Reconfigurable Networks: Enablers, Algorithms, Complexity

Stefan Schmid (University of Vienna)

Tutorial @ ITC 2019

Budapest, Hungary

A Great Time to Be a Networking Researcher!

Rhone and Arve Rivers, Switzerland

Credits: George Varghese.

Flexibilities: Along 3 Dimensions

Passau, Germany

Inn, Donau, Ilz

Flexibilities: Along 3 Dimensions

Passau, Germany

Inn, Donau, Ilz

Flexibilities: Along 3 Dimensions

u, Germany

Inn, Donau, Ilz

Flexibilities: Along 3 Dimensions

…, Germany

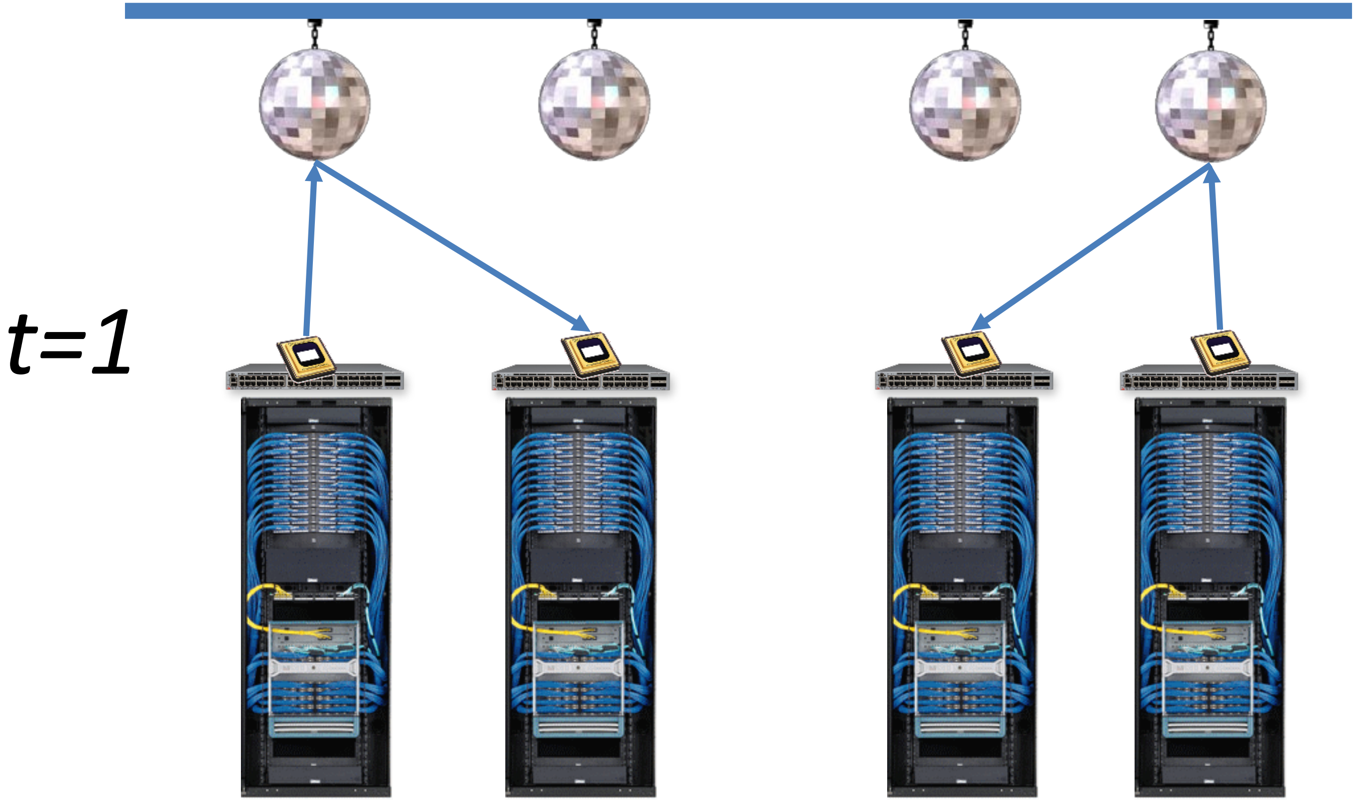
…Inn, Donau, Ilz

Enabling optical technologies for reconfigurable networks

Example: Manya Ghobadi et al.

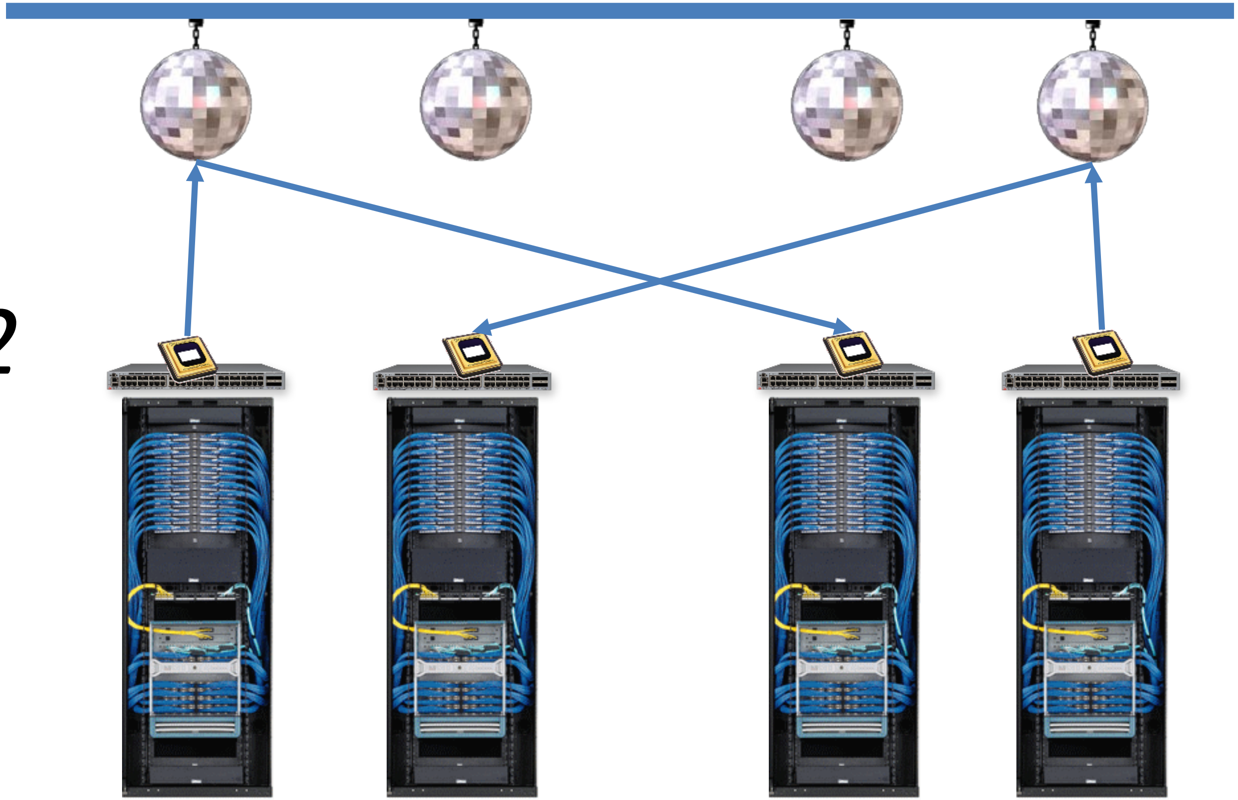
Kudos for some slides!

Example: ProjecToR



- Based on **free-space optics**

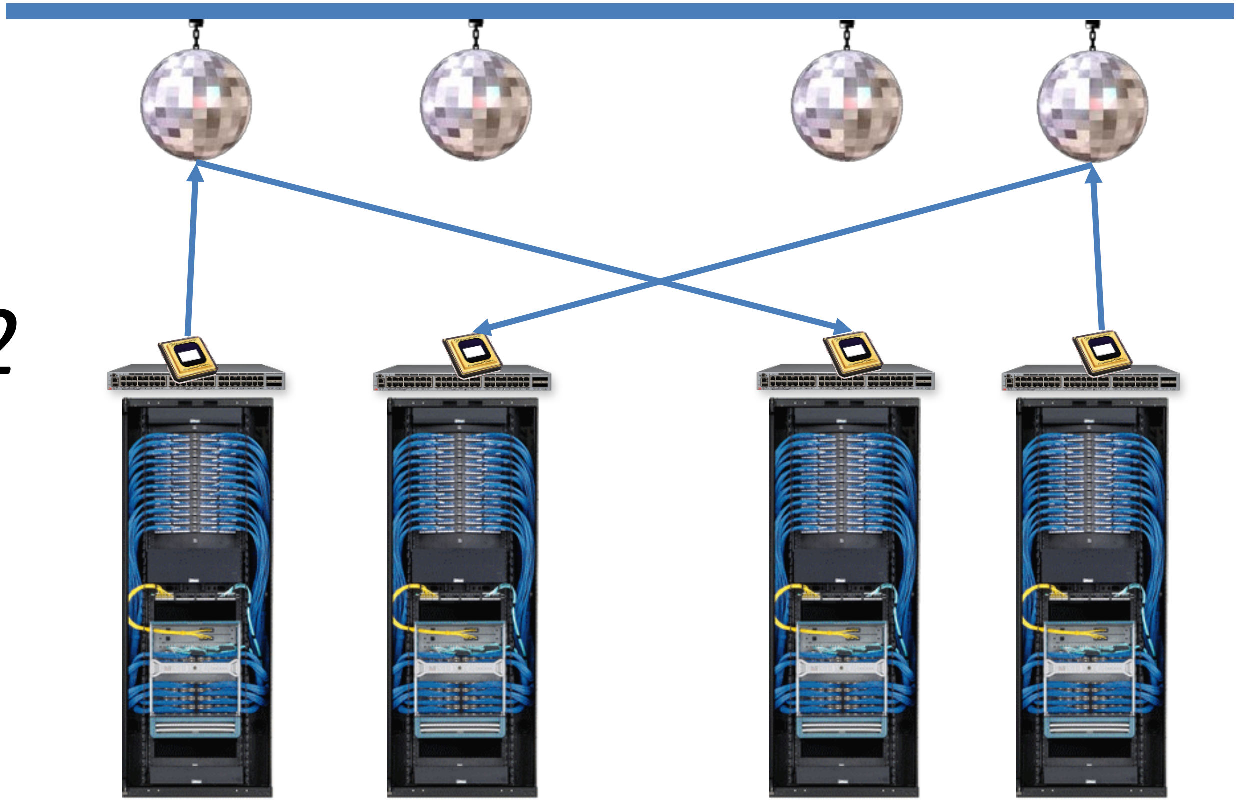
Example: ProjecToR



t=2

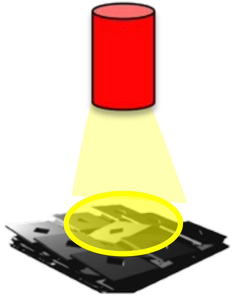
- Based on **free-space optics**

Example: ProjecToR



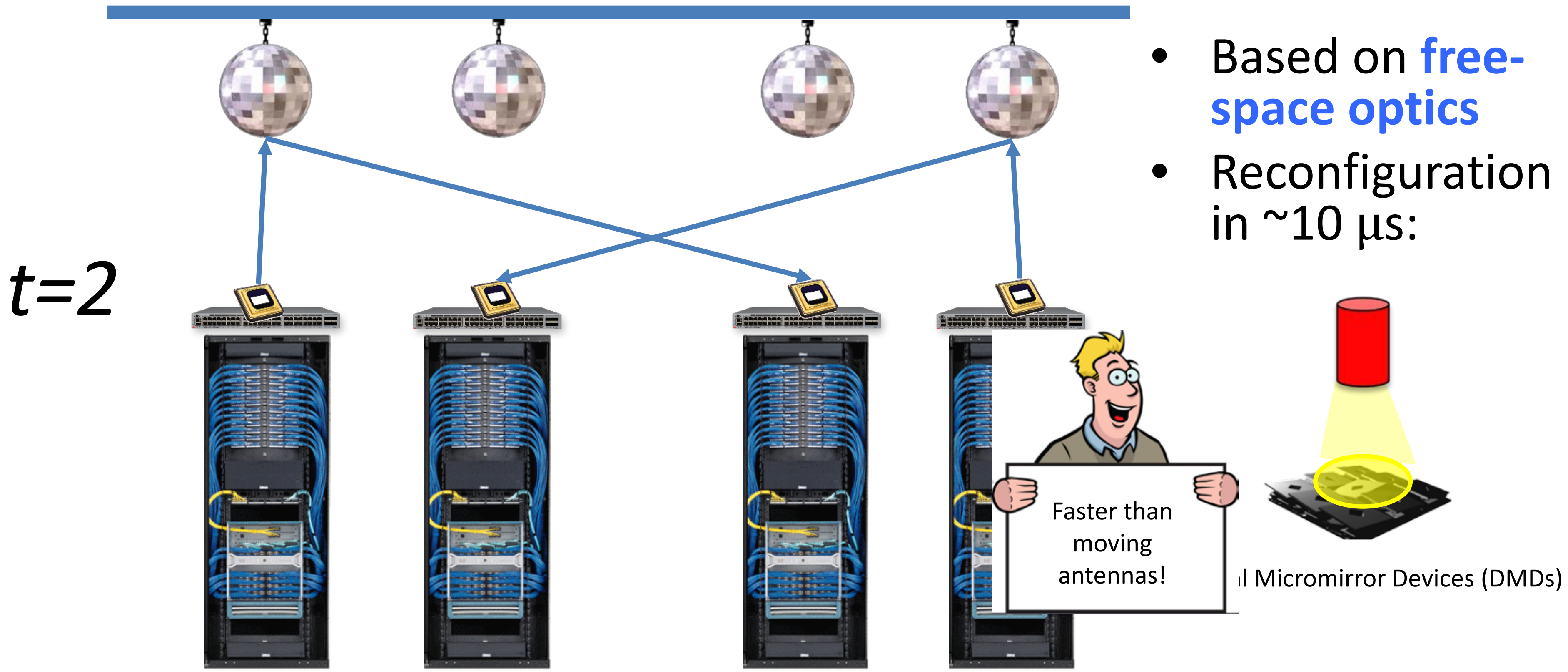
t=2

- Based on **free-space optics**
- Reconfiguration in ~10 μs:



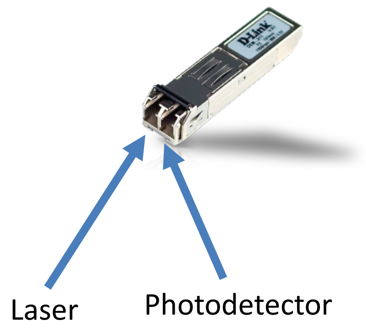
Digital Micromirror Devices (DMDs)

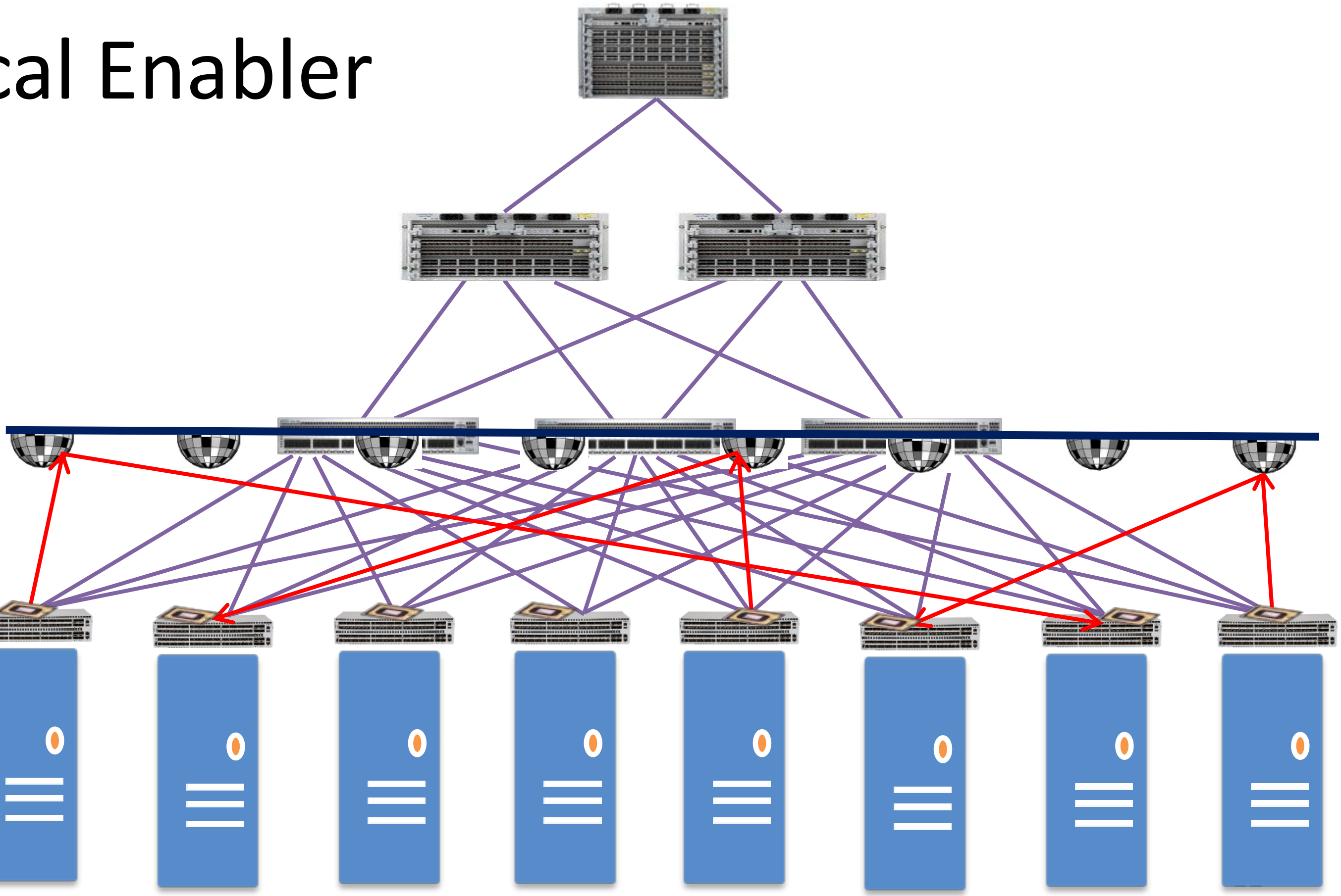
Example: ProjecToR



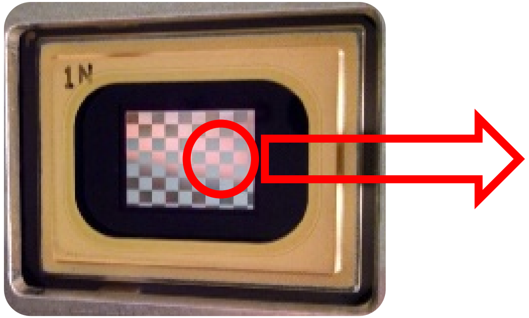
- Based on **free-space optics**
- Reconfiguration in ~10 μs:

ProjecToR in More Details: Technological Enabler

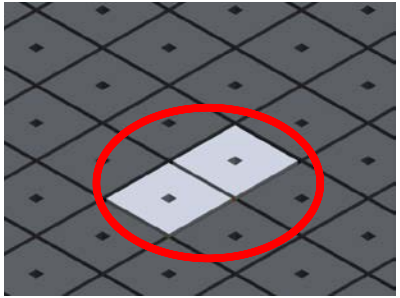




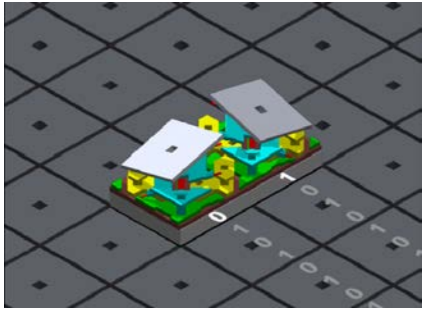
ProjecToR in More Details: DMDs



Texas Instruments



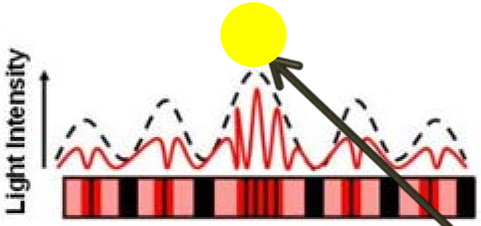
Array of micromirrors

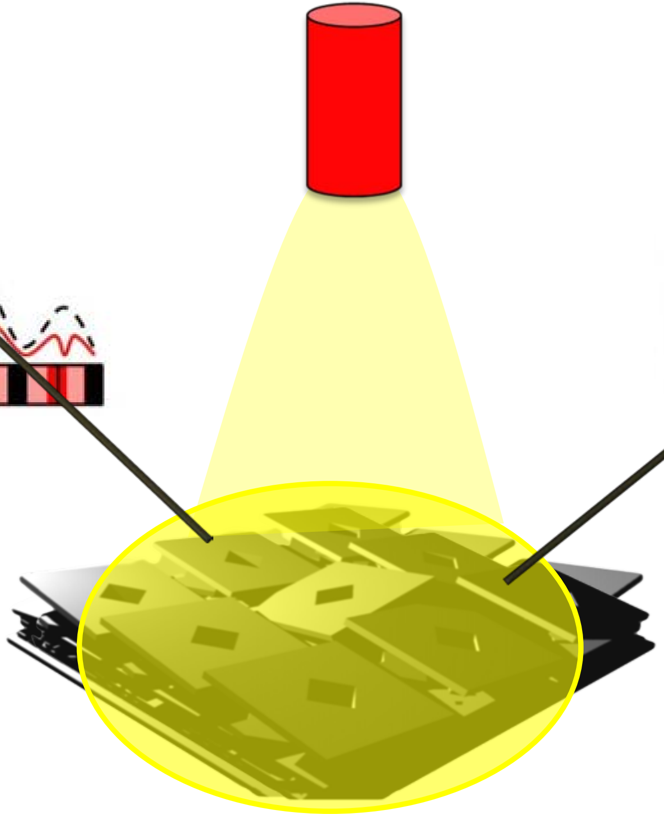


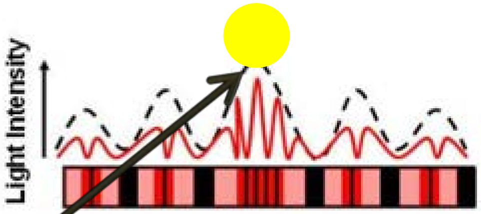
Memory cell

- Each micromirror can be turned on/off
- Essentially a **0/1-image**: e.g., array size 768 x 1024
- Direction of the diffracted light can be finely tuned

ProjecToR in More Details: DMDs to Redirect Light *Fast*



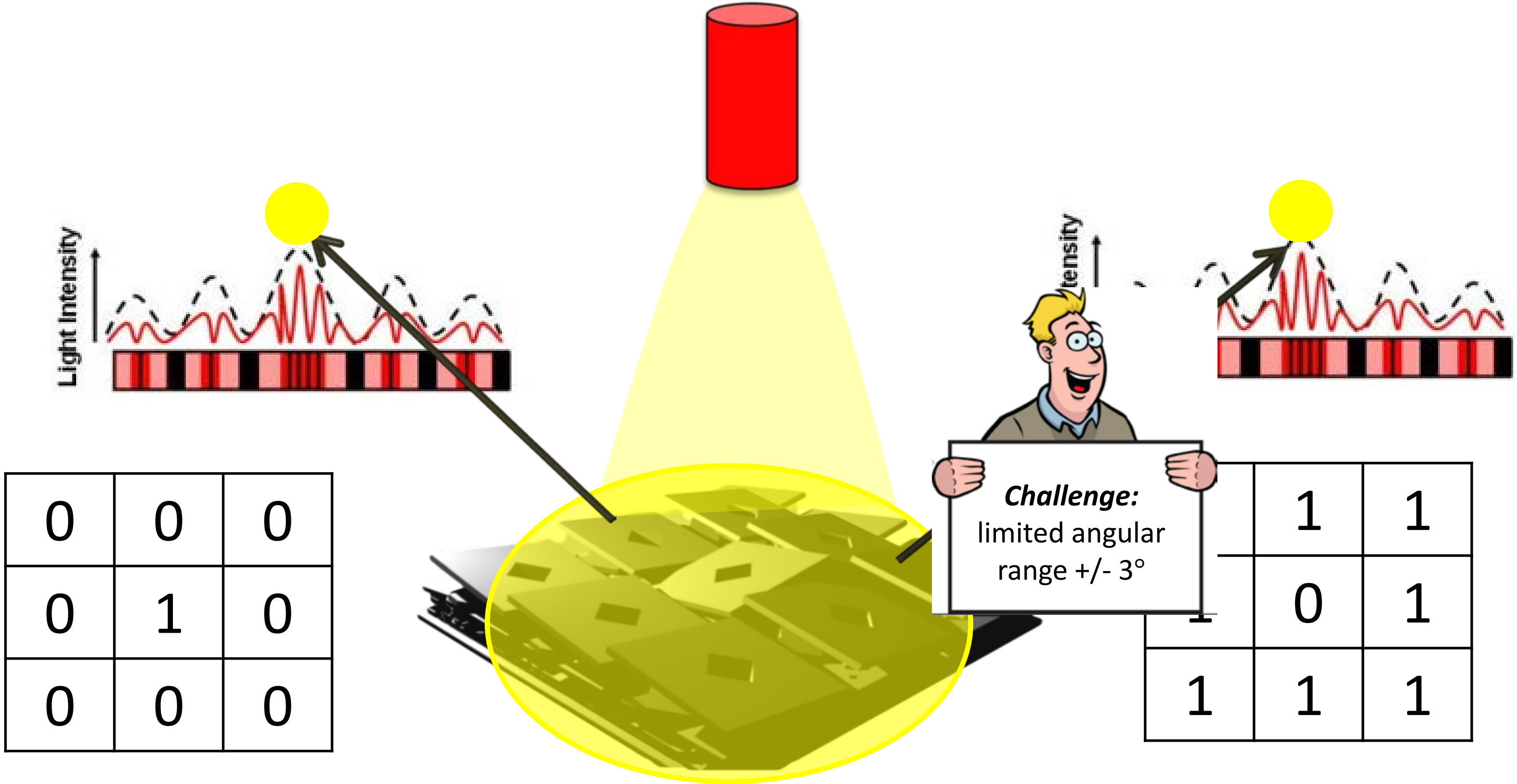




0	0	0
0	1	0
0	0	0

1	1	1
1	0	1
1	1	1

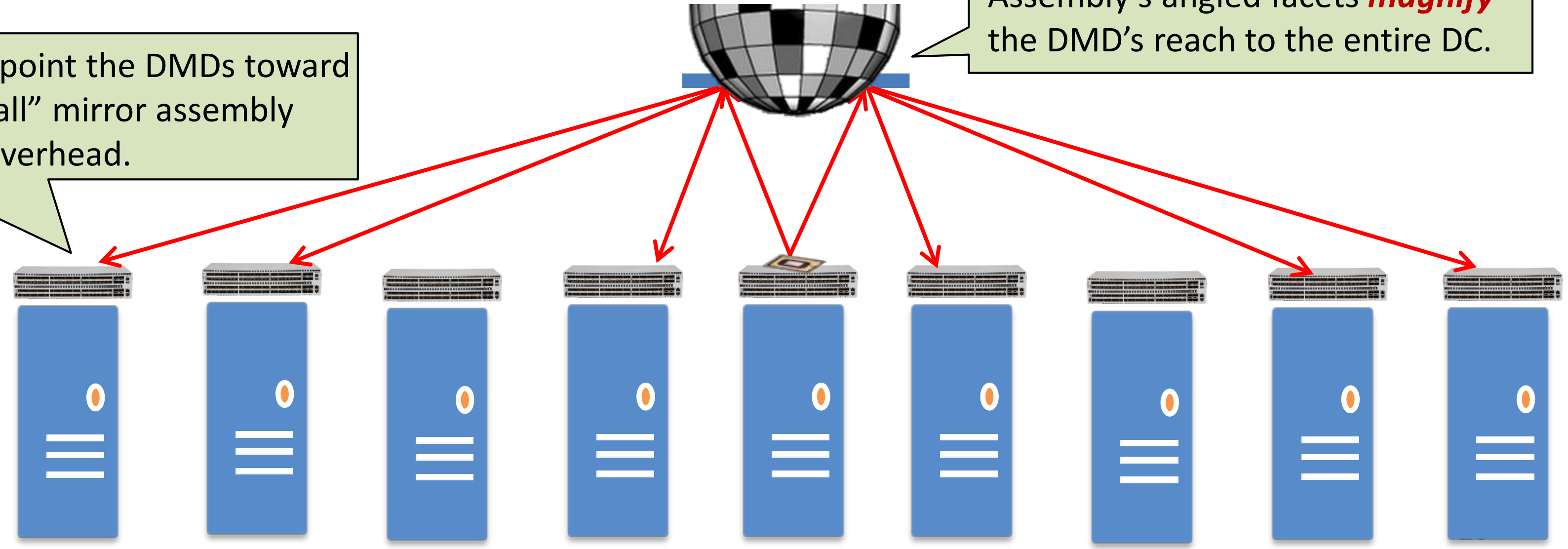
ProjecToR in More Details:
DMDs to Redirect Light *Fast*



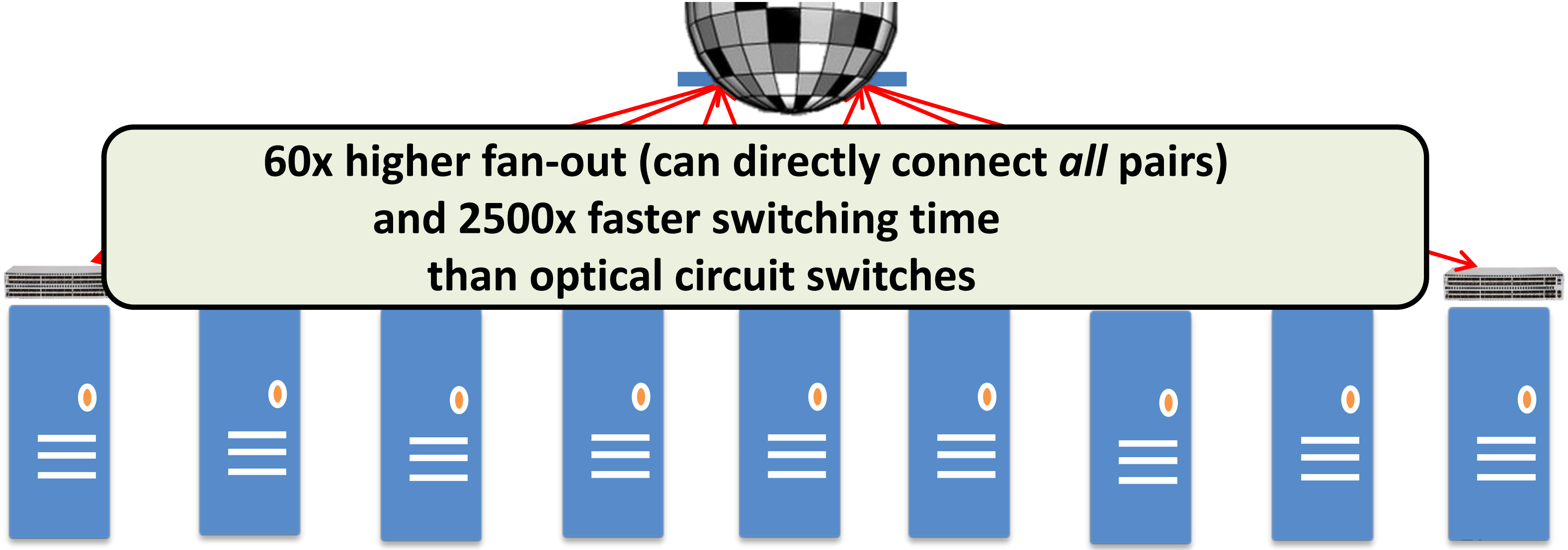
ProjecToR in More Details: Coupling DMDs with angled mirrors

Coupling: point the DMDs toward a "disco-ball" mirror assembly installed overhead.

Assembly's angled facets ***magnify*** the DMD's reach to the entire DC.

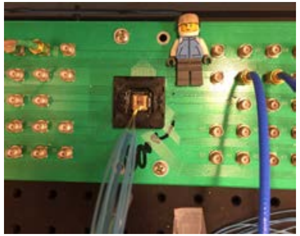


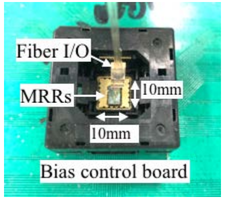
ProjecToR in More Details:
Coupling DMDs with angled mirrors



60x higher fan-out (can directly connect *all* pairs)
and 2500x faster switching time
than optical circuit switches

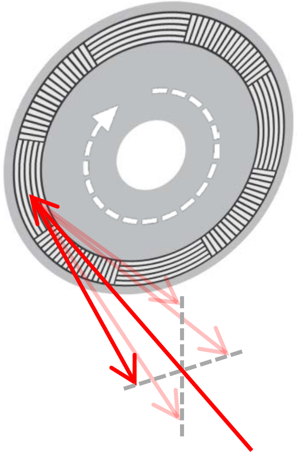
Other Technologies





Based on silicon photonics

2-NEMS



Rotating disks

Further reading:
Wade et al., A Bandwidth-Dense, Low Power Electronic-Photonic Platform and Architecture for Multi-Tbps Optical I/O [OFC'18]
Porter et al., "Integrating Microsecond Circuit Switching into the Data Center", Sigcomm'13

Timeline

Reconfiguration time: from milliseconds ***to microseconds*** (and decentralized).

Survey of Reconfigurable Data Center Networks. Foerster and Schmid. SIGACT News, 2019.

- 2009
  - *Flyways* [51]: Steerable antennas (narrow beamwidth at 60 GHz [78]) to serve hotspots
- 2010
  - *Helios* [33]/*c-Through* [98, 99]: Hybrid switch architecture, maximum matching (Edmond's algorithm [30]), single-hop reconfigurable connections ($O(10)ms$ reconfiguration time).
  - *Proteus* [21, 89]: k reconfigurable connections per ToR, multi-hop path stitching, multi-hop reconfigurable connections (weighted b-matching [69], edge-exchanges for connectivity [72], wavelength assignment via edge-coloring [67] on multigraphs)
- 2011
  - Extension of *Flyways* [51] to better handle practical concerns such as stability and interference for 60GHz links, along with greedy heuristics for dynamic link placement [45]
- 2012
  - *Mirror Mirror on the ceiling* [106]: 3D-beamforming (60 Ghz wireless), signals bounce off the ceiling
- 2013
  - *Mordia* [31, 32, 77]: Traffic matrix scheduling, matrix decomposition (Birkhoff-von-Neumann (BvN) [18, 97]), fiber ring structure with wavelengths ($O(10)\mu s$ reconfiguration time)
  - *SplayNets* [6, 76, 82]: Fine-grained and online reconfigurations in the spirit of self-adjusting datastructures (all links are reconfigurable), aiming to strike a balance between short route lengths and reconfiguration costs
- 2014
  - *REACToR* [56]: Buffer burst of packets at end-hosts until circuit provisioned, employs [77]
  - *Firefly* [14] Combination of Free Space Optics and Galvo/switchable mirrors (small fan-out)
- 2015
  - *Solstice* [57]: Greedy perfect matching based hybrid scheduling heuristic that outperforms BvN [77]
  - Designs for optical switches with a reconfiguration latency of $O(10)ns$ [3]
- 2016
  - *ProjecToR* [39]: Distributed Free Space Optics with digital micromirrors (high fan-out) [38] (Stable Matching [26]), goal of (starvation-free) low latency
  - *Eclipse* [95, 96]: $(1 - 1/e^{(1-\varepsilon)})$-approximation for throughput in traffic matrix scheduling (single-hop reconfigurable connections, hybrid switch architecture), outperforms heuristics in [57]
- 2017
  - *DAN* [7, 8, 11, 12]: Demand-aware networks based on reconfigurable links only and optimized for a demand snapshot, to minimized average route length and/or minimize load
  - *MegaSwitch* [23]: Non-blocking circuits over multiple fiber rings (stacking rings in [77] doesn't suffice)
  - *Rotornet* [63]: Oblivious cyclical reconfiguration w. selector switches [64] (Valiant load balancing [94])
  - *Tale of Two Topologies* [105]: Convert locally between Clos [24] topology and random graphs [87, 88]
- 2018
  - *DeepConf* [81]/*xWeaver* [102]: Machine learning approaches for topology reconfiguration
- 2019
  - Complexity classifications for weighted average path lengths in reconfigurable topologies [34, 35, 36]
  - *ReNet* [13] and *Push-Down-Trees* [9] providing statically and dynamically optimal reconfigurations
  - *DisSplayNets* [75]: fully decentralized *SplayNets*
  - *Opera* [60]: Maintaining expander-based topologies under (oblivious) reconfiguration

Such Fast Reconfigurability Enables Demand-Aware Networks (DANs)!

Why are self-adjusting networks useful?

	A	B	C	D
A	0	3	3	3
B	3	0	3	3
C	3	3	0	3
D	3	3	3	0

	A	B	C	D
A	0	0 / 6	0 / 6	0
B	0	0	0	0 / 7
C	0 / 8	0	0 / 6	0 / 7
D	0	0 / 12	0 / 8	0

In theory: traffic matrix uniform and static

In practice: **skewed** and **dynamic**

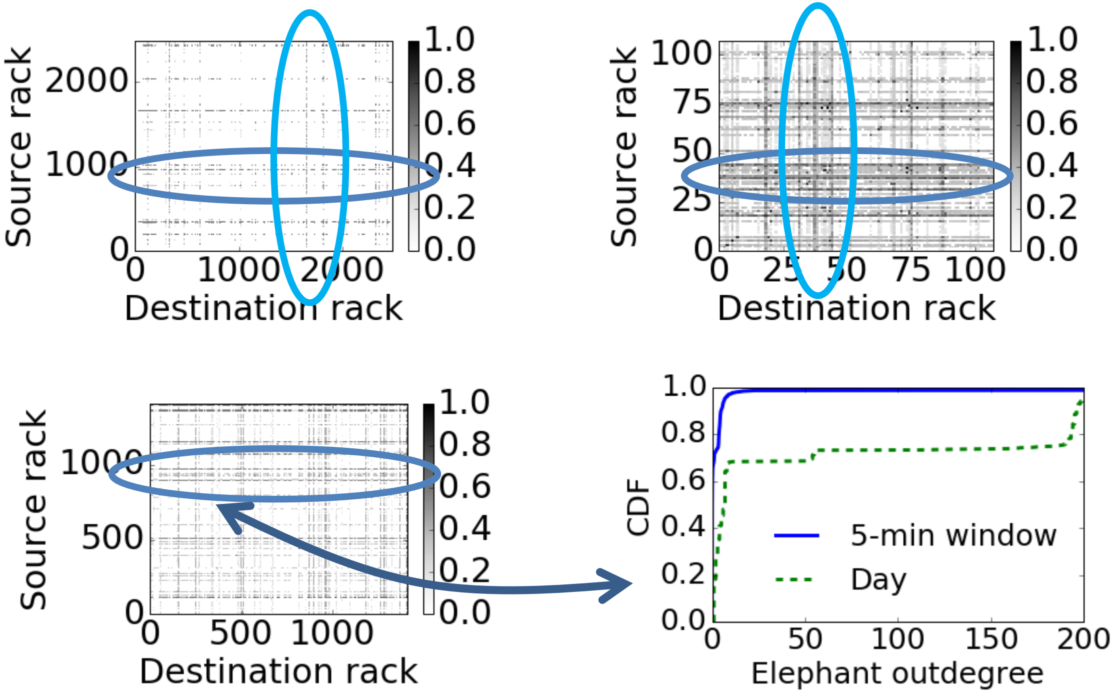
Empirical Motivation

Observation 1:

- Many rack pairs exchange ***little traffic***
- Only some ***hot rack pairs*** are active

Observation 2:

- Some source racks send large amounts of traffic ***to many other racks***



Microsoft data: 200K servers across 4 production clusters, cluster sizes: 100 - 2500 racks.
Mix of workloads: MapReduce-type jobs, index builders, database and storage systems.

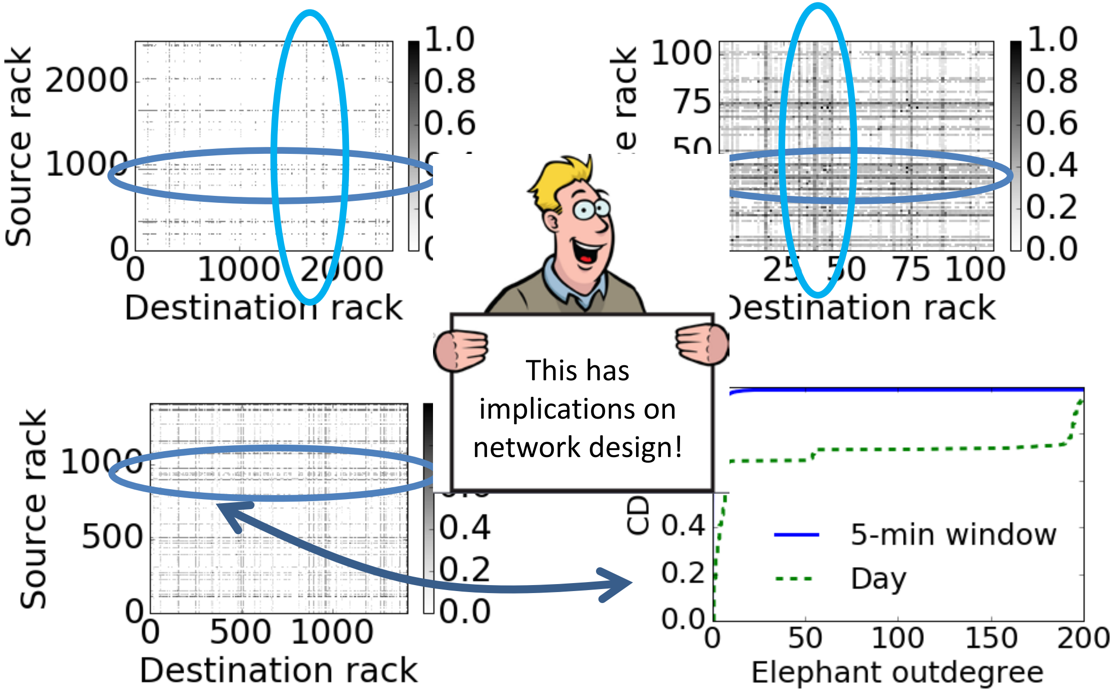
Empirical Motivation

Observation 1:

- Many rack pairs exchange ***little traffic***
- Only some ***hot rack pairs*** are active

Observation 2:

- Some source racks send large amounts of traffic ***to many other racks***



Microsoft data: 200K servers across 4 production clusters, cluster sizes: 100 - 2500 racks.
Mix of workloads: MapReduce-type jobs, index builders, database and storage systems.

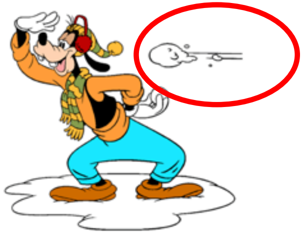
So what…?

You: I invented a great new reconfigurable network which allows to self-adjust to the demand it serves!

Boss: Okay, so how much better is your demand-aware network really compared to demand-oblivious networks!?

You: hmm…

A Simple Answer

Demand-Oblivious Networks = 

The *CS* Answer

- It depends…

The *CS* Answer

As always in computer science! ☺

- It depends…

The *CS* Answer

As always in computer science! ☺

- It depends…
- … on the demand!

We need **metrics**!

Roadmap

- Entropy: A metric for demand-aware networks?
  - Intuition
  - A lower bound
  - Algorithms achieving entropy bounds
- From static to dynamic demand-aware networks
  - Empirical motivation
  - A connection to self-adjusting datastructures

Roadmap

- Entropy: A metric for demand-aware networks?
  - Intuition
  - A lower bound
  - Algorithms achieving entropy bounds
- From static to dynamic demand-aware networks
  - Empirical motivation
  - A connection to self-adjusting datastructures

A Simple Example

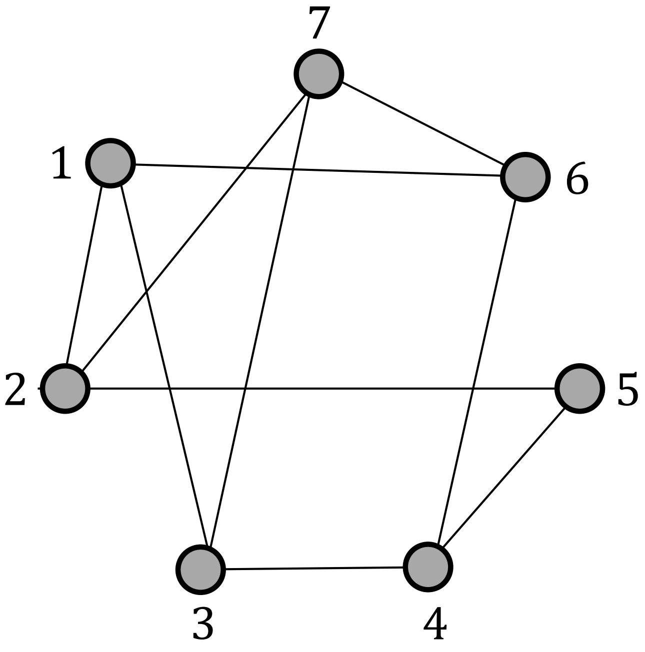
Input: Workload

Destinations

Sources	1	2	3	4	5	6	7
1	0	$\frac{2}{65}$	$\frac{1}{13}$	$\frac{1}{65}$	$\frac{1}{65}$	$\frac{2}{65}$	$\frac{3}{65}$
2	$\frac{2}{65}$	0	$\frac{1}{65}$	0	0	0	$\frac{2}{65}$
3	$\frac{1}{13}$	$\frac{1}{65}$	0	$\frac{2}{65}$	0	0	$\frac{1}{13}$
4	$\frac{1}{65}$	0	$\frac{2}{65}$	0	$\frac{4}{65}$	0	0
5	$\frac{1}{65}$	0	$\frac{3}{65}$	$\frac{4}{65}$	0	0	0
6	$\frac{2}{65}$	0	0	0	0	0	$\frac{3}{65}$
7	$\frac{3}{65}$	$\frac{2}{65}$	$\frac{1}{13}$	0	0	$\frac{3}{65}$	0

$\mathcal{D}$

Output: Constant-Degree DAN



N

Input:
Workload

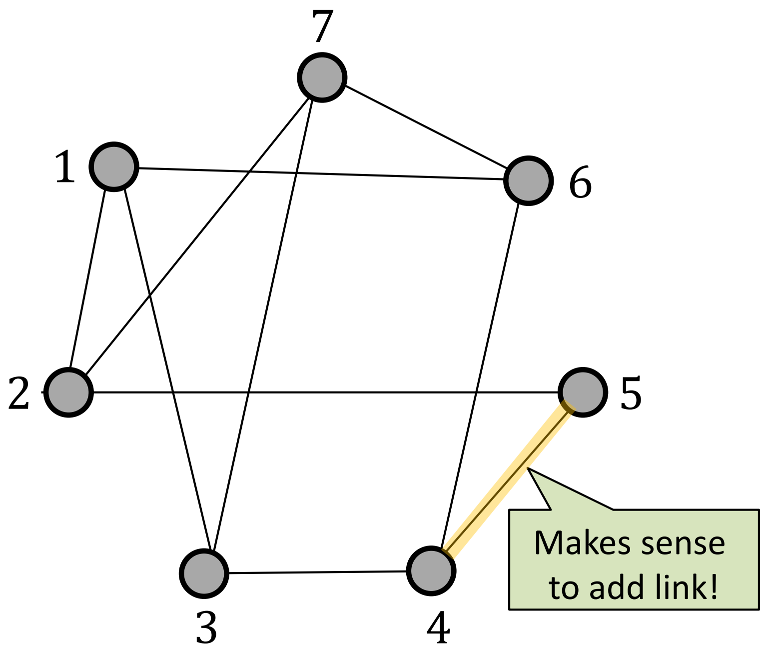
Destinations

Sources	1	2	3	4	5	6	7
1	0	$\frac{2}{65}$	$\frac{1}{13}$	$\frac{1}{65}$	$\frac{1}{65}$	$\frac{2}{65}$	$\frac{3}{65}$
2	$\frac{2}{65}$	0	$\frac{1}{65}$	0	0	0	$\frac{2}{65}$
3	$\frac{1}{13}$	$\frac{1}{65}$	0	$\frac{2}{65}$	[illegible]	[illegible]	[illegible]
4	$\frac{1}{65}$	0	$\frac{2}{65}$	0	$\frac{4}{65}$	0	0
5	$\frac{1}{65}$	0	$\frac{3}{65}$	$\frac{4}{65}$	0	0	0
6	$\frac{2}{65}$	0	0	0	0	0	$\frac{3}{65}$
7	$\frac{3}{65}$	$\frac{2}{65}$	$\frac{1}{13}$	0	0	$\frac{3}{65}$	0

Much from 4 to 5.

$\mathcal{D}$

Output:
Constant-Degree DAN



N

Input: Workload

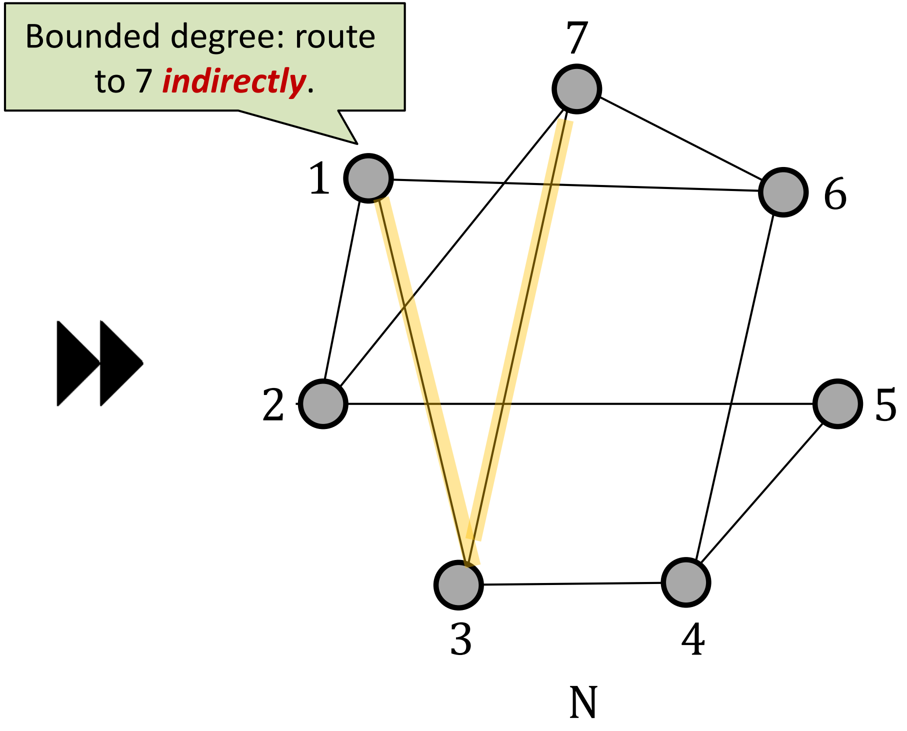
1 communicates to many.

Destinations

Sources	1	2	3	4	5	6	7
1	0	$\frac{2}{65}$	$\frac{1}{13}$	$\frac{1}{65}$	$\frac{1}{65}$	$\frac{2}{65}$	$\frac{3}{65}$
2	$\frac{2}{65}$	0	$\frac{1}{65}$	0	0	0	$\frac{2}{65}$
3	$\frac{1}{13}$	$\frac{1}{65}$	0	$\frac{2}{65}$	0	0	$\frac{1}{13}$
4	$\frac{1}{65}$	0	$\frac{2}{65}$	0	$\frac{4}{65}$	0	0
5	$\frac{1}{65}$	0	$\frac{3}{65}$	$\frac{4}{65}$	0	0	0
6	$\frac{2}{65}$	0	0	0	0	0	$\frac{3}{65}$
7	$\frac{3}{65}$	$\frac{2}{65}$	$\frac{1}{13}$	0	0	$\frac{3}{65}$	0

$\mathcal{D}$

Output: Constant-Degree DAN



Input: Workload

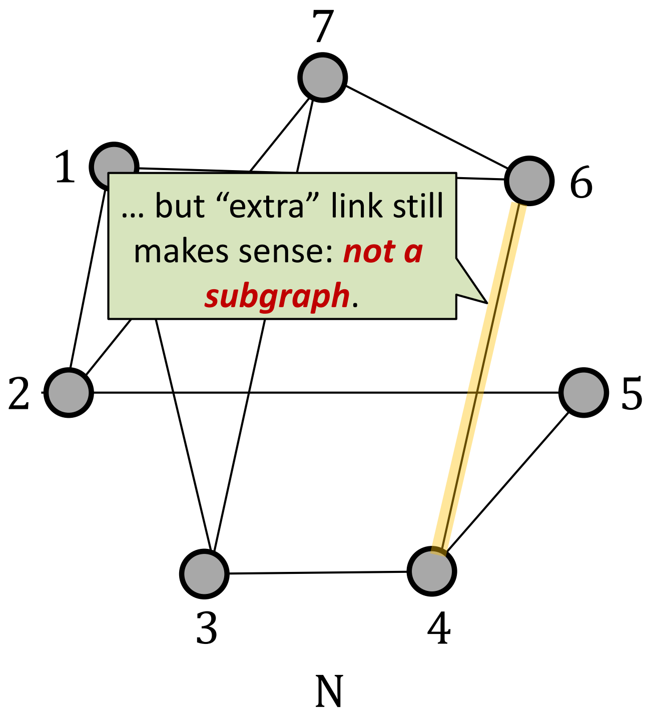
Output: Constant-Degree DAN

Destinations

Sources	1	2	3	4	5	6	7
1	0	$\frac{2}{65}$	$\frac{1}{13}$	$\frac{1}{65}$	$\frac{1}{65}$	$\frac{2}{65}$	$\frac{3}{65}$
2	$\frac{2}{65}$	0	$\frac{1}{65}$	0	0	0	$\frac{2}{65}$
3	$\frac{1}{13}$	$\frac{1}{65}$	0	$\frac{2}{65}$	0	0	$\frac{1}{13}$
4	$\frac{1}{65}$	0	$\frac{2}{65}$	0	$\frac{4}{65}$	0	0
5	$\frac{1}{65}$	0	$\frac{3}{65}$	$\frac{4}{65}$	0	[illegible]	[illegible]
6	$\frac{2}{65}$	0	0	0	0	0	[illegible] 65
7	$\frac{3}{65}$	$\frac{2}{65}$	$\frac{1}{13}$	0	0	$\frac{3}{65}$	0

4 and 6 don't communicate…

$\mathcal{D}$



Objective: Expected Route Length

DAN *N* of degree Δ

path length on *N*

$$\mathrm{ERL}(\mathcal{D},\mathrm{N}) = \sum_{(u,v)\in\mathcal{D}} p(u,v) \cdot d_N(u,v)$$

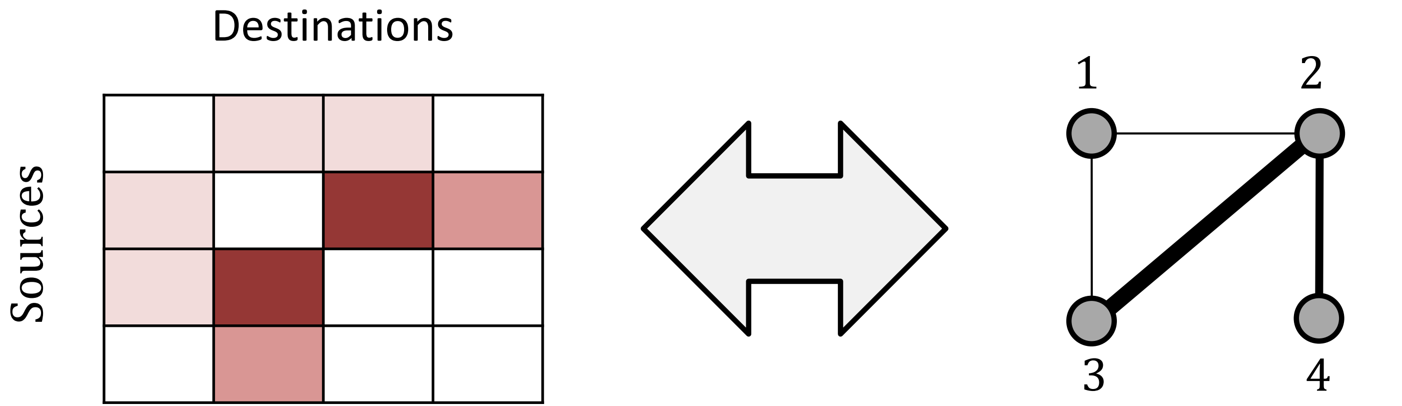
$\mathcal{D}[\mathbf{p}(\mathbf{i},\mathbf{j})]$: joint **distribution**

frequency

Remark

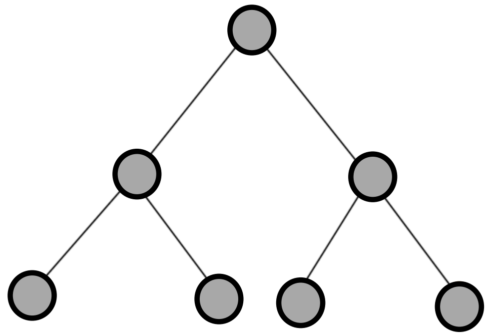
- Can represent demand matrix as a **demand graph**

sparse distribution $\mathcal{D}$ sparse graph G($\mathcal{D}$)



Some Examples

- DANs of Δ = 3:
  - E.g., complete binary **tree**
  - $d_N(u,v) \leq 2 \log n$
  - Can we do ***better*** than *log n*?

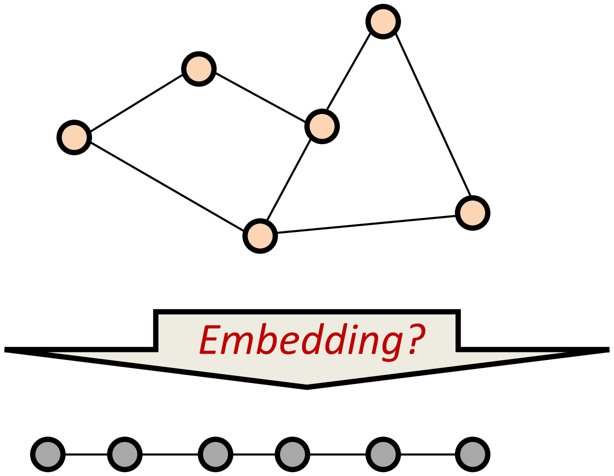


- DANs of Δ = 2:
  - E.g., set of **lines** and **cycles**

Remark: Hardness Proof

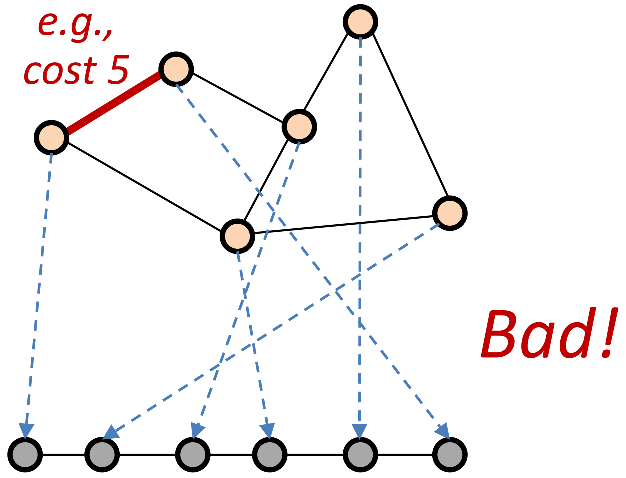
DAN design can be NP-hard

- **Example Δ = 2:** A Minimum Linear Arrangement (**MLA**) problem
  - A “Virtual Network Embedding Problem”, VNEP
  - *Minimize sum* of lengths of virtual edges



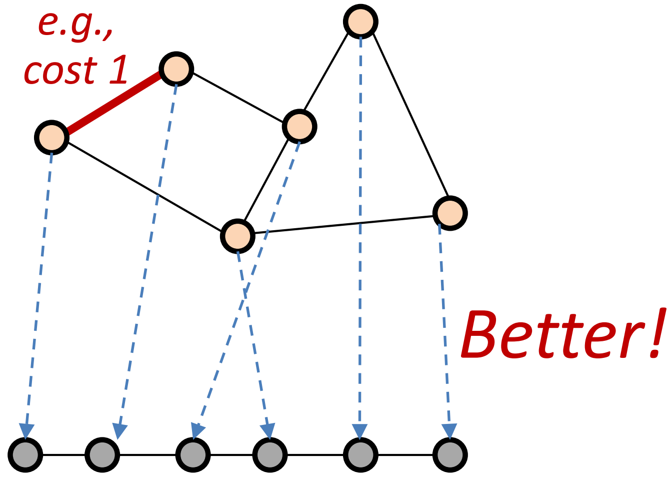
DAN design can be NP-hard

- **Example Δ = 2:** A Minimum Linear Arrangement (**MLA**) problem
  - A "Virtual Network Embedding Problem", VNEP
  - *Minimize sum* of lengths of virtual edges



DAN design can be NP-hard

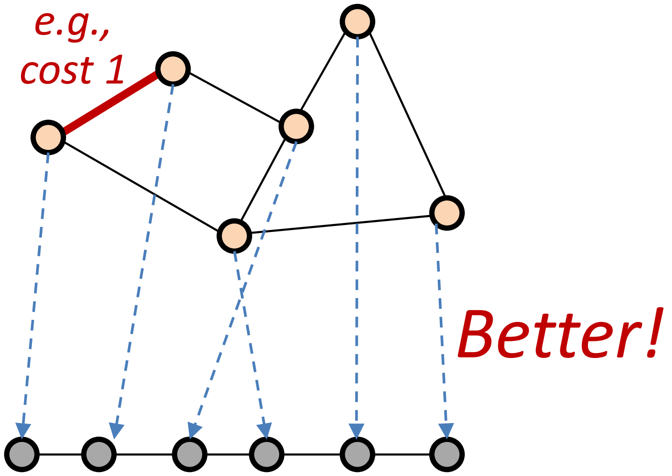
- **Example Δ = 2:** A Minimum Linear Arrangement (**MLA**) problem
  - A "Virtual Network Embedding Problem", VNEP
  - *Minimize sum* of lengths of virtual edges



DAN design can be NP-hard

- **Example Δ = 2:** A Mini … inear Arrangement (M … blem
  - A "Virtual N … bedding Problem", VNEP
  - *Minin* … lengths of virtual edges

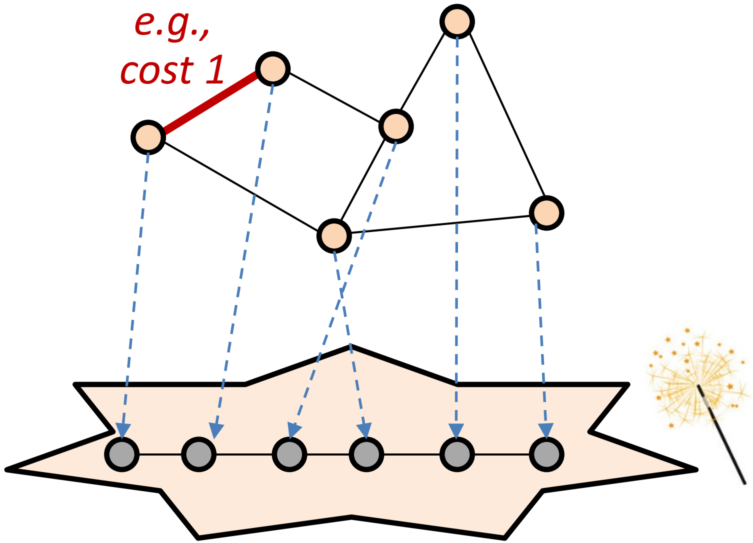
NP-hard, and so is DAN design.



DAN design can be NP-hard

- **Example Δ = 2:** A Mini inear Arrangement (M blem
  - A "Virtual N bedding Problem", VNEP
  - ***Minin*** engths of virtual edges

NP-hard, and so is DAN design.



- But what about > 2? ***Embedding*** problem still hard, but we have an additional **degree of freedom**:

Do topological flexibilities make problem easier or harder?!

A new knob for optimization!

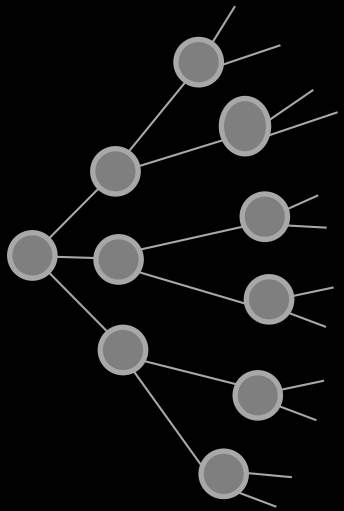
Rewinding the Clock: Degree-Diameter Tradeoff

Each network with n nodes and max degree $\Delta > 2$ must have a diameter of at least $\log(n)/\log(\Delta-1)-1$.

Example: constant Δ, $\log(n)$ diameter

Kudos to: Pedro Casas

Proof Idea



$1 \quad \Delta \quad \Delta(\Delta-1) \quad \ldots$

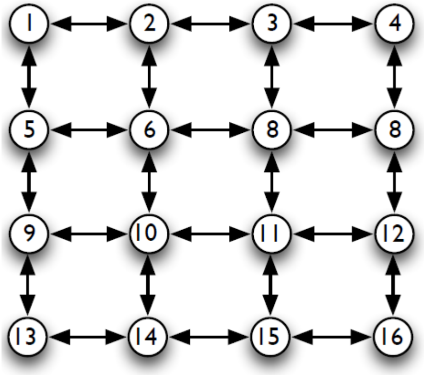
In k steps, reach at most $1+\sum \Delta(\Delta-1)^k$ nodes

Kudos to: Pedro Casas

Is there a better tradeoff in DANs?

Sometimes, DANs can be much better!

Example 1: low-degree demand

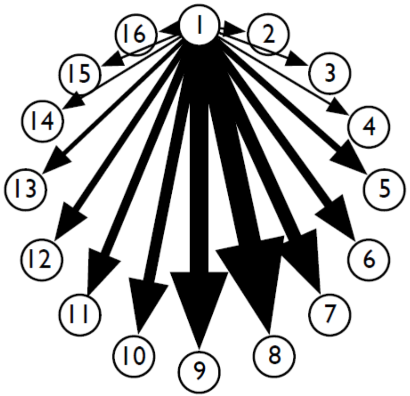


If **demand graph** is of degree Δ, it is trivial to design a **DAN** of degree Δ which achieves an ***expected route length of 1***.

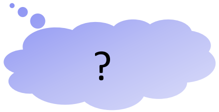
Just take DAN = demand graph!

Sometimes, DANs can be much better!

Example 2: skewed demand

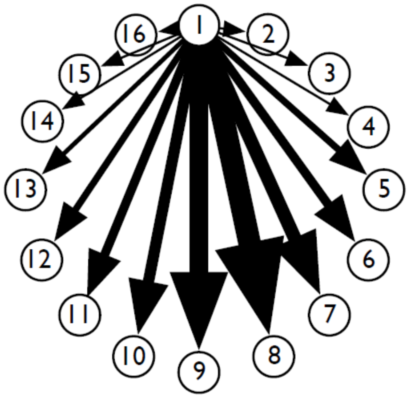


If **demand** is highly skewed, it is also possible to achieve an ***expected route length of O(1)*** in a constant-degree DAN.



Sometimes, DANs can be much better!

Example 2: skewed demand



If **demand** is highly skewed, it is also possible to achieve an ***expected route length of O(1)*** in a constant-degree DAN.

E.g., arrange neighbors of node 1 in a **Huffman** tree!

Toward Demand-Aware Networking: A Theory for Self-Adjusting Networks. Chen Avin and Stefan Schmid. ACM **SIGCOMM CCR**, October 2018

So on what does it depend?

So on what does it depend?

We argue (but still don‘t know!): on the **“entropy” of the demand**!

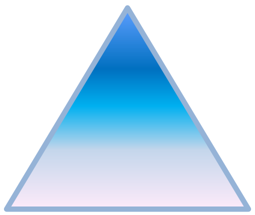
Intuition: Entropy Lower Bound

Lower Bound Idea: Leverage Coding or Datastructure

Destinations

Sources	1	2	3	4	5	6	7
1	0	$\frac{2}{65}$	$\frac{1}{13}$	$\frac{1}{65}$	$\frac{1}{65}$	$\frac{2}{65}$	$\frac{3}{65}$
2	$\frac{2}{65}$	0	$\frac{1}{65}$	0	0	0	$\frac{2}{65}$
3	$\frac{1}{13}$	$\frac{1}{65}$	0	$\frac{2}{65}$	0	0	$\frac{1}{13}$
4	$\frac{1}{65}$	0	$\frac{2}{65}$	0	$\frac{4}{65}$	0	0
5	$\frac{1}{65}$	0	$\frac{3}{65}$	$\frac{4}{65}$	0	0	0
6	$\frac{2}{65}$	0	0	0	0	0	$\frac{3}{65}$
7	$\frac{3}{65}$	$\frac{2}{65}$	$\frac{1}{13}$	0	0	$\frac{3}{65}$	0

- DAN just for a ***single (source) node 3***



- How good can this tree be? Cannot do better than Δ-ary **Huffman tree** for its destinations
- **Entropy** lower bound on ERL known for binary trees, e.g. ***Mehlhorn*** 1975

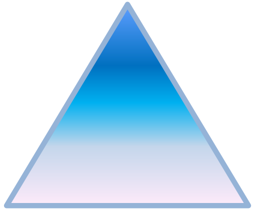
Lower Bound Idea: Leverage Coding or Datastructure

Destinations

Sources	1	2	3	4	5	6	7
1	0	$\frac{2}{65}$	$\frac{1}{13}$	$\frac{1}{65}$	$\frac{1}{65}$	$\frac{2}{65}$	$\frac{3}{65}$
2	$\frac{2}{65}$	0	$\frac{1}{65}$	0	0	0	$\frac{2}{65}$
3	$\frac{1}{13}$	$\frac{1}{65}$	0	$\frac{2}{65}$	0	0	$\frac{1}{13}$
4	$\frac{1}{65}$	0	$\frac{2}{65}$	0	$\frac{4}{65}$	0	0
5	$\frac{1}{65}$	0	$\frac{3}{65}$	$\frac{4}{65}$	0	0	0
6	$\frac{2}{65}$	0	0	0	0	0	$\frac{3}{65}$
7	$\frac{3}{65}$	$\frac{2}{65}$	$\frac{1}{13}$	0	0	$\frac{3}{65}$	0

An optimal "**ego-tree**" for this source!

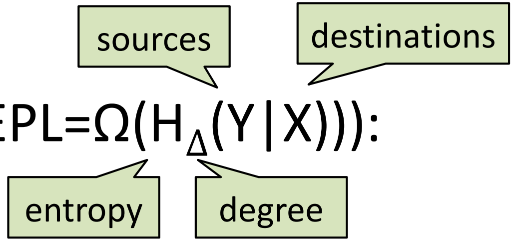
- DAN just for a ***single (source) node 3***



- How good can this tree be? Cannot do better than Δ-ary **Huffman tree** for its destinations
- **Entropy** lower bound on ERL known for binary trees, e.g. ***Mehlhorn*** 1975

So: Entropy of the *Entire* Demand

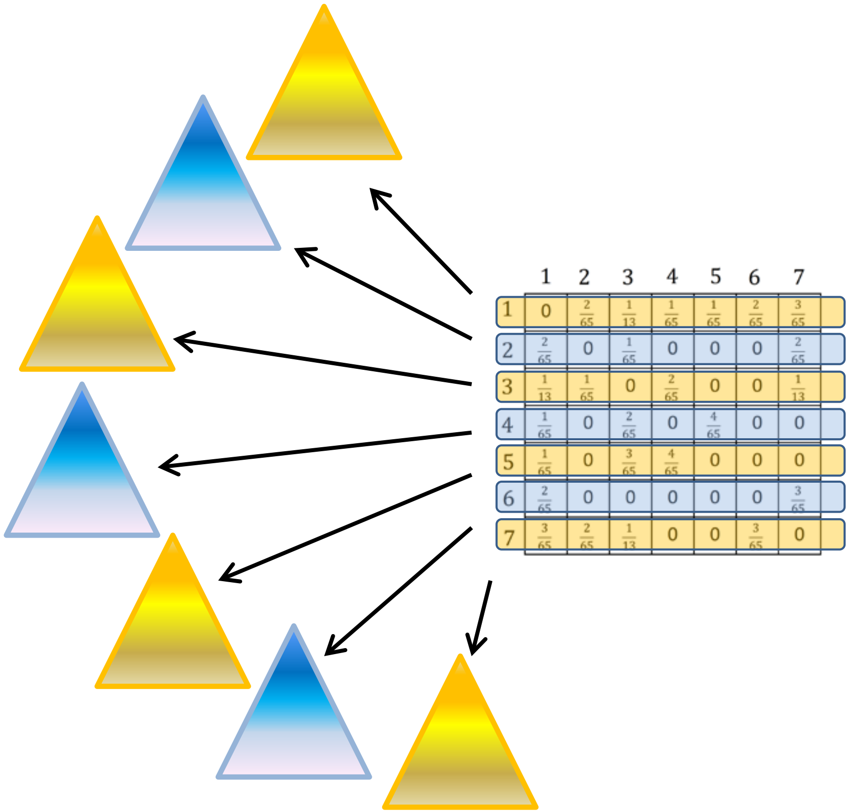
- **Proof idea** (EPL=Ω(H_Δ(Y|X))):



- Compute **ego-tree** for each source node

- Take ***union*** of all **ego-trees**

- Violates ***degree restriction*** but valid lower bound



	1	2	3	4	5	6	7
1	0	$\frac{2}{65}$	$\frac{1}{13}$	$\frac{1}{65}$	$\frac{1}{65}$	$\frac{2}{65}$	$\frac{3}{65}$
2	$\frac{2}{65}$	0	$\frac{1}{65}$	0	0	0	$\frac{2}{65}$
3	$\frac{1}{13}$	$\frac{1}{65}$	0	$\frac{2}{65}$	0	0	$\frac{1}{13}$
4	$\frac{1}{65}$	0	$\frac{2}{65}$	0	$\frac{4}{65}$	0	0
5	$\frac{1}{65}$	0	$\frac{3}{65}$	$\frac{4}{65}$	0	0	0
6	$\frac{2}{65}$	0	0	0	0	0	$\frac{3}{65}$
7	$\frac{3}{65}$	$\frac{2}{65}$	$\frac{1}{13}$	0	0	$\frac{3}{65}$	0

Entropy of the *Entire* Demand: Sources *and* Destinations

Do this in **both dimensions**:

EPL $\geq \Omega(\max\{H_\Delta(Y|X), H_\Delta(X|Y)\})$

$\Omega(H_\Delta(X|Y))$

	1	2	3	4	5	6	7
1	0	$\frac{2}{65}$	$\frac{1}{13}$	$\frac{1}{65}$	$\frac{1}{65}$	$\frac{2}{65}$	$\frac{3}{65}$
2	$\frac{2}{65}$	0	$\frac{1}{65}$	0	0	0	$\frac{2}{65}$
3	$\frac{1}{13}$	$\frac{1}{65}$	0	$\frac{2}{65}$	0	0	$\frac{1}{13}$
4	$\frac{1}{65}$	0	$\frac{2}{65}$	0	$\frac{4}{65}$	0	0
5	$\frac{1}{65}$	0	$\frac{3}{65}$	$\frac{4}{65}$	0	0	0
6	$\frac{2}{65}$	0	0	0	0	0	$\frac{3}{65}$
7	$\frac{3}{65}$	$\frac{2}{65}$	$\frac{1}{13}$	0	0	$\frac{3}{65}$	0

$\Omega(H_\Delta(Y|X))$

$\mathcal{D}$

Entropy of the *Entire* Demand: Sources *and* Destinations

Do this in **both dimensions**:

EPL ≥ $\Omega(\max\{H_\Delta(Y|X), H_\Delta(X|Y)\})$

$\Omega(H_\Delta(X|Y))$

	1	2	3	4	5	6	7
1	0	$\frac{2}{65}$	$\frac{1}{13}$	$\frac{1}{65}$	$\frac{1}{65}$	$\frac{2}{65}$	$\frac{3}{65}$
2	$\frac{2}{65}$	0	$\frac{1}{65}$	0	0	0	$\frac{2}{65}$
3	$\frac{1}{13}$	$\frac{1}{65}$	0	$\frac{2}{65}$	0	0	$\frac{1}{13}$
4	$\frac{1}{65}$	0	$\frac{2}{65}$	0	$\frac{4}{65}$	0	0
5	$\frac{1}{65}$	0	$\frac{3}{65}$	$\frac{4}{65}$	0	0	0
6	$\frac{2}{65}$	0	0	0	0	0	$\frac{3}{65}$
7	$\frac{3}{65}$	$\frac{2}{65}$	$\frac{1}{13}$	0	0	$\frac{3}{65}$	0

$\Omega(H_\Delta(Y|X))$

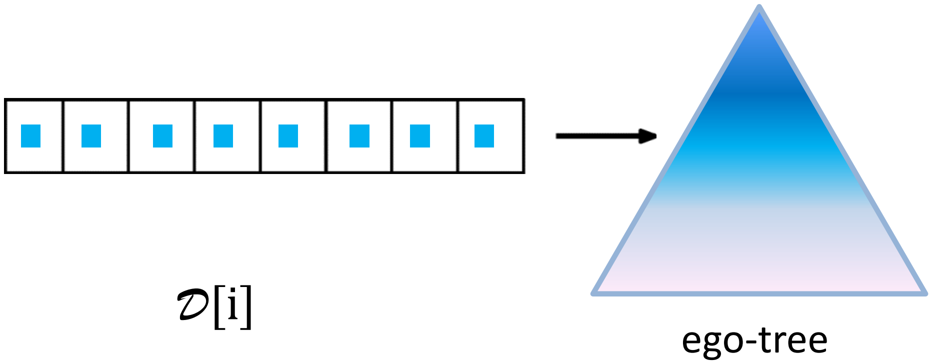
$\mathcal{D}$

Demand-Aware Network Designs of Bounded Degree. Chen Avin, Kaushik Mondal, and Stefan Schmid. **DISC**, 2017.

Achieving Entropy Limit: Algorithms

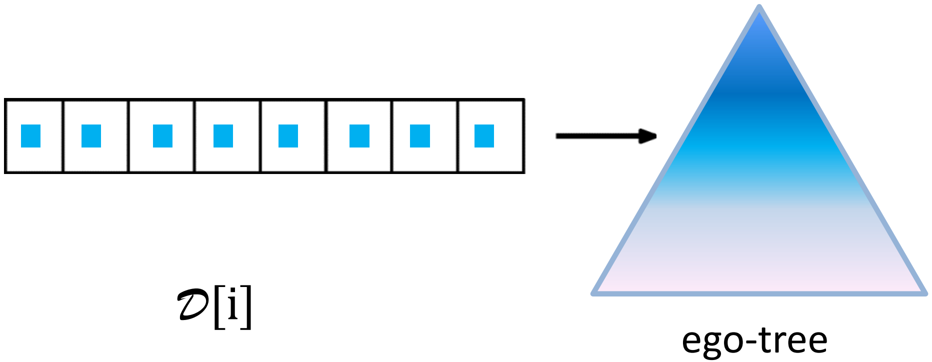
Ego-Trees Revisited

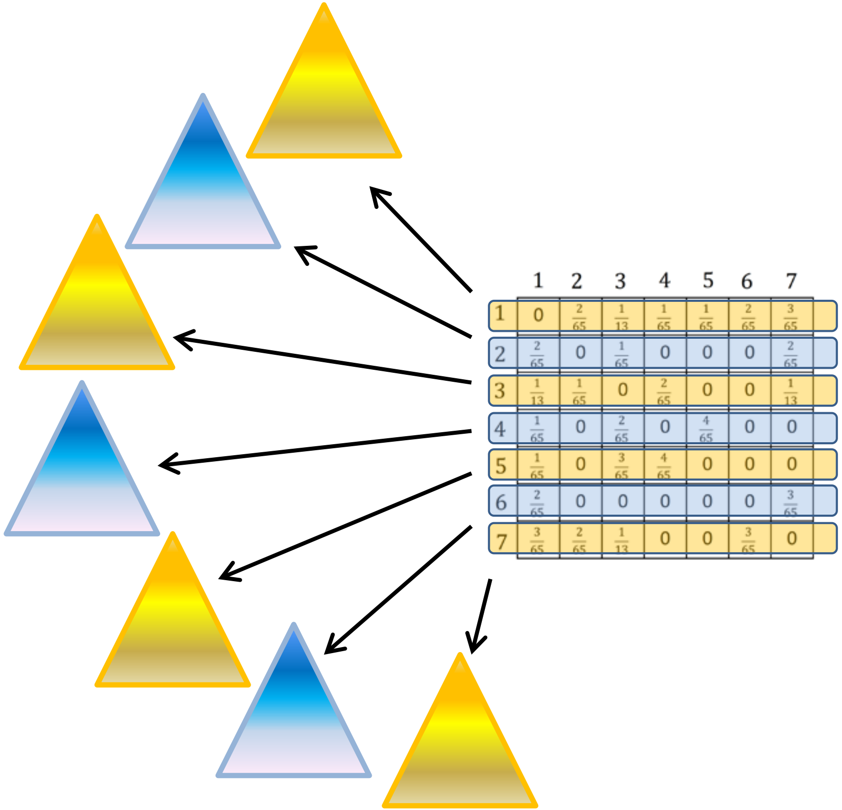
- ego-tree: optimal tree for a row (= given source)



Ego-Trees Revisited

- ego-tree: optimal tree for a row (= given source)



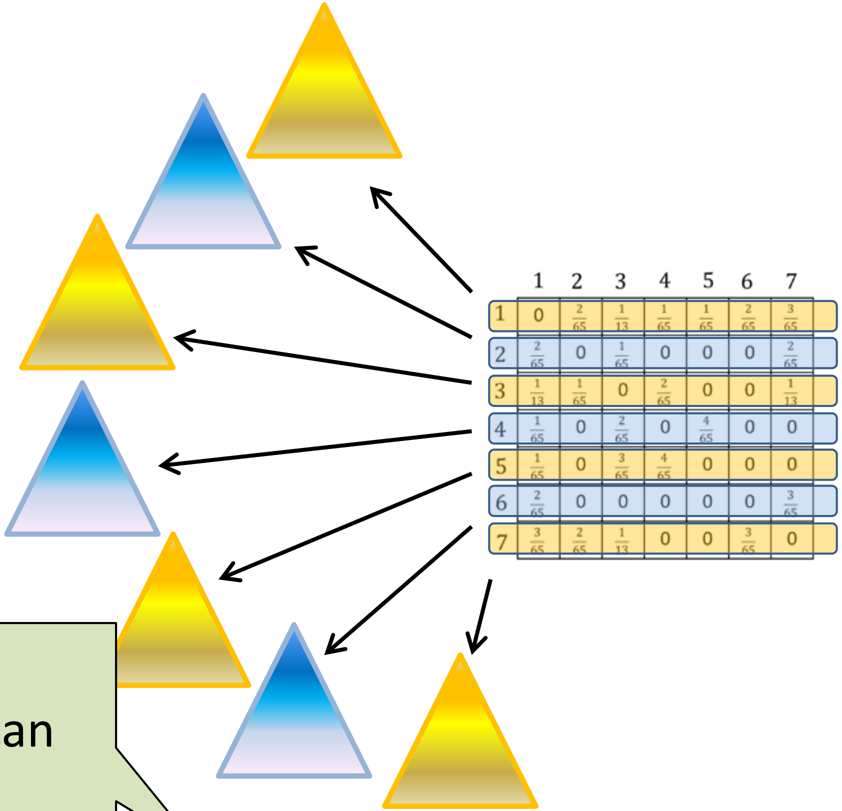


	1	2	3	4	5	6	7
1	0	$\frac{2}{65}$	$\frac{1}{13}$	$\frac{1}{65}$	$\frac{1}{65}$	$\frac{2}{65}$	$\frac{3}{65}$
2	$\frac{2}{65}$	0	$\frac{1}{65}$	0	0	0	$\frac{2}{65}$
3	$\frac{1}{13}$	$\frac{1}{65}$	0	$\frac{2}{65}$	0	0	$\frac{1}{13}$
4	$\frac{1}{65}$	0	$\frac{2}{65}$	0	$\frac{4}{65}$	0	0
5	$\frac{1}{65}$	0	$\frac{3}{65}$	$\frac{4}{65}$	0	0	0
6	$\frac{2}{65}$	0	0	0	0	0	$\frac{3}{65}$
7	$\frac{3}{65}$	$\frac{2}{65}$	$\frac{1}{13}$	0	0	$\frac{3}{65}$	0

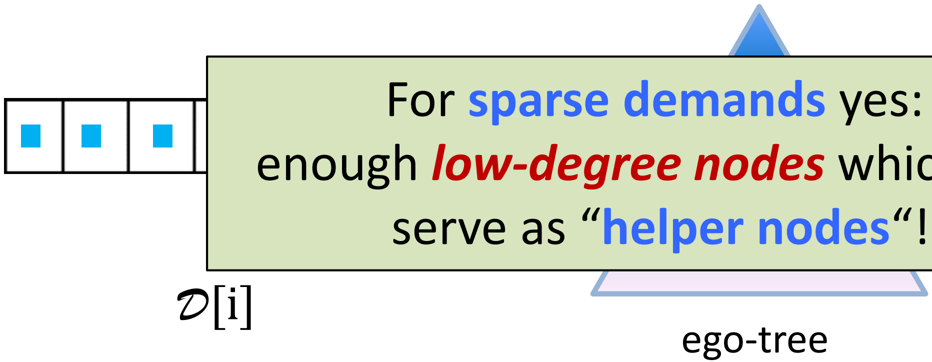
Can we merge the trees ***without distortion*** and ***keep degree low***?

Ego-Trees Revisited

- ego-tree: optimal tree for a row (= given source)



	1	2	3	4	5	6	7
1	0	$\frac{2}{65}$	$\frac{1}{13}$	$\frac{1}{65}$	$\frac{1}{65}$	$\frac{2}{65}$	$\frac{3}{65}$
2	$\frac{2}{65}$	0	$\frac{1}{65}$	0	0	0	$\frac{2}{65}$
3	$\frac{1}{13}$	$\frac{1}{65}$	0	$\frac{2}{65}$	0	0	$\frac{1}{13}$
4	$\frac{1}{65}$	0	$\frac{2}{65}$	0	$\frac{4}{65}$	0	0
5	$\frac{1}{65}$	0	$\frac{3}{65}$	$\frac{4}{65}$	0	0	0
6	$\frac{2}{65}$	0	0	0	0	0	$\frac{3}{65}$
7	$\frac{3}{65}$	$\frac{2}{65}$	$\frac{1}{13}$	0	0	$\frac{3}{65}$	0



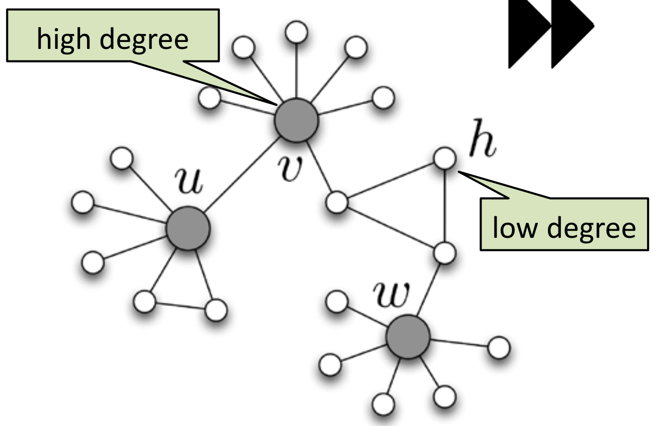
For **sparse demands** yes:
enough ***low-degree nodes*** which can serve as "**helper nodes**"!

Can we merge the trees ***without distortion*** and ***keep degree low***?

From Trees to Networks

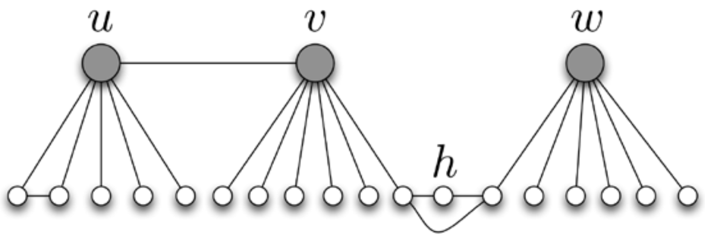
Idea: Degree Reduction

① **Demand graph**

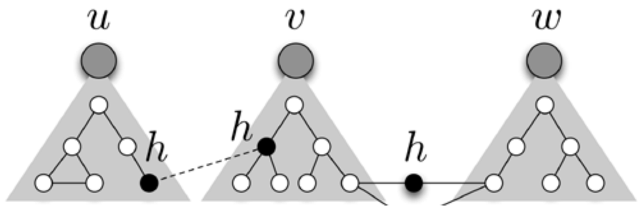


Taking union of ego-trees results in ***high degree***:
u and *v* will appear in many ego-trees

② **Hierarchical representation**



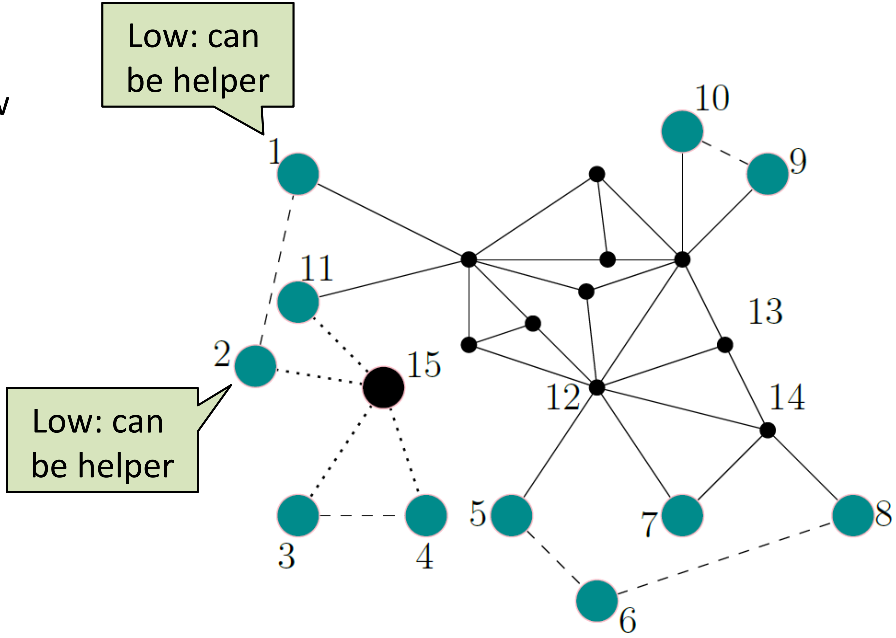
③ **Add low-degree nodes as helpers**



Node *h* ***helps edge (u, v)*** by participating in *ego-tree(u)* as a relay node toward *v* and *in ego-tree(v)* as a relay toward *u*

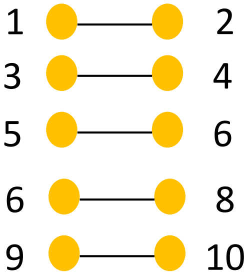
Algorithm: Degree Reduction

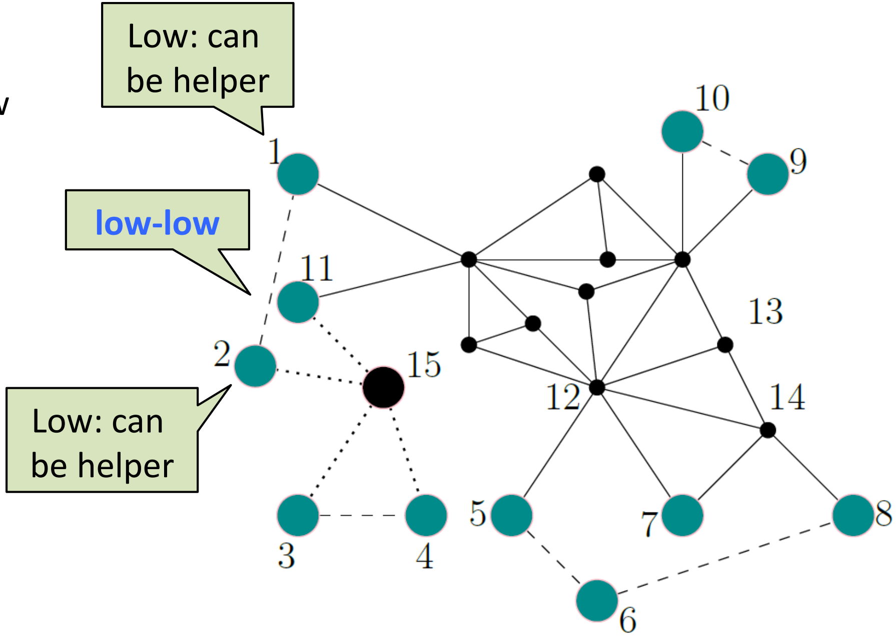
- Find low degree nodes
  - Half of the nodes of lowest degree: “below twice average degree”



Algorithm: Degree Reduction

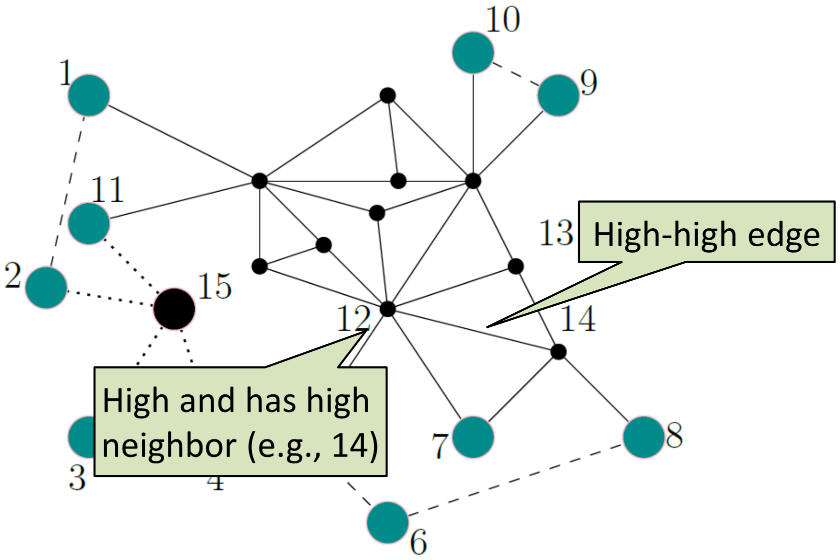
- Find low degree nodes
  - Half of the nodes of lowest degree: “below twice average degree”
- Put the low-low edges into DAN and remove from demand





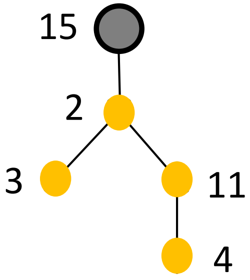
Algorithm: Degree Reduction

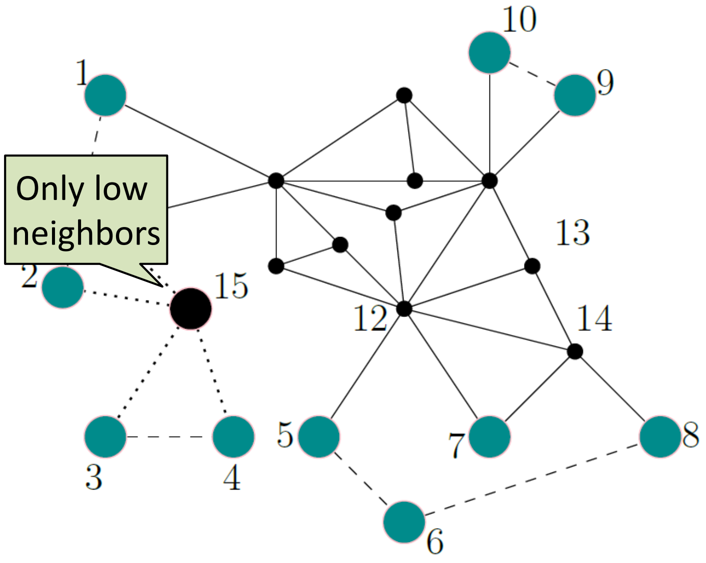
- Find low degree nodes
  - Half of the nodes of lowest degree: “below twice average degree”
- Put the low-low edges into DAN and remove from demand
- Mark high-high edges
  - Put (any) low degree nodes in between (e.g., 1 or 2): one is enough so distance increased by **+1**



Algorithm: Degree Reduction

- Find **low** degree nodes
  - Half of the nodes of lowest degree: “below twice average degree”
- **Put** the **low-low** edges into DAN and remove from demand
- Mark **high-high** edges
  - Put (any) **low degree** nodes in between (e.g., 1 or 2): one is enough so distance increased by **+1**
- Now high degree nodes have only low degree neighbors: make **tree**
  - Create optimal **binary tree** with low degree neighbors



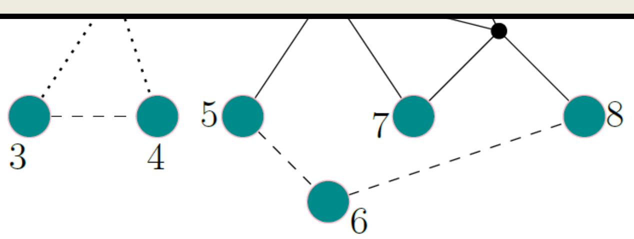


Algorithm: Degree Reduction

- Find **low** degree nodes
  - Half of the nodes of lowest degree: "below twice average degree"

Theorem [Asymptotic Optimality]: Helper node does not participate in many trees, so ***constant degree***, and ***constant distortion***.

- Now high degree nodes have only low degree neighbors: make **tree**
  - Create optimal **binary tree** with low degree neighbors

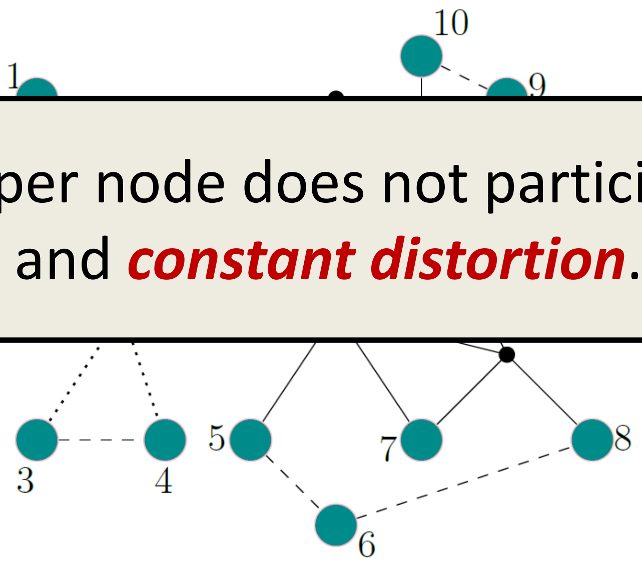


Algorithm: Degree Reduction

- Find **low** degree nodes
  - Half of the nodes of lowest degree: "below twice average degree"

Theorem [Asymptotic Optimality]: Helper node does not participate in many trees, so ***constant degree***, and ***constant distortion***.

- Now high degree nodes have only low degree neighbors: make **tree**
  - Create optimal **binary tree** with low degree neighbors



Demand-Aware Network Designs of Bounded Degree. Chen Avin, Kaushik Mondal, and Stefan Schmid. **DISC**, 2017.

DAN Design: Related to Spanners

Low-Distortion Spanners

- Classic problem: find sparse, distance-preserving (low-distortion) **spanner** of a graph

The „DAN“

The demand graph

Low-Distortion Spanners

- Classic problem: find sparse, distance-preserving (low-distortion) **spanner** of a graph
- But:
  - Spanners aim at low distortion among ***all pairs***; in our case, we are only interested in the ***local distortion***, 1-hop communication neighbors
  - We allow **auxiliary edges** (not a subgraph): similar to **geometric spanners**
  - We require ***constant degree***

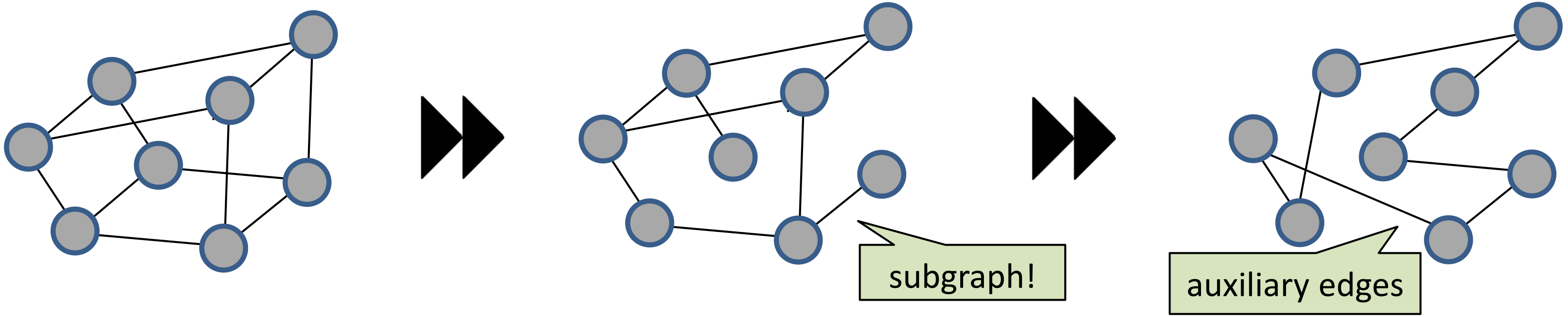
Yet: We can leverage the connection to spanners sometimes!

Theorem: If request distribution $\mathcal{D}$ is **regular and uniform**, and if we can find a constant distortion, linear sized (i.e., **constant**, **sparse**) spanner for this request graph: then we can design a constant degree DAN providing an ***optimal ERL*** (i.e., O(H(X|Y)+H(Y|X)).

r-regular *and* ***uniform*** **demand***:*

Sparse, ***irregular*** ***(constant)*** **spanner***:*

Constant *degree* ***optimal*** **DAN** (ERL at most ***log r***):

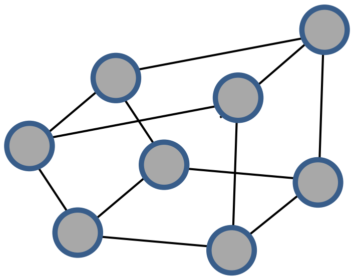


Yet: We can leverage the connection to spanners sometimes!

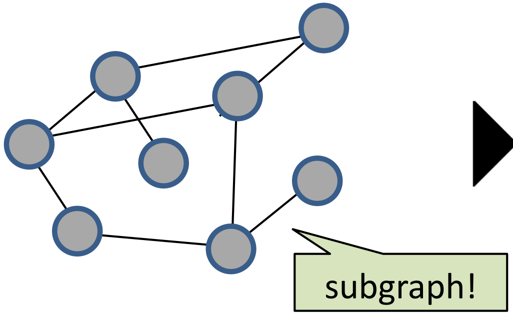
Theorem: If request distribution $\mathcal{D}$ is **regular and uniform**, and if we can find a constant distortion, linear sized (i.e., **constant, sparse**) spanner for this request graph: then we can design a constant degree DAN p[...] |X)).

Simply using our degree reduction trick again: now for spanner!

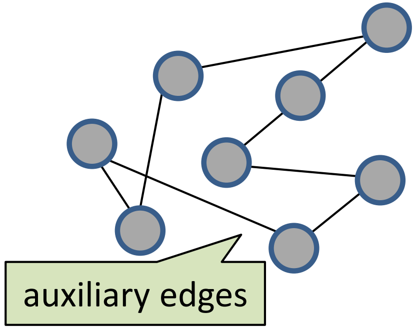
r-regular *and* ***uniform*** **demand***:*



Sparse, ***irregular (constant)*** **spanner***:*



Constant *degree* ***optimal*** **DAN** (ERL at most ***log r***):



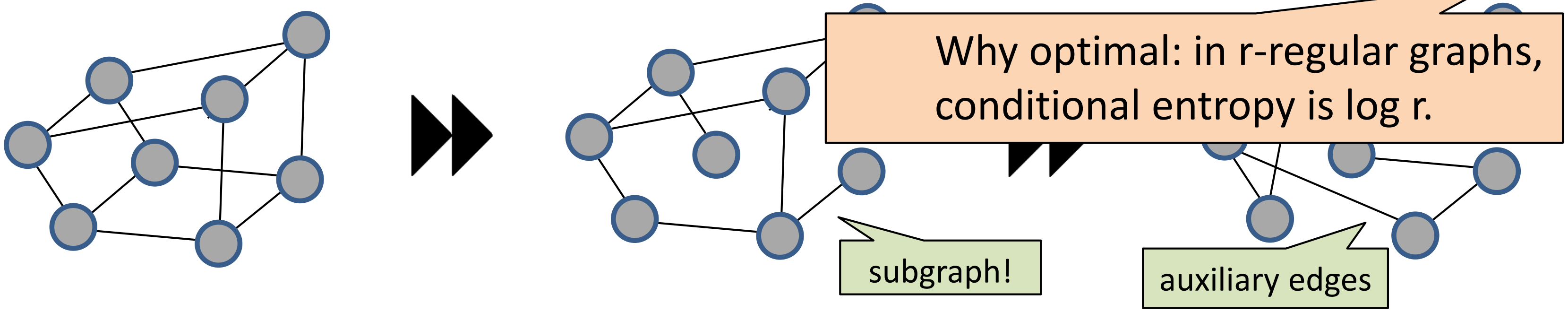
Yet: We can leverage the connection to spanners sometimes!

Theorem: If request distribution $\mathcal{D}$ is **regular and uniform**, and if we can find a constant distortion, linear sized (i.e., **constant**, **sparse**) spanner for this request graph: then we can design a constant degree DAN providing an ***optimal EPL*** (i.e., O(H(X|Y)+H(Y|X)).

r-regular *and* ***uniform*** **demand***:*

Sparse, ***irregular (constant)*** **spanner***:*

Constant *degree* ***optimal*** **DAN** (ERL at most ***log r***):



Proof Idea

- **Degree reduction** again, this time ***from sparse spanner*** (before: from sparse demand graph)

Corollaries

- Optimal DAN designs for
  - **Hypercubes** (with n log n edges)
  - **Chordal** graphs
  - Trivial: graphs with polynomial degree (**dense graphs**)
  - Graphs of **locally bounded doubling dimension**

Has sparse 3-spanner.

Has sparse O(1)-spanner.

We also know some more algos, e.g., for BSTs.

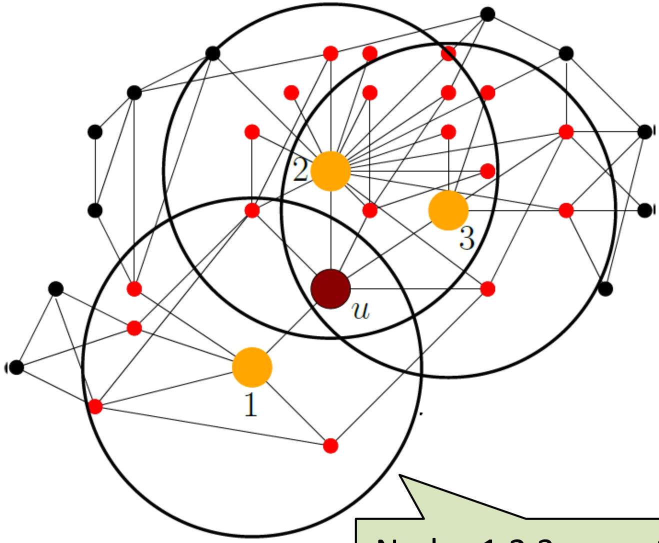
An Example: Demands of Locally-Bounded Doubling Dimension

- **LDD**: $G_{\mathcal{D}}$ has a **Locally-bounded Doubling Dimension** (LDD) iff all 2-hop neighbors are covered by 1-hop neighbors of just ***λ*** nodes
  - Note: care only about ***2-neighborhood***

We only consider 2 hops!

- Formally, $B(u, 2) \subseteq \bigcup_{i=1}^{\lambda} B(v_i, 1)$

- Challenge: can be of ***high degree***!



Nodes 1,2,3 cover 2-hop neighborhood of u.

DAN for Locally-Bounded Doubling Dimension

Lemma: There exists a sparse 9-(subgraph)spanner for LDD.

This ***implies optimal DAN***: still focus on regular and uniform!

Def. (ε-net): A subset V' of V is a **ε-net** for a graph $G = (V,E)$ if

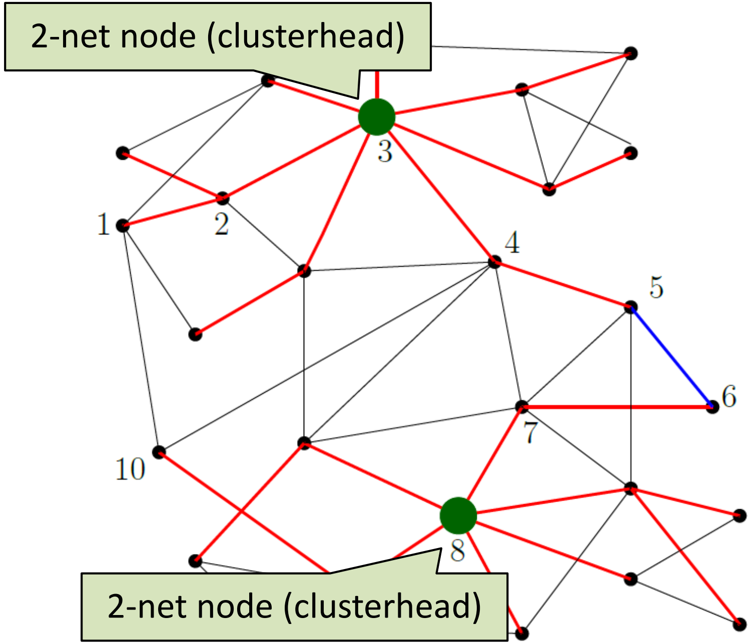
- V' sufficiently "independent": for every $u, v \in V'$, $d_G(u, v) > \boldsymbol{\varepsilon}$
- "dominating" V: for each $w \in V$, $\exists$ at least one $u \in V'$ such that, $d_G(u,w) \leq \boldsymbol{\varepsilon}$

9-Spanner for LDD (= optimal DAN)

Simple algorithm:

1. Find a 2-net

Easy: Select nodes into 2-net **one-by-one** in decreasing (remaining) degrees, **remove 2-neighborhood**. Iterate.



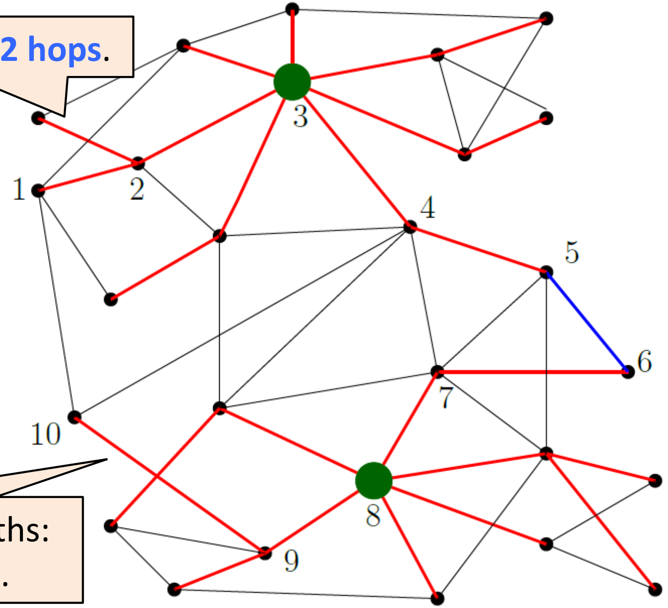
9-Spanner for LDD (= optimal DAN)

Simple algorithm:

1. Find a 2-net

2. Assign nodes to one of the closest 2-net nodes: tree

Assign: at most **2 hops**.



Union of these shortest paths: a **forest**. Add to **spanner**.

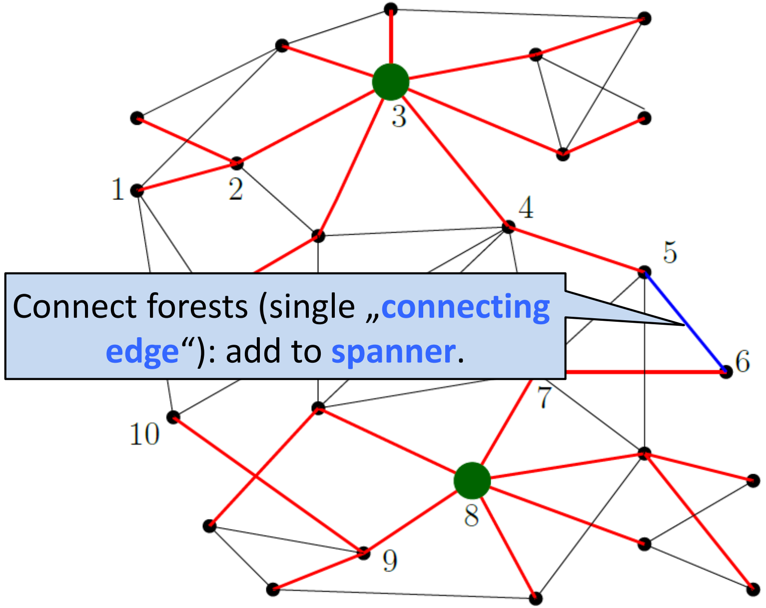
9-Spanner for LDD (= optimal DAN)

Simple algorithm:

1. Find a 2-net

2. Assign nodes to one of the closest 2-net nodes: tree

3. Join two clusters if there are edges in between



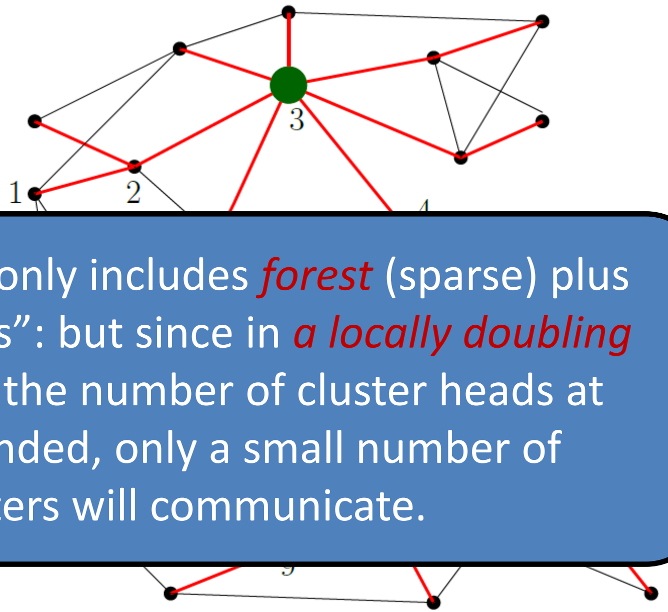
9-Spanner for LDD (= optimal DAN)

Distortion 9: *Short detour* via clusterheads: u,ch(u),x,y,ch(v),v

2. Assign nodes to one of the closest 2-net node

3. Join two clusters

edges in between



Sparse: Spanner only includes *forest* (sparse) plus "connecting edges": but since in *a locally doubling dimension graph* the number of cluster heads at distance 5 is bounded, only a small number of neighboring clusters will communicate.

So: How *much* structure/entropy is there?

How to ***measure*** it?
And which ***types of structures***? E.g., **temporal** structure in addition to **non-temporal** structure?
More ***tricky***!

Often only intuitions in the literature…

"less than 1% of the rack pairs account for 80% of the total traffic"

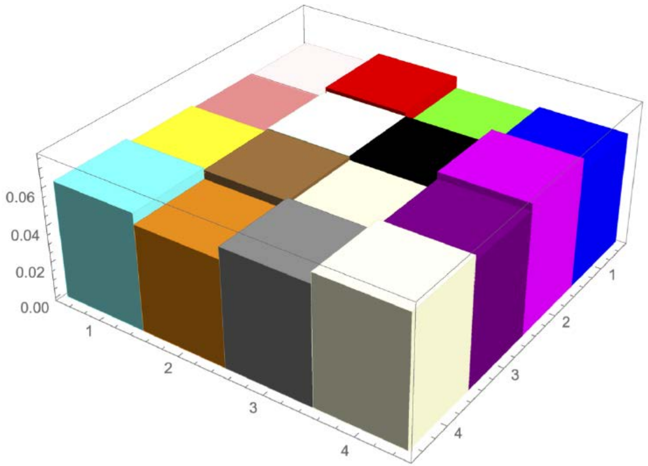
"only a few ToRs switches are hot and most of their traffic goes to a few other ToRs"

"over 90% bytes flow in elephant flows"

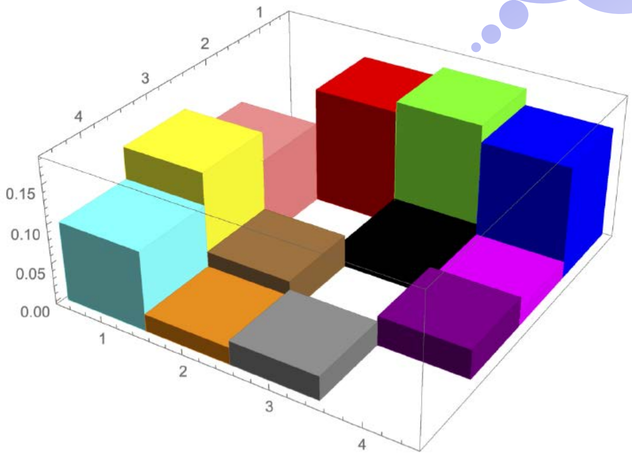
… and it *is* intuitive!

Non-temporal Structure

Color =
comm. pair



VS



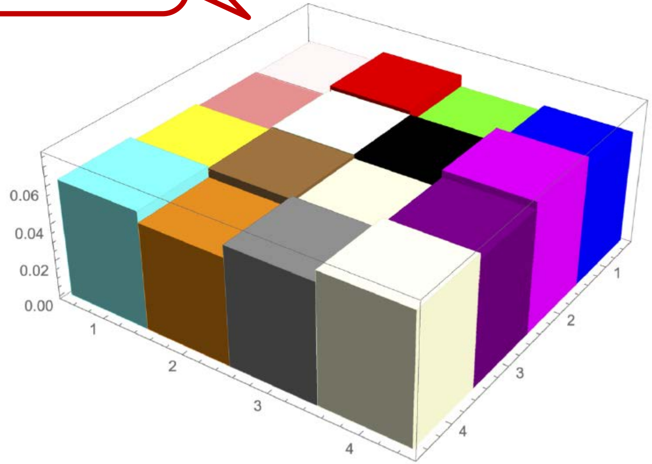
Traffic matrix of two different **distributed ML** applications (GPU-to-GPU):

Which one has *more structure*?

… and it *is* intuitive!

Non-temporal Structure

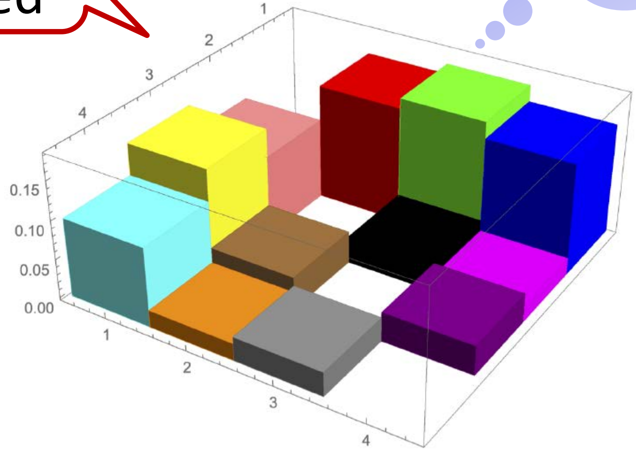
More uniform



VS

More skewed

Color = comm. pair

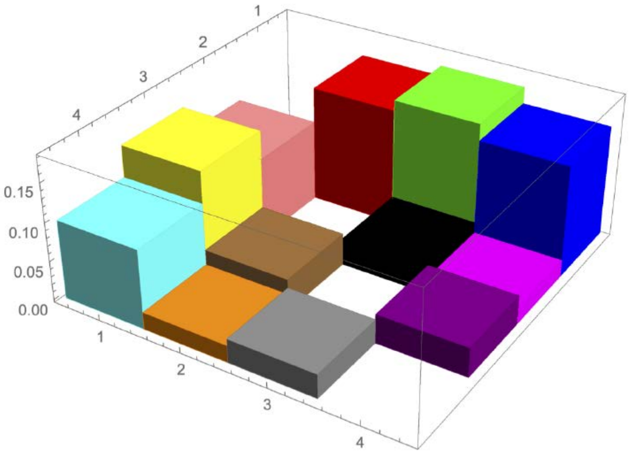


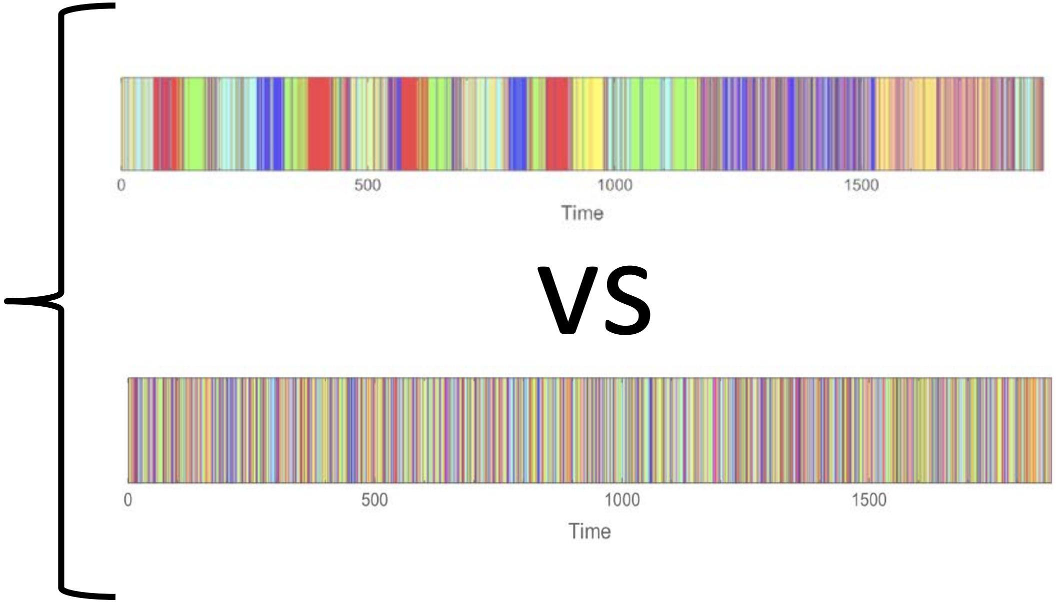
Traffic matrix of two different **distributed ML** applications (GPU-to-GPU):

Which one has *more structure*?

… and it *is* intuitive!

Temporal Structure



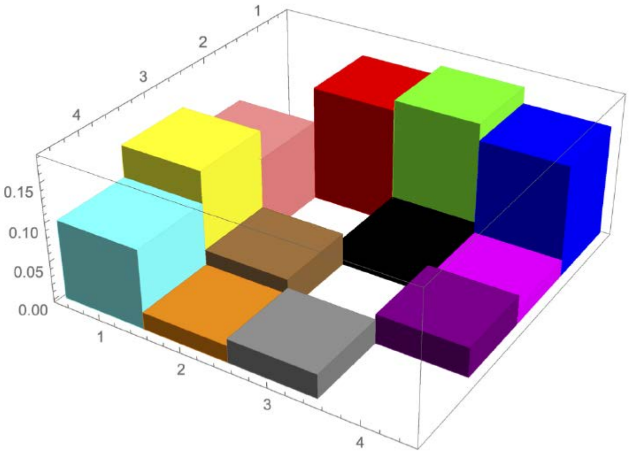


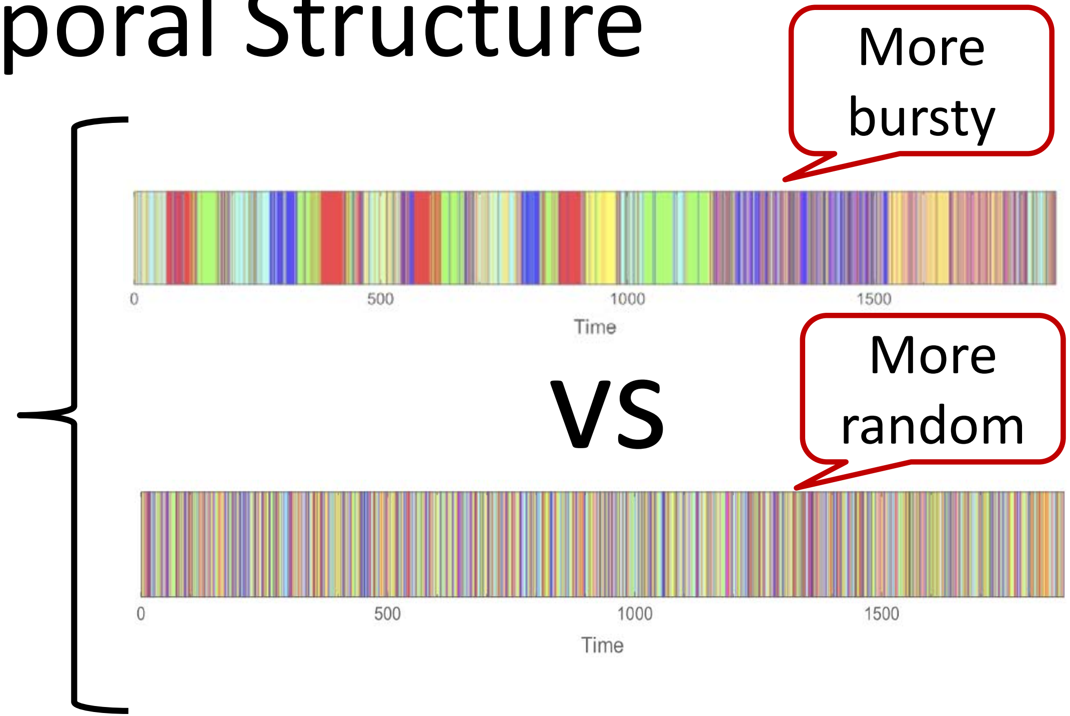
Two different ways to generate *same traffic matrix* (same non-temporal structure):

Which one has *more structure*?

… and it *is* intuitive!

Temporal Structure



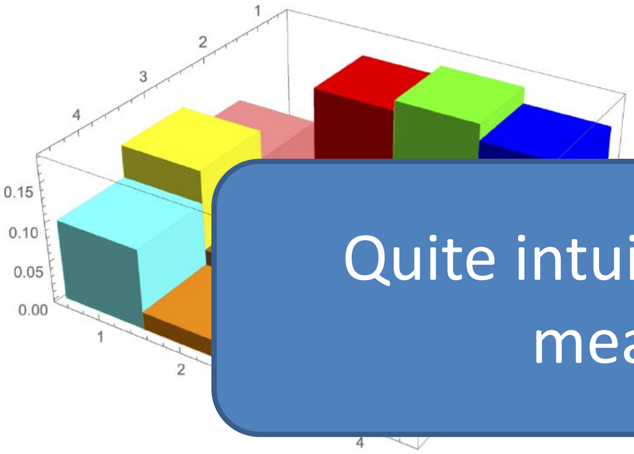


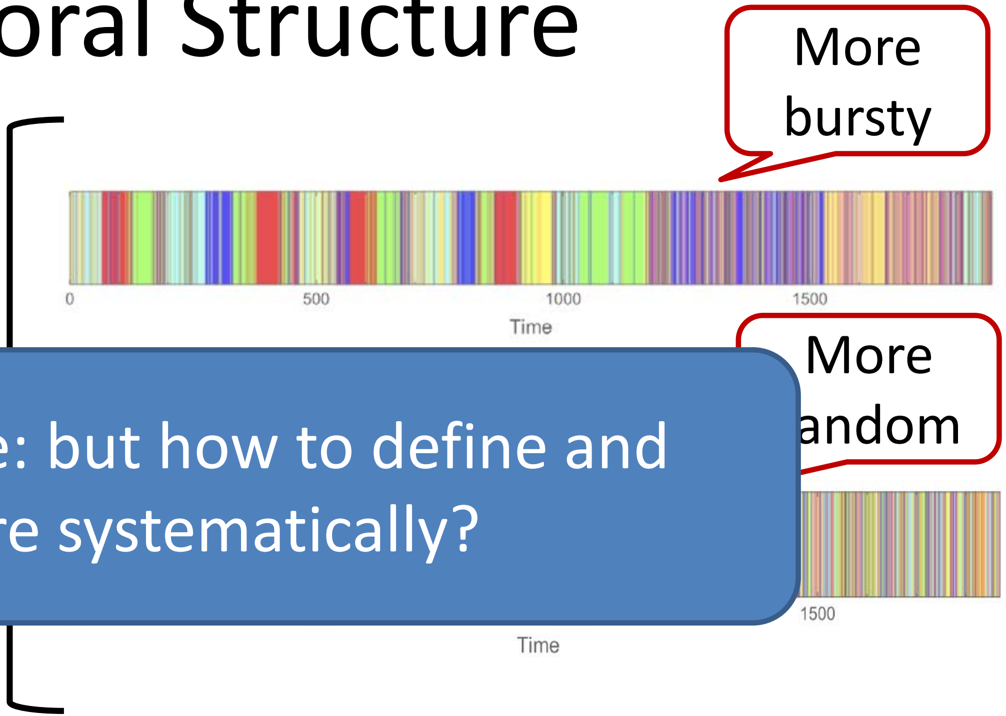
Two different ways to generate *same traffic matrix* (same non-temporal structure):

Which one has *more structure*?

… and it *is* intuitive!

Temporal Structure





Quite intuitive: but how to define and measure systematically?

Two different ways to generate *same traffic matrix* (same non-temporal structure):

Which one has *more structure*?

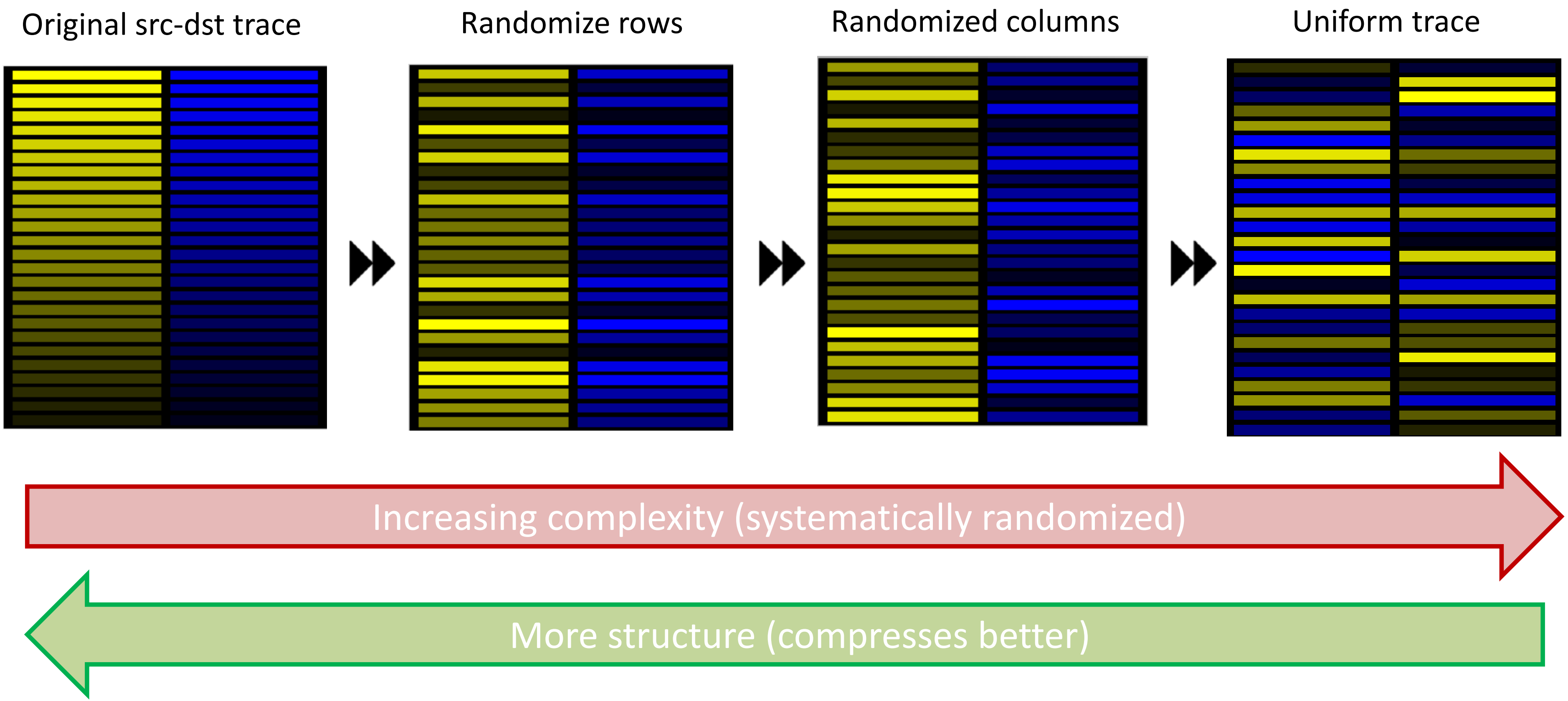
The Trace Complexity

- An **information-theoretic** approach: how can we ***measure the entropy*** (rate) of a traffic trace?

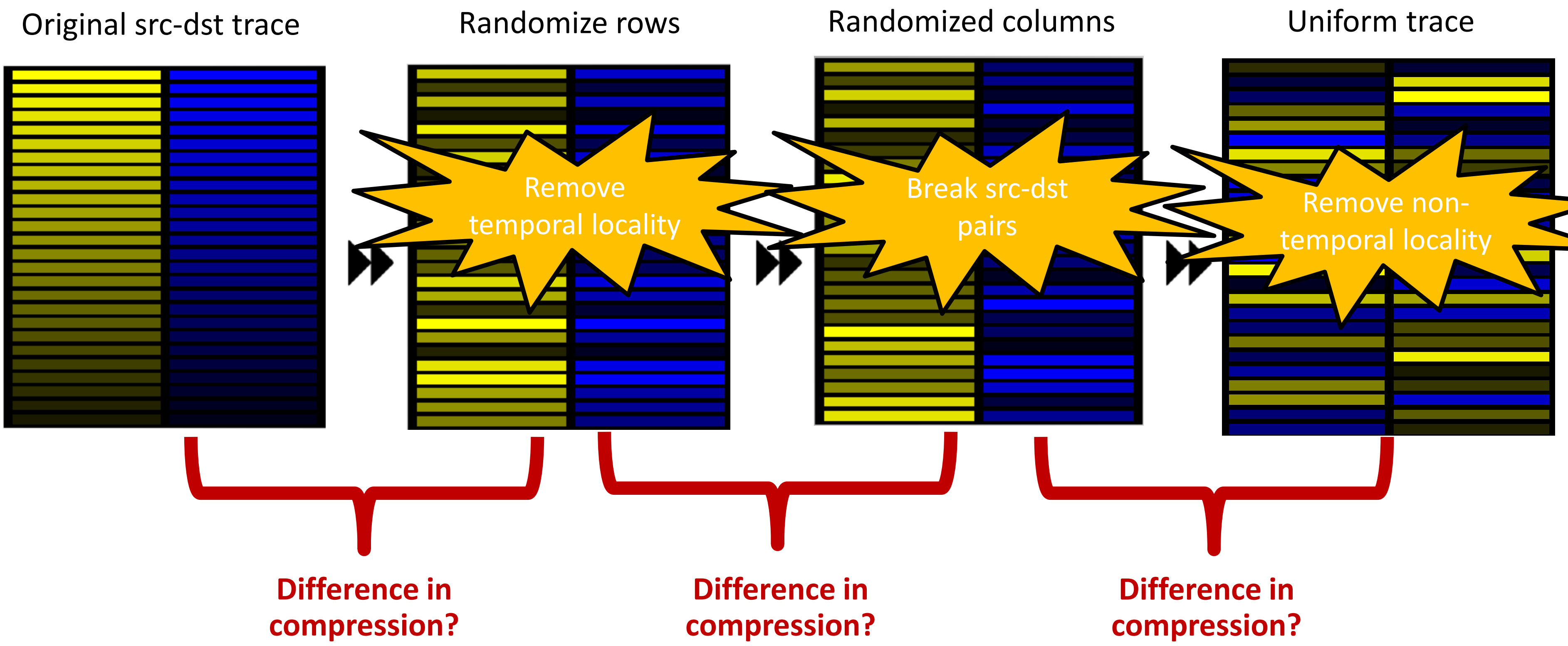
- Henceforth called the **trace complexity**

- Simple approximation: „**shuffle&compress**"
  - Remove structure by iterative *randomization*
  - Difference of compression *before and after* randomization: structure

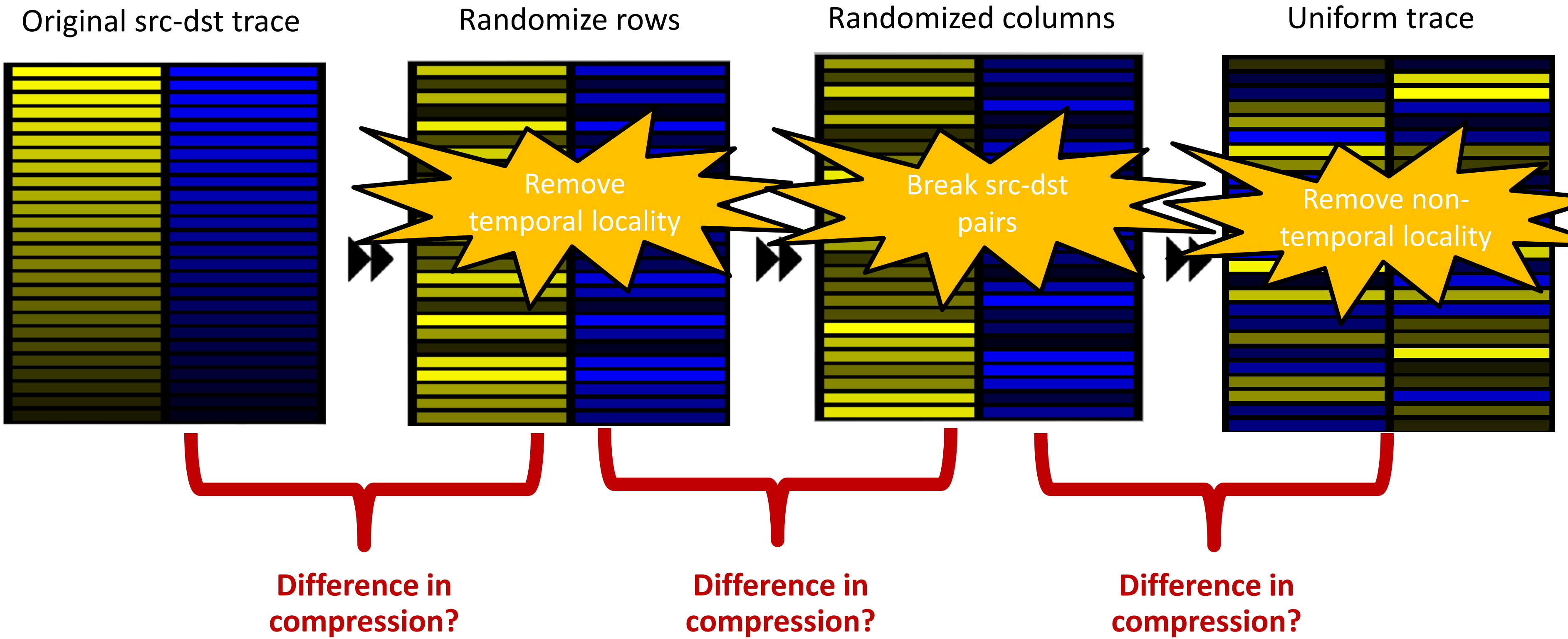
The Trace Complexity



The Trace Complexity

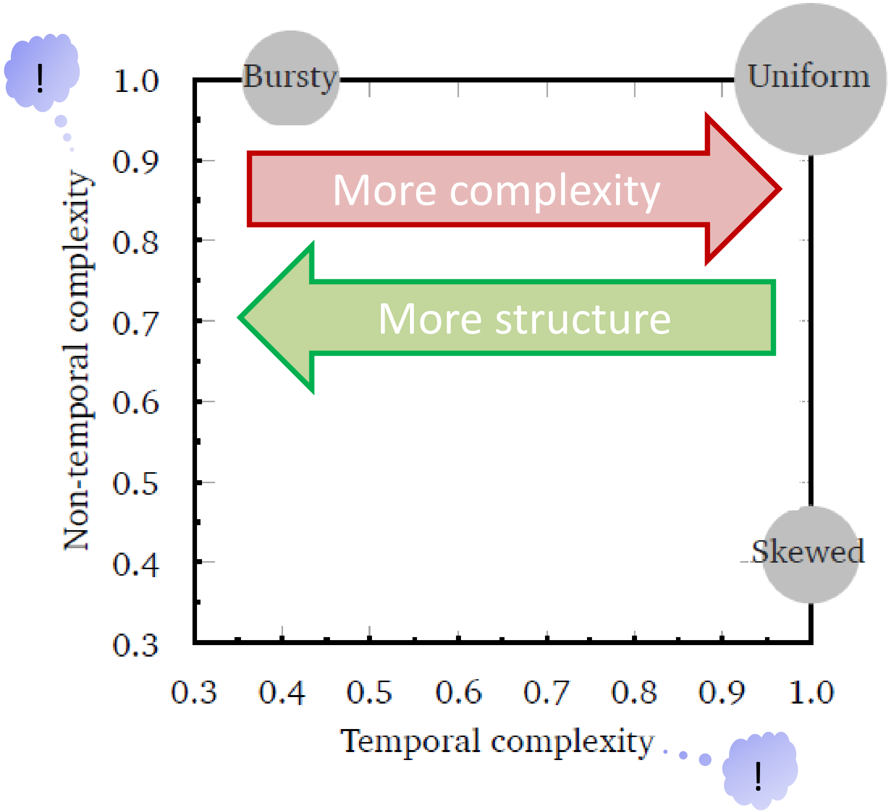


The Trace Complexity



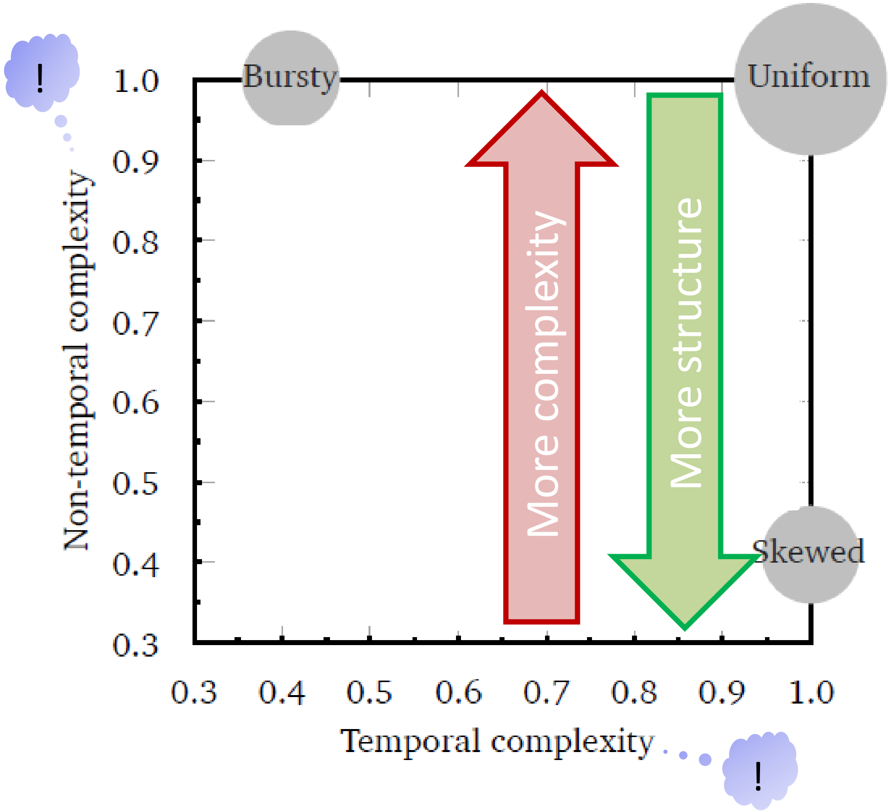
Can be used to define a „complexity map“!

The Complexity Map



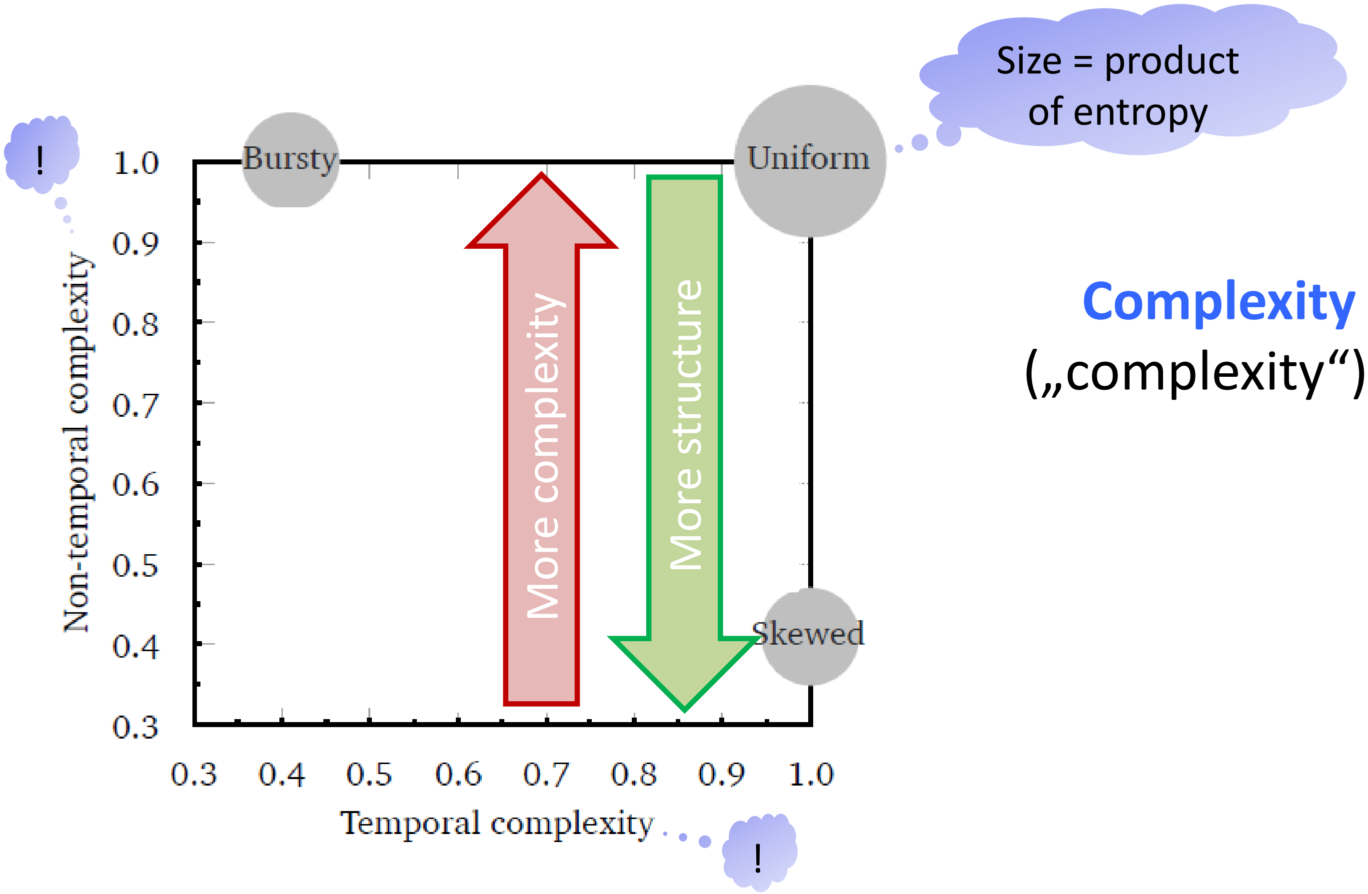
Complexity Map: Entropy („complexity“) of traffic traces.

The Complexity Map



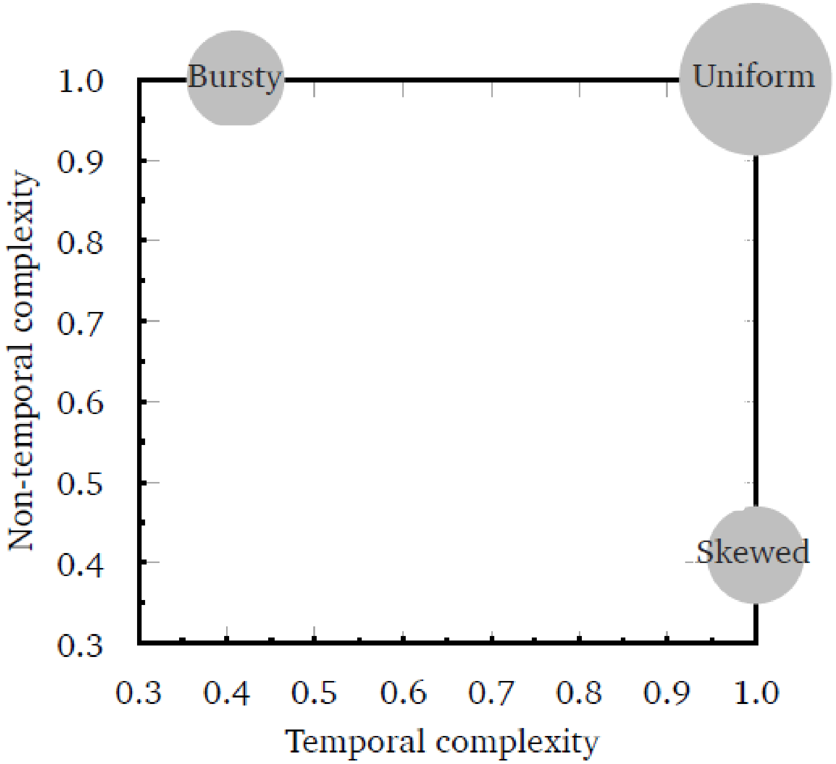
Complexity Map: Entropy („complexity“) of traffic traces.

The Complexity Map



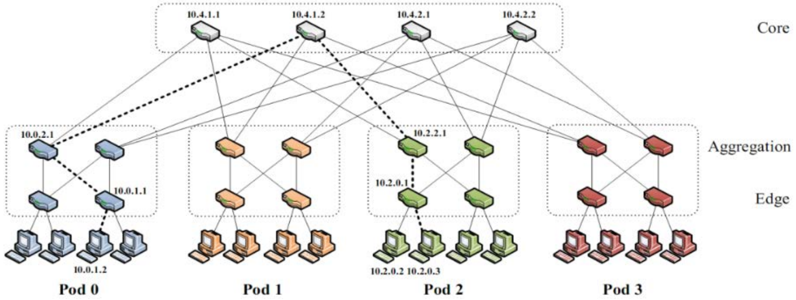
Complexity Map: Entropy („complexity“) of traffic traces.

The Complexity Map

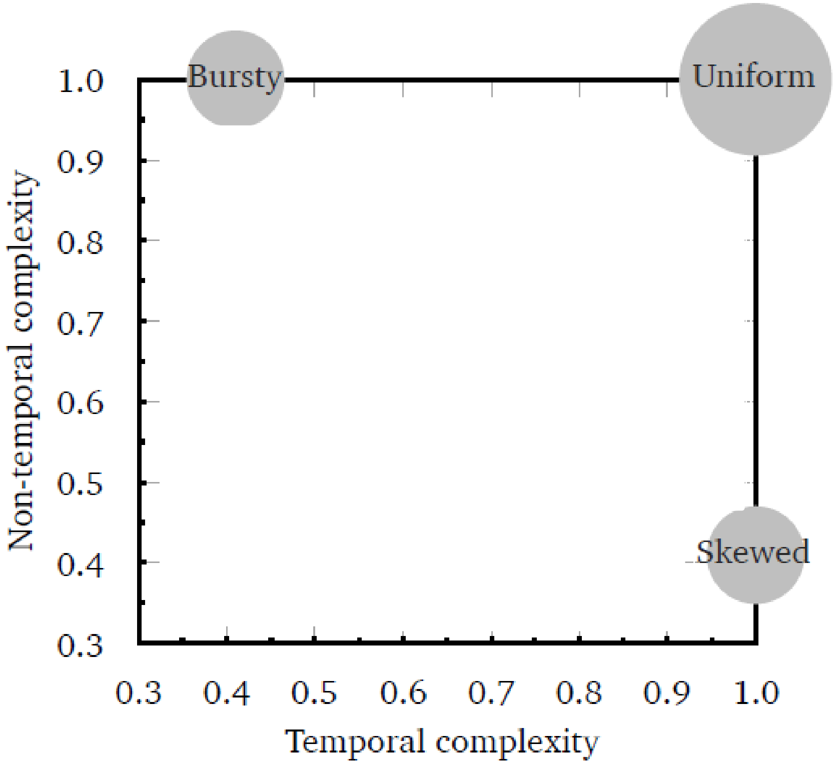


Uniform: Today's datacenters

- Traditional networks are optimized ***for the "worst-case"*** (**all-to-all** communication traffic)
- Example, fat-tree topologies: provide **full bisection bandwidth**

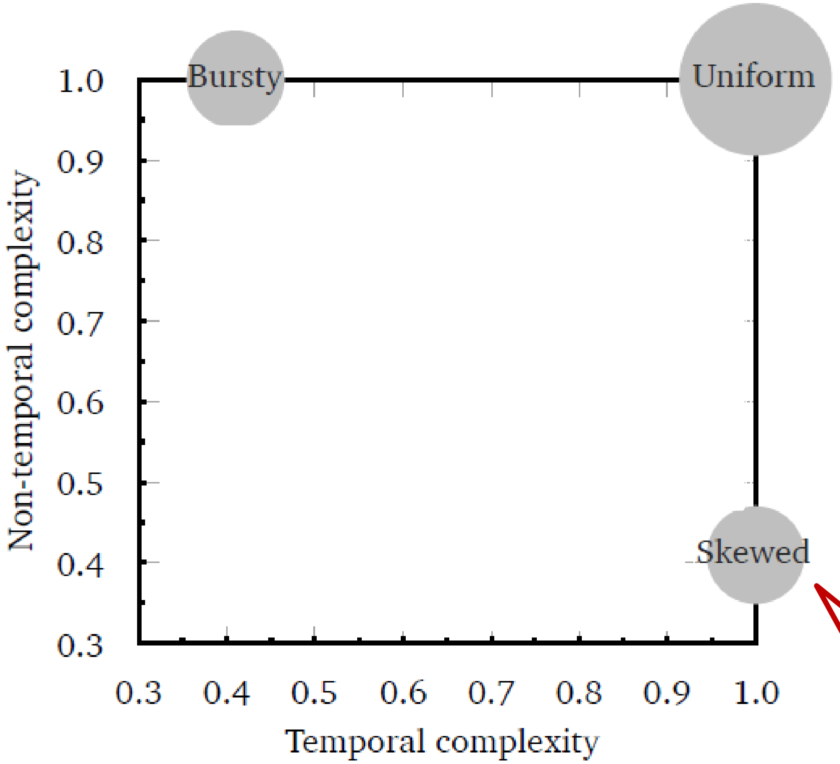


The Complexity Map



Good in the worst case ***but***: cannot leverage different **temporal** and **non-temporal** structures of traffic traces!

The Complexity Map

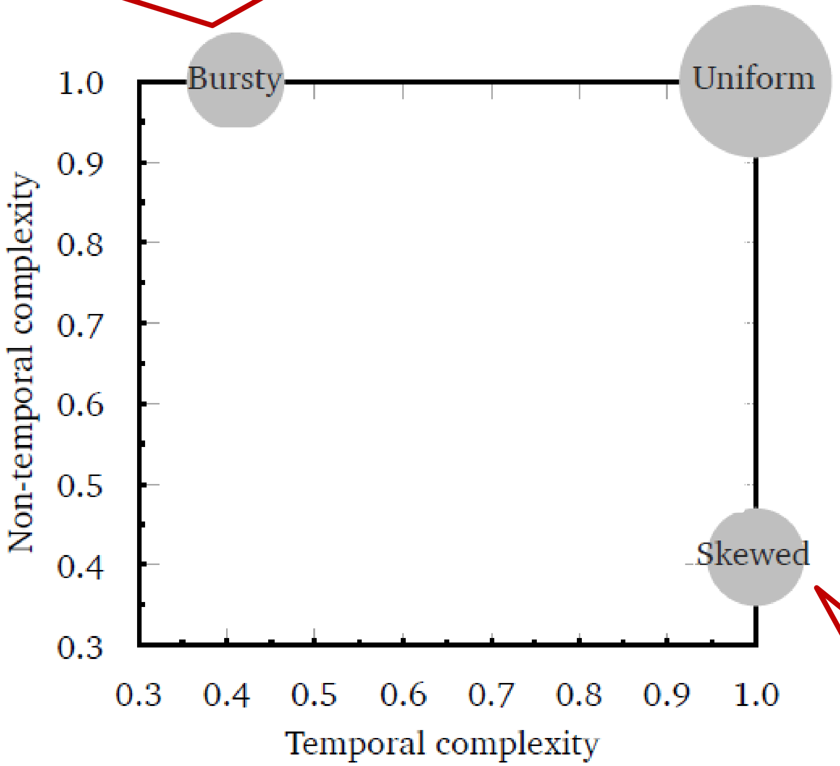


Good in the worst case ***but***: cannot leverage different **temporal** and **non-temporal** structures of traffic traces!

Non-temporal structure could be exploited already with ***static demand-aware networks***!

To exploit **temporal** structure, need ***adaptive demand-aware (“self-adjusting”) networks***.

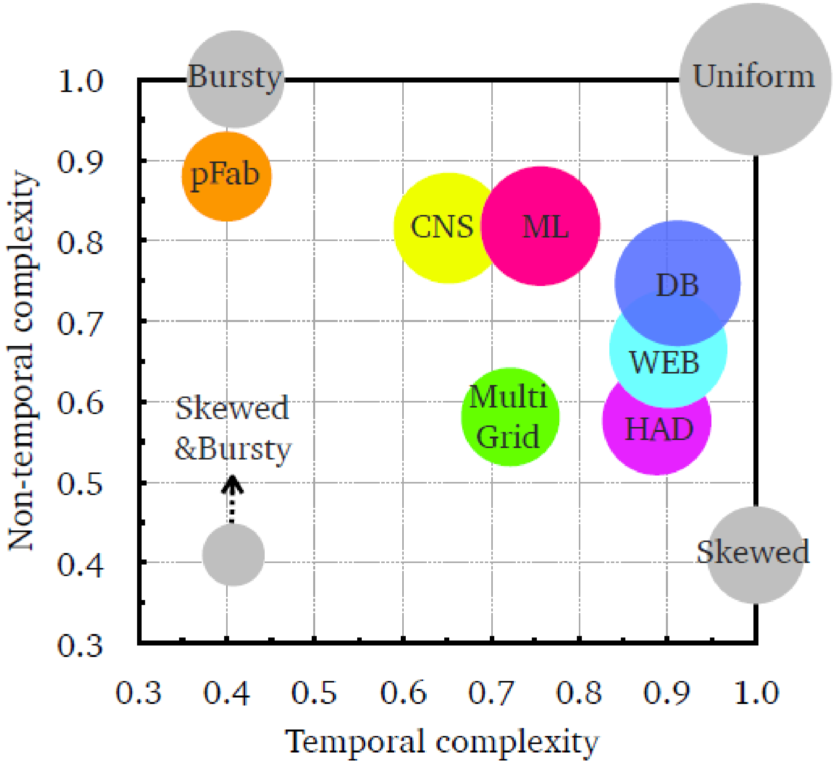
plexity Map



Good in the worst case ***but***: cannot leverage different **temporal** and **non-temporal** structures of traffic traces!

Non-temporal structure could be exploited already with ***static demand-aware networks***!

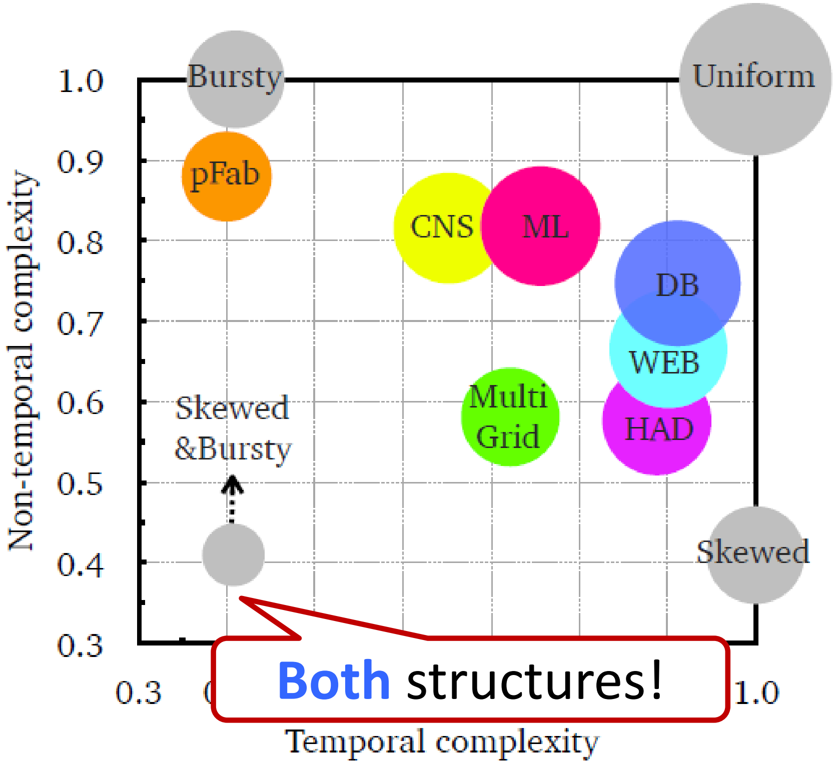
The Complexity Map



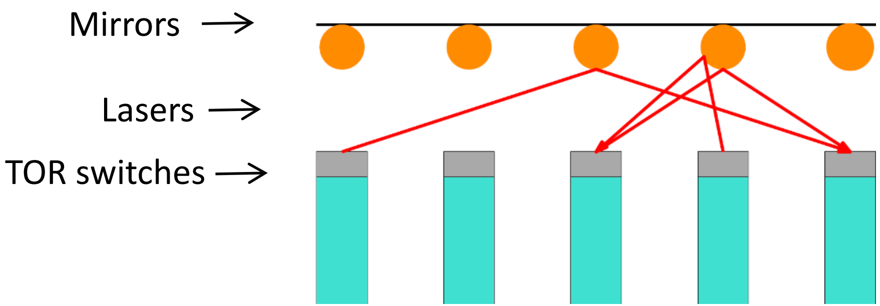
Observation: different applications feature quite significant (and different!) **temporal** and **non-temporal** structures.

- **Facebook** clusters: DB, WEB, HAD
- **HPC** workloads: CNS, Multigrid
- Distributed **Machine Learning** (ML)
- Synthetic traces like **pFabric**

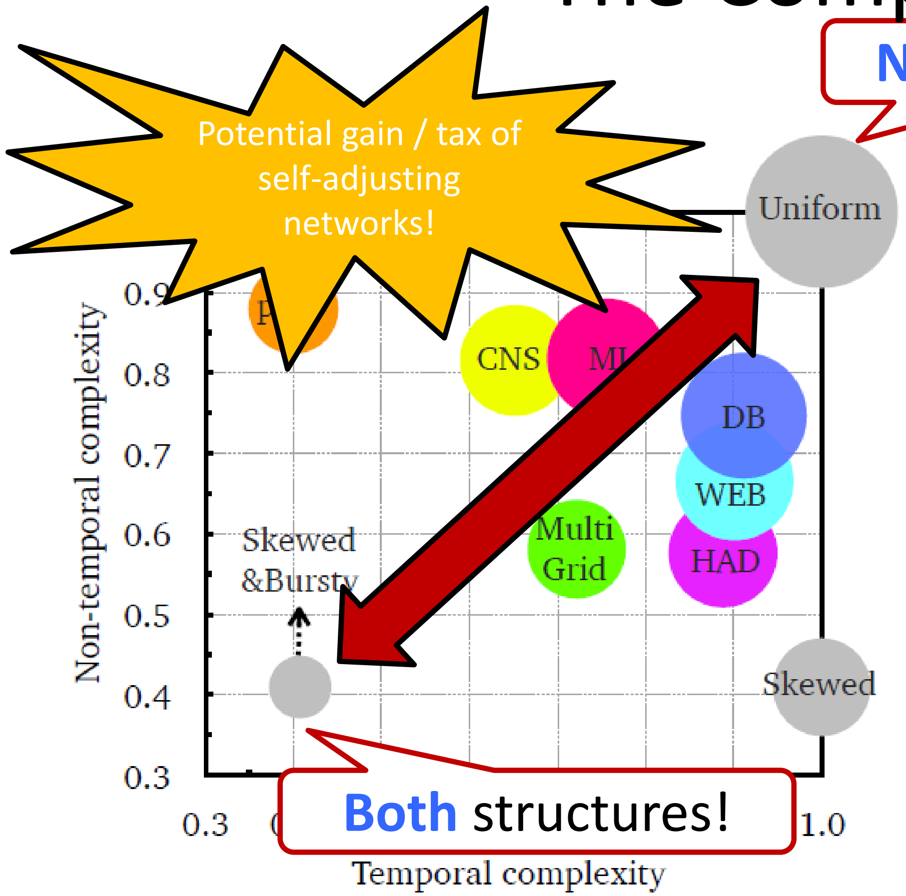
The Complexity Map



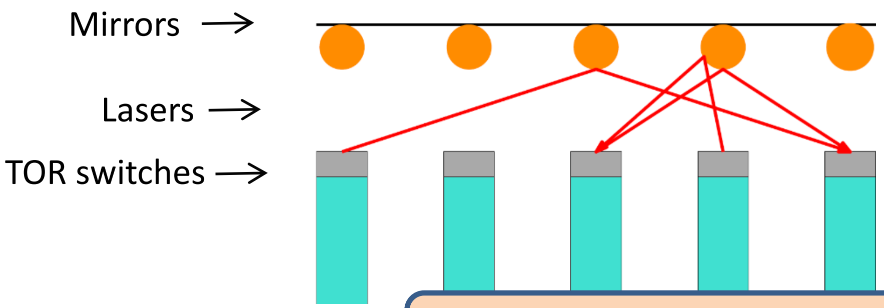
Goal: Design **self-adjusting networks** which leverage ***both*** dimensions of structure!



The Complexity Map



Goal: Design **self-adjusting networks** which leverage ***both*** dimensions of structure!



Measuring the Complexity of Packet Traces.
Avin, Ghobadi, Griner, Schmid. **ArXiv** 2019.

But: How to design DANs which also leverage *temporal structure*?

Inspiration from **self-adjusting datastructures** again!

Roadmap

- Entropy: A metric for demand-aware networks?
  – Empirical motivation
  – A lower bound
  – Algorithms achieving entropy bounds

- From static to dynamic demand-aware networks
  – A connection to self-adjusting datastructures

First: An Analogy

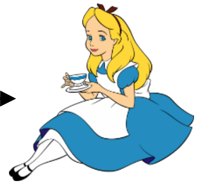
Static vs dynamic demand-aware networks!?

DANs vs **SANs**?

„Coming to ITC in Budapest?“

An Analogy to Coding

00110101…



if demand ***arbitrary*** and ***unknown***

worst case network:
Full BW

log diameter

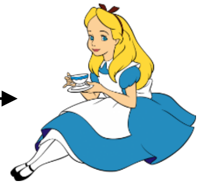
worst case coding:
00, 01, 10, 11

log # bits / symbol

„Coming to ITC in Budapest?“

An Analogy to Coding

01011…



if demand ***arbitrary*** and ***unknown***

worst case network:
Full BW

log diameter

worst case coding:
00, 01, 10, 11

log # bits / symbol

DAN!

if demand ***known*** and ***fixed***

static
Demand-Aware Nets

entropy?

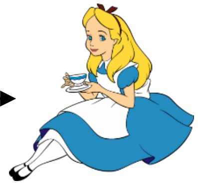
static Huffman:
1, 01, 001, 000

entropy / symbol

An Analogy to Coding

„Coming to ITC in Budapest?“

011…



if demand ***arbitrary*** and ***unknown***

worst case network:
Full BW

worst case coding:
00, 01, 10, 11

log diameter

log # bits / symbol

DAN!

Dynamic DANs:
Aka. **Self-Adjusting Networks** (SANs)!

SAN!

if demand ***known*** and ***fixed***

entropy?

entropy / symbol

static
Demand-Aware Nets

static Huffman:
1, 01, 001, 000

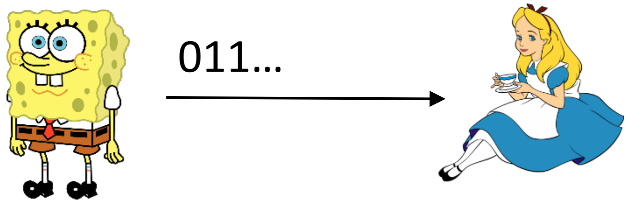
if demand ***unknown*** but ***reconfigurable***

dynamic
Demand-Aware Nets

dynamic
Huffman codes

An Analogy to Coding

„Coming to ITC in Budapest?“



if demand ***arbitrary*** and ***unknown***

worst case network:
Full BW

worst case coding:
00, 01, 10, 11

log diameter

log # bits / symbol

DAN!

SAN!

Dynamic DANs:
Aka. **Self-Adjusting Networks** (SANs)!

if demand ***known*** and ***fixed***

static
Demand-Aware Nets

static Huffman:
1, 01, 001, 000

Can exploit
spatial locality!

if demand ***unknown*** but ***reconfigurable***

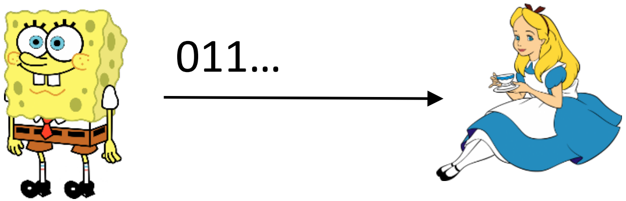
dynamic
Demand-Aware Nets

dynamic
Huffman codes

Additionally exploit
temporal locality!

„Coming to ITC in Budapest?“

An Analogy to Coding



if demand ***arbitrary*** and ***unknown***

worst case network:
Full BW

worst case coding:
00, 01, 10, 11

log diameter

log # bits / symbol

Dynamic DANs:
Aka. **Self-Adjusting Networks** (SANs)!

DAN!

SAN!

if demand ***know***

„Cheating“: need to know demand!

if demand ***unknown*** but ***r***

Need online algorithms!

Aware Nets

static Huffman:
1, 01, 001, 000

dynamic
Demand

Huf an codes

Can exploit
spati

Additionally exploit
temporal locality!

Analogous to *Datastructures*: Oblivious…

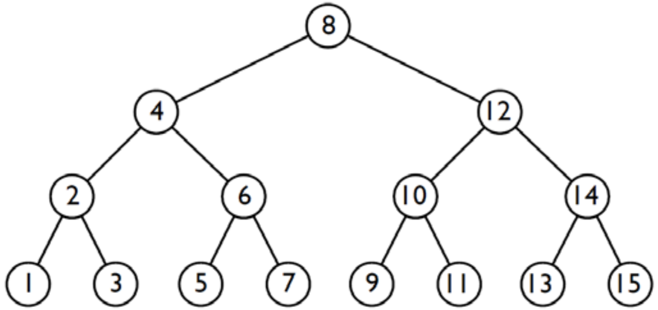
- Traditional, **fixed** BSTs do not rely on any assumptions on the demand
- Optimize for the **worst-case**
- Example **demand**:

  1,…,1,3,…,3,5,…,5,7,…,7,…,log(n),…,log(n)

  many *many* *many* *many* *many*

- Items stored at ***O(log n)*** from the root, **uniformly** and **independently** of their frequency

Corresponds to *max possible demand*!



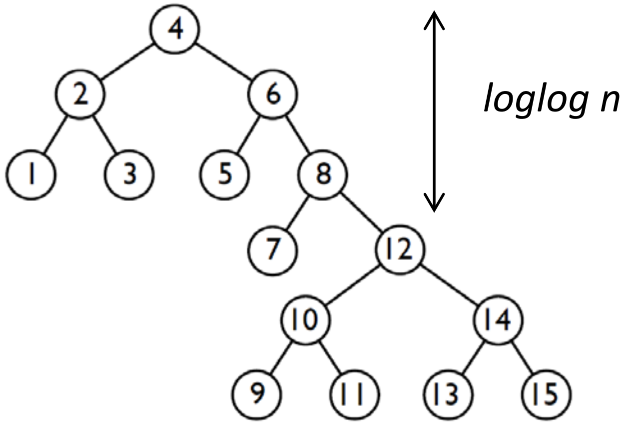
Many requests for leaf 1…

… then for leaf 3…

… Demand-Aware …

- **Demand-aware fixed** BSTs can take advantage of ***spatial locality*** of the demand
- E.g.: place frequently accessed elements close to the root
- E.g., **Knuth/Mehlhorn/Tarjan** trees
- Recall example **demand**: 1,…,1,3,…,3,5,…,5,7,…,7,…,log(n),…,log(n)
  - Amortized cost ***O(loglog n)***

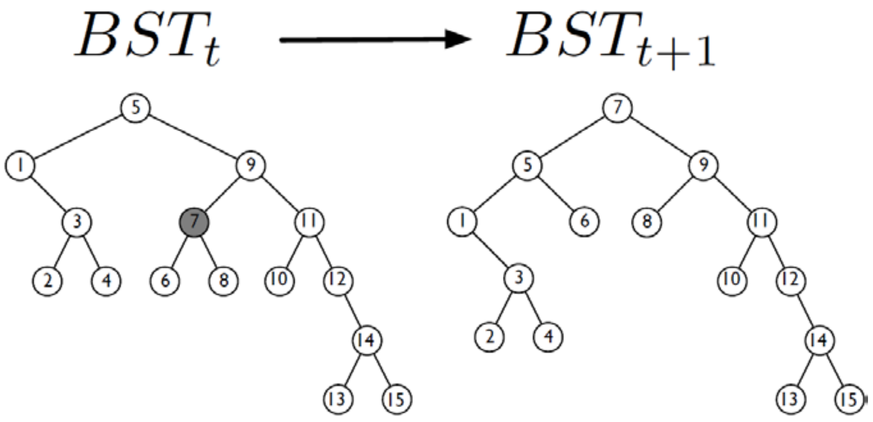
Amortized cost corresponds to ***empirical entropy of demand***!



… Self-Adjusting!

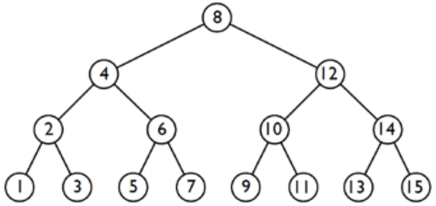
- **Demand-aware reconfigurable** BSTs can additionally take advantage of ***temporal locality***

- By moving accessed element to the root: amortized cost is ***constant***, i.e., O(1)
  - Recall example **demand**: 1,…,1,3,…,3,5,…,5,7,…,7,…,log(n),…,log(n)



Datastructures

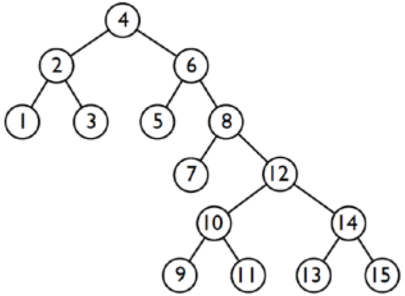
Oblivious



Lookup

O(log n)

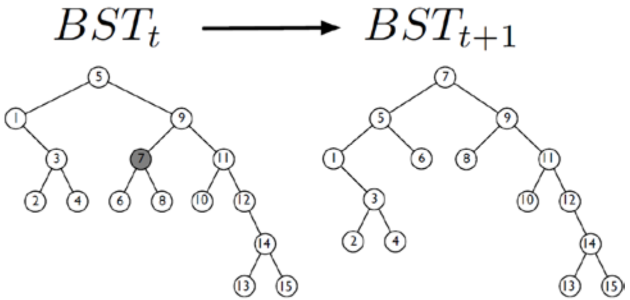
Demand-Aware



Exploit **spatial locality**:

empirical entropy O(loglog n)

Self-Adjusting

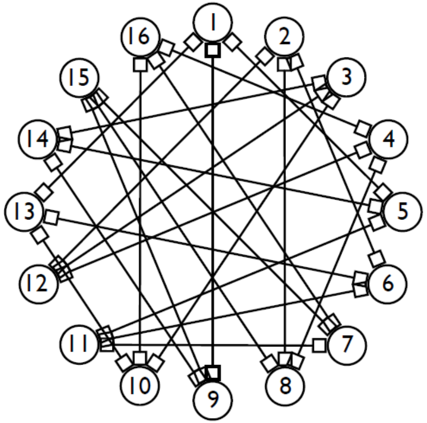


Exploit **temporal locality** as well:

O(1)

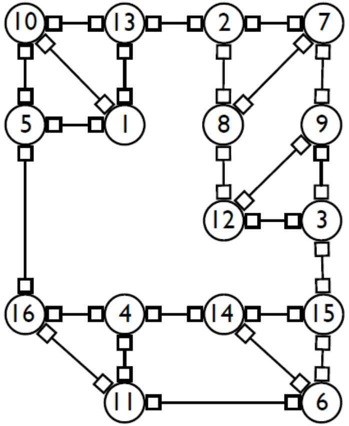
Analogously for Networks

Oblivious



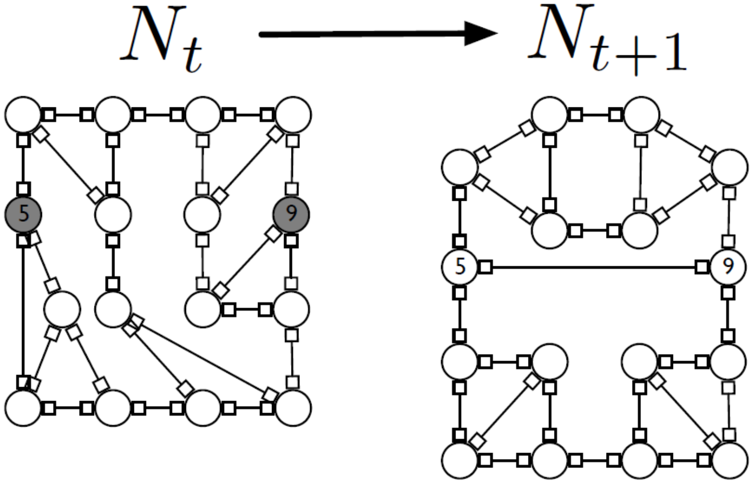
Const degree
(e.g., **expander**):
route lengths ***O(log n)***

DAN



Exploit **spatial locality**

SAN



Exploit **temporal locality** as well

Avin, S.: Toward Demand-Aware Networking: A Theory for Self-Adjusting Networks. **SIGCOMM CCR** 2018.

Now: Design of Self-Adjusting Networks (SANs)

Inspiration from **self-adjusting datastructures** again!

What’s the model?

What’s the model?

Again: *it depends…* ☺

The Problem Input

A **sequence σ** = (u1,v1), (u2,v2), (u3,v3)….

chosen arbitrarily

Chosen i.i.d. from initially unknown fixed distribution

The Problem Input

A **sequence σ** = (u1,v1), (u2,v2), (u3,v3)….

chosen arbitrarily

Chosen i.i.d. from initially unknown fixed distribution

revealed online

given offline

The Problem Input

A **sequence σ** = (u1,v1), (u2,v2), (u3,v3)….

chosen arbitrarily

Chosen i.i.d. from initially unknown fixed distribution

revealed online

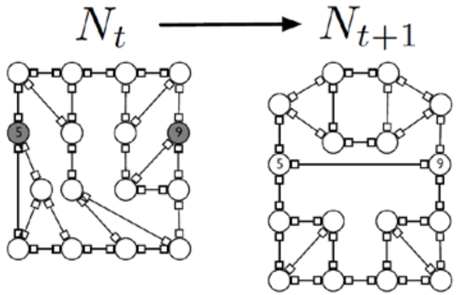
given offline

Other options: sequences of *snapshots*, generated according to *Markov process*, …

What’s the objective? Metric?

Also here: *it depends…* ☺

A Cost-Benefit Tradeoff



Short routes
High reconfiguration cost

Tradeoff

Low reconfiguration cost
Long routes

Basic question:

How often to reconfigure?

A Metric

Entropy of the demand again…

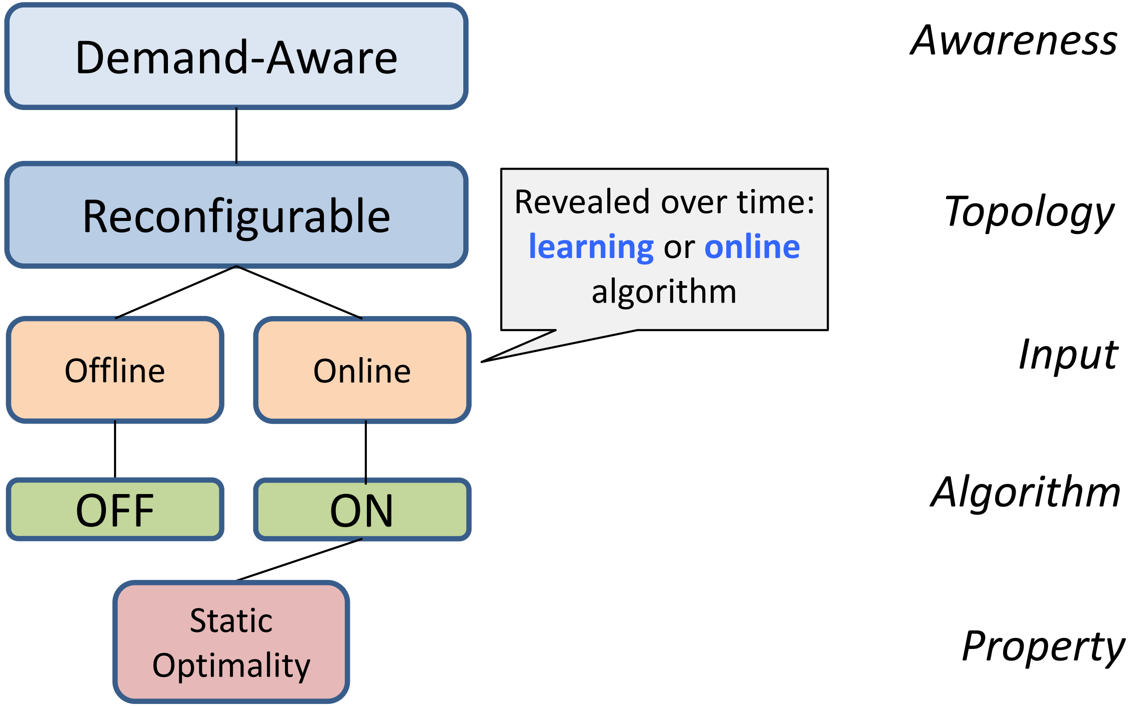
… but now **entropy rate** (entropy over time)!

A Taxonomy: Reconfigurable Networks

Static Optimality:

"***Not worse than static*** which knows demand ahead of time!"

ρ = Cost(ON)/Cost(STAT*) is constant.

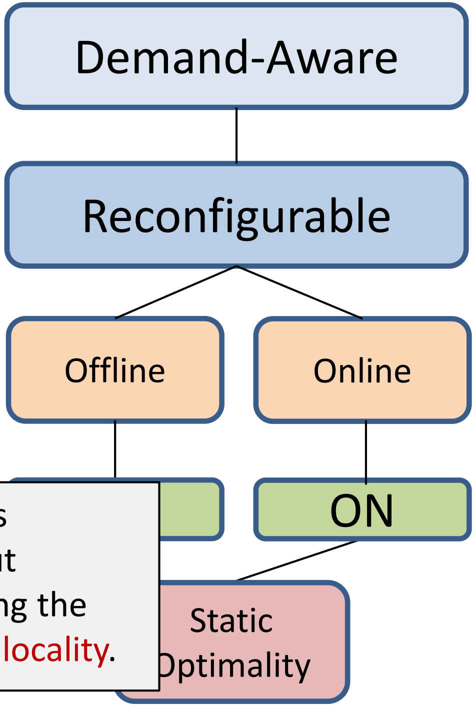


A Taxonomy: Reconfigurable Networks

Static Optimality:

“***Not worse than static*** which knows demand ahead of time!”

ρ = Cost(ON)/Cost(STAT*) is constant.



Revealed over time: **learning** or **online** algorithm

Awareness

Topology

Input

Algorithm

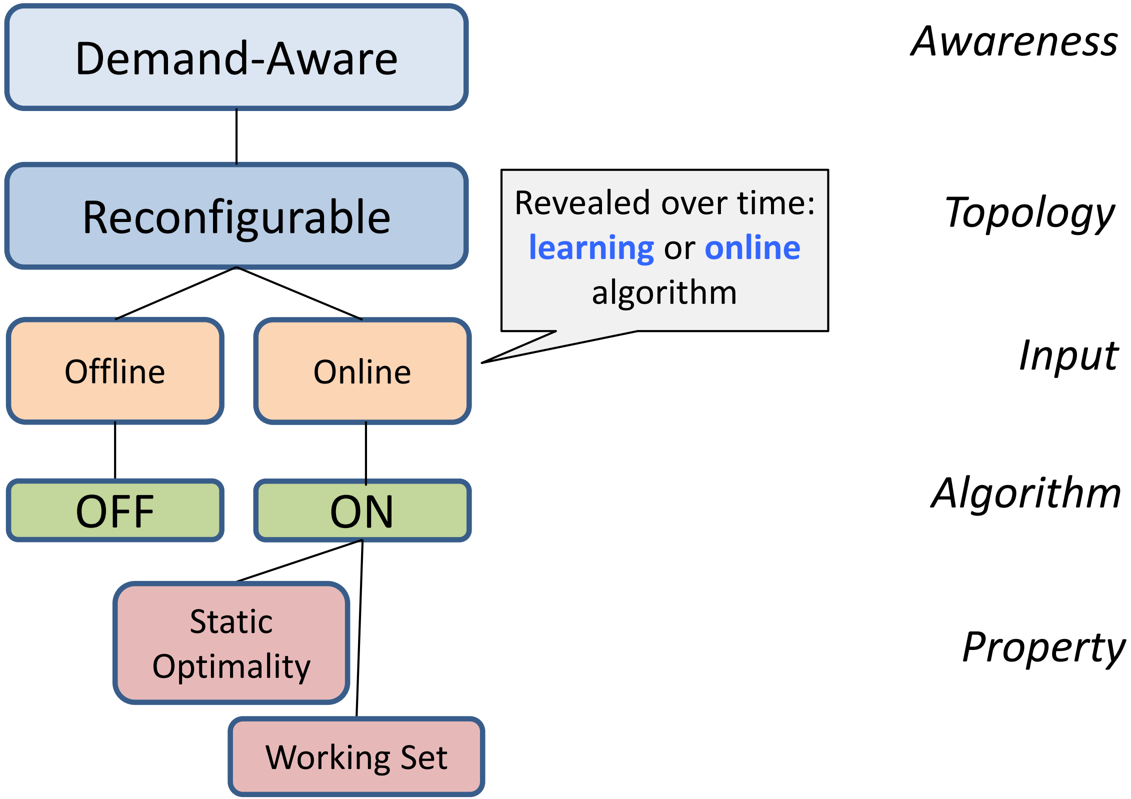
Property

Note: may be ***<<1***. ON has advantage of adjusting, but the disadvantage of not knowing the workload. E.g. if much temporal locality.

A Taxonomy: Reconfigurable Networks

Working Set Property:

"***Topological distance*** between nodes ***proportional to how recently they communicated***!"



A Taxonomy: Reconfigurable Networks

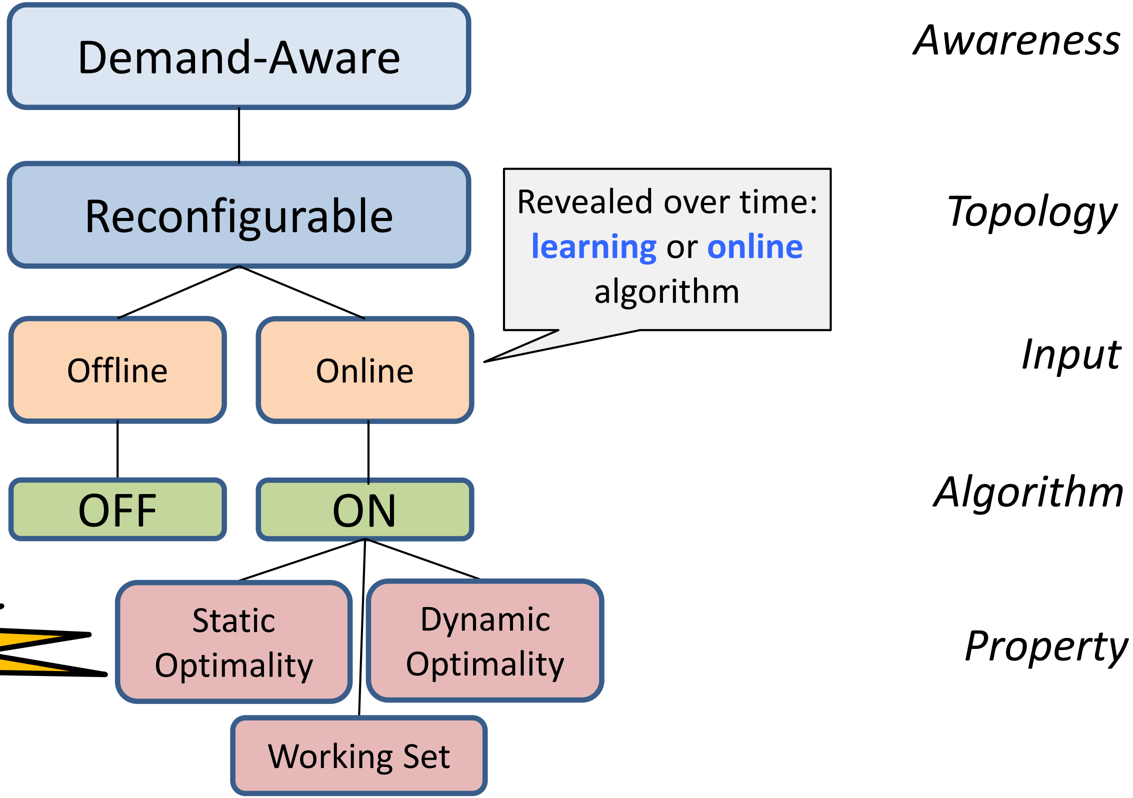
Dynamic Optimality:

"***No worse than an offline*** algorithm which ***knows the sequence***!"

ρ = Cost(ON)/Cost(OFF*) is constant.

Always >=1.

The holy grail!



Algorithms for Self-Adjusting Networks

Algorithms for Self-Adjusting Networks

Let us start with ***trees*** again:

Self-adjusting tree?

Algorithms for Self-Adjusting Networks

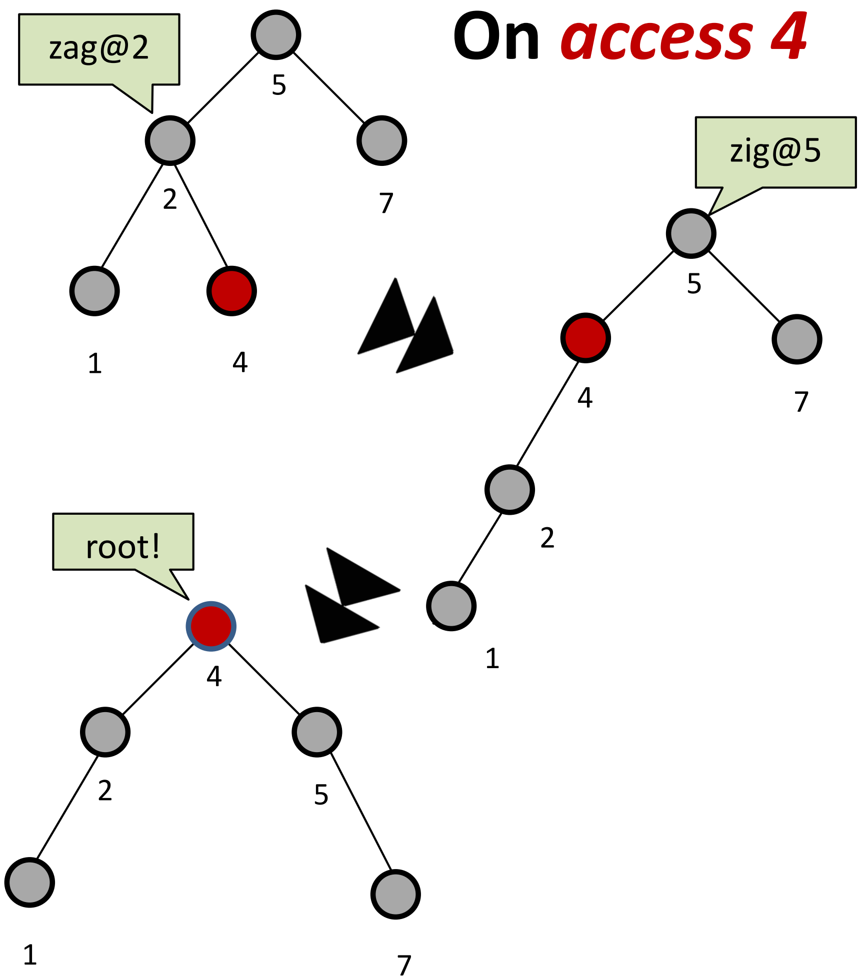
Let us start with ***trees*** again:

Self-adjusting tree?

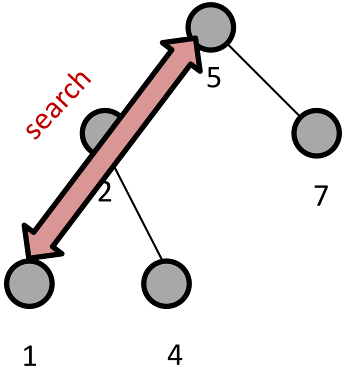
Use **self-adjusting BST**!

Recall: Splay Tree

- A Binary Search Tree (**BST**)
- Inspired by "**move-to-front**": move ***to root***!
- Self-adjustment: **zig**, **zigzig**, **zigzag**
  - Maintains ***search property***
- Many nice properties
  - **Static optimality**, **working set**, (static,dynamic) **fingers**, …

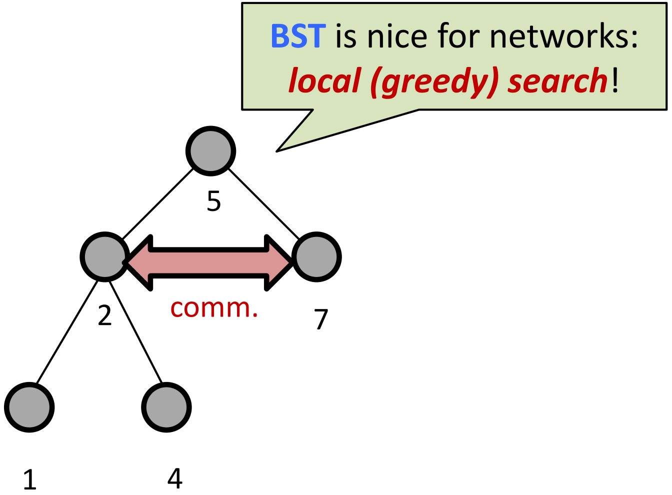


A Simple Idea: Generalize Splay Tree To *SplayNet*



Splay Tree

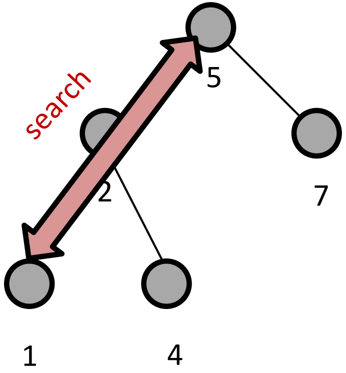
vs



SplayNet

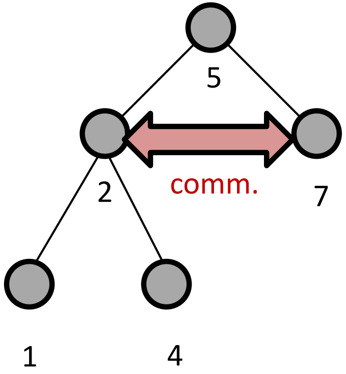
A Simple Idea:
Generalize Splay Tree To *SplayNet*

But how?



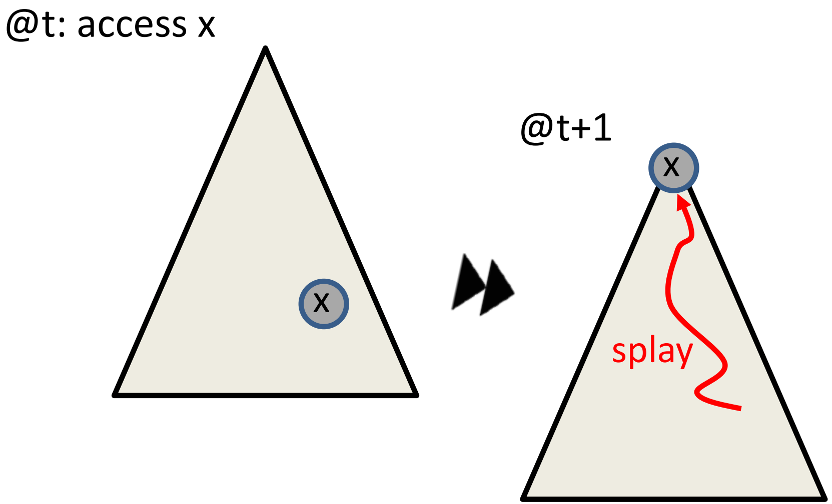
Splay Tree

VS

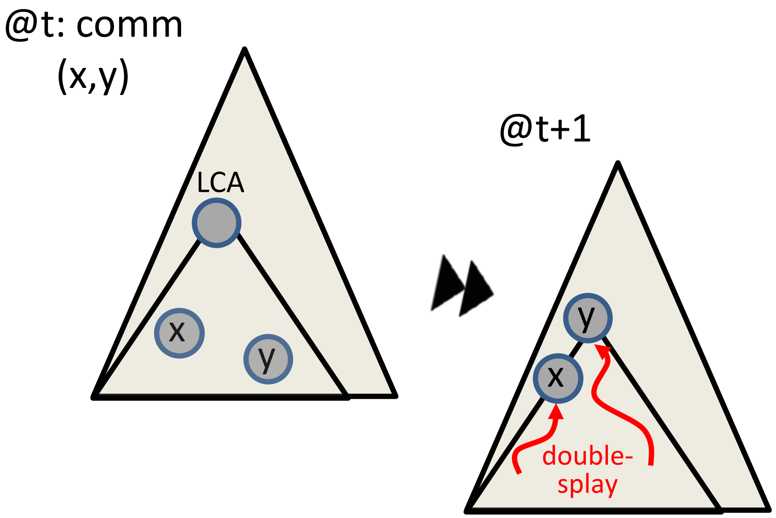


SplayNet

SplayNet: A Simple Idea

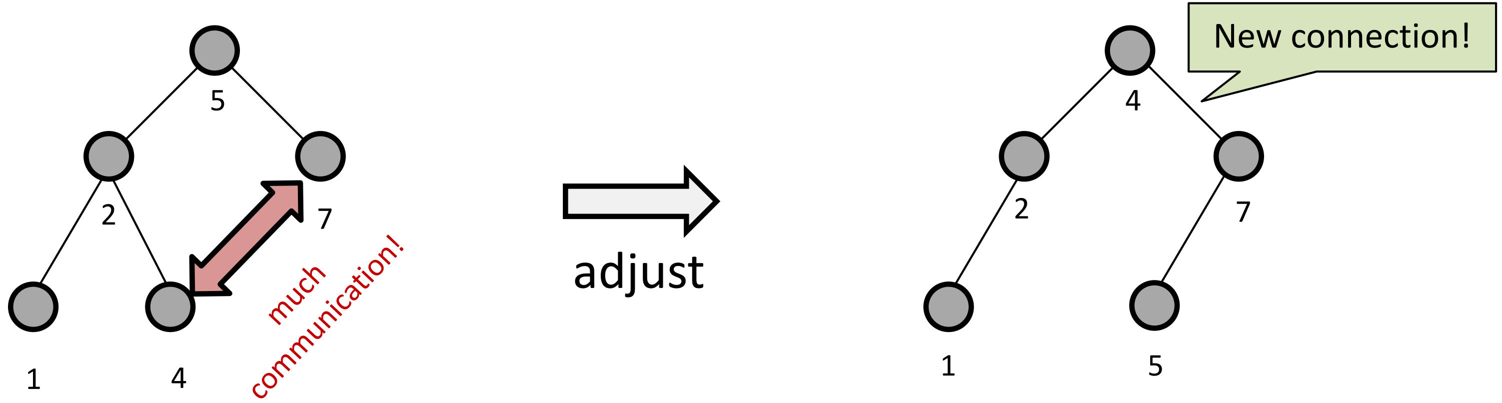


Splay Tree



SplayNet

Example



t=1

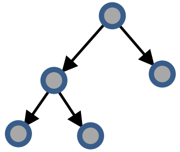
t=2

Challenges: How to **minimize reconfigurations**?
How to keep network **locally** routable?

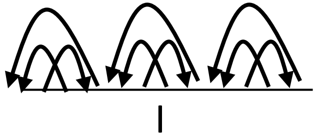
Properties of SplayNets

- **Statically optimal** if demand comes from a ***product distribution***
  - Product distribution: entropy equals conditional entropy, i.e., $H(X)+H(Y)=H(X|Y)+H(X|Y)$
- Converges to optimal static topology in
  - **Multicast scenario**: requests come from a binary tree as well
  - **Cluster scenario**: communication only within interval
  - **Laminated scenario** : communication is „non-crossing matching“

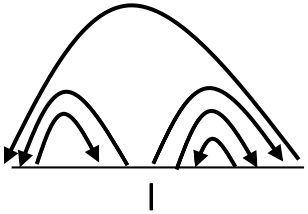
Multicast Scenario



Cluster Scenario



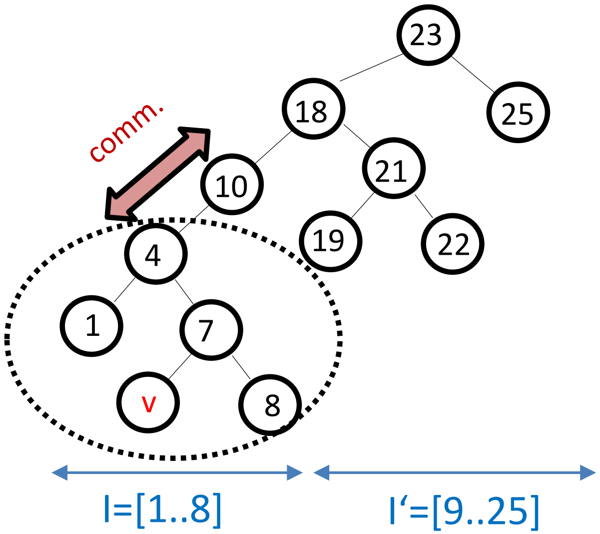
Laminated Scenario



Remark: Static SplayNet

Theorem: Optimal static SplayNet can be computed in polynomial-time (dynamic programming)

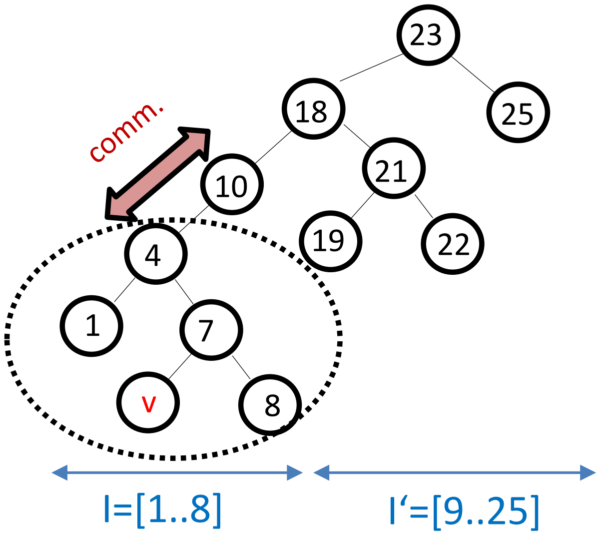
– **Unlike unordered tree?**



Remark: Static SplayNet

Theorem: Optimal static SplayNet can be computed in polynomial-time (dynamic programming)

- **Unlike unordered tree?**



SplayNet: Towards Locally Self-Adjusting Networks. Schmid et al. IEEE/ACM Transactions on Networking (**TON**), Volume 24, Issue 3, 2016.

Algorithms for Self-Adjusting Networks II

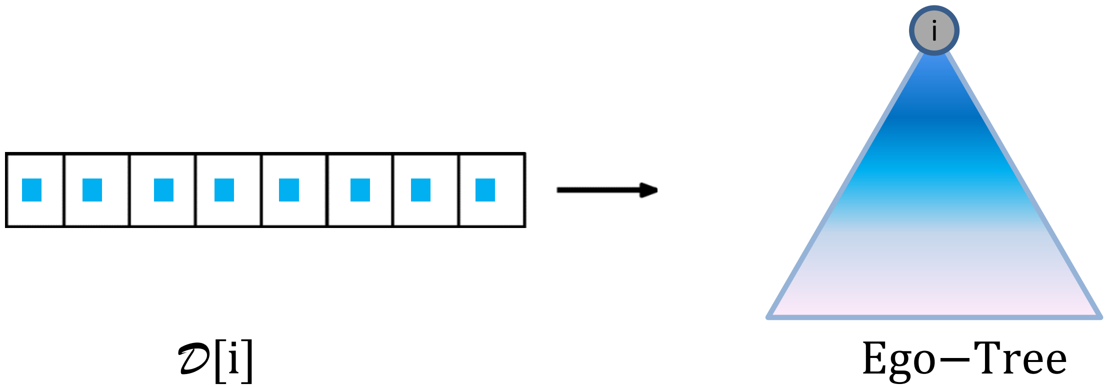
From trees to networks!

Algorithms for Self-Adjusting Networks II

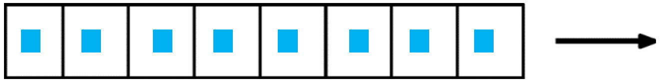

From trees to networks!

Ego-trees strike back!

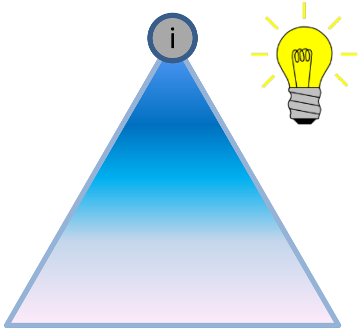
Total Recall: Ego-Trees!



Total Recall: Ego-Trees!



$\mathcal{D}$[i]



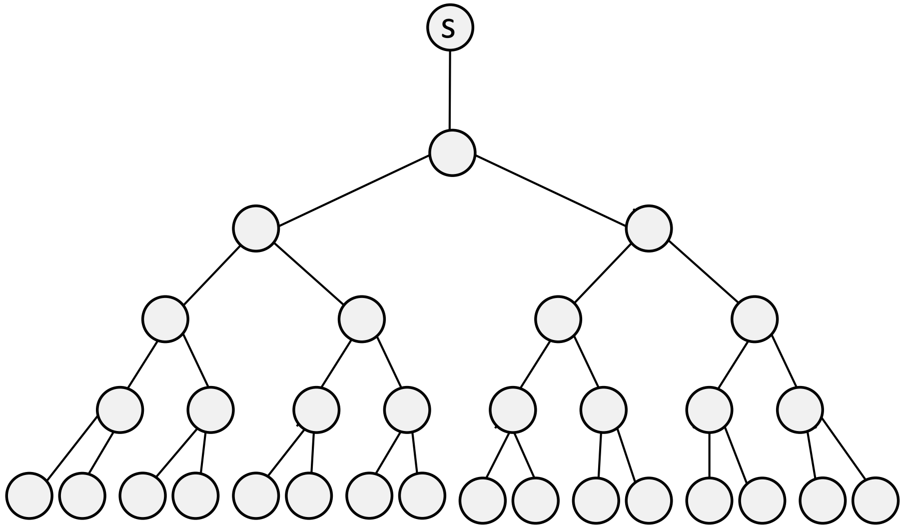
Ego−Tree

Idea: use our old approach but now let each node ***adjust its ego-tree***!

A Dynamic Ego-Tree:
Splay Tree

An Alternative Dynamic Ego-Tree: Push-Down Tree

- **Push-down tree:** a self-adjusting complete tree
- ***Dynamically optimal***
- Not ordered: requires **a map**

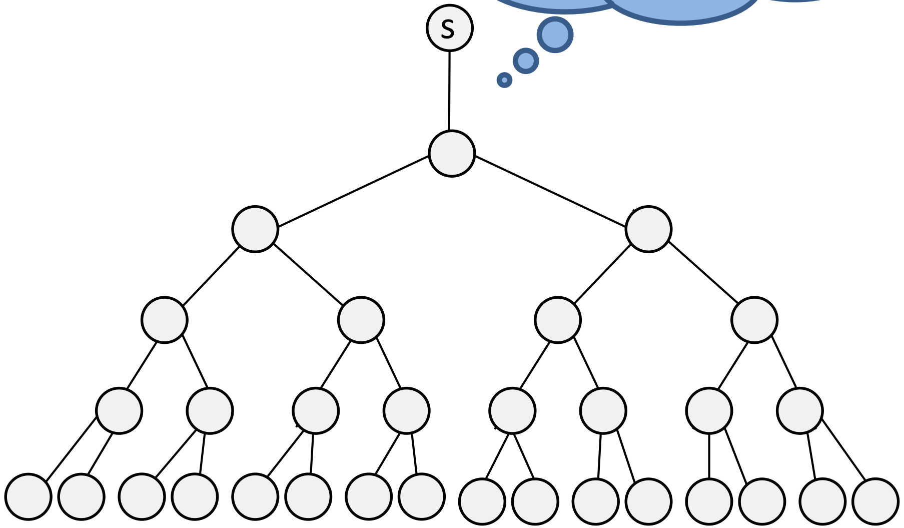


An Alternative Dynamic Ego-Tree: Push-Down Tree

Unordered!

- **Push-down tree:** a self-adjusting complete tree
- ***Dynamically optimal***
- Not ordered: requires **a map**

Equivalent: structure fix, moving nodes, not edges

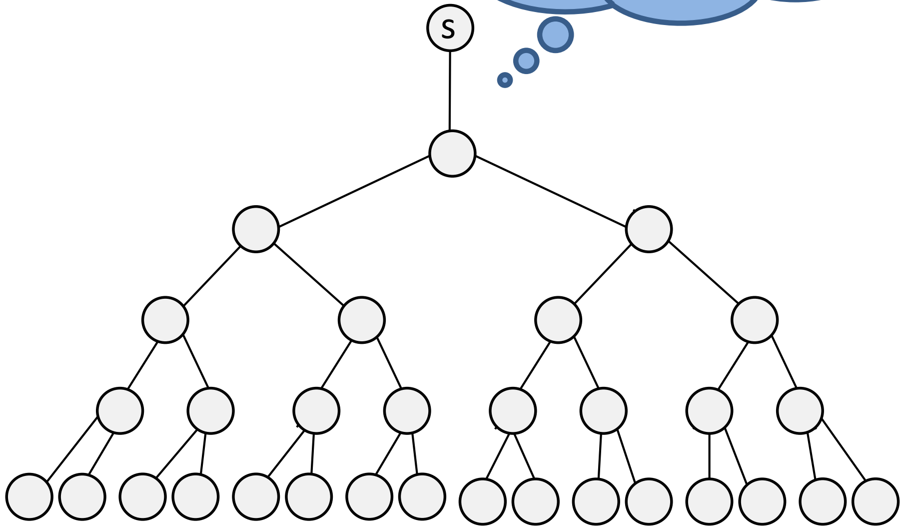


An Alternative Dynamic Ego-Tree: Push-Down Tree

Unordered!

Equivalent: structure fix, moving nodes, not edges

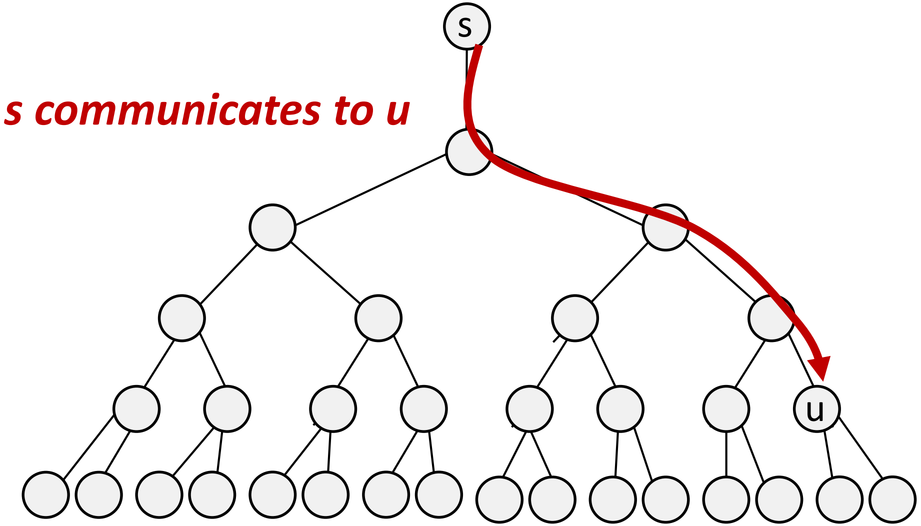
- **Push-down tree:** a self-adjusting complete tree
- ***Dynamically optimal***
- Not ordered: requires **a map**



A useful dynamic property: **Most-Recently Used (MRU)**!
Similar to **Working Set Property**: more recent communication Partners closer to source.

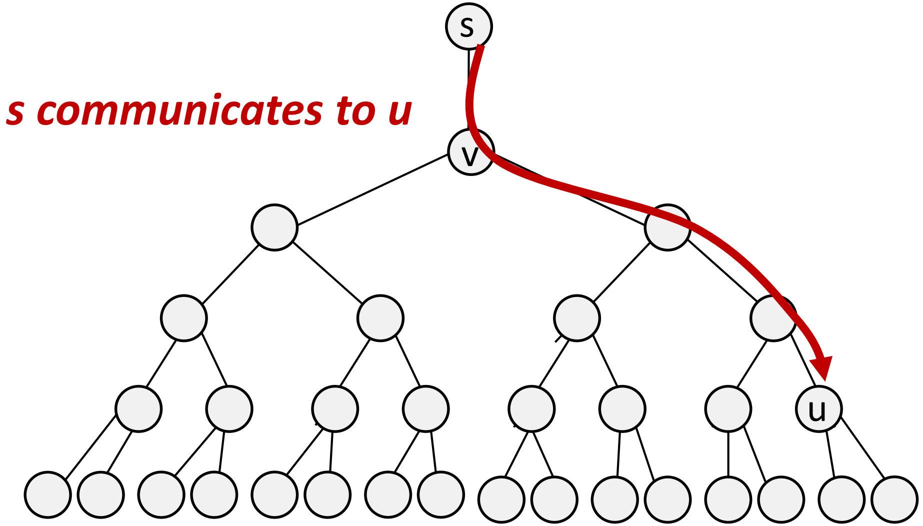
An Alternative Dynamic Ego-Tree: Push-Down Tree

- **Push-down tree:** a self-adjusting complete tree
- ***Dynamically optimal***
- Not ordered: requires **a map**



An Alternative Dynamic Ego-Tree: Push-Down Tree

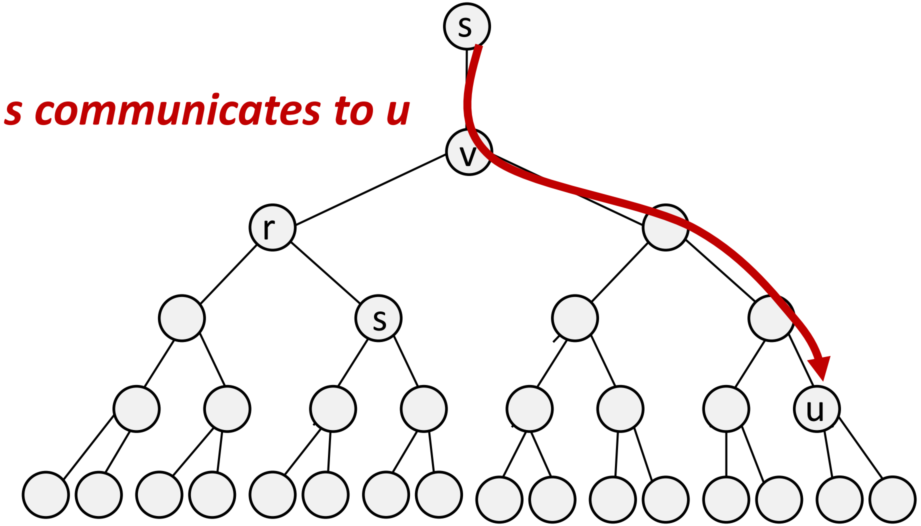
- **Push-down tree:** a self-adjusting complete tree
- ***Dynamically optimal***
- Not ordered: requires **a map**



Strict MRU requires: move u to root! But how? Cannot swap with v: v no longer MRU!

An Alternative Dynamic Ego-Tree: Push-Down Tree

- **Push-down tree:** a self-adjusting complete tree
- ***Dynamically optimal***
- Not ordered: requires **a map**

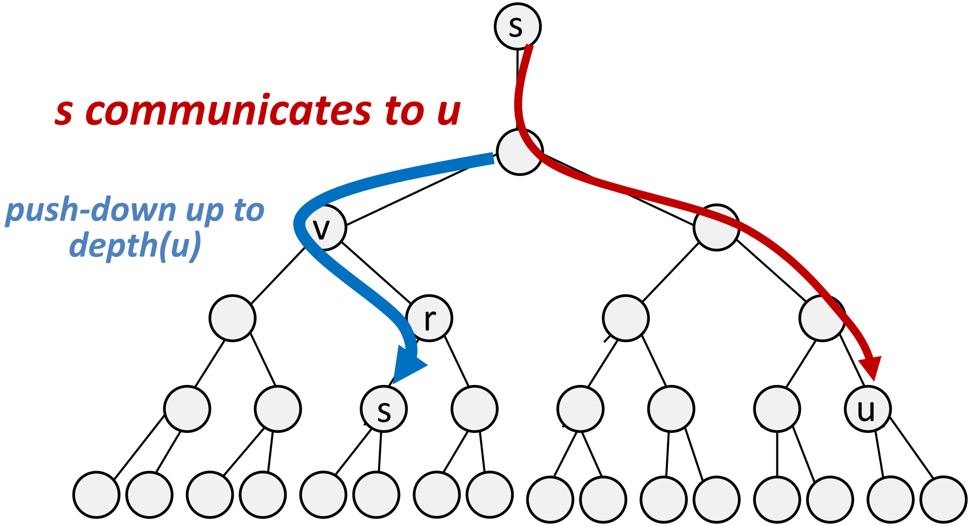


Strict MRU requires: move u to root! But how? Cannot swap with v: v no longer MRU!

Idea: Push v down, in a balanced manner, up to depth(u): left-right-left-right („rotate-push“)

An Alternative Dynamic Ego-Tree: Push-Down Tree

- **Push-down tree:** a self-adjusting complete tree
- ***Dynamically optimal***
- Not ordered: requires **a map**

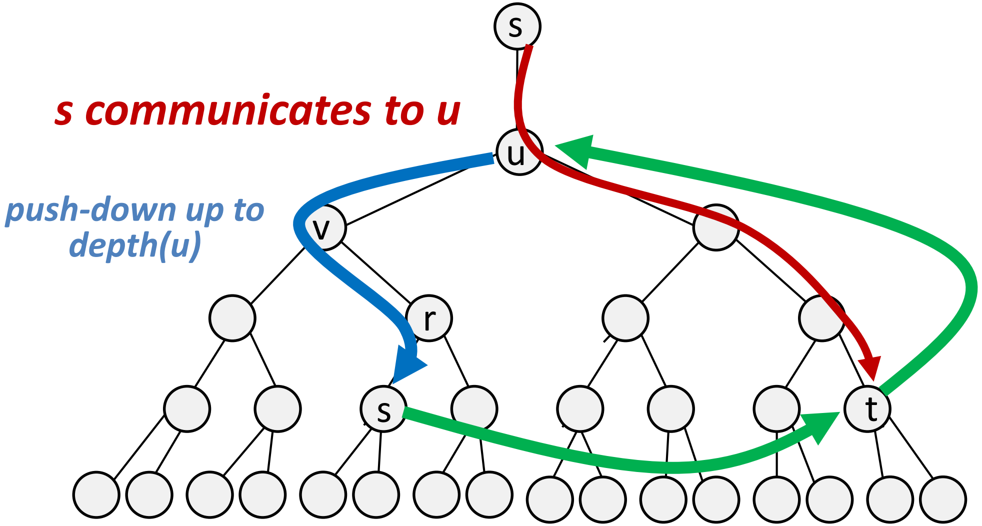


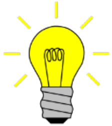
Strict MRU requires: move u to root! But how? Cannot swap with v: v no longer MRU!

Idea: Push v down, in a balanced manner, up to depth(u): left-right-left-right („rotate-push")

An Alternative Dynamic Ego-Tree: Push-Down Tree

- **Push-down tree:** a self-adjusting complete tree
- ***Dynamically optimal***
- Not ordered: requires **a map**

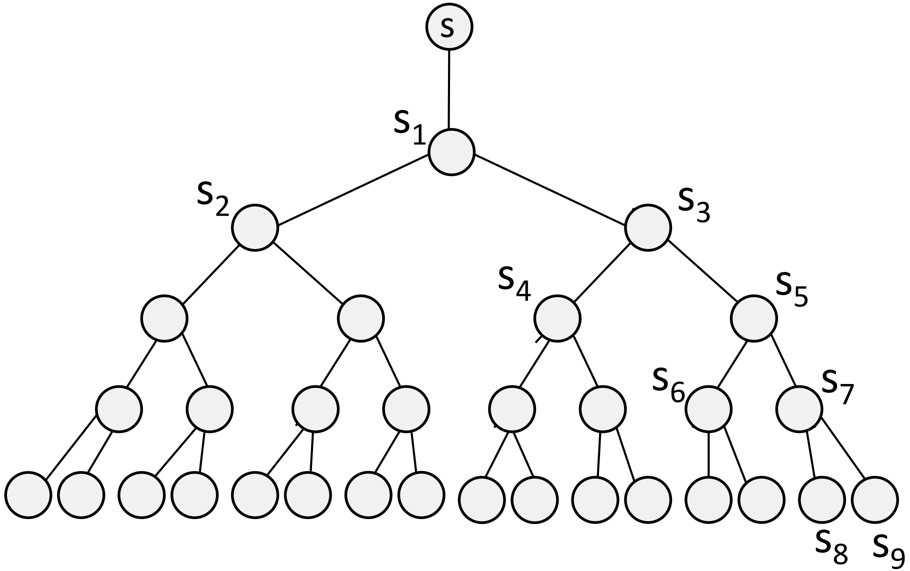




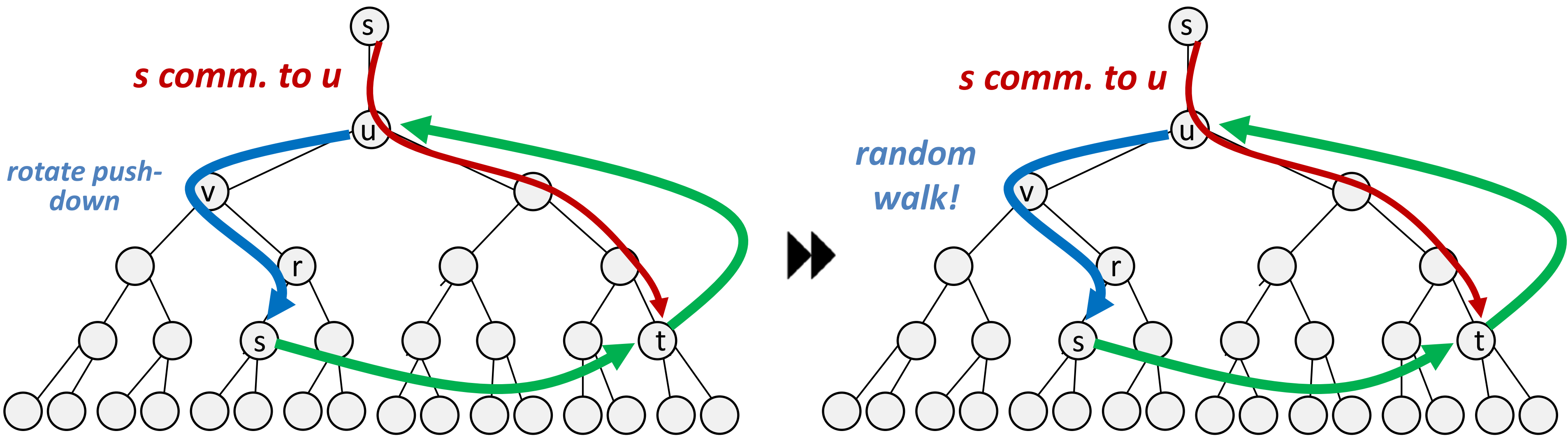
Then: promote u to available root, and t to u: at original depth!

Remarks

- Unfortunately, alternating push-down does ***not maintain MRU*** (working set) property
- Tree can ***degrade***, e.g.: sequence of requests from level 4,1,2,1,3,1,4,1

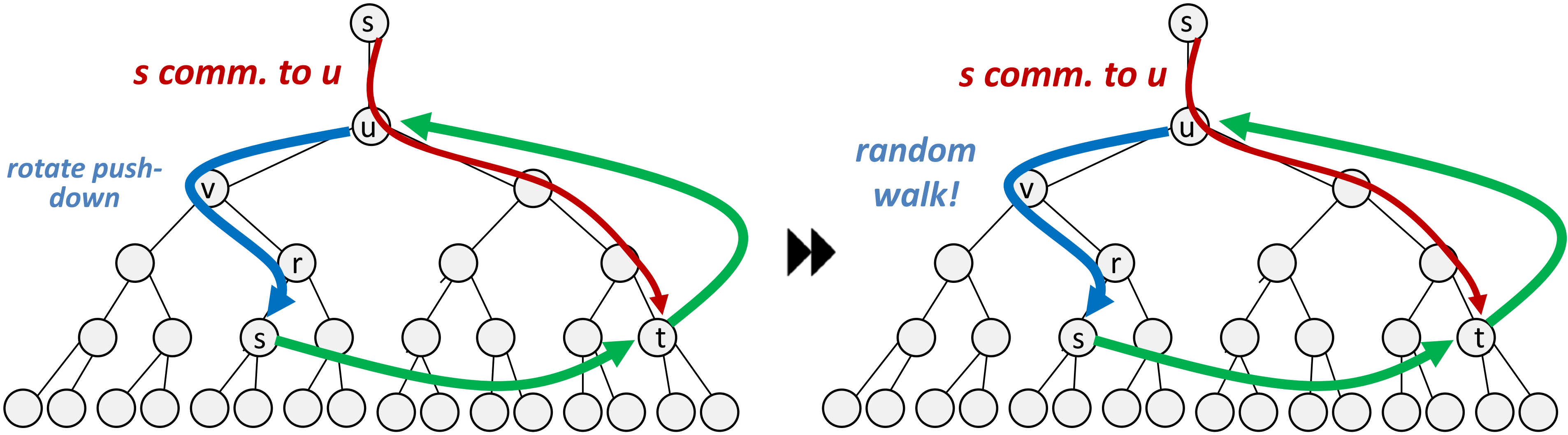


Solution: Random Walk



At least maintains approximate working set / MRU!

Solution: Random Walk



At least maint

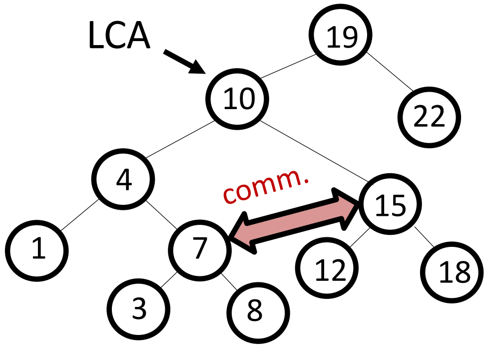
working

Push-Down Trees: Optimal Self-Adjusting Complete Trees
Chen Avin, Kaushik Mondal, and Stefan Schmid.
ArXiv Technical Report, July 2018.

Remark 1: Decentralized Algorithms

A "Simple" Decentralized Solution: Distributed SplayNet (*DiSplayNet*)

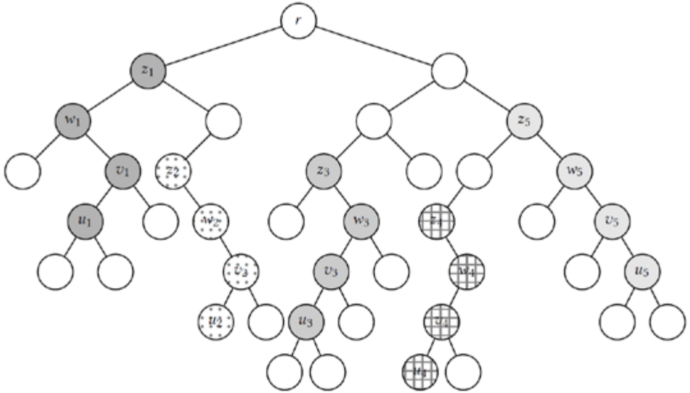
- SplayNet attractive: ordered BST supports **local routing**
  - Nodes ***maintain three ranges***: interval of left subtree, right subtree, upward
- If communicate (frequently): **double-splay** toward LCA
- Challenge: **concurrency**!
  - Access Lemma of splay trees no longer works: ***potential function*** does not „***telescope***" anymore: a concurrently rising node may push down another rising node again



SplayNet

DiSplayNet: Challenges

- DiSplayNet: Rotations (zig,zigzig,zigzag) are ***concurrent***
- To avoid conflict: distributed computation of **independent clusters**
- Still challenging:



	1	2	3	4	5	6	7	8	...	$i-6$	$i-5$	$i-4$	$i-3$	$i-2$	$i-1$	i
σ_1	✓	✓	✓	✓	-	-	-	-	...	-	-	-	-	-	-	-
σ_2	-	✗	✗	✗	✓	✓	✓	-	...	-	-	-	-	-	-	-
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
σ_{m-1}	-	-	-	-	-	-	-	-	...	✓	✓	-	-	-	-	-
σ_m	-	-	-	-	-	-	-	-	...	✗	✗	✓	✓	✓	✓	-

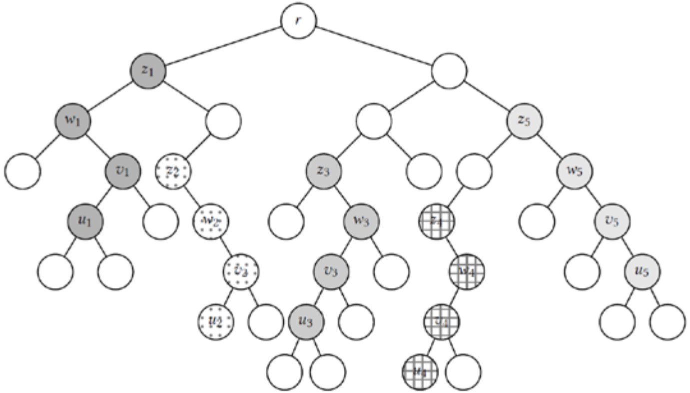
Sequential SplayNet: requests ***one after another***

	1	2	3	...	i	$i+1$	$i+2$	$i+3$	$i+4$	$i+5$	$i+6$	...	j	...	k
s_1	✓	✓	✓	...	✓	✓	✓	✓	✓	✓	...	-	✓	...	-
d_1	✓	✓	✓	...	✓	✓	✓	✓	✓	✓	...	-	✓	...	-
s_2	-	✓	✓	...	✓	✓	✓	✓	-	-	...	-	-	...	-
d_2	-	✓	✓	...	✓	✓	✗	✓	-	-	...	-	-	...	-
s_3	-	-	✓	...	✗	✗	✗	✗	✓	✗	✗	...	✓	...	-
d_3	-	-	✓	...	✗	✗	✗	✗	✗	✗	✗	...	✓	...	-

DiSplayNet: Analysis more challenging: potential function sum no longer **telescopic**. One request can “push-down” another.

DiSplayNet: Challenges

- DiSplayNet: Rotations (zig,zigzig,zigzag) are ***concurrent***
- To avoid conflict: distributed computation of **independent clusters**
- Still challenging:



Telescopic: max potential drop

	1	2	3	4	5	6	7	8	...	$i-6$	$i-5$	$i-4$	$i-3$	$i-2$	$i-1$	i
σ_1	✓	✓	✓	✓				-	...	-	-	-	-	-	-	-
σ_2	-	✗	✗	✗	✓	✓	✓	-	...	-	-	-	-	-	-	-
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
σ_{m-1}	-	-	-	-	-	-	-	-	...	✓	✓					-
σ_m	-	-	-	-	-	-	-	-	...	✗	✗	✓	✓	✓	✓	-

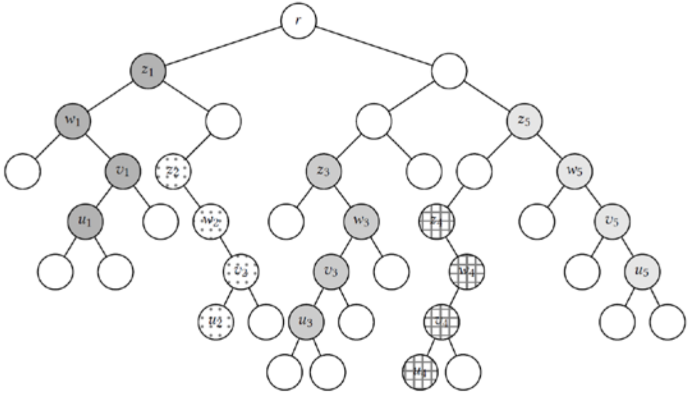
Sequential SplayNet: requests ***one after another***

	1	2	3	...	i	$i+1$	$i+2$	$i+3$	$i+4$	$i+5$	$i+6$	...	j	...	k
s_1	✓	✓	✓	...	✓	✓	✓	✓	✓	✓	...	-	✓	...	-
d_1	✓	✓	✓	...	✓	✓	✓	✓	✓	✓	...	-	✓	...	-
s_2	-	✓	✓	...	✓	✓	✓	✓	-	-	...	-	-	...	-
d_2	-	✓	✓	...	✓	✓	✗	✓	-	-	...	-	-	...	-
s_3	-	-	✓	...	✗	✗	✗	✗	✓	✗	✗	...	✓	...	-
d_3	-	-	✓	...	✗	✗	✗	✗	✗	✗	✗	...	✓	...	-

DiSplayNet: Analysis more challenging: potential function sum no longer **telescopic**. One request can "push-down" another.

DiSplayNet: Challenges

- DiSplayNet: Rotations (zig,zigzig,zigzag) are ***concurrent***
- To avoid conflict: distributed computation of **independent clusters**
- Still challenging:



Telescopic: max potential drop

	1	2	3	4	5	6	7	8	...	$i-6$	$i-5$	$i-4$	$i-3$	$i-2$	$i-1$	i
σ_1	✓	✓	✓	✓				-	...	-	-	-	-	-	-	-
σ_2	-	✗	✗	✗	✓	✓	✓	-	...	-	-	-	-	-	-	-
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
σ_{m-1}	-	-	-	-	-	-	-	-	...	✓	✓					-
σ_m	-	-	-	-	-	-	-	-	...	✗	✗	✓	✓	✓	✓	-

	1	2	3	...	i	$i+1$	$i+2$	$i+3$	$i+4$	$i+5$	$i+6$	...	j	...	k
s_1	✓	✓	✓	...	✓	✓	✓	✓	✓	✓	...	-	✓	...	-
d_1	✓	✓	✓	...	✓	✓	✓	✓	✓	✓	...	-	✓	...	-
s_2	-	✓	✓	...	✓	✓	✓	✓	-	-	...	-	-	...	-
d_2	-	✓	✓	...	✓	✓	✗	✓	-	-	...	-	-	...	-
s_3	-	-	✓	...	✗	✗	✗	✗	✓	✗	✗	...	✓	...	-
d_3	-	-	✓	...	✗	✗	✗	✗	✗	✗	✗	...	✓	...	-

Sequential SplayNet: re

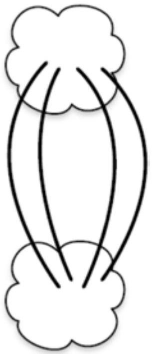
Distributed Self-Adjusting Tree Networks. Bruna Peres, Otavio Augusto de Oliveira Souza, Olga Goussevskaia, Chen Avin, and Stefan Schmid. IEEE **INFOCOM**, 2019.

Remark 2: Accounting for Congestion

A Tradeoff?!

Short routes:
congestion

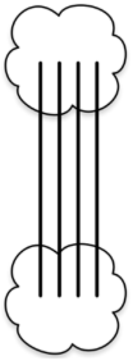
vs



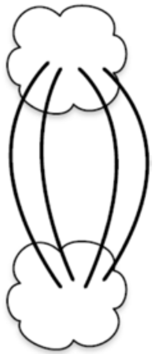
Low congestion:
long routes

A Tradeoff?!

Short routes:
congestion

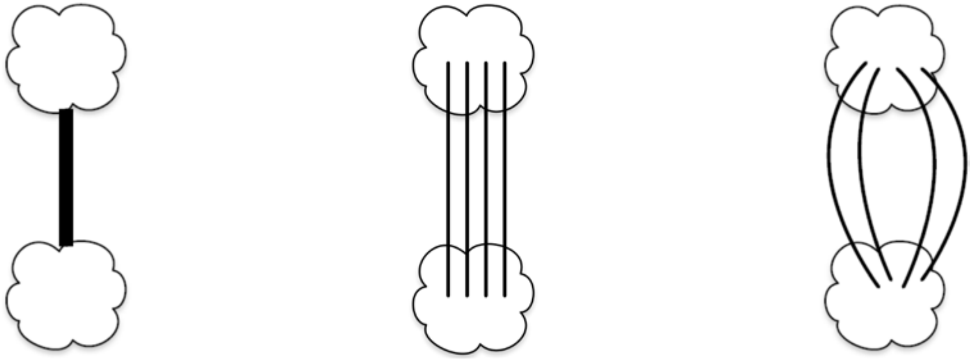


Or both?



Low congestion:
long routes

A Tradeoff?!



Short routes: congestion

Or both?

Low congestion: long routes

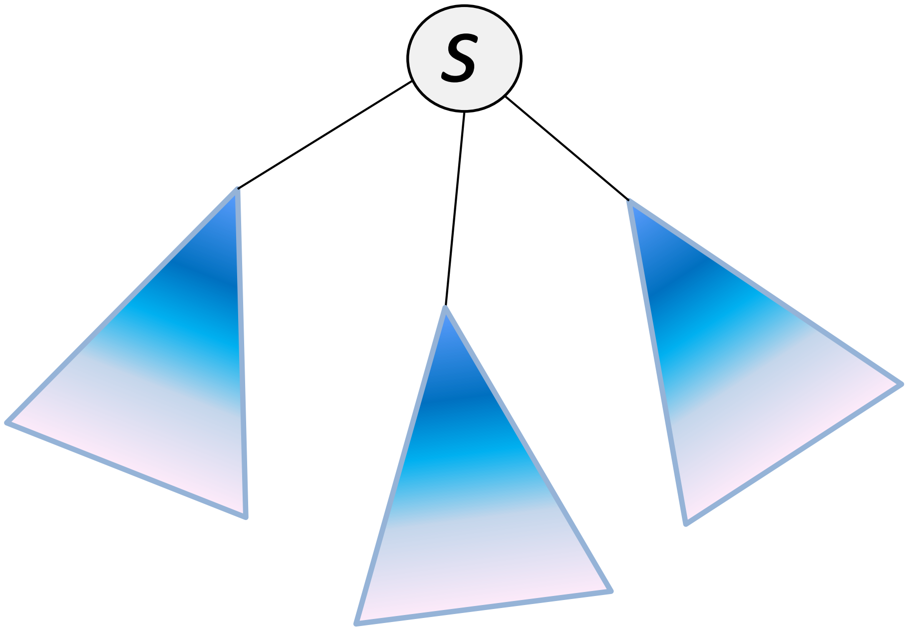
Ego-Tree++!

- Idea: place destination nodes greedily across subtrees s.t. ***congestion balanced***

- … while preserving distance

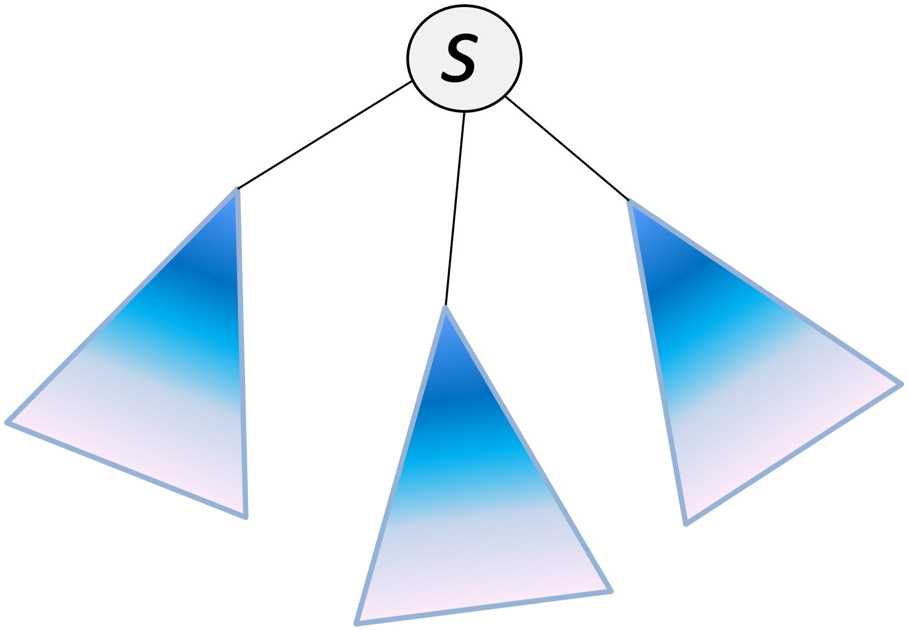
- Trees can have different sizes but ***similar mass***!

- Bicriteria guarantee



Ego-Tree++!

- Idea: place destination nodes greedily across subtrees s.t. ***congestion balanced***
- … while preserving distance
- Trees can have different sizes but ***similar mass***!
- Bicriteria guarantee



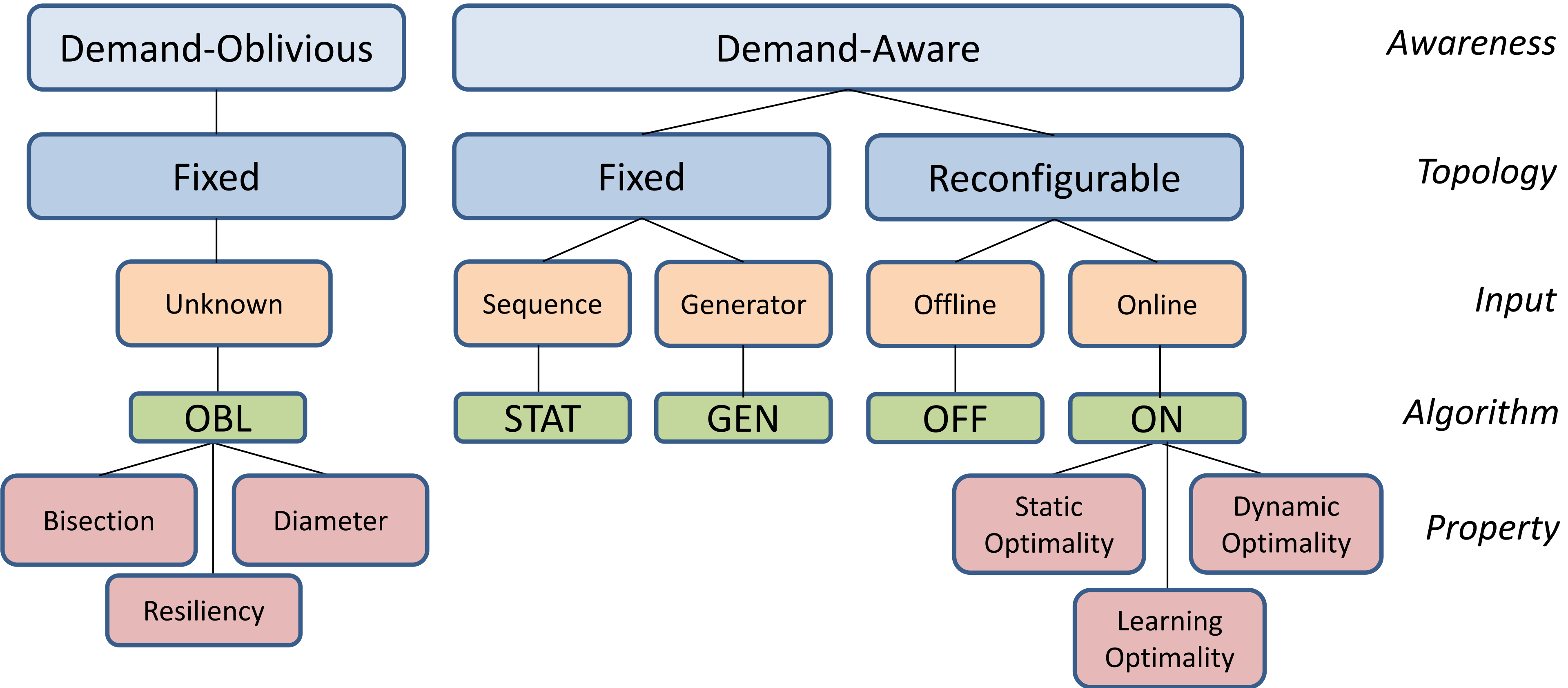
Demand-Aware Network Design with Minimal Congestion and Route Lengths. Chen Avin, Kaushik Mondal, and Stefan Schmid. IEEE **INFOCOM** 2019.

Roadmap

- Entropy: A metric for demand-aware networks?
  - Intuition
  - A lower bound
  - Algorithms achieving entropy bounds
- From static to dynamic demand-aware networks
  - Empirical motivation
  - A connection to self-adjusting datastructures

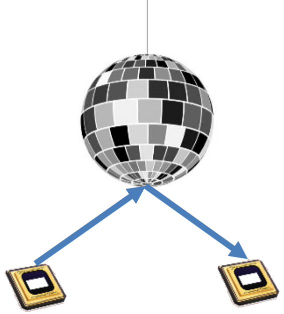
Uncharted Landscape!

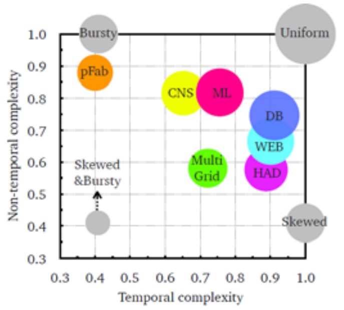
Toward Demand-Aware Networking: A Theory for Self-Adjusting Networks. **SIGCOMM CCR,** 2018.



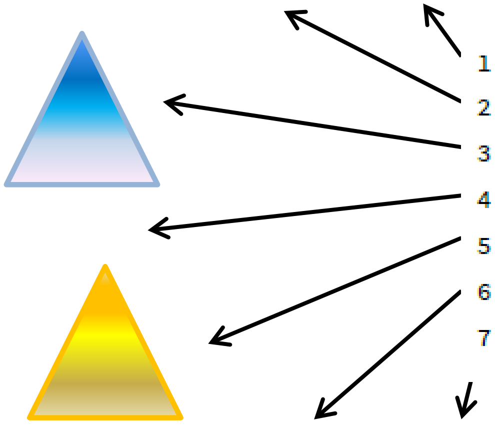
Open Questions

- Optimal static and dynamic topologies? Approximation algorithms?
- Scalable control plane?
- Cross-layer aspects
- Metrics: just the beginning!
- We need more data: how structured and predictable are workloads?
- How to convince operators?





Thank you! Questions?



	1	2	3	4	5	6	7
1	0	2/65	1/13	1/65	1/65	2/65	3/65
2	2/65	0	1/65	0	0	0	2/65
3	1/13	1/65	0	2/65	0	0	1/13
4	1/65	0	2/65	0	4/65	0	0
5	1/65	0	3/65	4/65	0	0	0
6	2/65	0	0	0	0	0	3/65
7	3/65	2/65	1/13	0	0	3/65	0

Further Reading

A survey!

Survey of Reconfigurable Data Center Networks: Enablers, Algorithms, Complexity
Klaus-Tycho Foerster and Stefan Schmid.
SIGACT News, June 2019.

Toward Demand-Aware Networking: A Theory for Self-Adjusting Networks (Editorial)
Chen Avin and Stefan Schmid.
ACM SIGCOMM Computer Communication Review (**CCR**), October 2018.

Measuring the Complexity of Network Traffic Traces
Chen Griner, Chen Avin, Manya Ghobadi, and Stefan Schmid.
arXiv, 2019.

Demand-Aware Network Design with Minimal Congestion and Route Lengths
Chen Avin, Kaushik Mondal, and Stefan Schmid.
38th IEEE Conference on Computer Communications (**INFOCOM**), Paris, France, April 2019.

Distributed Self-Adjusting Tree Networks
Bruna Peres, Otavio Augusto de Oliveira Souza, Olga Goussevskaia, Chen Avin, and Stefan Schmid.
38th IEEE Conference on Computer Communications (**INFOCOM**), Paris, France, April 2019.

Efficient Non-Segregated Routing for Reconfigurable Demand-Aware Networks
Thomas Fenz, Klaus-Tycho Foerster, Stefan Schmid, and Anaïs Villedieu.
IFIP Networking, Warsaw, Poland, May 2019.

DaRTree: Deadline-Aware Multicast Transfers in Reconfigurable Wide-Area Networks
Long Luo, Klaus-Tycho Foerster, Stefan Schmid, and Hongfang Yu.
IEEE/ACM International Symposium on Quality of Service (**IWQoS**), Phoenix, Arizona, USA, June 2019.

Demand-Aware Network Designs of Bounded Degree
Chen Avin, Kaushik Mondal, and Stefan Schmid.
31st International Symposium on Distributed Computing (**DISC**), Vienna, Austria, October 2017.

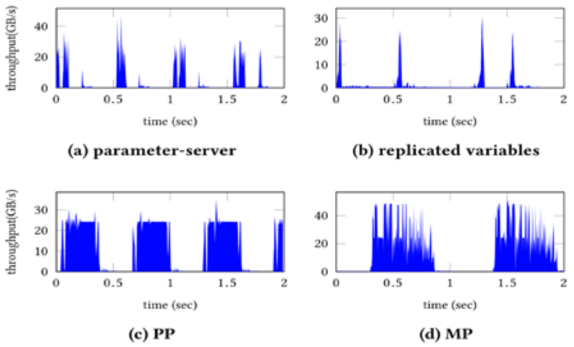
SplayNet: Towards Locally Self-Adjusting Networks
Stefan Schmid, Chen Avin, Christian Scheideler, Michael Borokhovich, Bernhard Haeupler, and Zvi Lotker.
IEEE/ACM Transactions on Networking (**TON**), Volume 24, Issue 3, 2016. Early version: IEEE **IPDPS** 2013.

Characterizing the Algorithmic Complexity of Reconfigurable Data Center Architectures
Klaus-Tycho Foerster, Monia Ghobadi, and Stefan Schmid.
ACM/IEEE Symposium on Architectures for Networking and Communications Systems (**ANCS**), Ithaca, New York, USA, July 2018.

How Predictable is Traffic?

Even if reconfiguration fast, control plane (e.g., data collection) can become a bottleneck. However, many good examples:

- Machine learning applications
- Trend to disaggregation (specialized racks)
- Datacenter communication dominated by elephant flows
- Etc.



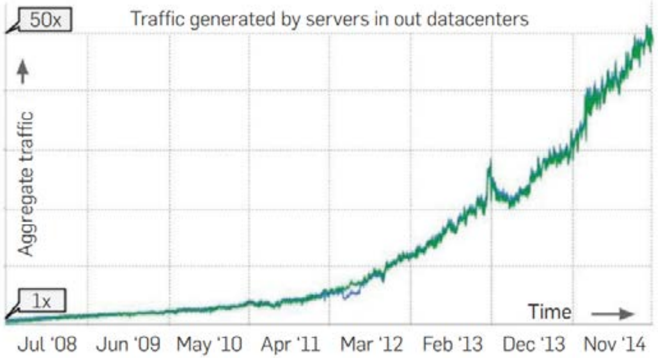
ML workload (GPU to GPU):

deep convolutional neural network

Predictable from their dataflow graph

Explosive Growth of Demand…

Batch processing, web services, **distributed ML**, ….: ***data-centric applications*** are distributed and interconnecting network is ***critical***

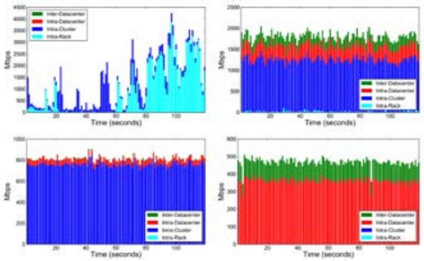


Source: Jupiter Rising. SIGCOMM 2015.

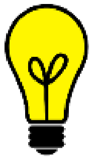
Aggregate server traffic in **Google's datacenter fleet**

… But Much Structure!

Facebook

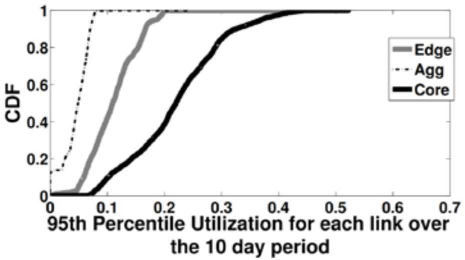


Inside the Social Network's (Datacenter) Network @ SIGCOMM 2015



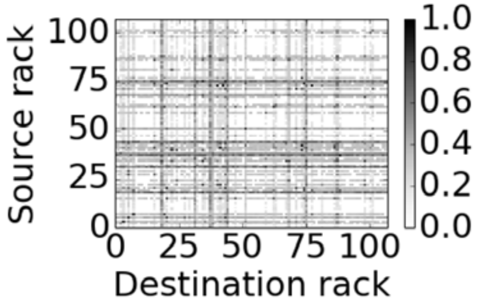
Spatial (***sparse!***) and temporal **locality**

Benson et al.



Understanding Data Center Traffic Characteristics @ WREN 2009

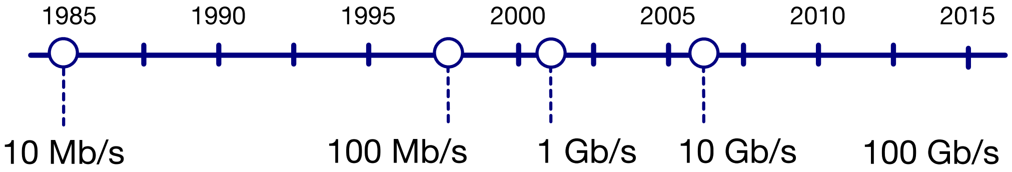
Microsoft



ProjecToR @ SIGCOMM 2016

Historical Motivation: Growth of Hyperscale Datacenters

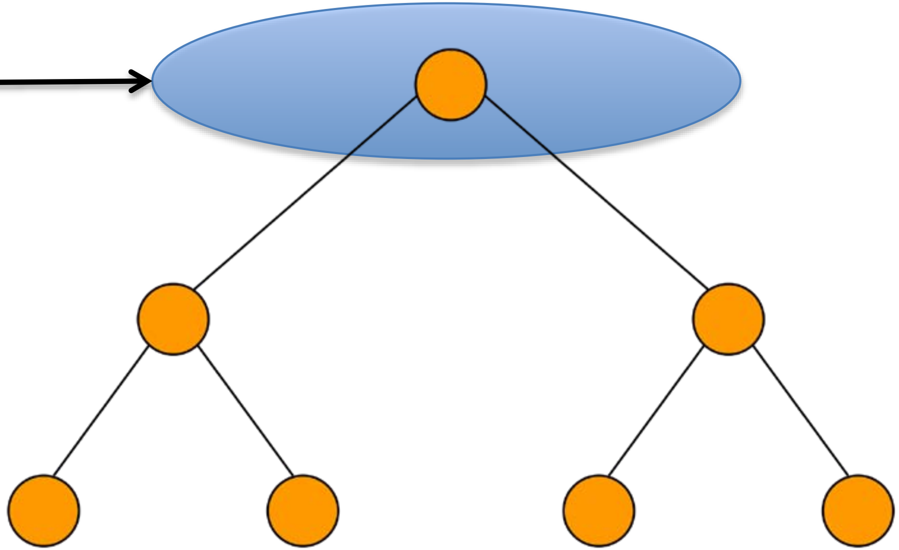
Network problem:
connecting >100,000 servers



Kudos to George Porter
for some slides.

How to move from 1 Gb/s to 10 Gb/s?

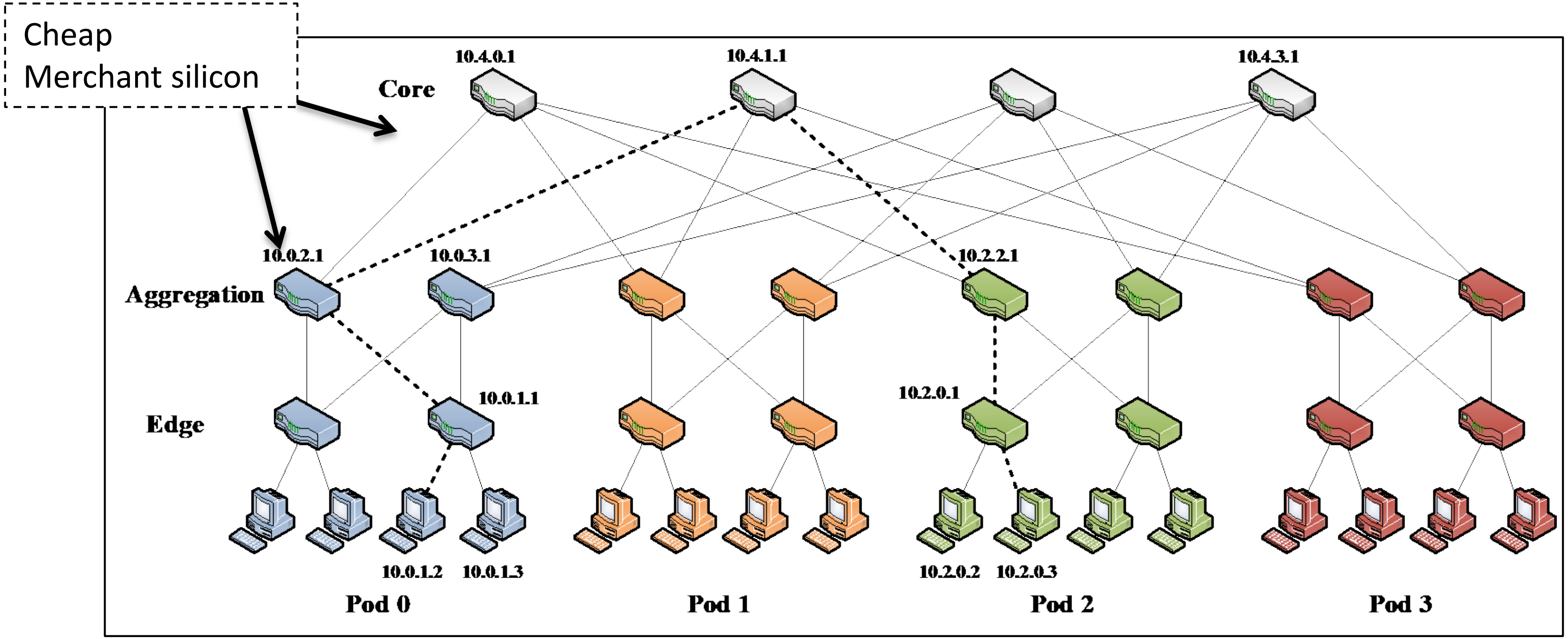
Can't buy sufficiently fast core switches!



100,000 x 10 Gb/s = 1 Pb/s

Kudos to George Porter for some slides.

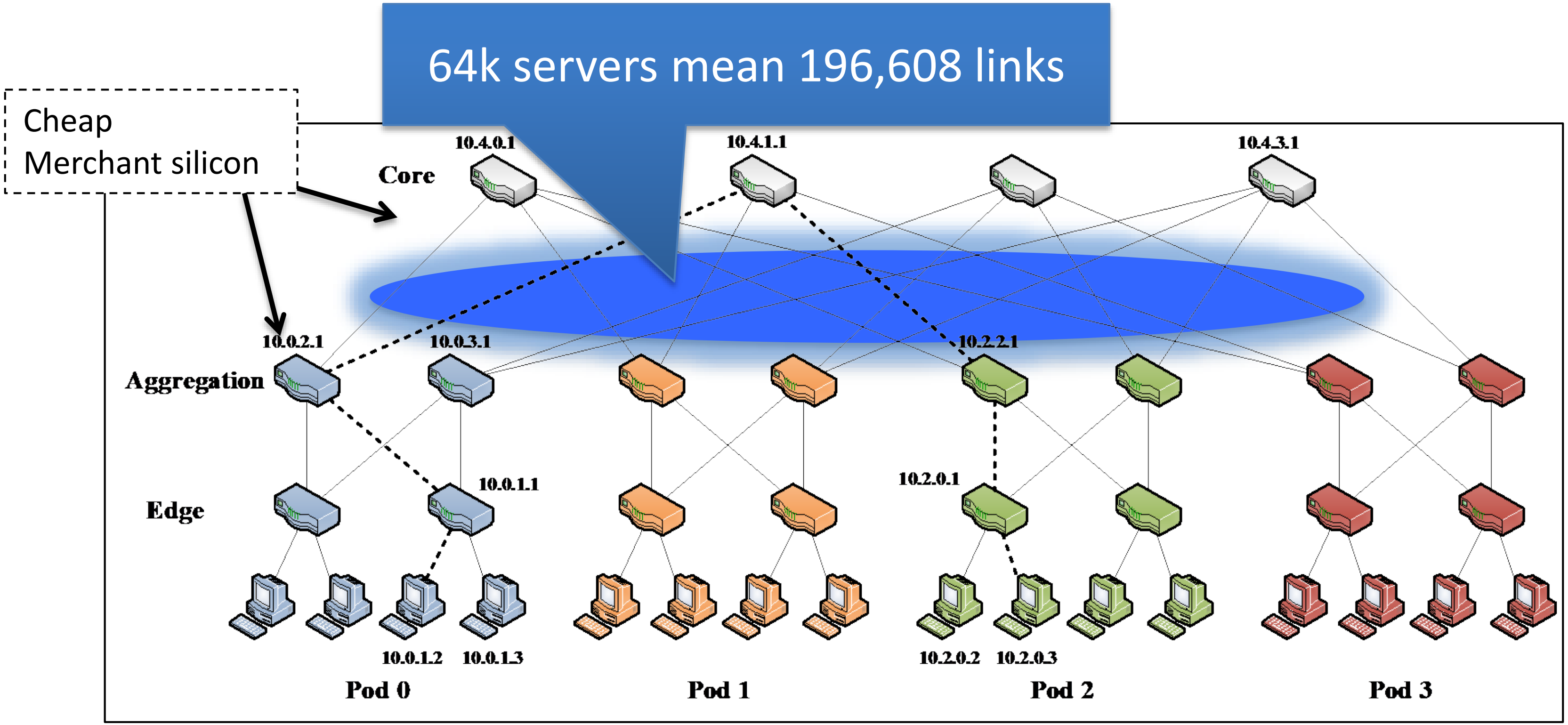
Scale-out Datacenters (2009)



Scale out: more switches and cables!

Kudos to George Porter
for some slides.

Scale-out Datacenters (2009)



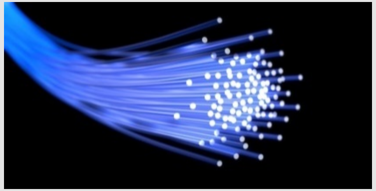
Scale out: more switches and cables!

Kudos to George Porter
for some slides.

Proliferation of Optical Tranceivers a Growing Problem

Optical links

1,000 + Gb/s –
1,000 + meters



Single mode fiber

10 Gb/s –
2,000 meters

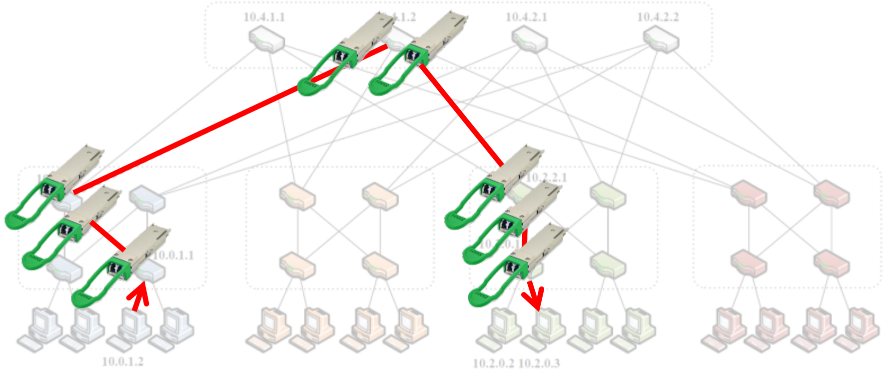
SFP+ transceiver

1 Gb/s – 100m

10 Gb/s – 10 meters

Electrical links

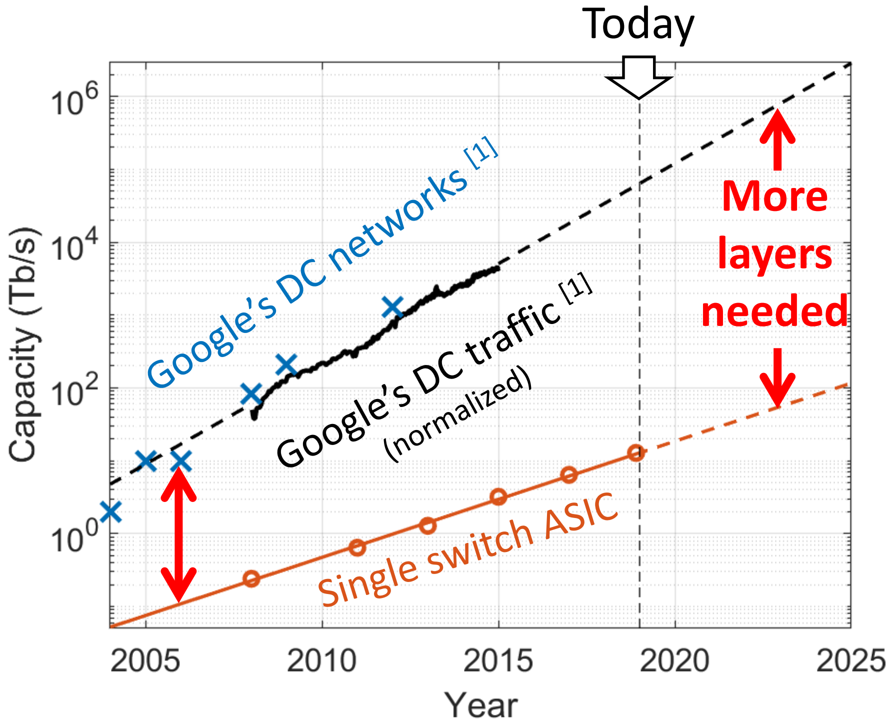
Datacenter Network



For every device attached to the network, there are multiple transceivers in the network: ***high energy and cost***

Kudos to George Porter
for some slides.

Scaling Limitations Today



- Increasing difficulty getting data in/out of the chip
- More fabric layers = higher cost & power

0.64 TB/s

5.12 TB/s

12.8 TB/s

Kudos to George Porter for some slides.