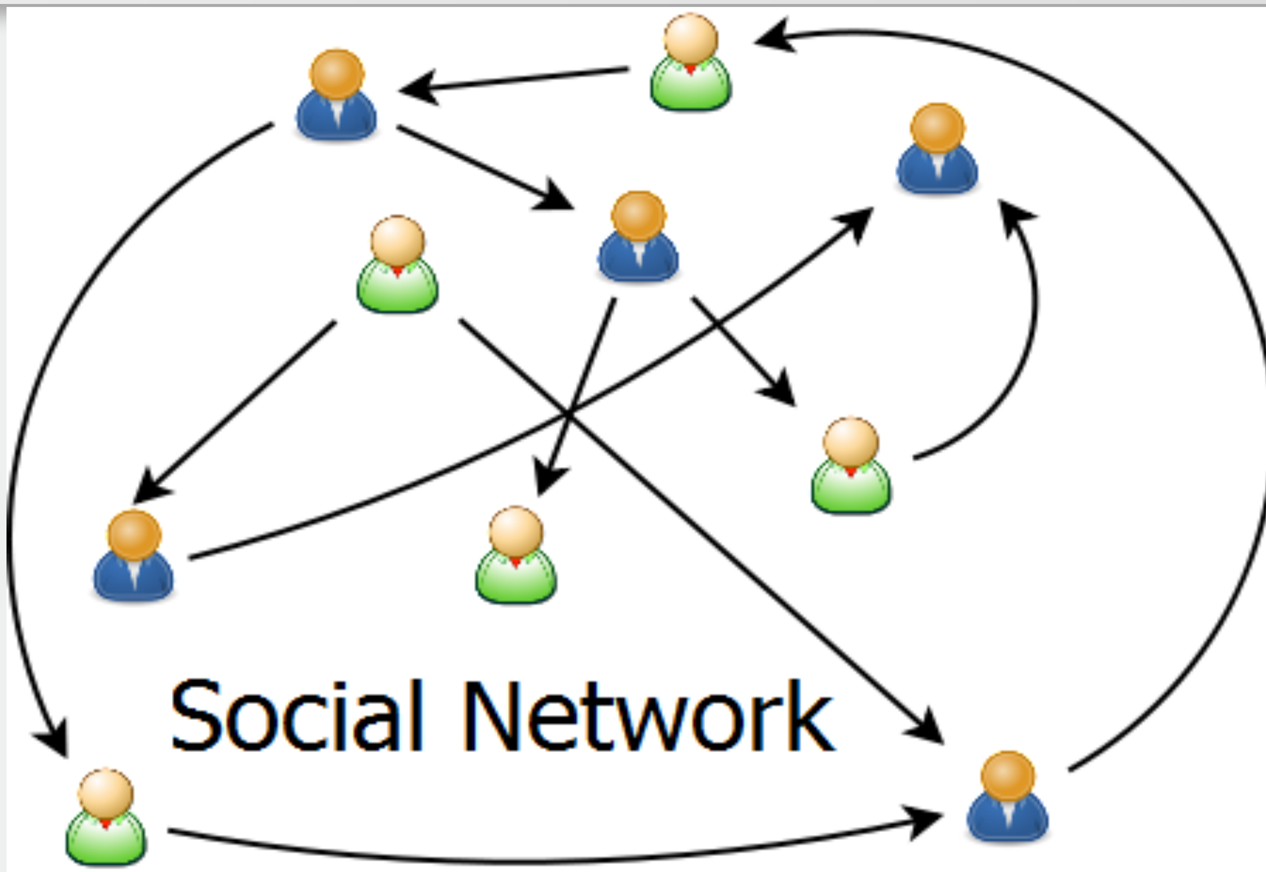
Order/Radix Problem: Towards Low End-to-End Latency Interconnection Networks

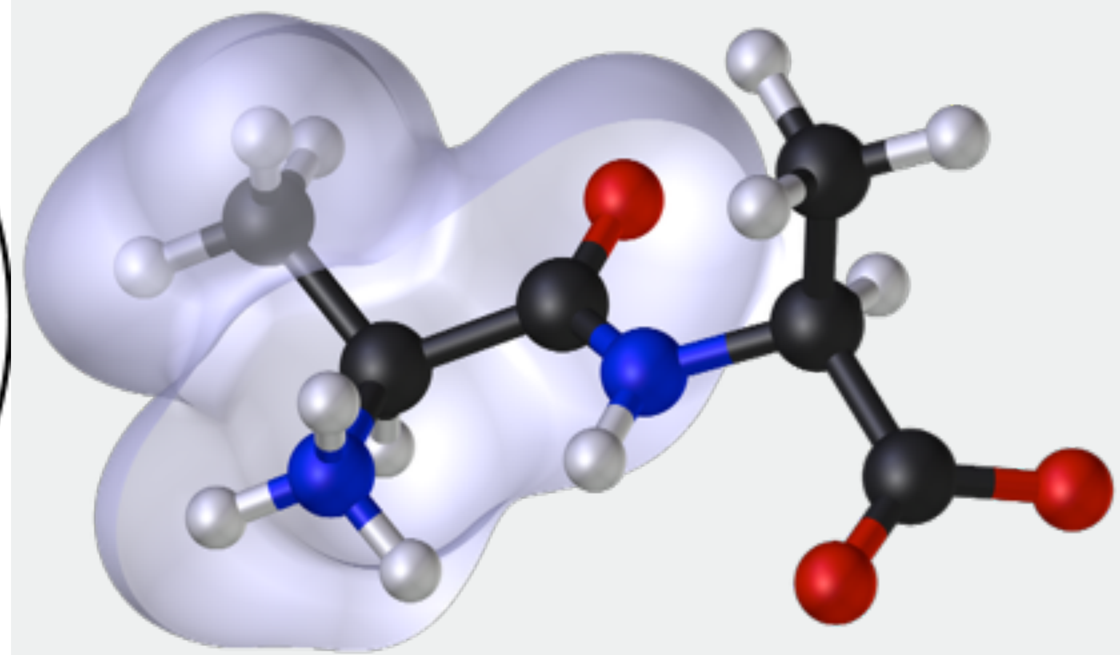
Ryota Yasudo[1], Michihiro Koibuchi[2], Koji Nakano[3], Hiroki Matsutani[1], and Hideharu Amano[1]

[1]Keio University, Japan
[2]National Institute of Informatics, Japan
[3]Hiroshima University, Japan

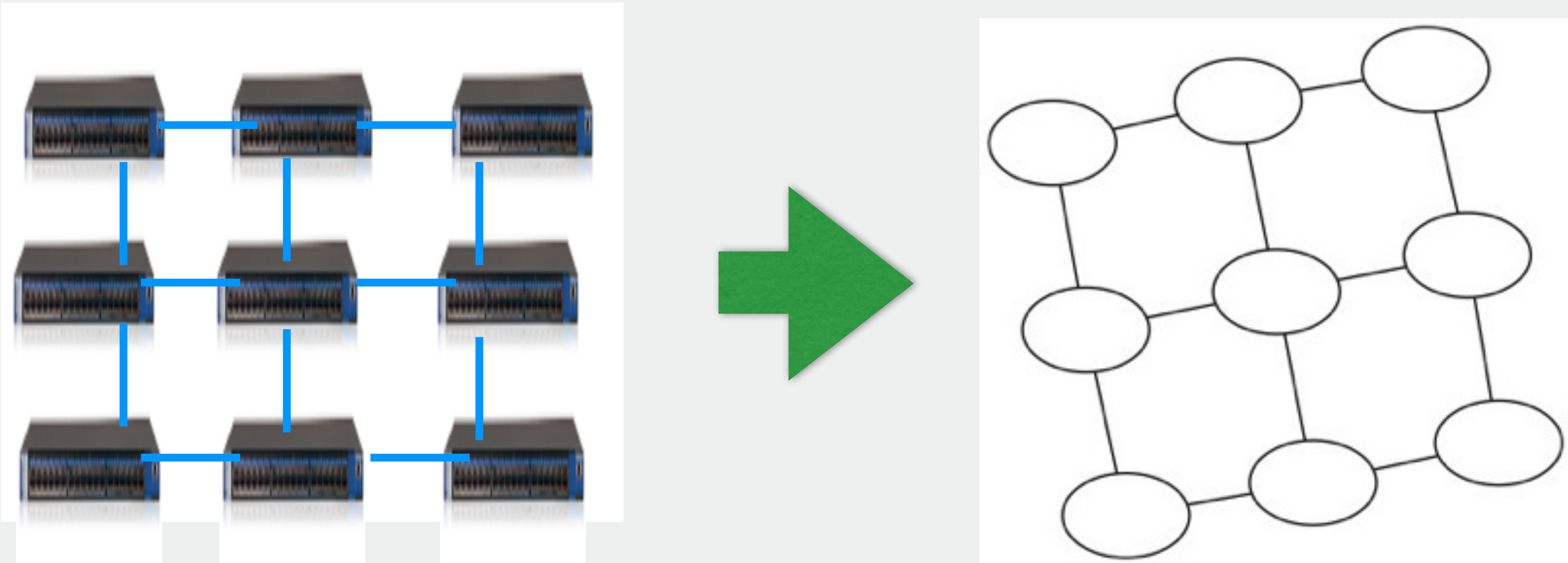
Presentation at International Conference on Parallel Processing 2017 on 16 August, 2017 @ Bristol, United Kingdom

Graph is everywhere



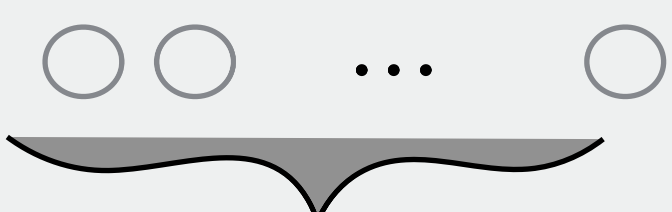


An interconnection network is also a graph

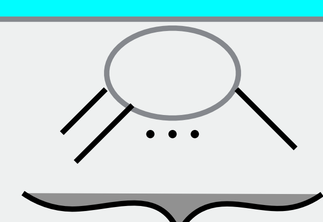


Important topological properties for interconnection networks

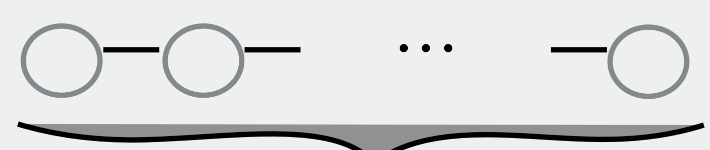
Order	Degree	Diameter



How many nodes?



How many links per node?

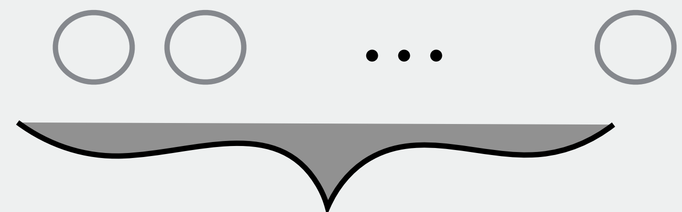


Shortest path length between farthest nodes

Classical problem:
The Degree/Diameter Problem

Optimize (maximize):

Order

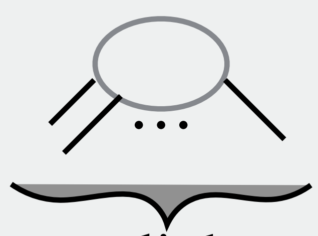


How many nodes?

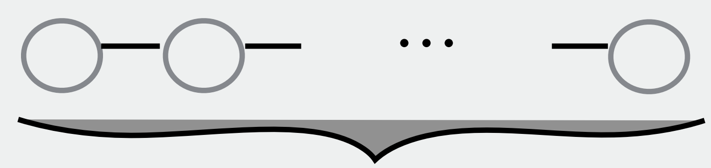
Subject to:

Degree

Diameter



How many links per node?



Shortest path length
between farthest nodes

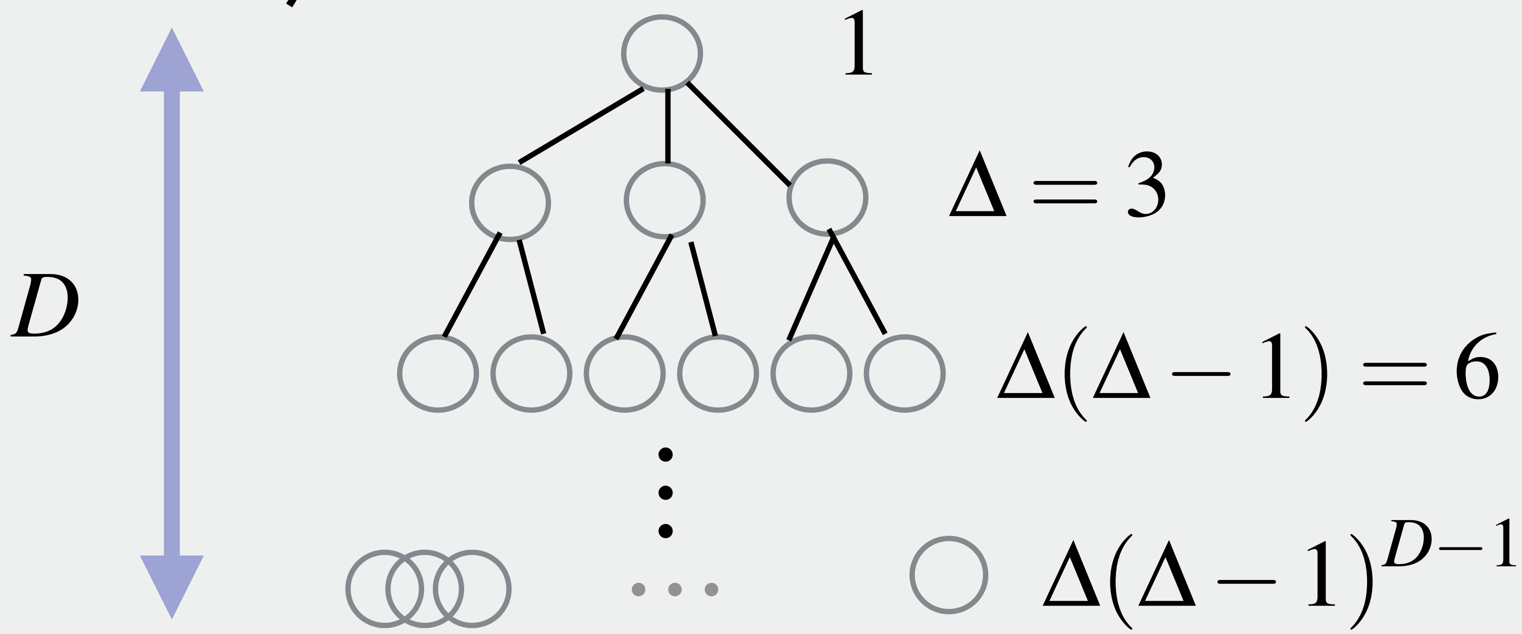
Summarised in Combinatorics Wiki

http://combinatoricswiki.org/wiki/The_Degree/Diameter_Problem

The Moore graph (optimum graph)

Δ : Degree, D : Diameter

ex) $\Delta = 3$



Edward F. Moore (1925-2003)

Upper bound on the order (called the *Moore bound*):

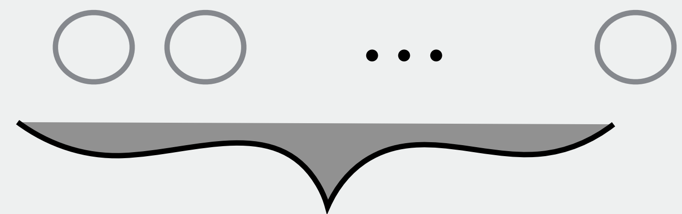
$$\boxed{1 + \Delta \sum_{i=0}^{D-1} (\Delta - 1)^i}$$

Shortcoming of the Degree/Diameter Problem

Optimize (maximize):

Practically,
the order should be fixed

Order

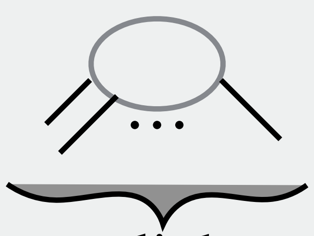


How many nodes?

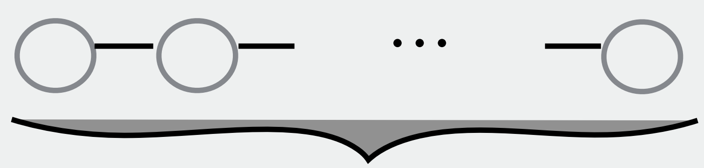
Subject to:

Degree

Diameter



How many links per node?



Shortest path length
between farthest nodes

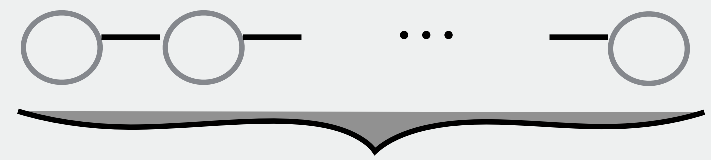
Summarised in Combinatorics Wiki

http://combinatoricswiki.org/wiki/The_Degree/Diameter_Problem

The Order/Degree Problem (ODP)

Optimize (minimize):

Diameter

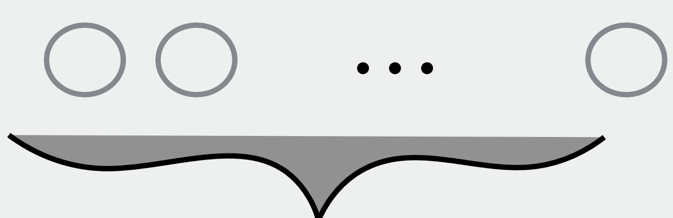


Shortest path length
between farthest nodes

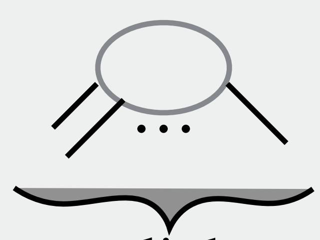
<u>Practically,</u>
<u>the order should be fixed</u>

Subject to:

Order | **Degree**



How many nodes?



How many links per node?

Graph Golf:
ODP competition

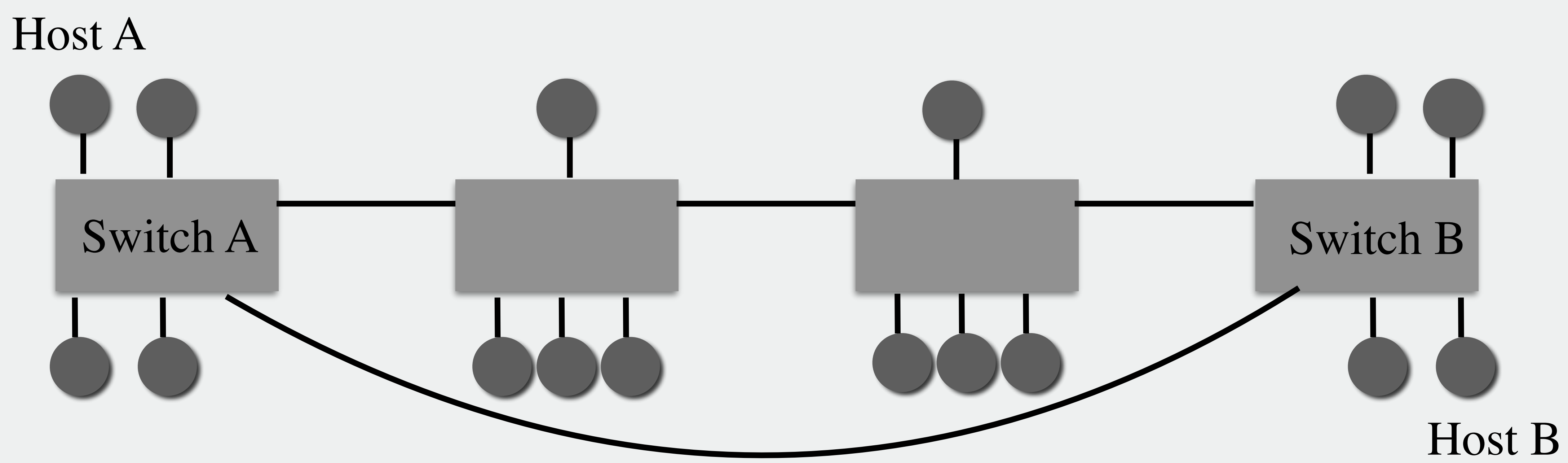
<u>http://research.nii.ac.jp/graphgolf/</u>

Mapping???

- Node ⇔ Switch?
  - But <u># of switches</u> are NOT essential
- Ordinary graph ignores <u># of **hosts**</u>! 
  - <u># of **hosts**</u> should be fixed

Network consists of switches and hosts

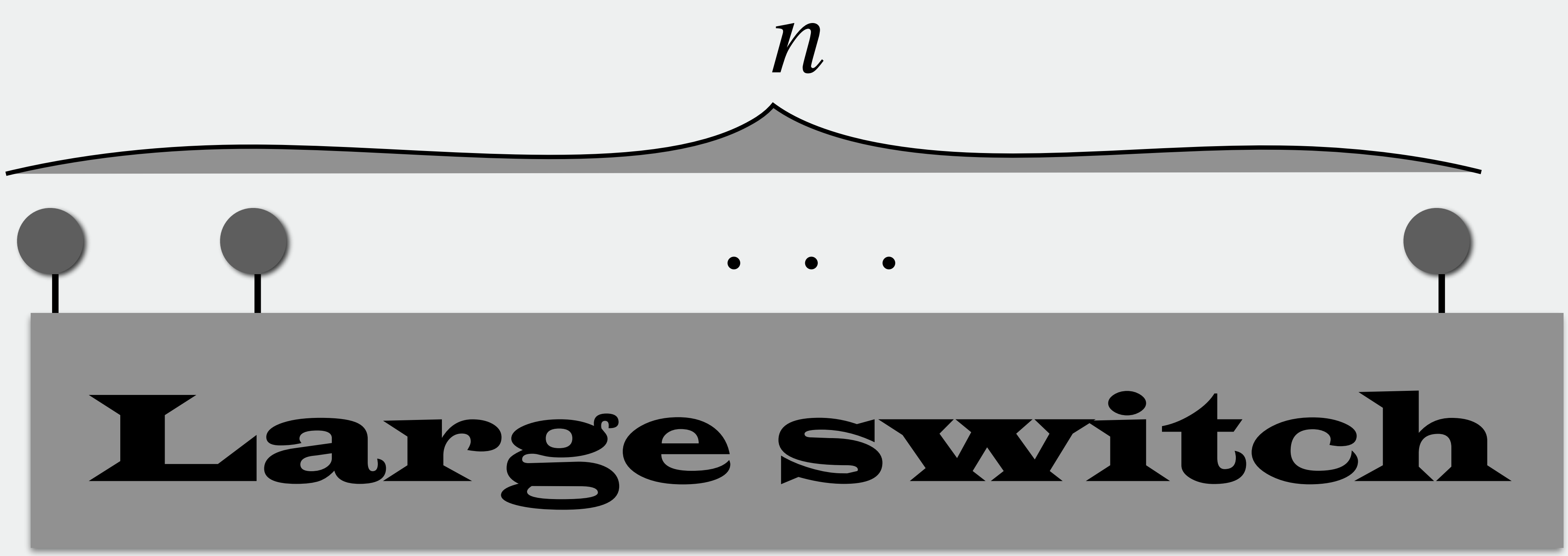
A host-switch graph



Our Goal:

To minimise *host-to-host average shortest path length (h-ASPL)*

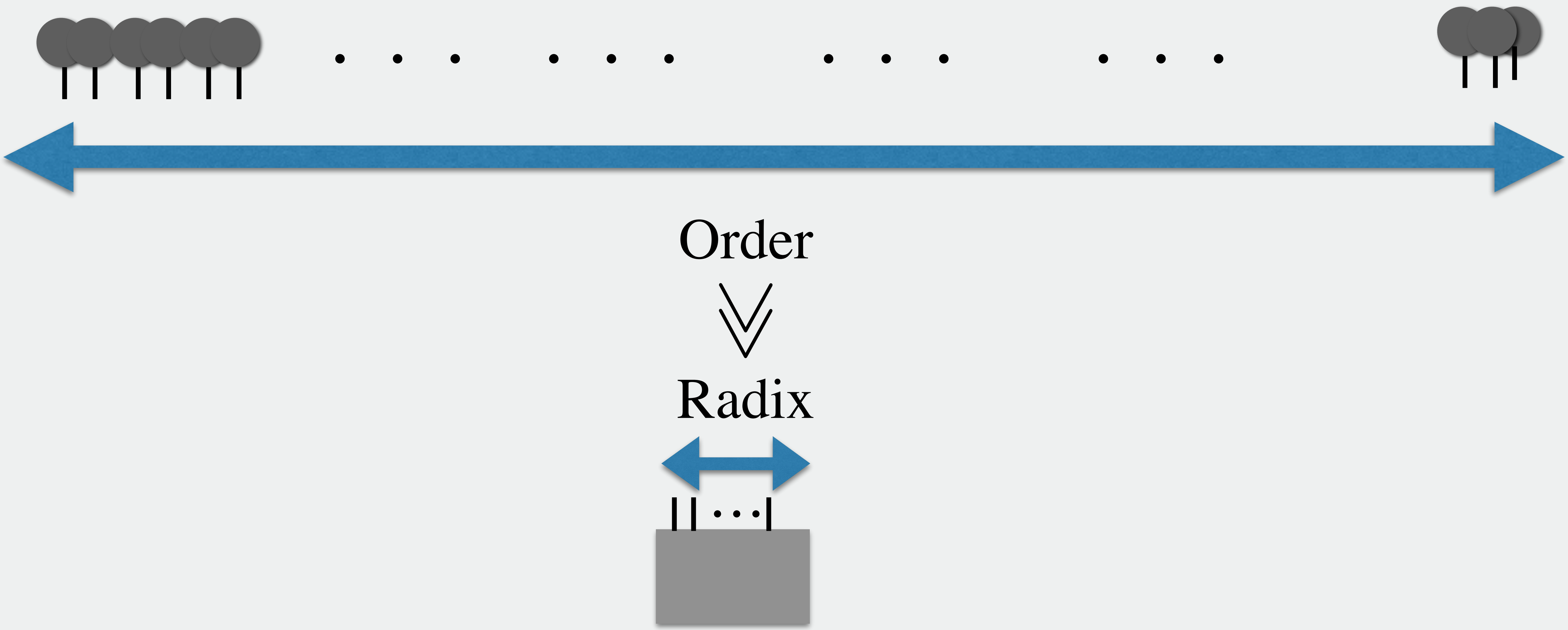
Let's connect *n* hosts



In practical, however, *radix* **(# of ports of a switch) is limited**

In practical situations, Order $\gg$ Radix

Order rapidly increases as technology advances

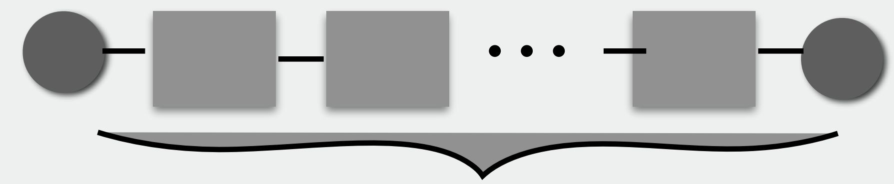


Designing high-radix switch requires high cost,
so radix is limited

The *Order*/*Radix* *Problem* (*ORP*)

Optimize (minimize):

h-ASPL

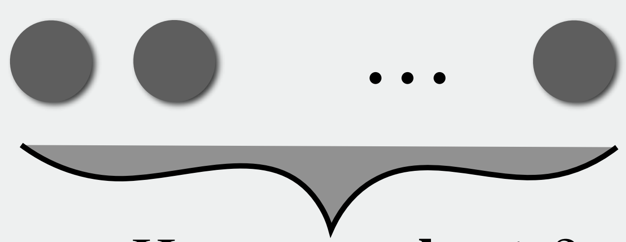


host-to-host
average shortest path length

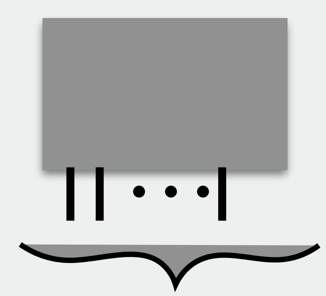
Subject to:

Order

Radix



How many **hosts**?

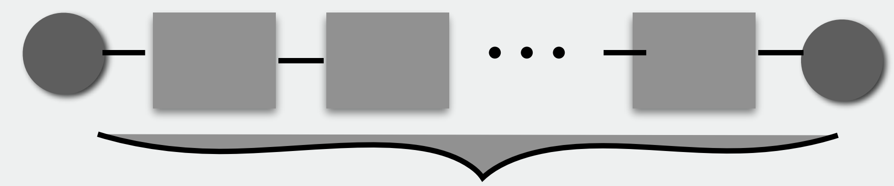


How many links **per switch**?

The *Order/Radix Problem* (ORP)

Optimize (minimize):

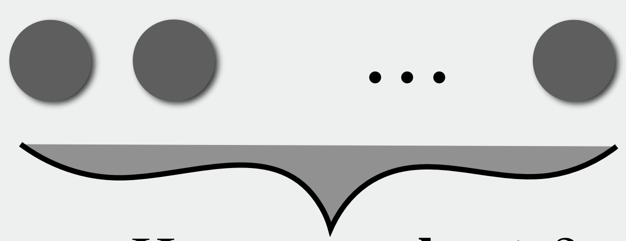
h-ASPL

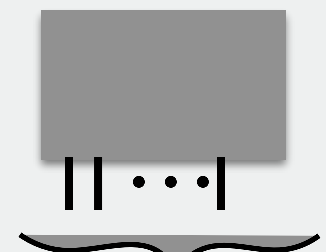


Subject to:

Order

Radix



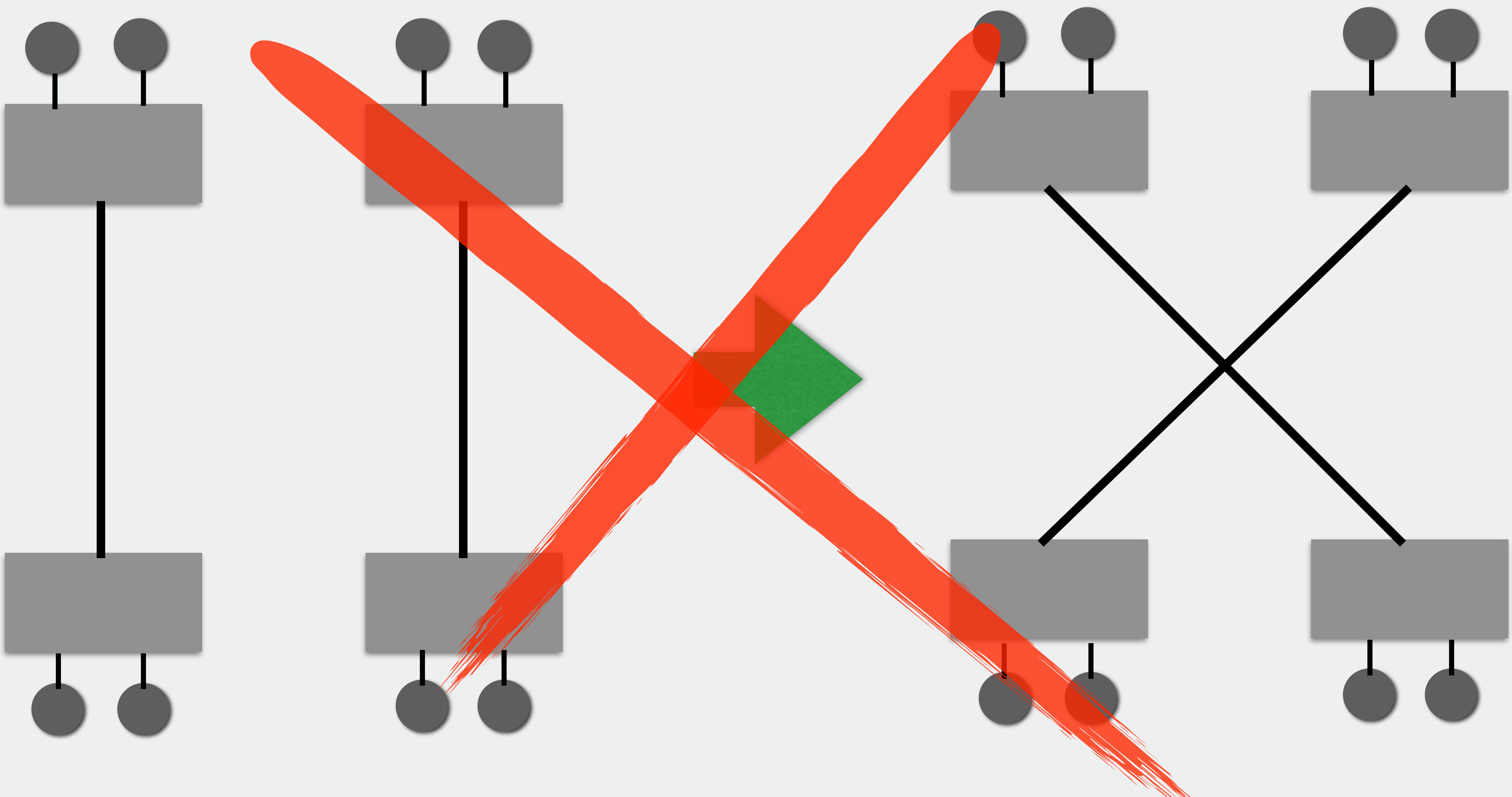


Important questions:

Q1. ***How many switches should be used?***

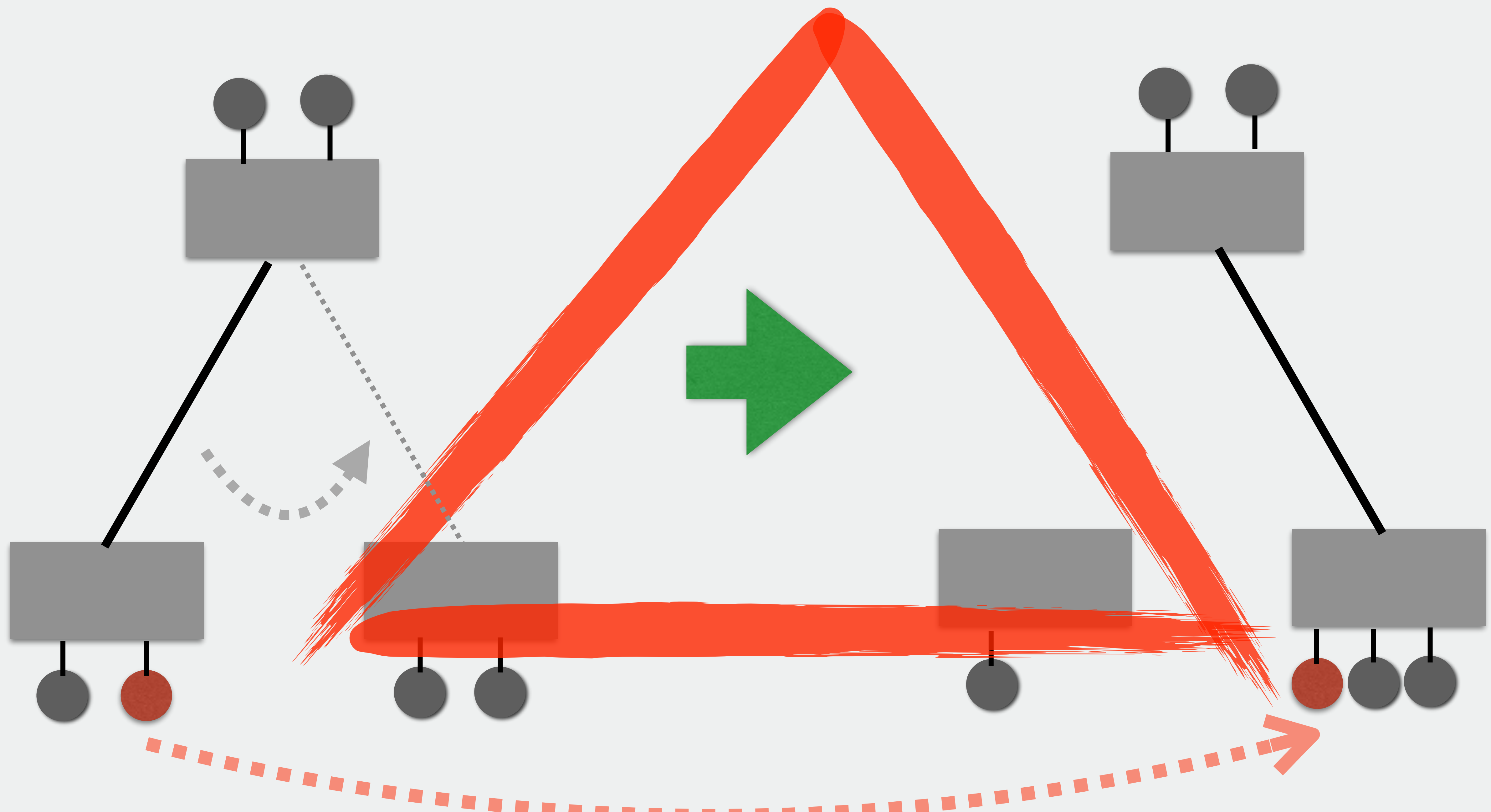
Q2. ***Should hosts be connected uniformly, or non-uniformly?***

Existing technique for ODP: 2-opt



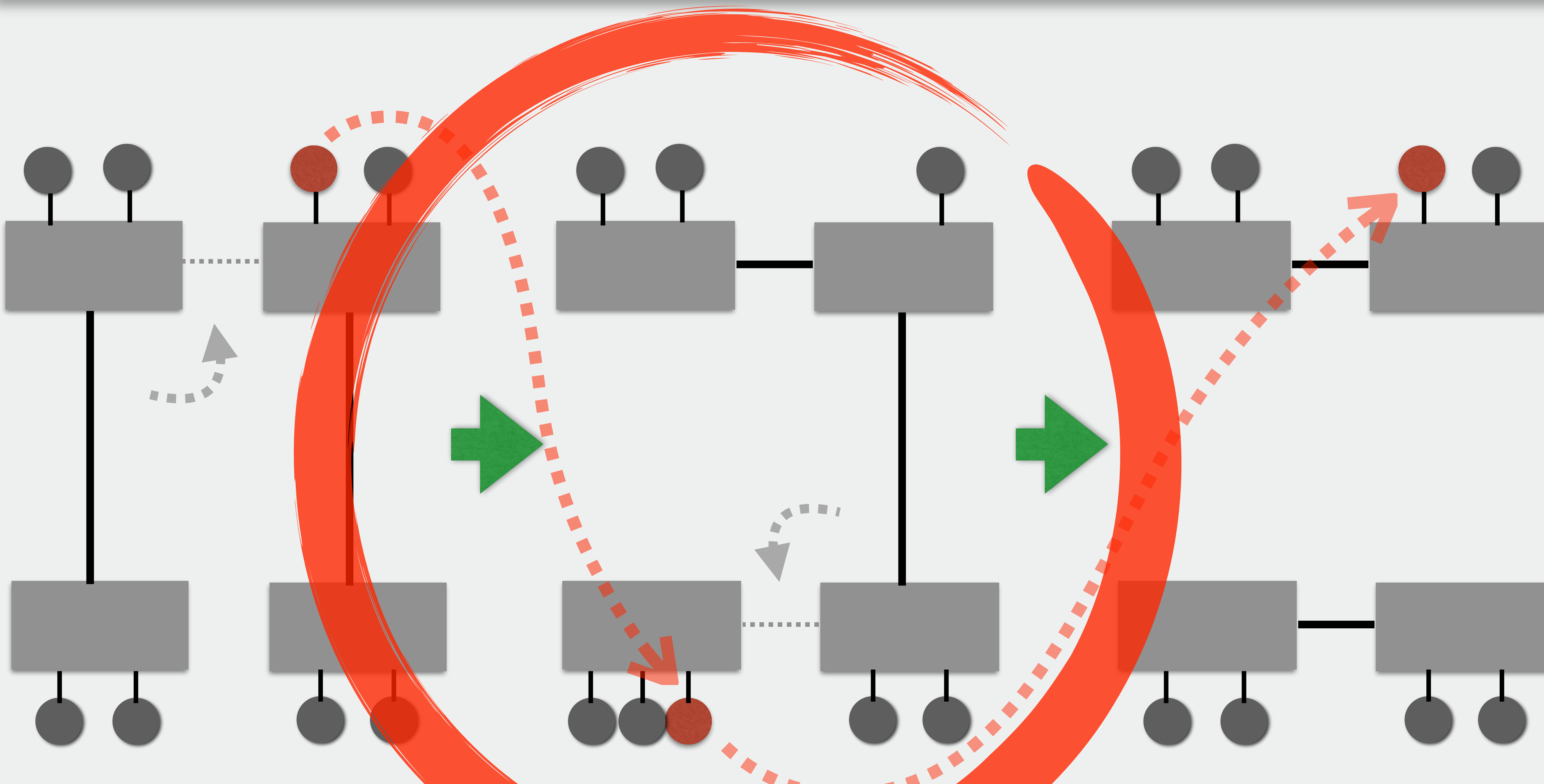
- # of hosts connected to each switch **never** changes!

Swing operation



- # of hosts connected to each switch **always** changes!

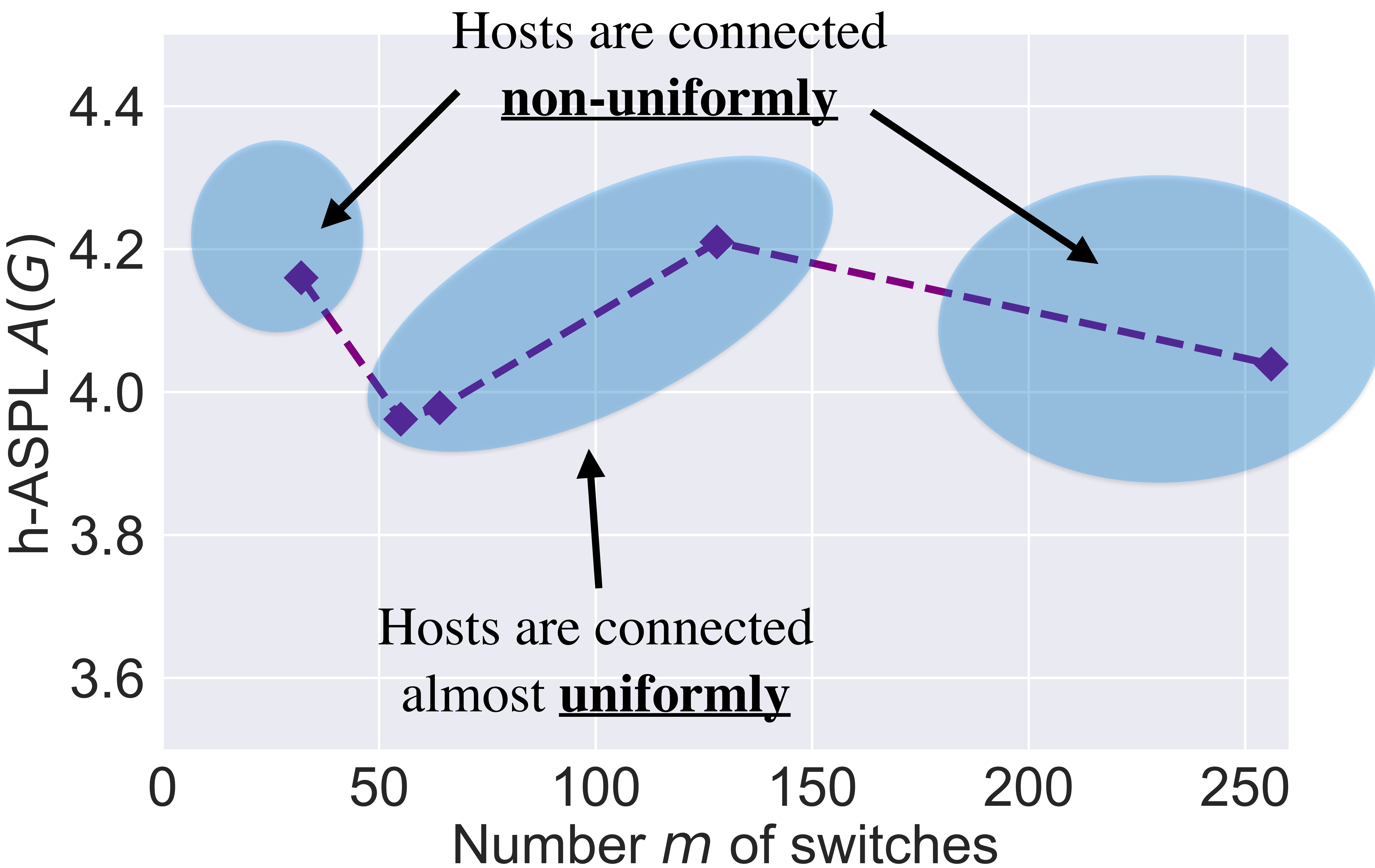
2-neighbour swing operation



1-neighbour solution ≡ swing solution

2-neighbour solution ≡ 2-opt solution

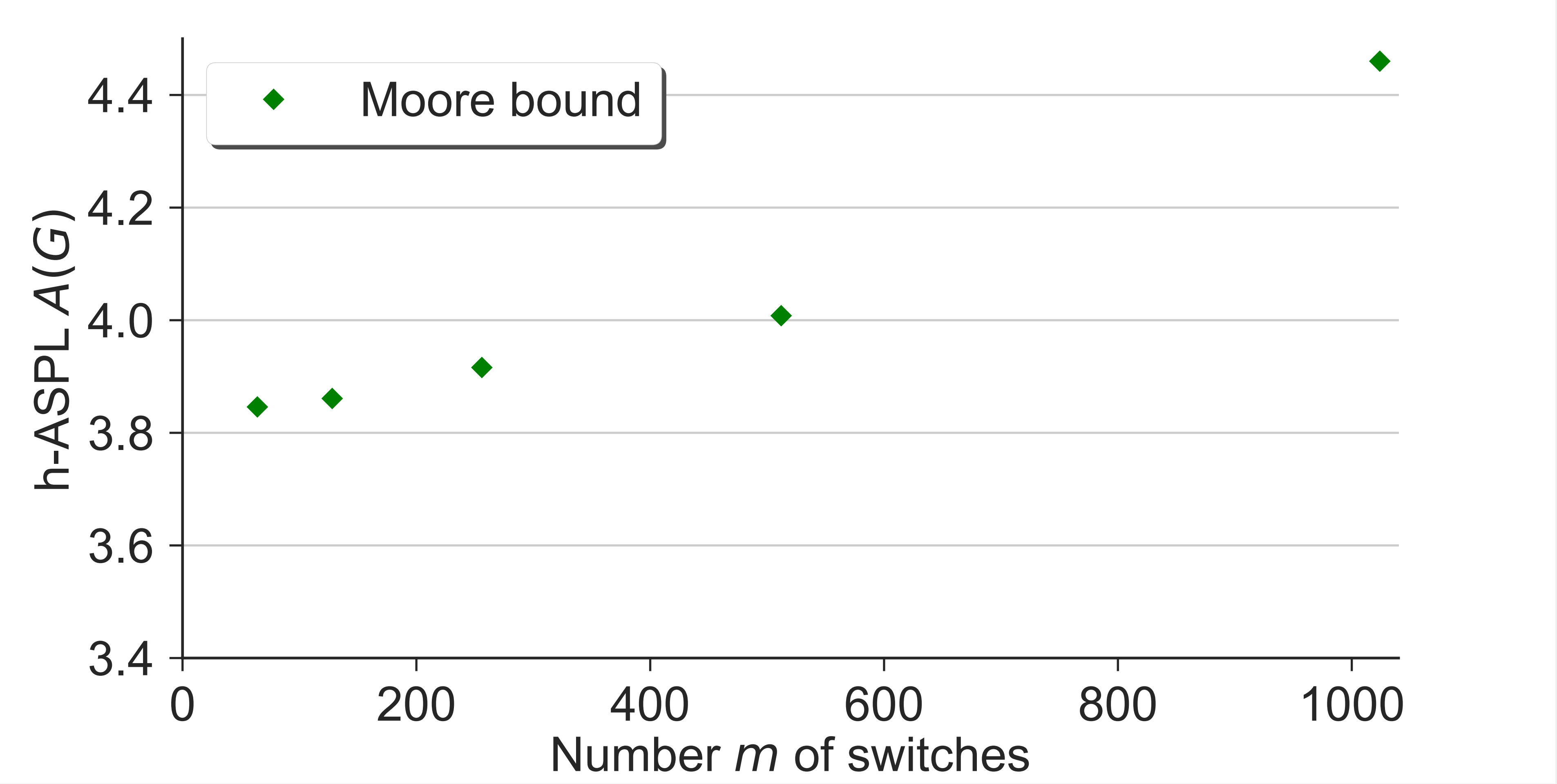
Relationship between h-ASPL and # of switches



Again, let's consider the Moore graph

- Lower bound on the h-ASPL can be calculated by the Moore graph consisting of only switches <u>if we assume each switch has fixed number of hosts.</u>

Edward F. Moore (1925-2003)

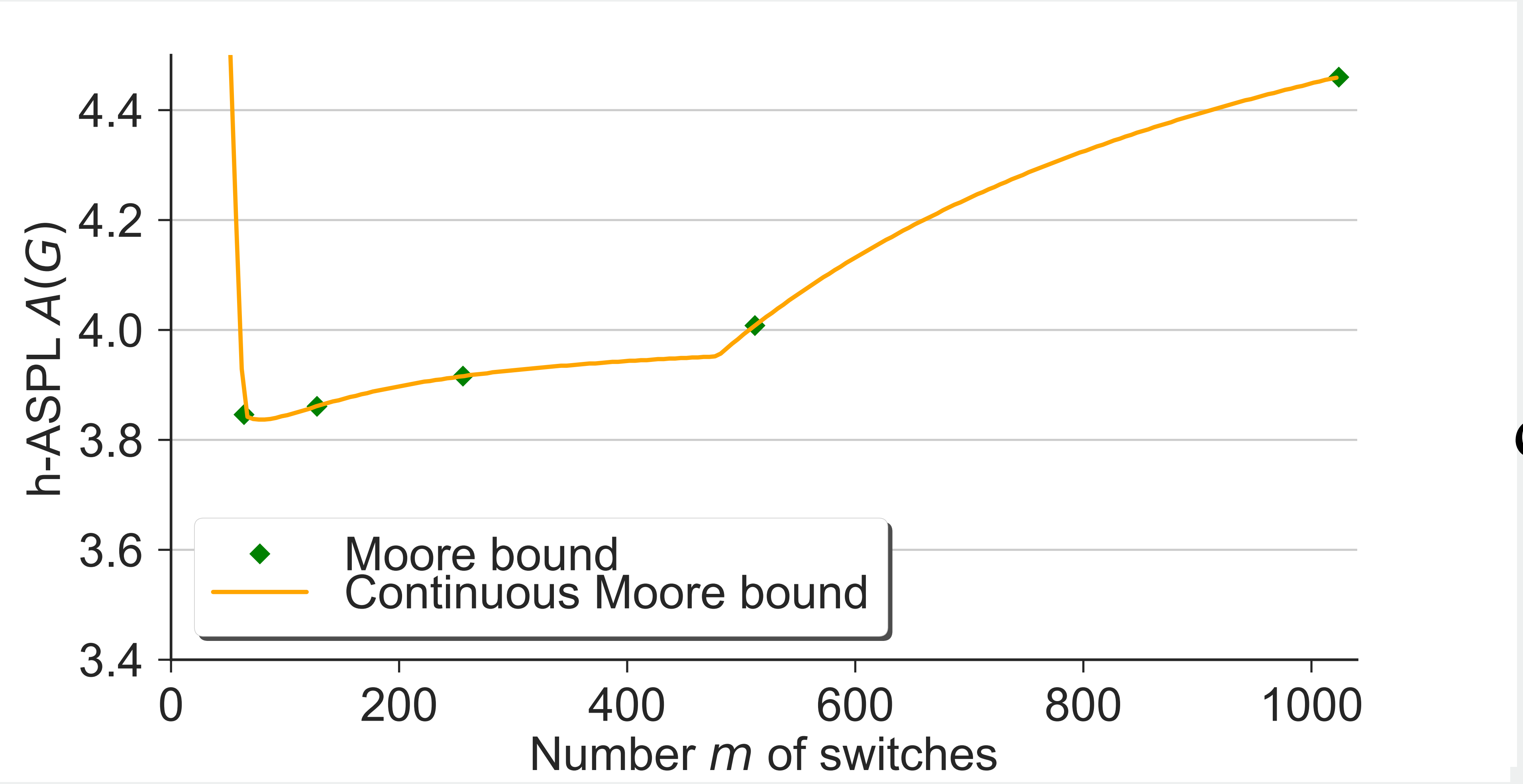


# of hosts must be natural number

The *continuous Moore bound*

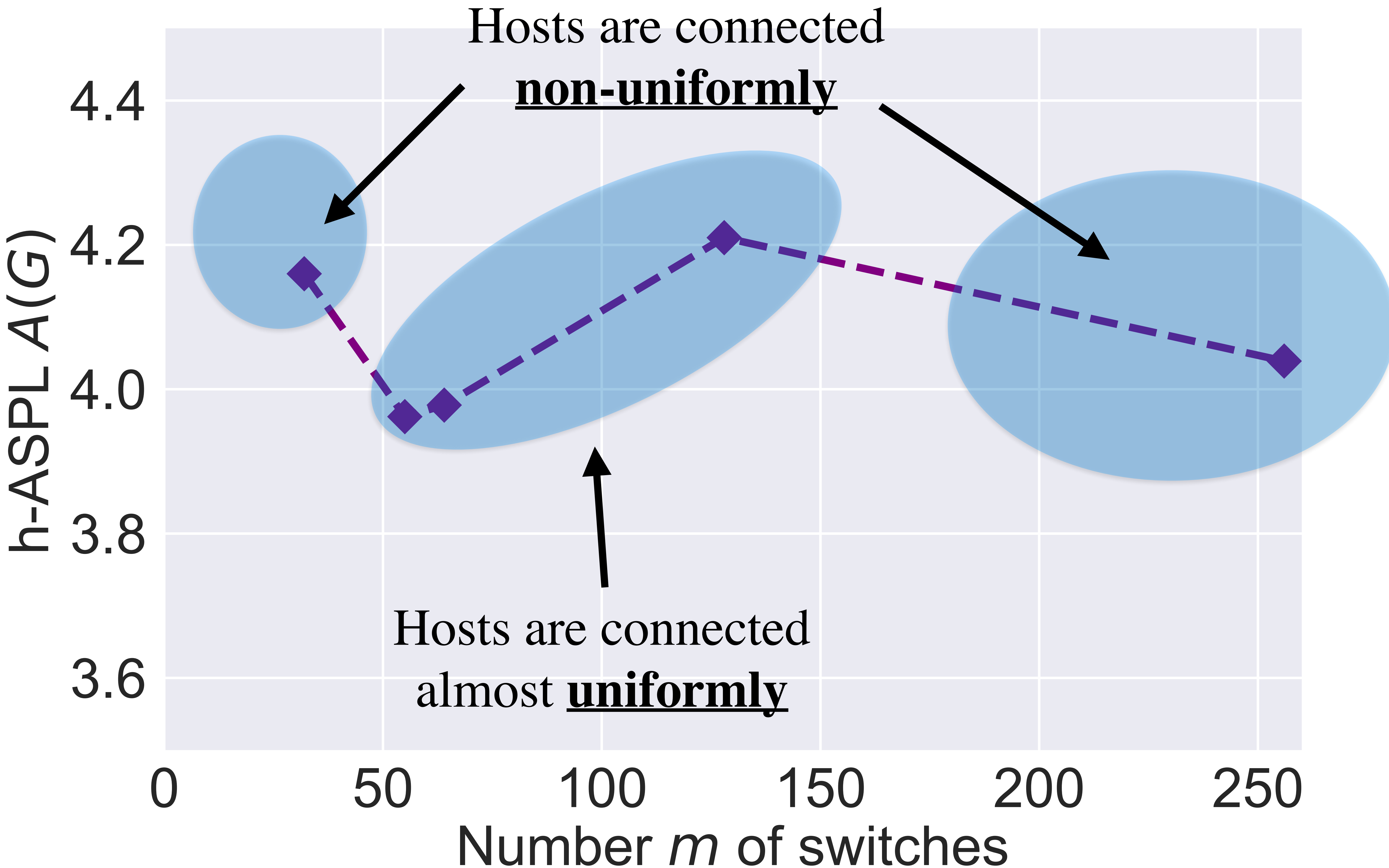
- Lower bound on the h-ASPL can be calculated by the Moore graph consisting of only switches if we assume each switch has fixed number of hosts.

Edward F. Moore (1925-2003)

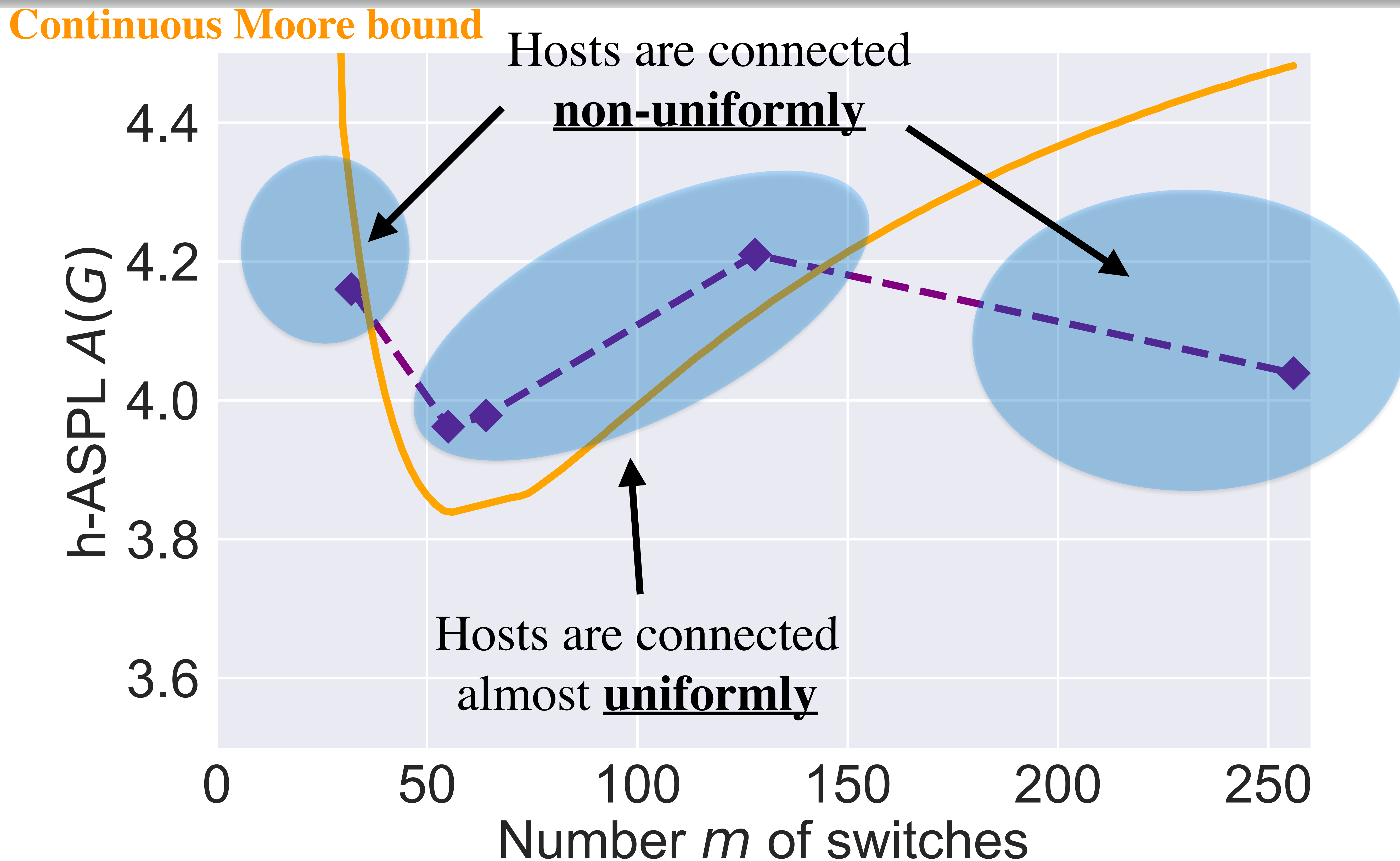


\# of hosts does NOT need to be natural number

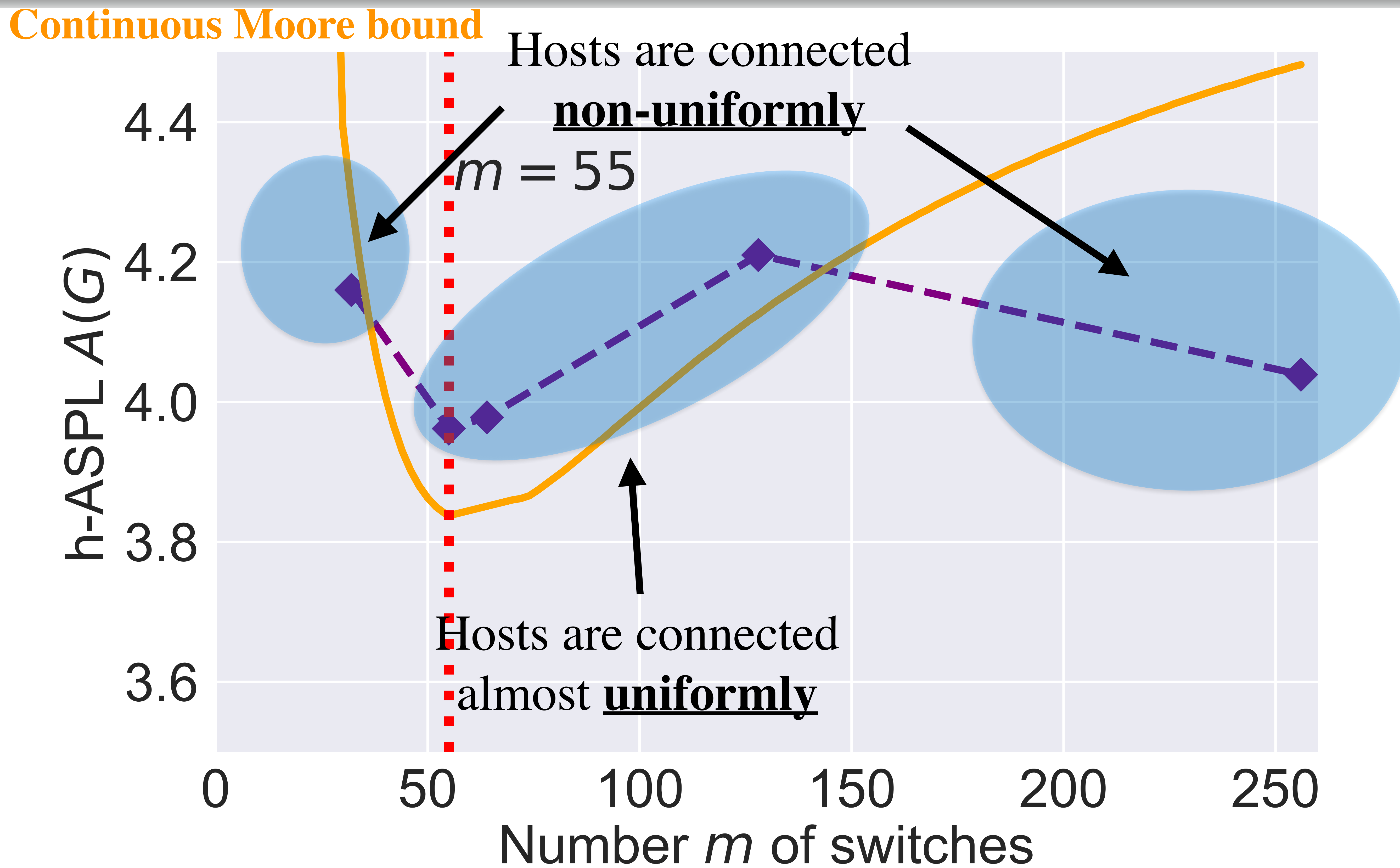
Relationship between h-ASPL and # of switches



Relationship between h-ASPL and # of switches



Relationship between h-ASPL and # of switches



Answers to the questions

Important questions:

Q1. ***How many switches should be used?***

Q2. ***Should hosts be connected uniformly, or NON-uniformly?***

Empirical answers:

A1. The number such that the continuous Moore bound becomes minimum.

A2. Hosts should be connected uniformly.

Comparison with existing topologies

- The torus, the dragonfly, and the fat-tree
- Picked up from interconnection networks used in supercomputers ranked in TOP500

TOP 10 Sites for June 2017

For more information about the sites and systems in the list, click on the links or view the complete list.

1-100 | 101-200 | 201-300 | 301-400 | 401-500

Rank	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway , NRCPC National Supercomputing Center in Wuxi China	10,649,600	93,014.6	125,435.9	15,371
2	**Tianhe-2 (MilkyWay-2)** - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31S1P , NUDT National Super Computer Center in Guangzhou China	3,120,000	33,862.7	54,902.4	17,808
3	**Piz Daint** - Cray XC50, Xeon E5-2690v3 12C 2.6GHz, Aries interconnect , NVIDIA Tesla P100 , Cray Inc. Swiss National Supercomputing Centre (CSCS) Switzerland	361,760	19,590.0	25,326.3	2,272

https://www.top500.org/lists/2017/06/

Overview of comparison

- Performance, Power consumption, Cost breakdowns (including switch and cable costs)
- We construct a topology by as optimised host-switch graph with the same order and radix for each existing topology.
- Based on two experiments

Experiment 1: SimGrid simulation

- SimGrid discrete event simulator
  - SMPI simulates unmodified MPI applications
  - NAS parallel benchmark
- Networks with 1024 hosts
  - 5-ary 3-torus
  - Dragonfly with diameter 5
  - 16-ary fat-tree

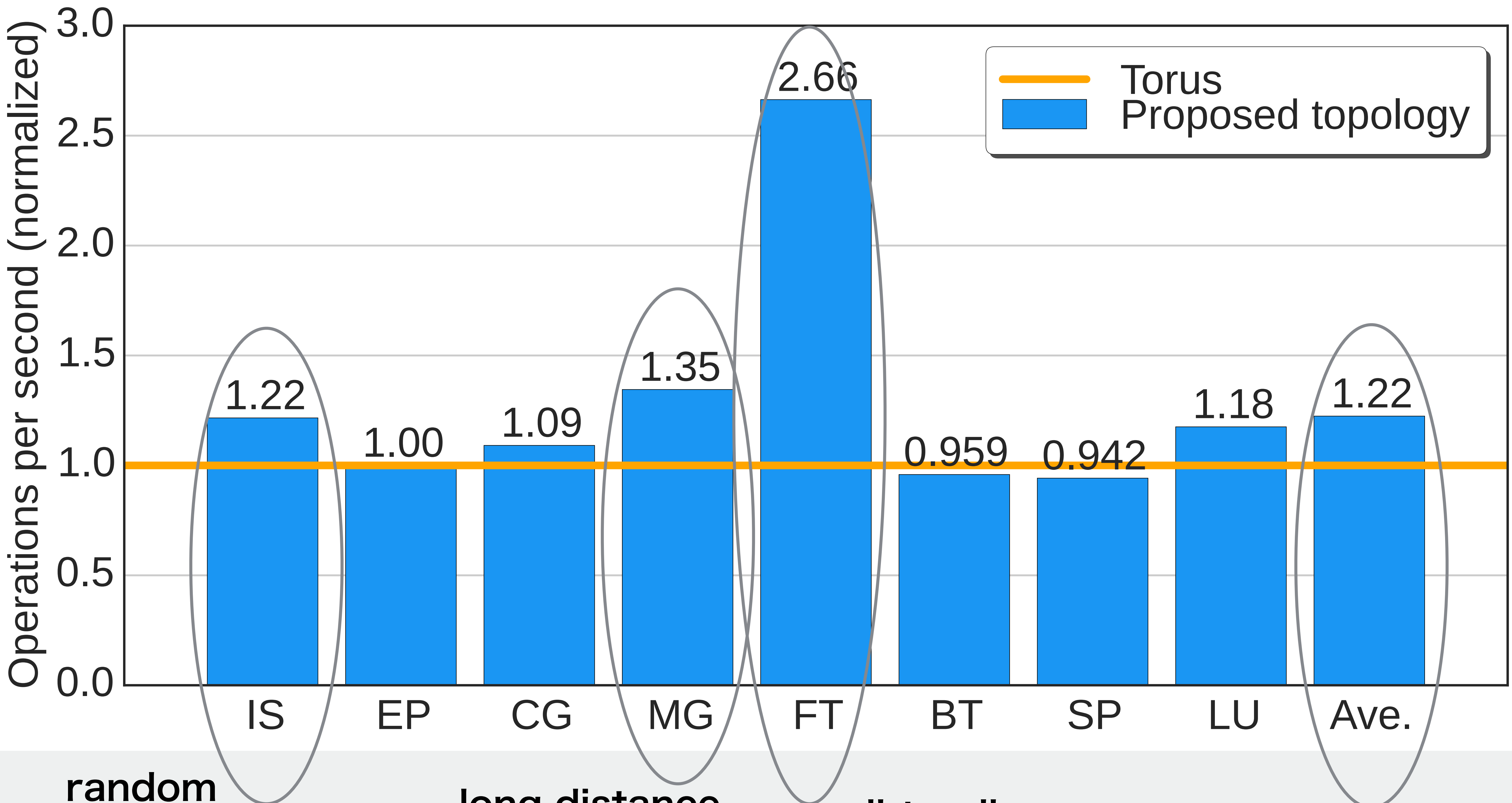
Experiment 2: Modelling

- Models of Mellanox InfiniBand switches/cables.
    - As with [Besta and Hoefler, 2014]

[Besta and Hoefler 2014] "**Slim fly: A cost effective low-diameter network topology**," SC, Nov. 2014, pp. 348–359.

- Based on 60cm x 210 cm floorplan

Performance comparison with Torus



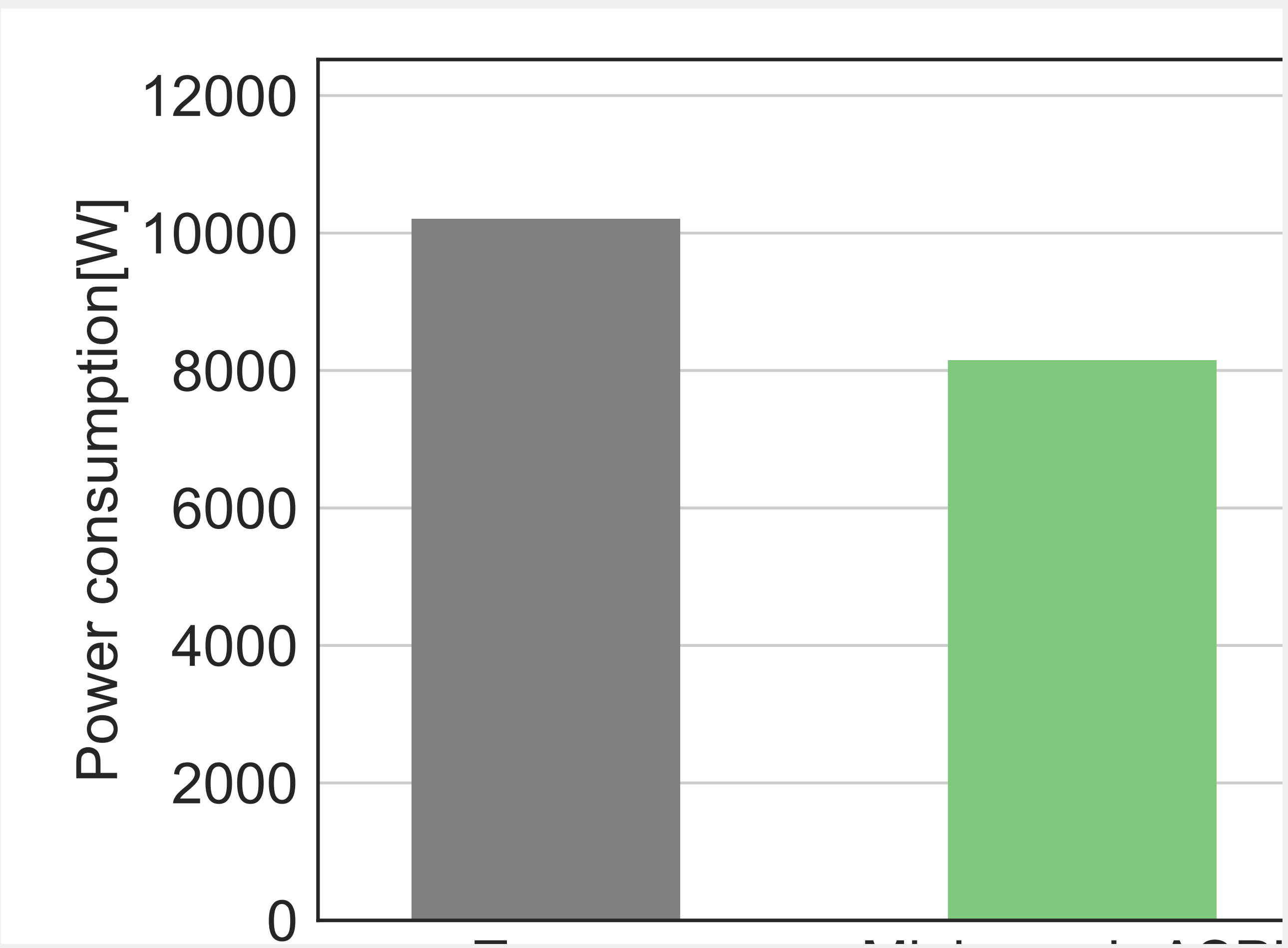
random
memory accesses

long distance
communications

all-to-all

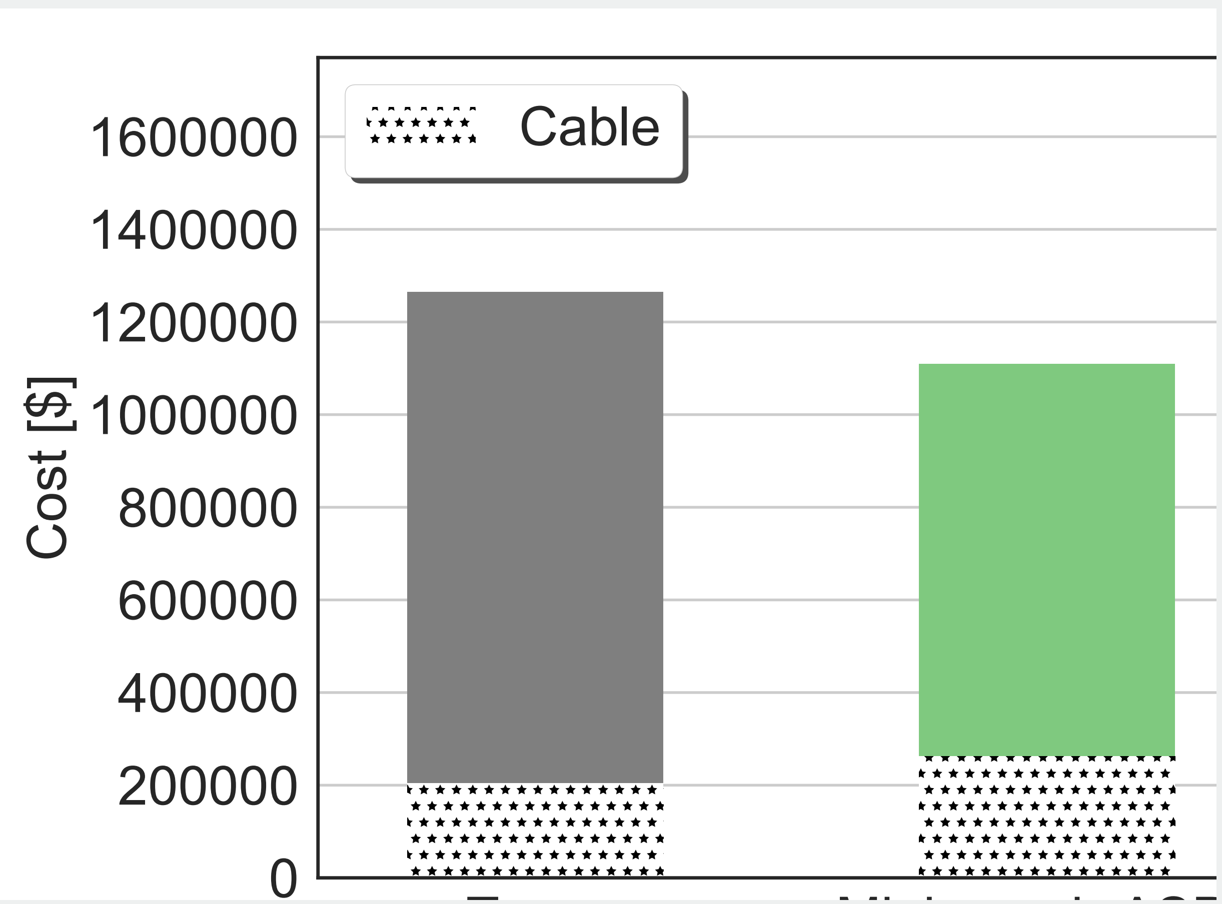
Power/costs comparison with Torus

Power consumption



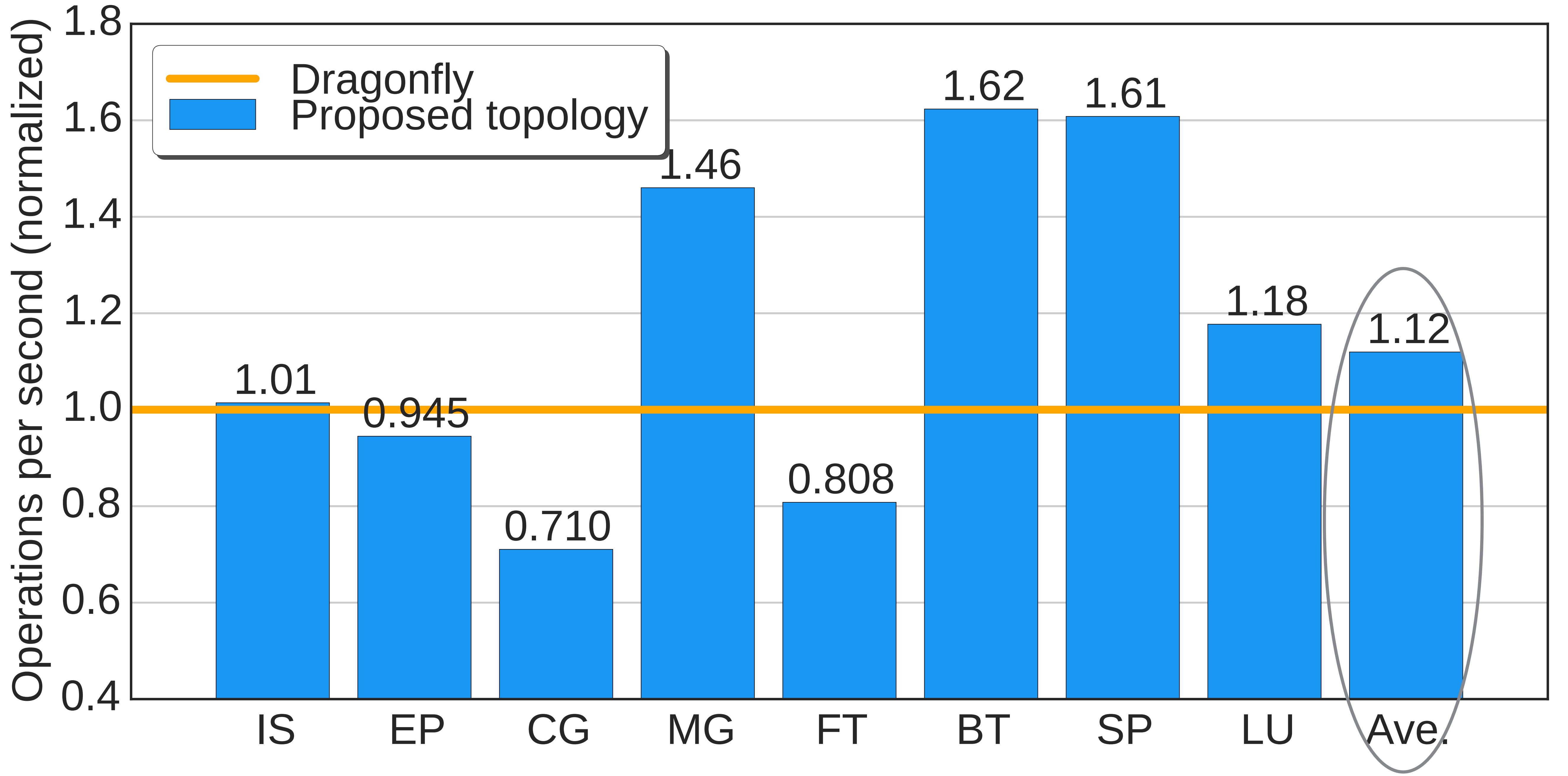
Torus Proposed

Cost breakdowns



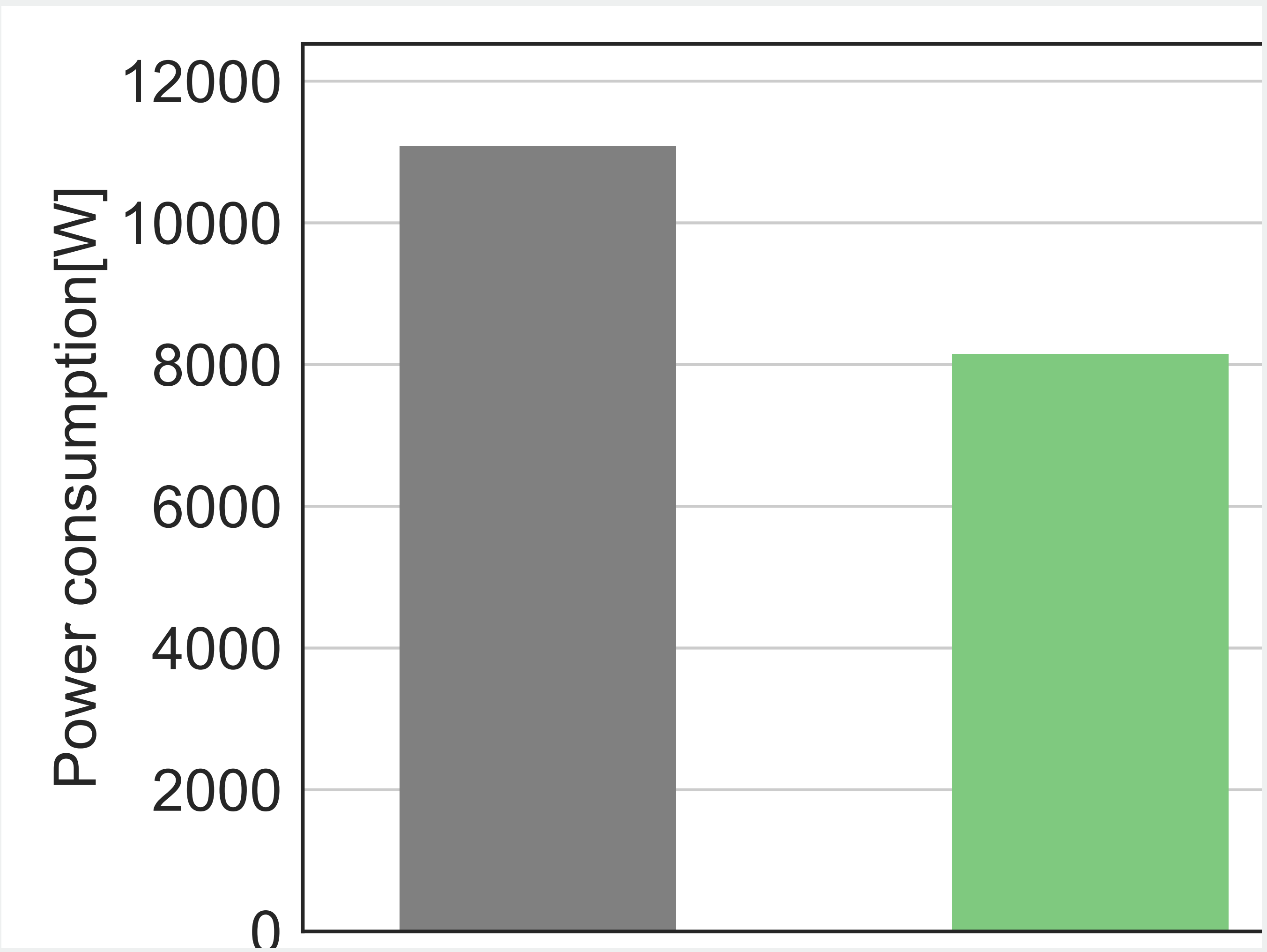
Torus Proposed

Performance comparison with Dragonfly



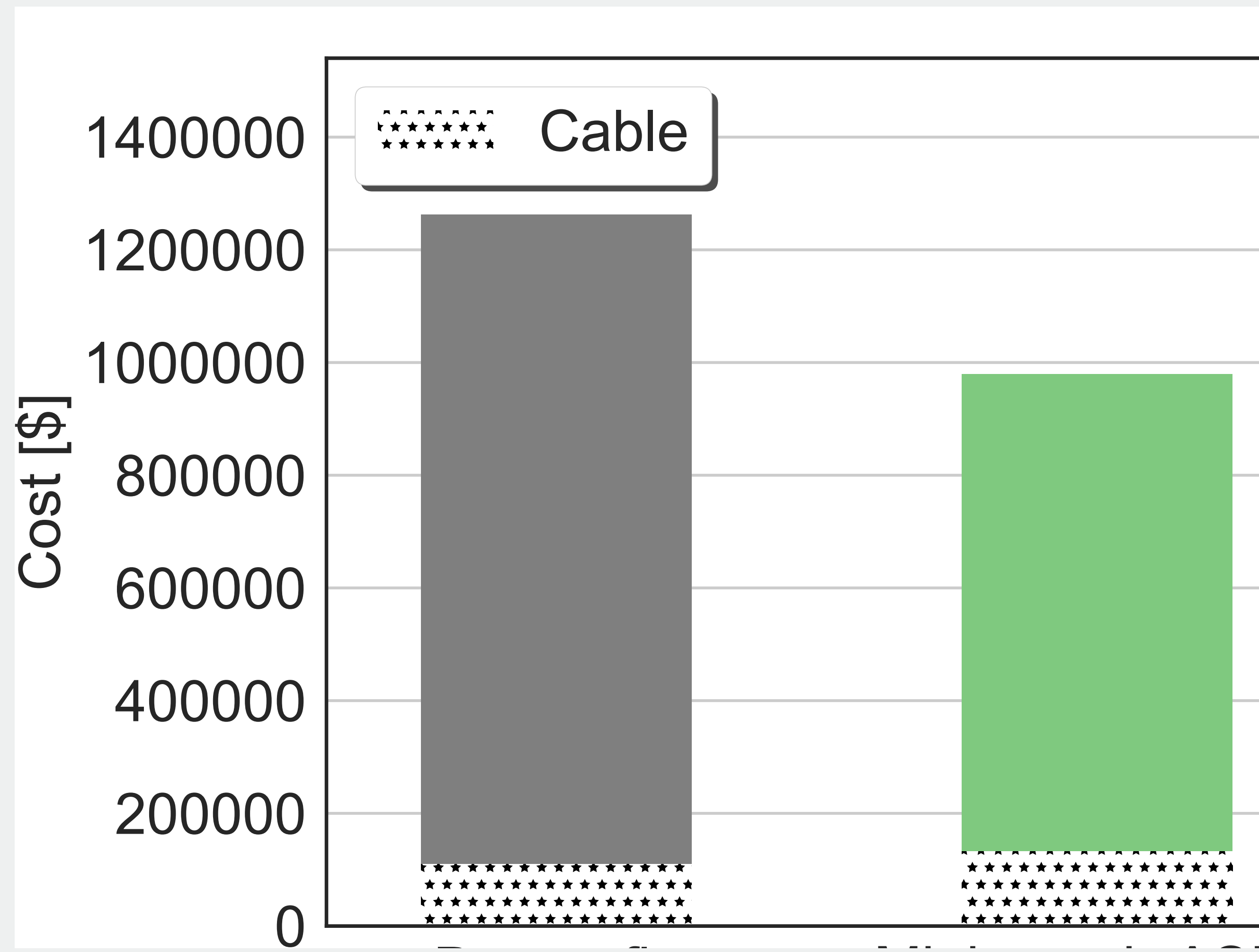
Power/costs comparison with Dragonfly

Power consumption



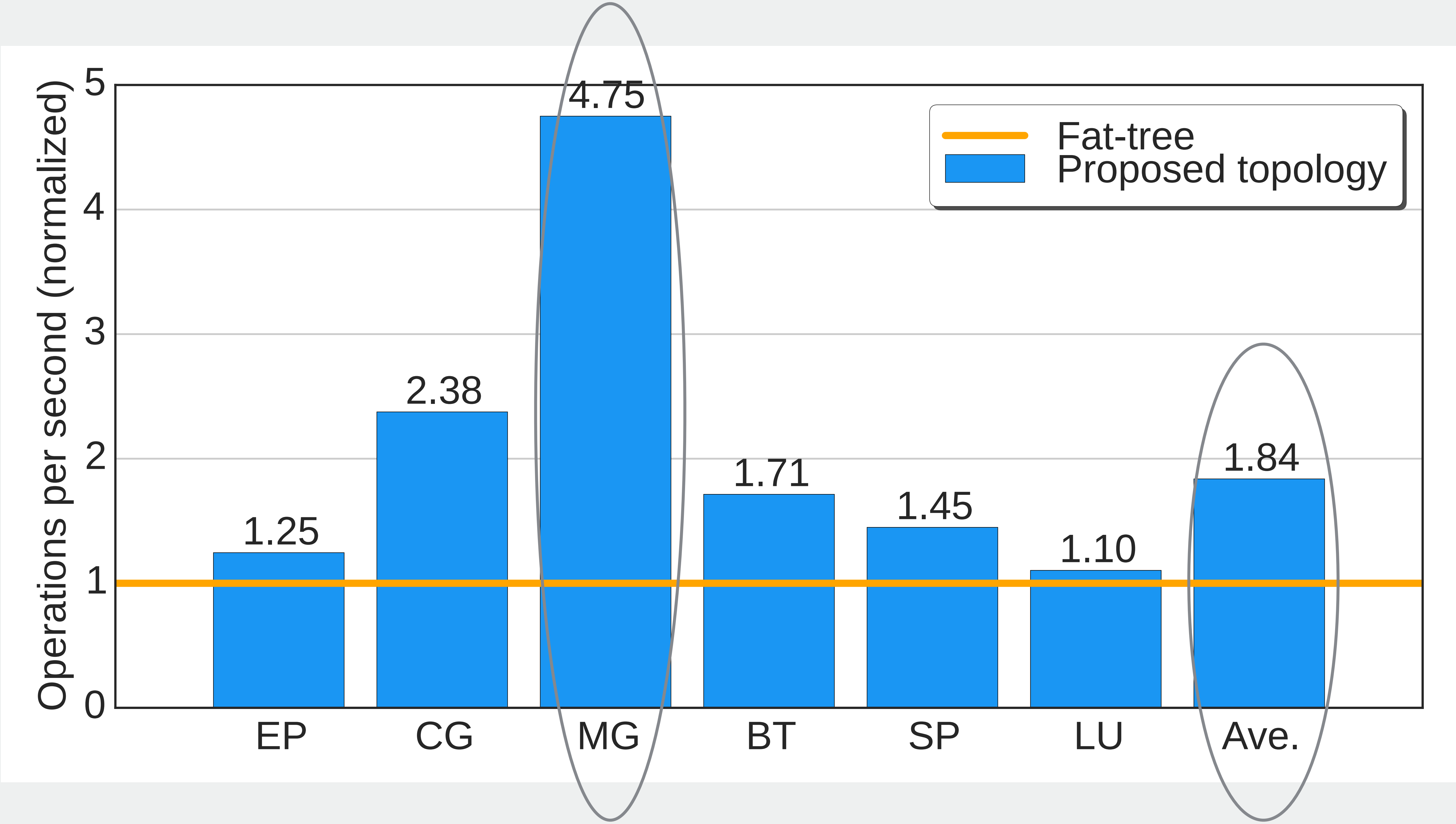
Dragonfly Proposed

Cost breakdowns



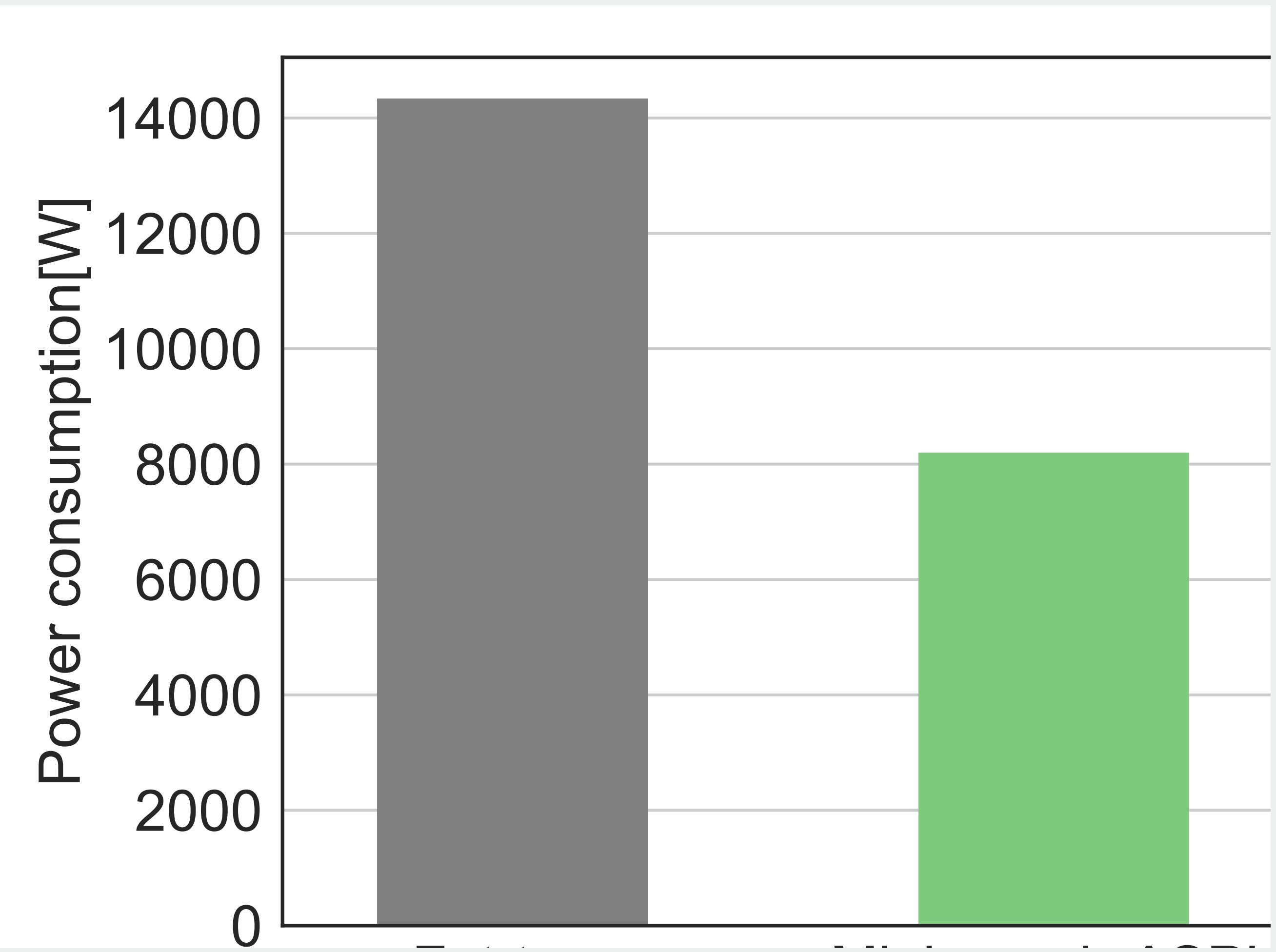
Dragonfly Proposed

Performance comparison with Fat-tree



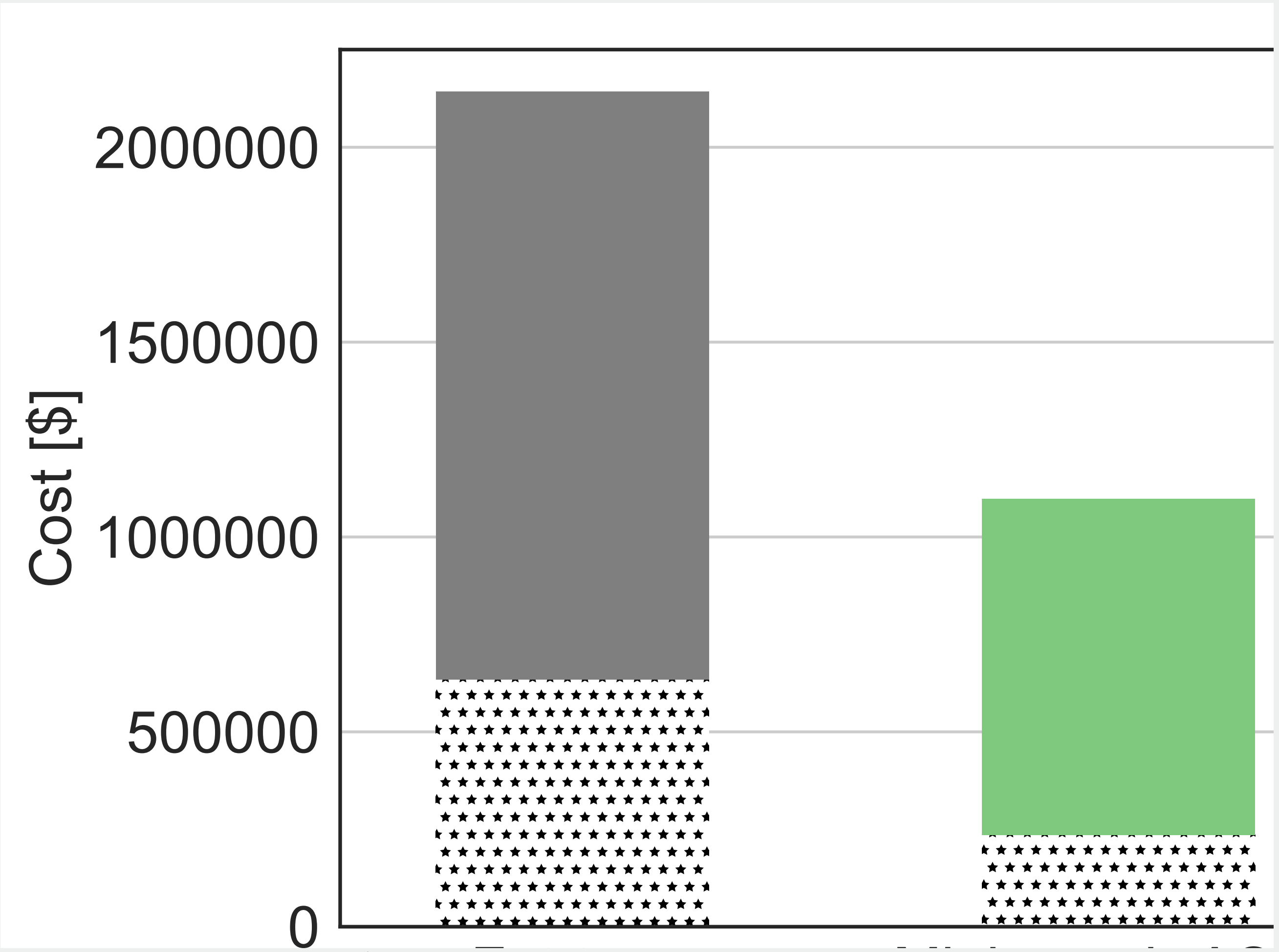
Power/costs comparison with Fat-tree

Power consumption



Fat-tree Proposed

Cost breakdowns



Fat-tree Proposed

Conclusions

- A host-switch graph
- The order/radix problem
- Our solution:
  - Reducing h-ASPL with 2-neighbour operation
  - Approximation of the optimal number of switches by using the continuous Moore bound
- Our topologies attain 12%-84% faster MPI execution with lower power/costs