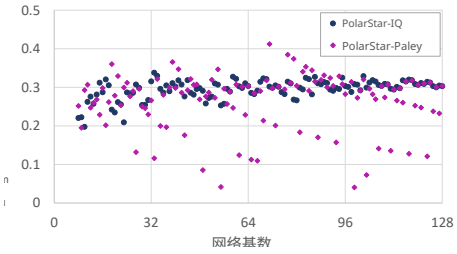


图13：采用 METIS 近似计算的 PolarStar 与 Inductive-Quad 及 Paley 超节点的最小二分。[32]

图13展示了 PolarStar 的二分割规模如何随基数和超节点拓扑变化。采用 Inductive-Quad 和 Paley 超节点的 PolarStar，其二分割中边的平均占比分别为 29.5% 和 26.6%。前者还提供了更稳定的二分带宽，适用于多种基数。这是因为 Inductive-Quad 有更多可行基数，并能更好地在结构图和超节点之间分配基数。相比之下，Paley 图可选的基数有限，可能导致 PolarStar 的超节点较大而结构图较小。这种网络将

由于大多数链路都集中在密集的超节点子图内部，因此二分带宽较小。

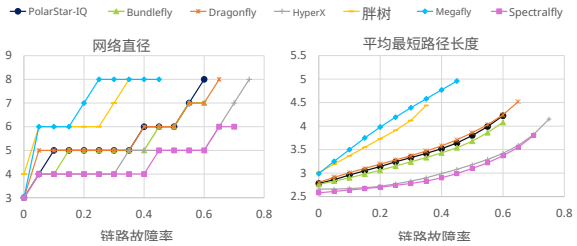


图14：随机链路故障下的网络直径和平均路径长度。对于胖树和 Megafly，我们仅考虑具有端点的节点之间的距离。为清晰起见，此处未展示 PolarStar-Paley 的行为，因为它与 PolarStar-IQ 相似。

11.2 容错性

为评估容错性，我们模拟 100 种随机链路故障场景，直至发生断连。我们随机选取一次断连率处于中位数的模拟，并在图14中报告网络直径和平均最短路径长度的变化。

PolarStar 与 Bundlefly 在 60% 的断连率下具有相似的韧性。Dragonfly 的断连率更高，为 65%。不过，在故障率较低时，Dragonfly 的直径和平均最短路径长度会迅速增加。这很可能是因为一旦全局链路发生故障，对应组之间的最短路径就必须经过其他组。由于连接密度更高，HyperX 和 Spectralfly 是最具韧性的直径为3的拓扑，尽管它们的可扩展性较差（图1）。图14中的所有直连拓扑，其断连率都高于间接拓扑胖树和 Megafly。与 Dragonfly 一样，Megafly 的每对组之间只有一条全局链路。因此，只要有 5% 的链路发生故障，其直径就会增至 6，且平均最短路径长度的增长速度快于胖树。

参考文献

[1] Marcel Abas。2017年。基于凯莱图的直径为二的大型网络。载于《在线计算机科学会议》。施普林格，225–233。[2] Jung Ho Ahn、Nathan Binkert、Al Davis、Moray McLaren 和 Robert S Schreiber。2009年。HyperX：高效大规模网络的拓扑、路由与封装。载于《高性能计算、网络、存储与分析会议论文集》。1–11。[3] Yoshinari Awaji、Kunimasa Saitoh 和 Shoichiro Matsuo。2013年。《光纤通信 VIB：第13章。采用多芯光纤的传输系统》。爱思唯尔公司章节。[4] Jun Ho Bahn 和 Nader Bagherzadeh。2008年。片上互连网络的通用流量模型。《片上网络架构》（2008年），22。[5] Eiichi Bannai 和 Tatsuro Ito。1973年。关于有限穆尔图。《东京大学理学部纪要》第1辑A：数学 20（1973年），191–208。[6] J.C. Bermond、C. Delorme 和 G. Farhi。1982年。具有给定度数和直径的大型图 III。《离散数学年刊》13（1982年），23–32。[7] Maciej Besta 和 Torsten Hoefler。2014年。Slim Fly：一种经济高效的低直径网络拓扑。载于《SC'14：高性能计算、网络、存储与分析国际会议论文集》。IEEE，348–359。[8] Maciej Besta 和 Torsten Hoefler。2015年。利用硬件事务内存和主动消息加速非规则计算。载于《第24届高性能并行与分布式计算国际研讨会论文集》。161–172。[9] Mark S Birrittella、Mark Debbage、Ram Huggahalli、James Kunz、Tom Lovett、Todd Rimmer、Keith D Underwood 和 Robert C Zak。2015年。Intel® Omni-Path 架构：实现可扩展的高性能互连网络。载于《2015年 IEEE 第23届高性能互连年度研讨会》。IEEE，1–9。[10] Dhananjay Brahme、Onkar Bhardwaj 和 Vipin Chaudhary。2013年。SymSig：一种适用于高性能计算集群的低延迟互连拓扑。载于《第20届高性能计算年度国际会议》。IEEE，462–471。[11] W. G. Brown。1966年。关于不包含汤姆森图的图。《加拿大数学公报》。 9, 3 1966 281 285 10 4153 1966 036 2 ( ), –。[12] https://doi.org/ . /CMB- - - Cristóbal Camarero、Carmen Martínez、Enrique Vallejo 和 Ramón Beivide。2016年。射影网络：大型并行计算机系统的拓扑。《IEEE 并行与分布式系统汇刊》28, 7（2016年），2003–2016。[13] Charles Q Choi。2022年。世界首台百亿亿次级超级计算机的核心。https://spectrum.ieee.org/frontier-exascale-supercomputer。[14] C. Dalfó。2019年。关于缺失的穆尔图的综述。《线性代数及其应用》（2019年）。[15] R. M. Damerell。1973年。关于穆尔图。《剑桥哲学学会会刊》74（1973年），227–236。[16] Hoang-Vu Dang、Roshan Dathathri、Gurbinder Gill、Alex Brooks、Nikoli Dryden、Andrew Lenharth、Loc Hoang、Keshav Pingali 和 Marc Snir。2018年。用于分布式图分析的轻量级通信运行时。载于《2018年 IEEE 并行与分布式处理国际研讨会（IPDPS）》。IEEE，980–989。[17] Stuart Daudlin、Anthony Rizzo、Nathan C Abrams、Sunwoo Lee、Devesh Khilwani、Vaishnavi Murthy、James Robinson、Terence Collier、Alyosha Molnar 和 Keren Bergman。2021年。用于太比特级密集波分复用互连的 3D 集成多芯片模块收发器。载于《光纤通信，OFC 2021》。[18] Aleyah Dawkins、Kelly Isham、Ales Kubicek、Kartik Lakhotia 和 Laura Monroe。2024年。星乘积网络上的边不相交生成树。arXiv 预印本 arXiv:2403.12231（2024年）。[19] Jeffrey Dean 和 Luiz André Barroso。2013年。规模化下的长尾。《美国计算机协会通讯》56（2013年），74–80。http://cacm.acm.org/magazines/2013/2/160173-the-tail-at-scale/fulltext[20] Jack Dongarra。2020年。《富士通"富岳"系统报告》。技术报告 ICL-UT-20-06。田纳西大学诺克斯维尔分校。[21] Paul Erdős 和 Alfred Rényi。1962年。图论中的一个问题。《匈牙利科学院数学研究所出版物》7A（1962年），623–641。[22] Paul L. Erdos 和 Alfréd Rényi。1963年。非对称图。《匈牙利科学院数学学报》14（1963年），295–315。[23] Mario Flajslik、Eric Borch 和 Mike A Parker。2018年。Megafly：一种面向百亿亿次级系统的拓扑。载于《高性能计算（ISC 高性能计算大会 2018）》。施普林格，289–310。[24] Denis Foley 和 John Danskin。2017年。超高性能 Pascal GPU 与 NVLink 互连。《IEEE Micro》37, 2（2017年），7–17。[25] Bernard Gold 和 Charles M Rader。1969年。《信号数字处理》。《信号数字处理》（1969年）。[26] Simon David Hammond、Karl Scott Hemmert、Michael J Levenhagen、Arun F Rodrigues 和 Gwendolyn Renae Voskuilen。2015年。Ember：面向百亿亿次级计算的通信模式参考。技术报告。桑迪亚国家实验室（SNL-NM），阿尔伯克基，新墨西哥州（美国）。[27] Karl Scott Hemmert。2018年。《Merlin 元素库深入解析》。技术报告。桑迪亚国家实验室（SNL-NM），阿尔伯克基，新墨西哥州（美国）。[28] A. J. Hoffman 和 R. R. Singleton。1960年。关于直径为2和3的穆尔图。《IBM 研究与开发杂志》4, 5（1960年），497–504。https: //doi.org/10.1147/rd.45.0497 [29] Adolfy Hoisie、Olaf Lubeck 和 Harvey Wasserman。2007年。超大规模分布式系统中波前算法的性能分析。载于广域网与高性能计算研讨会。施普林格，171–187。[30] Nan Jiang、Daniel U Becker、George Michelogiannakis、James Balfour、Brian Towles、David E Shaw、John Kim 和 William J Dally。2013年。一种详细且灵活的周期精确网络芯片模拟器。载于2013年IEEE系统与软件性能分析国际研讨会（ISPASS）。IEEE。[31] Christoforos Kachris 和 I oannis Tomkos。2012年。数据中心光互连综述。IEEE通信调查与教程 14，4（2012年），1021–1036。[32] George Karypis 和 Vipin Kumar。2009年。MeTis：非结构化图划分与稀疏矩阵排序系统，4.0版。https://github.com/KarypisLab/ METIS。[33] G. Kathareios、C. Minkenberg、B. Prisacari、G. Rodriguez 和 Torsten Hoefler。2015年。经济高效的直径二拓扑：分析与评估。载于高性能计算、网络、存储与分析国际会议论文集（美国得克萨斯州奥斯汀）。美国计算机协会。[34] Gordon Keeler。 [n. d.]。ERI计划专题组——第二阶段概述。2019年DARPA ERI峰会。[35] John Kim、William J Dally 和 Dennis Abts。2007年。Flattened butterfly：一种经济高效的高基数网络拓扑。载于第34届计算机体系结构年度国际研讨会论文集。126–137。[36] John Kim、Wiliam J. Dally、Steve Scott 和 Dennis Abts。2008年。技术驱动的高度可扩展 Dragonfly拓扑。载于第35届ISCA论文集。IEEE计算机学会，美国华盛顿哥伦比亚特区。[37] Kartik Lakhotia、Maciej Besta、Laura Monroe、Kelly Isham、Patrick Iff、Torsten Hoefler 和 Fabrizio Petrini。2022年。PolarFly：一种经济高效且灵活的低直径拓扑。载于高性能计算、网络、存储与分析国际会议论文集。IEEE，1–15。[38] Kartik Lakhotia、Kelly Isham、Laura Monroe、Maciej Besta、Torsten Hoefler 和 Fabrizio Petrini。2023年。PolarFly上基于多个生成树的网络内Allreduce。载于第35届ACM算法与体系结构并行性研讨会论文集。165–176。[39] Kartik Lakhotia、Fabrizio Petrini、Rajgopal Kannan 和 Viktor Prasanna。2021年。利用Intel PIUMA上的网络内归约加速Allreduce。IEEE Micro 42，2（2021年），44–52。[40] Fei Lei、Dezun Dong、Xiangke Liao 和 José Duato。2020年。Bundlefly：一种适用于多核光纤的低直径拓扑。载于第34届ACM超算国际会议论文集。[41] Fei Lei、Dezun Dong、Xiangke Liao、Xing Su 和 Cunlu Li。2016年。Galaxyfly：适用于大规模互连网络的一类新型灵活基数低直径拓扑。载于2016年超算国际会议论文集（土耳其伊斯坦布尔）（ICS '16）。美国计算机协会，美国纽约州纽约市，文章24，共12页。https://doi.org/10.1145/2925426. 2926275[42] Charles E Leiserson。1985年。胖树：面向硬件高效超级计算的通用网络。IEEE计算机汇刊 100，10（1985年），892–901。[43] Charles E. Leiserson。1985年。胖树：面向硬件高效超级计算的通用网络。IEEE计算机汇刊 34，10（1985年10月），892–901。[44] Dongsheng Li、Xicheng Lu 和 Jinshu Su。2004年。Kautz拓扑与DHT方案的图论分析。载于IFIP网络与并行计算国际会议。施普林格，308–315。[45] Lightmatter。 [n. d.]。https://lightmatter.co/。[46] E Loz、H. Pérez-Rosés 和 G.Pineda-Villavicencio。2010年。一般图的度-直径问题。http://www.combinatoricswiki.org/wiki/The_Degree_Diameter_Problem_for_General_Graphs。[47] Robert J. McEliece。1987年。面向计算机科学家与工程师的有限域。施普林格，美国马萨诸塞州波士顿。[48] Brendan D. McKay、Mirka Miller 和 Jozef Širán。1998年。关于直径为二且最大度给定的大图的一则注记。组合论杂志B辑（1998年）。[49] Hans Meuer、Erich Strohmaier、Jack Dongarra、Horst Simon 和 Martin Meuer。2023年。TOP500：榜单。https://top500.org/lists/top500/。[50] R E Paley。1933年。论正交矩阵。数学与物理杂志 12（1933年），311–320。[51] Gregory F Pfister。2001。InfiniBand架构简介。高性能海量存储与并行I/O 42，617-632（2001年），10。[52] Karl E Prikopa、Wilfried N Gansterer 和 Elias Wimmer。2016年。基于全归约操作的线性最小二乘并行迭代改进求解器。Parallel Comput. 57（2016年），167–184。[53] Rolf Rabenseifner。2004年。集合归约操作的优化。载于计算科学国际会议，第3036卷。1–9。[54] Arun F Rodrigues、K Scott Hemmert、Brian W Barrett、Chad Kersey、Ron Oldfield、Marlo Weston、Rolf Risen、Jeanine Cook、Paul Rosenfeld、Elliot Cooper-Balis 等人。2011年。结构仿真工具包。ACM SIGMETRICS性能评估评论 38 , 4（2011）, 37–42。

[55] Hamid Sarbazi-Azad、Mohamed Ould-Khaoua 和 Lewis M. Mackenzie。2001年。位反转流量存在时超立方体中的通信延迟。并行计算27, 13（2001年），1801–1816。[56] Alexander Sergeev 和 Mike Del Balso。2018年。Horovod：在 TensorFlow 中快速、轻松地进行分布式深度学习。*arXiv* 预印本 *arXiv:1802.05799*（2018年）。[57] Alexander Shpiner、Zachy Haramaty、Saar Eliad、Vladimir Zdornov、Barak Gafni 和 Eitan Zahavi。2017年。Dragonfly+：用于扩展数据中心的低成本拓扑。载于*2017* 年 *IEEE* 第三届百亿亿次计算与大数据时代高性能互连网络国际研讨会（*HiPINEB*）。IEEE，1–8。[58] Ankit Singla、Chi-Yao Hong、Lucian Popa 和 P. Brighten Godfrey。2012年。Jellyfish：随机构建数据中心网络。载于第九届 *USENIX NSDI* 会议论文集（加利福尼亚州圣何塞）。USENIX协会，伯克利，加利福尼亚州，美国，17–17。[59] Lawrence C Stewart 和 David Gingold。2006年。新一代集群互连。白皮书，*SiCortex Inc*（2006年）。

[60] OlaTorudbakken 和 Ashok V Krishnamoorthy。2013年。一种 50 Tbps 光缆连接的 InfiniBand 数据中心交换机。载于光纤通信会议。Optica出版集团，OTu3H–1。[61] Mark Wade、Erik Anderson、Shaha Ardalan、Pavan Bhargava、Sidney Buchbinder、Michael Davenport、John Fini、Haiwei Lu、Chen Li 和 Roy Meade。2020年。Tera-PHY：一种用于低功耗、高带宽封装内光学输入/输出的芯粒技术。*IEEE Micro* 40，2（2020年），63–71。[62] Robert Wilber。1989年。使用旋转访问二叉搜索树的下界。*SIAM*计算杂志 18，1（1989年），56–67。[63] Stephen Young、Sinan Aksoy、Jesun Firoz、Roberto Gioiosa、Tobias Hagge、Mark Kempton、Juan Escobedo 和 Mark Raugas。2022年。SpectralFly：作为灵活高效互连网络的拉马努金图。载于*2022*年 *IEEE* 国际并行与分布式处理研讨会（*IPDPS*）。1040–1050。