Leveraging Heterogeneity to Reduce the Cost of Data Center Upgrades

Andy Curtis

joint work with:
S. Keshav
Alejandro López-Ortiz
Tommy Carpenter
Mustafa Elsheikh

University of Waterloo


Data centers change over time

Data centers constantly evolve

- 63% of Data Center Knowledge readers are either in the midst of data center expansion projects or have just completed a new facility
- 59% continue to build and manage their data centers in-house

http://www.datacenterknowledge.com/archives/2010/08/16/data-center-industry-expansion-in-full-swing/

Network upgrade motivation

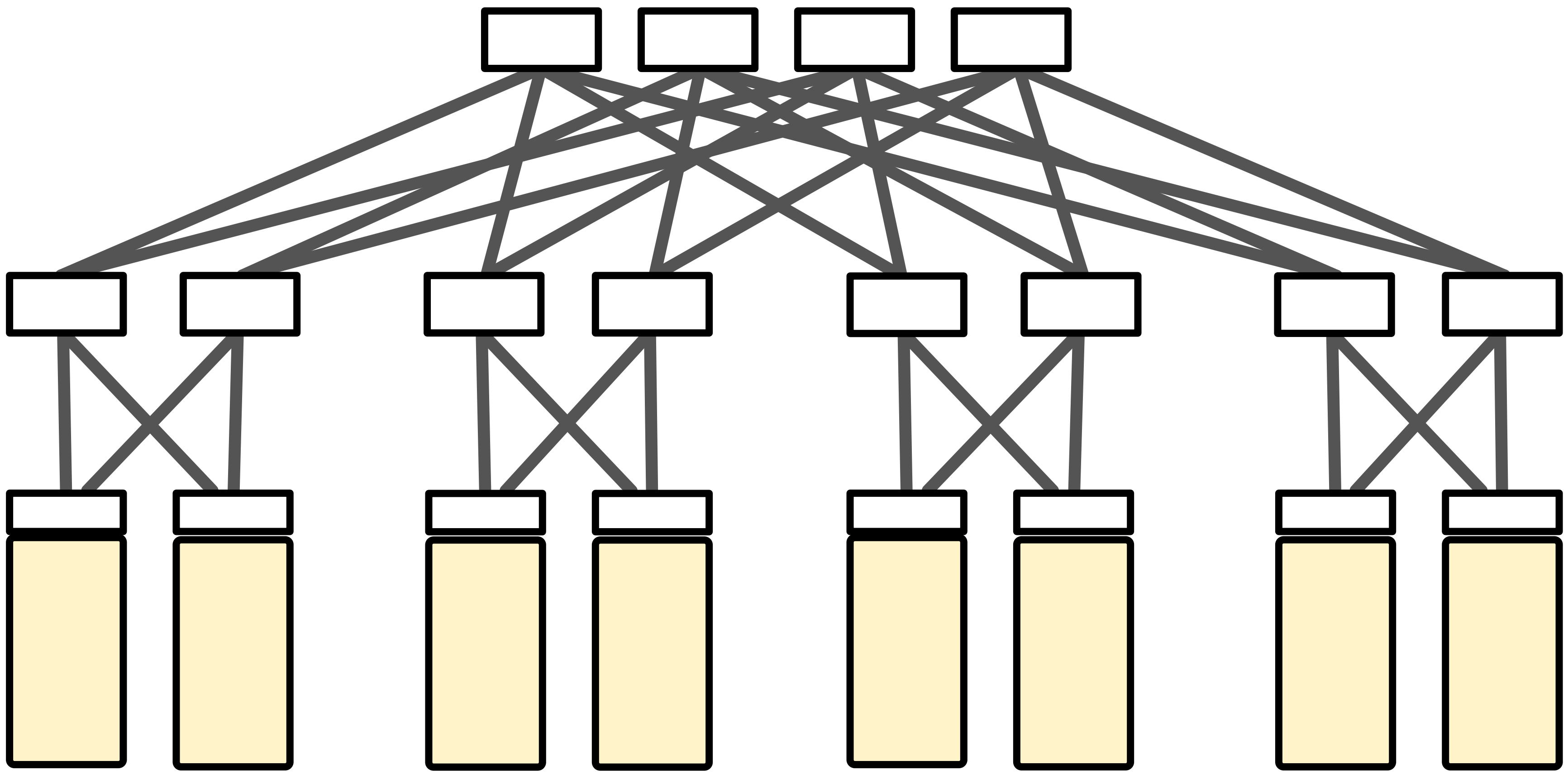
Network upgrade motivation

- Several prior solutions for greenfield data centers
  - VL2, flattened butterfly, HyperX, BCube, DCell, Al-Fares *et al.*, MDCube

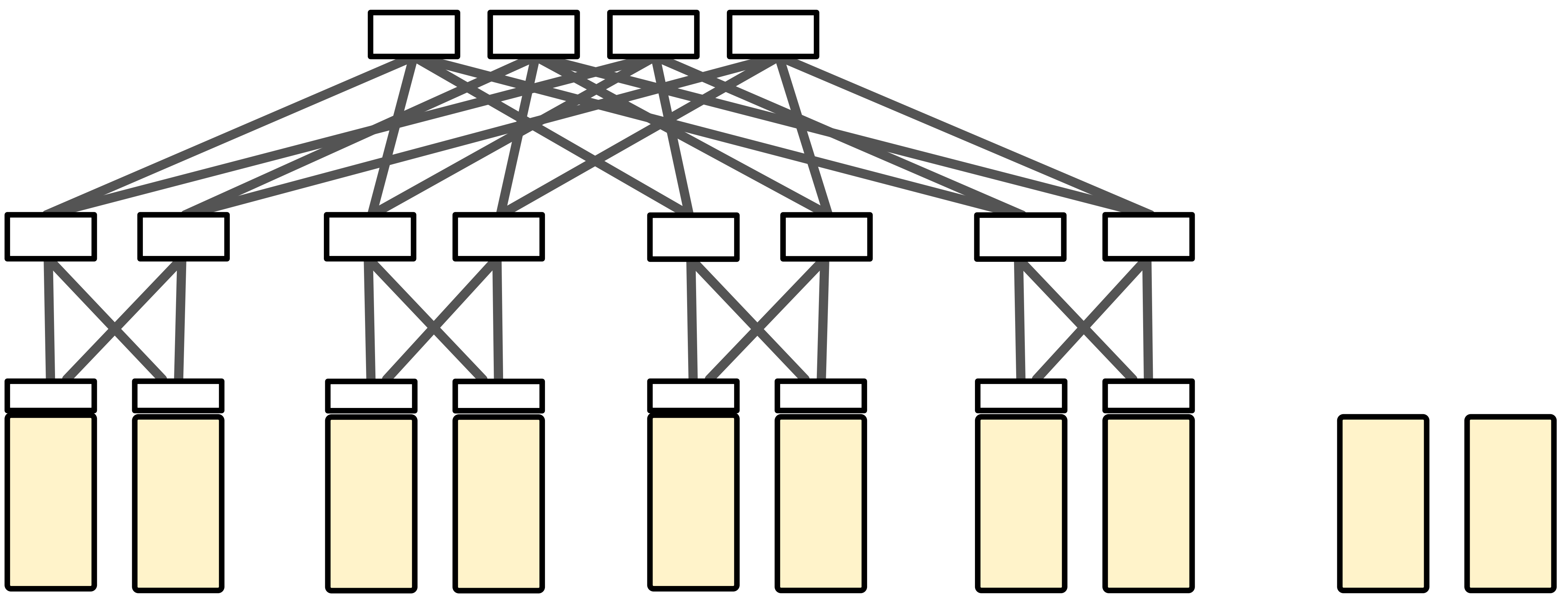
Network upgrade motivation

- Several prior solutions for greenfield data centers
  - VL2, flattened butterfly, HyperX, BCube, DCell, Al-Fares *et al.*, MDCube
- What about legacy data centers?

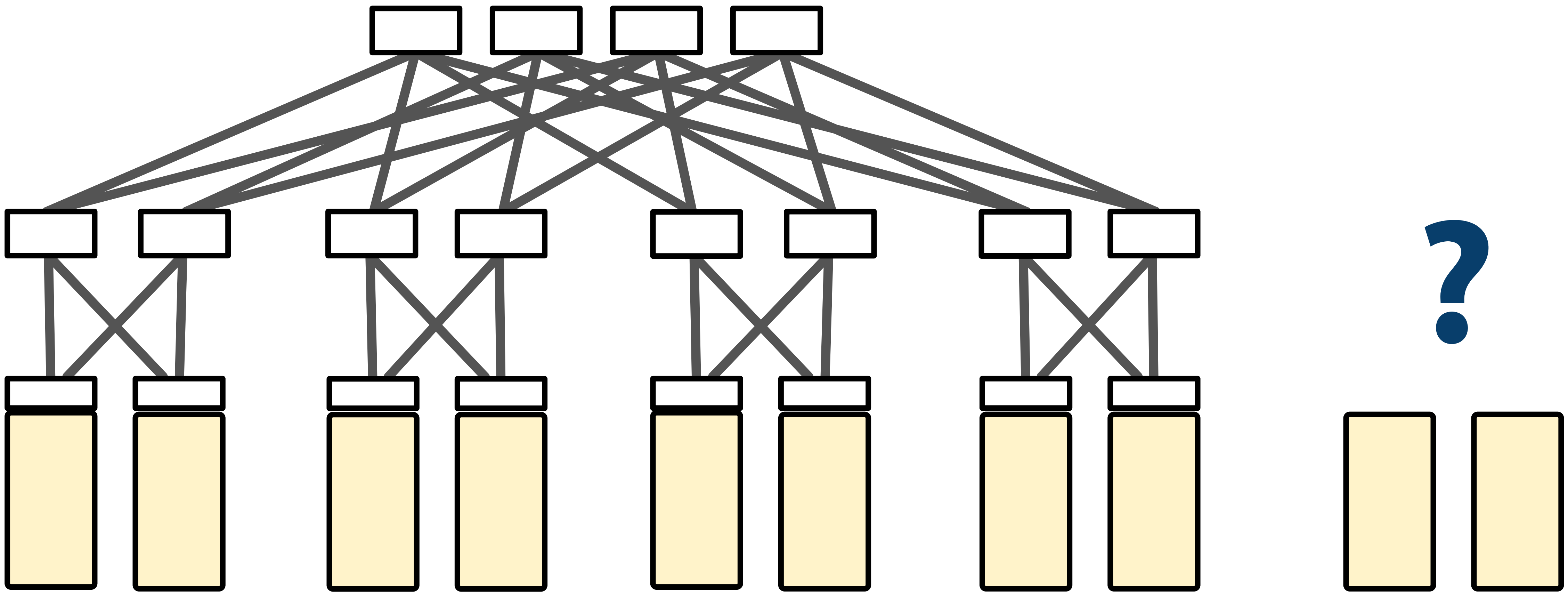
Existing topologies are not flexible enough



Existing topologies are not flexible enough



Existing topologies are not flexible enough

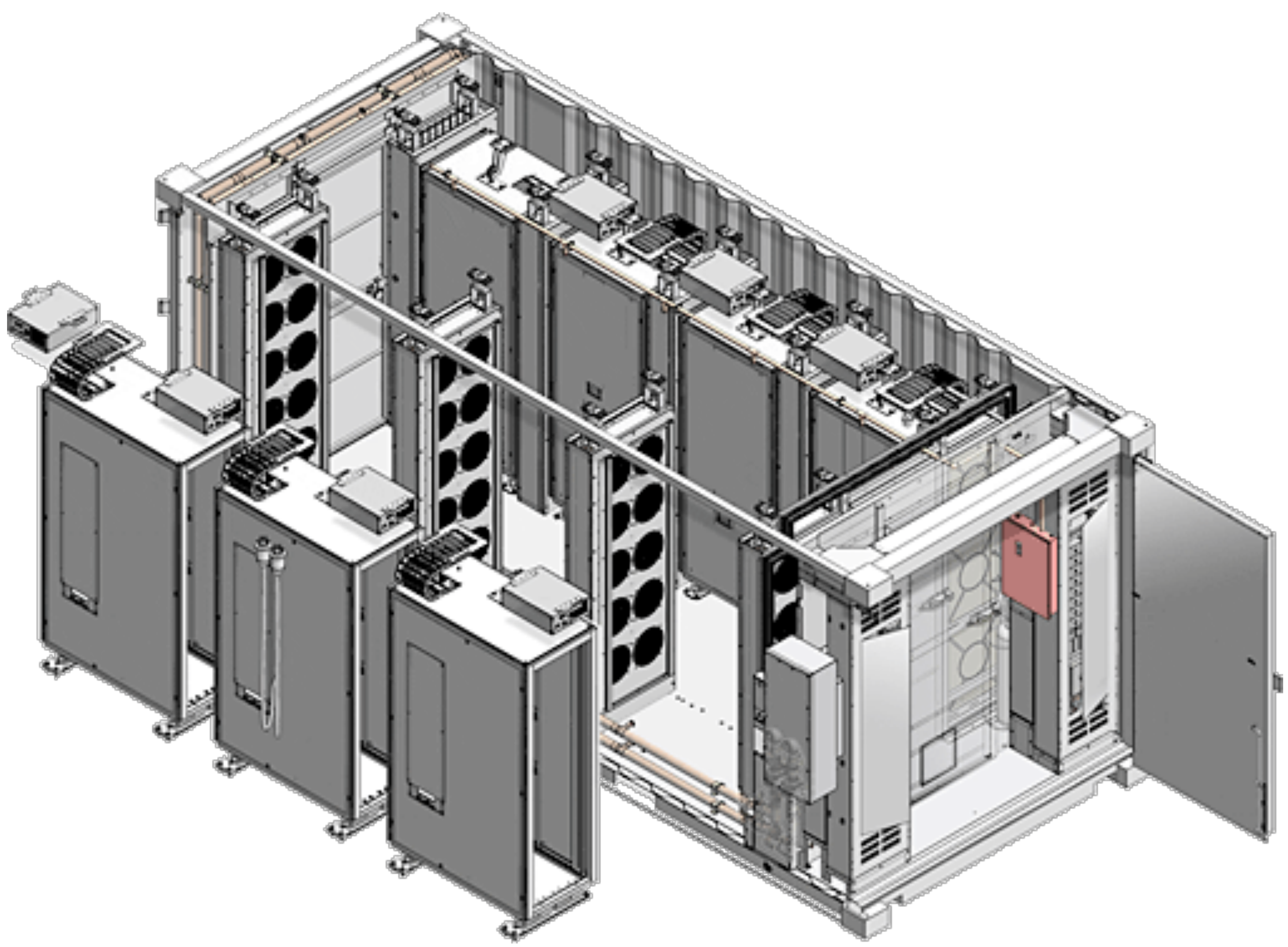


Goal

It should be easy and cost-effective to add capacity to a data center network

Challenging problem

- Designing a data center expansion or upgrade isn't easy
  - Huge design space
  - Many constraints



Problem 1

- It's hard to analyze and understand heterogeneous topologies

Problem 2

- How to design an upgraded topology?

Problem 1

- High performance network topologies are based on rigid constructions
  - Homogeneous switches
  - Prescribed switch radix
  - Single link rate

Problem 1

- High performance network topologies are based on rigid constructions
  - Homogeneous switches
  - Prescribed switch radix
  - Single link rate

Solutions:

1. ***develop theory of heterogeneous Clos networks***

2. ***explore unstructured data center network topologies***

Two solutions:

LEGUP: output is a heterogeneous Clos network

[Curtis, Keshav, López-Ortiz; CoNEXT 2010]

REWIRE: designs unstructured DCN topologies

[Curtis et al.; INFOCOM 2012]

Two solutions:

LEGUP: output is a heterogeneous Clos network

[Curtis, Keshav, López-Ortiz; CoNEXT 2010]

REWIRE: designs unstructured DCN topologies

[Curtis et al.; INFOCOM 2012]

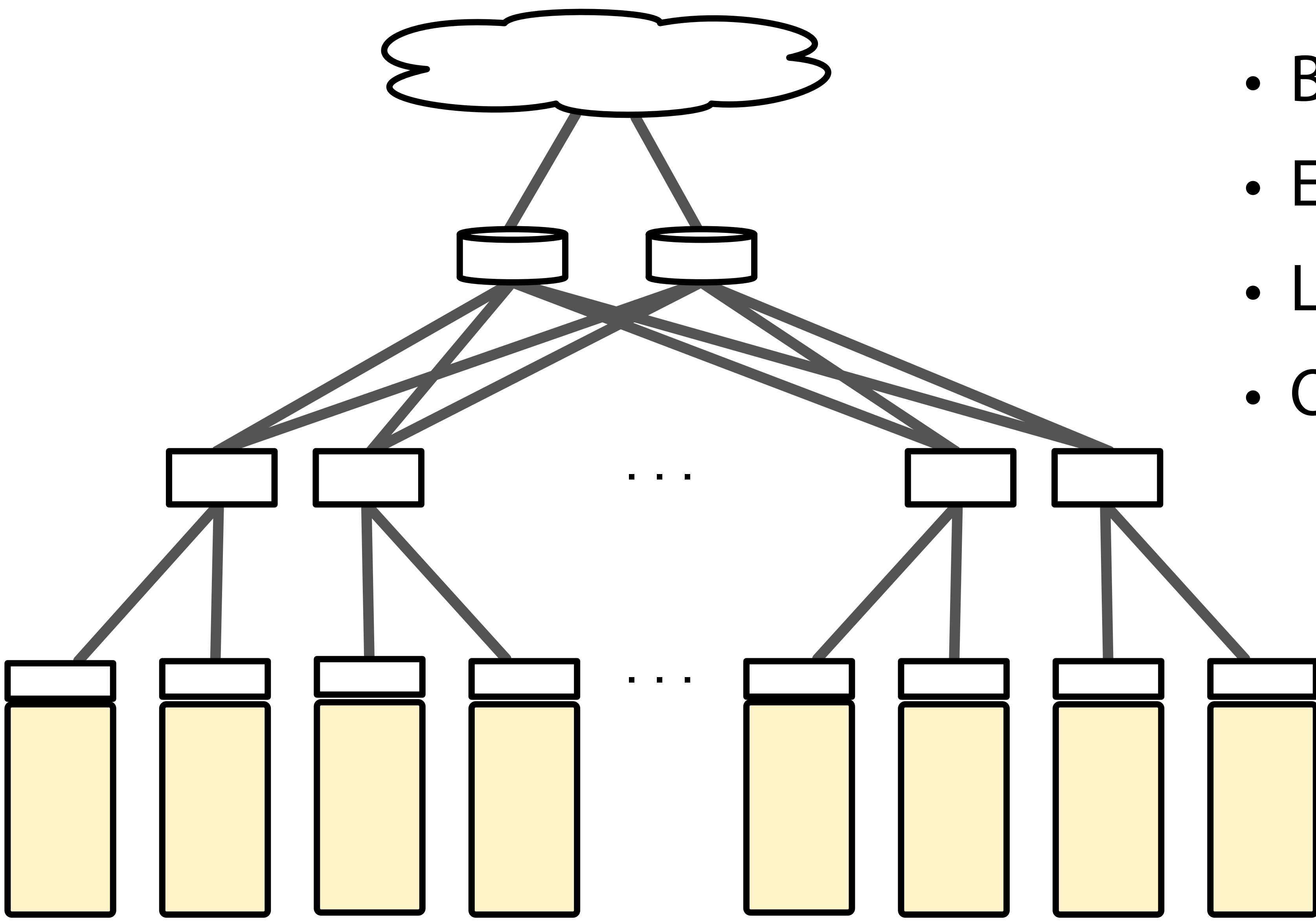
LEGUP in brief:

LEGUP designs upgraded/expanded networks for legacy data center networks

LEGUP in brief:

LEGUP designs upgraded/expanded networks for legacy data center networks

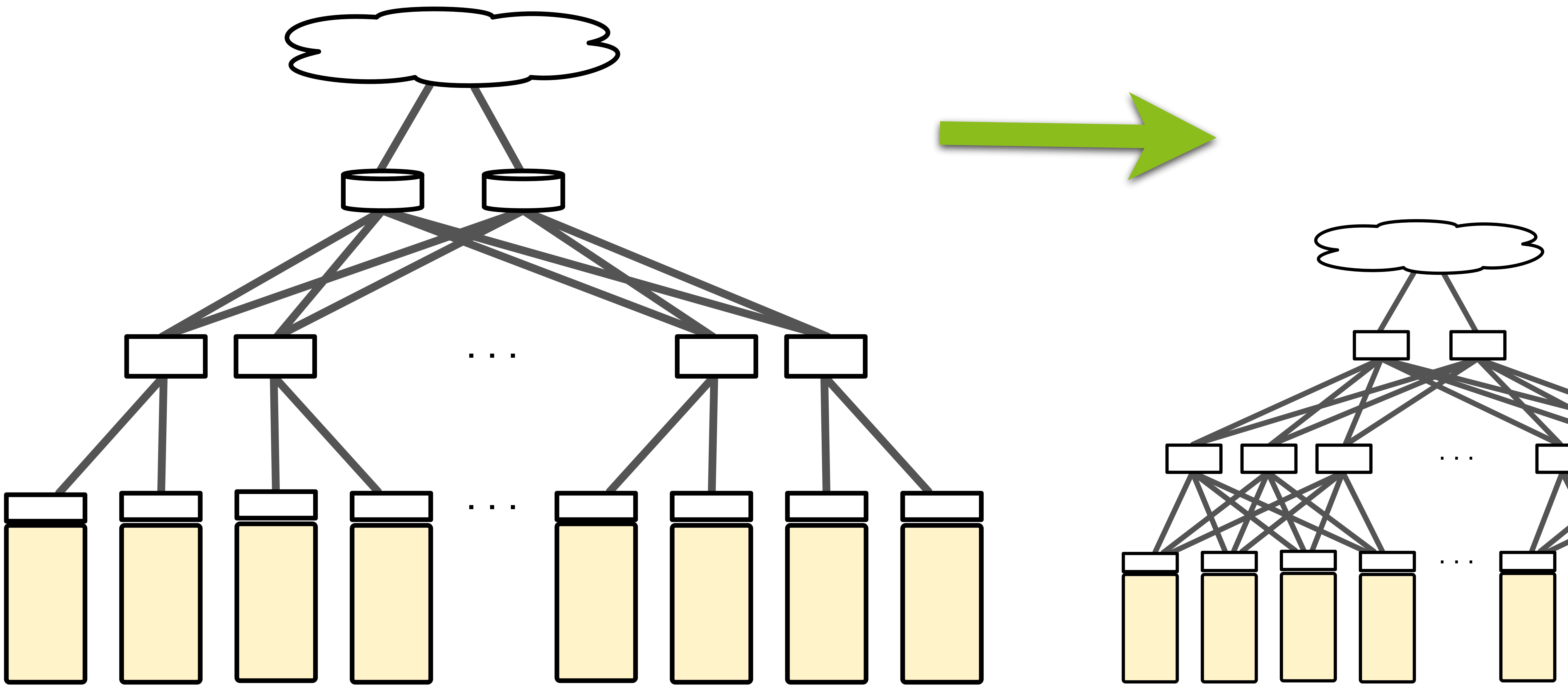
Input



- Budget
- Existing network topology
- List of switches & line cards
- Optional: data center model

LEGUP in brief:

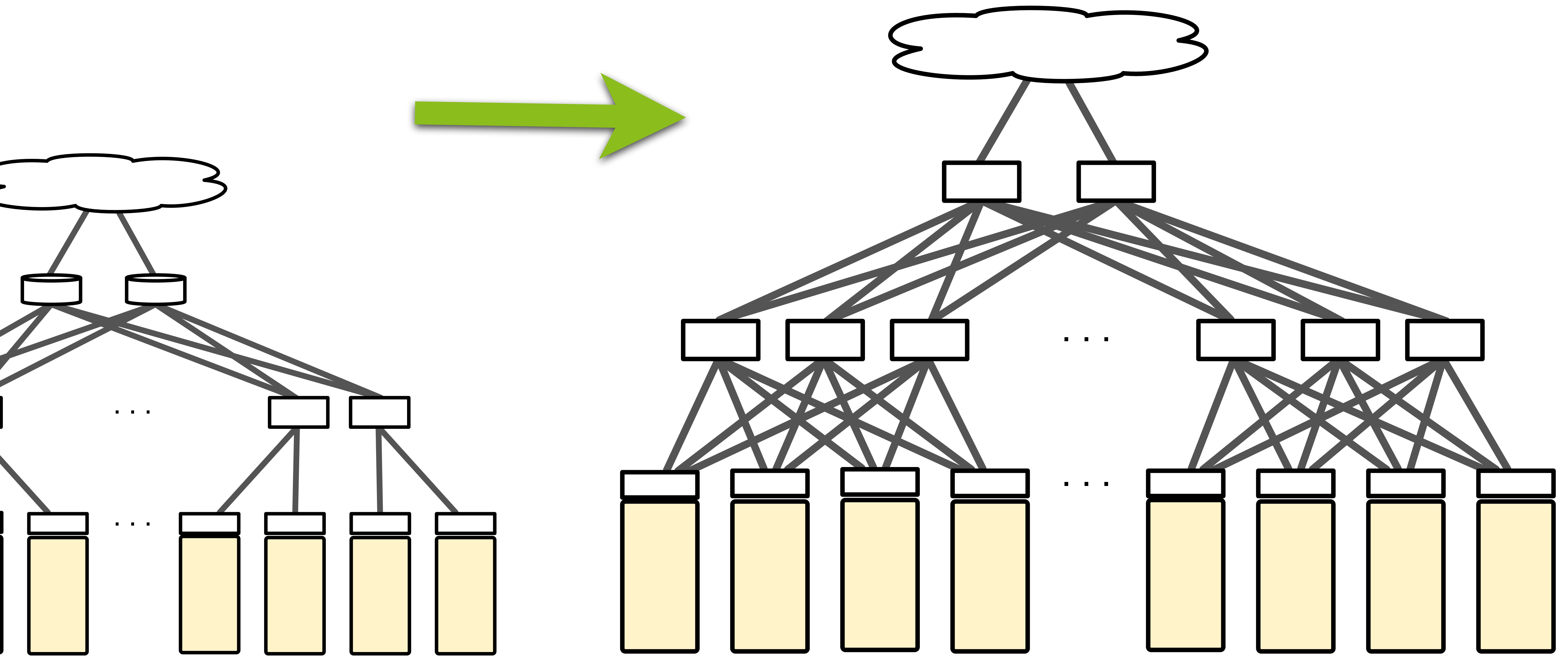
LEGUP designs upgraded/expanded networks for legacy data center networks



Input ***Output***

LEGUP in brief:

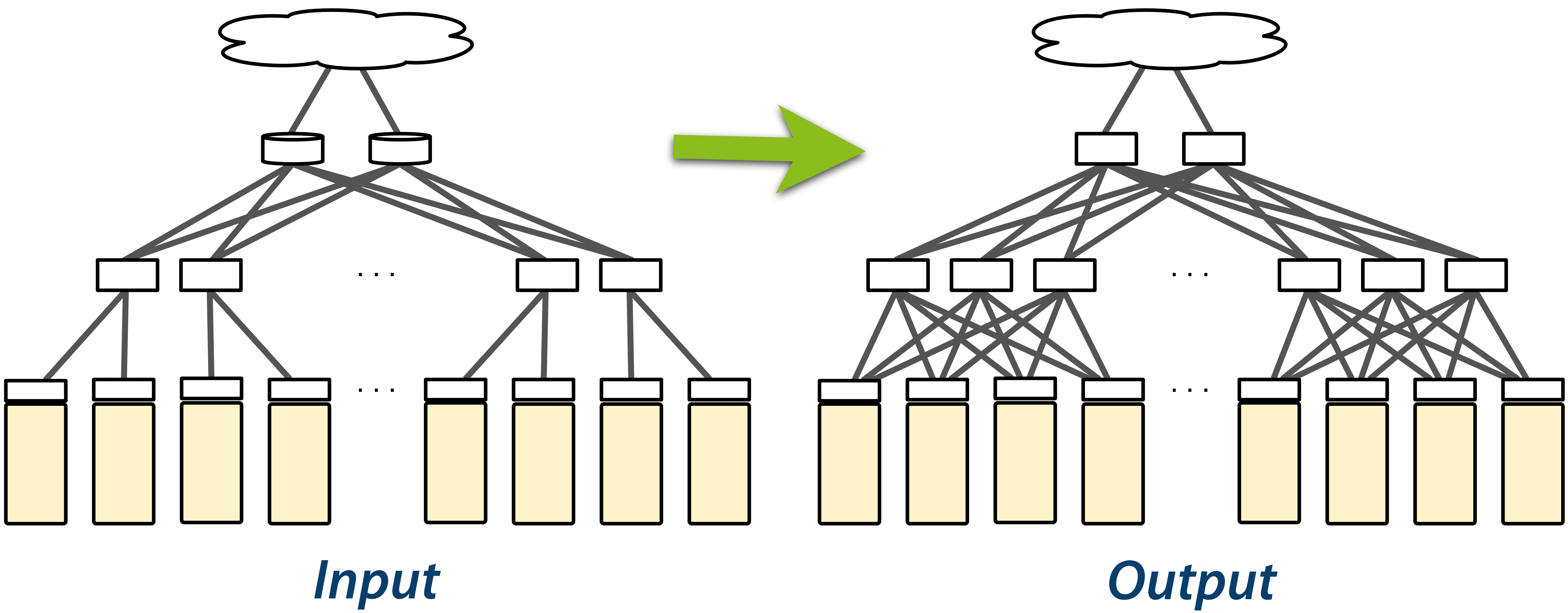
LEGUP designs upgraded/expanded networks for legacy data center networks



Input ***Output***

LEGUP in brief:

LEGUP designs upgraded/expanded networks for legacy data center networks



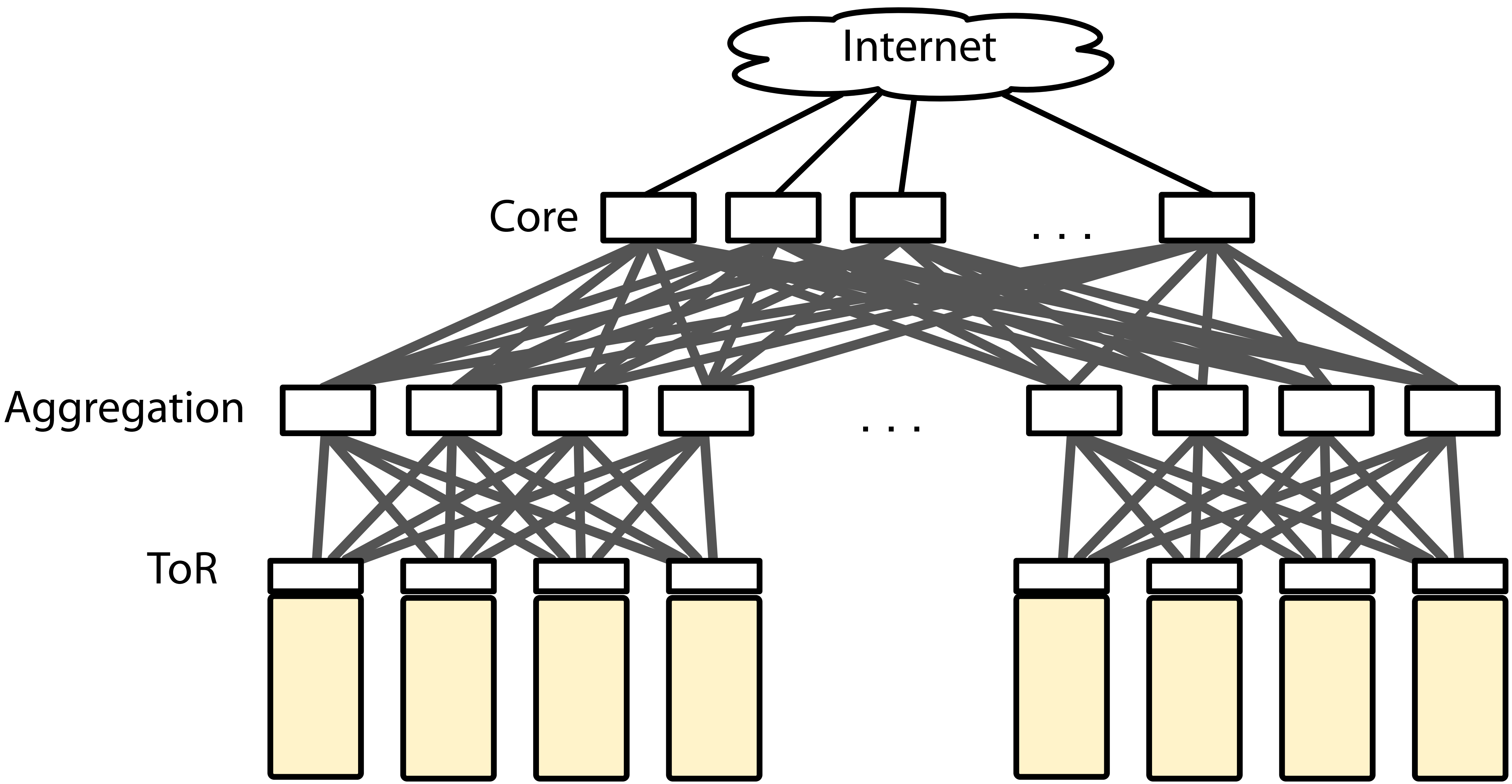
Difficult optimization problem

Difficult optimization problem

First pass: limit solution space by finding only *heterogeneous Clos networks*

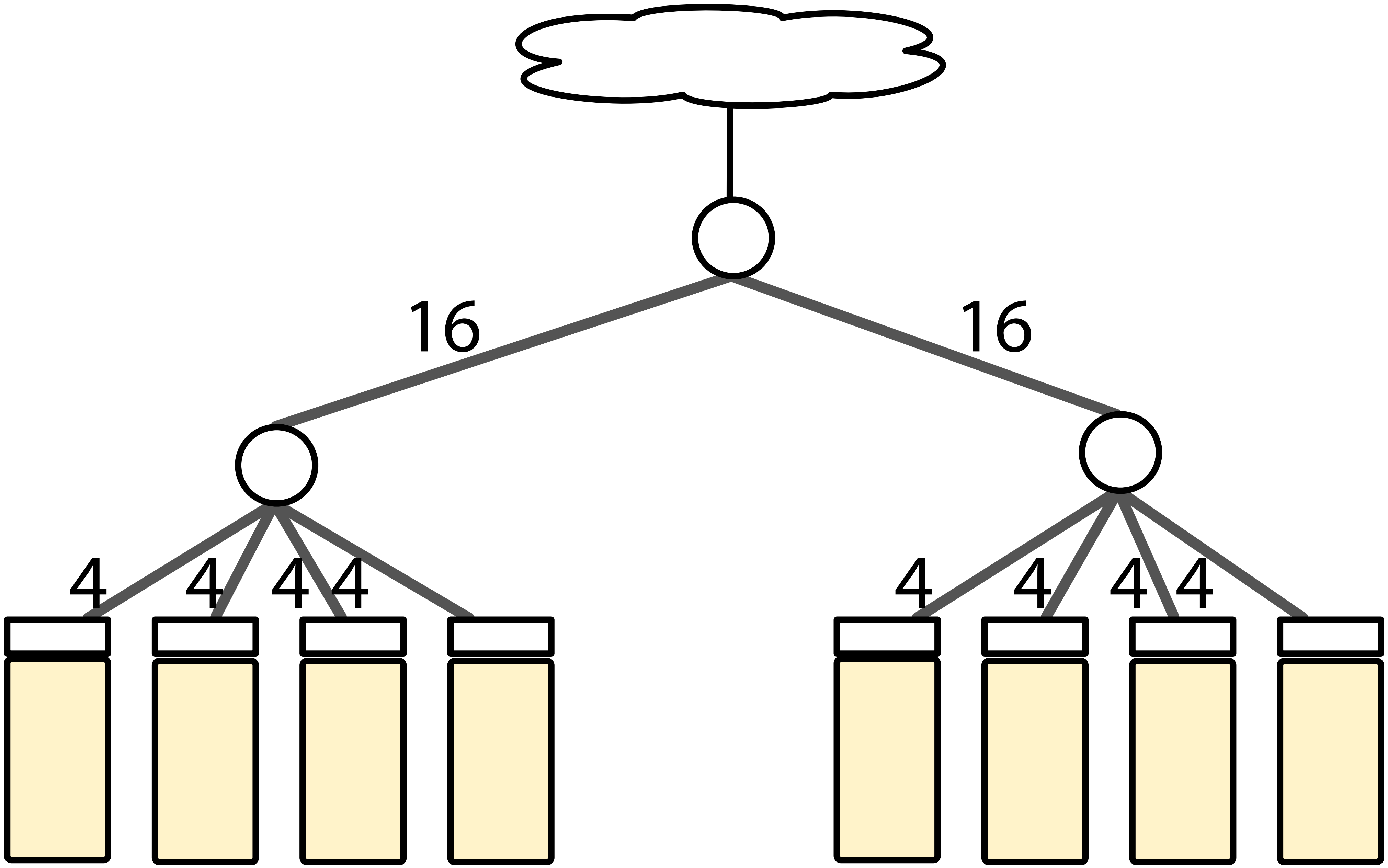
Clos networks

This is a *physical realization* of a Clos network



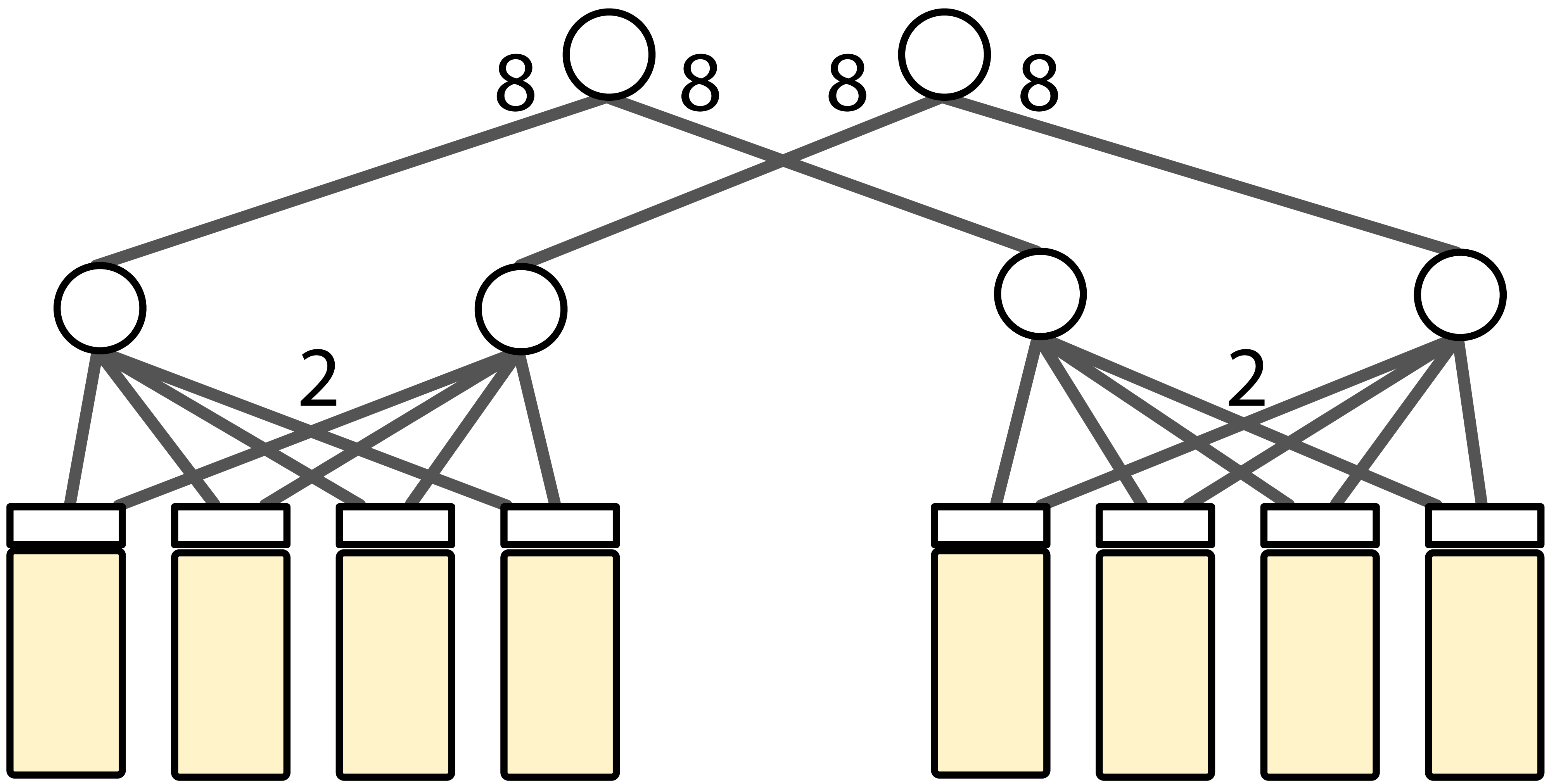
Clos networks

We can find a *logical topology* for this network



Heterogeneous Clos networks

Logical topology is a forest



Theoretical contributions

**optimal = uses same link capacity an equivalent stage Clos network*

Theoretical contributions

Lemma 1: How to construct all optimal logical forests for a set of switches

**optimal = uses same link capacity an equivalent stage Clos network*

Theoretical contributions

Lemma 1: How to construct all optimal logical forests for a set of switches

Lemma 2: How to build a physical realization from a logical forest

**optimal = uses same link capacity an equivalent stage Clos network*

Theoretical contributions

Lemma 1: How to construct all optimal logical forests for a set of switches

Lemma 2: How to build a physical realization from a logical forest

Theorem: A characterization of heterogeneous Clos networks

**optimal = uses same link capacity an equivalent stage Clos network*

Theoretical contributions

Lemma 1: How to construct all optimal logical forests for a set of switches

Lemma 2: How to build a physical realization from a logical forest

Theorem: A characterization of heterogeneous Clos networks

This is the first optimal heterogeneous topology

**optimal = uses same link capacity an equivalent stage Clos network*

Problem 1

- It's hard to analyze and understand heterogeneous topologies *more later...*

Problem 2

- How to design an upgraded topology?

Problem 1

- It’s hard to analyze and understand heterogeneous topologies

Problem 2

- How to design an upgraded topology?

heterogeneous Clos

Problem 2

Upgraded network should:

- Maximize performance, minimize cost
- Be realized in the target data center
- Incorporate existing network equipment if it makes sense

Approach: use optimization

LEGUP algorithm

- Branch and bound search of solution space
  - Heuristics to map switches to a rack
- See paper for details
- Time is bottleneck in algorithm
  - Exponential in number of switch types and (worst-case) in number ToRs
  - 760 server data center: 5–10 minutes to run algorithm
  - 7600 server data center: 1–2 days
  - But can be parallelized

LEGUP summary

- Developed theory of heterogeneous Clos networks
- Implemented LEGUP design algorithm
- On our data center, we see substantial cost savings: spend less than half as much money as a fat-tree for same performance

Two solutions:

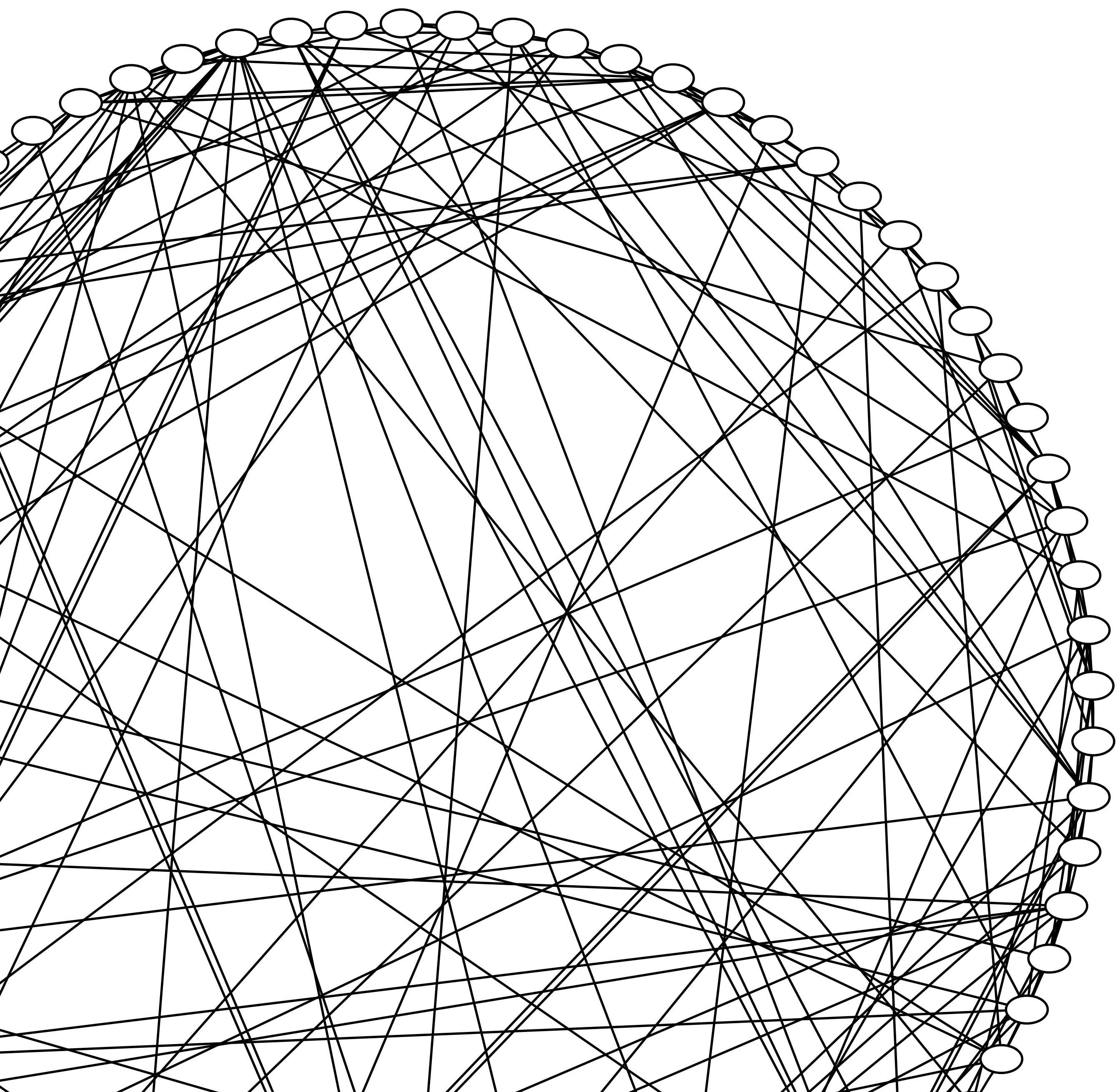
LEGUP: output is a heterogeneous Clos network

[Curtis, Keshav, López-Ortiz; CoNEXT 2010]

REWIRE: designs unstructured DCN topologies

[Curtis et al.; INFOCOM 2012]

Can we do better with unstructured networks?



Problem

- Now we have an even harder network design problem

Problem

- Now we have an even harder network design problem

Approach

- Use local search heuristics to find a “good enough” solution

REWIRE

Uses simulated annealing to find a network that:

- Maximizes performance

Subject to:

- The budget
- Physical constraints of the data center model (thermal, power, space)
- No topology restrictions

REWIRE

Uses simulated annealing to find a network that:

- Maximizes performance **_Bisection bandwidth - Diameter_**

Subject to:

- The budget
- Physical constraints of the data center model (thermal, power, space)
- No topology restrictions

REWIRE

Uses simulated annealing to find a network that:

- Maximizes performance

Subject to:

- The budget
- Physical constraints of the data center model (thermal, power, space)
- No topology restrictions

$$\textit{Costs} = \textit{new cables} + \textit{moved cables} + \textit{new switches}$$

Simulated annealing algorithm

- *At each iteration, computes*
  - Performance of candidate solution
  - If accept this solution, then
    - Compute next neighbor to consider

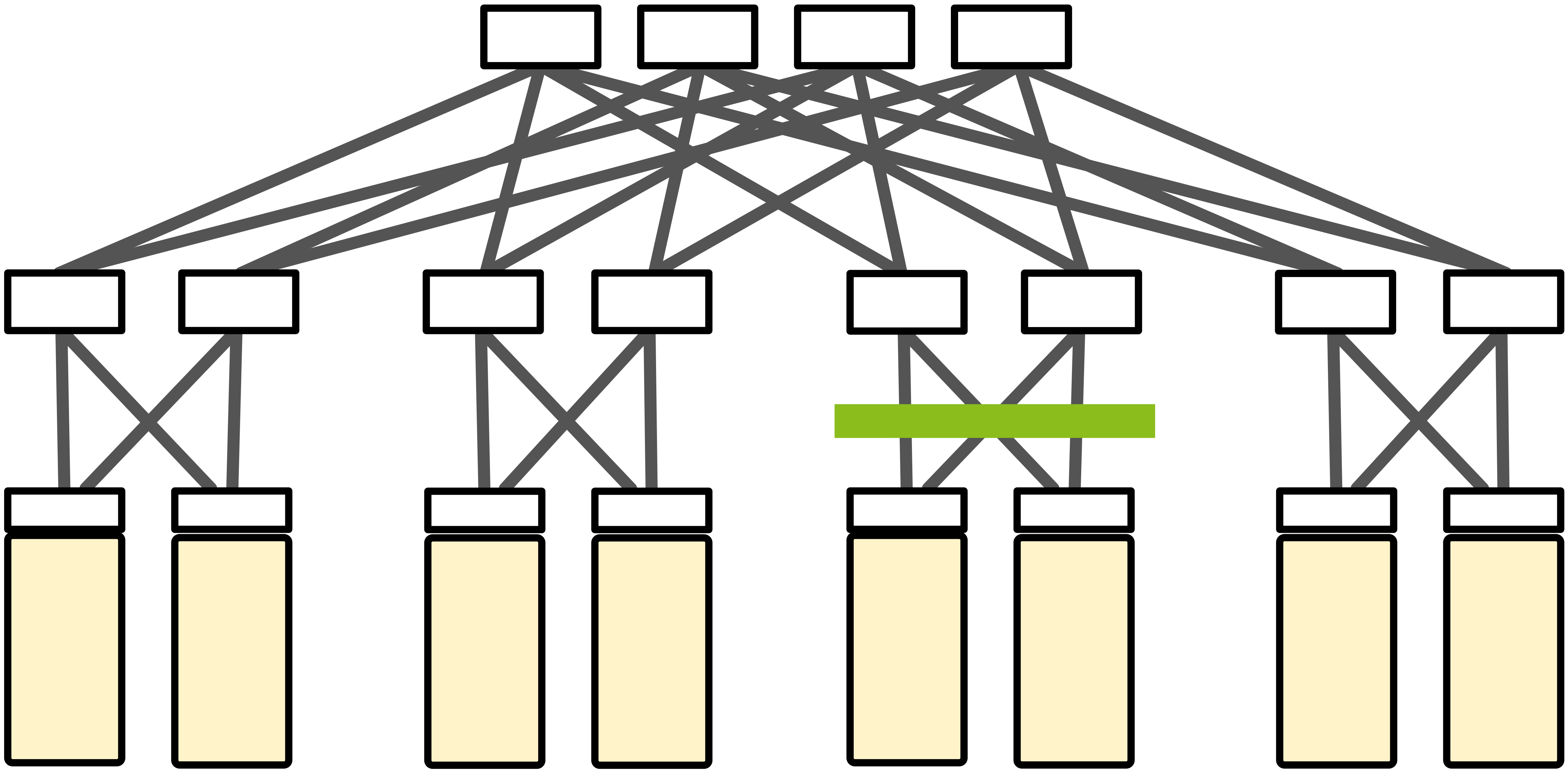
Simulated annealing algorithm

- *At each iteration, computes*
  - **Performance of candidate solution**

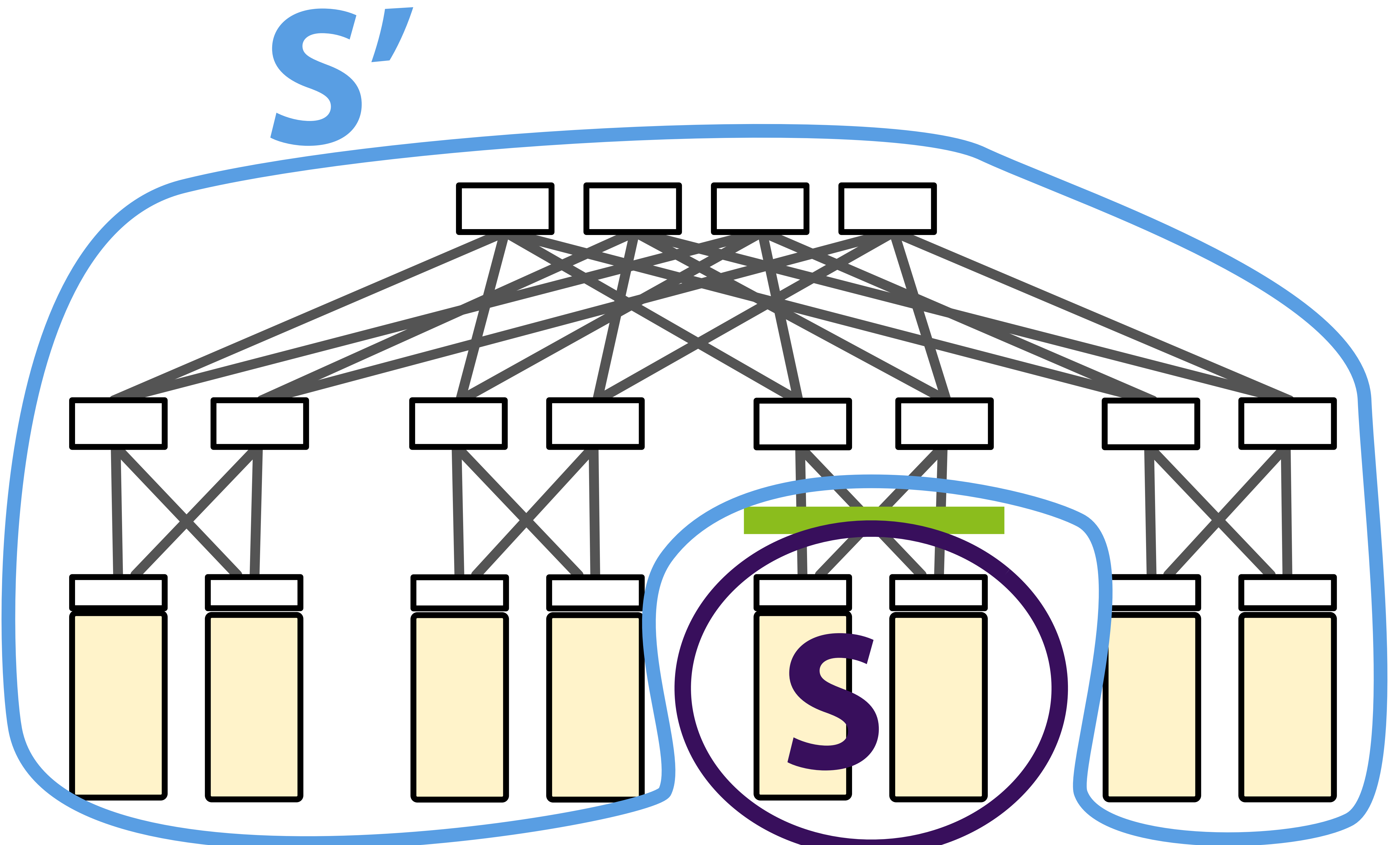
No known algorithm to find the bisection bandwidth of an arbitrary network!

Bisection bandwidth computation

Easy for a single cut

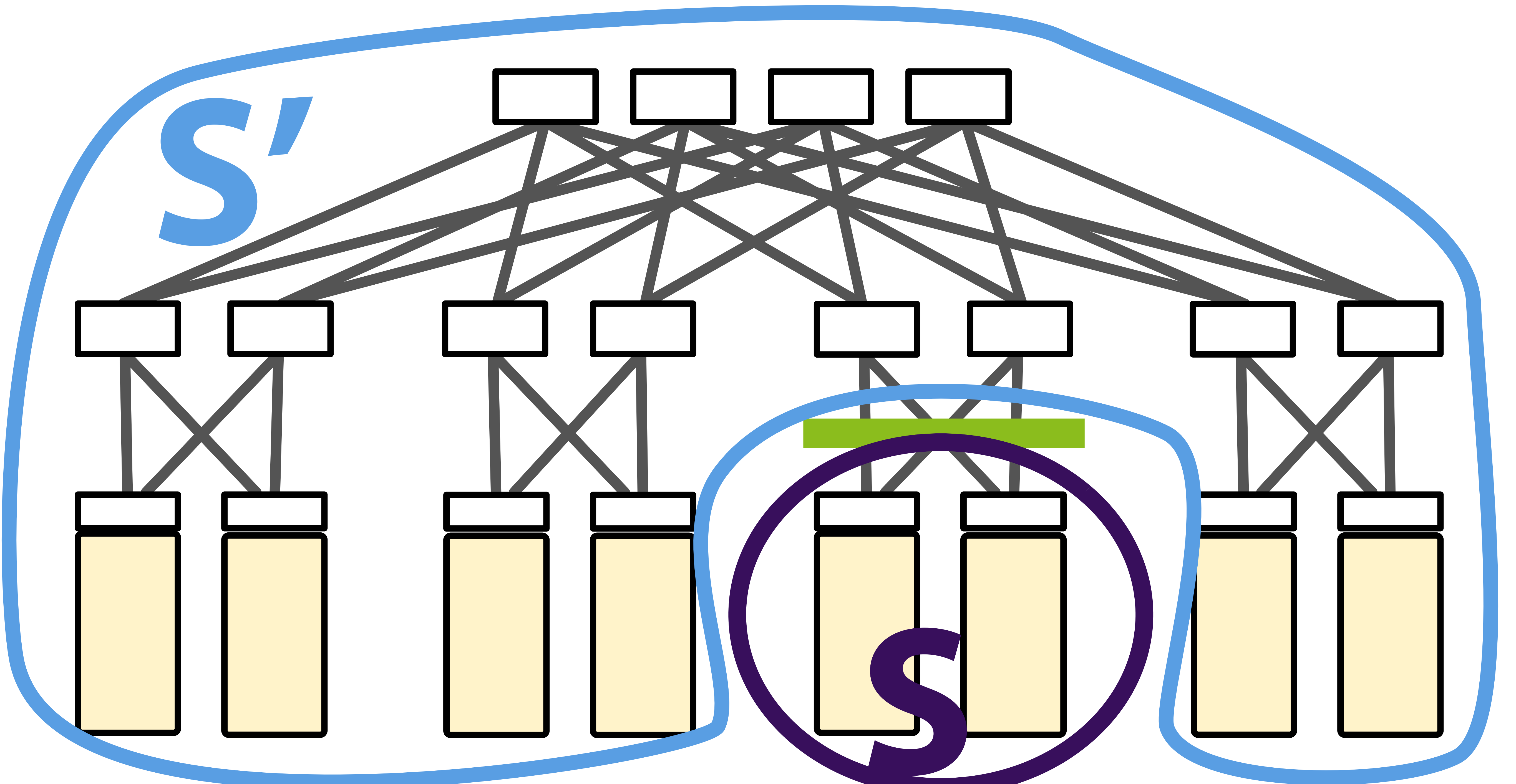


Bisection bandwidth computation



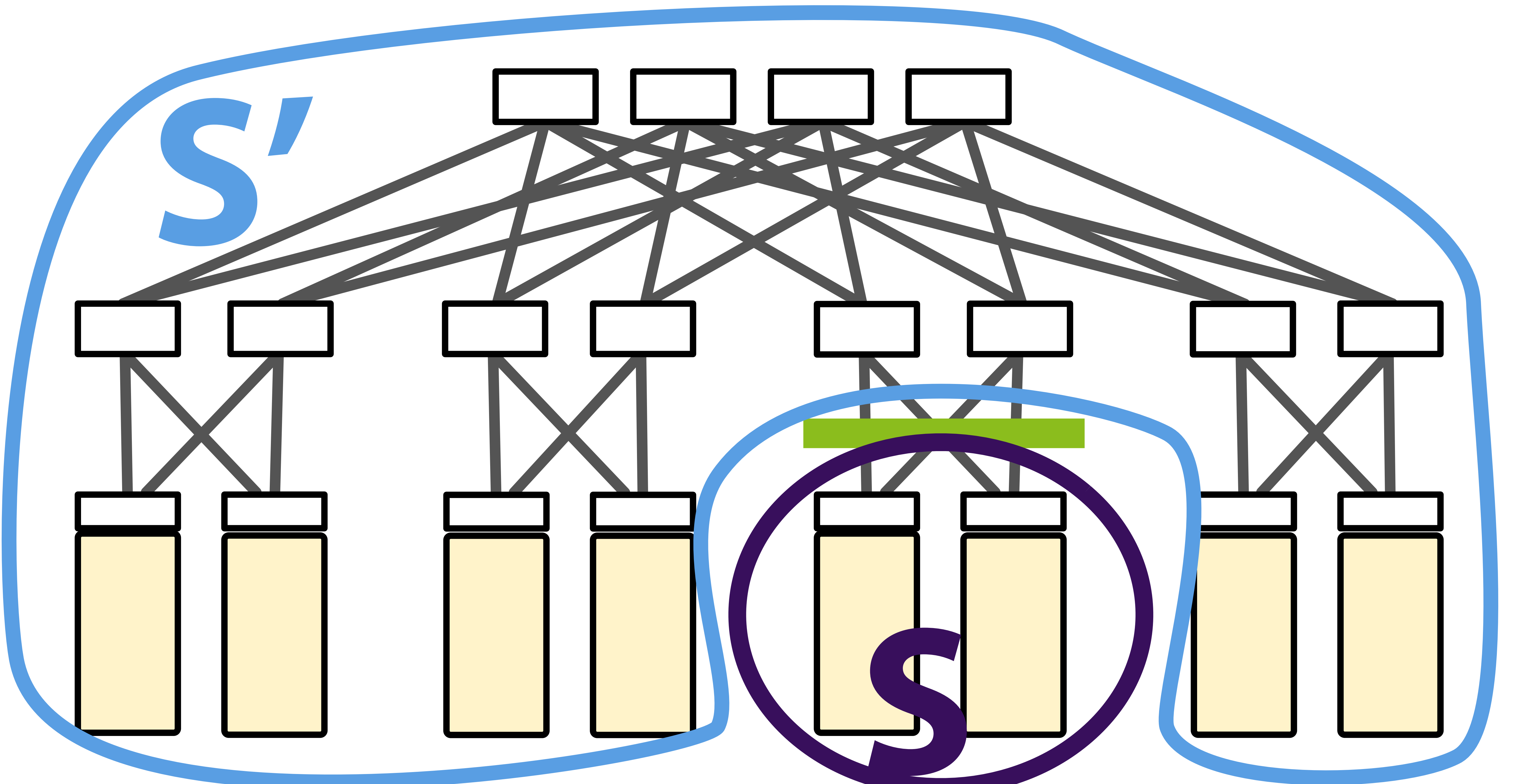
Bisection bandwidth computation

$$bw(S,S') = \frac{\text{link cap}(S,S')}{\min\{\text{ server rates}(S), \text{server rates}(S')\ \}}$$



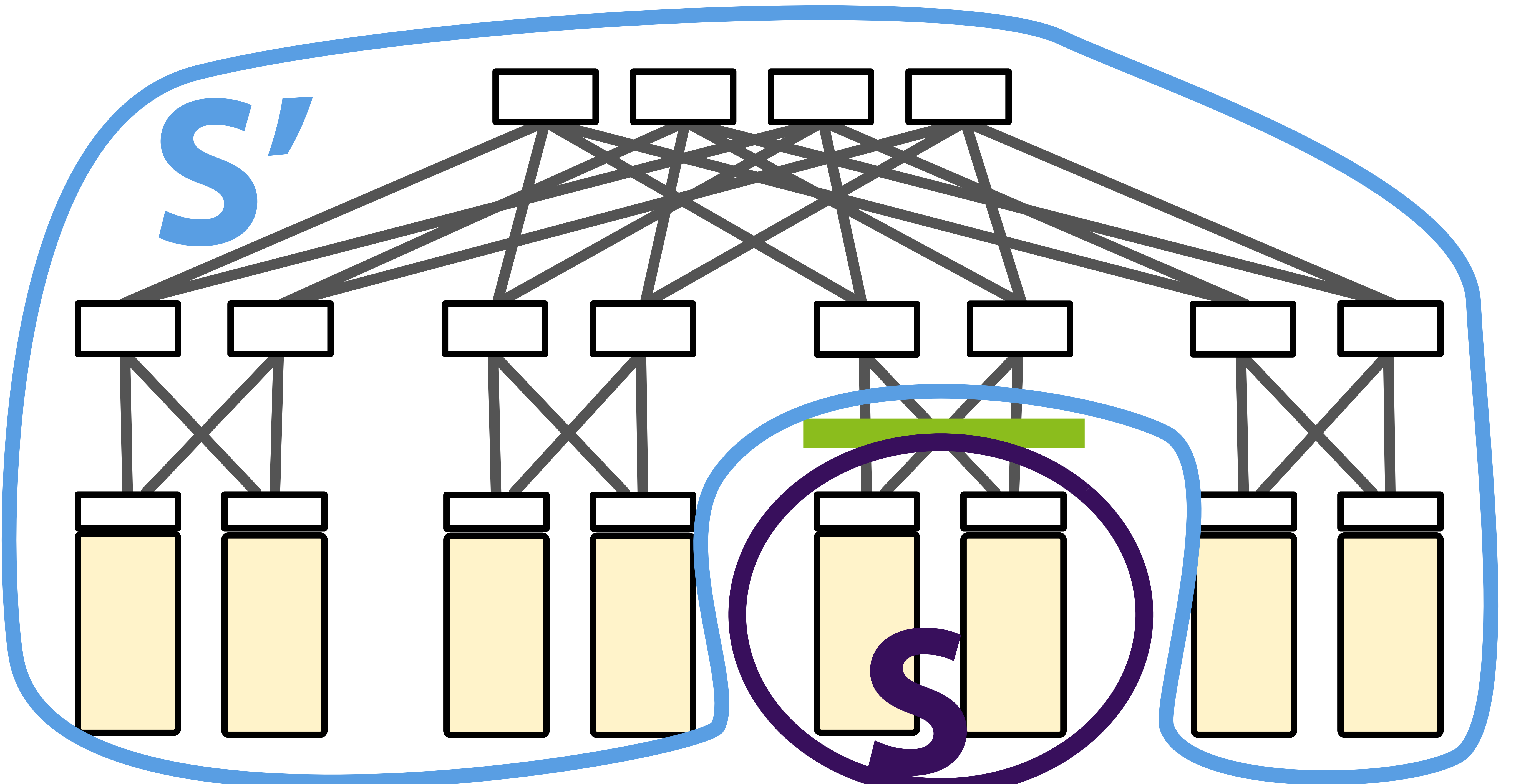
Bisection bandwidth computation

$$\text{bw}(S,S') = \frac{4}{\min\{2, 6\}}$$



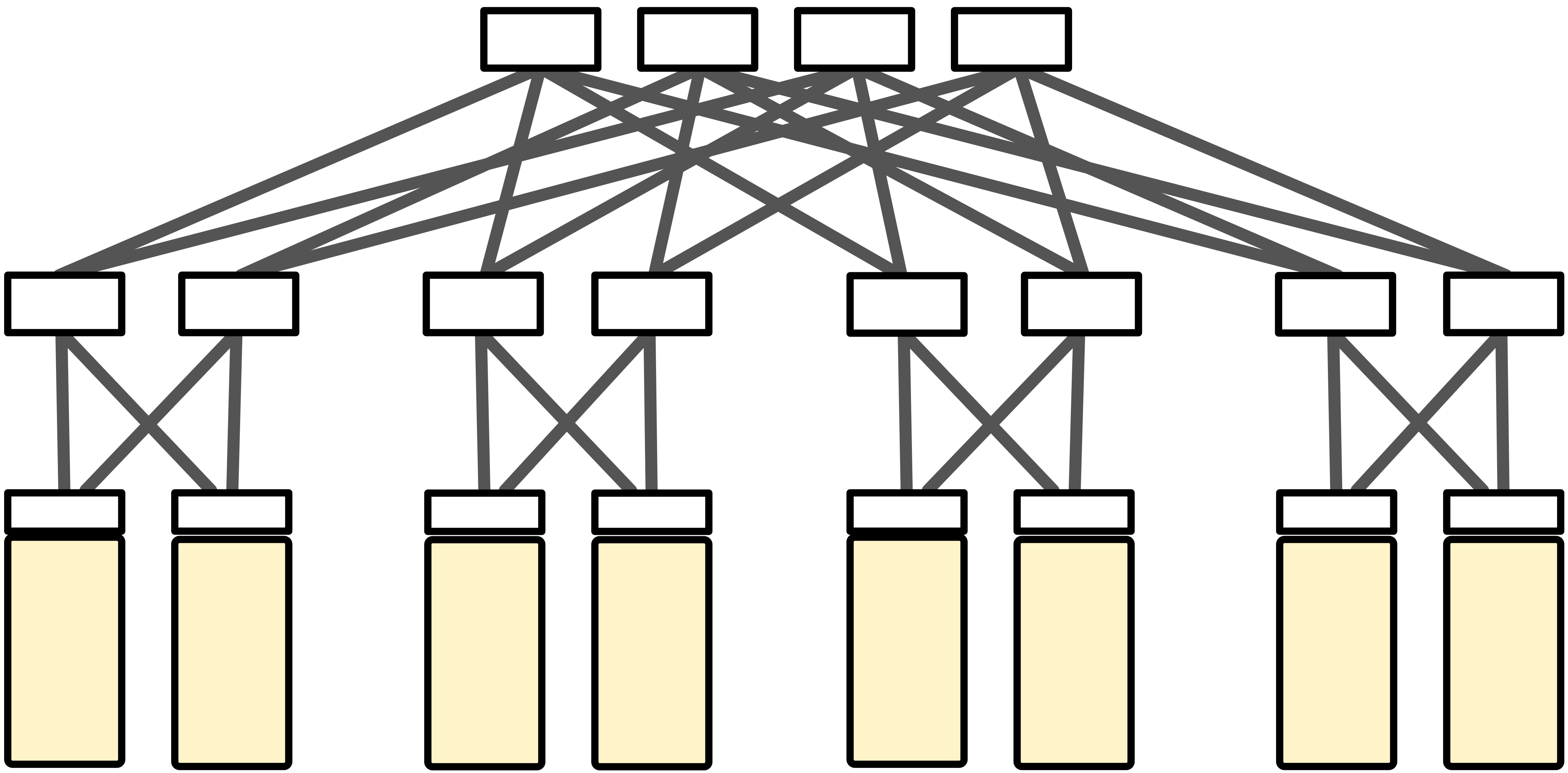
Bisection bandwidth computation

Then bisection bandwidth is the min over all cuts



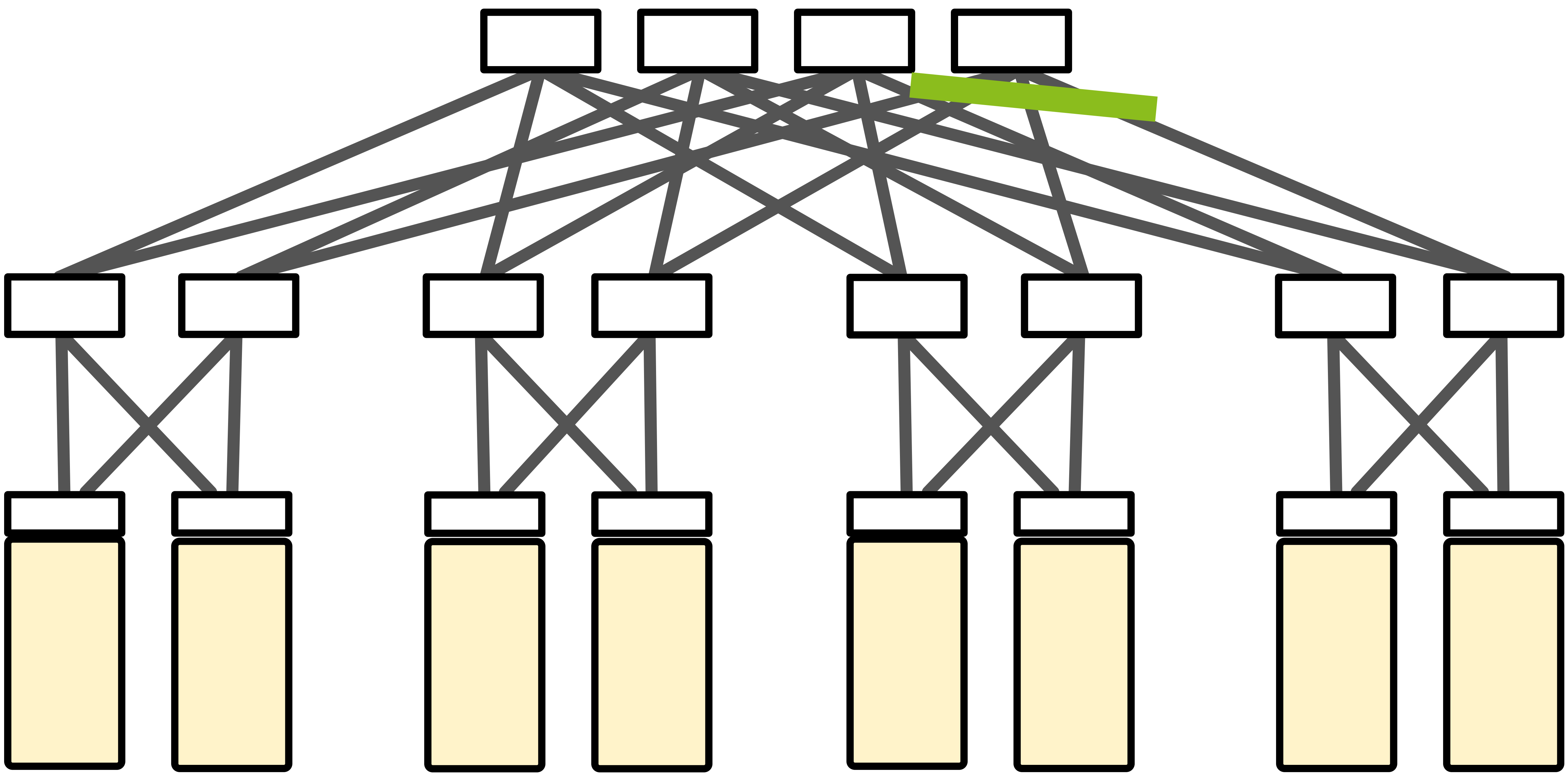
Bisection bandwidth computation

- Easy on tree-like topologies because there are O(n) cuts



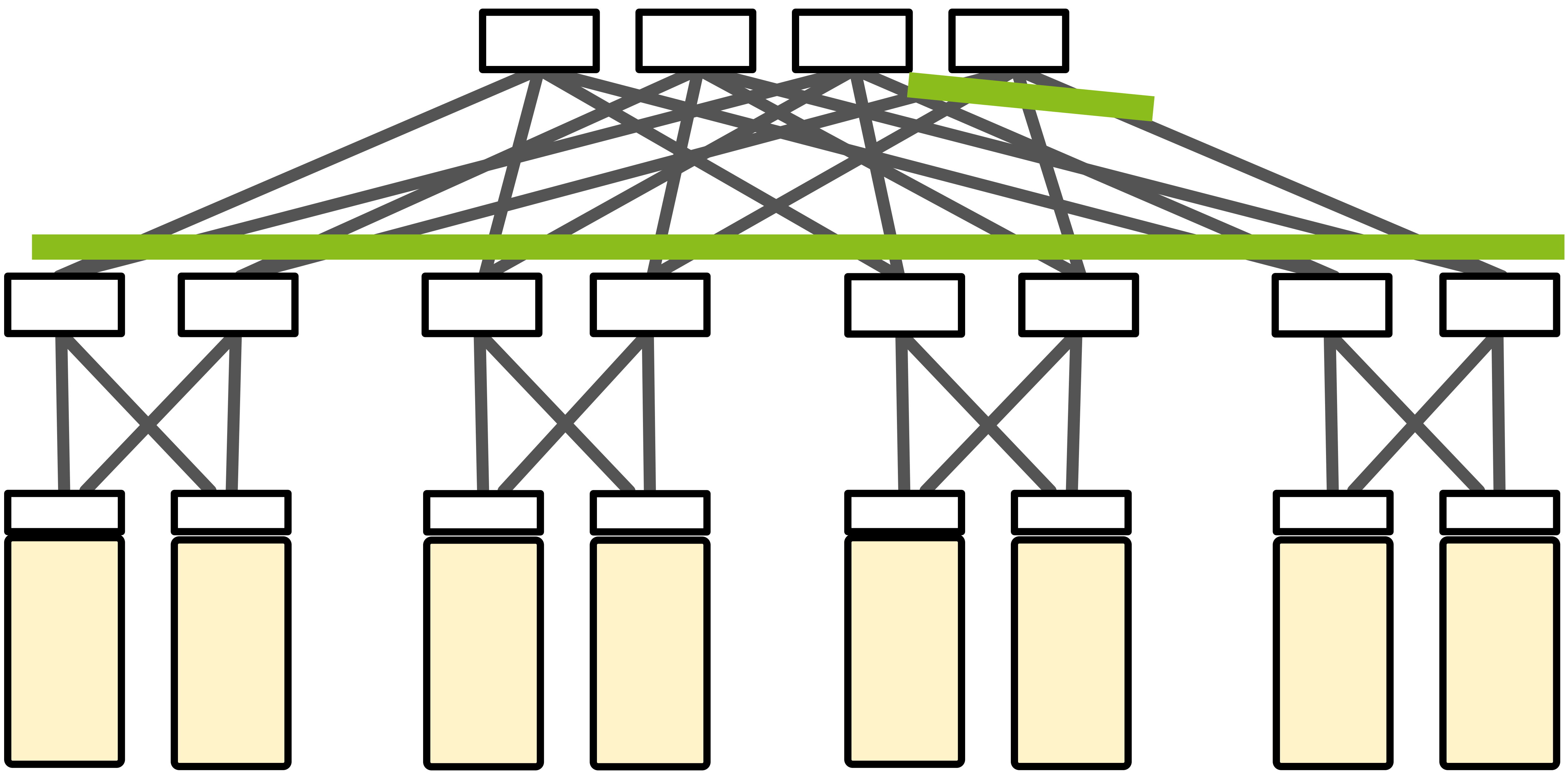
Bisection bandwidth computation

- Easy on tree-like topologies because there are O(n) cuts



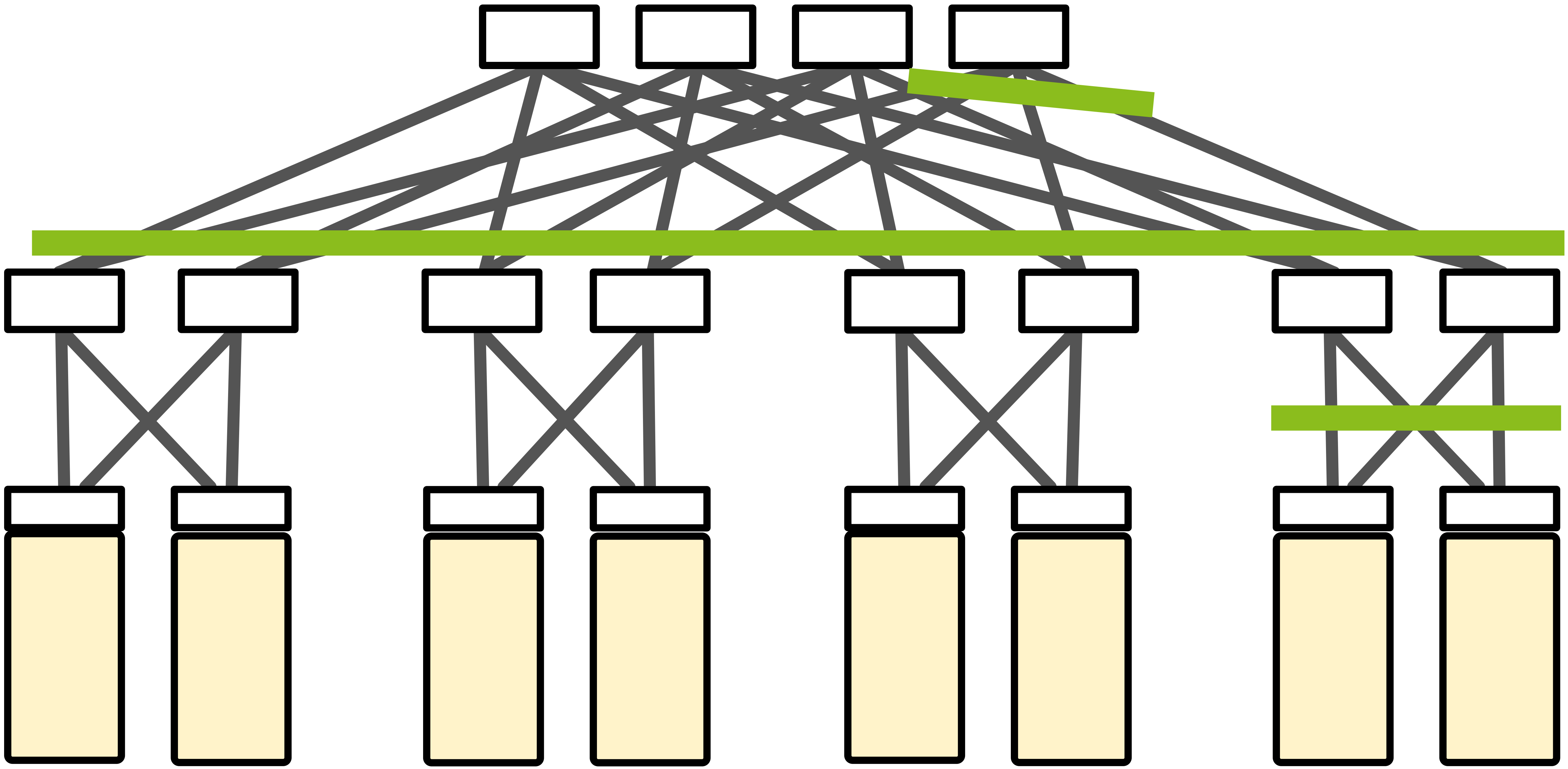
Bisection bandwidth computation

- Easy on tree-like topologies because there are O(n) cuts



Bisection bandwidth computation

- Easy on tree-like topologies because there are O(n) cuts



Bisection bandwidth computation

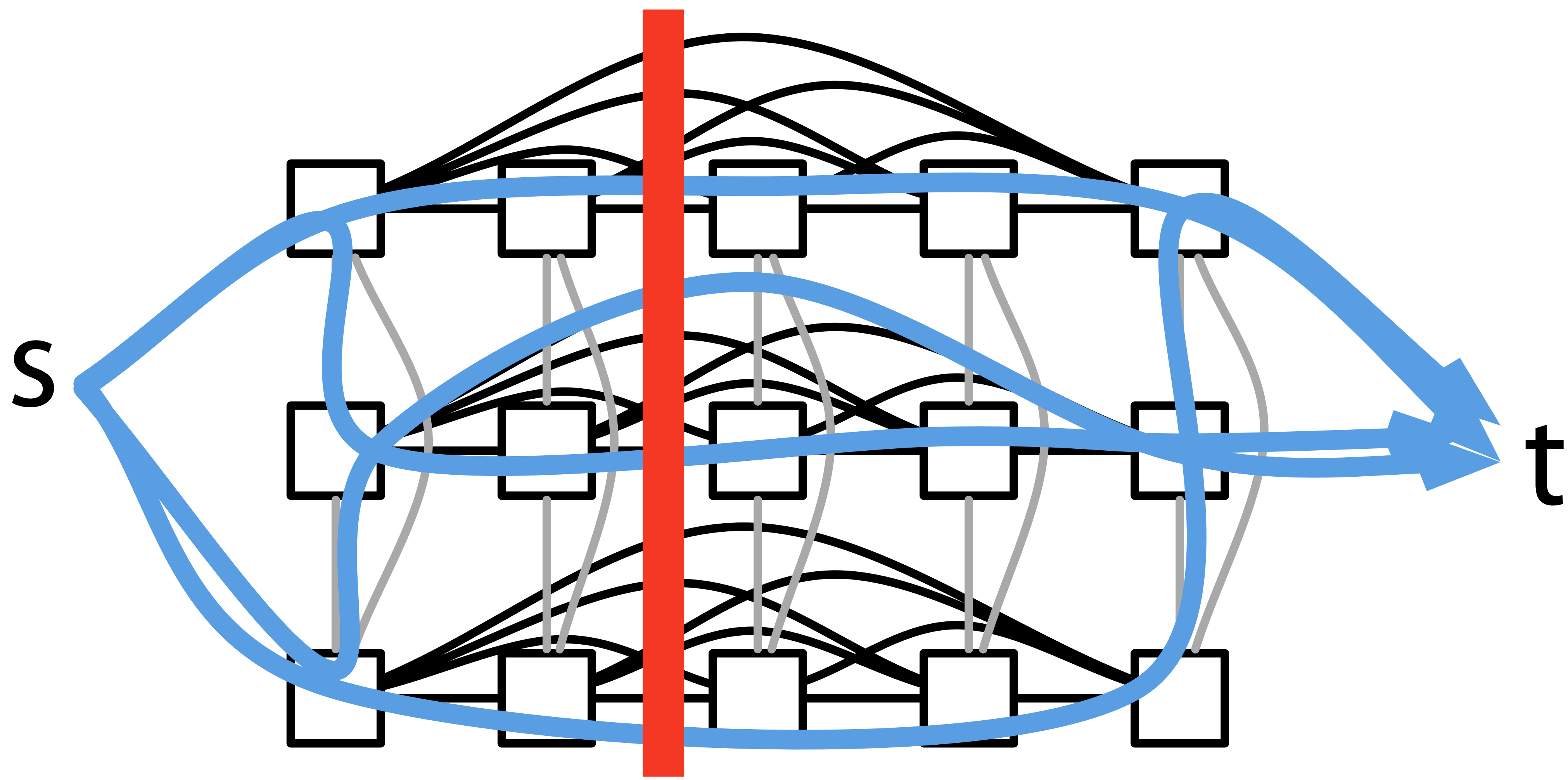
Bisection bandwidth computation

Exponentially many cuts on arbitrary topologies

Bisection bandwidth computation

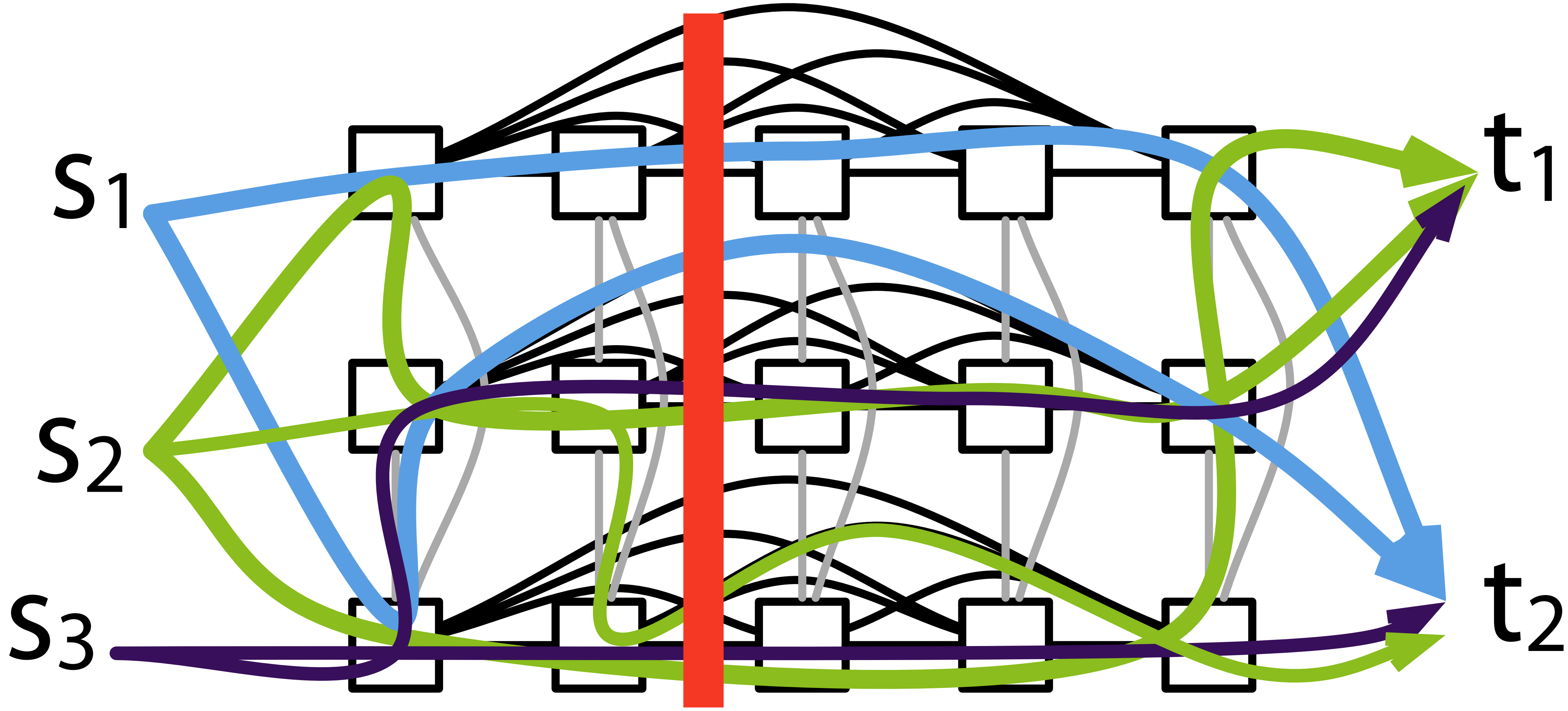
Exponentially many cuts on arbitrary topologies

Need: A min-cut, max-flow type theorem for multi-commodity flow



Bisection bandwidth computation

Need: A min-cut, max-flow type theorem for multi-commodity flow



Bisection bandwidth computation

Bisection bandwidth computation

Theorem [Curtis and López-Ortiz, INFOCOM 2009]**:**

A network can feasibly route all traffic matrices feasible under the server NIC rates using multipath routing iff all its cuts have bandwidth ≥ a sum dependent on α_i for all nodes i

Bisection bandwidth computation

Theorem [Curtis and López-Ortiz, INFOCOM 2009]**:**

A network can feasibly route all traffic matrices feasible under the server NIC rates using multipath routing iff all its cuts have bandwidth ≥ a sum dependent on α_i for all nodes i

We can compute the α_i values using linear programming

[Kodialam et al. INFOCOM 2006]

Bisection bandwidth computation

Theorem [Curtis and López-Ortiz, INFOCOM 2009]**:**
A network can feasibly route all traffic matrices feasible under the server NIC rates using multipath routing iff all its cuts have bandwidth ≥ a sum dependent on α_i for all nodes i

We can compute the α_i values using linear programming

[Kodialam et al. INFOCOM 2006]

These two theoretical results give us a polynomial-time algorithm to find the bisection bandwidth of an arbitrary network

Evaluation

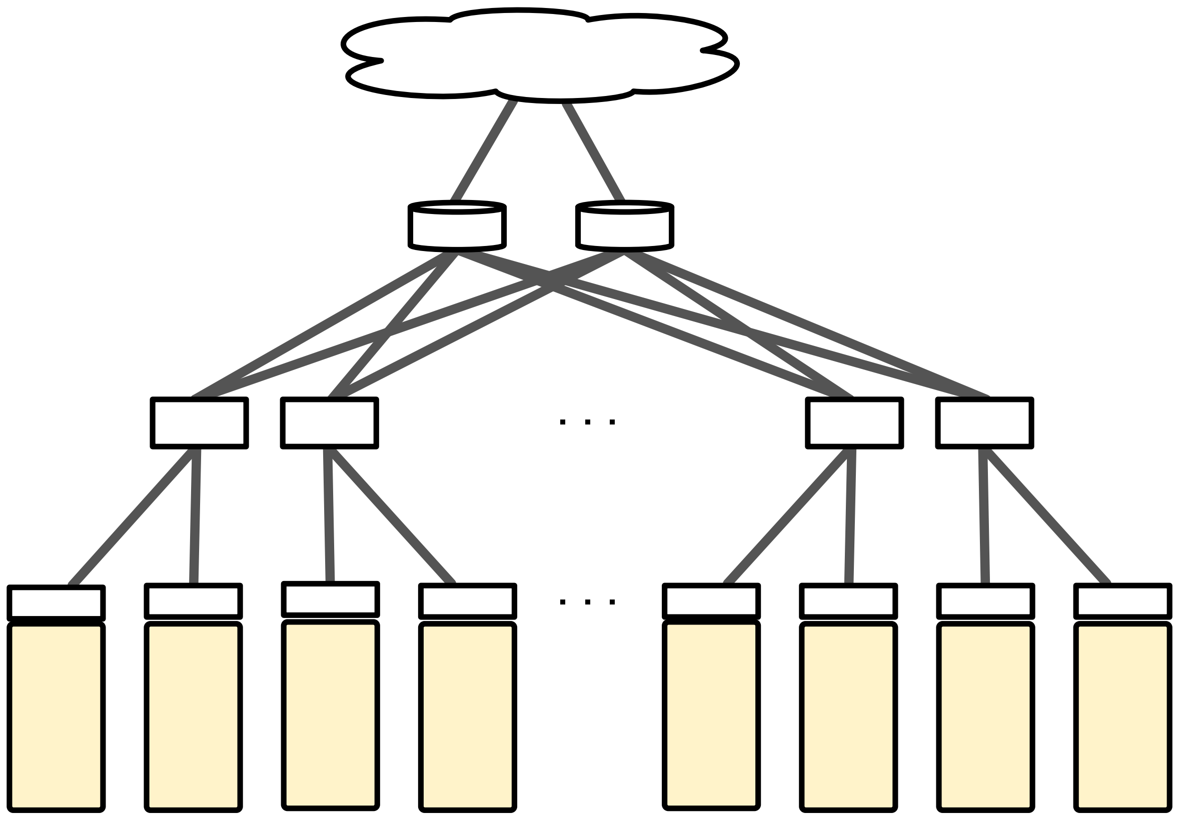
How much performance do we gain with heterogeneous network equipment?

Evaluation

- U of Waterloo School of Computer Science data center as input
- Three scenarios:
  - Upgrading the network (see paper)
  - Expansion by adding servers
  - Greenfield data center

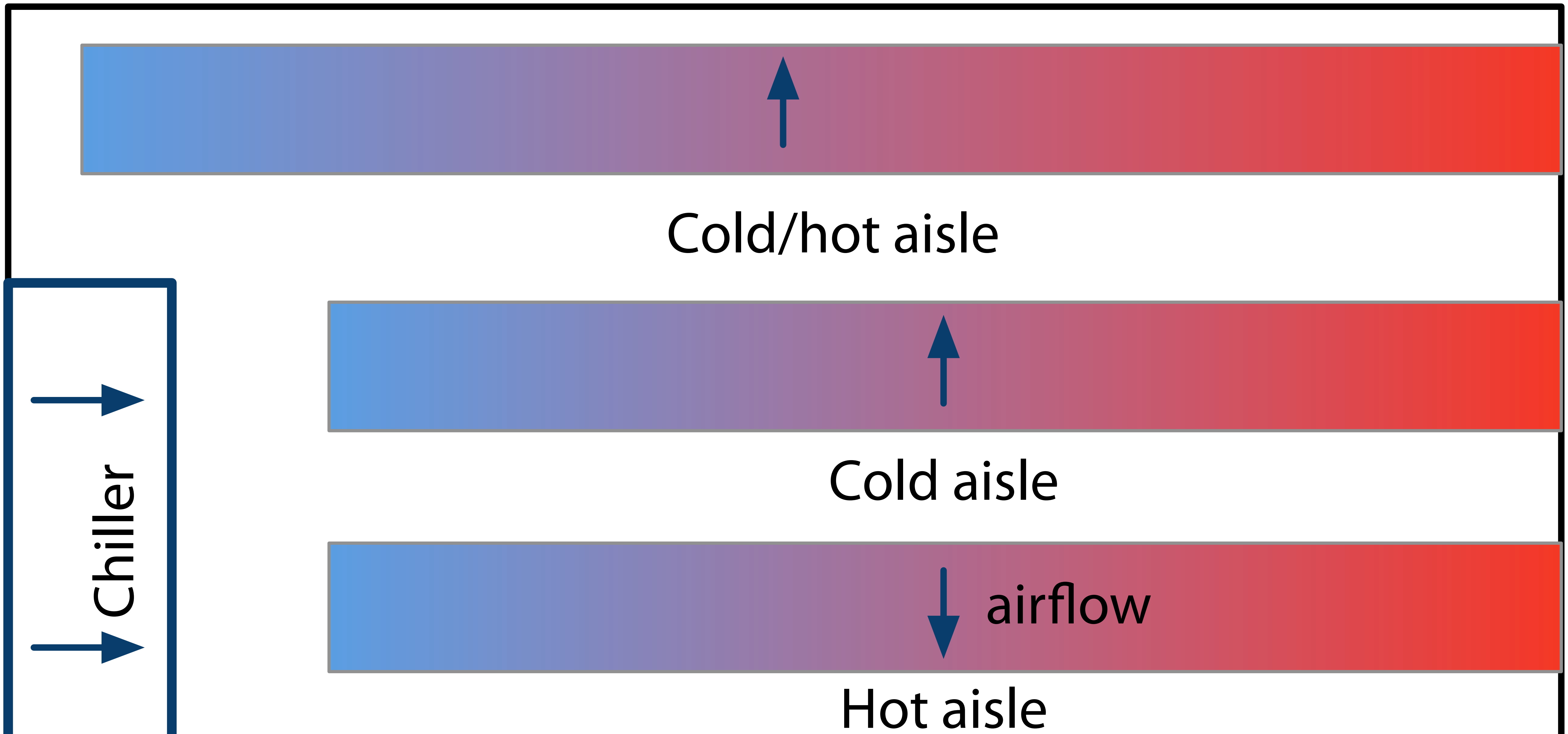
Evaluation: input

- SCS data center topology
  - 19 edge switches, 760 servers
  - Heterogeneous edge switches
  - All aggregation switches are HP 5406 models



Evaluation: input

- The data center handles air poorly.
  So, we add thermal constraints modeling this



Evaluation: cost model

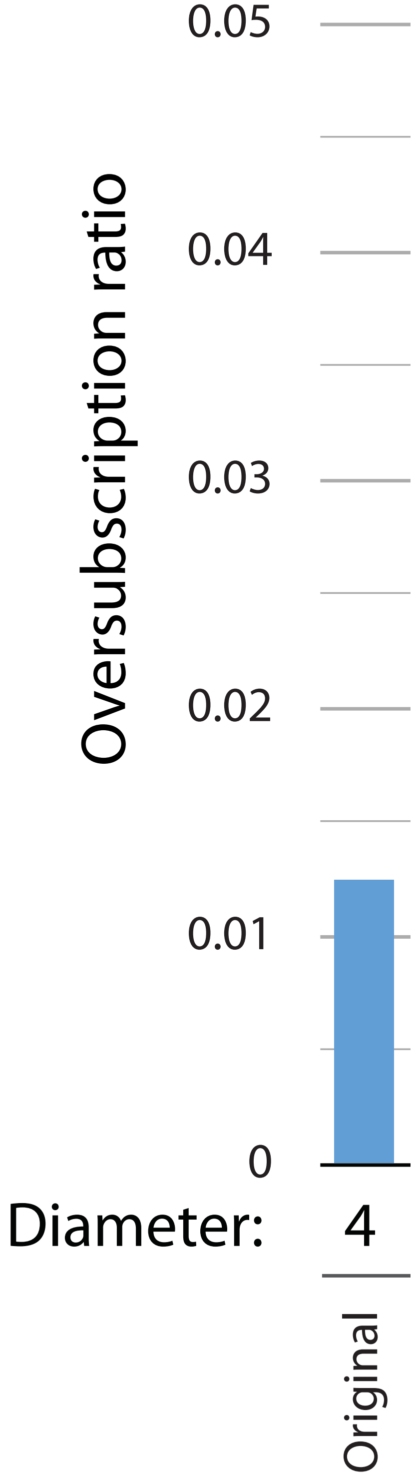
1 Gb ports	10 Gb ports	Watts	Cost ($)
24		100	250
48		150	1,500
48	4	235	5,000
	24	300	6,000
	48	600	10,000
	144	5000	75,000

Rate	Short ($)	Medium ($)	Long ($)
1 Gb	5	10	20
10 Gb	50	100	200
Install cost	10	20	50

Evaluation: comparison methods

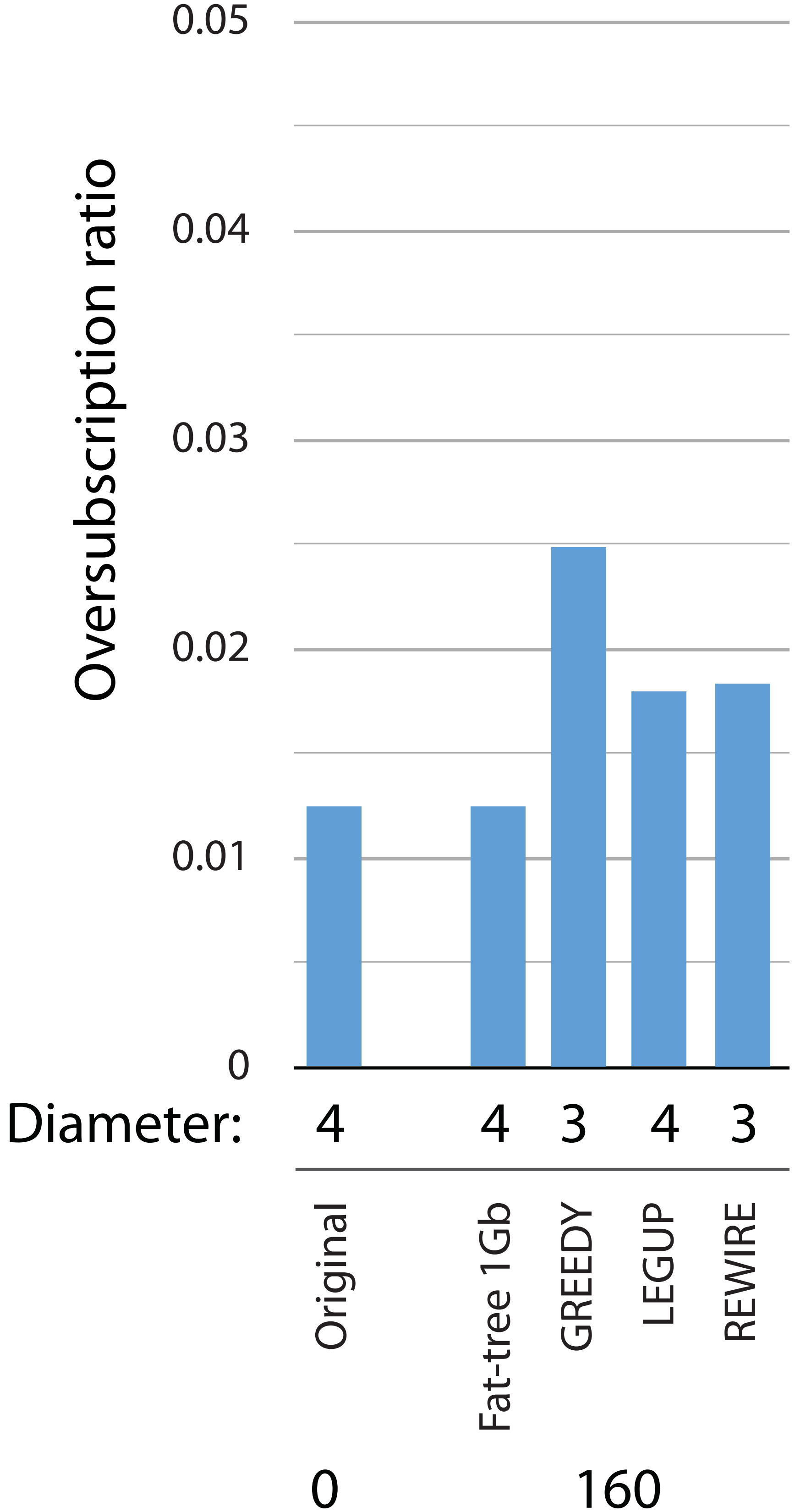
- Generalized fat-tree
  - Bounded best-case performance
- Greedy algorithm
  - Finds link addition that improves performance the most, adds it, and repeats
- Random graph
  - Proposed by Singla et al., HotCloud 2011 as data center network topology

Expanding the Waterloo SCS data center

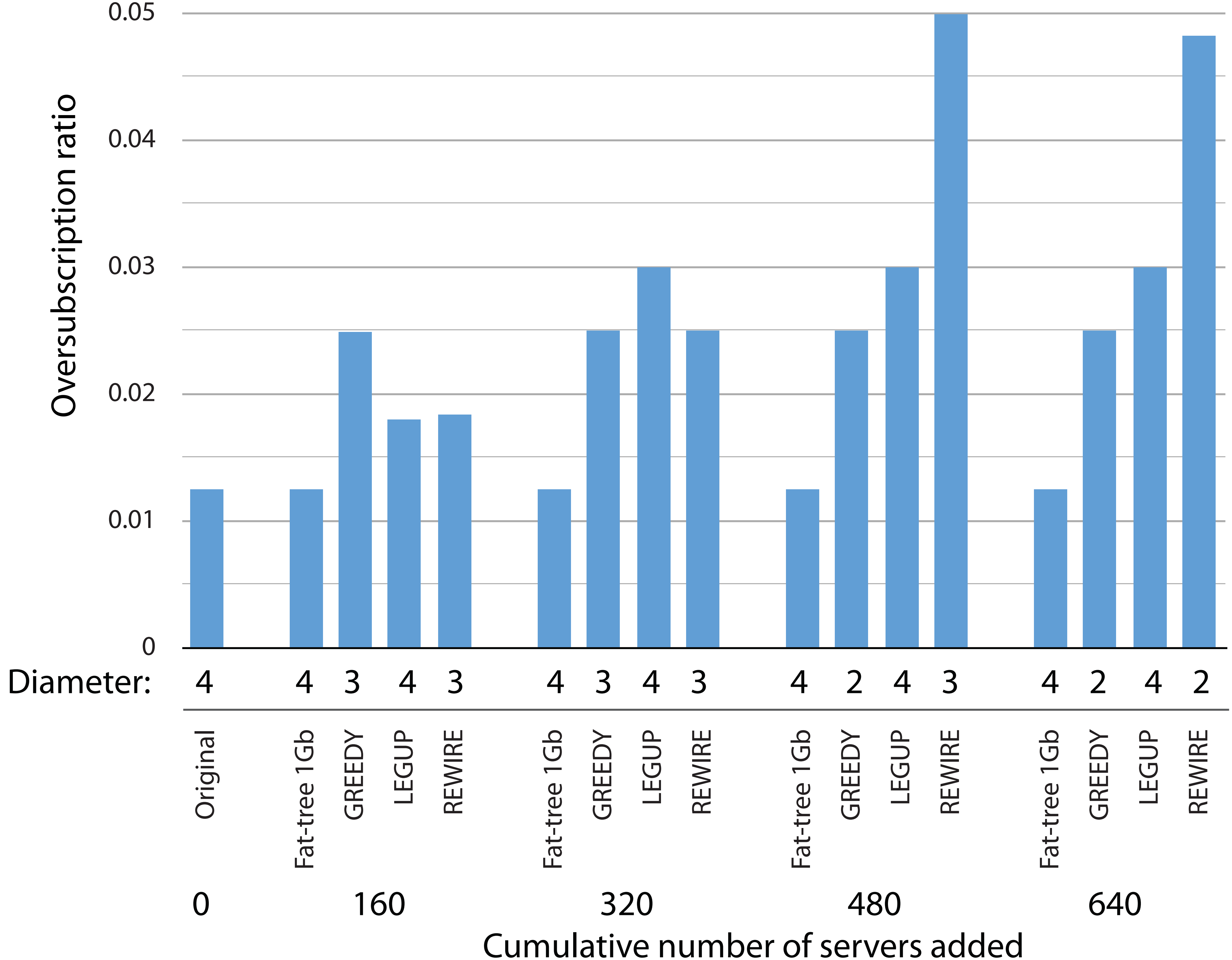


Starting servers = 760

Expanding the Waterloo SCS data center



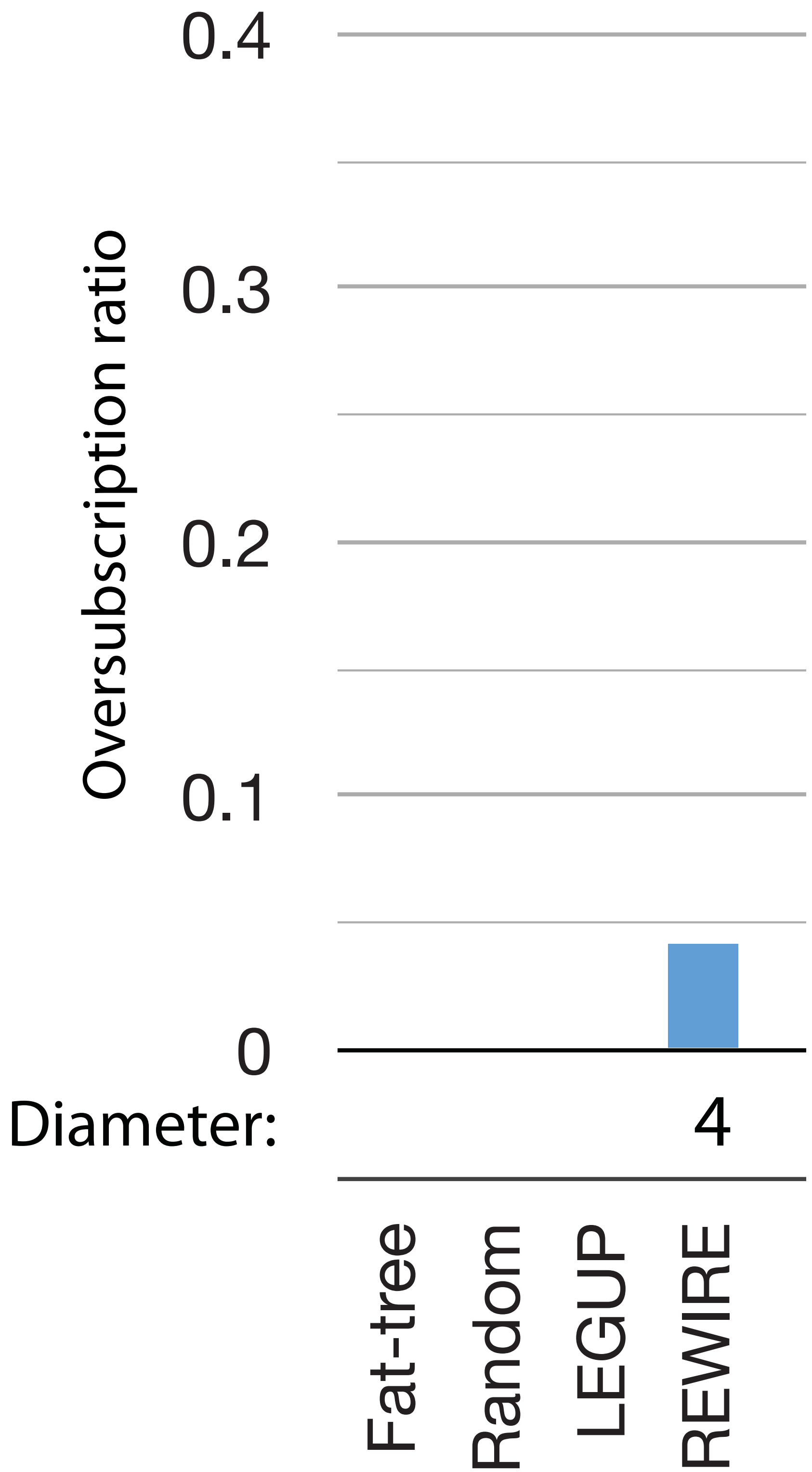
Expanding the Waterloo SCS data center



Greenfield network design

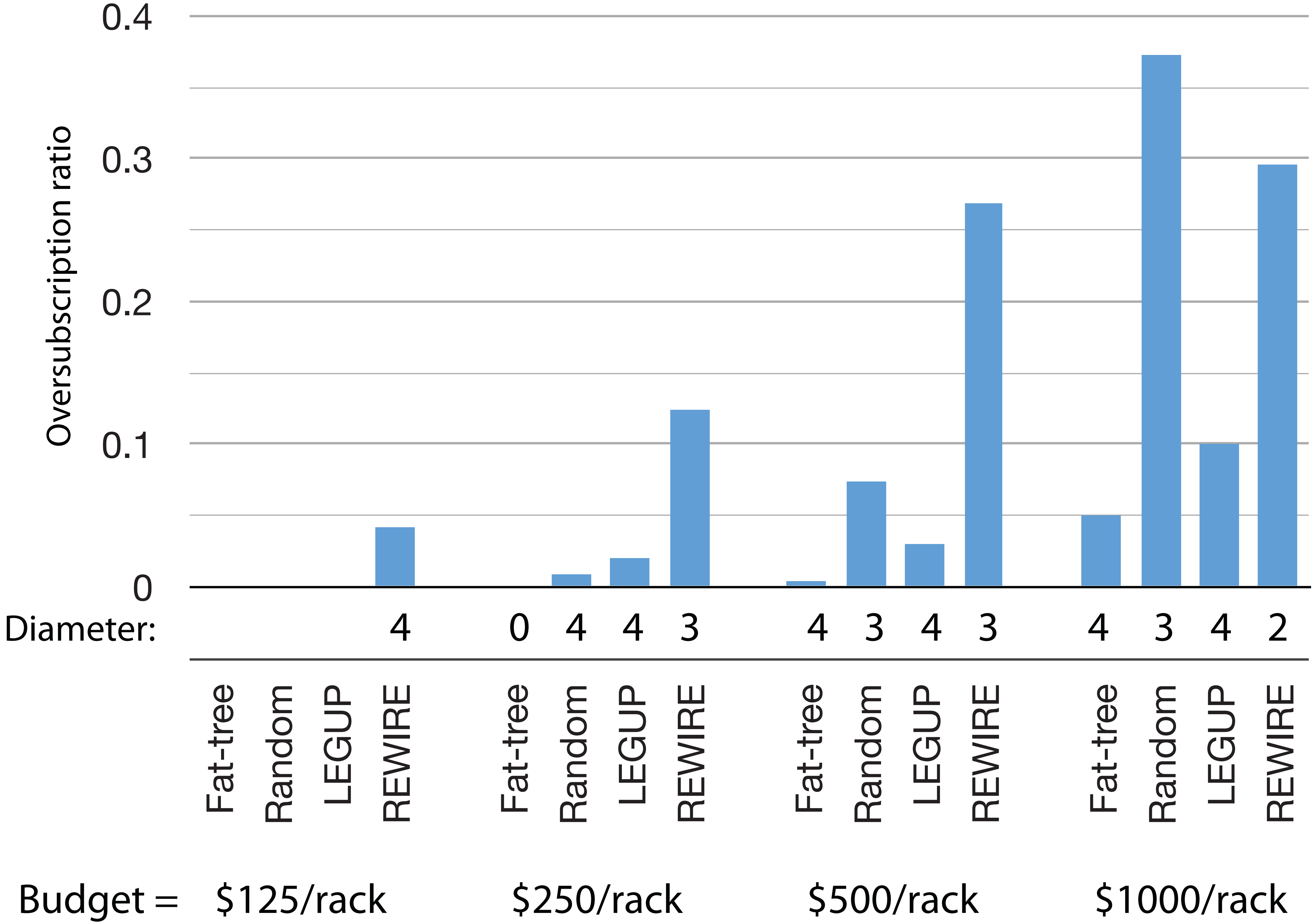
- 1920 servers
- Edges switches have 48 gigabit ports
  - Assume 24 servers per rack

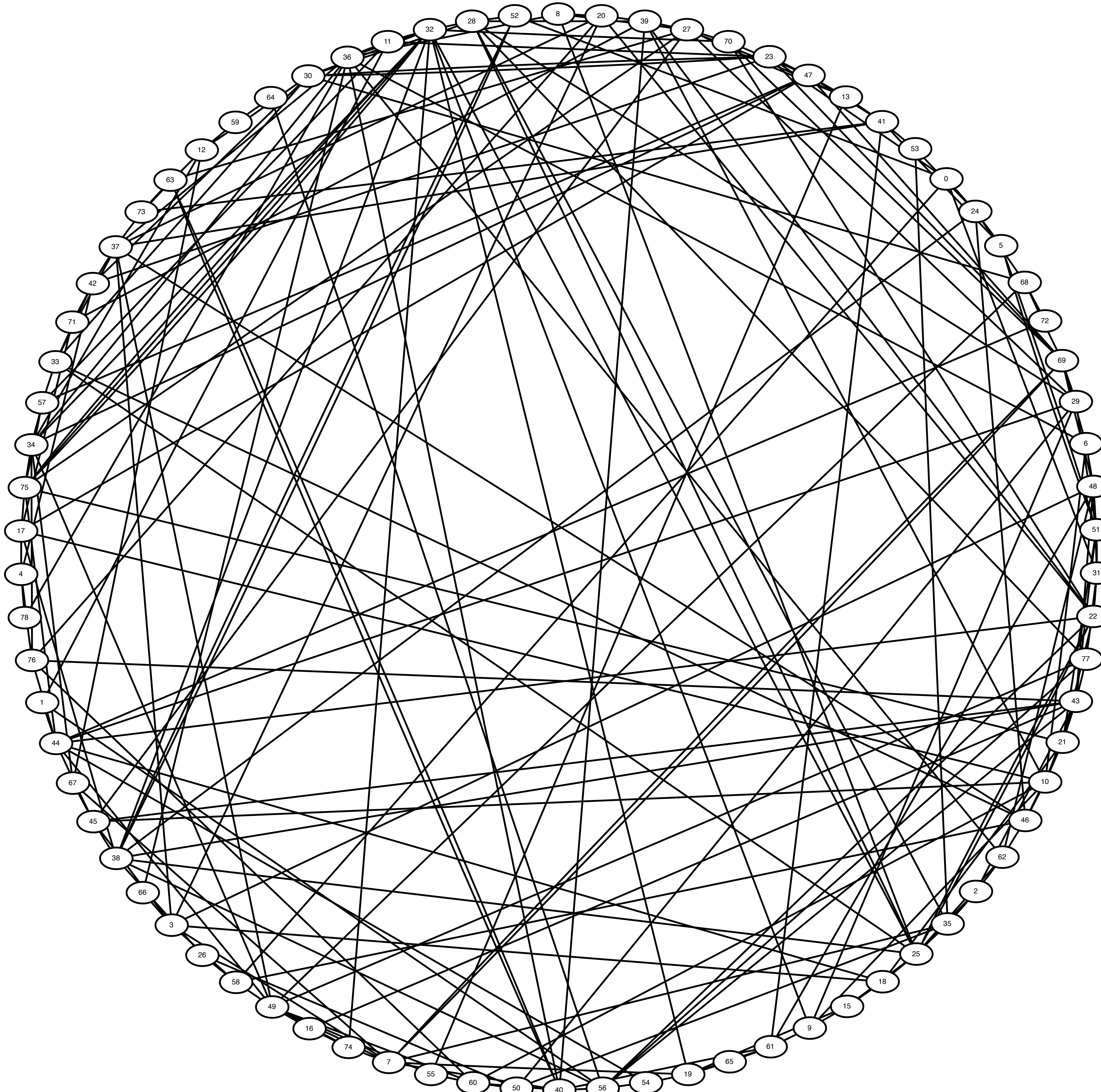
Greenfield network design



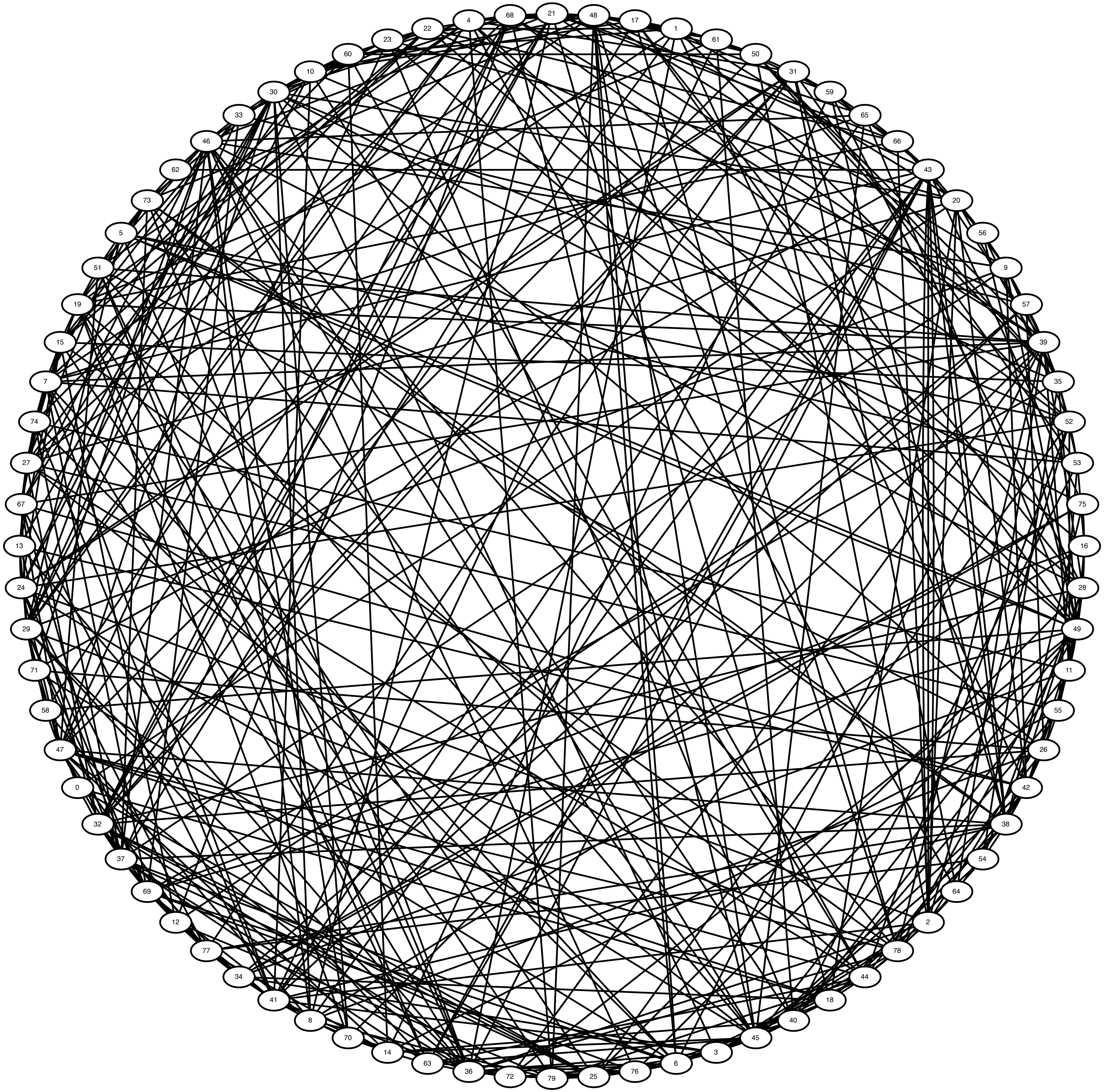
Budget = $125/rack

Greenfield network design





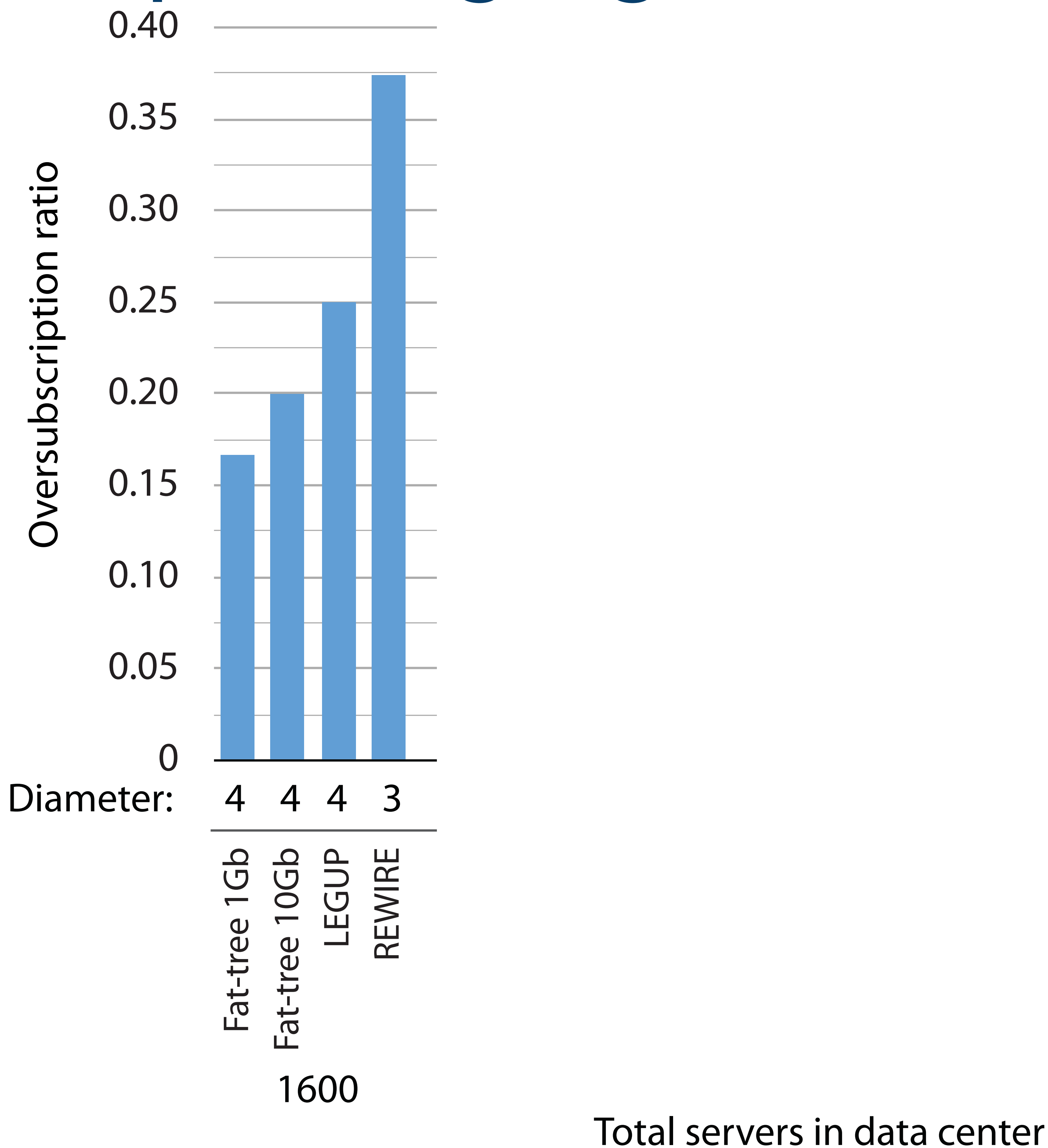
8
20
52
39
28
27
32
70
11
23
36
47
30
13
64
41
59
53
12
0
63
24
73
5
37
68
42
72
71
69
33
29
57
6
34
48
75
51
17
31
4
22
78
77
76
43
1
21
44
10
67
46
45
62
38
2
66
35
3
25
26
18
58
15
49
9
16
61
74
65
7
19
55
54
60
56
50
40


4
68
21
48
17
1
22
23
61
60
50
10
31
30
59
33
65
46
66
62
43
73
20
5
56
51
9
19
57
15
39
7
35
74
52
27
53
67
75
13
16
24
28
29
49
71
11
58
55
47
26
0
42
32
38
37
54
69
64
12
2
77
78
34
44
41
18
8
40
70
45
14
3
63
6
36
76
72
25
79

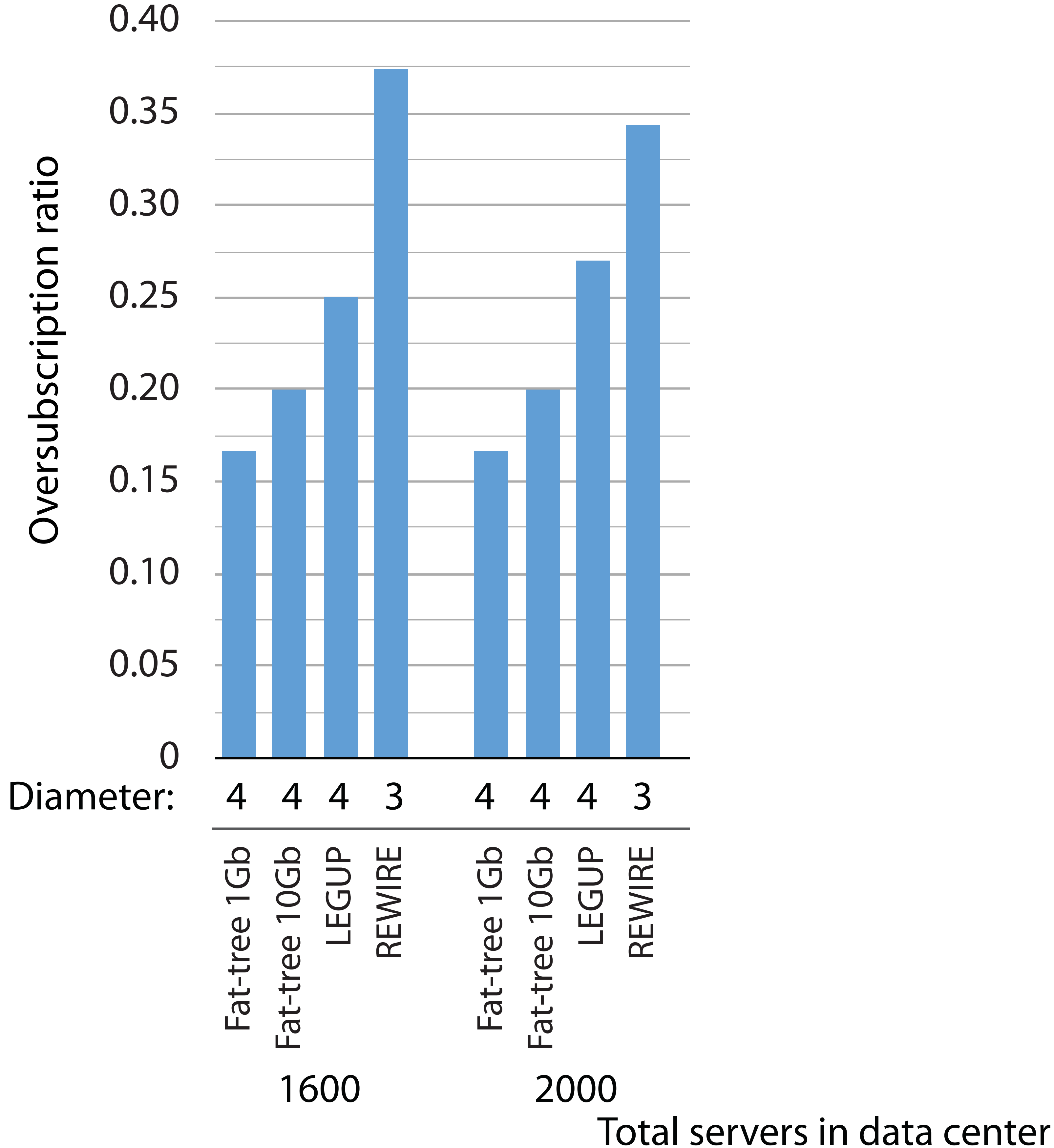
Greenfield network design

- Expanding a greenfield network
- 1600 servers initially
  - Grow by increments of 400 servers (10 racks)
  - $6000/rack budget

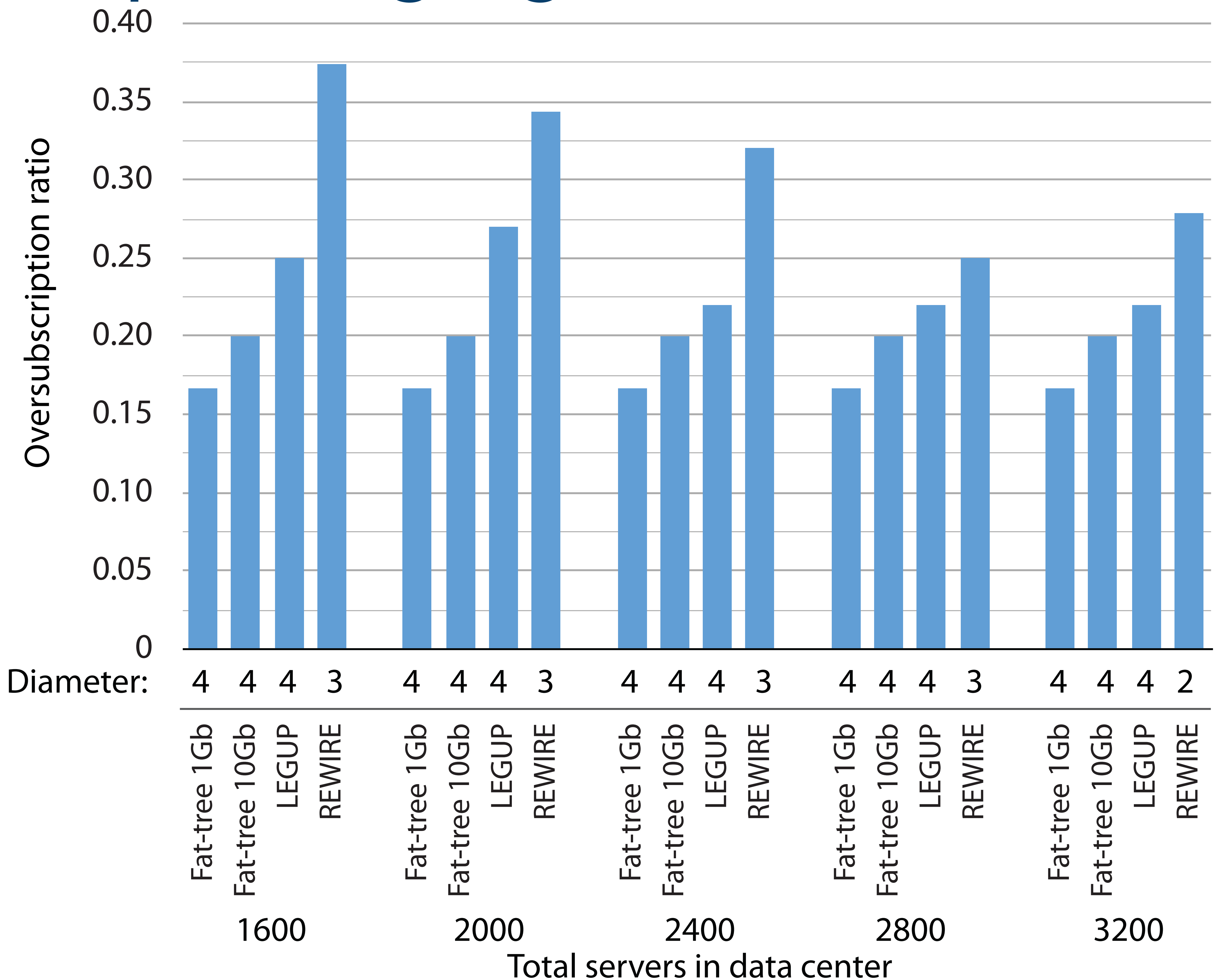
Expanding a greenfield network



Expanding a greenfield network



Expanding a greenfield network



Are unstructured topologies worth it?

- Higher performance
  - Up to 10x more bisection bandwidth than heterogeneous Clos for same cost
  - Lower latency
    (can get 2 hops between racks instead of 4)
- But difficult to manage
  - Cost to build/manage is unclear
  - Need to use Multipath TCP [Raiciu et al. SIGCOMM 2011] or SPAIN [Mudigonda et al., NSDI 2010] to effectively use available bandwidth

REWIRE future work

- Structural constraints on topology
  - Generalize greenfield topology design framework of Mudigonda et al.,USENIX ATC 2011
- Bisection bandwidth computation algorithm numerically unstable
- Scale local search approach to larger networks
- Relationship between spectral gap and bisection bandwidth?

Conclusions

- Best practices are not enough for data center upgrades
- Need theory to understand and effectively build heterogeneous networks
- Implemented LEGUP and REWIRE, optimization algorithms to design heterogeneous DCNs