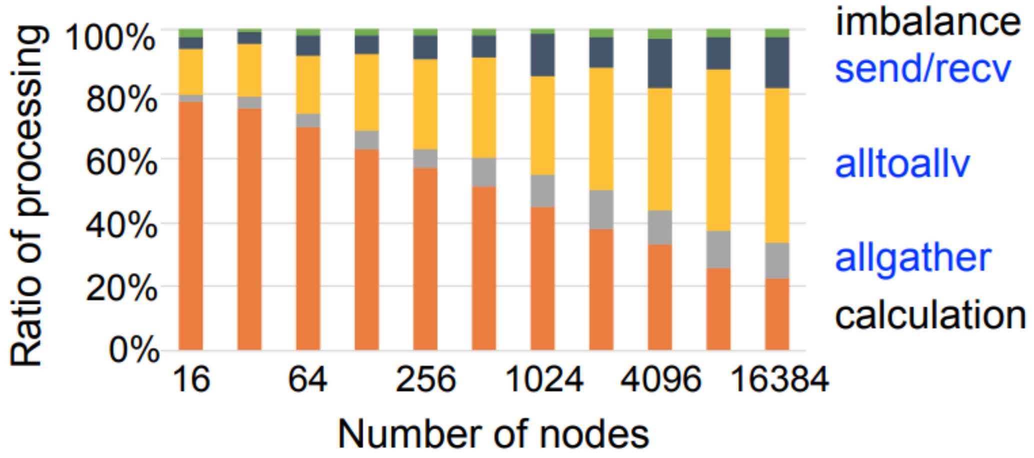
SC22
Dallas, TX | hpc accelerates.

PolarFly: A Cost-Effective and Flexible Low-Diameter Topology

Kartik Lakhotia, Maciej Besta, Laura Monroe, Kelly Isham, Patrick Iff, Torsten Hoefler, Fabrizio Petrini

Motivation

- Application Performance and Scalability
  - Large systems + sparse applications bottlenecked by network bandwidth

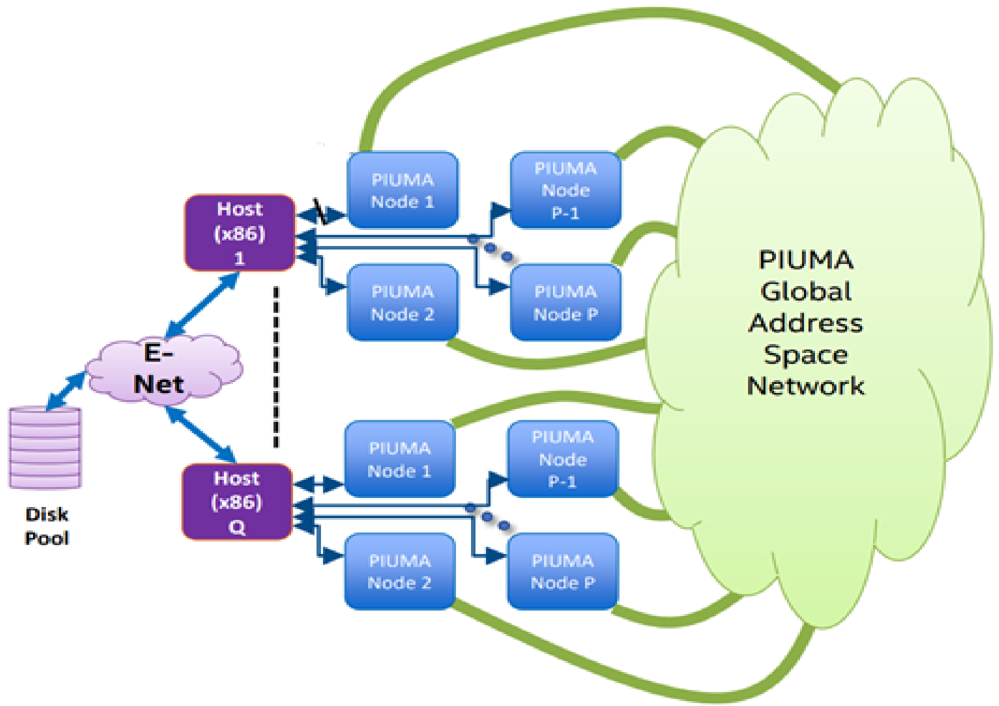


Graph 500 benchmarking on Fugaku[1]

[1] https://www.hpci-office.jp/invite2/documents2/meeting_A64FX_201209/Graph500.pdf

Motivation

- Application Performance and Scalability
  - Large systems + sparse applications bottlenecked by network bandwidth
  - Global Memory Space Programming Models need low-latency communication

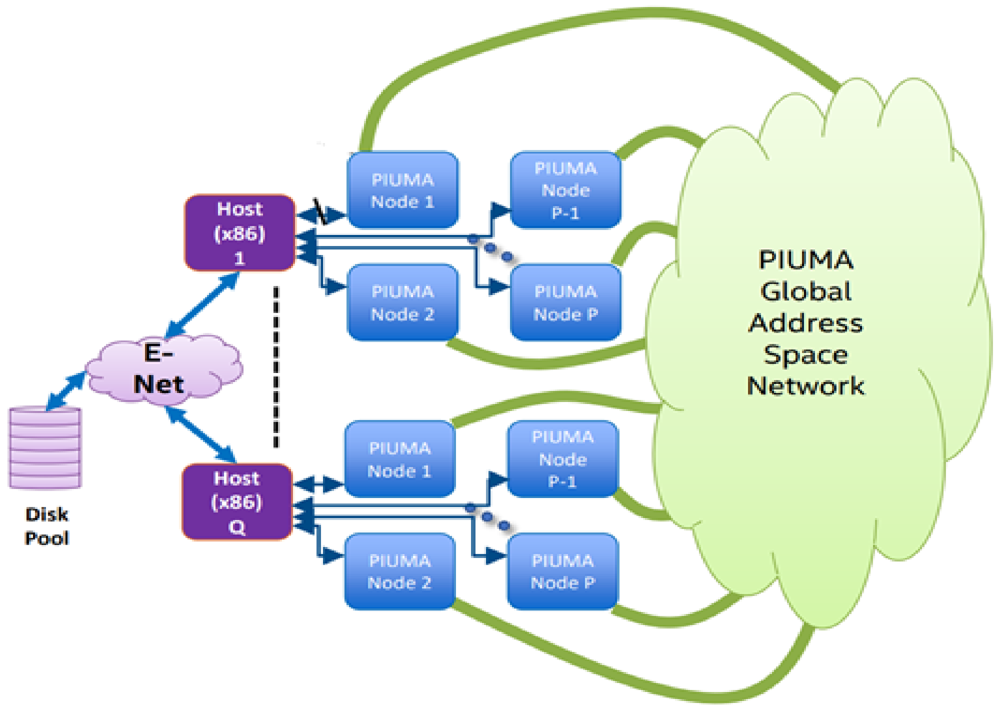


Intel PIUMA system features a GAS model[2]

[1] https://www.hpci-office.jp/invite2/documents2/meeting_A64FX_201209/Graph500.pdf
[2] Aananthakrishnan, Sriram, et al. "PIUMA: programmable integrated unified memory architecture." *arXiv preprint 2020*.

Motivation

- Application Performance and Scalability
  - Large systems + sparse applications bottlenecked by network bandwidth
  - Global Memory Space Programming Models need low-latency communication
- Network Cost[3]
  - $10\mathrm{K} - 50\mathrm{K}$ endpoints $\rightarrow 10\mathrm{M} - 100\mathrm{M}\ \$$



Intel PIUMA system features a GAS model[3]

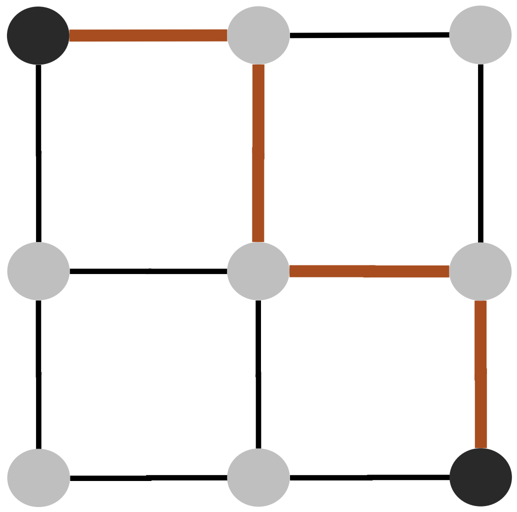
[1] https://www.hpci-office.jp/invite2/documents2/meeting_A64FX_201209/Graph500.pdf
[2] Aananthakrishnan, Sriram, et al. "PIUMA: programmable integrated unified memory architecture." *arXiv preprint 2020*.
[3] Besta, Maciej, and Torsten Hoefler. "Slim fly: A cost effective low-diameter network topology." *Supercomputing, 2014*.

Topological Requirements

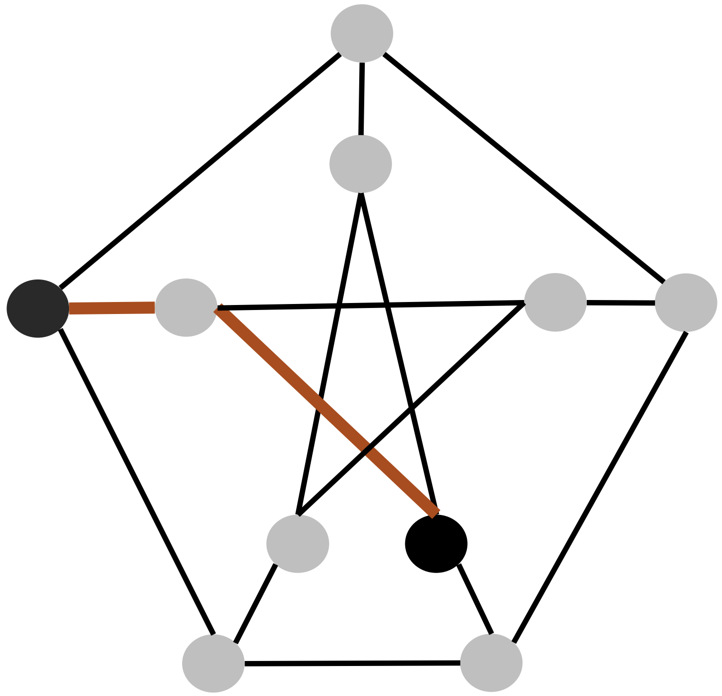
- Topology – graph, how do you connect routers?

Topological Requirements

- Topology – graph, how do you connect routers?
- Low-Diameter
  - Impacts latency



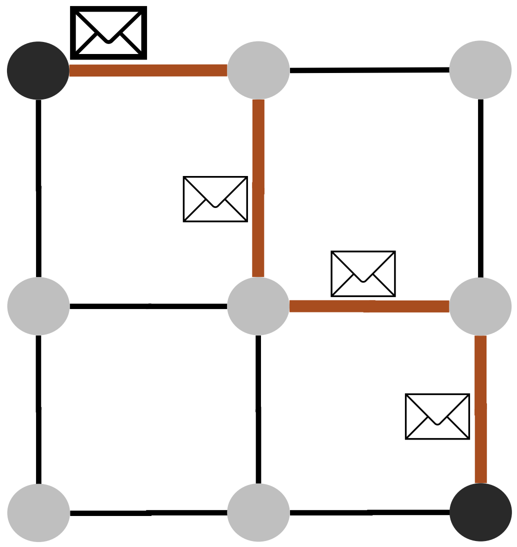
Diameter-4



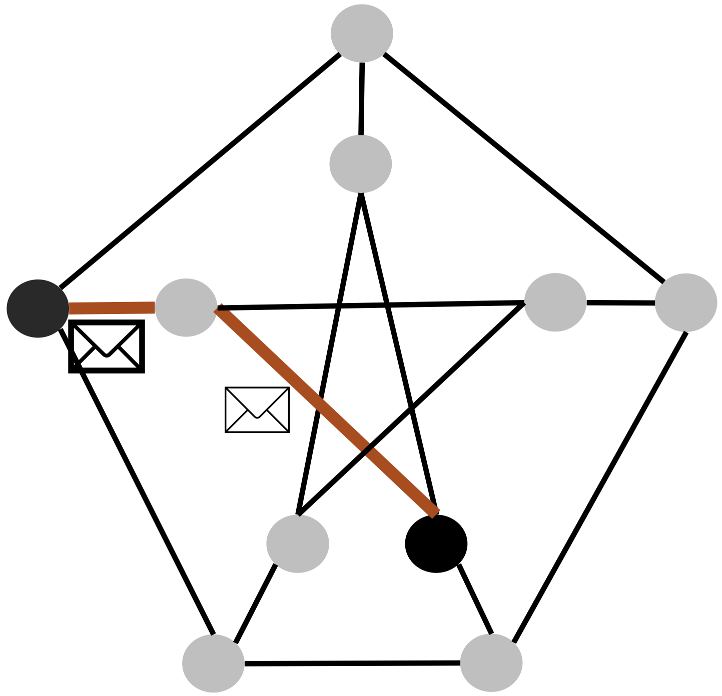
Diameter-2

Topological Requirements

- Topology – graph, how do you connect routers, graph?
- Low-Diameter
  - Impacts latency
  - Impacts injection bandwidth



Diameter-4



Diameter-2

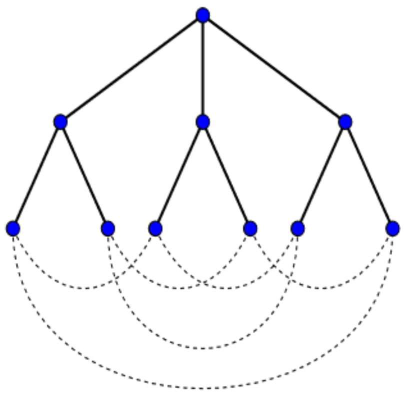
Topological Requirements

- High Scalability - connect numerous nodes

Topological Requirements

- High Scalability - connect numerous nodes
- Moore bound – formal optimality for direct networks
  - Maximum vertices for degree d + diameter k

$$1 + d\sum_{i=0}^{k-1}(d-1)^i.$$



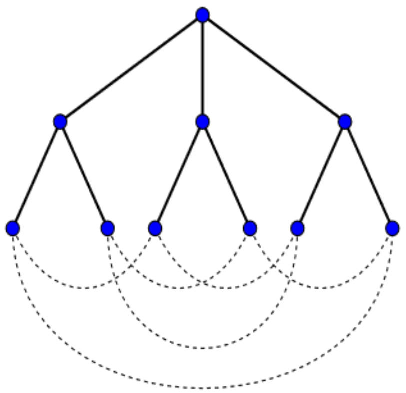
Moore Bound construction:
degree d = 3, diameter k = 2

Topological Requirements

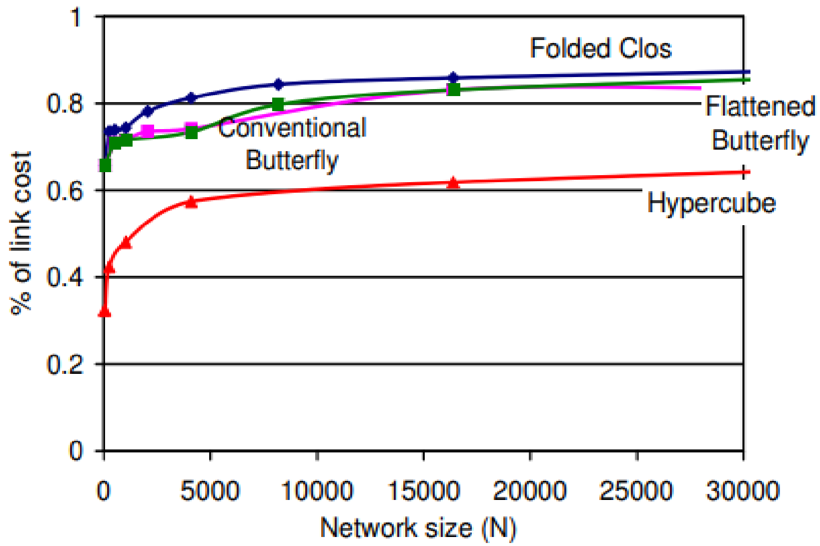
- High Scalability - connect numerous nodes
- Moore bound – formal optimality for direct networks
  - Maximum vertices for degree d + diameter k

$$1 + d\sum_{i=0}^{k-1}(d-1)^i.$$

- Moore Bound Efficiency reduces cost
  - Same scale, lower radix, less cables and IO ports



Moore Bound construction:
degree d = 3, diameter k = 2

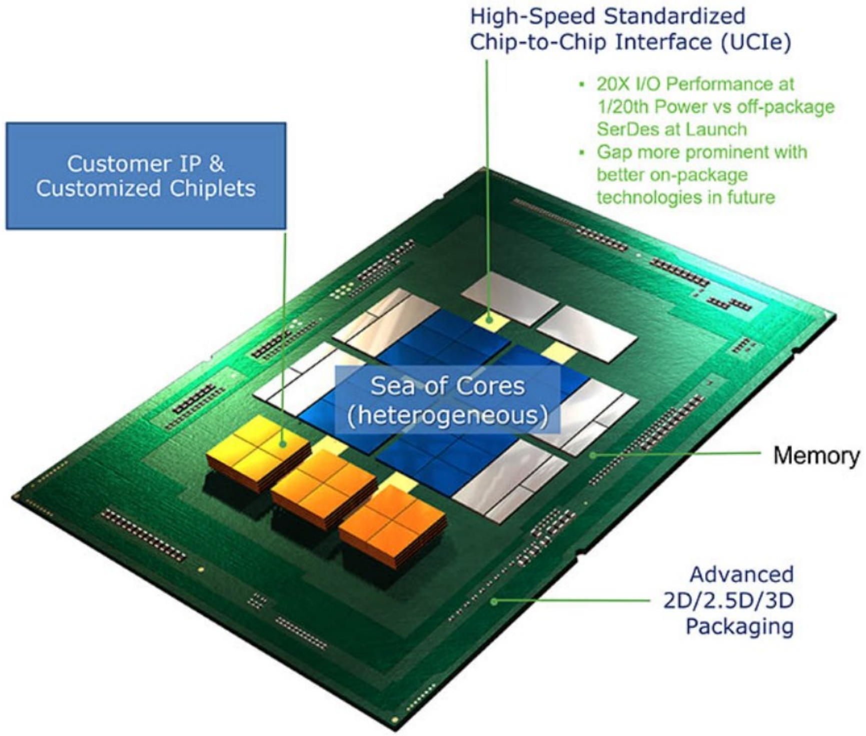


Fraction of Cable Cost in Network[1]

[1] J. Kim et al. (2007), *Flattened Butterfly: A Cost-Efficient Topology for High-Radix Networks*, ISCA'07

Technological Amplifiers

- Co-packaged Optics
  - Compute and router glued together
  - Low latency, high bandwidth



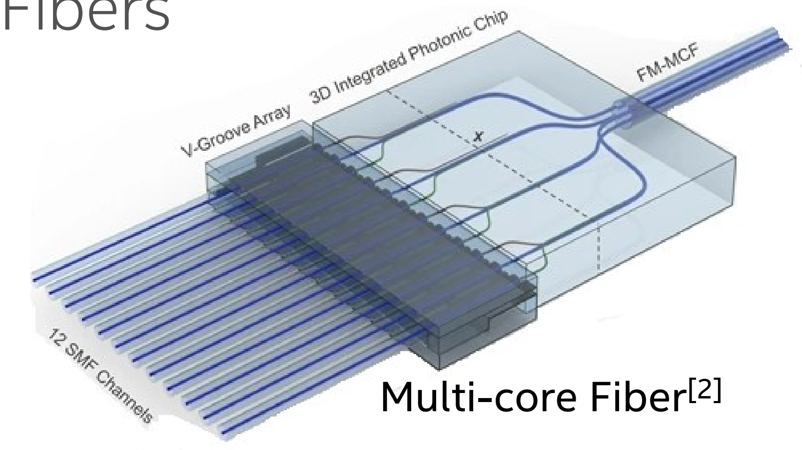
Universal Chiplet Interconnect Express UCIe 1.0[1]

[1] https://www.servethehome.com/this-intel-silicon-photonics-connector-is-a-huge-deal

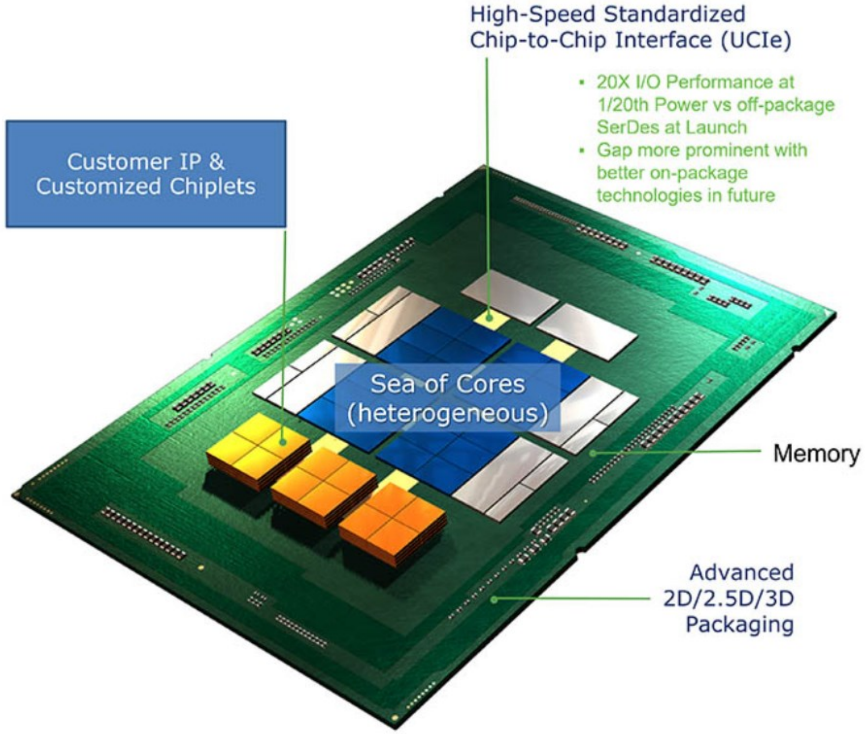
Technological Amplifiers

- Co-packaged Optics
  - Compute and router glued together
  - Low latency, high bandwidth

- Multi-core Fibers



Multi-core Fiber[2]

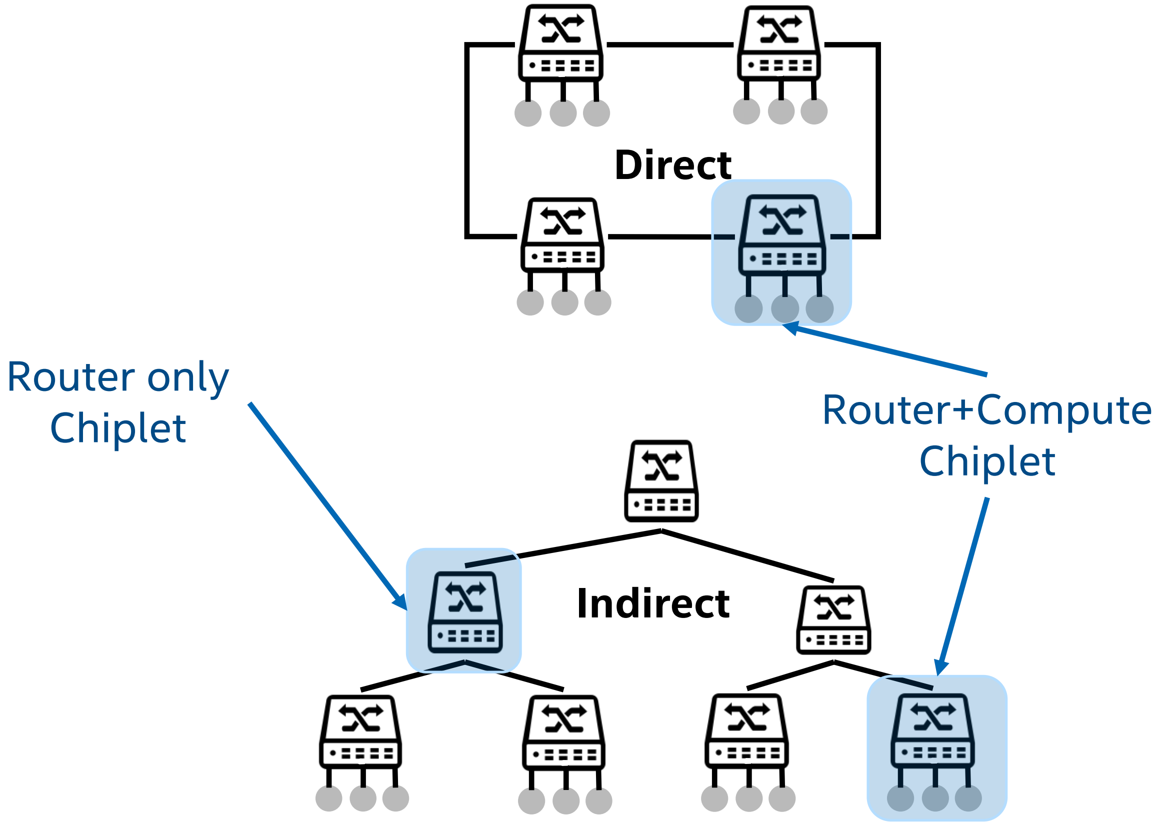


Universal Chiplet Interconnect Express UCIe 1.0[1]

[1] https://www.servethehome.com/this-intel-silicon-photonics-connector-is-a-huge-deal/
[2] Riesen, Nicolas, et al. "Monolithic mode-selective few-mode multicore fiber multiplexers." *Scientific Reports* 7.1.

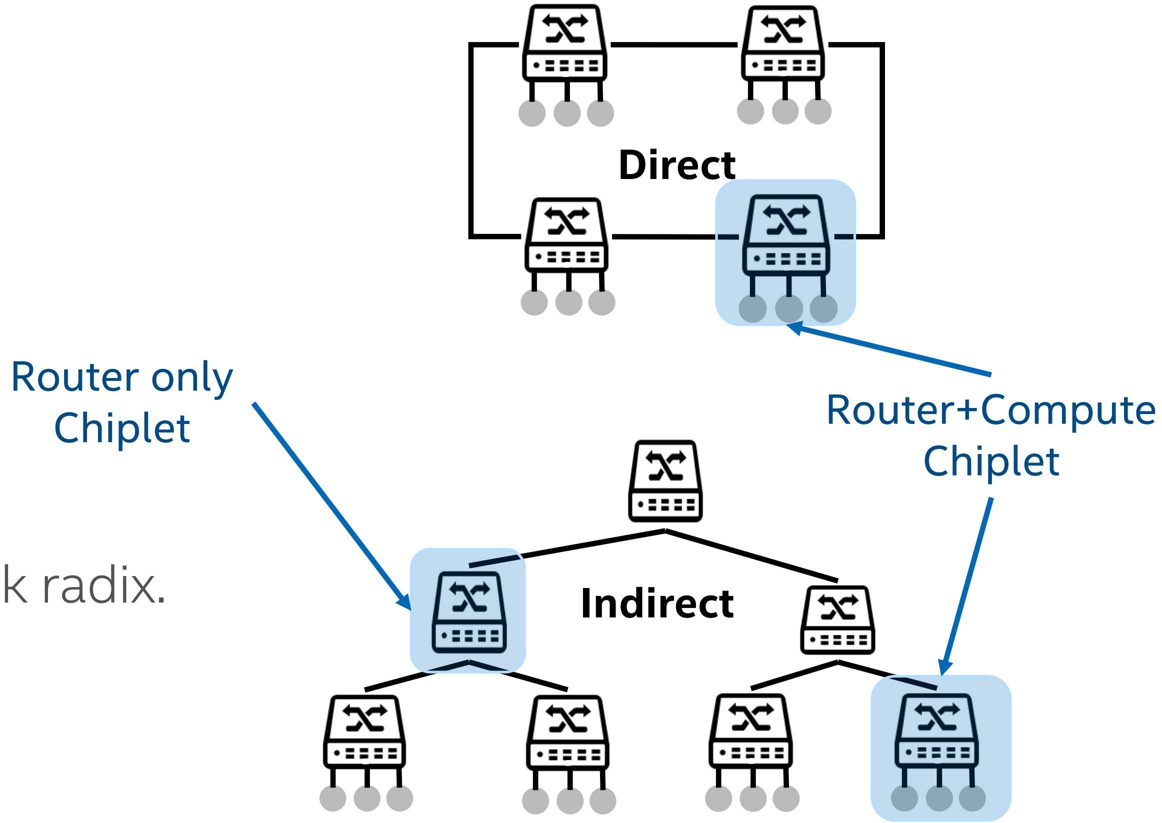
Impact on Network Design

- Direct vs Indirect Networks
  - Direct cheaper for co-packaged networks.



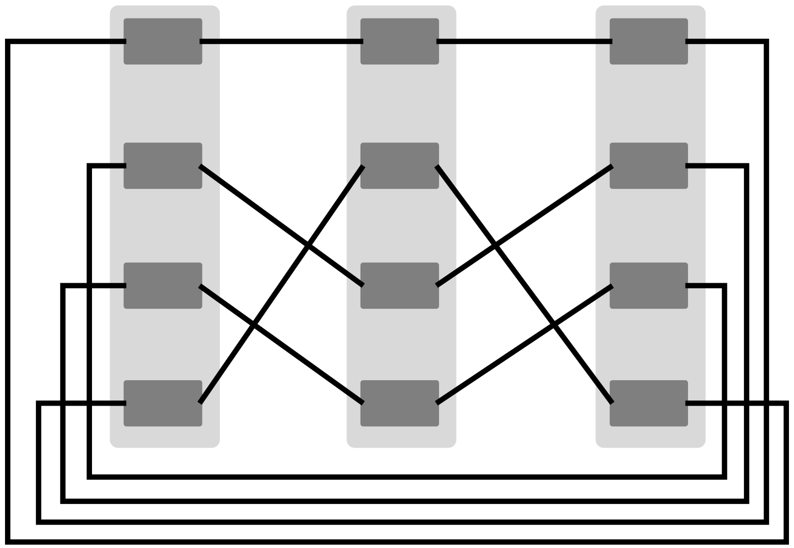
Impact on Network Design

- Direct vs Indirect Networks
  - Direct cheaper for co-packaged networks.
- Flexibility – many feasible radixes
  - Co-packaged networks: router radix = network radix.

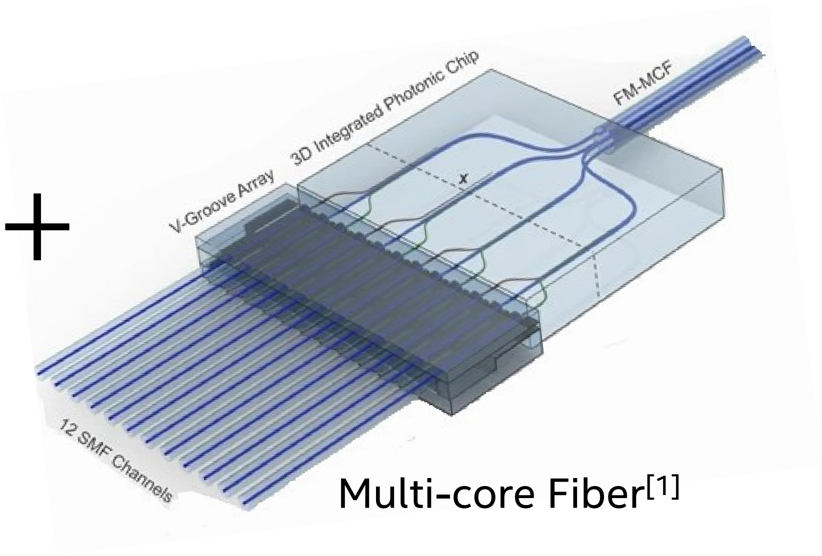


Impact on Network Design

- Modular and Bundlable Layout
  - Reduces deployment complexity and cost



+

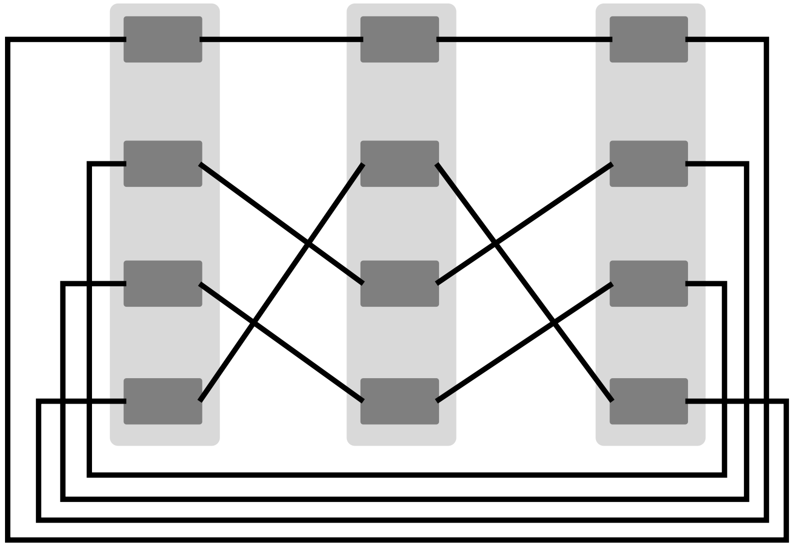


Multi-core Fiber[1]

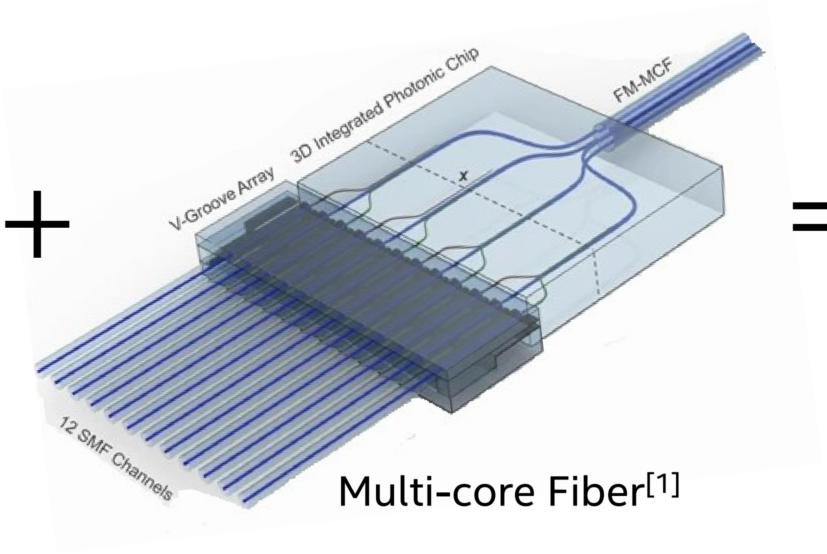
[1] Pic Credits: Riesen, Nicolas, et al. "Monolithic mode-selective few-mode multicore fiber multiplexers." *Scientific Reports* 7.1.

Impact on Network Design

- Modular and Bundlable Layout
  - Reduces deployment complexity and cost

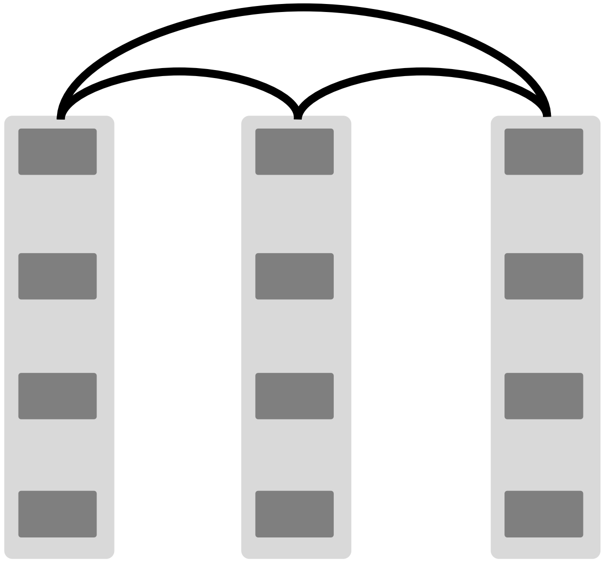


+



Multi-core Fiber[1]

=



[1] Pic Credits: Riesen, Nicolas, et al. "Monolithic mode-selective few-mode multicore fiber multiplexers." *Scientific Reports* 7.1.

PolarFly: A Scalable Diameter-2 Topology

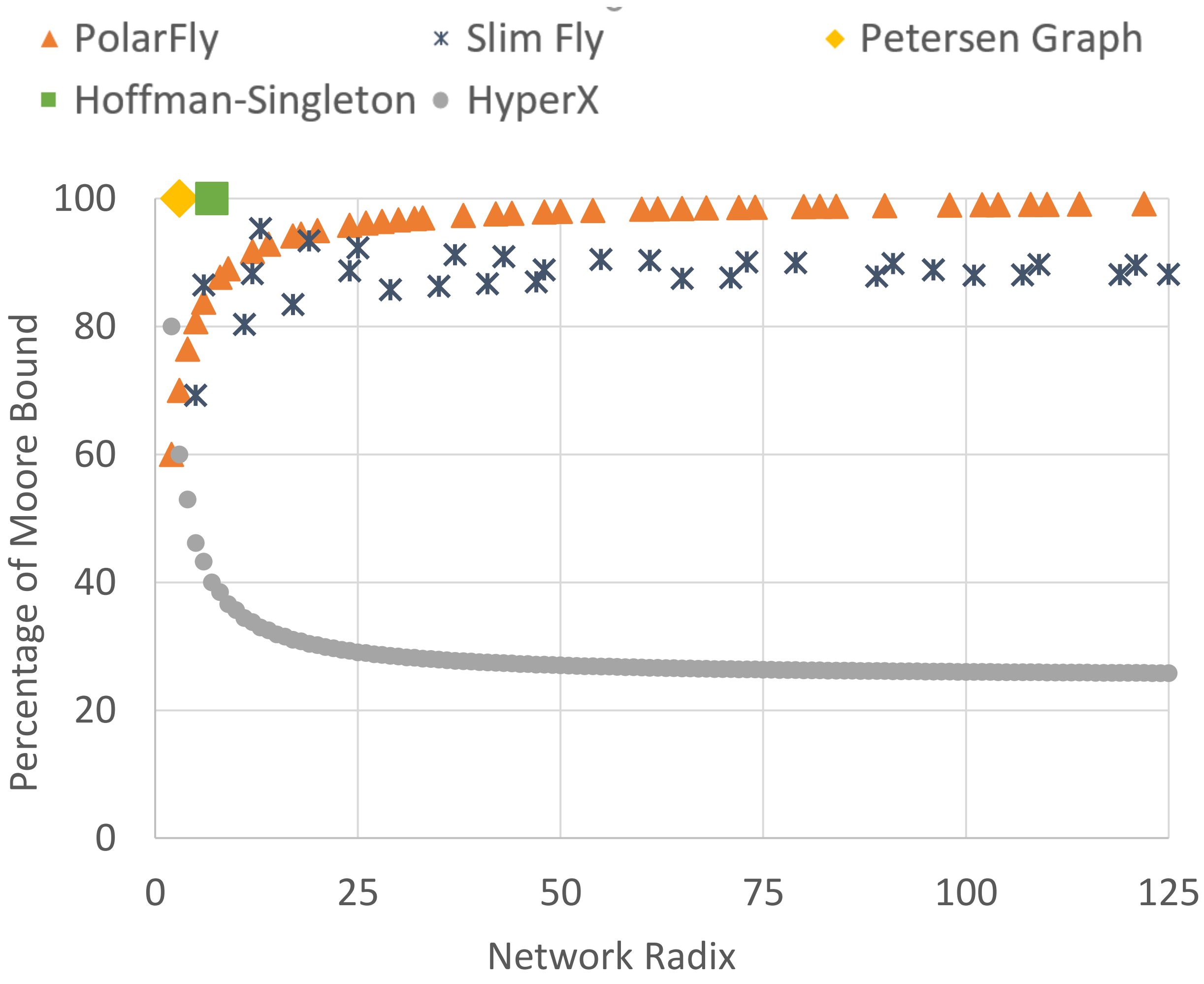
- A formal mathematical approach to network scaling and optimality
- Based on Erdős-Rényi polarity graph ER_q
  - Discovered independently by Erdős-Rényi (1962) and by Brown (1966)
  - Degree $= q + 1$ where q is a prime power
- Direct network, diameter $= 2$
  - *Lowest possible* for a non-complete graph

PolarFly: A Scalable Diameter-2 Topology

- Order of $ER_q = q^2 + q + 1$
  - Moore-bound $= q^2 + 2q + 2$
  - $\lim_{q \to \infty} \frac{q^2+q+1}{q^2+2q+2} \to 1$, asymptotically optimal

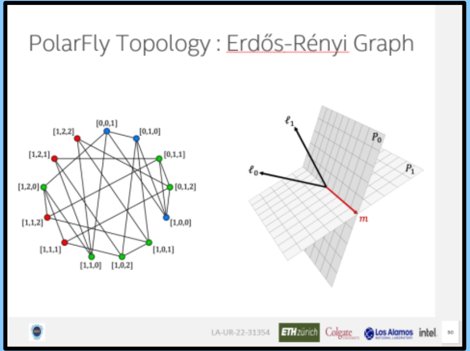
PolarFly: A Scalable Diameter-2 Topology

- Order of $ER_q = q^2 + q + 1$
  - Moore-bound $= q^2 + 2q + 2$
  - $\lim_{q \to \infty} \frac{q^2+q+1}{q^2+2q+2} \to 1$, asymptotically optimal
  - More than 95% efficiency for radix ≥20
  - $\sim 4\times$ more scalable than 2D HyperX

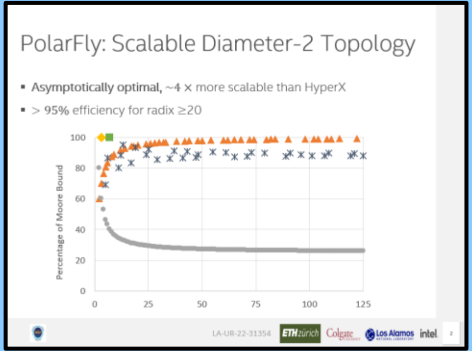


Overview

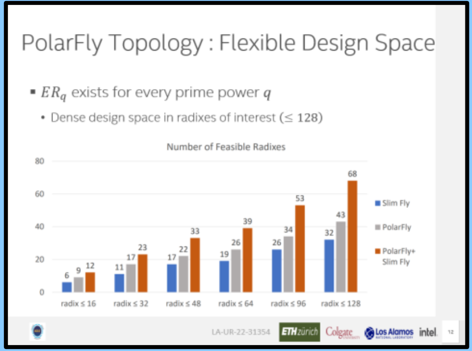
PolarFly Topology



Construction



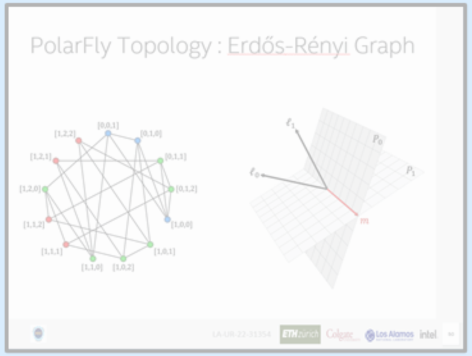
Scalability

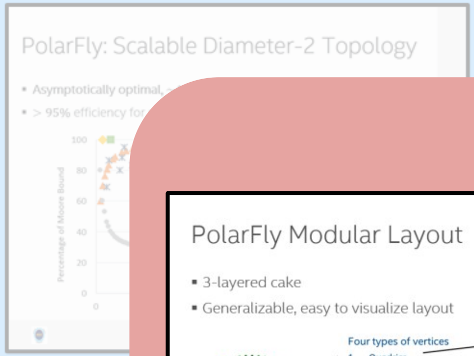


Design Space

Overview

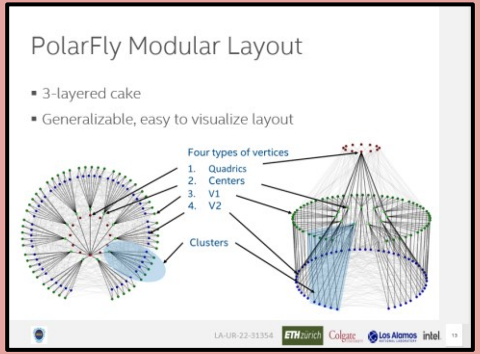
PolarFly Topology



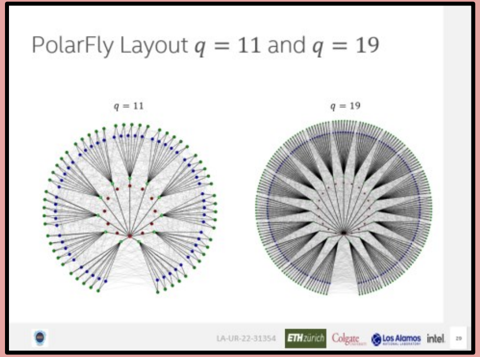


Construction

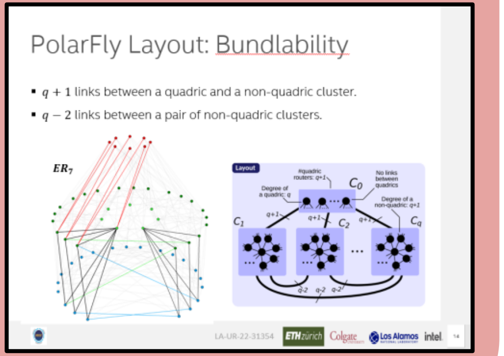
PolarFly Layout



Modular

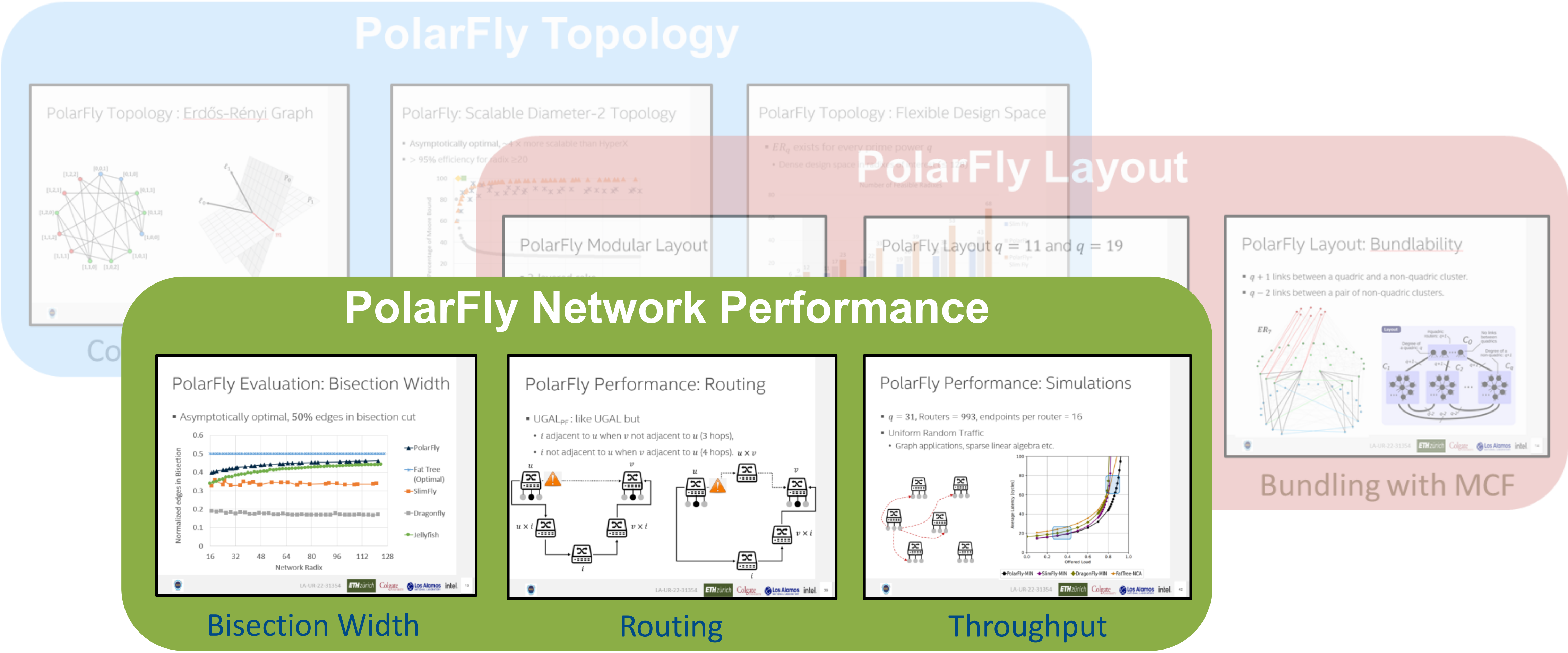


Generalized



Bundling with MCF

Overview


PolarFly Topology
PolarFly Topology : Erdős-Rényi Graph
PolarFly: Scalable Diameter-2 Topology
PolarFly Topology : Flexible Design Space
PolarFly Layout
PolarFly Modular Layout
PolarFly Layout $q = 11$ and $q = 19$
PolarFly Layout: Bundlability
$q+1$ links between a quadric and a non-quadric cluster.
$q-2$ links between a pair of non-quadric clusters.
Bundling with MCF
PolarFly Network Performance
PolarFly Evaluation: Bisection Width
Asymptotically optimal, 50% edges in bisection cut
Normalized edges in Bisection
Network Radix
PolarFly
Fat Tree (Optimal)
SlimFly
Dragonfly
Jellyfish
PolarFly Performance: Routing
UGAL$_{PF}$: like UGAL but
i adjacent to u when v not adjacent to u (3 hops),
i not adjacent to u when v adjacent to u (4 hops).
PolarFly Performance: Simulations
$q = 31$, Routers $= 993$, endpoints per router $= 16$
Uniform Random Traffic
Graph applications, sparse linear algebra etc.
Bisection Width
Routing
Throughput

Overview

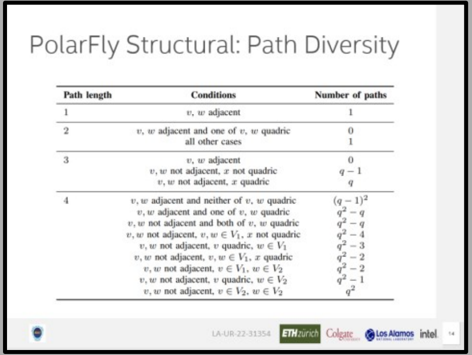
PolarFly Topology

PolarFly Topology : Erdős-Rényi Graph

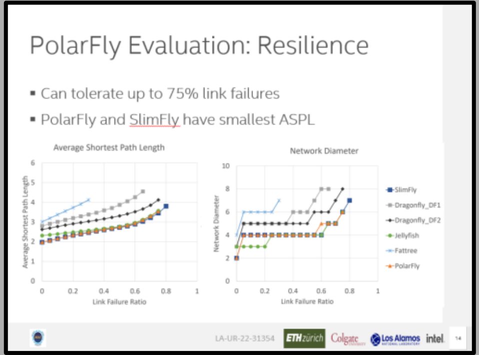
PolarFly: Scalable Diameter-2 Topology

PolarFly Topology : Flexible Design Space

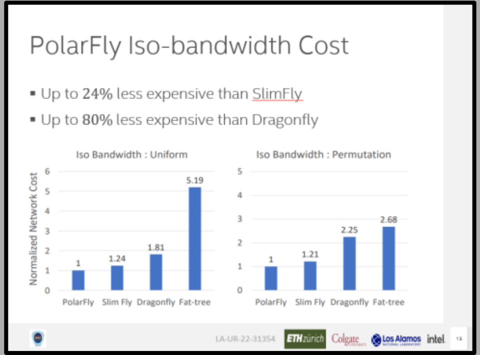
More on PolarFly



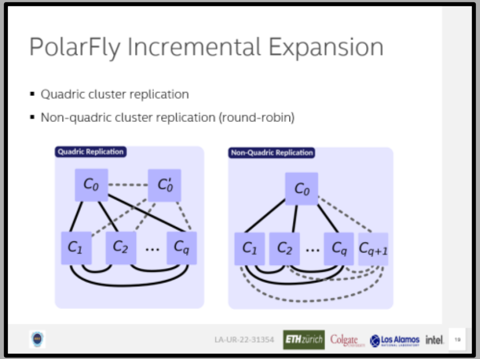
Path Diversity



Resilience

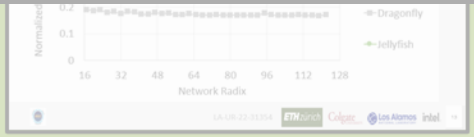


Cost-effectiveness



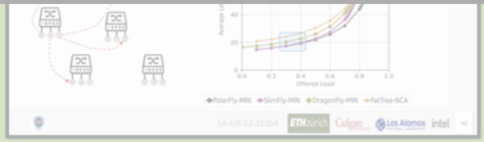
Expandability

.....



Bisection Width

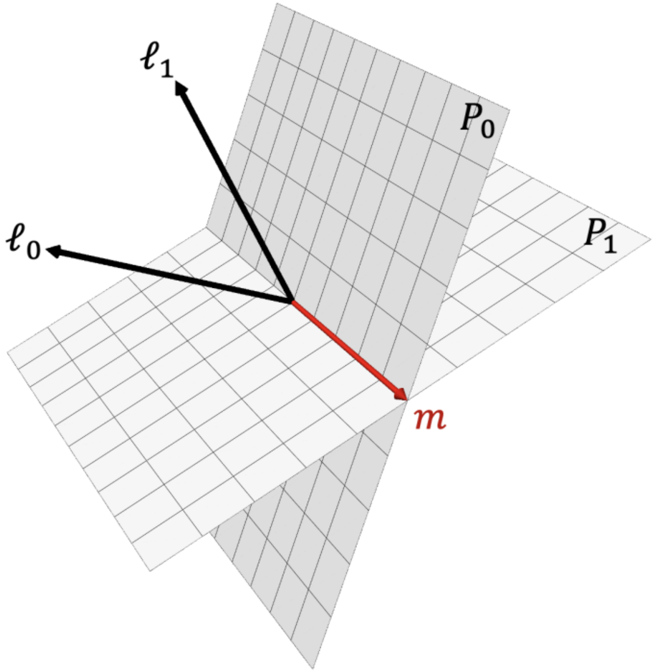
Routing



Throughput

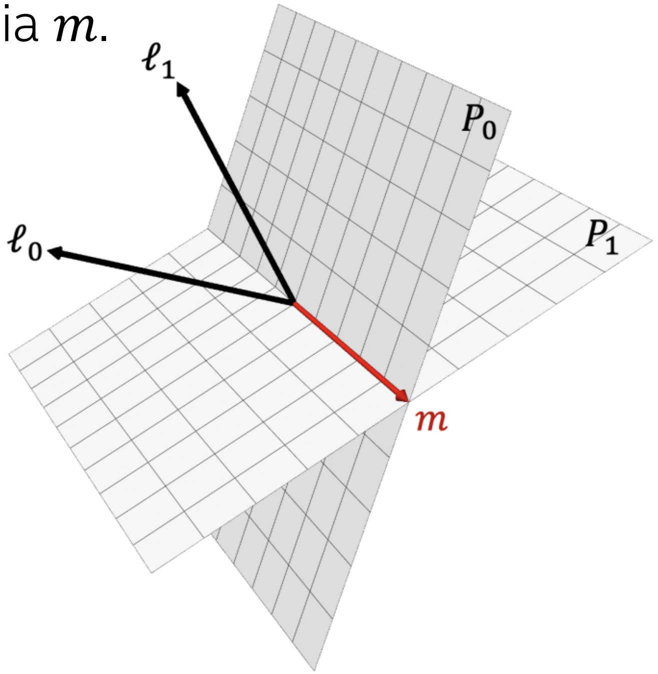
PolarFly Topology : Erdős-Rényi Graph

- If $l_0 \neq l_1$ are any two vectors, there is a vector m orthogonal to both.
  - m is the cross-product.



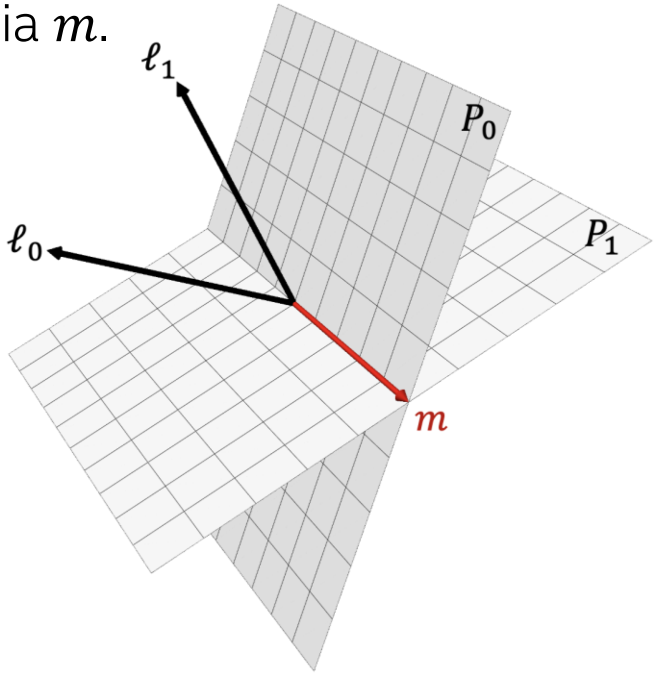
PolarFly Topology : Erdős-Rényi Graph

- If $l_0 \neq l_1$ are any two vectors, there is a vector m orthogonal to both.
  - m is the cross-product.
- What if a graph's edges expressed dot-product orthogonality
  - (l_0, m) and (m, l_1) are edges in the graph, so you can get from l_0 to l_1 via m.
  - This graph has **diameter $\mathbf{2}$**.



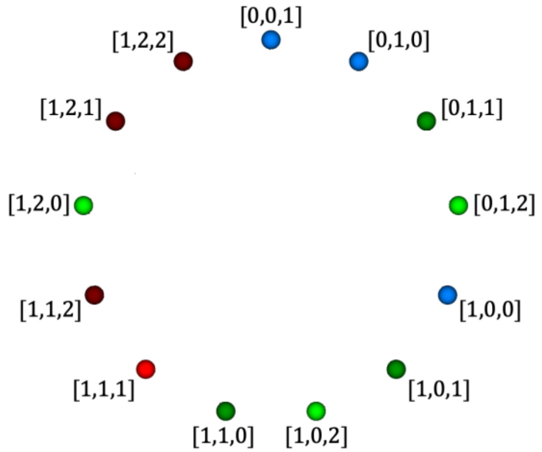
PolarFly Topology : Erdős-Rényi Graph

- If $l_0 \neq l_1$ are any two vectors, there is a vector m orthogonal to both.
  - m is the cross-product.
- What if a graph's edges expressed dot-product orthogonality
  - (l_0, m) and (m, l_1) are edges in the graph, so you can get from l_0 to l_1 via m.
  - This graph has **diameter $\mathbf{2}$**.
- Orthogonality is scale invariant
  - Vertices of $ER_q \leftarrow$ non-$\mathbf{0}$ left-normalized vectors from $\mathbb{F}_q^3$.
  - Degree $\leftarrow q+1$, # Vertices $\leftarrow q^2+q+1$, *very close to Moore bound*.



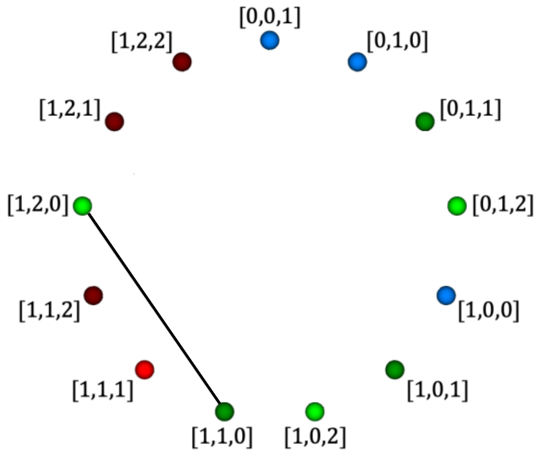
Erdős-Rényi Topology Example : ER_3

- 13 left-normalized 3-vectors in $\mathbb{F}_q^3$.
- v and w are adjacent iff $v_0w_0 + v_1w_1 + v_2w_2 \equiv 0 \pmod 3$.



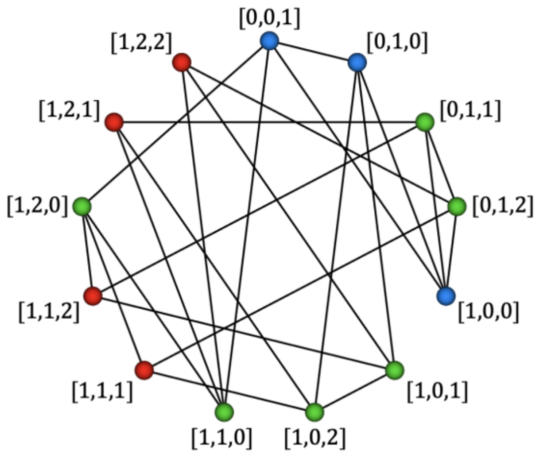
Erdős-Rényi Topology Example : ER_3

- 13 left-normalized 3-vectors in $\mathbb{F}_q^3$.
- v and w are adjacent iff $v_0 w_0 + v_1 w_1 + v_2 w_2 \equiv 0 \pmod 3$.
  - For example, $[1,1,0] \cdot [1,2,0] = 1 + 2 + 0 = 0$ (mod 3).



Erdős-Rényi Topology Example : ER_3

- 13 left-normalized 3-vectors in $\mathbb{F}_q^3$.
- v and w are adjacent iff $v_0 w_0 + v_1 w_1 + v_2 w_2 \equiv 0$ (mod 3).
  - For example, $[1{,}1{,}0] \cdot [1{,}2{,}0] = 1 + 2 + 0 = 0$ (mod 3).
- Some vectors are self-orthogonal
  - For example, $[1{,}1{,}1] \cdot [1{,}1{,}1] = 1 + 1 + 1 = 0$ (mod 3).
  - These are called ***quadrics*** (colored red).

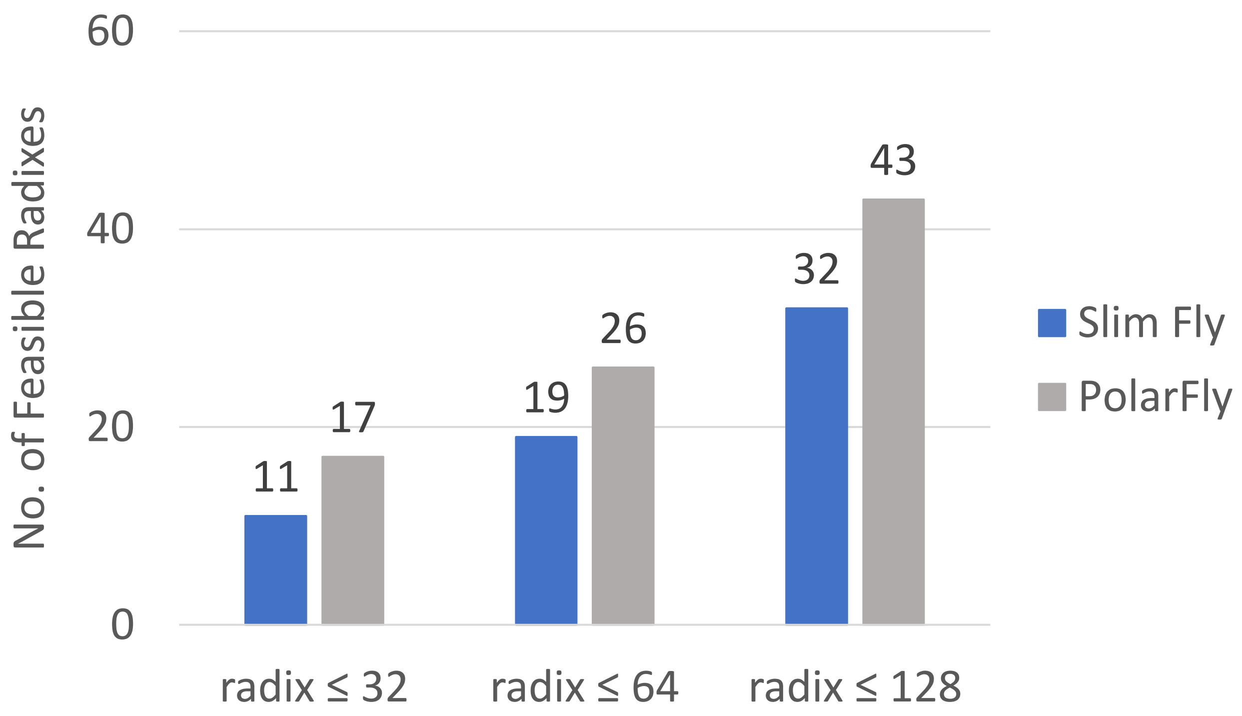


PolarFly Topology : Flexible Design Space

- ER_q exists for every prime power q

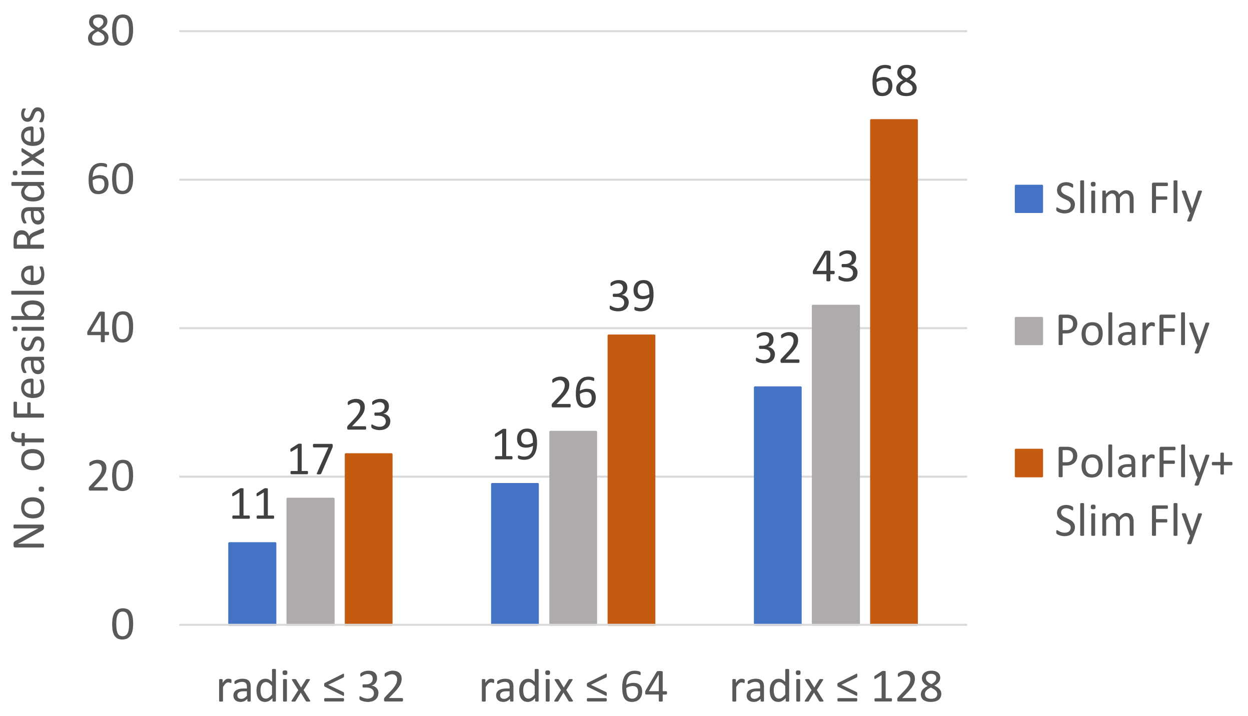
PolarFly Topology : Flexible Design Space

- ER_q exists for every prime power q
  - Dense design space in radixes of interest (≤ 128)



PolarFly Topology : Flexible Design Space

- ER_q exists for every prime power q
  - Dense design space in radixes of interest (≤ 128)
- Complemented by SlimFly
  - $> 50\%$ radixes covered



PolarFly Topology Summary and what next?

- So we have a good low-diameter topology with potential.
  - Diameter = 2
  - Highly scalable, asymptotically approaches Moore bound ***very quickly***.
  - Simple Construction, ***flexible design space.***

PolarFly Topology Summary and what next?

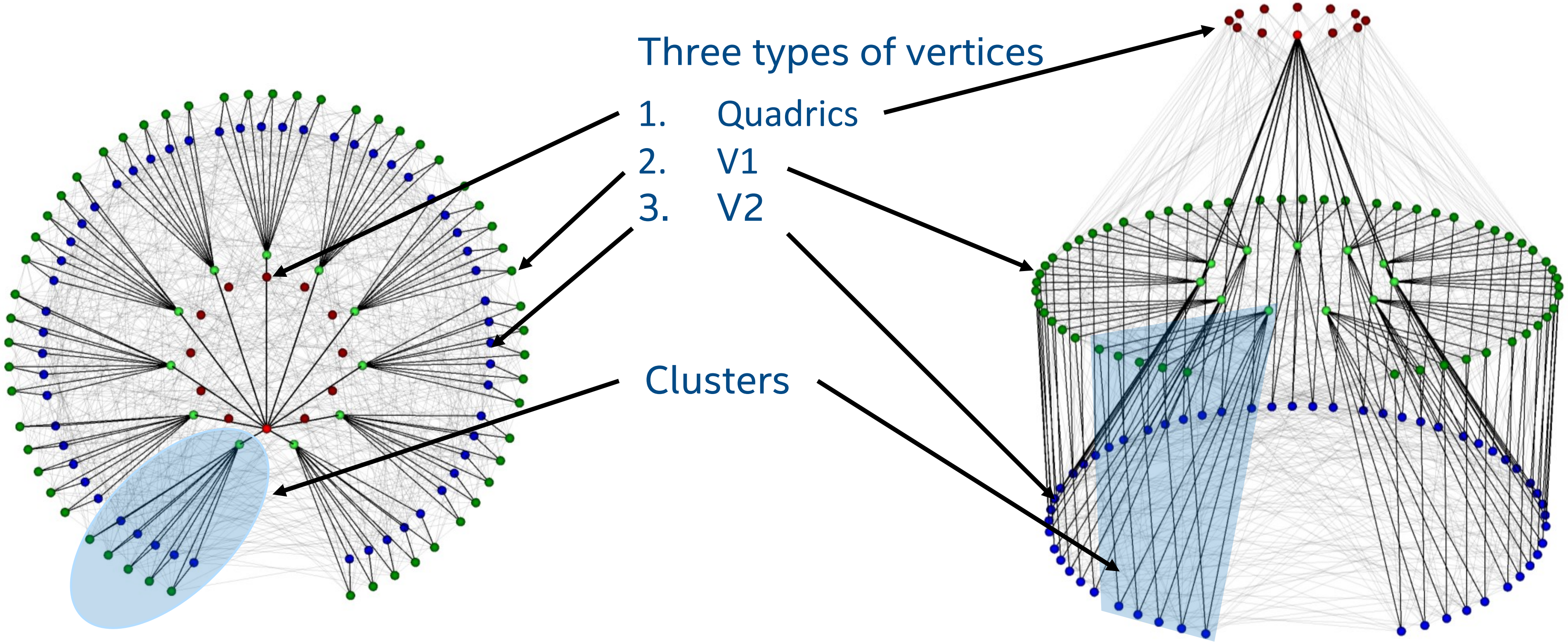
- So we have a good low-diameter topology with potential.
  - Diameter = 2
  - Highly scalable, asymptotically approaches Moore bound ***very quickly***.
  - Simple Construction, ***flexible design space.***
- How do we lay it out to make a usable network?

PolarFly Topology Summary and what next?

- So we have a good low-diameter topology with potential.
  - Diameter = 2
  - Highly scalable, asymptotically approaches Moore bound ***very quickly***.
  - Simple Construction, ***flexible design space.***
- How do we lay it out to make a usable network?
- PolarFly topology has a lot of mathematical structure – that helps

PolarFly Layout : Overview

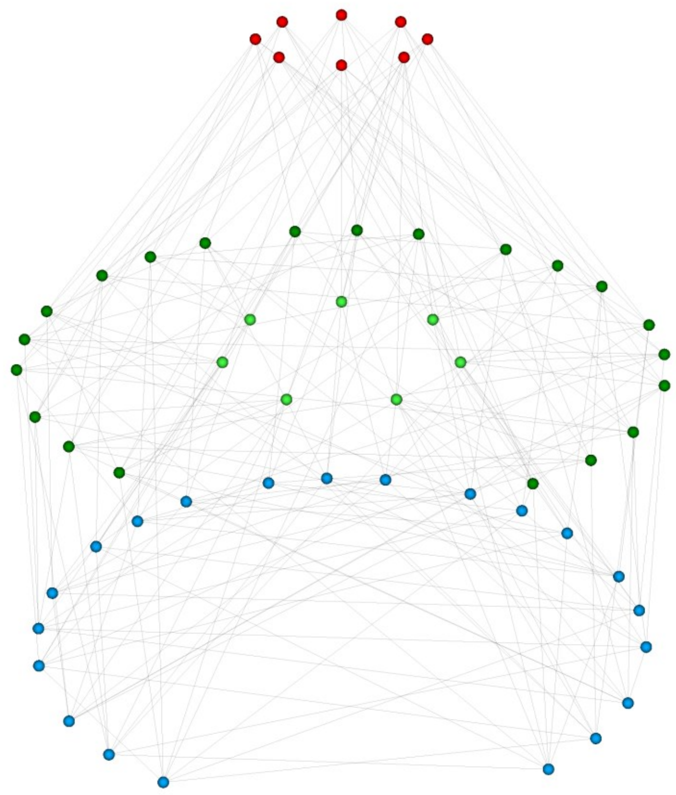
- 3-layered cake
- Generalizable, easy to visualize layout



A Modular Layout for PolarFly

- All self-orthogonal quadrics (red) form a cluster.

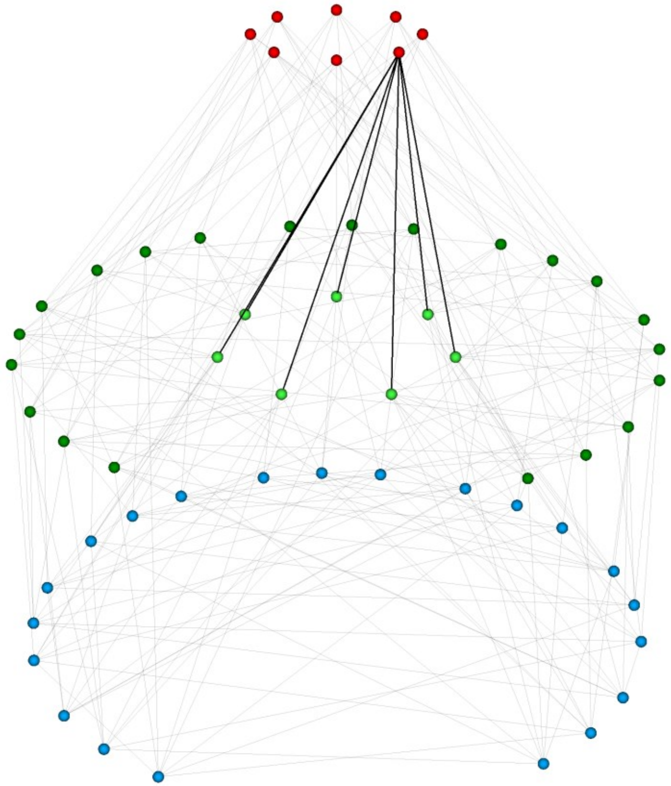
ER_7



A Modular Layout for PolarFly

- All self-orthogonal quadrics (red) form a cluster.
- Pick one as the starter quadric l.
  - Take all vectors c orthogonal to l.
  - These are q cluster centers.

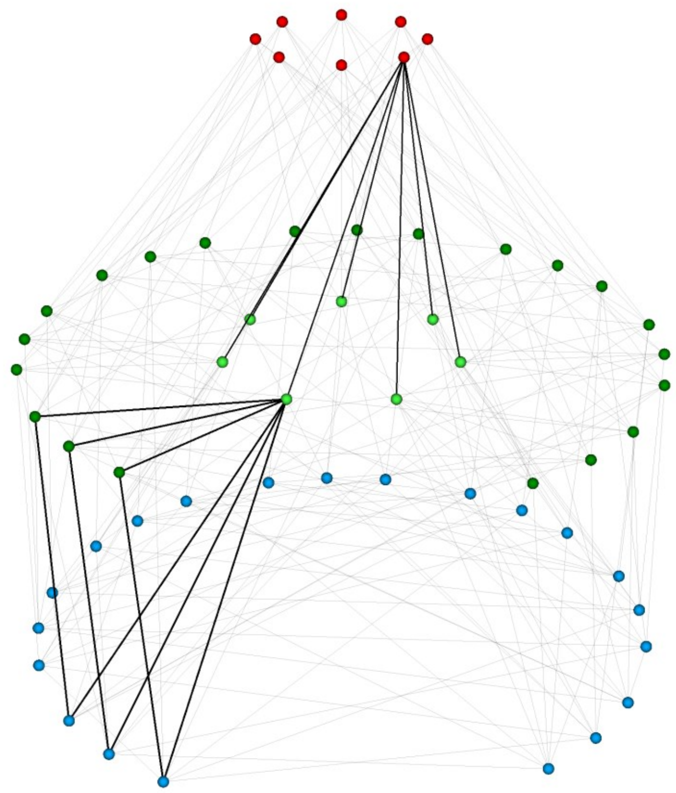
$\boldsymbol{ER_7}$



A Modular Layout for PolarFly

- All self-orthogonal quadrics (red) form a cluster.
- Pick one as the starter quadric l.
  - Take all vectors c orthogonal to l.
  - These are q cluster centers.
- Each center c starts its own cluster.
  - All q non-quadric vectors v orthogonal to c.

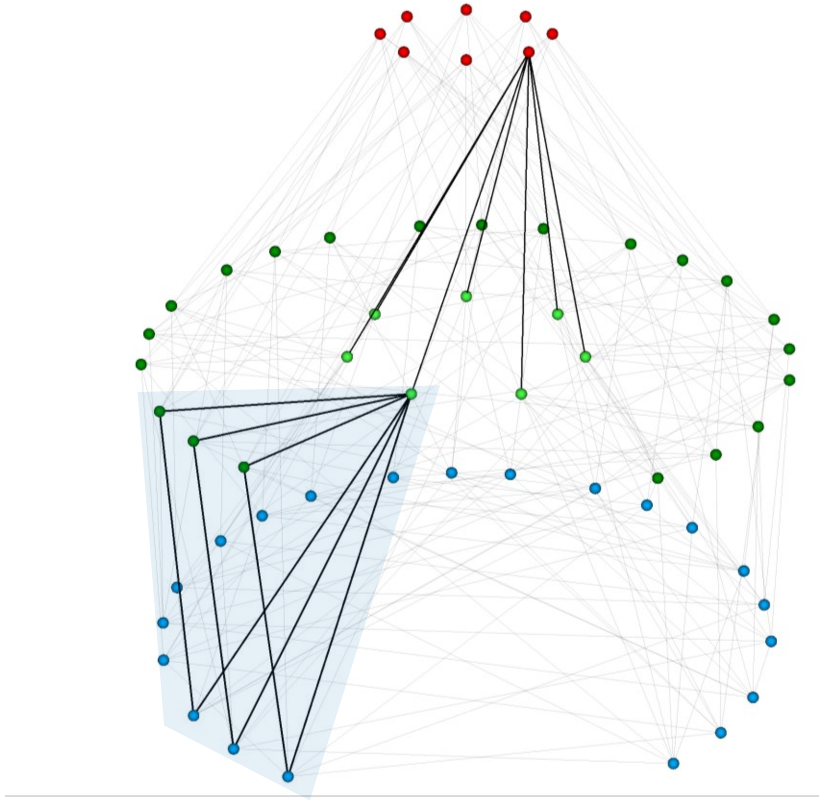
ER_7



PolarFly Layout Properties

- A non-quadric cluster induces $\frac{q-1}{2}$ edge disjoint triangles

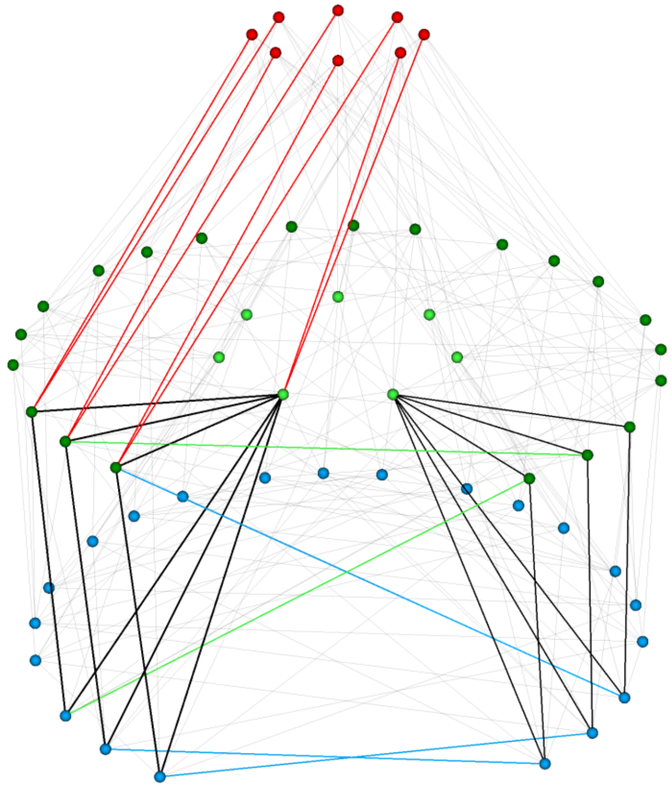
$\boldsymbol{ER_7}$



PolarFly Layout Properties

- A non-quadric cluster induces $\frac{q-1}{2}$ edge disjoint triangles.
- Inter-cluster Connectivity :
  - $q + 1$ links between a quadric and a non-quadric cluster.
  - $q - 2$ links between a pair of non-quadric clusters.
  - Can *bundle* into multi-core fibers.

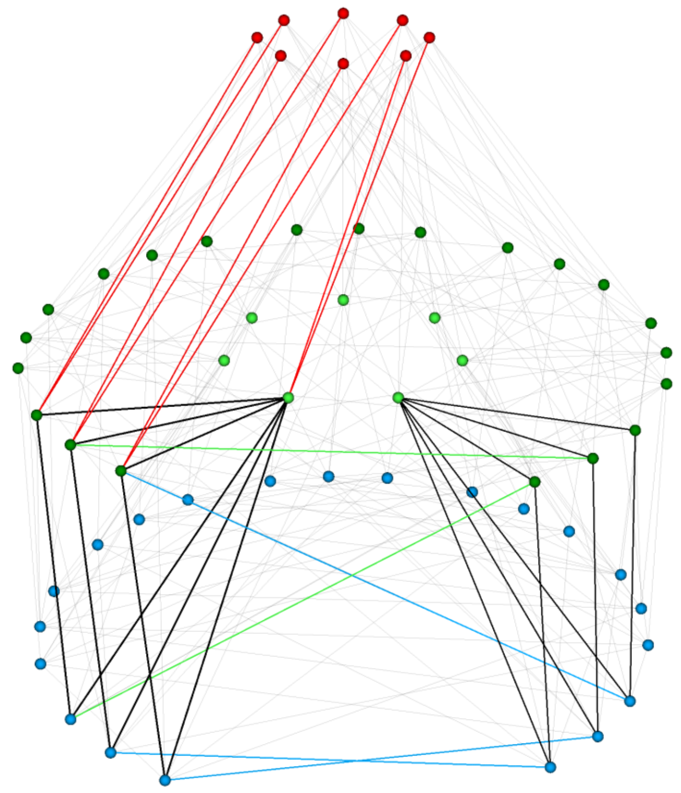
$\boldsymbol{ER_7}$



PolarFly Layout Properties

- A non-quadric cluster induces $\frac{q-1}{2}$ edge disjoint triangles.
- Inter-cluster Connectivity :
  - $q+1$ links between a quadric and a non-quadric cluster.
  - $q-2$ links between a pair of non-quadric clusters.
  - Can *bundle* into multi-core fibers.

$\boldsymbol{ER_7}$



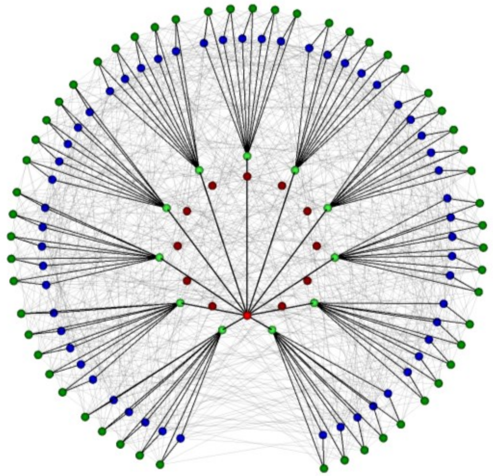
PolarFly Topology and Layout Summary

- Good Diameter-2 topology
  - Near-optimal scalability, state-of-the-art.
  - Flexible design space.

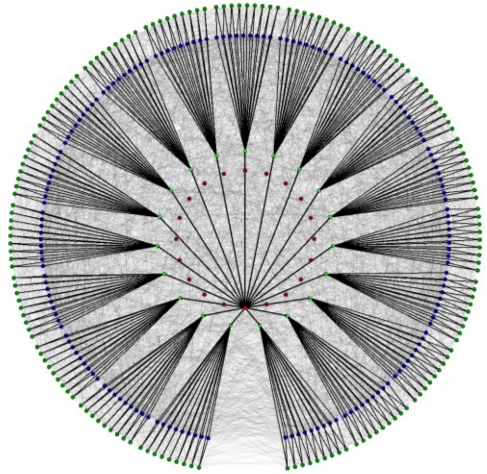
PolarFly Topology and Layout Summary

- Good Diameter-2 topology
  - Near-optimal scalability, state-of-the-art.
  - Flexible design space.
- Generalized and Modular layout.
  - Bundling into Multi-core Fiber – simplify cabling, reduce cost.

$q = 11$



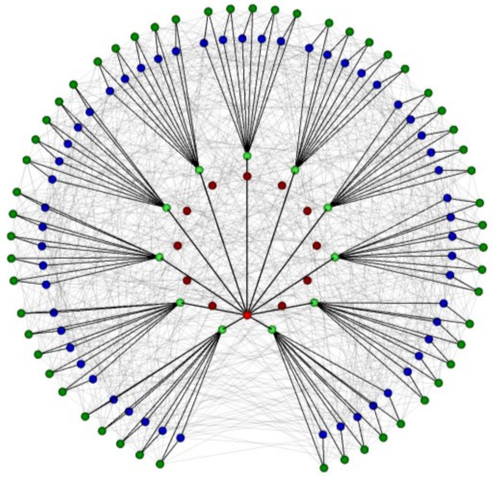
$q = 19$



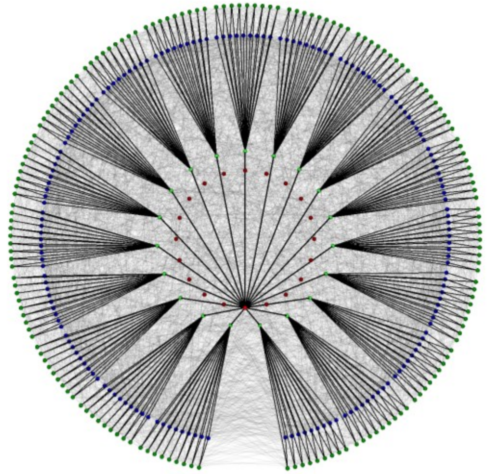
PolarFly Topology and Layout Summary

- Good Diameter-2 topology
  - Near-optimal scalability, state-of-the-art.
  - Flexible design space.
- Generalized and Modular layout.
  - Bundling into Multi-core Fiber – simplify cabling, reduce cost.
- How would it perform as a network??
  - Bisection width, Throughput?

$q = 11$

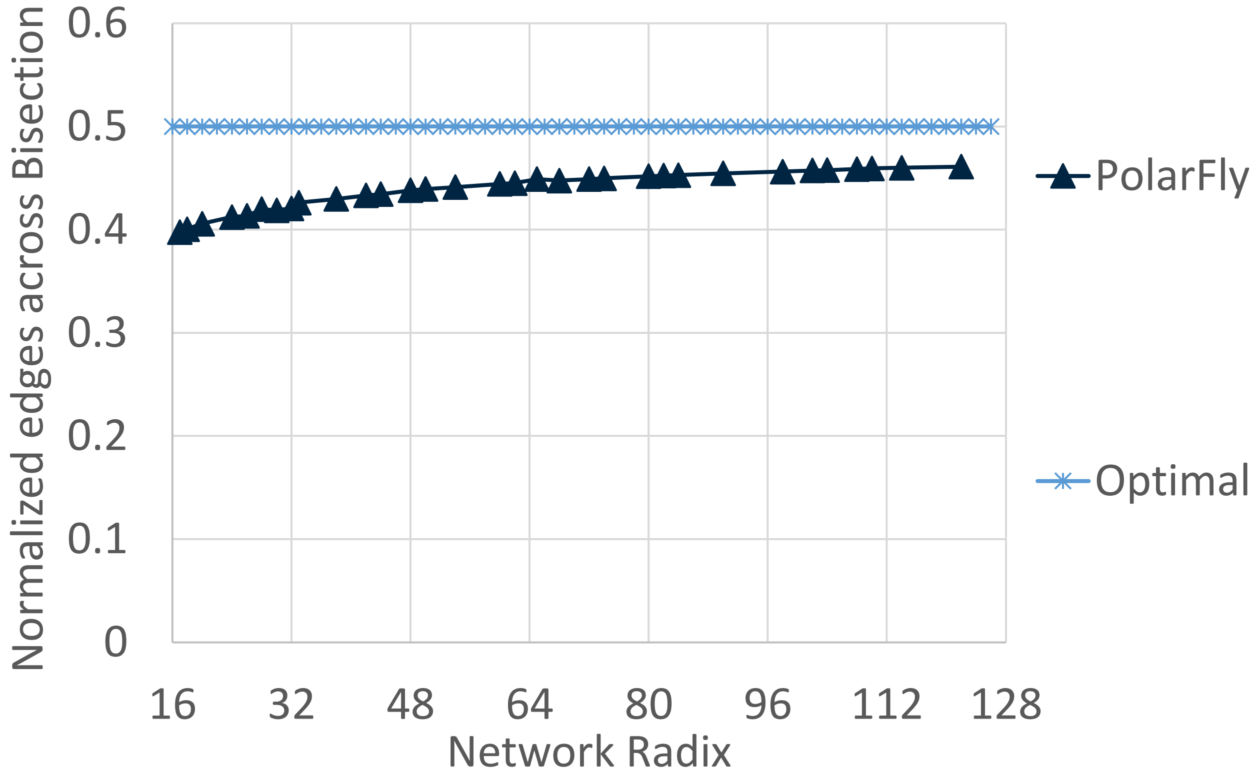


$q = 19$



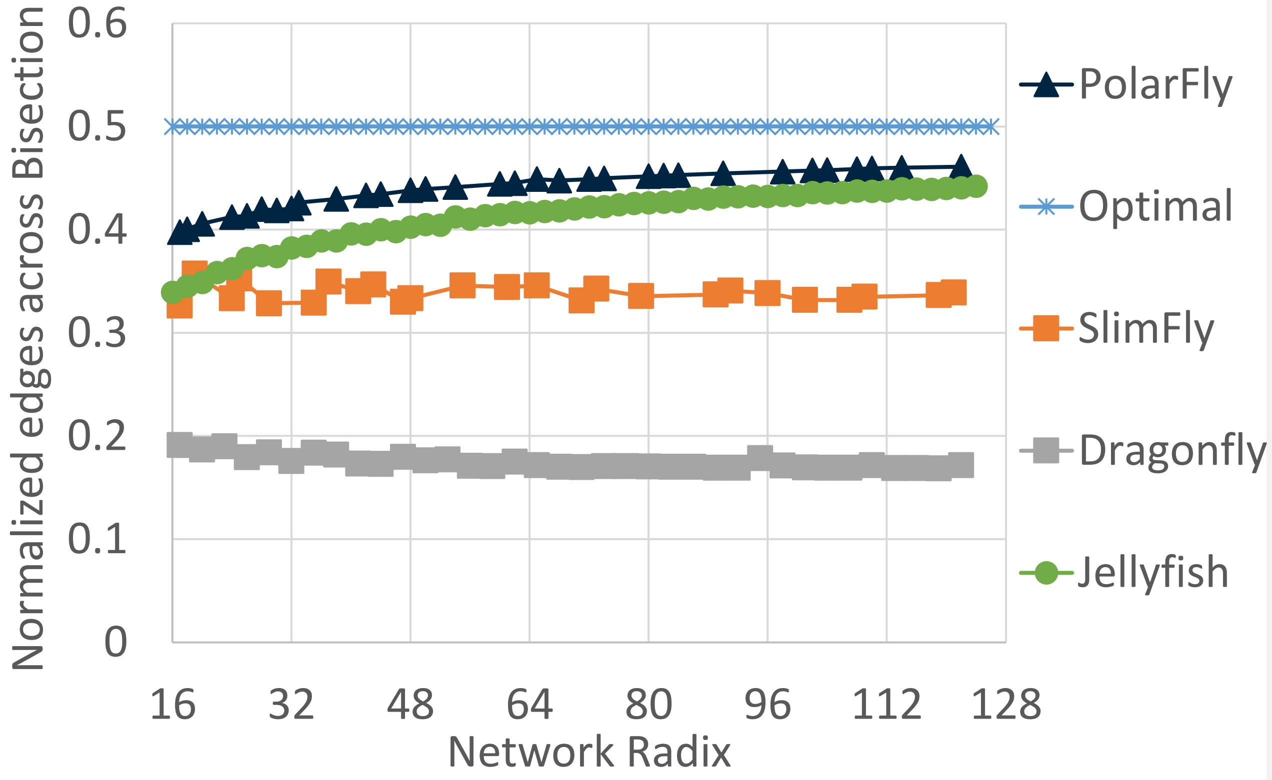
PolarFly Evaluation: Bisection Width

- Asymptotically optimal
  - 50% edges in bisection cut



PolarFly Evaluation: Bisection Width

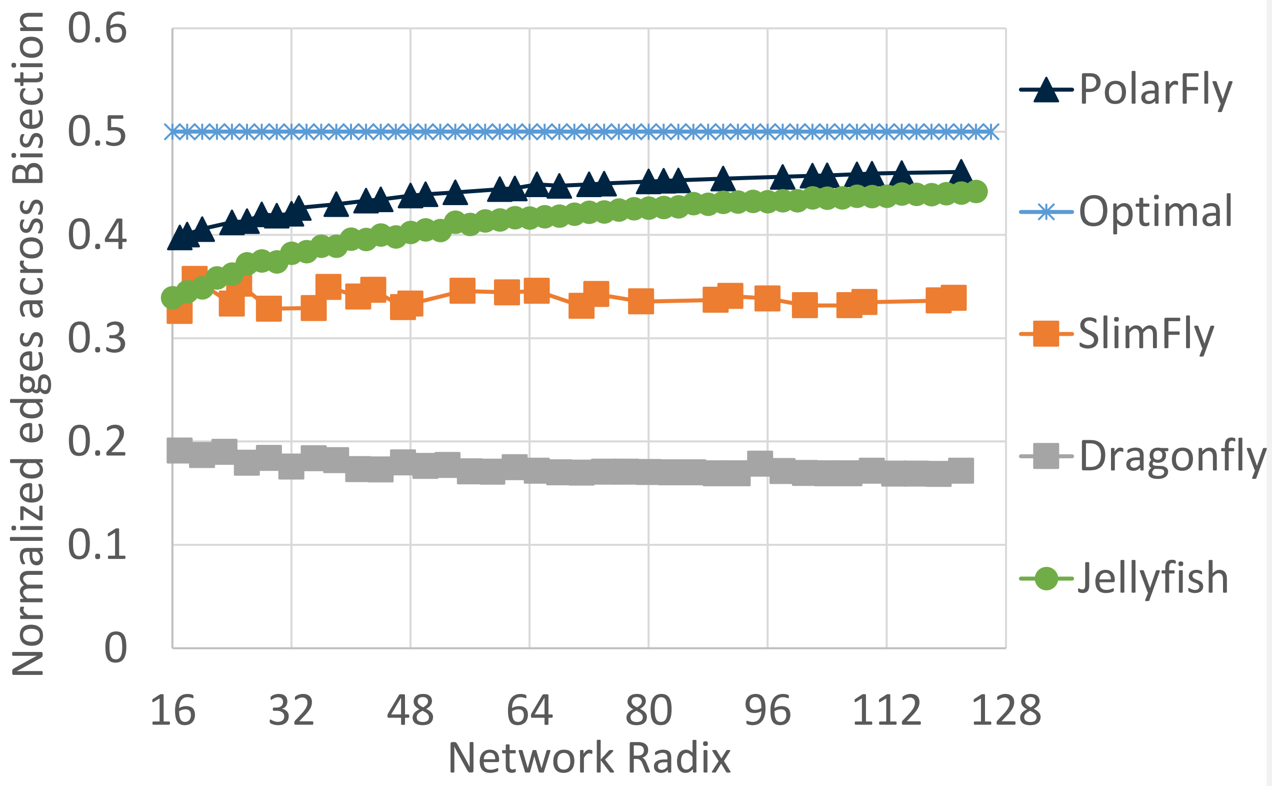
- Asymptotically optimal
  - 50% edges in bisection cut
- Higher fraction of edges across bisection than *any* direct network[1]
  - 28% geomean higher than SlimFly



[1] Karypis, George, and Vipin Kumar. "METIS: A software package for partitioning unstructured graphs, partitioning meshes, and computing fill-reducing orderings of sparse matrices." (1997).

PolarFly Evaluation: Bisection Width

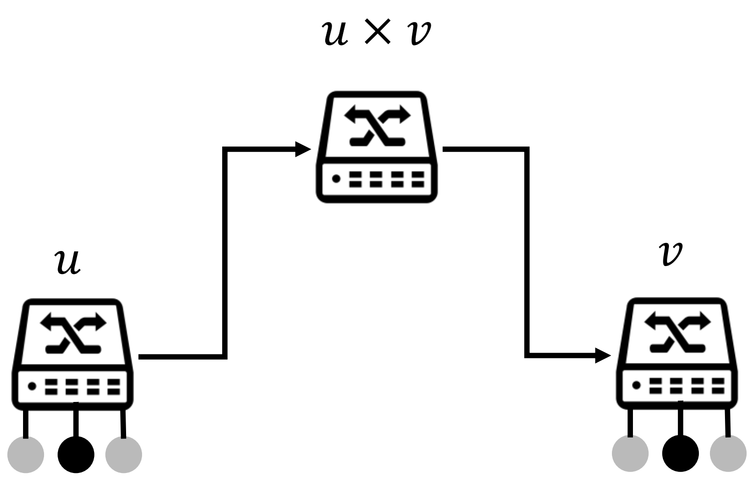
- Asymptotically optimal
  - 50% edges in bisection cut
- Higher fraction of edges across bisection than *any* direct network[1]
  - 28% geomean higher than SlimFly
- ↑ scalability → ↑ expansion
  - Any vertex subset has lot of edges to other half



[1] Karypis, George, and Vipin Kumar. "METIS: A software package for partitioning unstructured graphs, partitioning meshes, and computing fill-reducing orderings of sparse matrices." (1997).

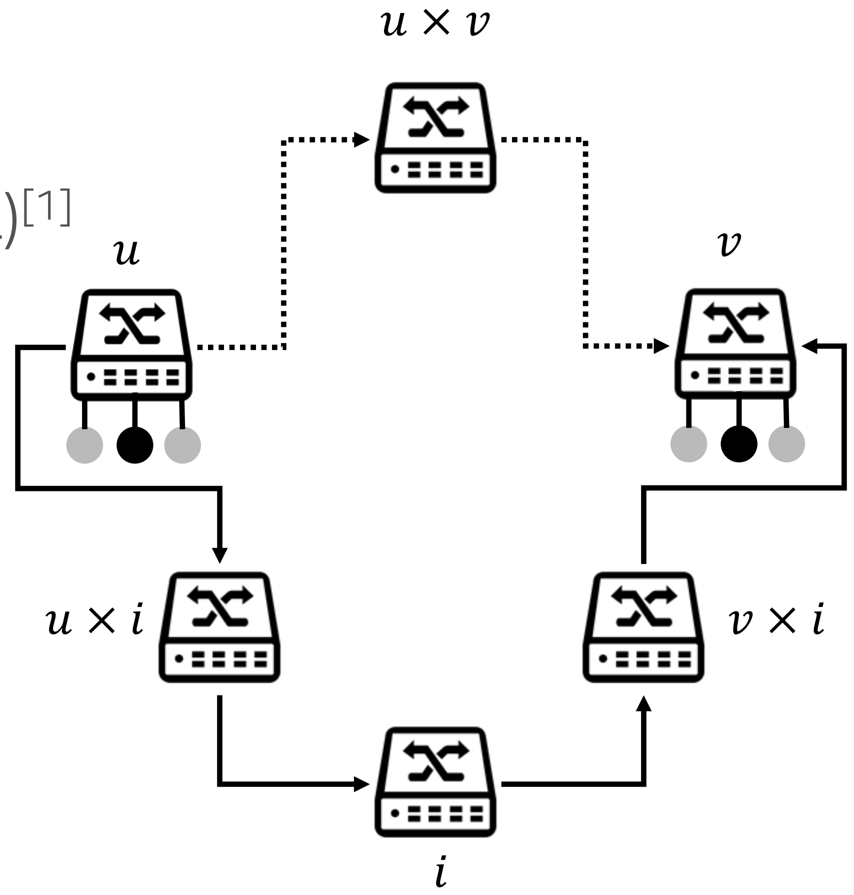
PolarFly Performance: Routing

- Source Router u, Destination Router v
- Minpath Routing (MIN) : $u \rightarrow u \times v \rightarrow v$



PolarFly Performance: Routing

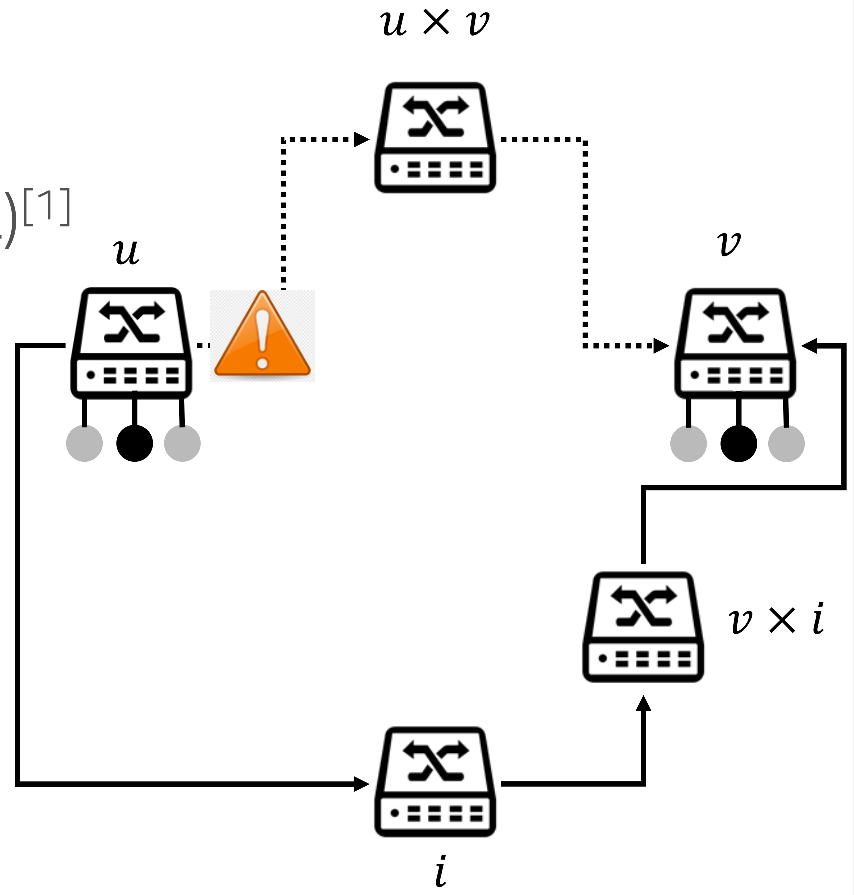
- Source Router u, Destination Router v
- Minpath Routing (MIN) : $u \rightarrow u \times v \rightarrow v$
- Universal Globally-Adaptive Load-Balancing (UGAL)[1]
  - MIN + option to misroute via random router i
  - Latency estimates using local queues



[1] A. Singh. Load-Balanced Routing in Interconnection Networks. PhD thesis, Stanford University, 2005

PolarFly Performance: Routing

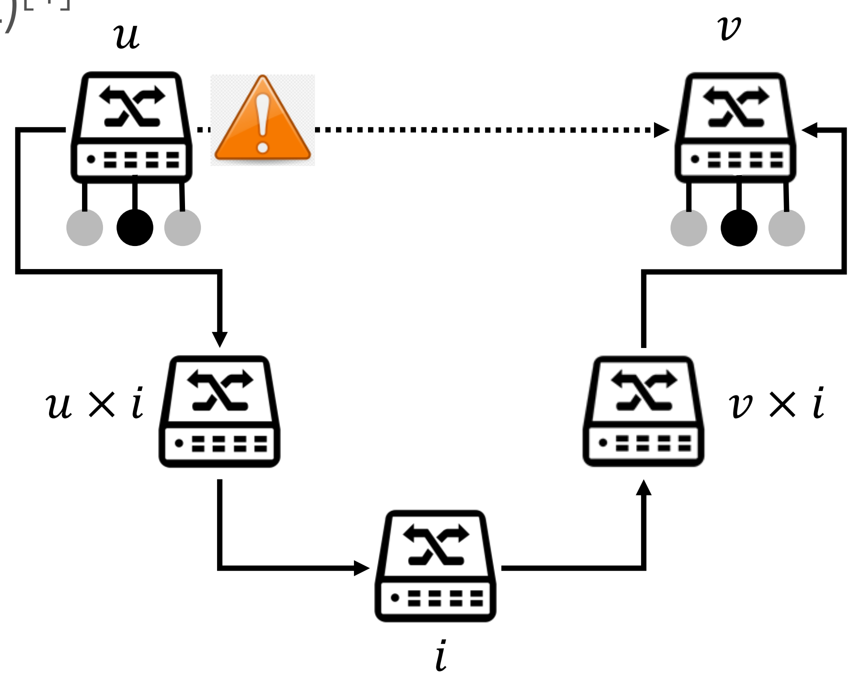
- Source Router u, Destination Router v
- Minpath Routing (MIN) : $u \rightarrow u \times v \rightarrow v$
- Universal Globally-Adaptive Load-Balancing (UGAL)[1]
  - MIN + option to misroute via random router i.
  - Latency estimates using local queues.
- UGAL_{PF} : like UGAL with adaptation threshold, but
  - i adjacent to u when v not adjacent to u (3 hops),



[1] A. Singh. Load-Balanced Routing in Interconnection Networks. PhD thesis, Stanford University, 2005

PolarFly Performance: Routing

- Source Router u, Destination Router v
- Minpath Routing (MIN) : $u \rightarrow u \times v \rightarrow v$
- Universal Globally-Adaptive Load-Balancing (UGAL)[1]
  - MIN + option to misroute via random router i
  - Latency estimates using local queues
- UGAL_{PF} : like UGAL with adaptation threshold, but
  - i adjacent to u when v not adjacent to u (3 hops),
  - i not adjacent to u when v adjacent to u (4 hops).



[1] A. Singh. Load-Balanced Routing in Interconnection Networks. PhD thesis, Stanford University, 2005

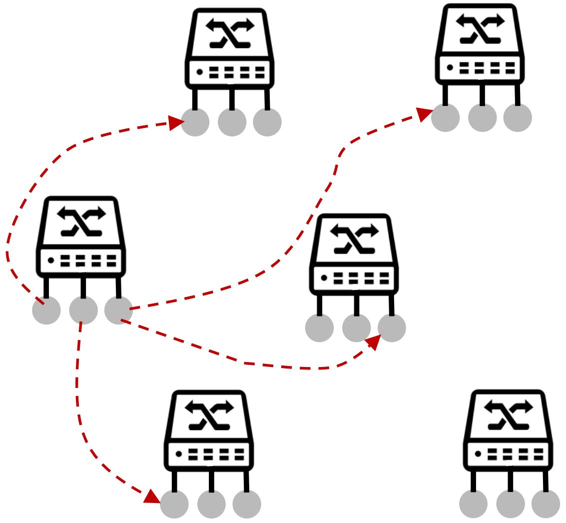
PolarFly Performance: Cycle-accurate Simulations[1]

- $q = 31$, Routers $= 993$, endpoints per router = 16

[1] Jiang, Nan, et al. "A detailed and flexible cycle-accurate network-on-chip simulator." *2013 IEEE international symposium on performance analysis of systems and software (ISPASS).*

PolarFly Performance: Cycle-accurate Simulations[1]

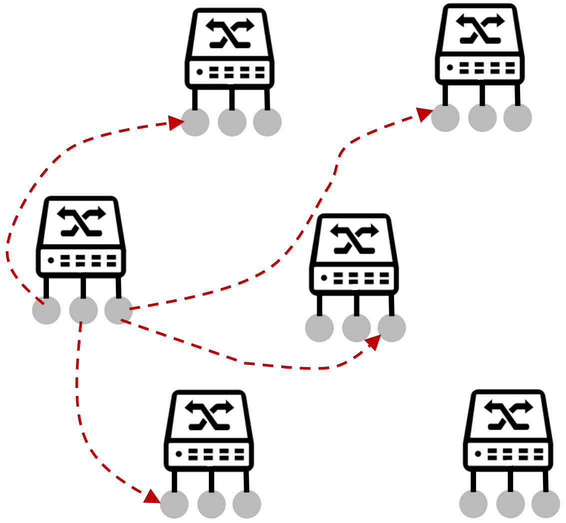
- $q = 31$, Routers $= 993$, endpoints per router = 16
- Uniform Random Traffic
  - Graph applications, sparse linear algebra etc.

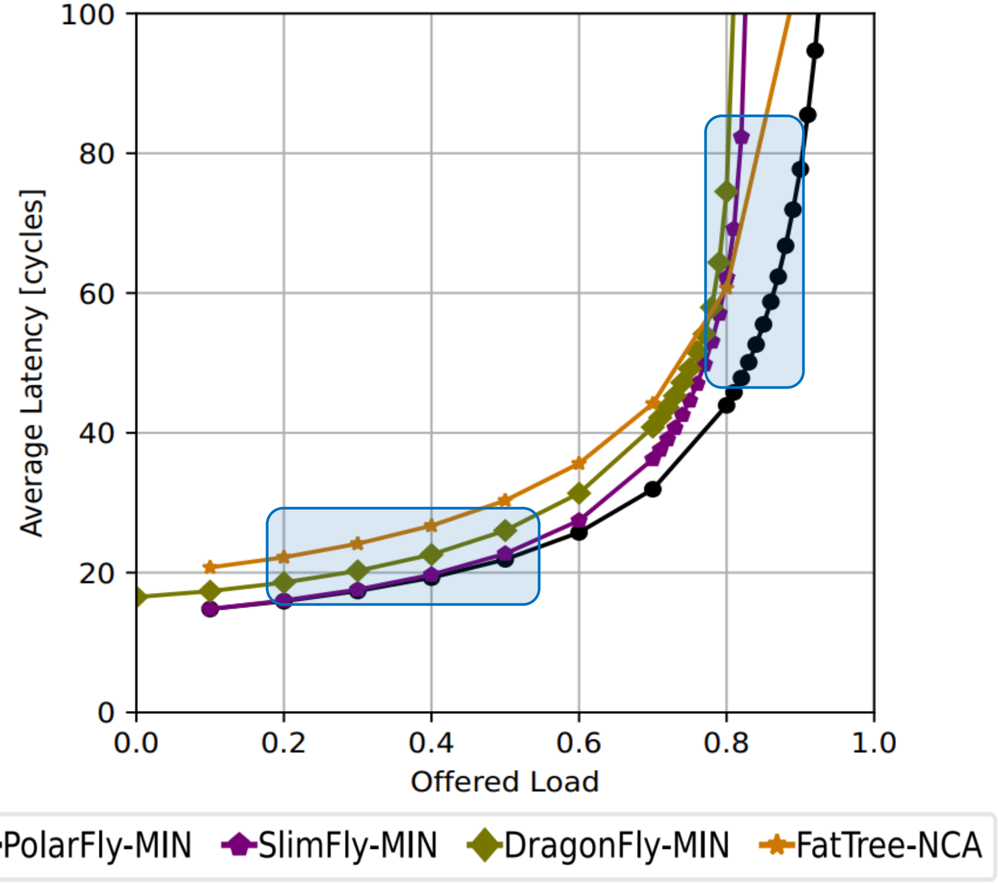


[1] Jiang, Nan, et al. "A detailed and flexible cycle-accurate network-on-chip simulator." *2013 IEEE international symposium on performance analysis of systems and software (ISPASS).*

PolarFly Performance: Cycle-accurate Simulations[1]

- $q = 31$, Routers $= 993$, endpoints per router = 16
- Uniform Random Traffic
  - Graph applications, sparse linear algebra etc.

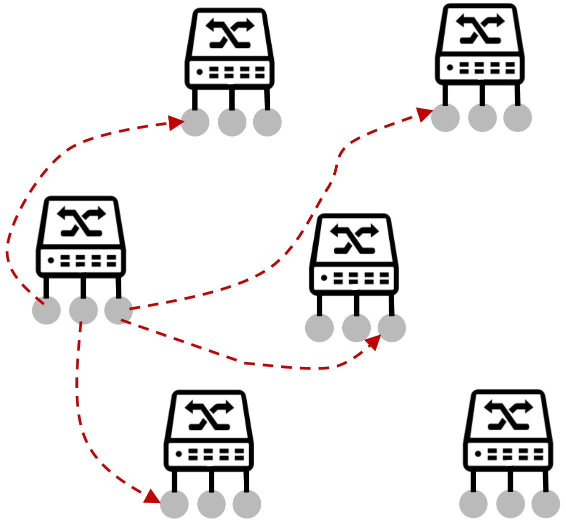


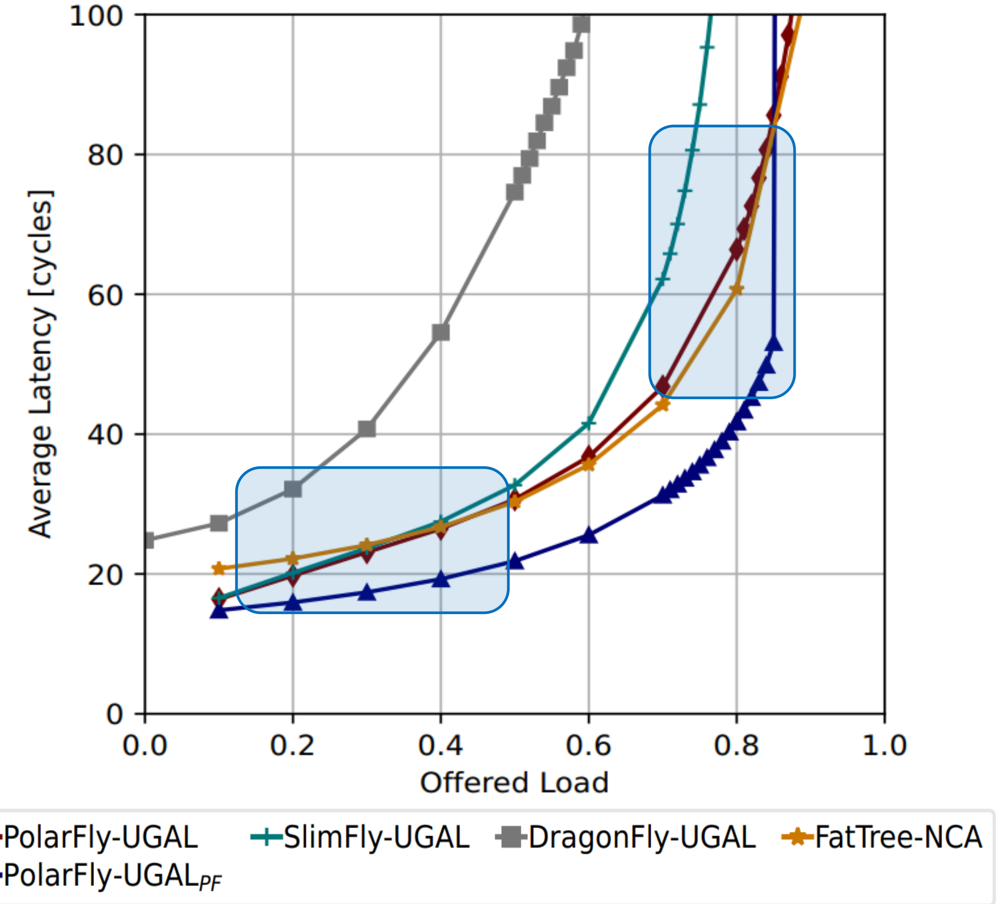


[1] Jiang, Nan, et al. "A detailed and flexible cycle-accurate network-on-chip simulator." *2013 IEEE international symposium on performance analysis of systems and software (ISPASS).*

PolarFly Performance: Cycle-accurate Simulations[1]

- $q = 31$, Routers $= 993$, endpoints per router $= 16$
- Uniform Random Traffic
  - Graph applications, sparse linear algebra etc.

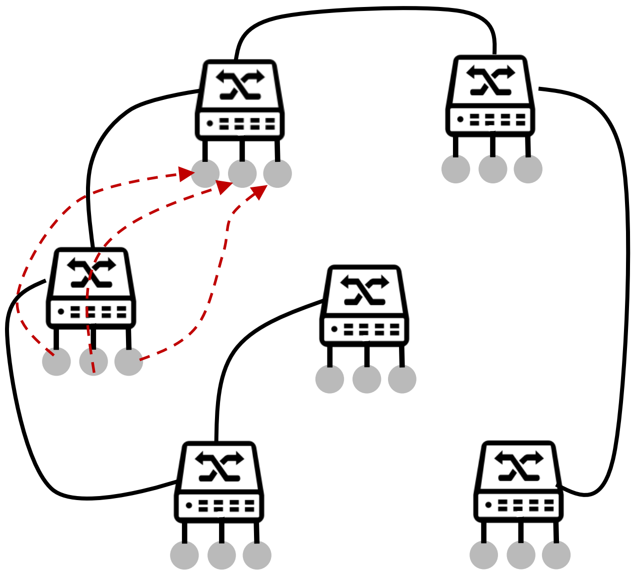




[1] Jiang, Nan, et al. "A detailed and flexible cycle-accurate network-on-chip simulator." *2013 IEEE international symposium on performance analysis of systems and software (ISPASS).*

PolarFly Performance: Cycle-accurate Simulations[1]

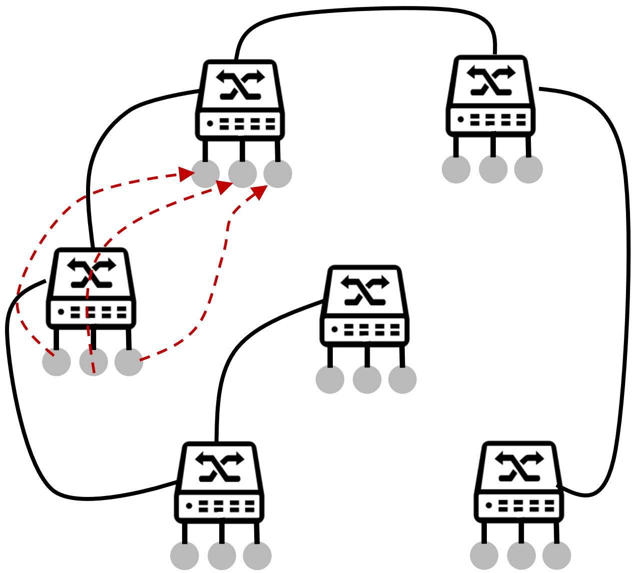
- Adversarial Traffic Pattern
  - All traffic from a router goes to one neighbor
  - Adaptive misrouting takes 4-hops

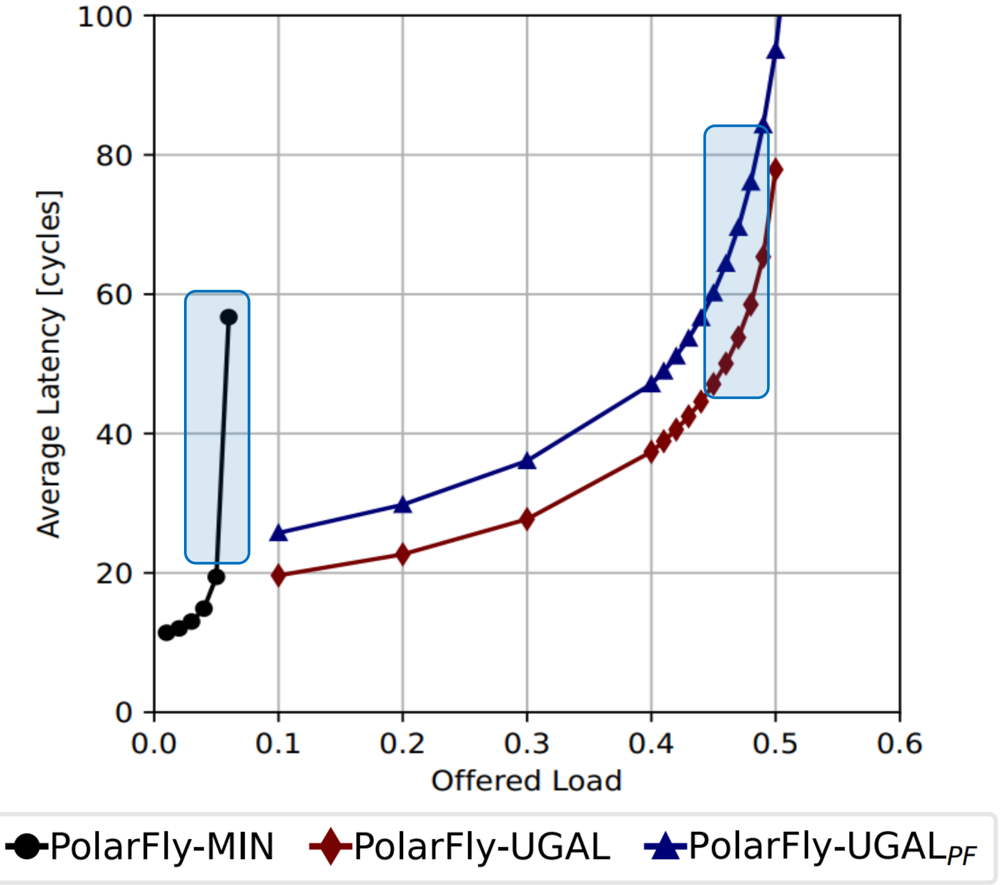


[1] Jiang, Nan, et al. "A detailed and flexible cycle-accurate network-on-chip simulator." *2013 IEEE international symposium on performance analysis of systems and software (ISPASS).*

PolarFly Performance: Cycle-accurate Simulations[1]

- Adversarial Traffic Pattern
  - All traffic from a router goes to one neighbor
  - Adaptive misrouting takes 4-hops





[1] Jiang, Nan, et al. "A detailed and flexible cycle-accurate network-on-chip simulator." *2013 IEEE international symposium on performance analysis of systems and software (ISPASS).*

PolarFly Summary

- Formal mathematical approach for scalable network design
  - Diameter-2 topology with near-optimal scalability

PolarFly Summary

- Formal mathematical approach for scalable network design
  - Diameter-2 topology with near-optimal scalability
- Modular layout amenable to bundling

PolarFly Summary

- Formal mathematical approach for scalable network design
  - Diameter-2 topology with near-optimal scalability
- Modular layout amenable to bundling
- High Performance, asymptotically optimal bisection width

PolarFly Summary

- Formal mathematical approach for scalable network design
  - Diameter-2 topology with near-optimal scalability
- Modular layout amenable to bundling
- High Performance, asymptotically optimal bisection width
- More in the paper
  - Iso bandwidth cost per node: 24% and 80% lower than Slim Fly and Dragonfly
  - Structural Analysis : Large non-minimal path diversity, resilient to link failures
  - Expandability: Incremental growth by cluster replication

Thank you!

kartik.lakhotia@intel.com