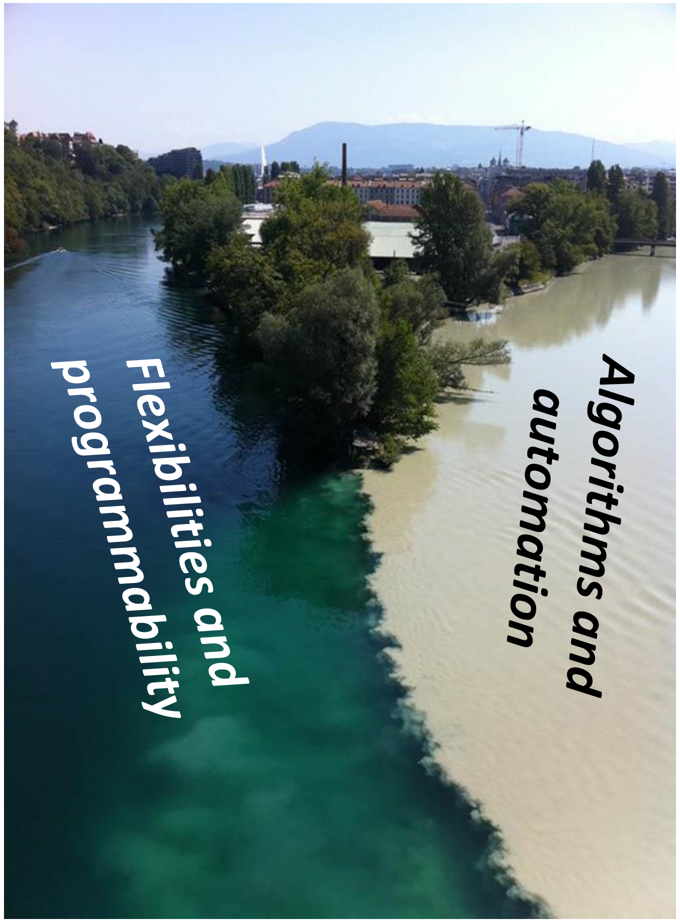
Self-Driving Networks: Use Cases, Approaches, and Research Challenges

Stefan Schmid

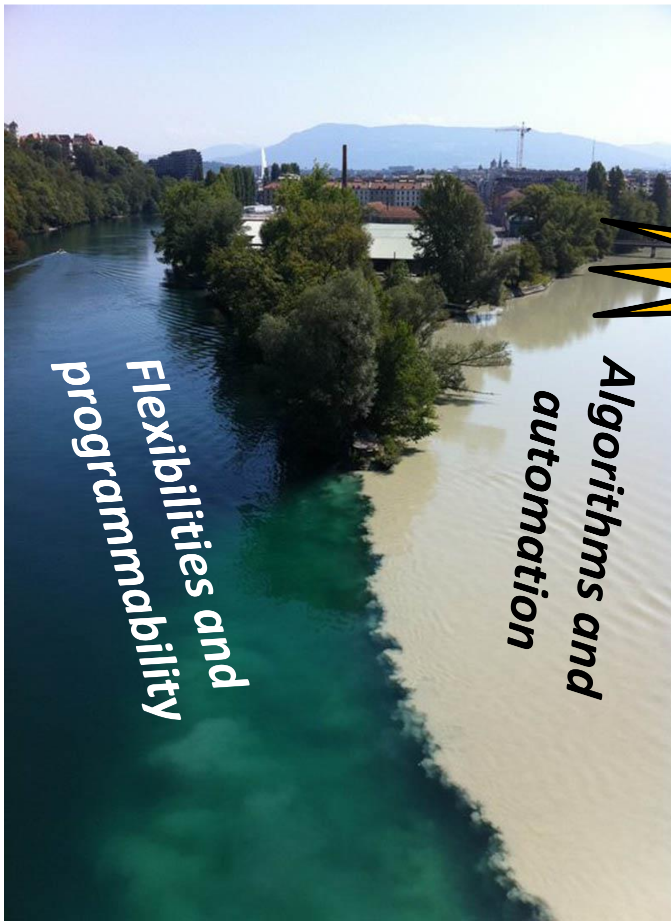
"We cannot direct the wind,
but we can adjust the sails."
(Folklore)

Acknowledgements:

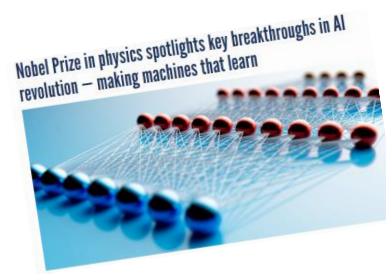
It`s a Great Time to Be a Networking Researcher!

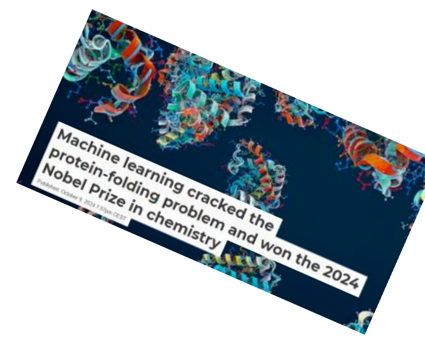


It`s a Great Time to Be a Networking Researcher!



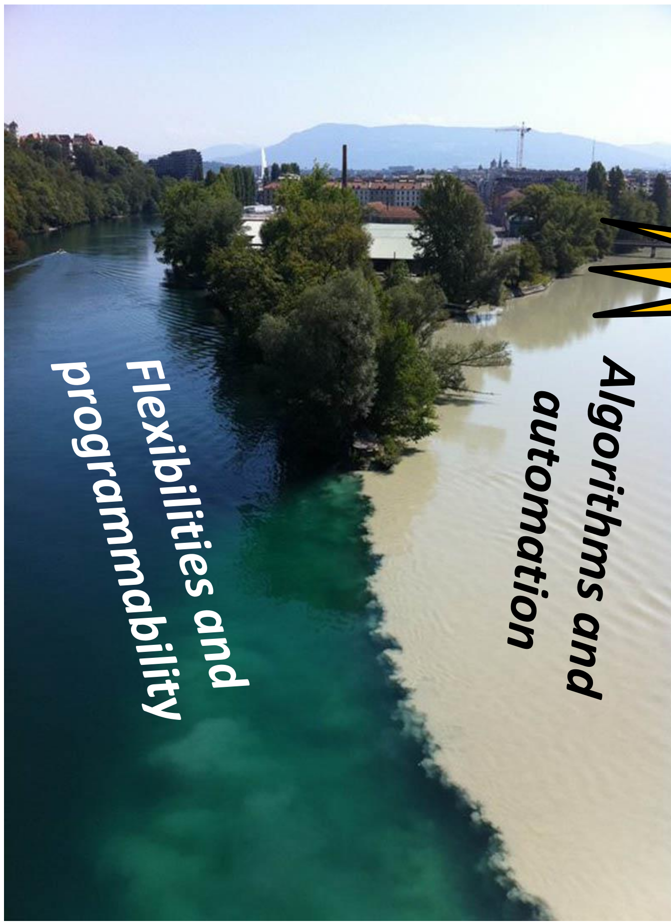
AI/ML everywhere!



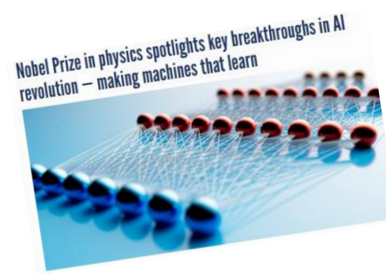


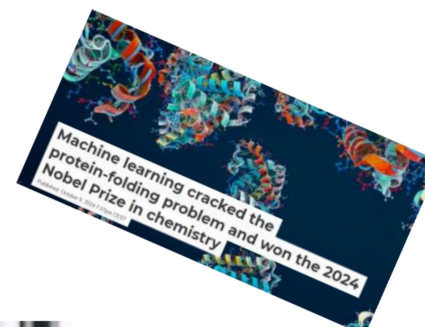
Credits: George Varghese

It`s a Great Time to Be a Networking Researcher!



AI/ML everywhere!





Credits: George Varghese

It`s a Great Time to Be a Networking Researcher!

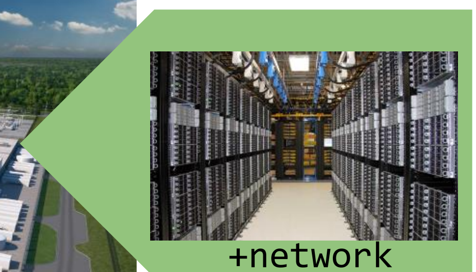
It`s a Great Time to Be a Networking Researcher!

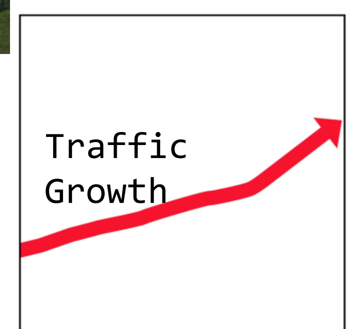
Enables and motivates **self-driving networks**!

Innovations Needed!

Explosive Traffic

Datacenters (“hyper-scale”)



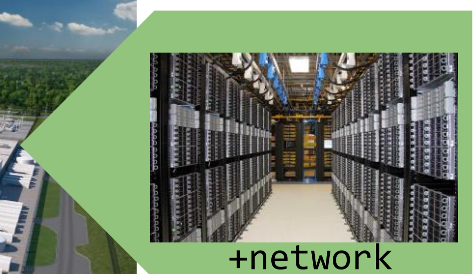


Interconnecting networks:
a **critical infrastructure**
of our digital society.

Innovations Needed!

Explosive Traffic

Datacenters ("hyper-scale")



Interconnecting networks:
a **critical infrastructure**
of our digital society.

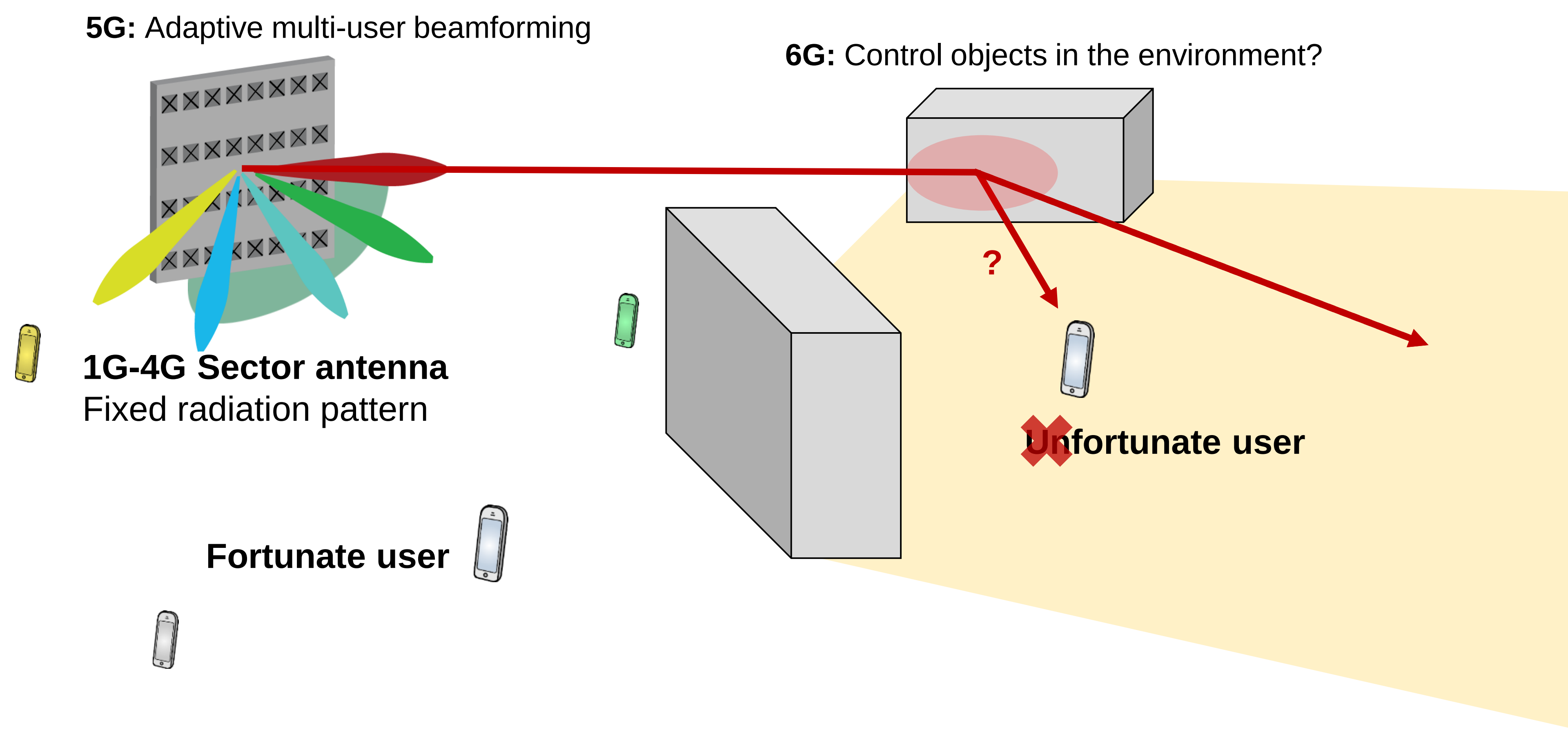
Credits: Marco Chiesa

Fast growing traffic also in…

… wireless and mobile

From generation to generation more…

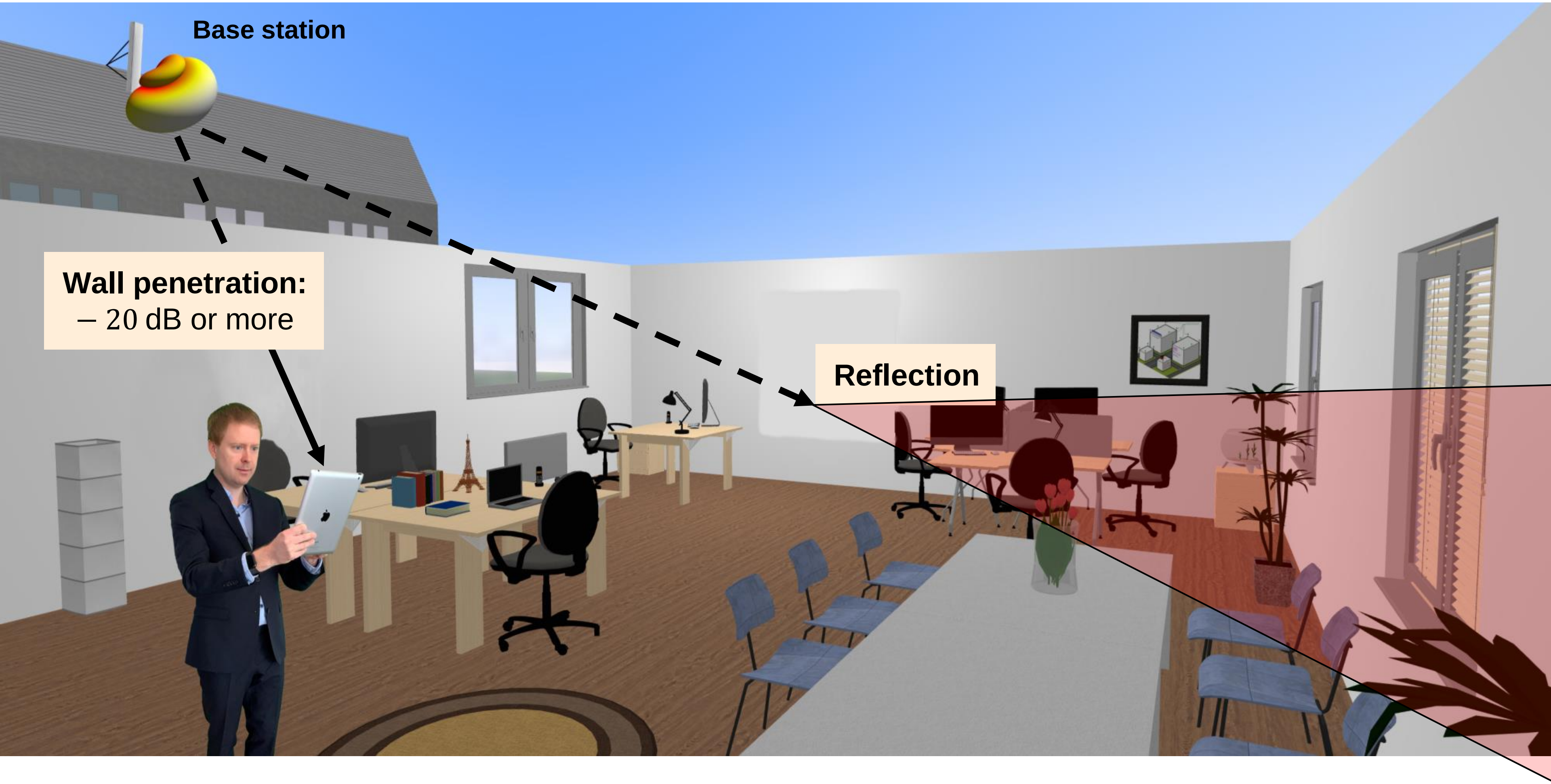
Exciting Flexibilities



credit: Emil Björnson, Christos Liaskos

Traditionally limited by

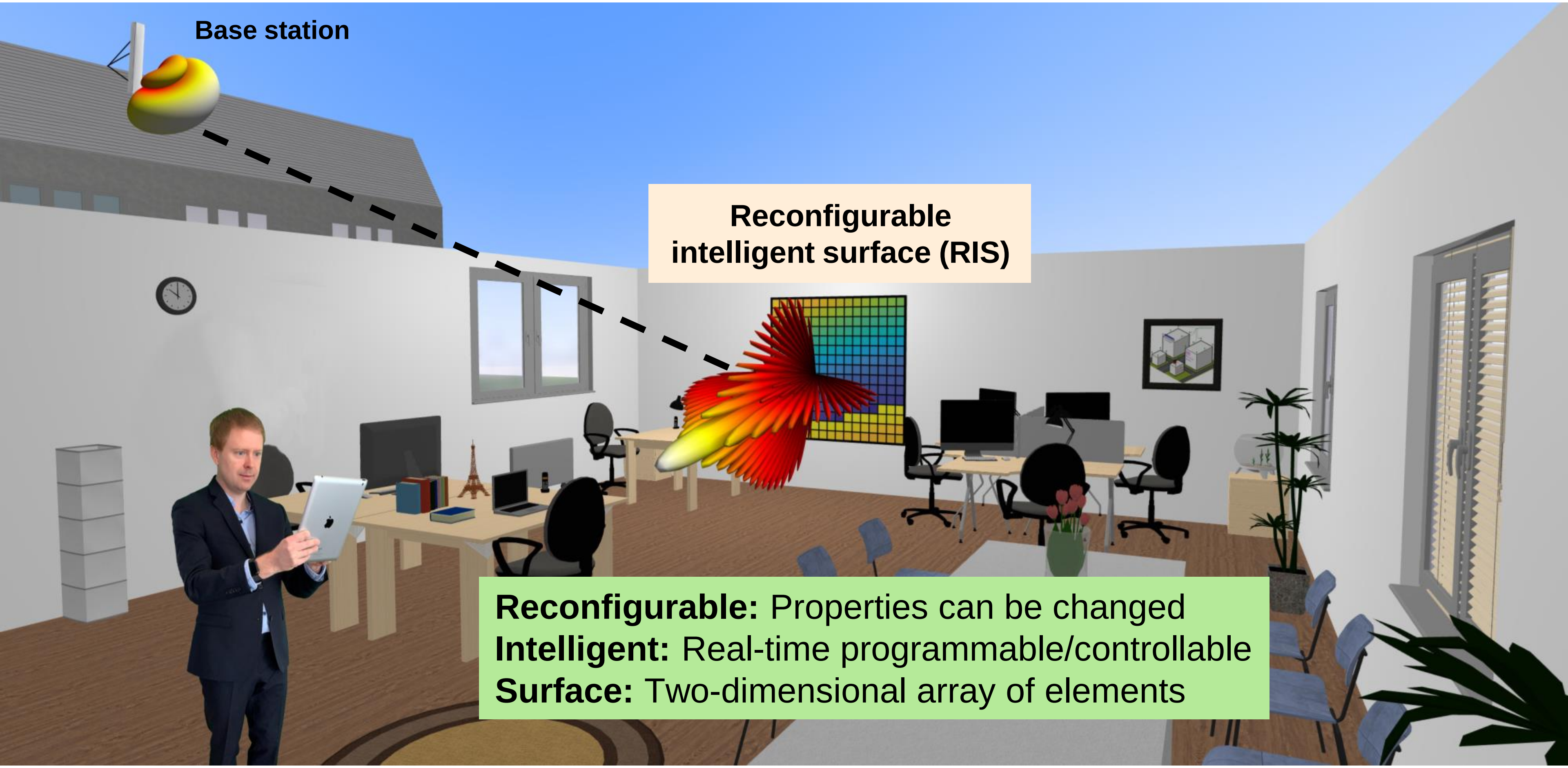
Line of Sight Only



credit: Emil Björnson

Reconfigurable Intelligent Surfaces: Extend to

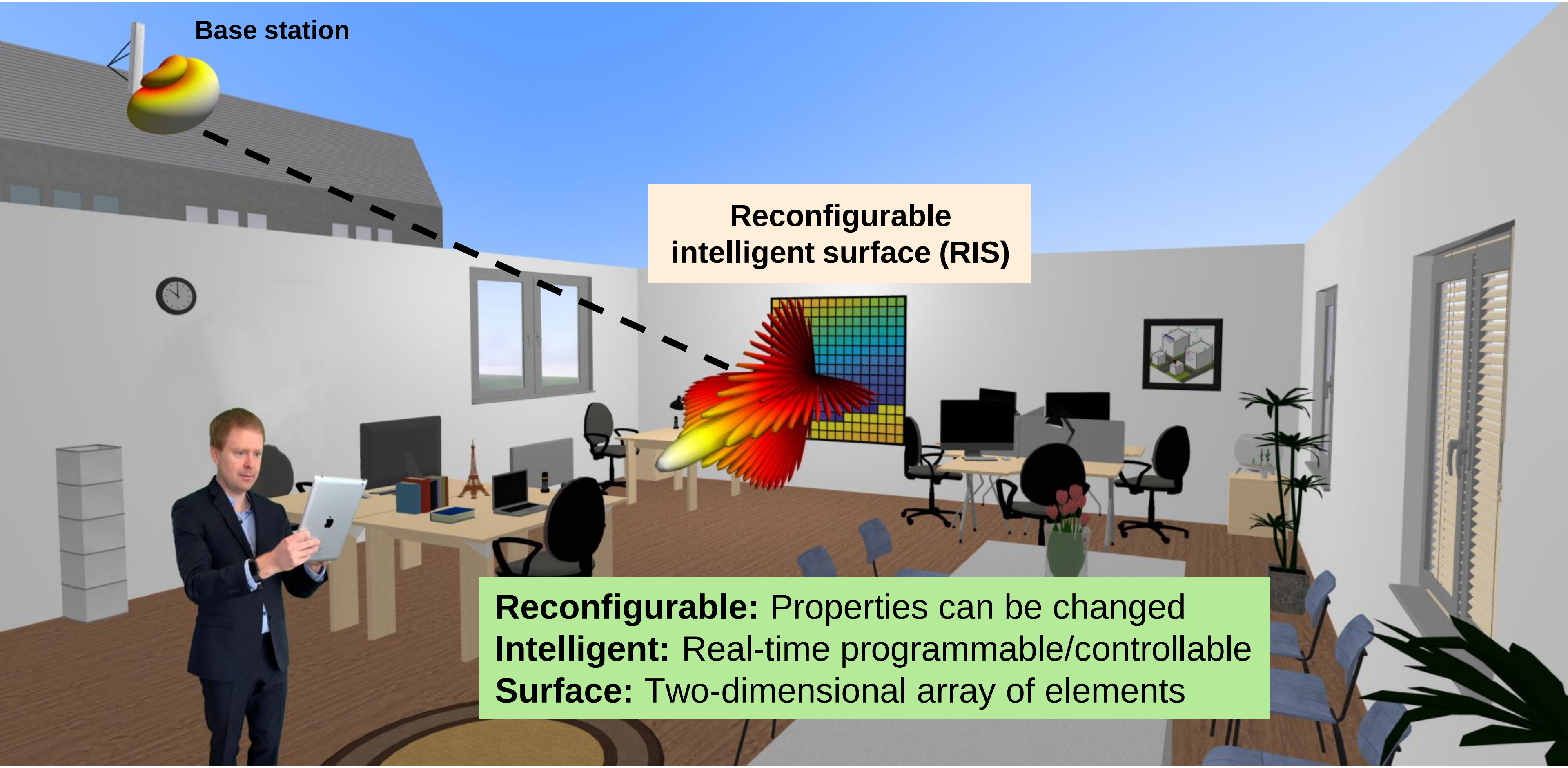
Virtual Line of Sight



credit: Emil Björnson

Reconfigurable Intelligent Surfaces: Extend to

Virtual Line of Sight



Reconfigurable: Properties can be changed
Intelligent: Real-time programmable/controllable
Surface: Two-dimensional array of elements

Literature: Software-Defined Reconfigurable Intelligent Surfaces: From Theory to End-to-End Implementation. Liaskos et al. Proceedings IEEE, 2022.

Great opportunities but come with…

Challenges

- With growing ***demand*** for networks, also increasing ***dependability***
- Important step toward dependable networks: ***modelling***…
- … and ***automated design*** (also using formal methods)!
- Contributions from IEEE CAMAD community critical!

It’s high time for computer-aided designs!

Reality vs Requirements

Today, dependability requirements stand in contrast with reality:

Countries disconnected

Data Centre ▸ Networks

Google routing blunder sent Japan's Internet dark on Friday

Another big BGP blunder

By Richard Chirgwin 27 Aug 2017 at 22:35 40 SHARE ▼

Last Friday, someone in Google fat-thumbed a border gateway protocol (BGP) advertisement and sent Japanese Internet traffic into a black hole.

The trouble began when The Chocolate Factory “leaked” a big route table to Verizon, the result of which was traffic from Japanese giants like NTT and KDDI was sent to Google on the expectation it would be treated as transit.

Passengers stranded

British Airways' latest Total Inability To Support Upwardness of Planes* caused by Amadeus system outage

Stuck on the ground awaiting a load sheet? Here's why

By Gareth Corfield 19 Jul 2018 at 11:16 109 SHARE ▼

BA flights around the world were grounded as a result of the Amadeus outage

Even 911 affected

Officials: Human error to blame in Minn. 911 outage

According to a press release, CenturyLink told department of public safety that human error by an employee of a third party vendor was to blame for the outage

Aug 16, 2018

Duluth News Tribune

SAINT PAUL, Minn. — The Minnesota Department of Public Safety Emergency Communication Networks division was told by its 911 provider that an Aug. 1 outage was caused by human error.

Even tech-savvy companies struggle:

Go Daddy.com github SOCIAL CODING amazon web services

It’s high time for computer-aided designs!

Reality vs Requirements

Today, dependability requirements stand in contrast with reality:

Countries disconnected

Data Centre ▸ Networks

Google routing blunder sent Japan's Internet dark on Friday

Another big BGP blunder

By Richard Chirgwin 27 Aug 2017 at 22:35 40 SHARE ▼

Last Friday, someone in Google fat-thumbed a border gateway protocol (BGP) advertisement and sent Japanese Internet traffic into a black hole.

The trouble began when The Chocolate Factory “leaked” a big route table to Verizon, the result of which was traffic from Japanese giants like NTT and KDDI was sent to Google on the expectation it would be treated as transit.

Passengers stranded

British Airways' latest Total Inability To Support Upwardness of Planes* caused by Amadeus system outage

Stuck on the ground awaiting a load sheet? Here's why

By Gareth Corfield 19 Jul 2018 at 11:16 109 SHARE ▼

BA flights around the world were grounded as a result of the Amadeus outage

Even 911 affected

Officials: Human error to blame in Minn. 911 outage

According to a press release, CenturyLink told department of public safety that human error by an employee of a third party vendor was to blame for the outage

Aug 16, 2018

Duluth News Tribune

SAINT PAUL, Minn. — The Minnesota Department of Public Safety Emergency Communication Networks division was told by its 911 provider that an Aug. 1 outage was caused by human error.

Mainly: human errors!

Even tech-savvy companies struggle:

Go Daddy.com github SOCIAL CODING amazon web services

It’s high time for computer-aided designs!

Reality vs Requirements

Today, dependability requirements stand in contrast with reality:

Countries disconnected

Data Centre ▸ Networks

Google routing blunder sent Japan's Internet dark on Friday

Another big BGP blunder

By Richard Chirgwin 27 Aug 2017 at 22:35 40 SHARE ▼

Last Friday, someone in Google fat-thumbed a border gateway protocol (BGP) advertisement and sent Japanese Internet traffic into a black hole.

The trouble began when The Chocolate Factory “leaked” a big route table to Verizon, the result of which was traffic from Japanese giants like NTT and KDDI was sent to Google on the expectation it would be treated as transit.

Passengers stranded

British Airways' latest Total Inability To Support Upwardness of Planes* caused by Amadeus system outage

Stuck on the ground awaiting a load sheet? Here's why

By Gareth Corfield 19 Jul 2018 at 11:16 109 SHARE ▼

BA flights around the world were grounded as a result of the Amadeus outage

Even 911 affected

Officials: Human error to blame in Minn. 911 outage

According to a press release, CenturyLink told department of public safety that human error by an employee of a third party vendor was to blame for the outage

Aug 16, 2018

Duluth News Tribune

SAINT PAUL, Minn. — The Minnesota Department of Public Safety Emergency Communication Networks division was told by its 911 provider that an Aug. 1 outage was caused by human error.

Mainly: human errors!

Even tech-savvy companies struggle:

GoDaddy.com github SOCIAL CODING amazon web services

Wireless particularly challenging to model!

Roadmap

- Performance: Self-adjusting datacenter networks
- Modelling: How to model workloads, such as ML workloads?
- Dependability: Self-correcting MPLS networks
- More Use cases for self-driving networks

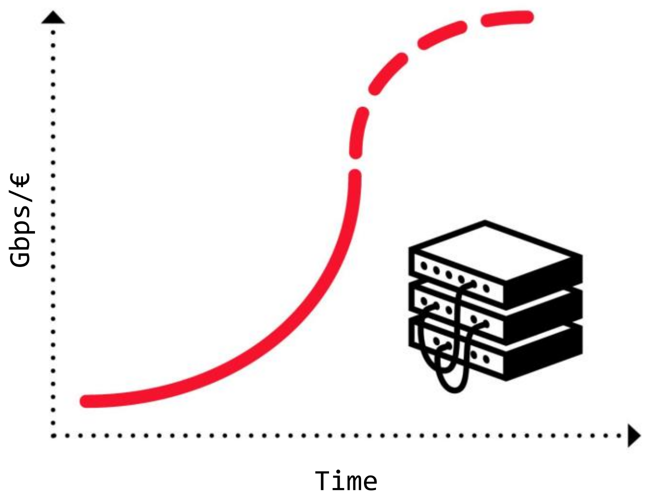
Datacenters Today

Huge Infrastructure, Inefficient Use

⇢ Network equipment reaching capacity limits
  ⇀ Transistor density rates stalling
  ⇀ “End of **Moore‘s Law** in networking”

⇢ Hence: more equipment, larger networks

⇢ Resource intensive and: **inefficient**



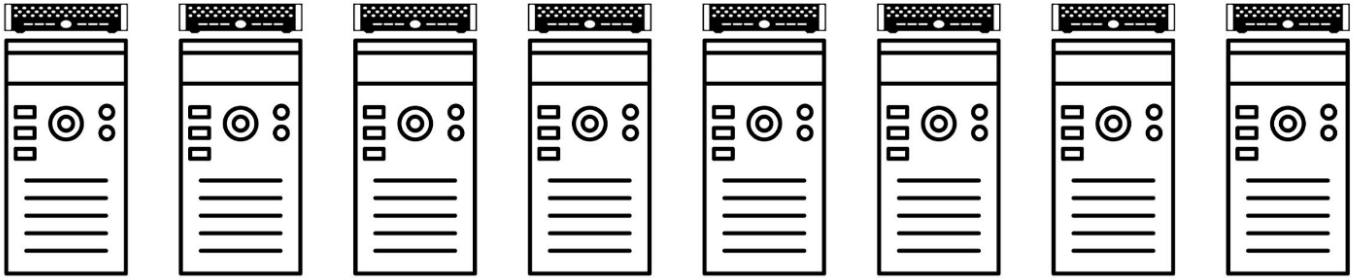
[1] Source: Microsoft, 2019

Annoying for companies, **opportunity** for researchers!

Root Cause

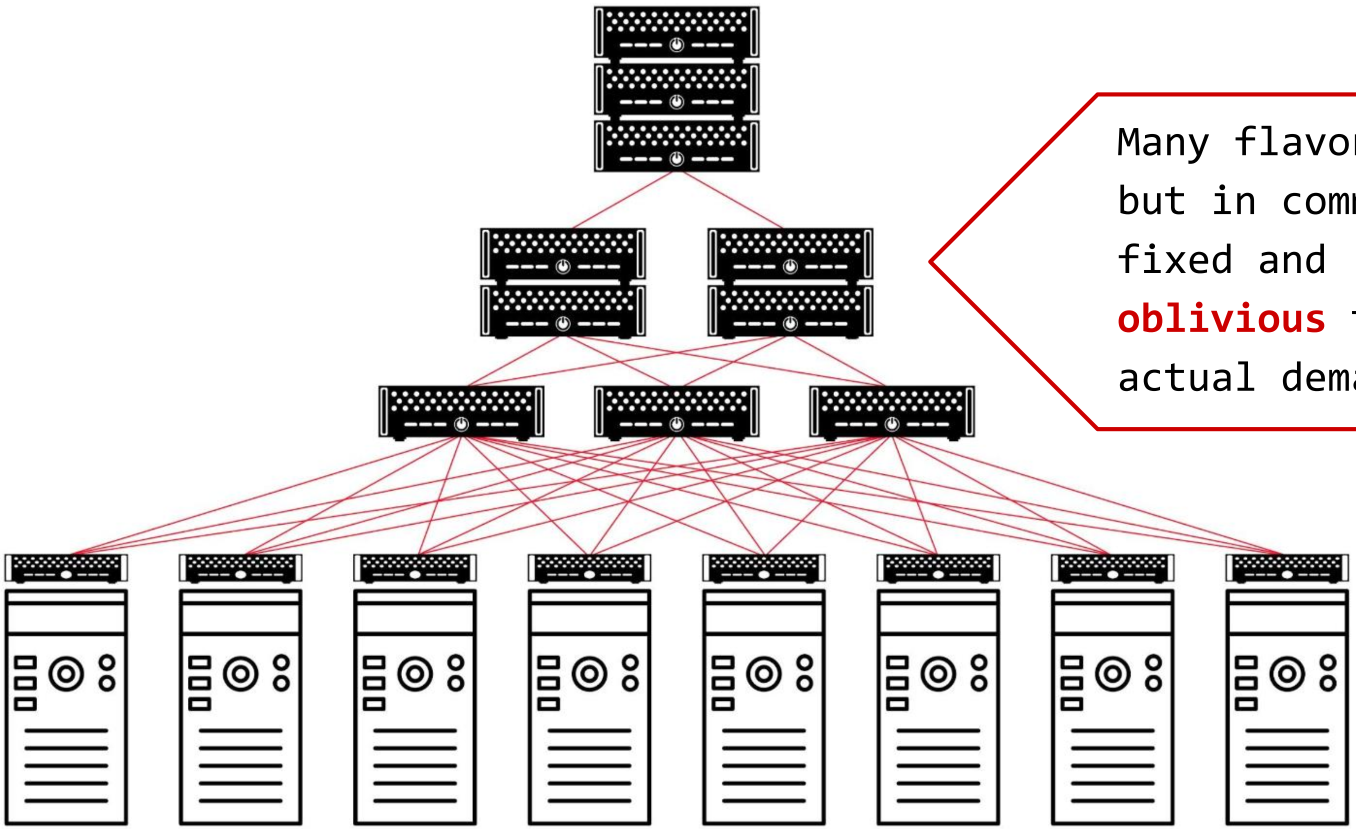
Fixed and Demand-Oblivious Topology

How to interconnect?



Root Cause

Fixed and Demand-Oblivious Topology

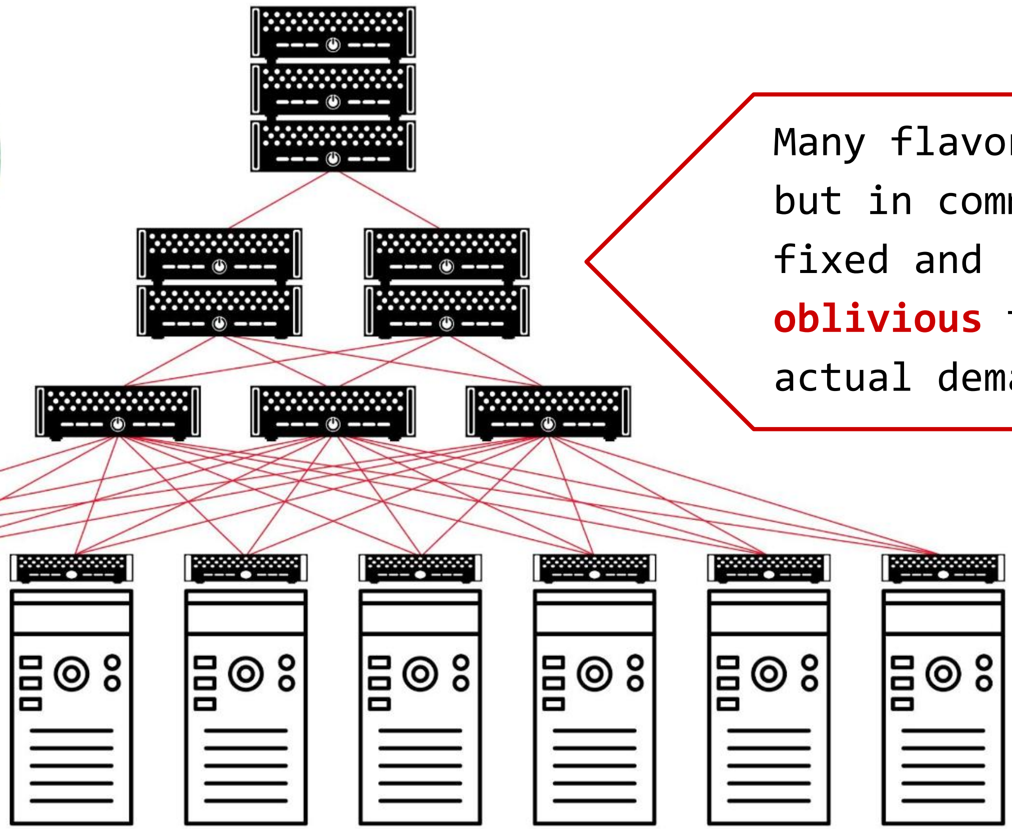


Many flavors, but in common: fixed and **oblivious** to actual demand.

Root Cause

Fixed and Demand-Oblivious Topology

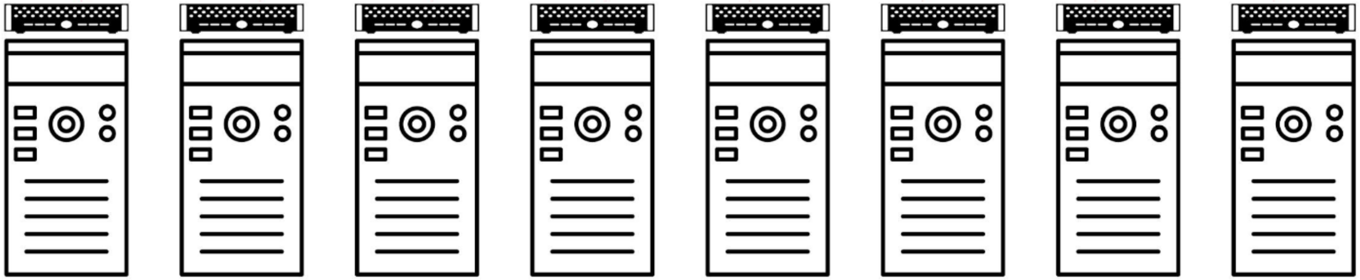
Highway which ignores actual traffic: frustrating!



Many flavors, but in common: fixed and **oblivious** to actual demand.

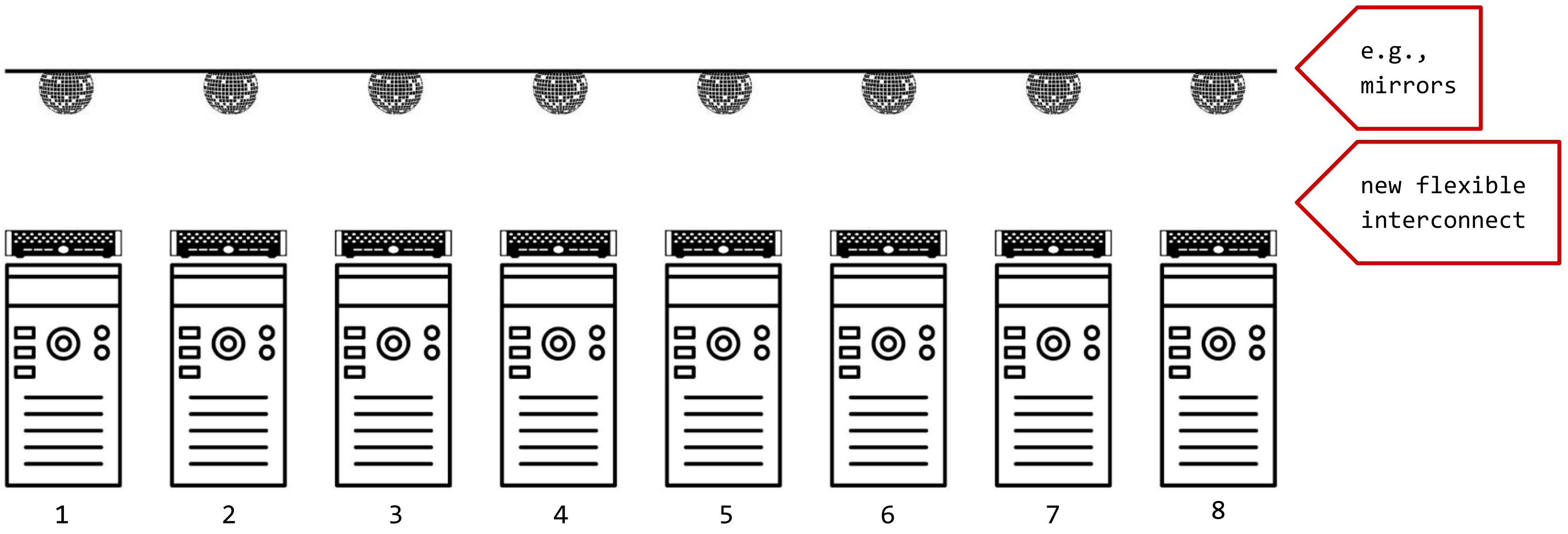
A Vision

Flexible and Demand-Aware Topologies



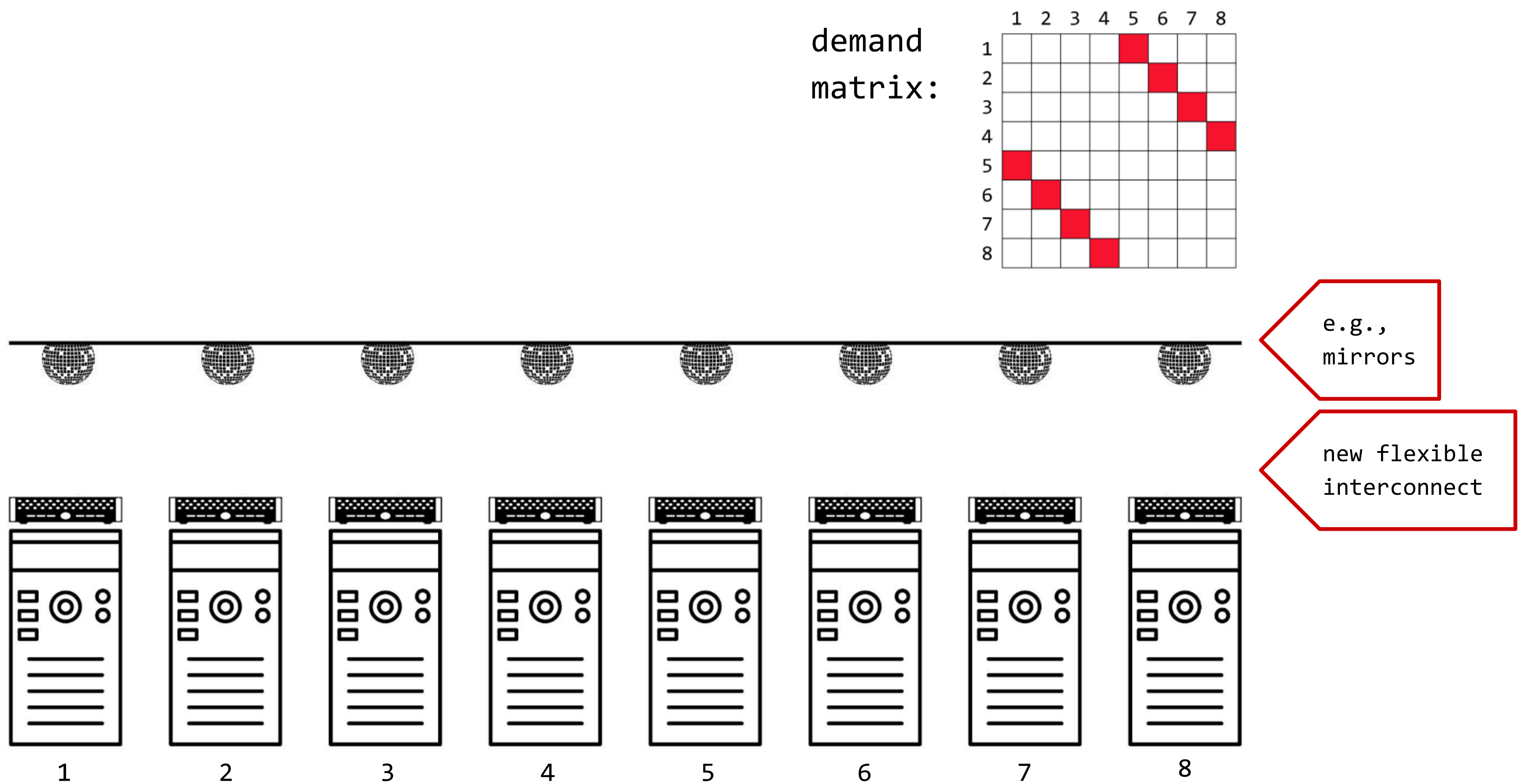
A Vision

Flexible and Demand-Aware Topologies



A Vision

Flexible and Demand-Aware Topologies

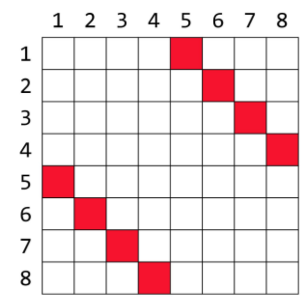


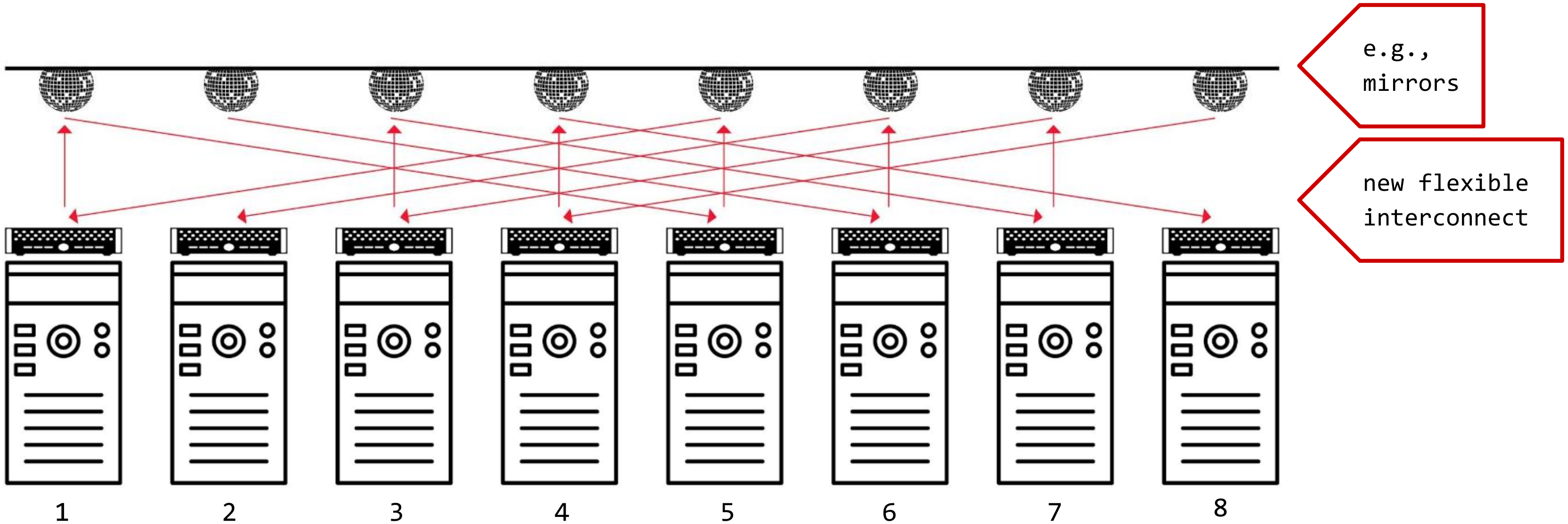
A Vision

Flexible and Demand-Aware Topologies

Matches demand

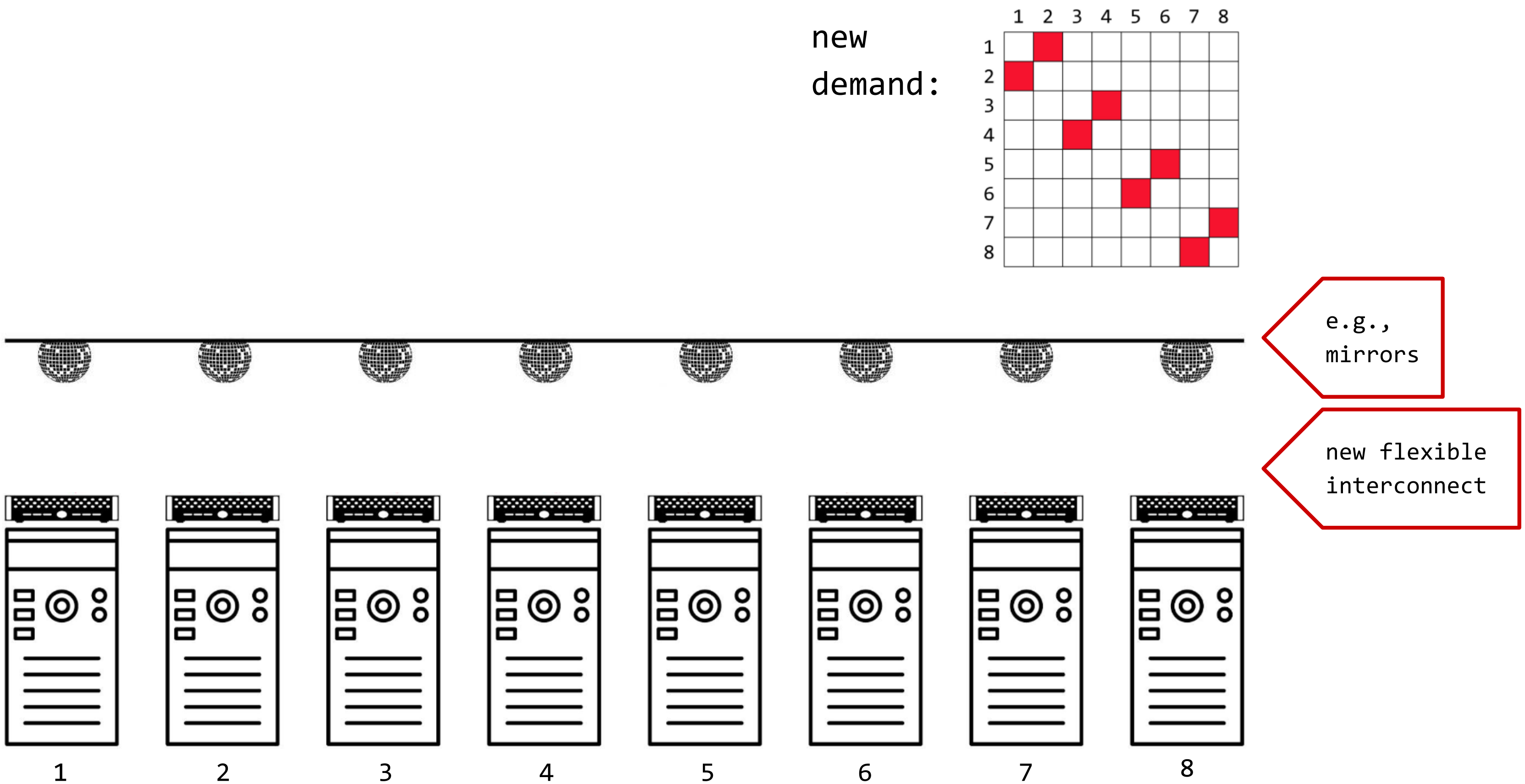
demand matrix:





A Vision

Flexible and Demand-Aware Topologies

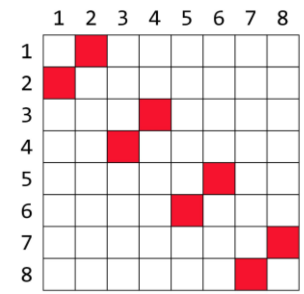


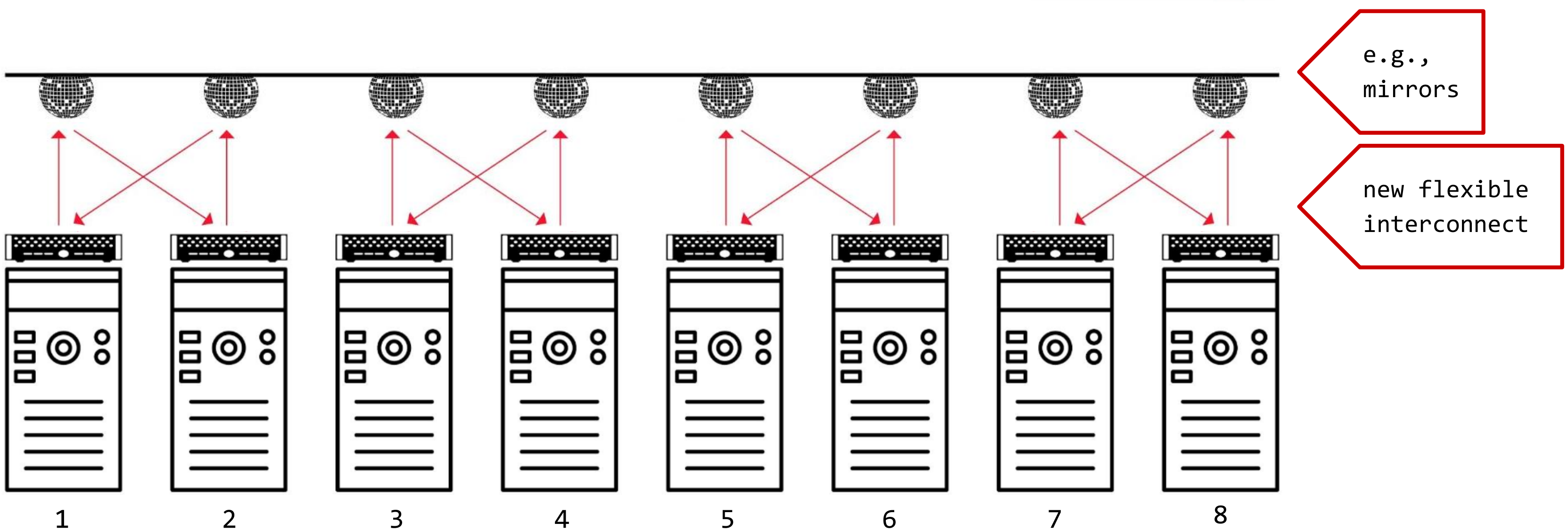
A Vision

Flexible and Demand-Aware Topologies

Matches demand

new demand:



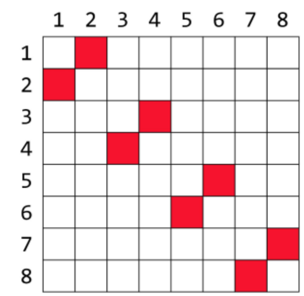


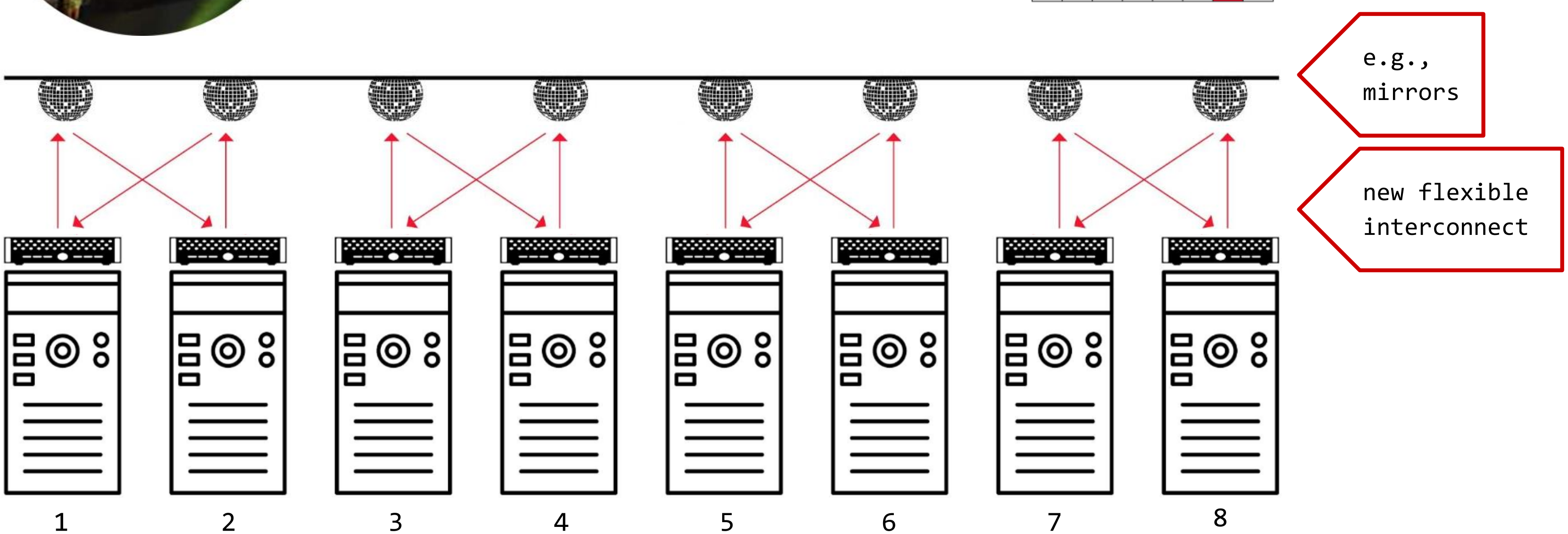
A Vision

Flexible and Demand-Aware Topologies

Self-Adjusting Networks

new demand:





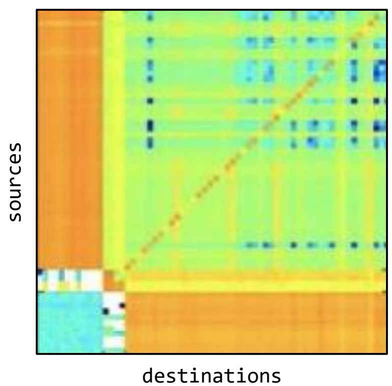
The Motivation

Much Structure in the Demand

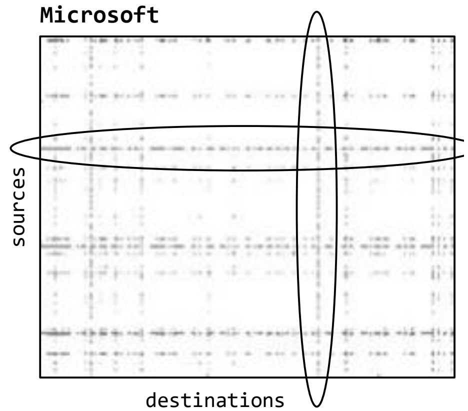
Empirical studies:

traffic matrices sparse and skewed

Facebook

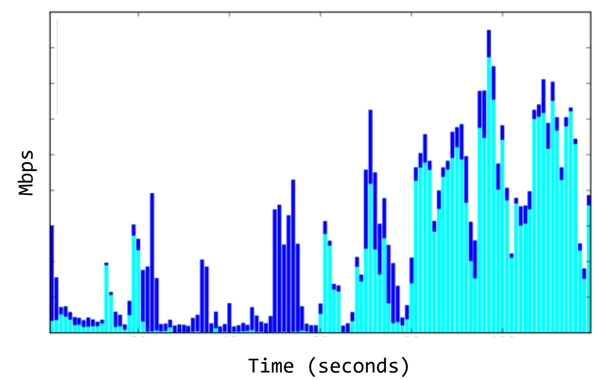


Microsoft



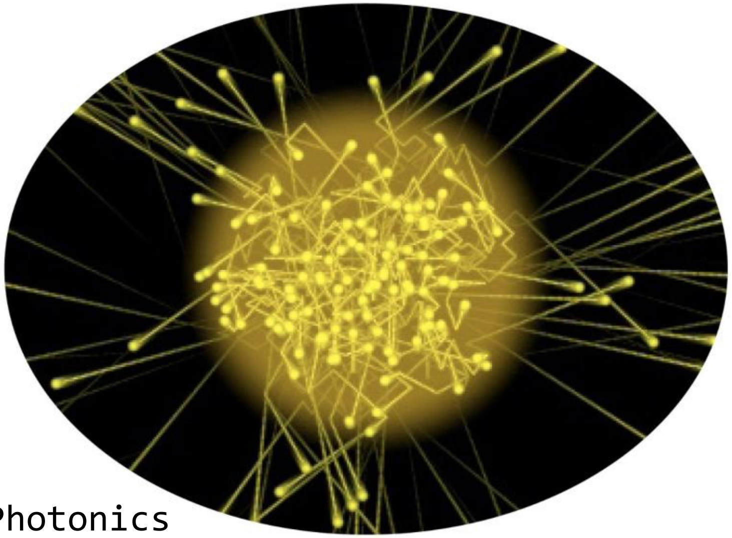
traffic bursty over time

Facebook



The hypothesis: can be exploited.

Sounds Crazy?
Emerging Enabling Technology.



Photonics

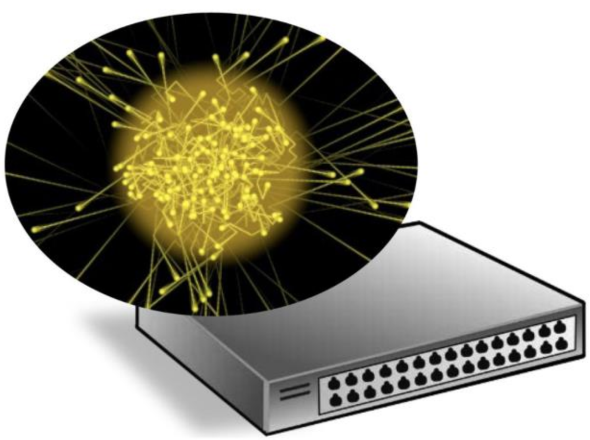
H2020:
“Photonics one of only five key enabling technologies for future prosperity.”

US National Research Council:
“Photons are the new Electrons.”

Enabler

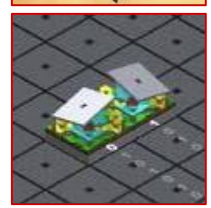
Novel Reconfigurable Optical Switches

⇢ **Spectrum** of prototypes
- ⇀ Different sizes, different reconfiguration times
- ⇀ From our ACM **SIGCOMM** workshop OptSys



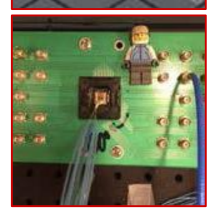
Prototype 1

Moving antenna (ms)



Prototype 2

Moving mirrors (mus)



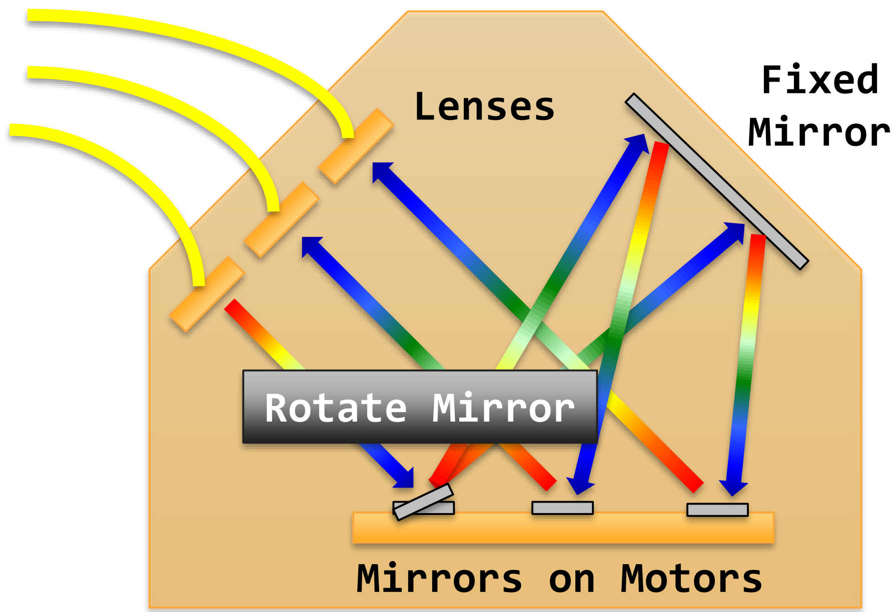
Prototype 3

Changing lambdas (ns)

Example

Optical Circuit Switch

- ⇢ Optical Circuit Switch rapid adaption of physical layer
  - ⇀ Based on rotating mirrors



Optical Circuit Switch

By Nathan Farrington, SIGCOMM 2010

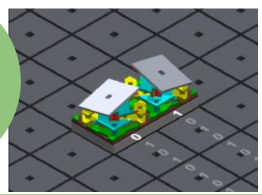
First Deployments

E.g., Google

The Big Picture

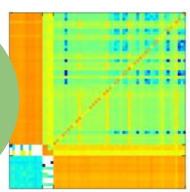
New!

Flexibility



More!

Structure



Self-Adjusting Networks

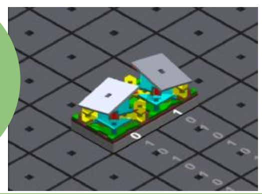
Efficiency

Now is the time!

The Big Picture

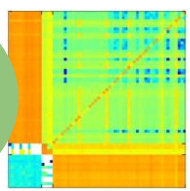
New!

Flexibility



More!

Structure



Self-Adjusting Networks

Now is the time!

Efficiency

Missing: Theoretical **foundations** of demand-aware, self-adjusting networks.

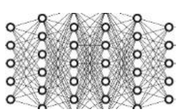
Unique Position

Demand-Aware, Self-Adjusting Systems

Everywhere, but mainly in software

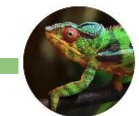
Algorithmic trading

Recommender systems

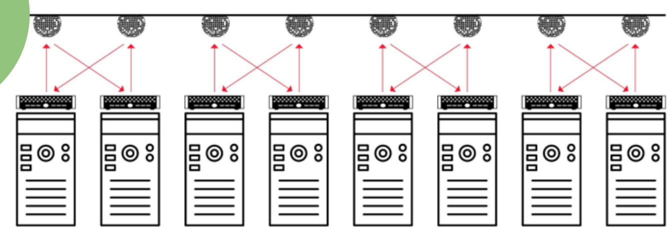


Neural networks

VS



Our focus in this talk: in hardware



First basic question:

How to measure and model structure in workloads?

A first insight: related to entropy.

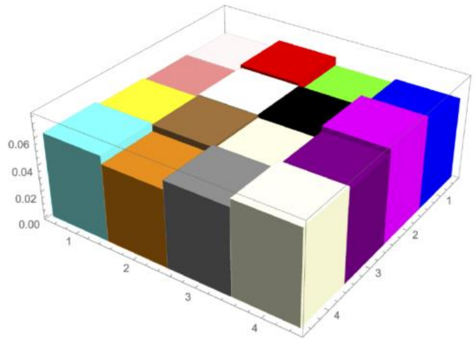
Intuition

Which demand has more structure?

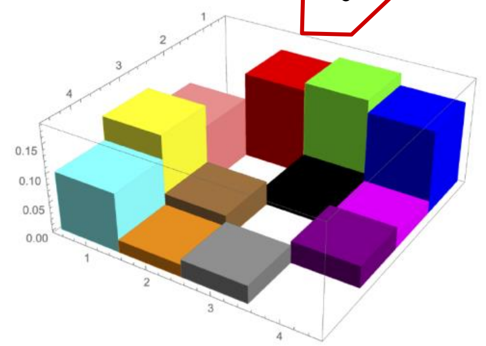
⇢ Traffic matrices of two different distributed ML applications

- ⇀ GPU-to-GPU

Color = communication pair



VS



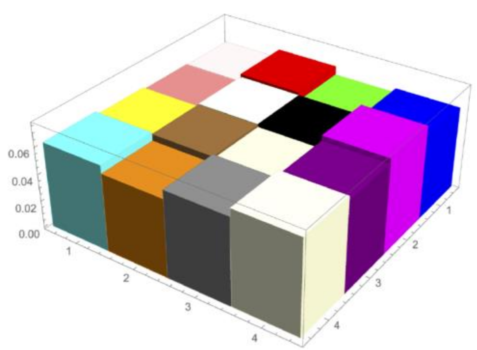
Intuition

Which demand has more structure?

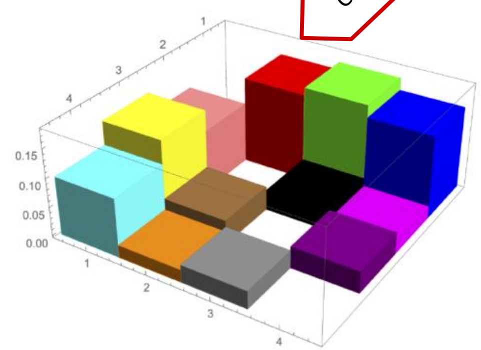
⇢ Traffic matrices of two different distributed ML applications

→ GPU-to-GPU

Color = communication pair



VS



More uniform

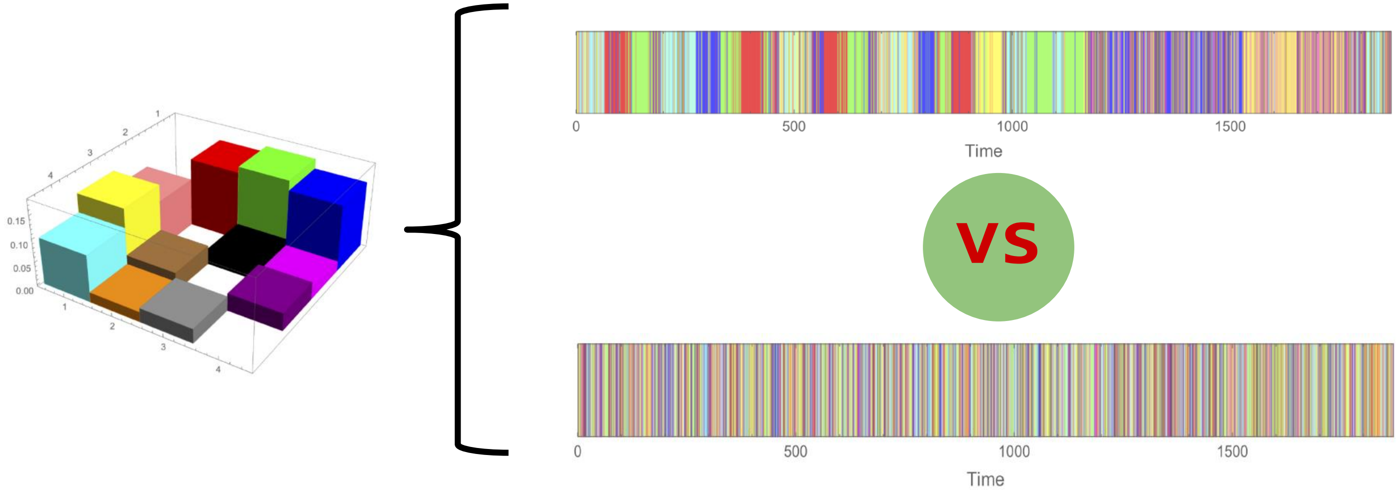
More structure

Intuition

Spatial vs temporal structure

⇢ Two different ways to generate same traffic matrix:
  ⇀ Same non-temporal structure

⇢ Which one has more structure?

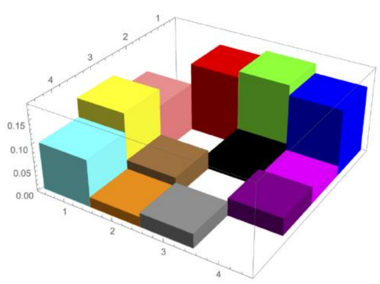


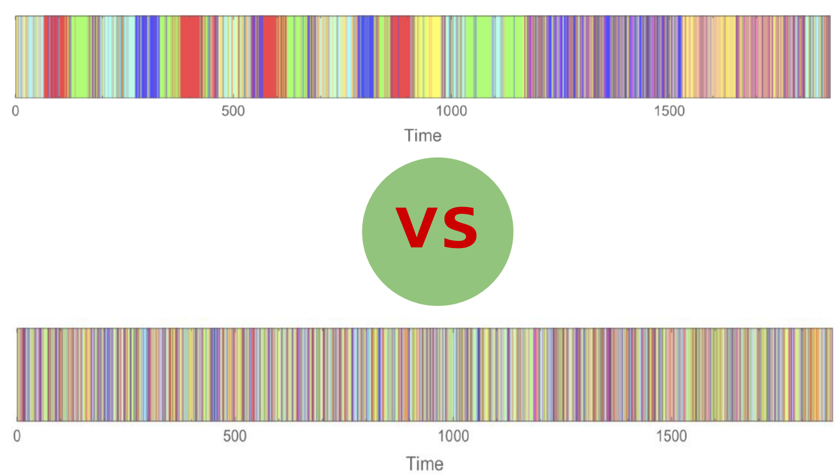
Intuition

Spatial vs temporal structure

- Two different ways to generate same traffic matrix:
  - Same non-temporal structure

- Which one has more structure?



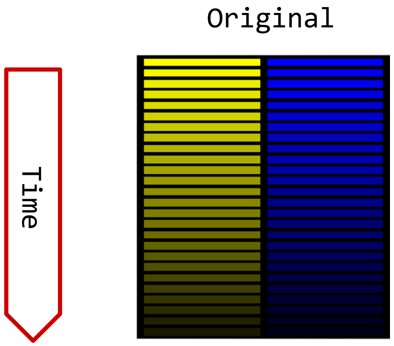


Systematically?

Trace Complexity

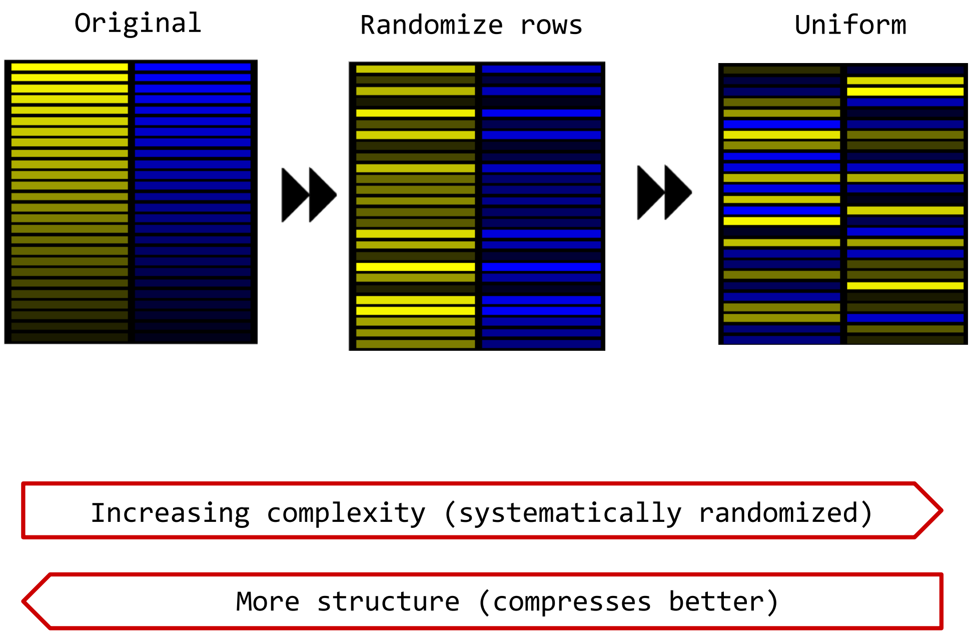
Information-Theoretic Approach

“Shuffle&Compress”



Trace Complexity

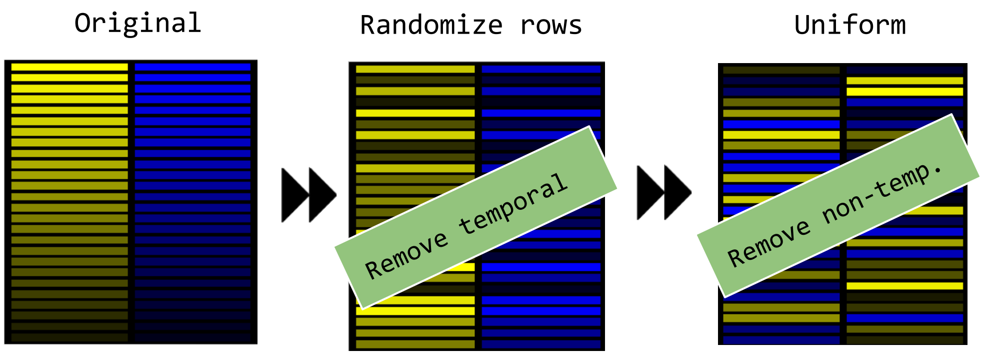
Information-Theoretic Approach
“Shuffle&Compress”



Trace Complexity

Information-Theoretic Approach

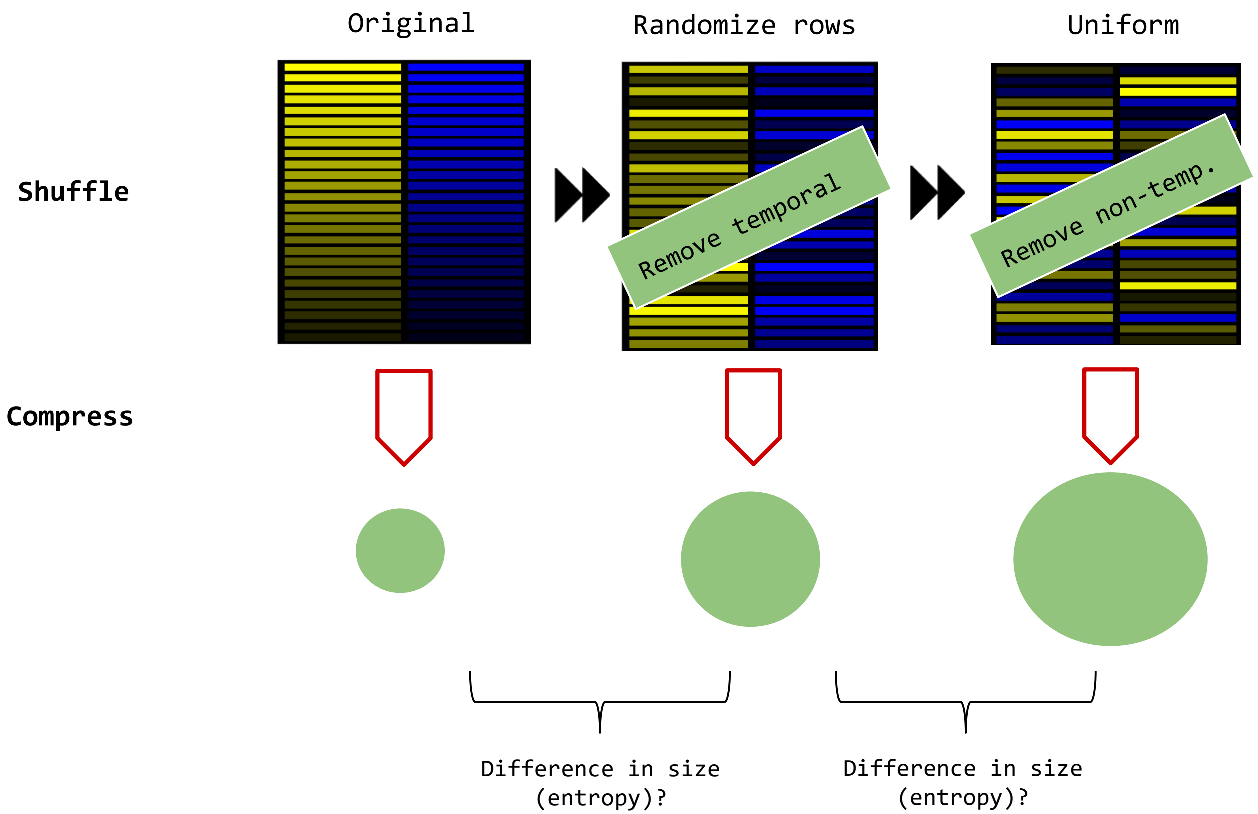
“Shuffle&Compress”



Trace Complexity

Information-Theoretic Approach

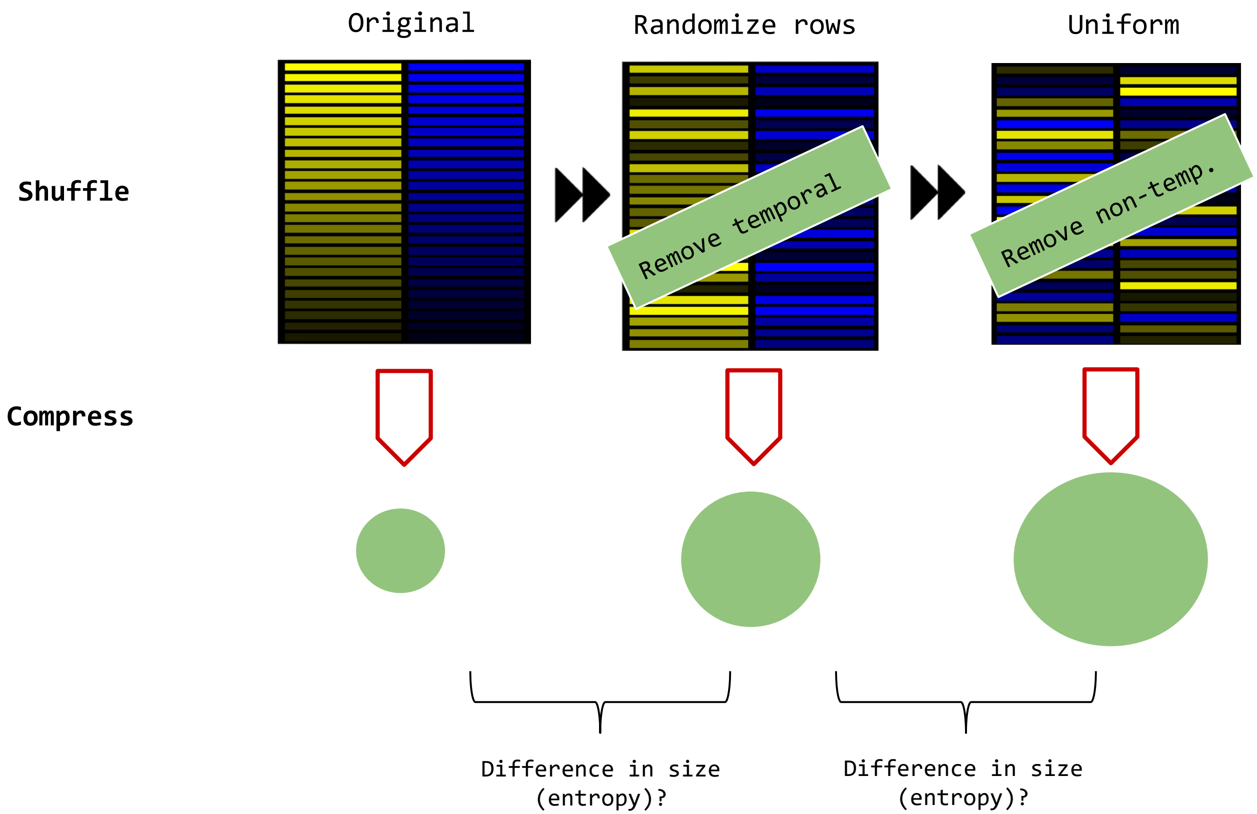
“Shuffle&Compress”



Trace Complexity

Information-Theoretic Approach

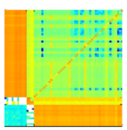
“Shuffle&Compress”

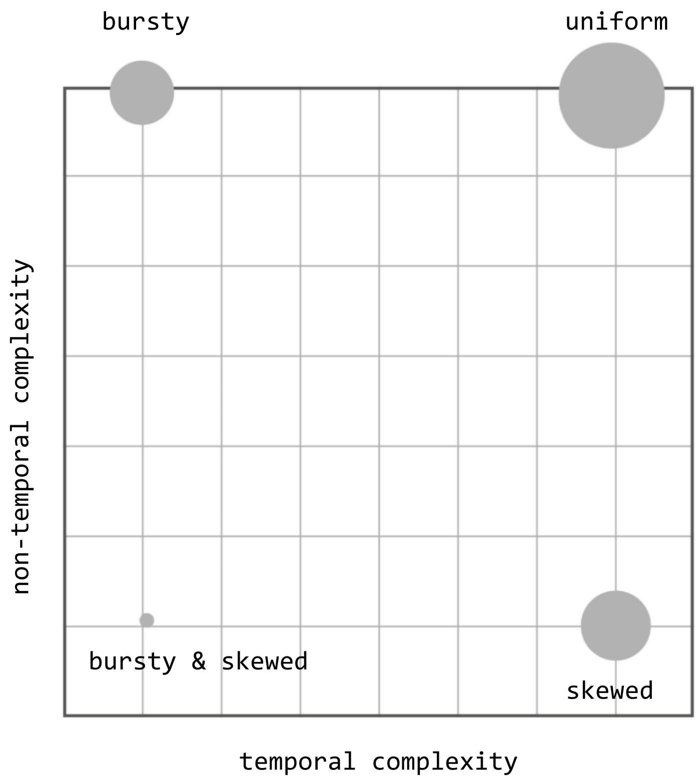


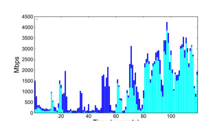
Can be used to define 2-dimensional complexity map!

Avin et al. (Sigmetrics'2020)

Complexity Map





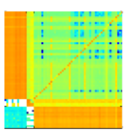


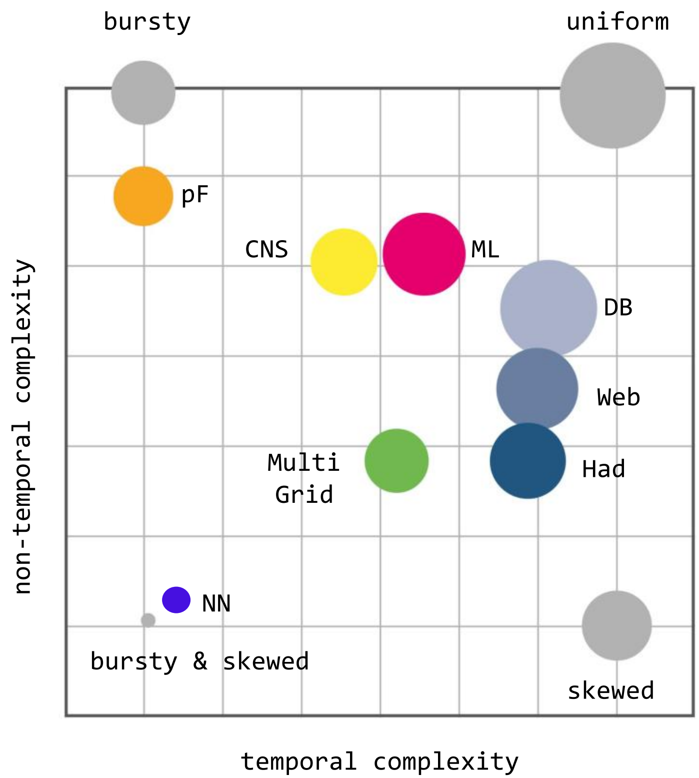
No structure

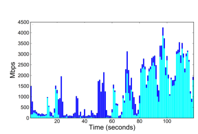
Our **approach:** iterative **randomization and compression** of trace to identify dimensions of structure.

Avin et al. (Sigmetrics’2020)

Complexity Map





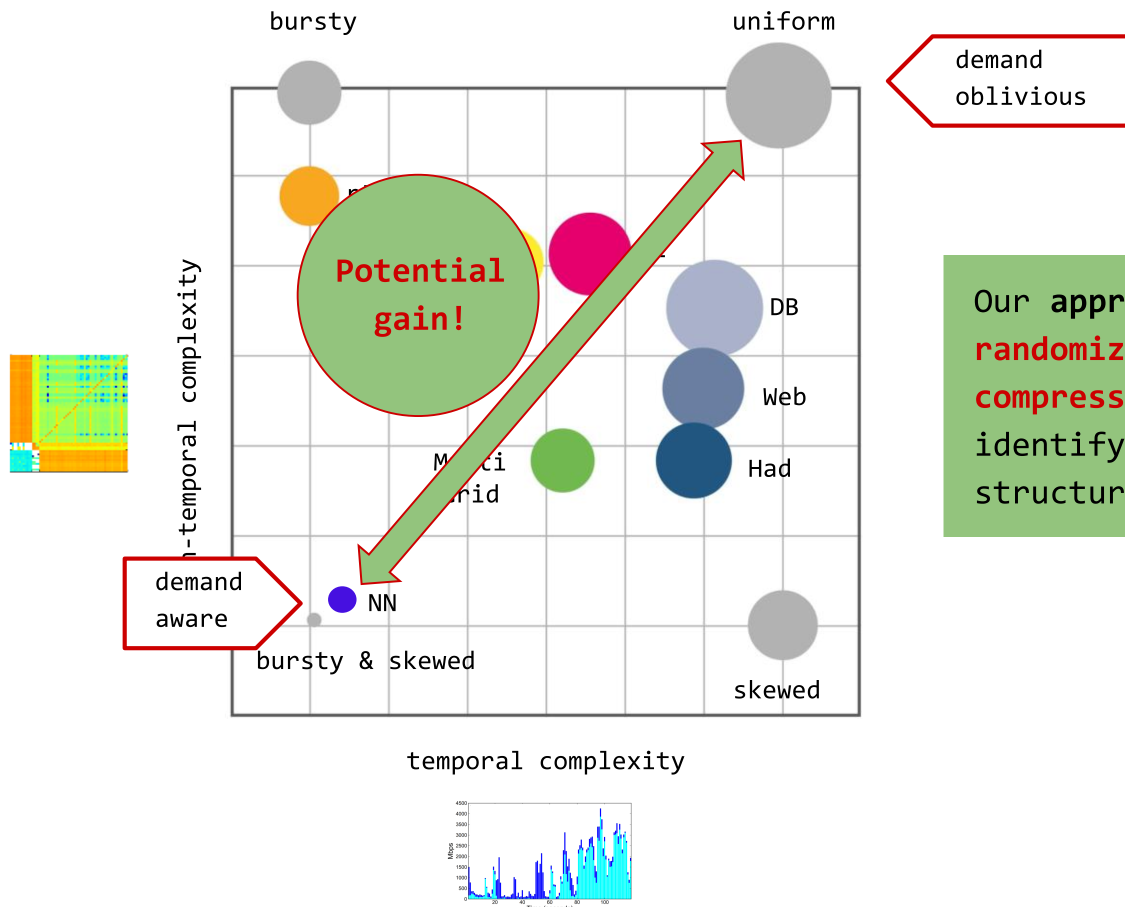


Our **approach**: iterative **randomization and compression** of trace to identify dimensions of structure.

Different structures!

Avin et al. (Sigmetrics'2020)

Complexity Map

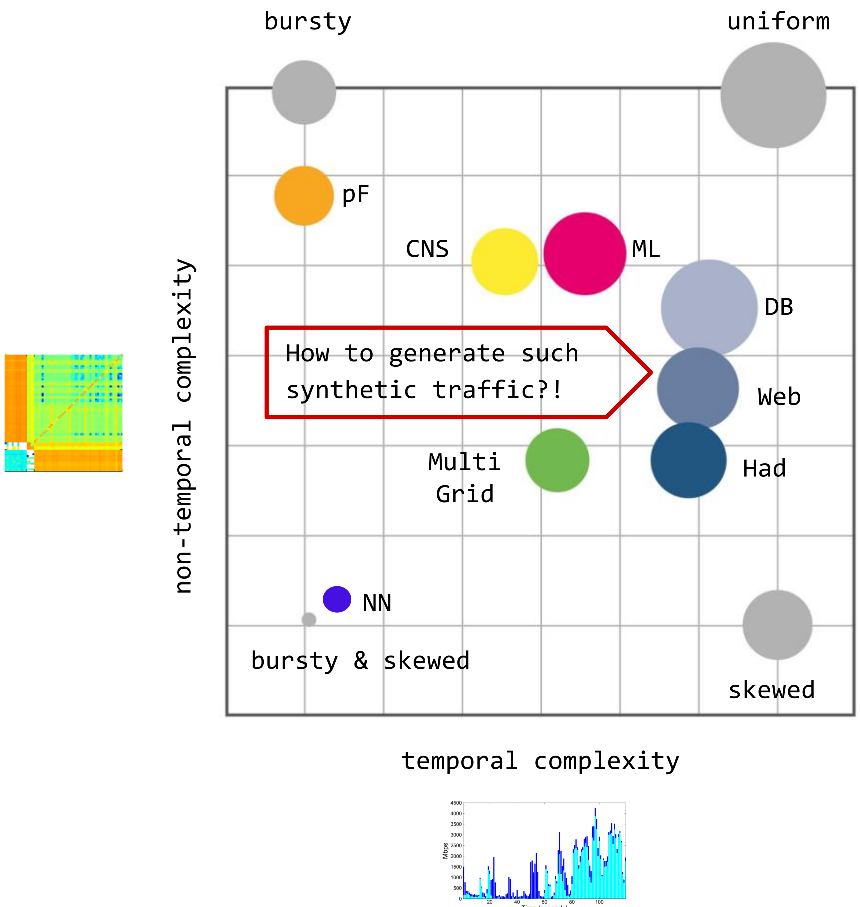


Our **approach**: iterative **randomization and compression** of trace to identify dimensions of structure.

Literature: On the Complexity of Traffic Traces and Implications. Avin et al., ACM SIGMETRICS, 2020.

Avin et al. (Sigmetrics’2020)

Complexity Map

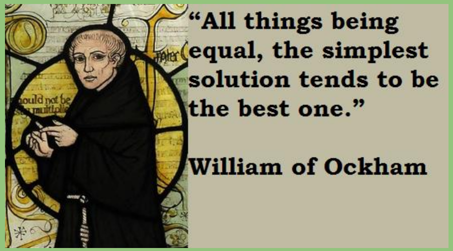


Our **approach**: iterative **randomization and compression** of trace to identify dimensions of structure.

Literature: On the Complexity of Traffic Traces and Implications. Avin et al., ACM SIGMETRICS, 2020.

From Analysis to

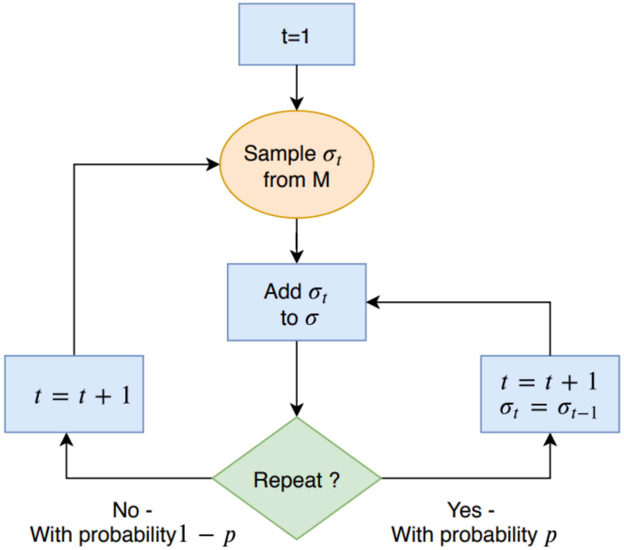
Synthesis



- Complexity map is just 2-dimensional: many ways to synthesize any point on map

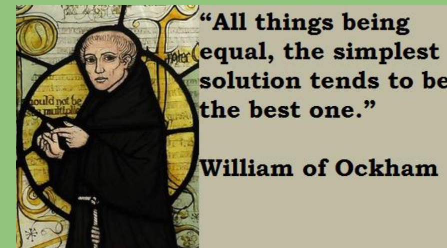
- Most simple ("Occam's razor"):
  - ***Spatial distribution:*** empirical traffic matrix M (or synthetic distribution, e.g. Zipf)
  - ***Temporal distribution:*** repeat with probability p (can be computed analytically from data)

- Resulting ***Markov process*** generates corresponding disk on complexity map
  - ***Stationary distribution*** corresponds to M
  - Temporary pattern matches ***entropy rate***



From Analysis to

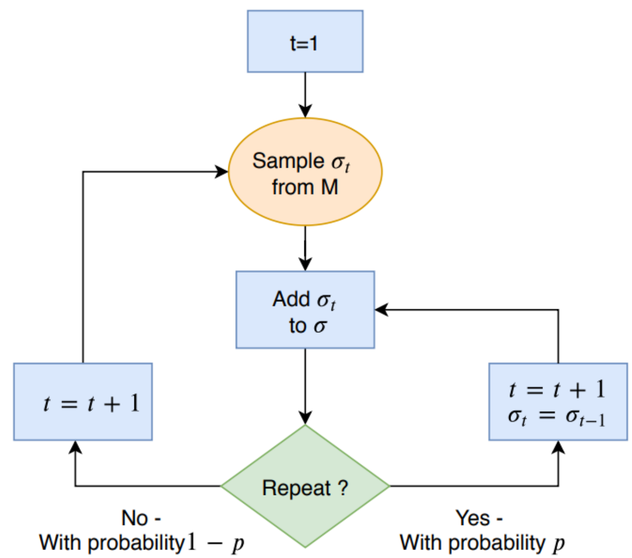
Synthesis



- Complexity map is just 2-dimensional: many ways to synthesize any point on map

- Most simple ("Occam's razor"):
  - ***Spatial distribution:*** empirical traffic matrix M (or synthetic distribution, e.g. Zipf)
  - ***Temporal distribution:*** repeat with probability p (can be computed analytically from data)

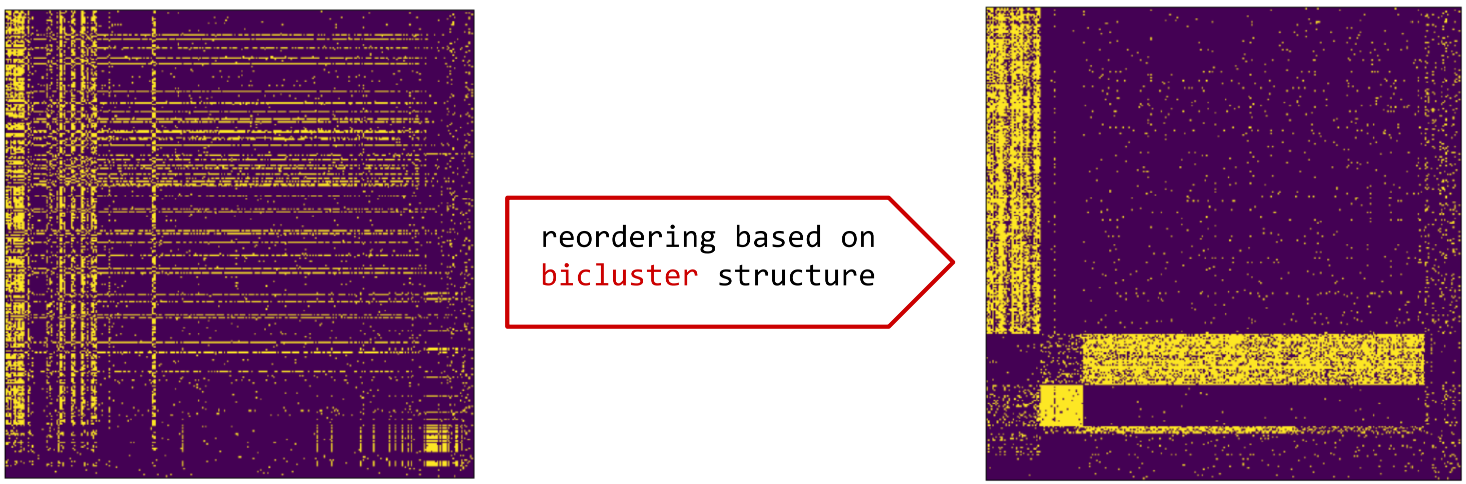
- Resulting ***Markov process*** generates corresponding disk on complexity map
  - ***Stationary distribution*** corresponds to M
  - Temporary pattern matches ***entropy rate***



Literature: On the Complexity of Traffic Traces and Implications. Avin et al., ACM SIGMETRICS, 2020.

Traffic is also clustered:

Small Stable Clusters



Opportunity: *exploit* with little reconfigurations!

Literature: Analyzing the Communication Clusters in Datacenters. Foerster et al. WWW Conference, 2023.

Further Reading

[On the Complexity of Traffic Traces and Implications]
Chen Avin, Manya Ghobadi, Chen Griner, and Stefan Schmid.
ACM **SIGMETRICS** and ACM Performance Evaluation Review (**PER**), Boston, Massachusetts, USA, June 2020.

[Analyzing the Communication Clusters in Datacenters]
Klaus-Tycho Foerster, Thibault Marette, Stefan Neumann, Claudia Plant, Ylli Sadikaj, Stefan Schmid, and Yllka Velaj.
The Web Conference (**WWW**), Austin, Texas, USA, April 2023.

[Network Traffic Characteristics of Machine Learning Frameworks Under the Microscope]
Johannes Zerwas, Kaan Aykurt, Stefan Schmid, and Andreas Blenk. 17th International Conference on Network and Service Management (**CNSM**), Izmir, Turkey, October 2021.

Website: trace-collection.net

The Natural Question:

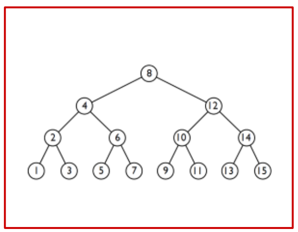
Given This Structure, What Can Be Achieved? Metrics and Algorithms?

Also depends on entropy of the demand!

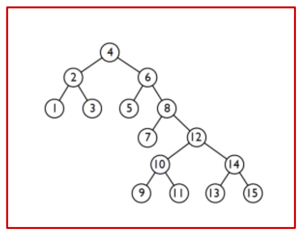
Insight:

Connection to Datastructures

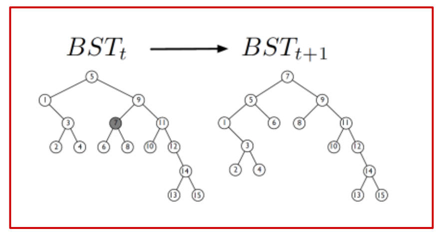
Traditional BST



Demand-aware BST



Self-adjusting BST

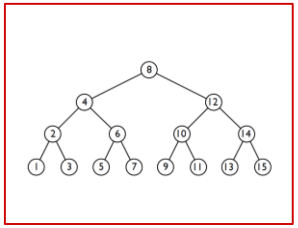


More structure: improved **access cost**

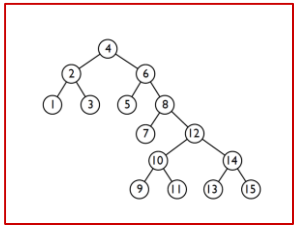
Insight:

Connection to Datastructures & Coding

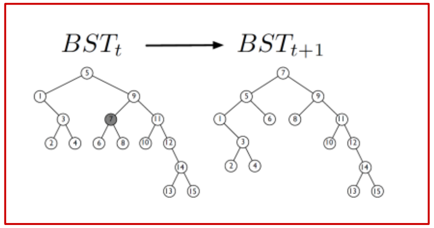
Traditional BST
(Worst-case coding)



Demand-aware BST
(Huffman coding)



Self-adjusting BST
(Dynamic Huffman coding)



More structure: improved **access cost** / shorter **codes**

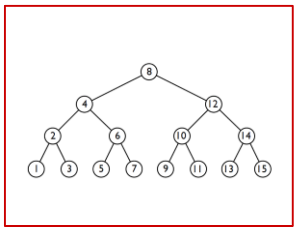
Insight:

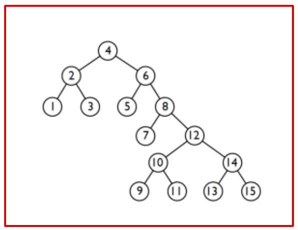
Connection to Datastructures & Coding

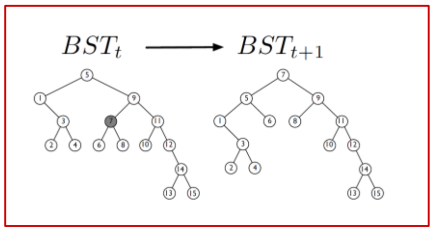
Traditional BST
(Worst-case coding)

Demand-aware BST
(Huffman coding)

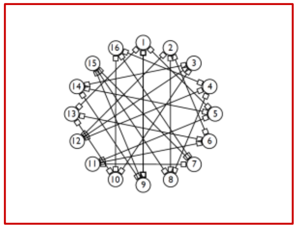
Self-adjusting BST
(Dynamic Huffman coding)

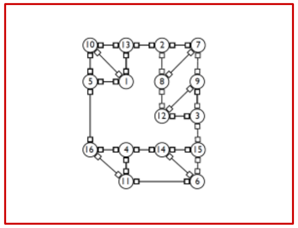


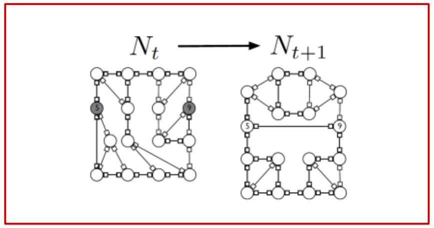




More structure: improved **access cost** / shorter **codes**







Similar **benefits**?

Insight:

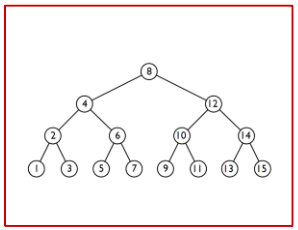
Connection to Datastructures & Coding

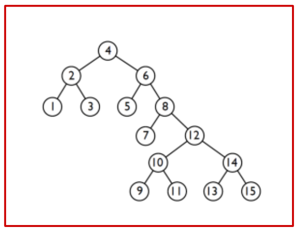
Traditional BST
(Worst-case coding)

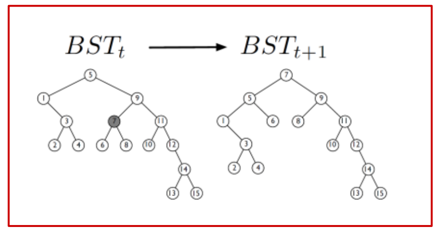
Demand-aware BST
(Huffman coding)

Self-adjusting BST
(Dynamic Huffman coding)

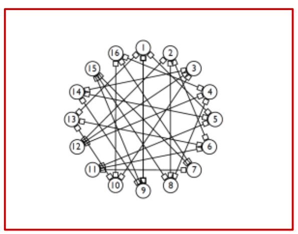
More than
an analogy!

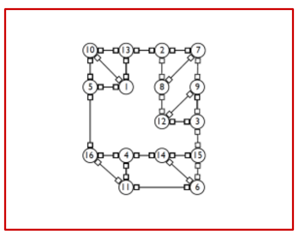


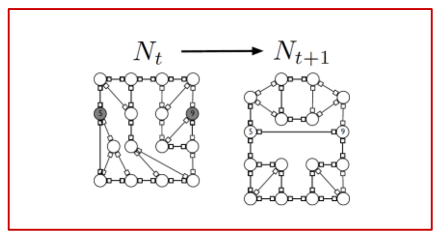




More structure: improved **access cost** / shorter **codes**







Similar **benefits**?

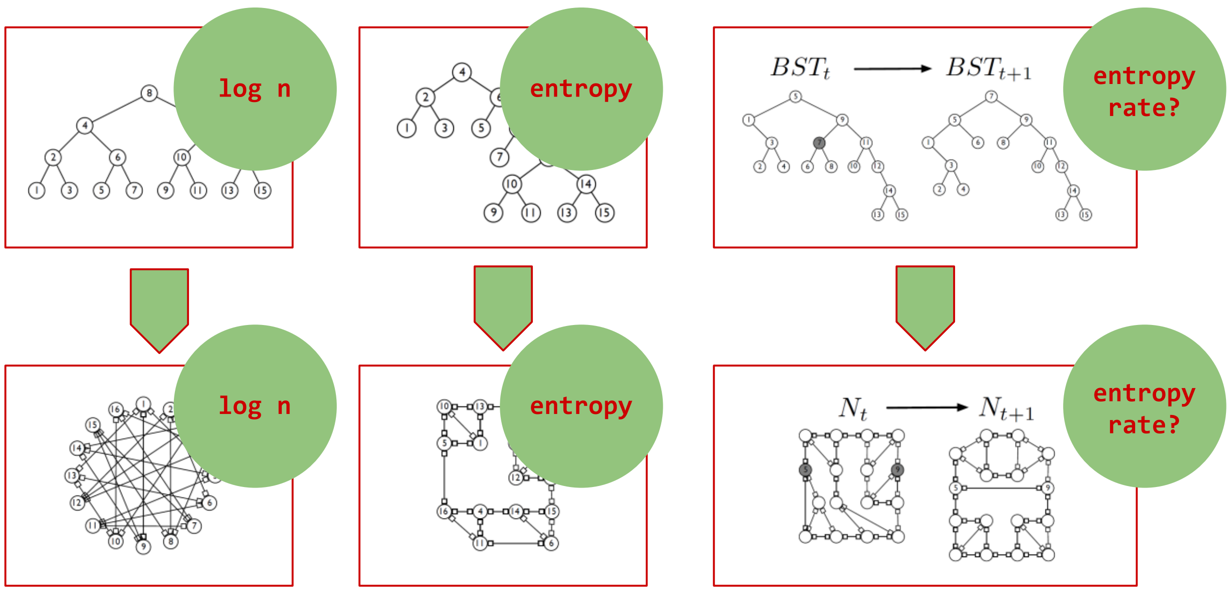
Insight:

Connection to Datastructures & Coding

Traditional BST
(Worst-case coding)

Demand-aware BST
(Huffman coding)

Self-adjusting BST
(Dynamic Huffman coding)



Reduced expected **route lengths**!

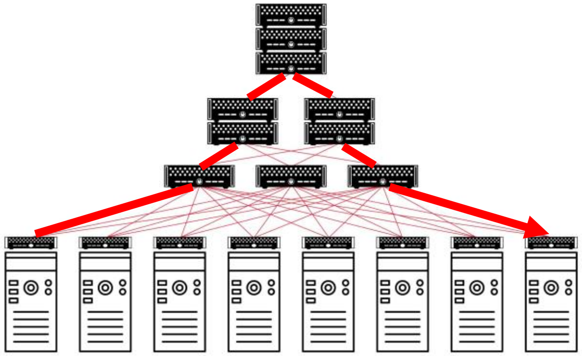
More than
an analogy!

Generalize methodology:
... and transfer entropy bounds and algorithms of data-structures to networks.

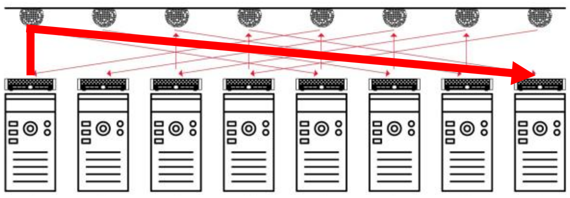
First result:
Demand-aware networks of asymptotically optimal route lengths.

Reality more complicated

→ Self-adjusting networks may be really useful to serve large flows (elephant flows): avoiding multi-hop routing



vs

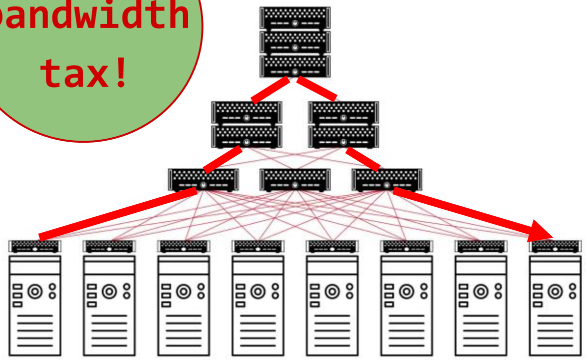


6 hops **1 hop**

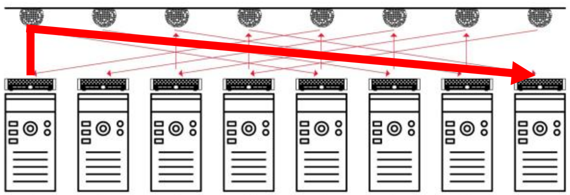
Reality more complicated

→ Self-adjusting networks may be really useful to serve large flows (elephant flows): avoiding multi-hop routing

bandwidth tax!



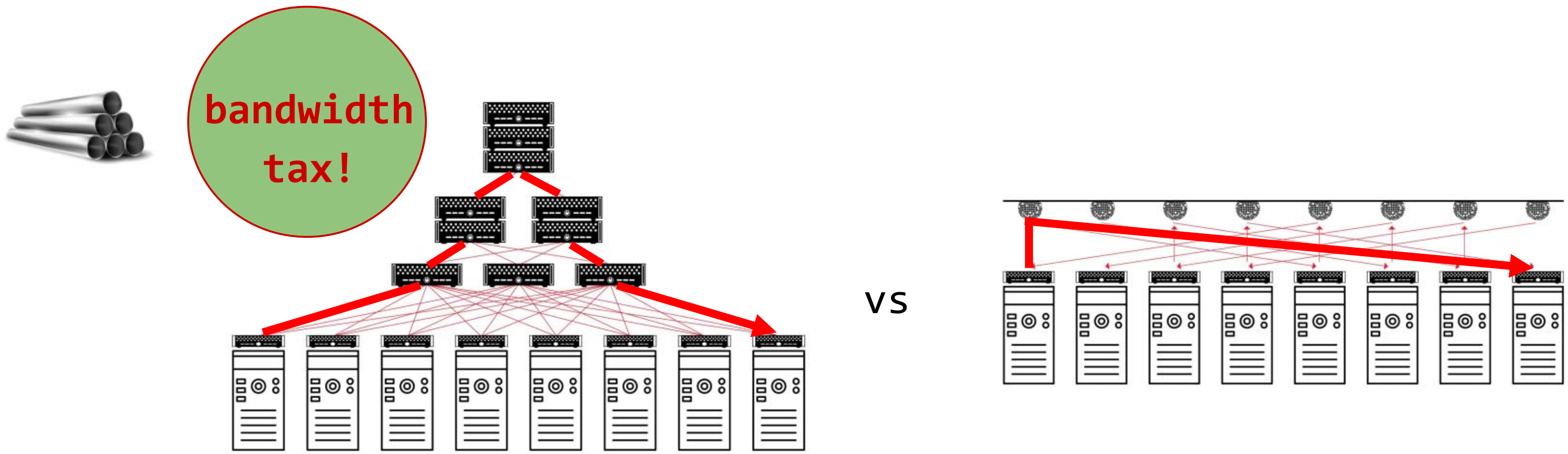
vs



6 hops **1 hop**

Reality more complicated

→ Self-adjusting networks may be really useful to serve large flows (elephant flows): avoiding multi-hop routing

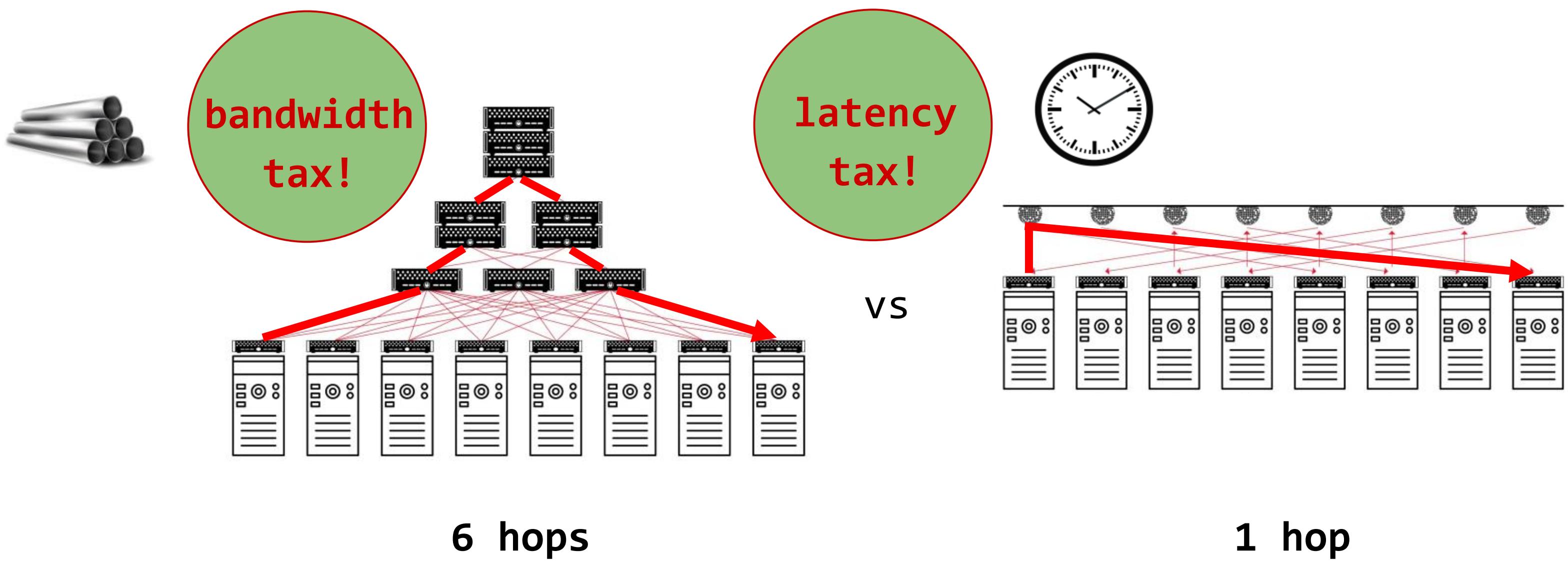


6 hops **1 hop**

→ However, requires optimization and adaption, which takes time

Reality more complicated

→ Self-adjusting networks may be really useful to serve large flows (elephant flows): avoiding multi-hop routing



→ However, requires optimization and adaption, which takes time

Indeed, it is more complicated than that…

Challenge: Traffic Diversity

Diverse patterns:

⇀ Shuffling/Hadoop: all-to-all

⇀ All-reduce/ML: ring or tree traffic patterns

  ⇀ Elephant flows

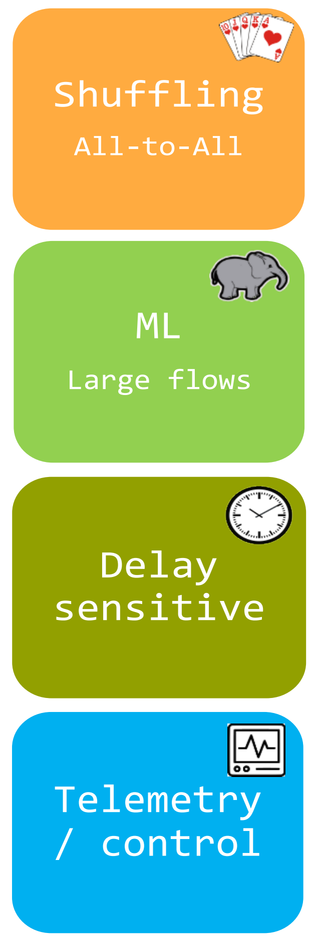
⇀ Query traffic: skewed

  ⇀ Mice flows

⇀ Control traffic: does not evolve but has non-temporal structure

Diverse requirements:

⇀ ML is bandwidth hungry, small flows are latency-sensitive



Opportunity: Tech Diversity

Diverse topology components:

→ demand-oblivious and
demand-aware

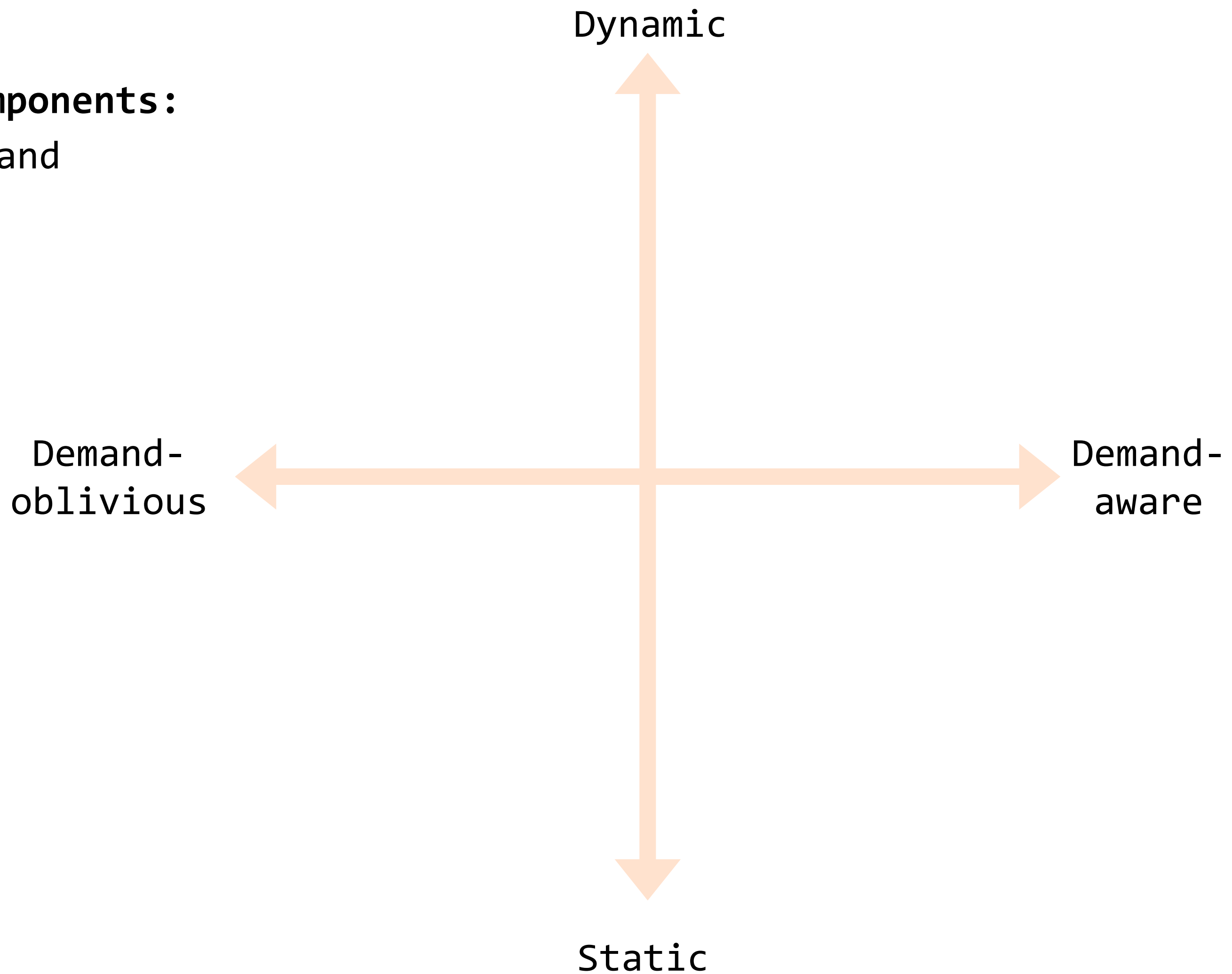
Demand-oblivious ⟷ Demand-aware

Opportunity: Tech Diversity

Diverse topology components:

⇀ demand-oblivious and
  demand-aware

⇀ static vs dynamic

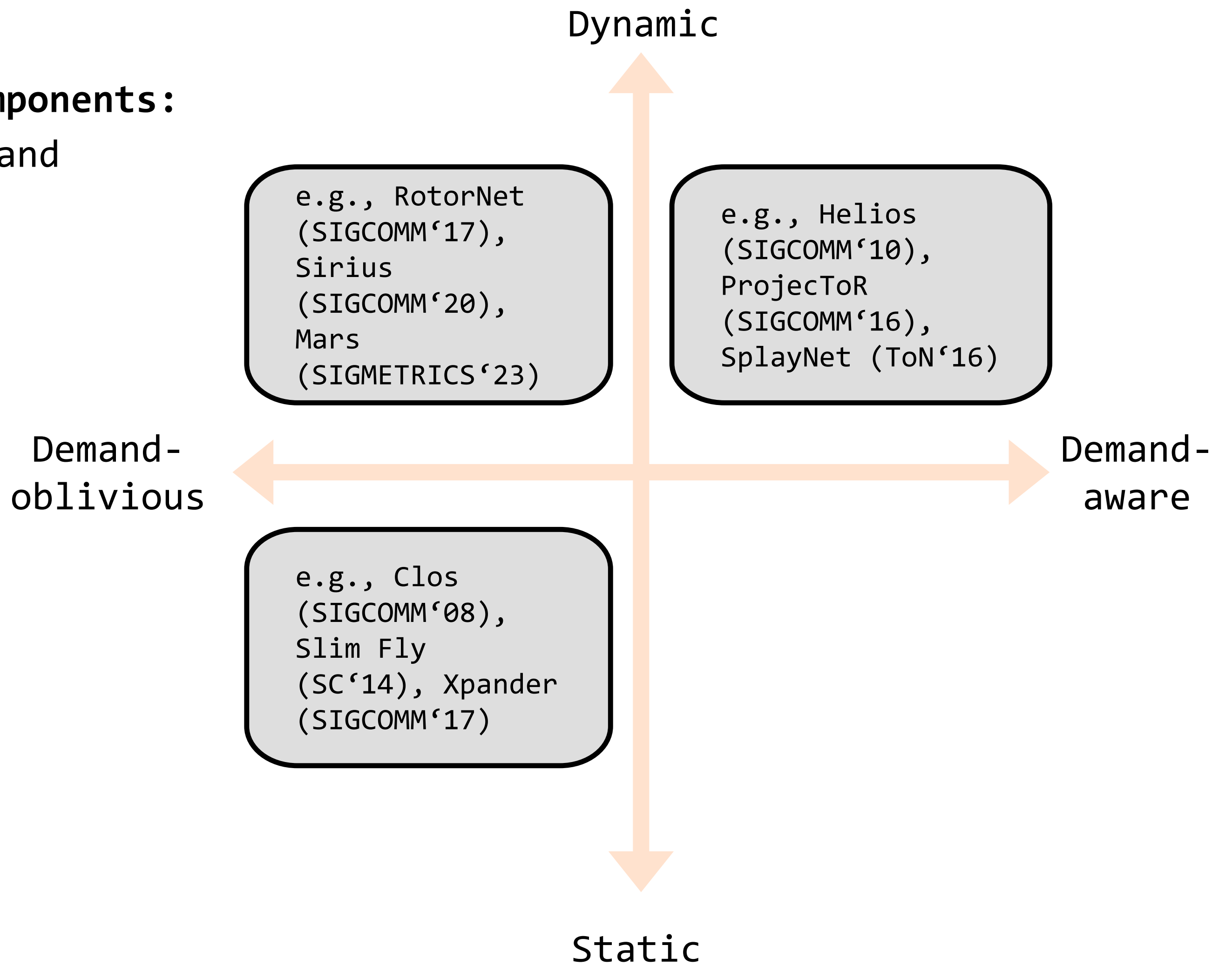


Opportunity: Tech Diversity

Diverse topology components:

⇀ demand-oblivious and demand-aware

⇀ static vs dynamic

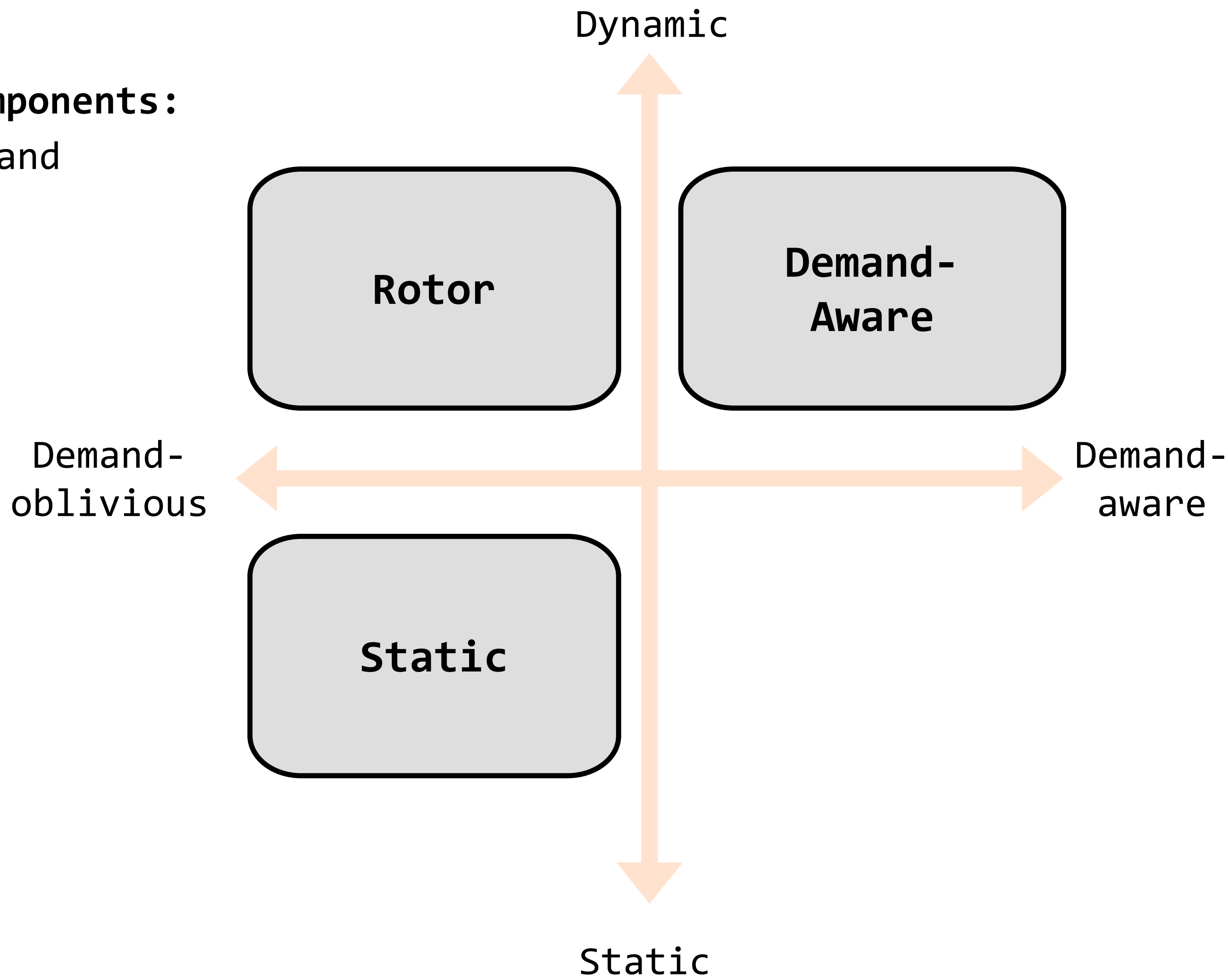


Opportunity: Tech Diversity

Diverse topology components:

⇀ demand-oblivious and demand-aware

⇀ static vs dynamic



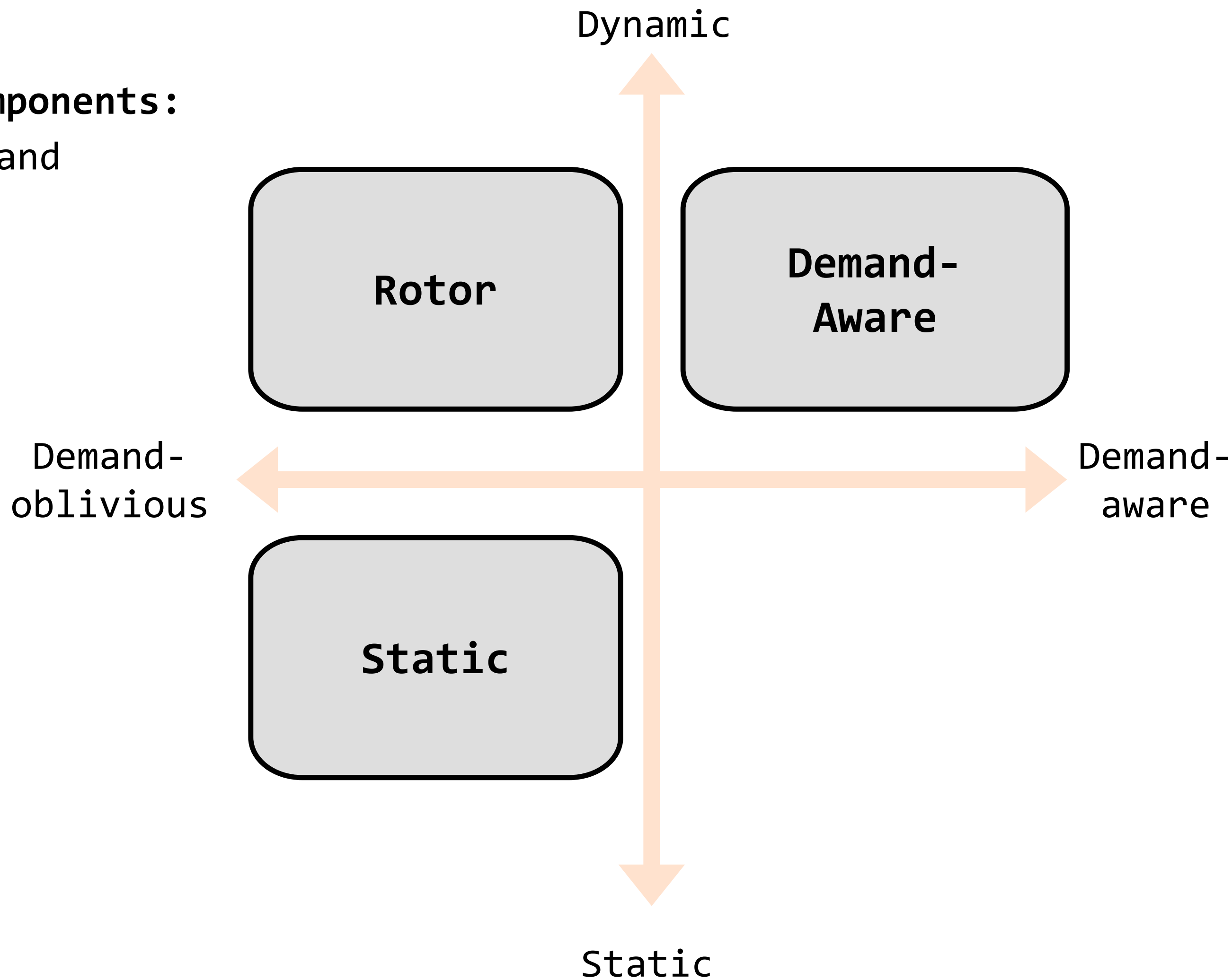
Opportunity: Tech Diversity

Diverse topology components:

⇀ demand-oblivious and demand-aware

⇀ static vs dynamic

Which approach is best?

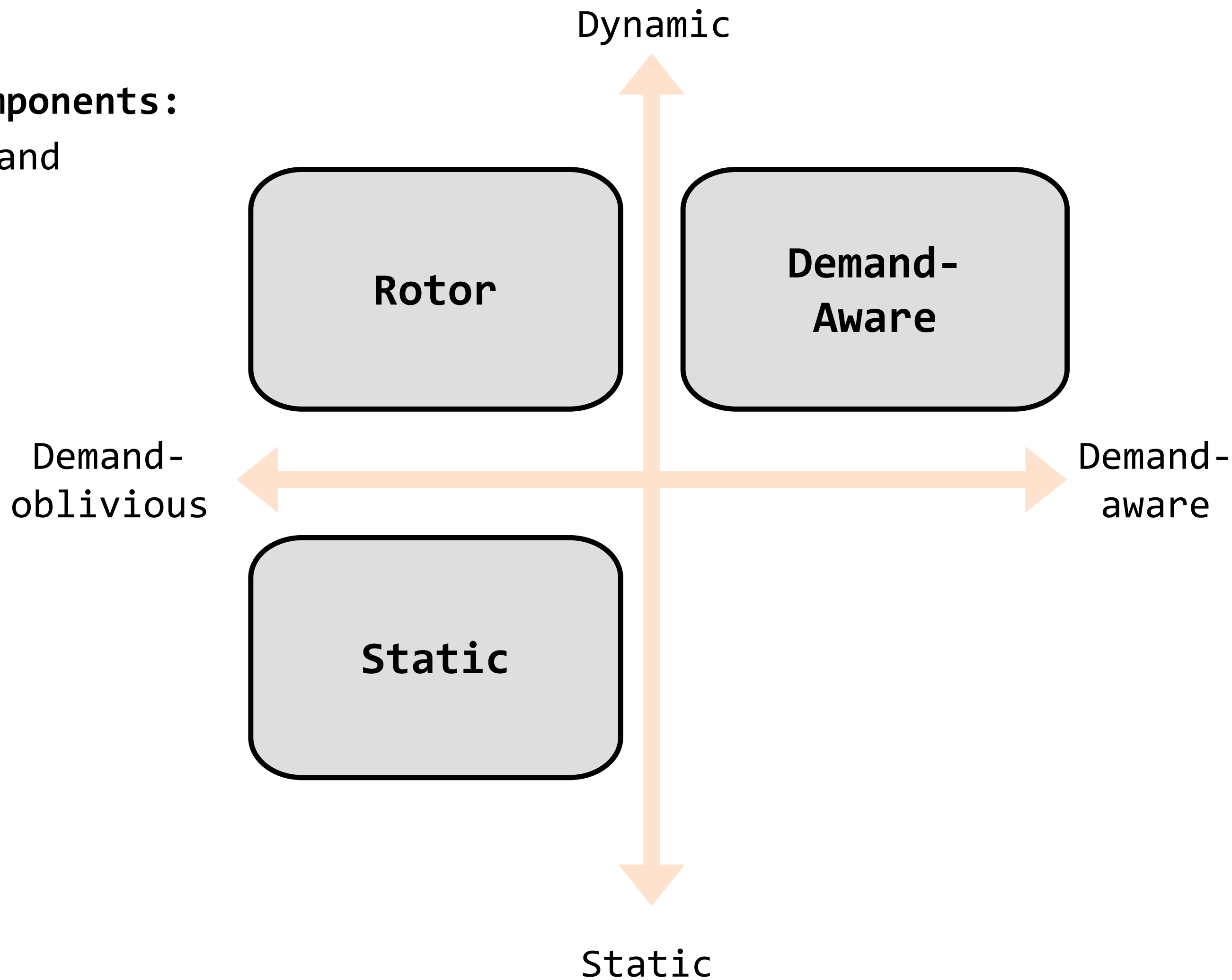


Opportunity: Tech Diversity

Diverse topology components:

⇀ demand-oblivious and demand-aware

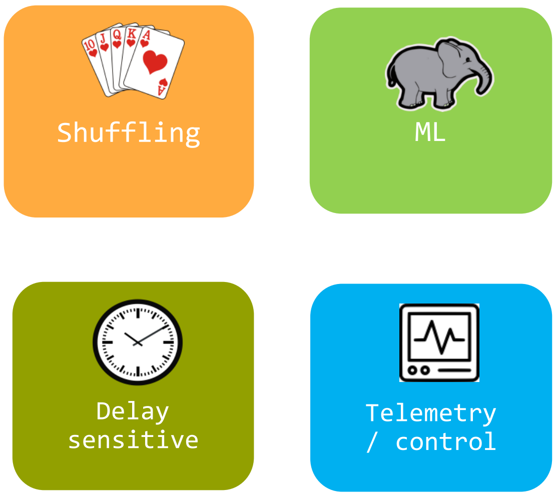
⇀ static vs dynamic



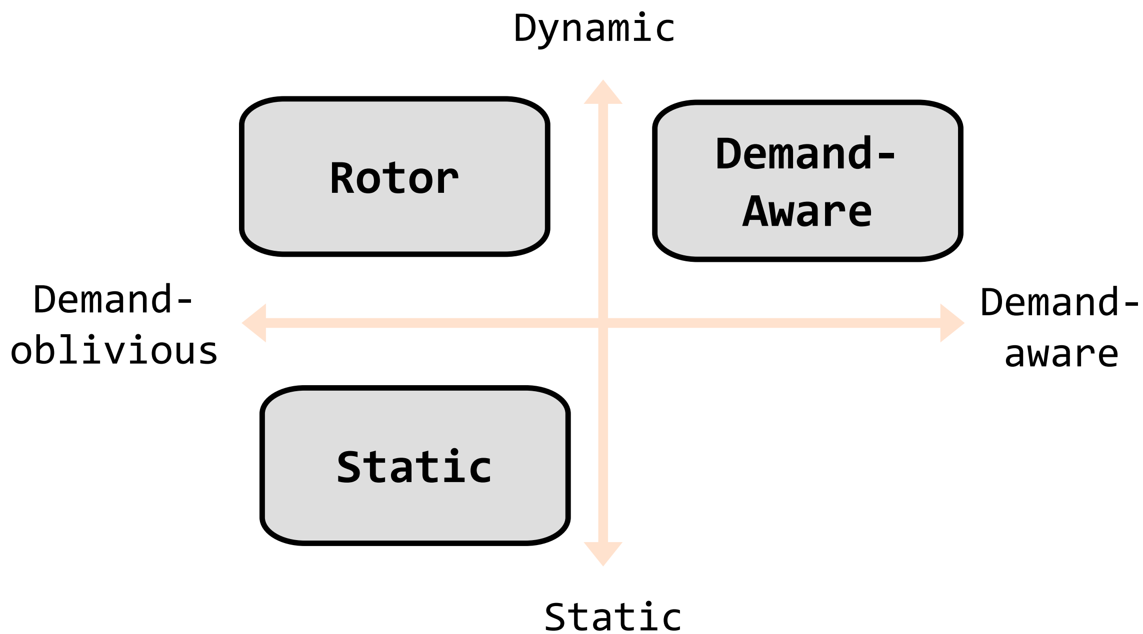
Which approach is best?

As always in CS: It depends…

Examples: Match or Mismatch?

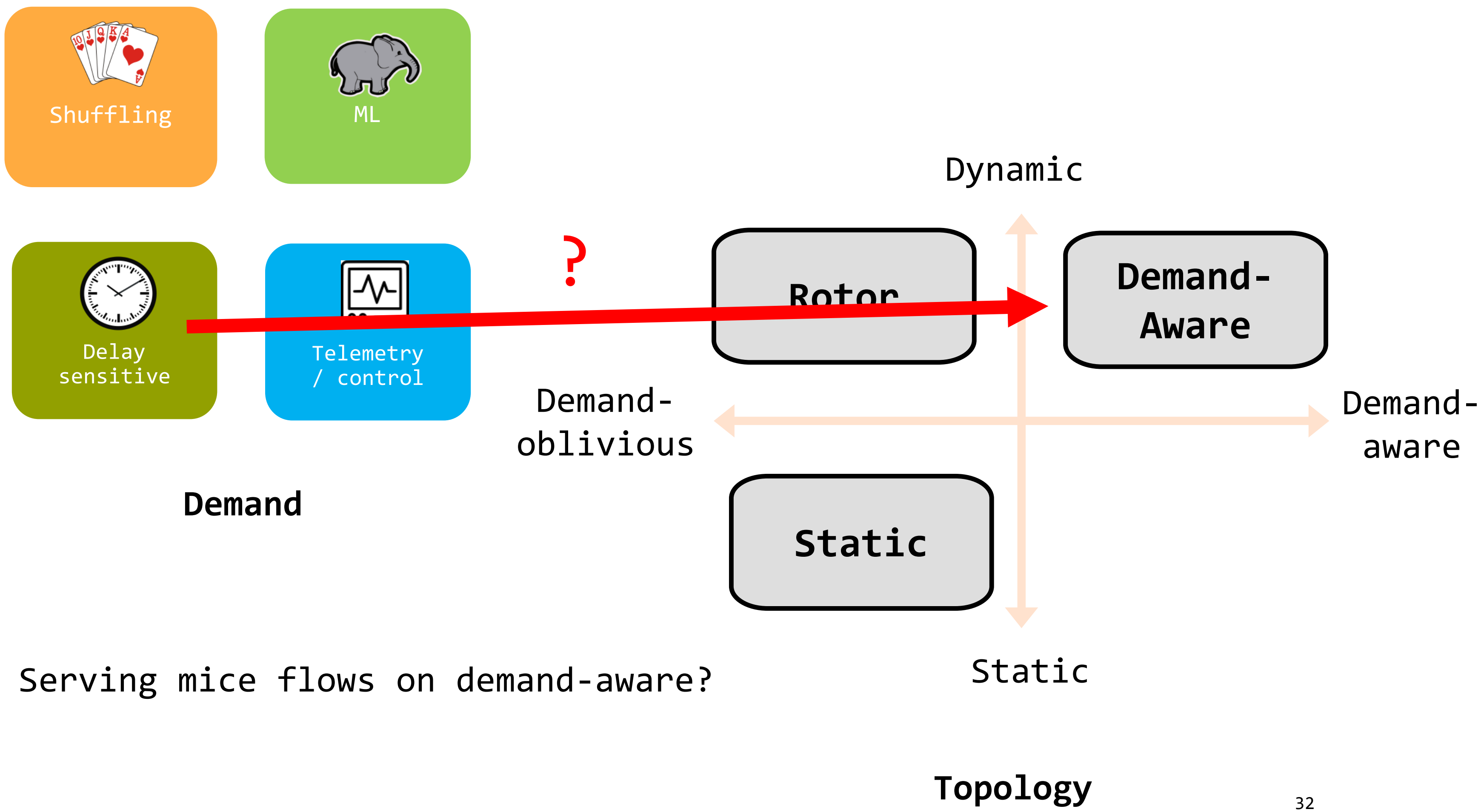


Demand



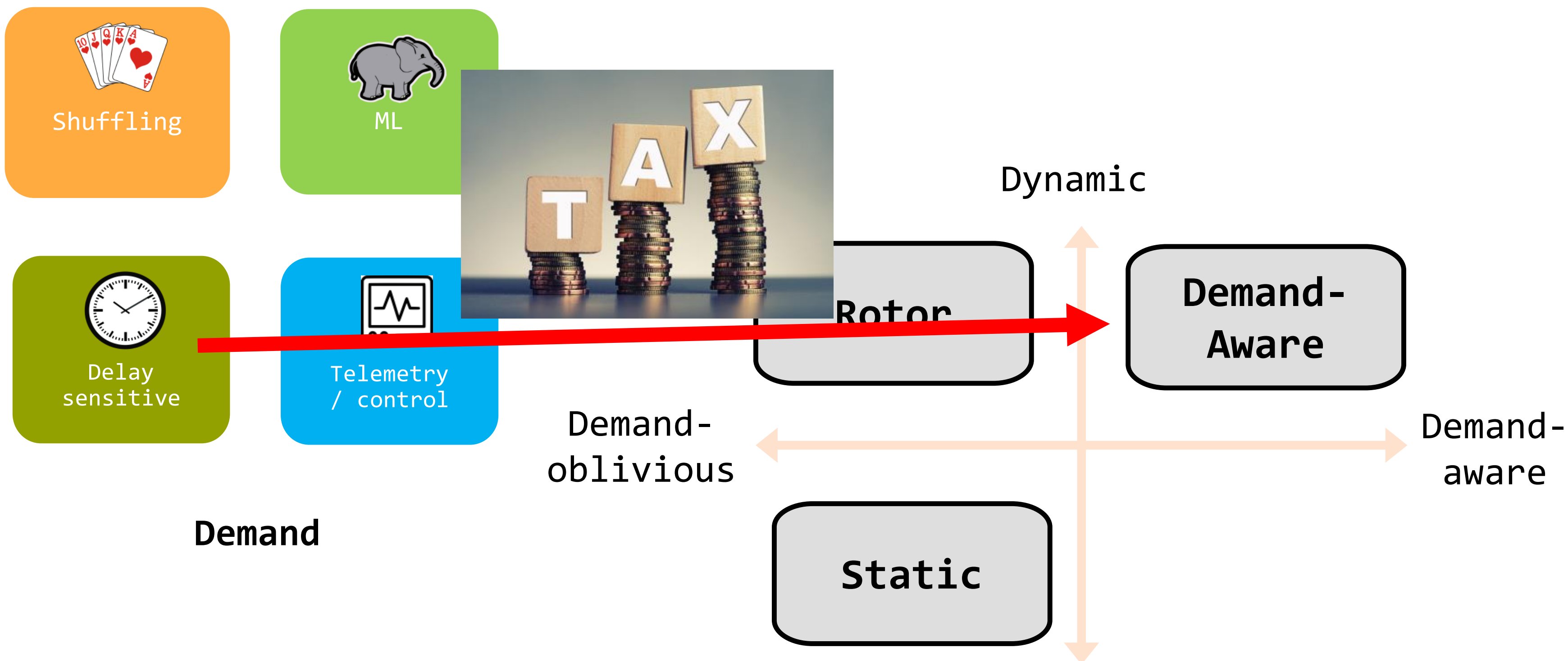
Topology

Examples:
Match or Mismatch?



Serving mice flows on demand-aware?

Examples: Match or Mismatch?



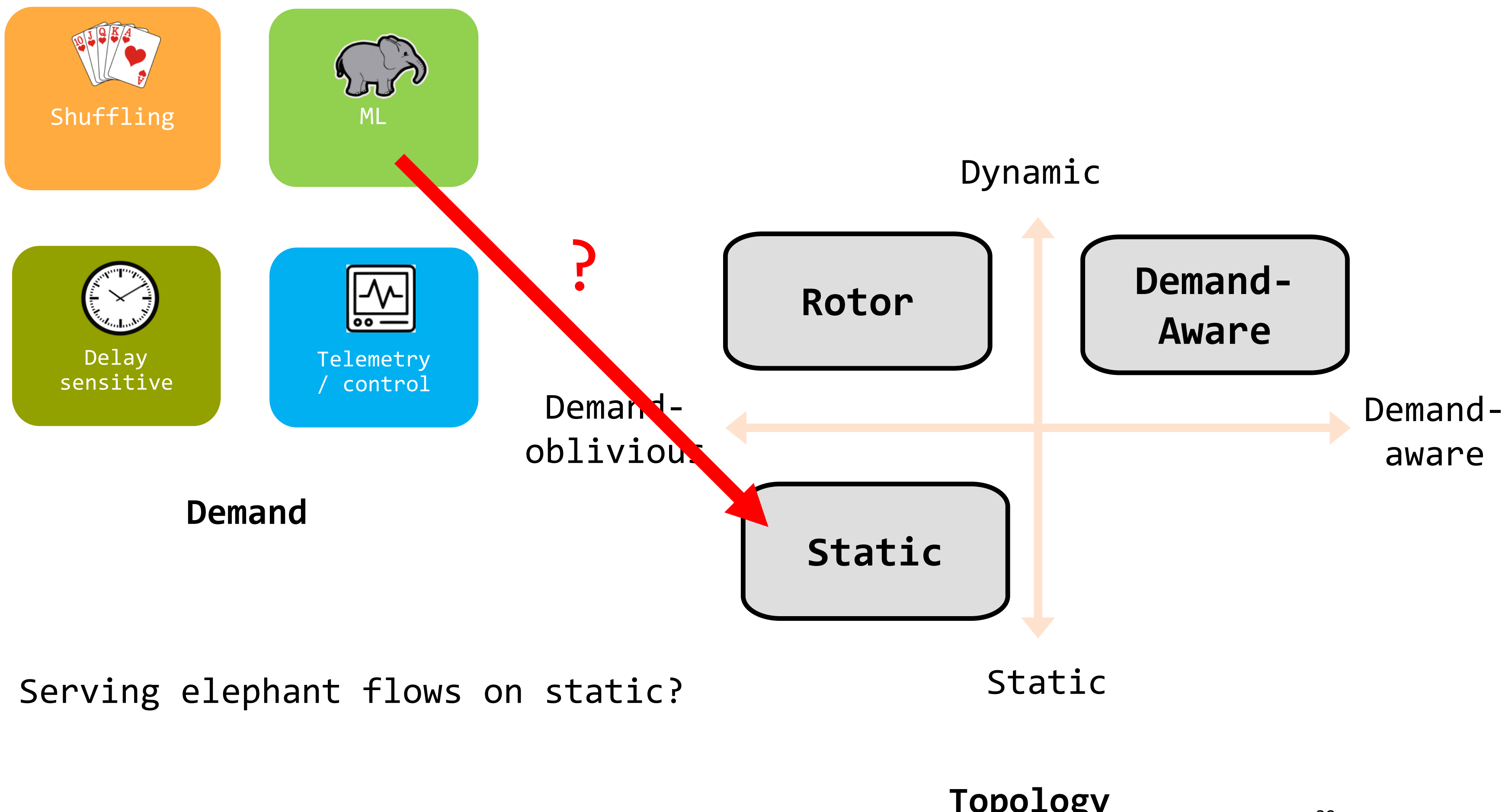
Serving mice flows on demand-aware?

Bad idea! Latency tax.

Static

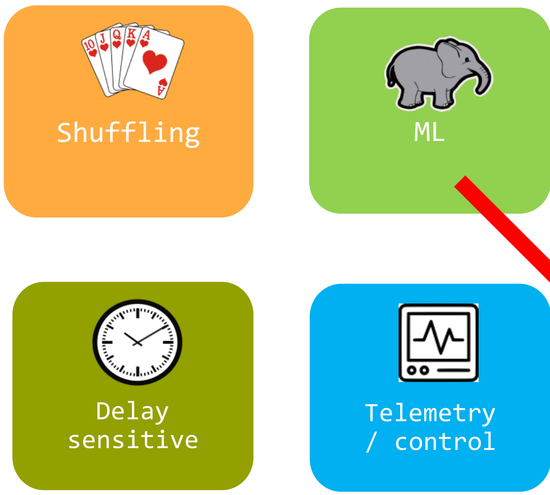
Topology

Examples:
Match or Mismatch?

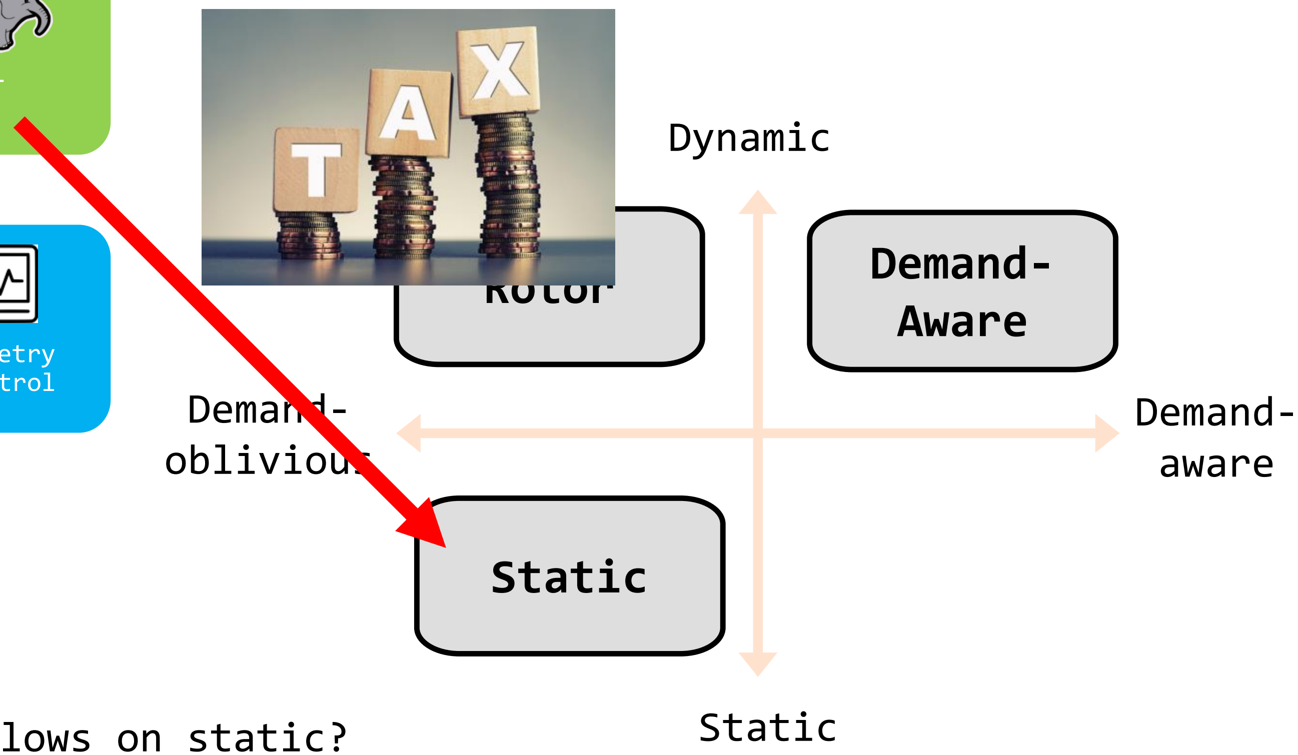


Serving elephant flows on static?

Examples:
Match or Mismatch?



Demand

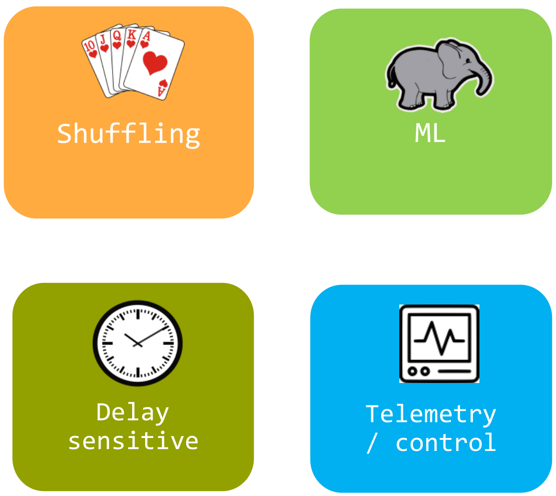


Topology

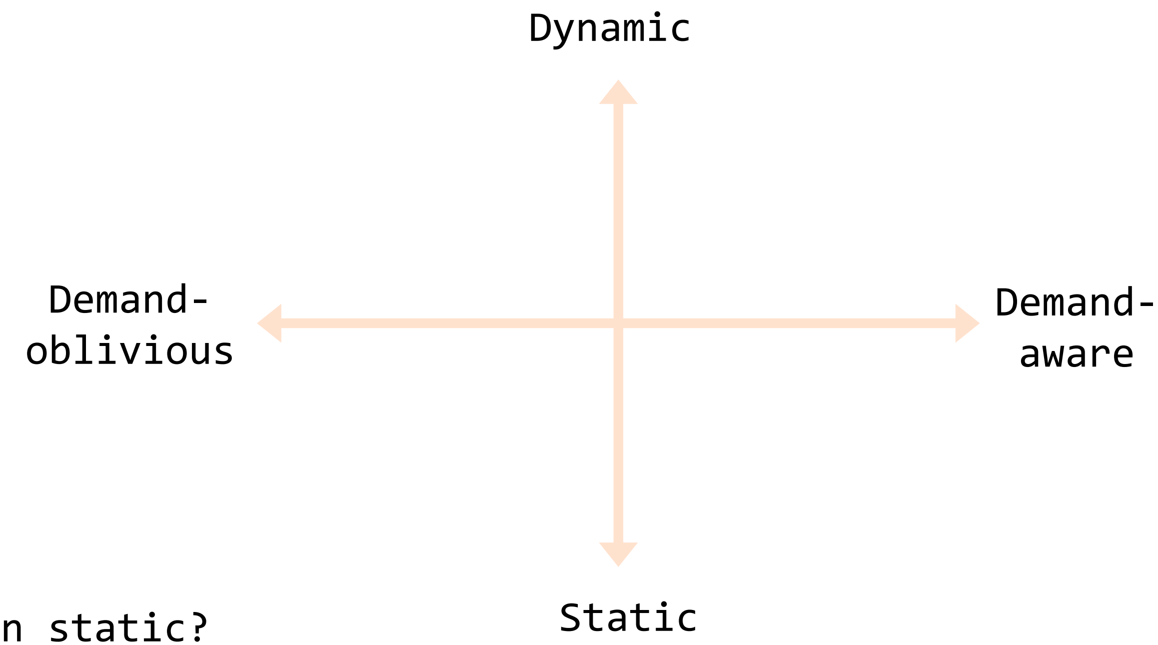
Serving elephant flows on static?

Bad idea! Bandwidth tax.

Examples:
Match or Mismatch?



Demand

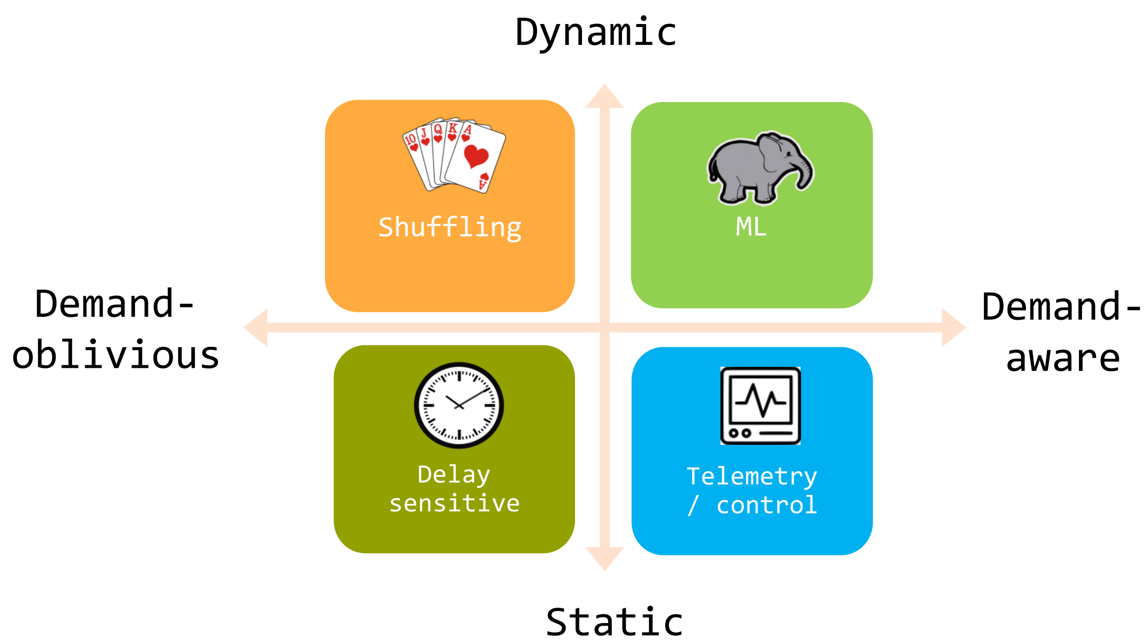


Serving elephant flows on static?

Bad idea! Bandwidth tax.

Topology

Optimal Solution: It's a Match!



We have a first approach:

Cerberus*** serves traffic on the “best topology”! (Optimality open)

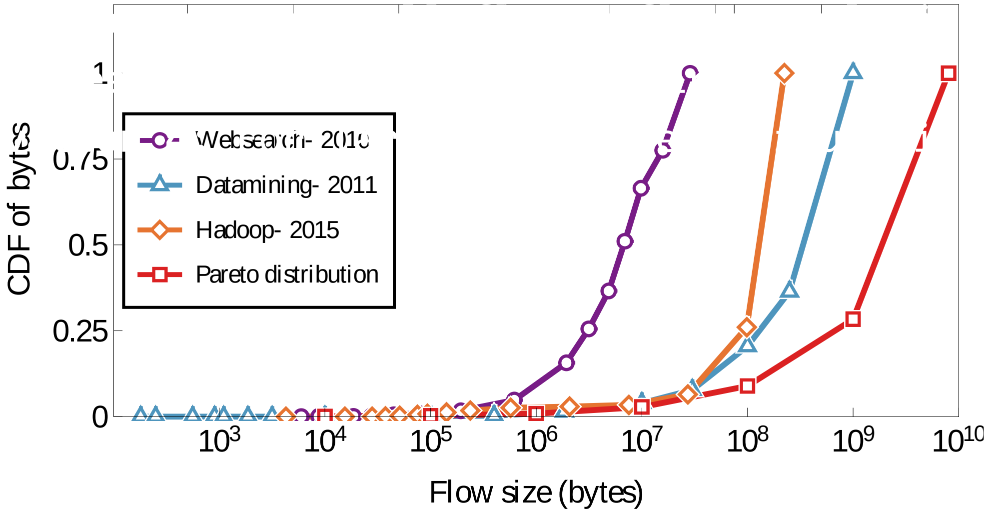
* Cerberus: The Power of Choices in Datacenter Topology Design. Griner et al. ACM SIGMETRICS, 2022.

Flow Size Matters

On what should topology type depend? We argue: flow size.

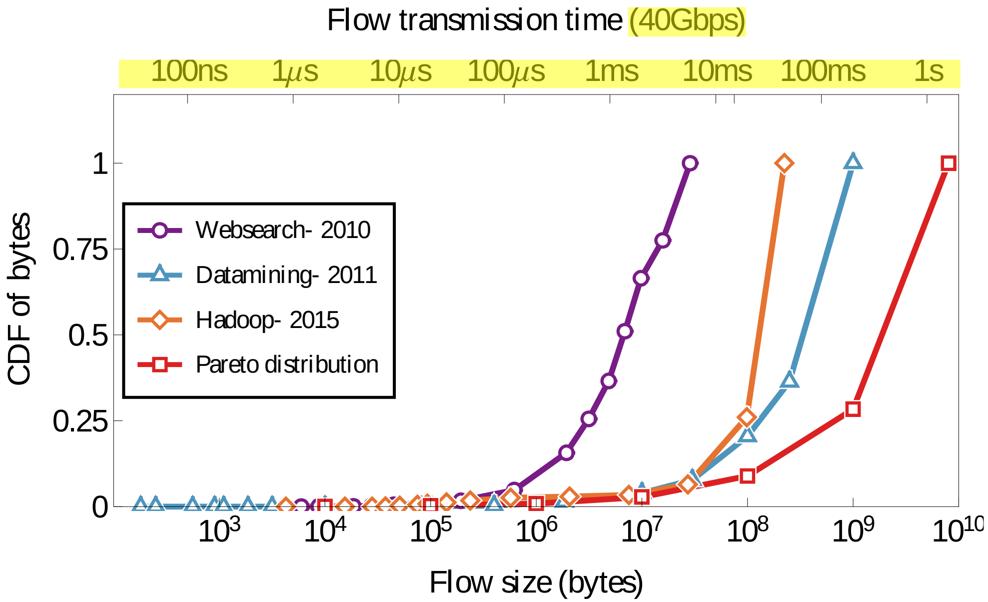
Flow Size Matters

On what should topology type depend? We argue: flow size.



⇢ **Observation 1:** Different apps have different flow size distributions.

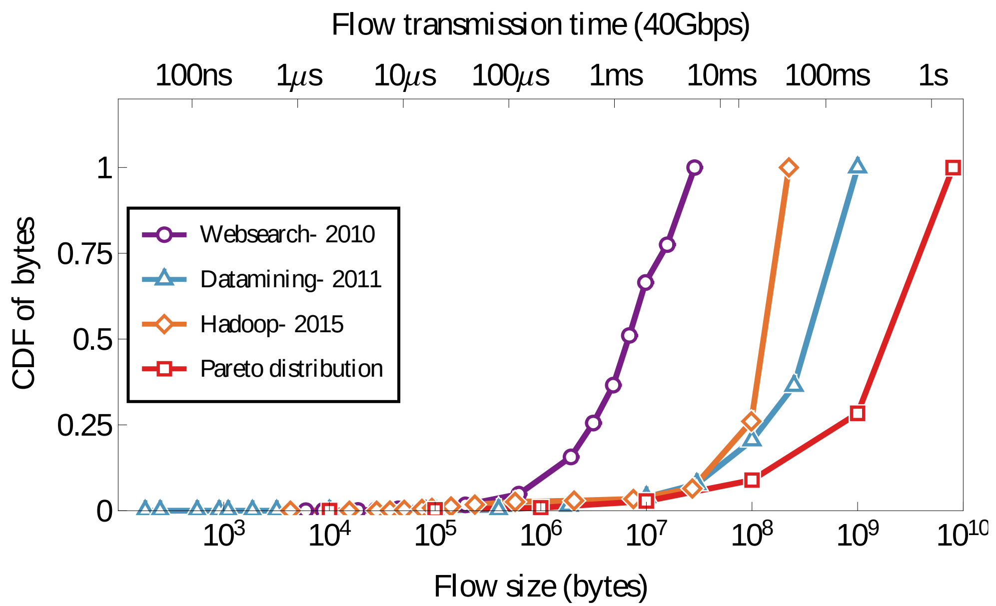
Flow Size Matters



⇢ **Observation 1:** Different apps have different flow size distributions.

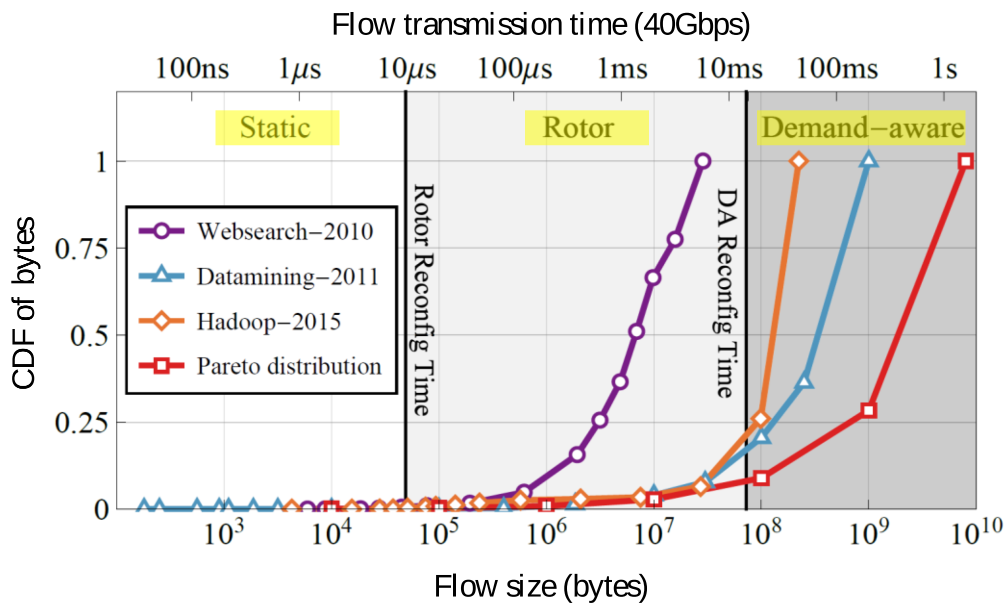
⇢ **Observation 2:** The transmission time of a flow depends on its size.

Flow Size Matters



- **Observation 1:** Different apps have different flow size distributions.
- **Observation 2:** The transmission time of a flow depends on its size.
- **Observation 3:** For small flows, flow completion time suffers if network needs to be reconfigured first.
- **Observation 4:** For large flows, reconfiguration time may amortize.

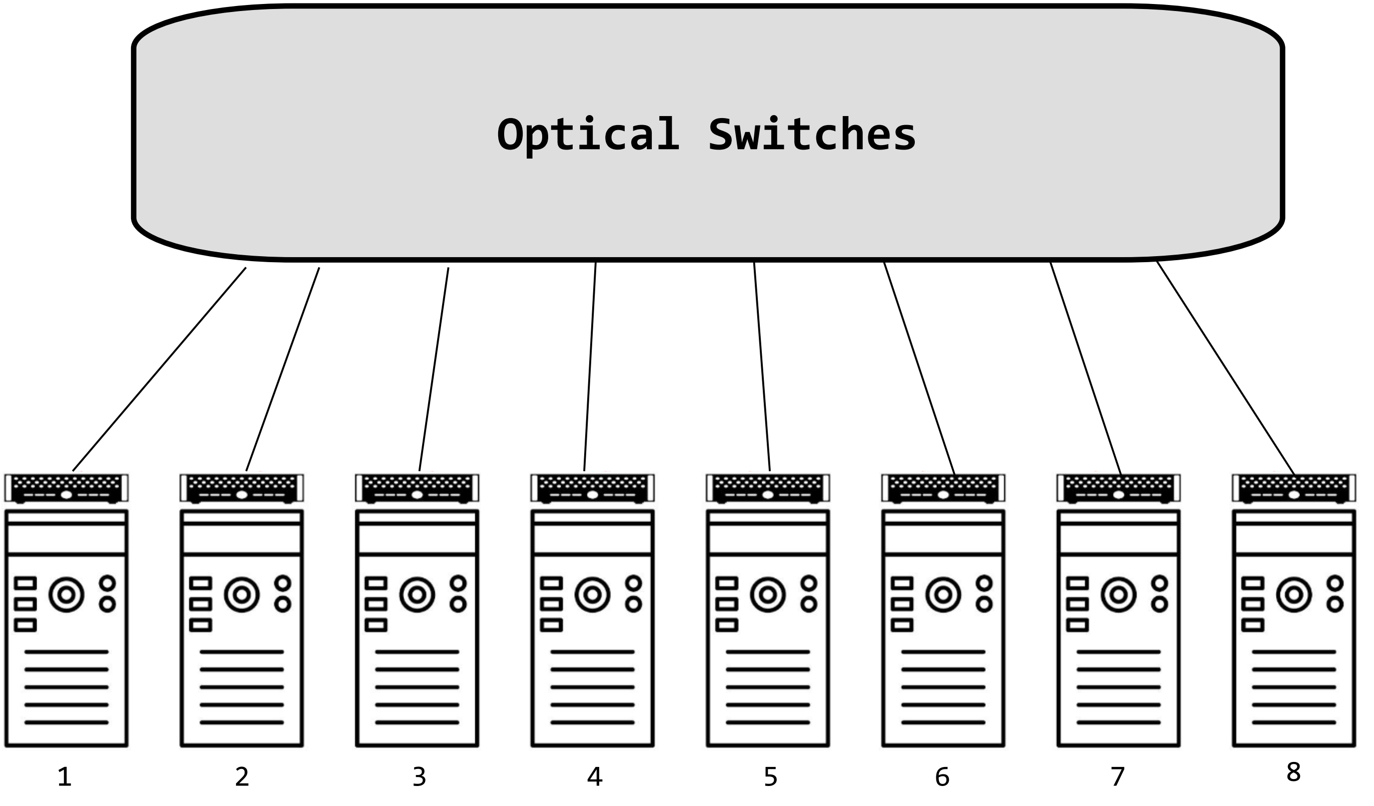
Flow Size Matters



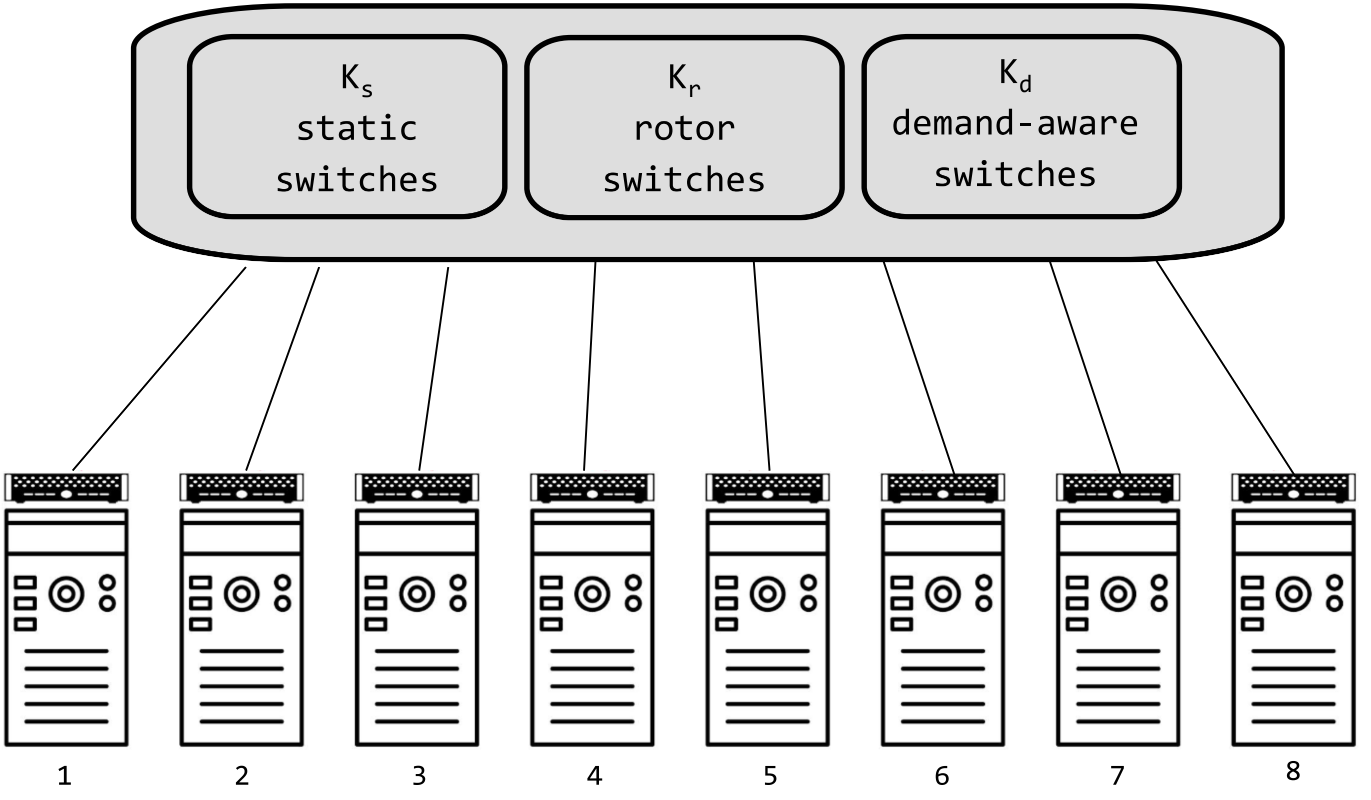
It's a Match!

- **Observation 1:** Different apps have different flow size distributions.
- **Observation 2:** The transmission time of a flow depends on its size.
- **Observation 3:** For small flows, flow completion time suffers if network needs to be reconfigured first.
- **Observation 4:** For large flows, reconfiguration time may amortize.

Cerberus

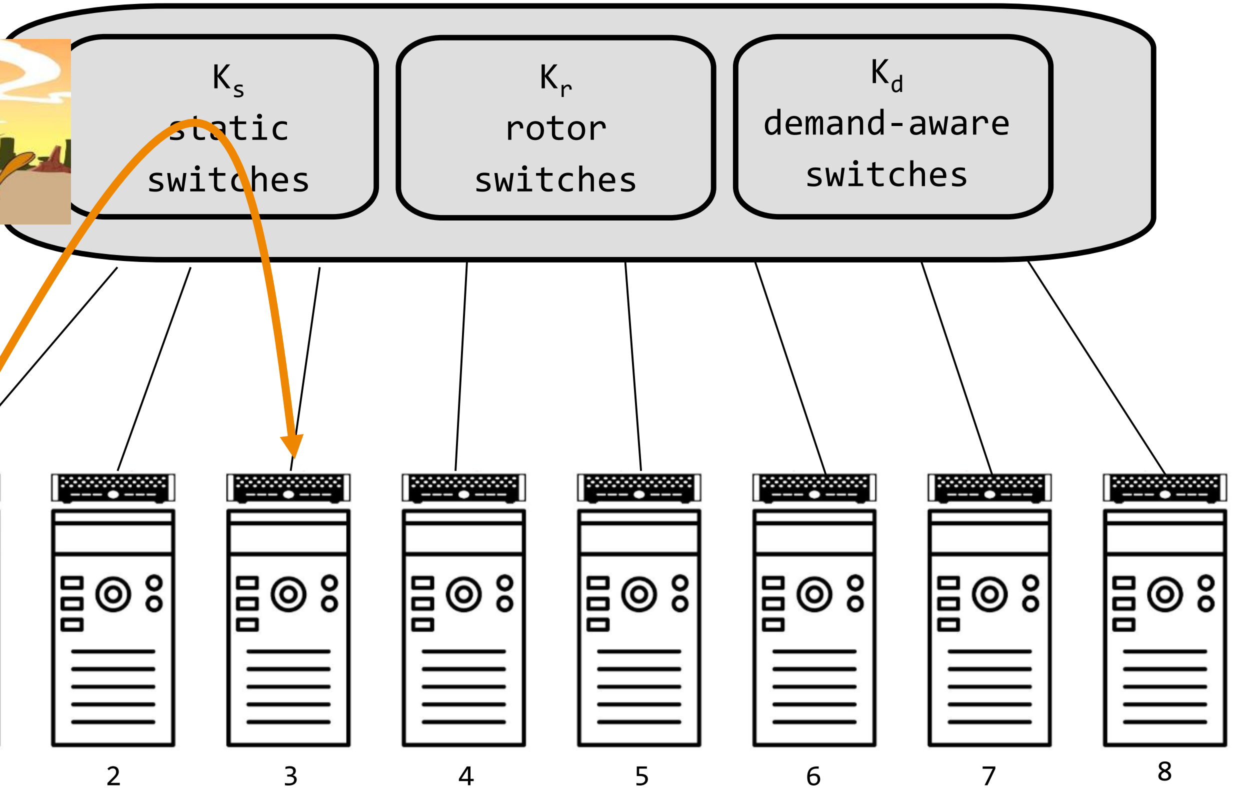


Cerberus



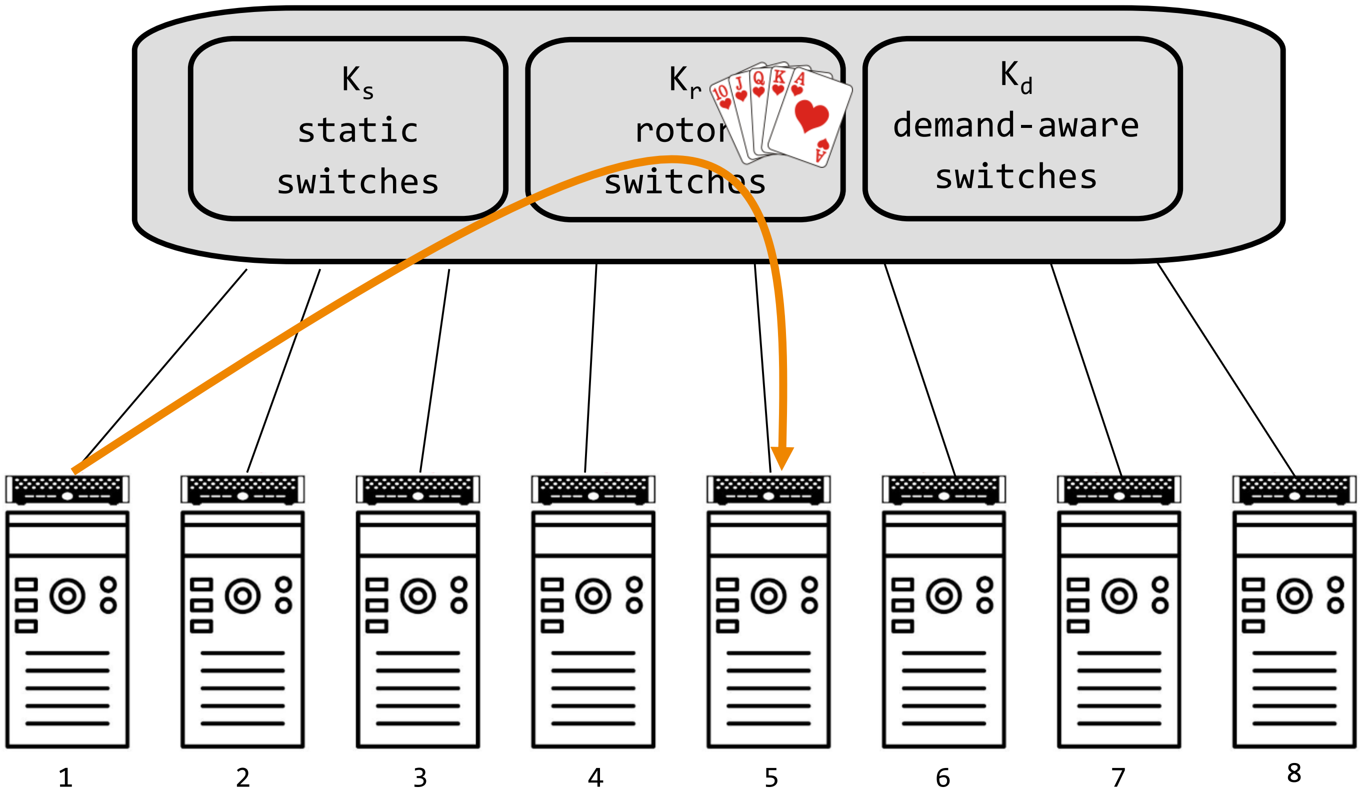
K_s
static
switches
K_r
rotor
switches
K_d
demand-aware
switches
1
2
3
4
5
6
7
8

Cerberus



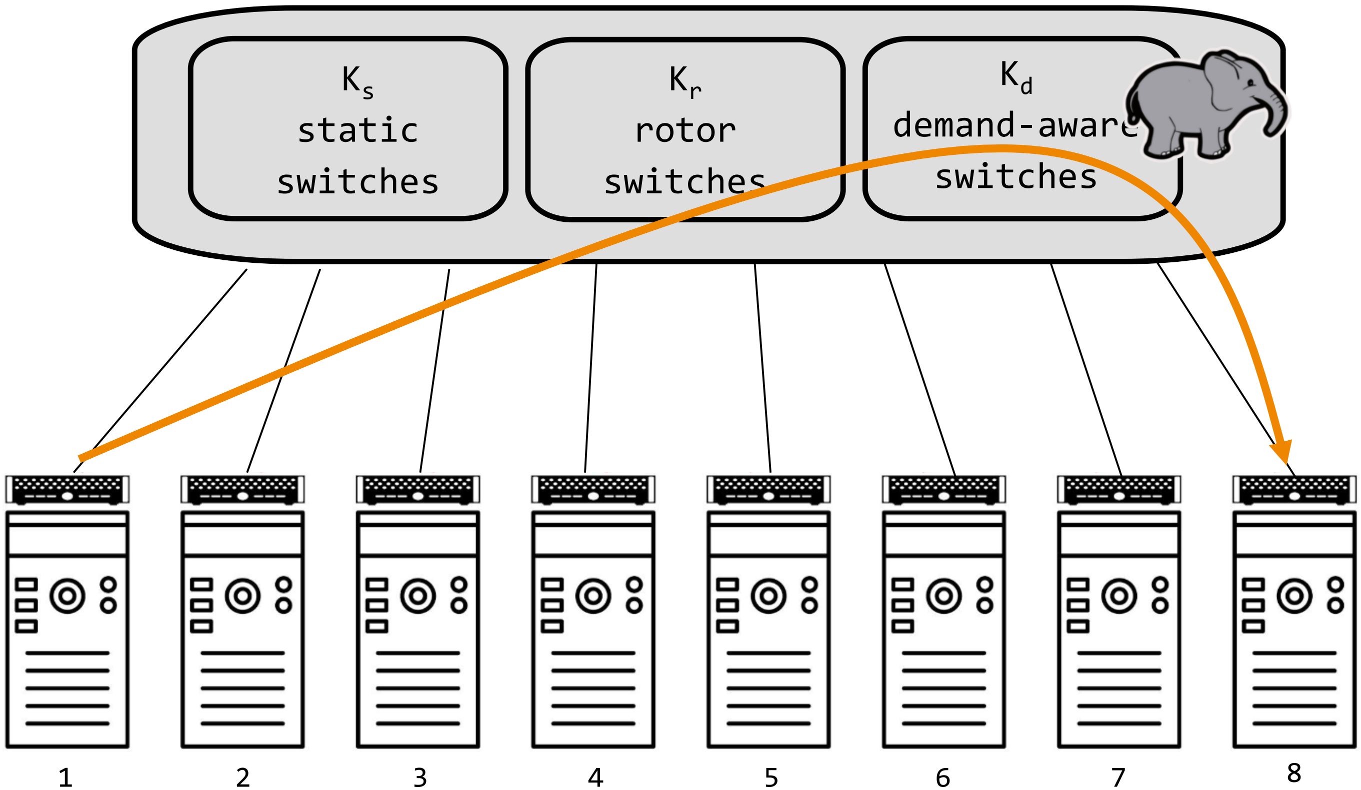
Scheduling: Small flows go via static switches…

Cerberus



Scheduling: … medium flows via rotor switches…

Cerberus



Scheduling: … and large flows via demand-aware switches
(if one available, otherwise via rotor).

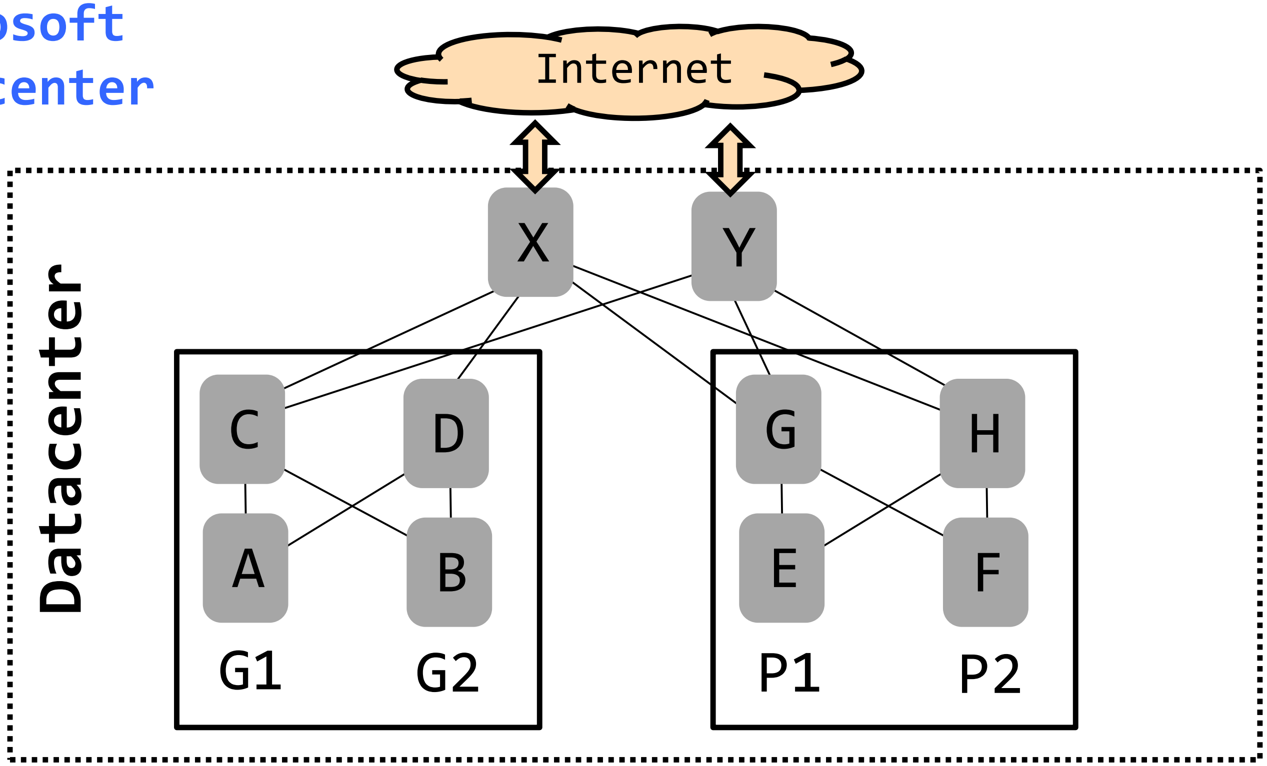
Roadmap

- Performance: Self-adjusting datacenter networks
- Modelling: How to model workloads, such as ML workloads?
- Dependability: Self-correcting MPLS networks
- More Use cases for self-driving networks

Challenge: Complexity

Especially Under Failures (Policy Compliance)

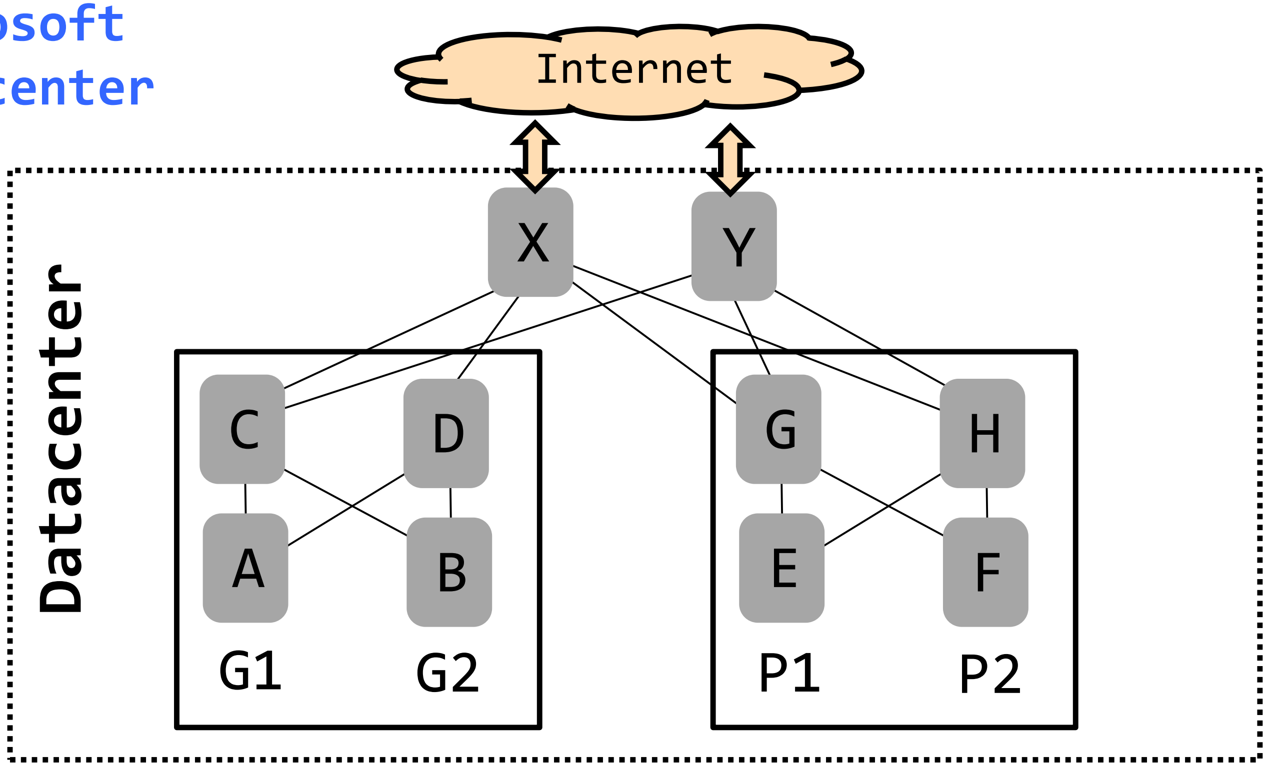
Example: BGP in **Microsoft datacenter**



Challenge: Complexity

Especially Under Failures (Policy Compliance)

Example: BGP in **Microsoft datacenter**



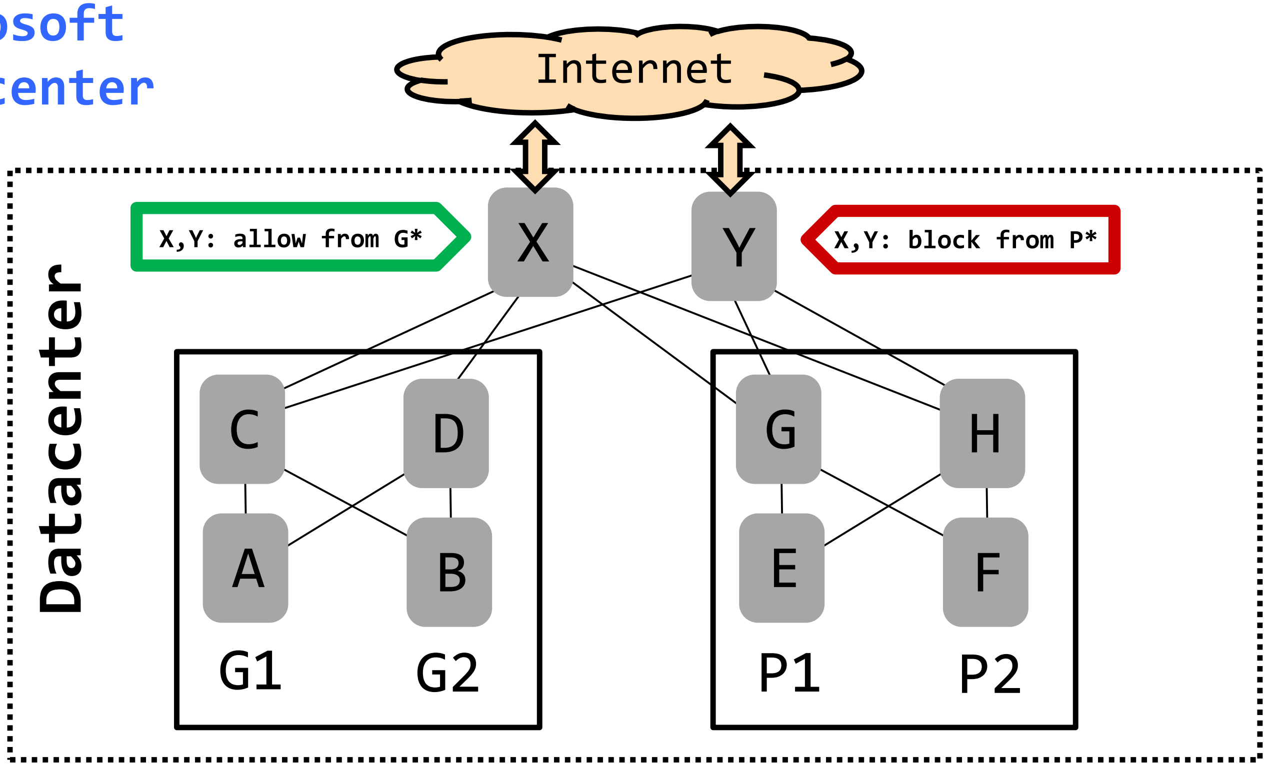
Cluster with globally reachable services

Cluster with internally accessible services

Challenge: Complexity

Especially Under Failures (Policy Compliance)

Example: BGP in **Microsoft datacenter**



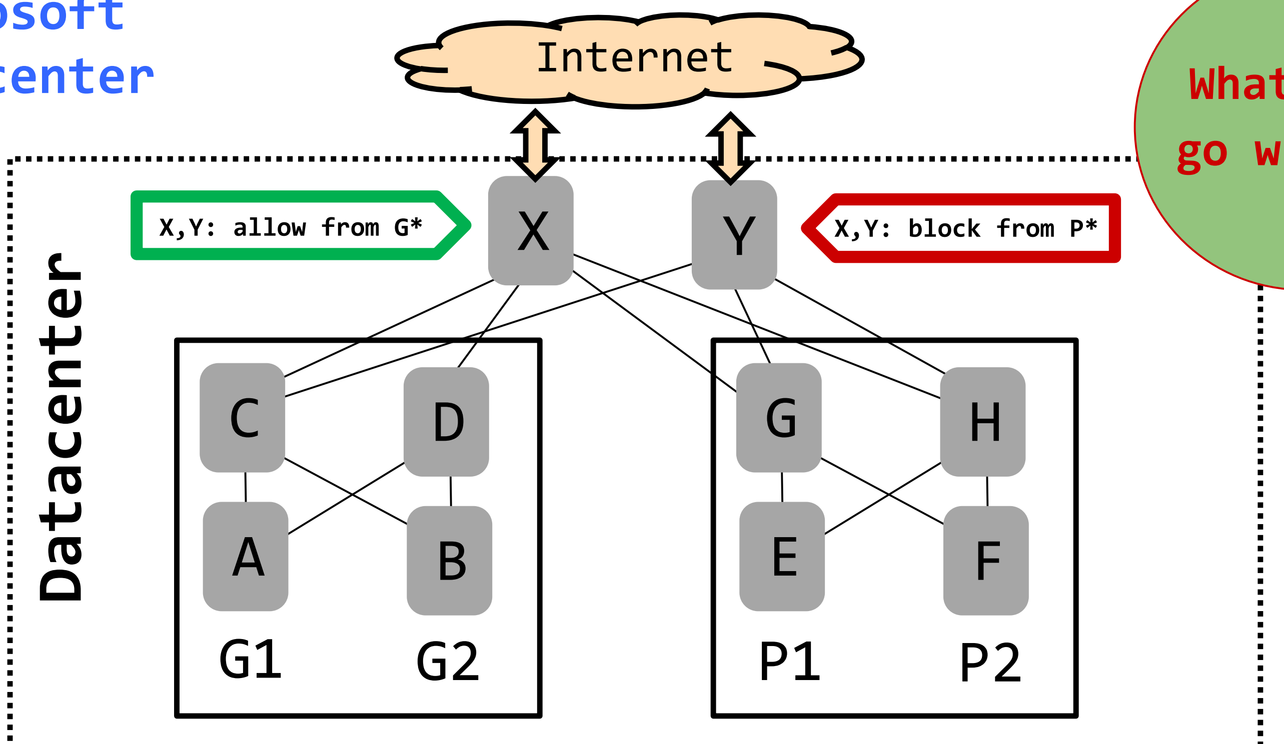
Cluster with globally reachable services

Cluster with internally accessible services

Challenge: Complexity

Especially Under Failures (Policy Compliance)

Example: BGP in **Microsoft datacenter**



What can go wrong?

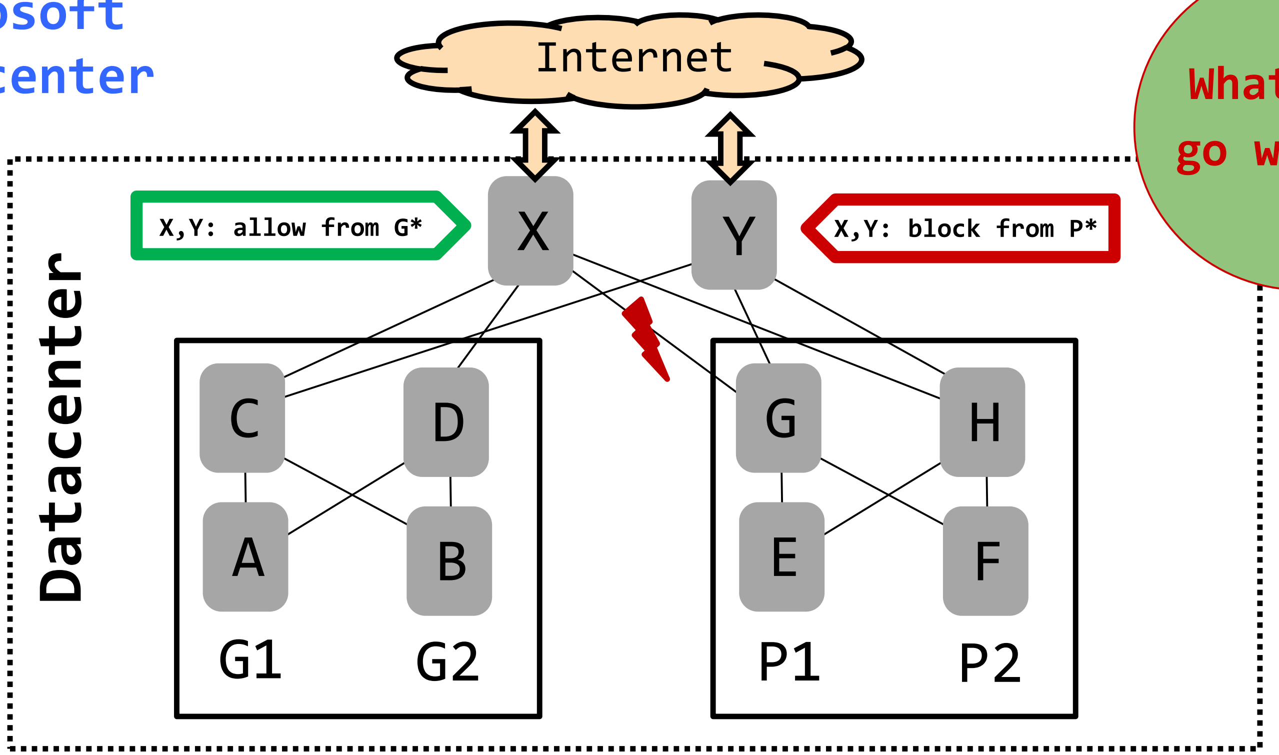
Cluster with globally reachable services

Cluster with internally accessible services

Challenge: Complexity

Especially Under Failures (Policy Compliance)

Example: BGP in **Microsoft datacenter**

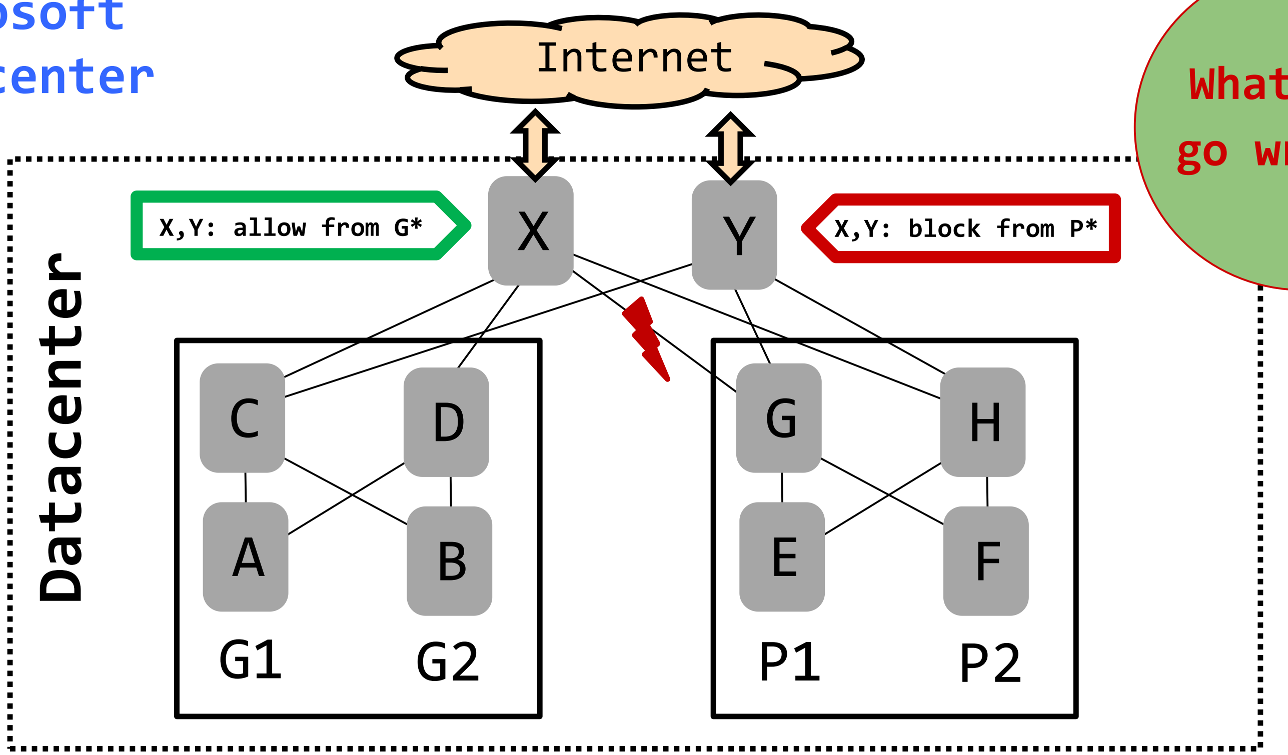


What can go wrong?

Challenge: Complexity

Especially Under Failures (Policy Compliance)

Example: BGP in **Microsoft datacenter**



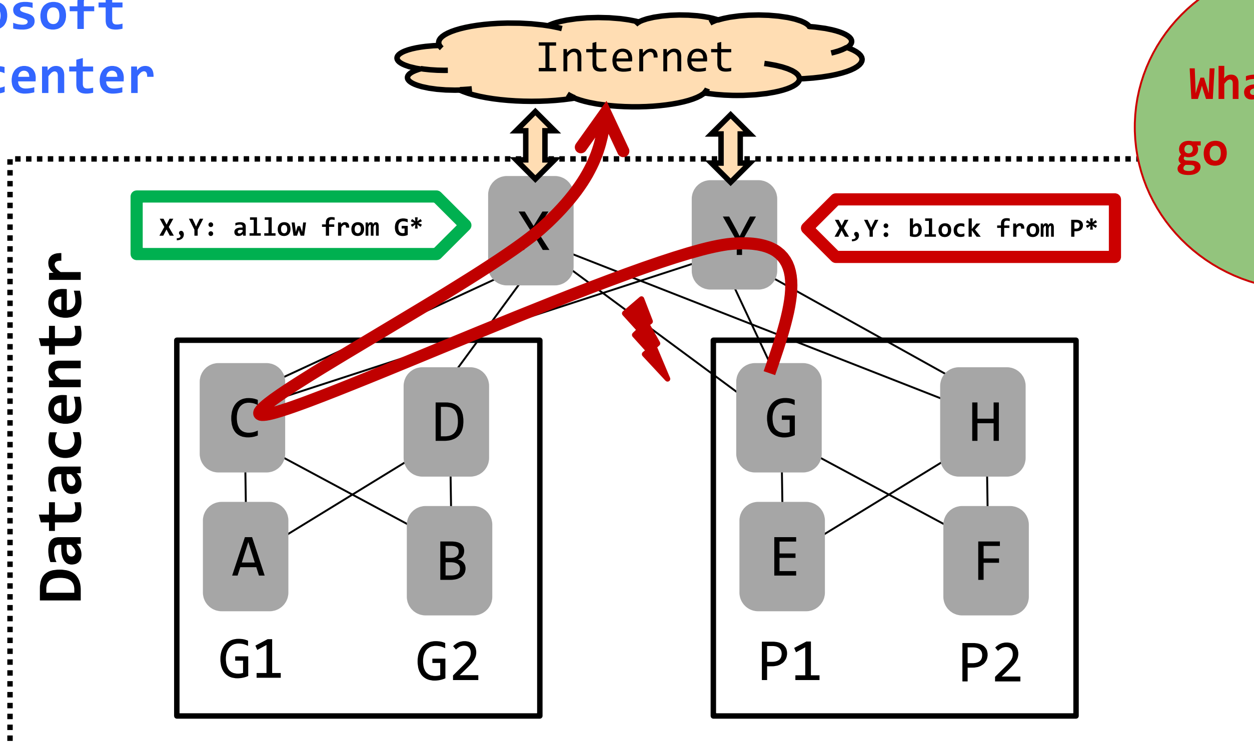
What can go wrong?

If link (G,X) fails and traffic from G is rerouted via Y and C to X:
X announces (does not block) G and H as it comes from C. (Note: BGP.)

Challenge: Complexity

Especially Under Failures (Policy Compliance)

Example: BGP in **Microsoft datacenter**



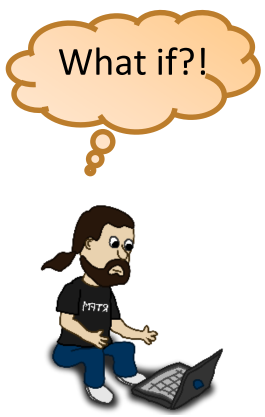
What can go wrong?

If link (G,X) fails and traffic from G is rerouted via Y and C to X:
X announces (does not block) G and H as it comes from C. (Note: BGP.)

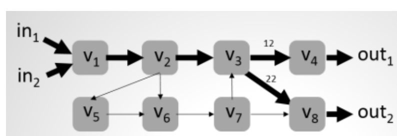
Dependable Networks with

Automated Whatif Analysis

⇢ Formal methods good for verifying networks! E.g., P-Rex for MPLS (Jensen et al. CoNEXT’19)

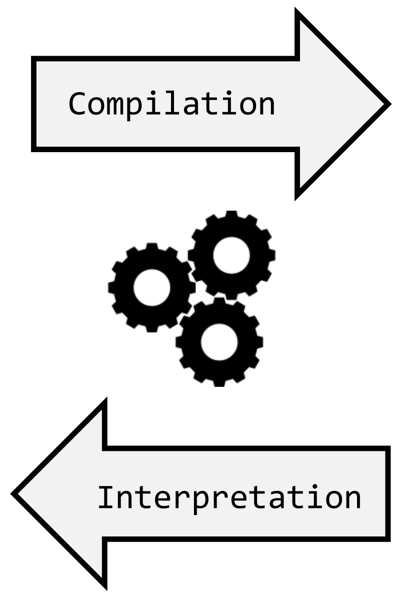


FT	In-I	In-Label	Out-I	op
τ_{v_1}	in_1	$\bot$	(v_1, v_2)	$push(10)$
	in_2	$\bot$	(v_1, v_2)	$push(20)$
τ_{v_2}	(v_1, v_2)	10	(v_2, v_3)	$swap(11)$
	(v_1, v_2)	20	(v_2, v_3)	$swap(21)$
τ_{v_3}	(v_2, v_3)	11	(v_3, v_4)	$swap(12)$
	(v_2, v_3)	21	(v_3, v_8)	$swap(22)$
	(v_7, v_3)	11	(v_3, v_4)	$swap(12)$
	(v_7, v_3)	21	(v_3, v_8)	$swap(22)$
τ_{v_4}	(v_3, v_4)	12	out_1	pop
τ_{v_5}	(v_2, v_5)	40	(v_5, v_6)	pop
τ_{v_6}	(v_2, v_6)	30	(v_6, v_7)	$swap(31)$
	(v_5, v_6)	30	(v_6, v_7)	$swap(31)$
	(v_5, v_6)	61	(v_6, v_7)	$swap(62)$
	(v_5, v_6)	71	(v_6, v_7)	$swap(72)$
τ_{v_7}	(v_6, v_7)	31	(v_7, v_3)	pop
	(v_6, v_7)	62	(v_7, v_3)	$swap(11)$
	(v_6, v_7)	72	(v_7, v_8)	$swap(22)$
τ_{v_8}	(v_3, v_8)	22	out_2	pop
	(v_7, v_8)	22	out_2	pop



local FFT	Out-I	In-Label	Out-I	op
τ_{v_2}	(v_2, v_3)	11	(v_2, v_6)	$push(30)$
	(v_2, v_3)	21	(v_2, v_6)	$push(30)$
	(v_2, v_6)	30	(v_2, v_5)	$push(40)$

global FFT	Out-I	In-Label	Out-I	op
τ'_{v_2}	(v_2, v_3)	11	(v_2, v_6)	$swap(61)$
	(v_2, v_3)	21	(v_2, v_6)	$swap(71)$
	(v_2, v_6)	61	(v_2, v_5)	$push(40)$
	(v_2, v_6)	71	(v_2, v_5)	$push(40)$



$$pX \Rightarrow qXX$$
$$pX \Rightarrow qYX$$
$$qY \Rightarrow rYY$$
$$rY \Rightarrow r$$
$$rX \Rightarrow pX$$

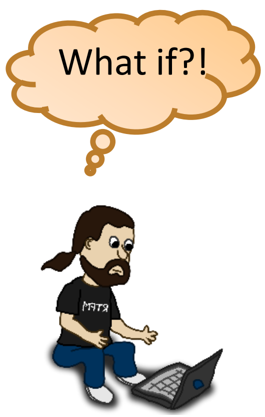
Router **configurations**
(Cisco, Juniper, etc.)

Formal language
which supports
automated analysis

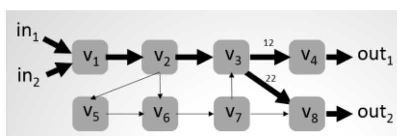
Dependable Networks with

Automated Whatif Analysis

⇢ Formal methods good for verifying networks! E.g., P-Rex for MPLS (Jensen et al. CoNEXT'19)

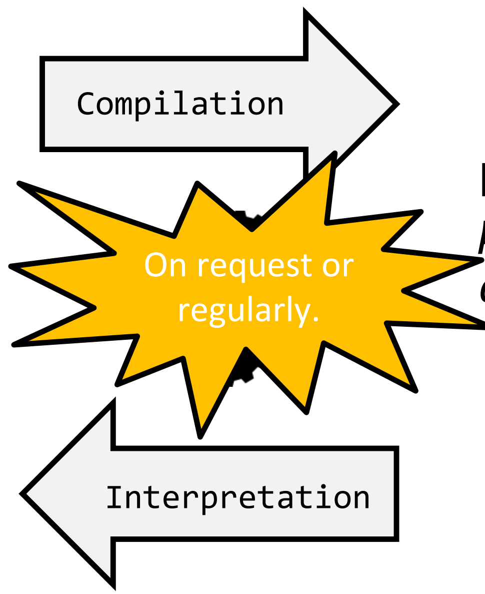


FT	In-I	In-Label	Out-I	op
τ_{v_1}	in_1	$\perp$	(v_1, v_2)	$push(10)$
	in_2	$\perp$	(v_1, v_2)	$push(20)$
τ_{v_2}	(v_1, v_2)	10	(v_2, v_3)	$swap(11)$
	(v_1, v_2)	20	(v_2, v_3)	$swap(21)$
τ_{v_3}	(v_2, v_3)	11	(v_3, v_4)	$swap(12)$
	(v_2, v_3)	21	(v_3, v_8)	$swap(22)$
	(v_7, v_3)	11	(v_3, v_4)	$swap(12)$
	(v_7, v_3)	21	(v_3, v_8)	$swap(22)$
τ_{v_4}	(v_3, v_4)	12	out_1	pop
τ_{v_5}	(v_2, v_5)	40	(v_5, v_6)	pop
τ_{v_6}	(v_2, v_6)	30	(v_6, v_7)	$swap(31)$
	(v_5, v_6)	30	(v_6, v_7)	$swap(31)$
	(v_5, v_6)	61	(v_6, v_7)	$swap(62)$
	(v_5, v_6)	71	(v_6, v_7)	$swap(72)$
τ_{v_7}	(v_6, v_7)	31	(v_7, v_3)	pop
	(v_6, v_7)	62	(v_7, v_3)	$swap(11)$
	(v_6, v_7)	72	(v_7, v_8)	$swap(22)$
τ_{v_8}	(v_3, v_8)	22	out_2	pop
	(v_7, v_8)	22	out_2	pop



local FFT	Out-I	In-Label	Out-I	op
τ_{v_2}	(v_2, v_3)	11	(v_2, v_6)	$push(30)$
	(v_2, v_3)	21	(v_2, v_6)	$push(30)$
	(v_2, v_6)	30	(v_2, v_5)	$push(40)$

global FFT	Out-I	In-Label	Out-I	op
τ'_{v_2}	(v_2, v_3)	11	(v_2, v_6)	$swap(61)$
	(v_2, v_3)	21	(v_2, v_6)	$swap(71)$
	(v_2, v_6)	61	(v_2, v_5)	$push(40)$
	(v_2, v_6)	71	(v_2, v_5)	$push(40)$



$$pX \Rightarrow qXX$$
$$pX \Rightarrow qYX$$
$$qY \Rightarrow rYY$$
$$rY \Rightarrow r$$
$$rX \Rightarrow pX$$

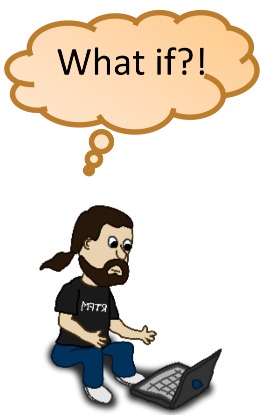
Router **configurations**
(Cisco, Juniper, etc.)

Formal language
which supports
automated analysis

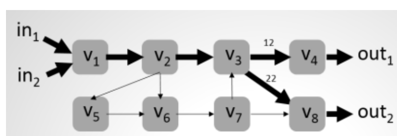
Dependable Networks with

Automated Whatif Analysis

⇢ Formal methods good for verifying networks! E.g., P-Rex for MPLS (Jensen et al. CoNEXT’19)

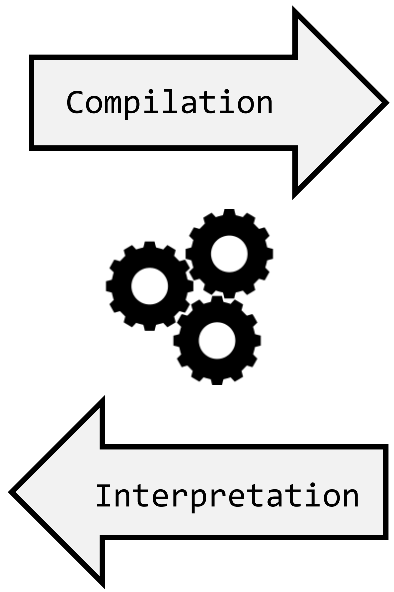


FT	In-I	In-Label	Out-I	op
τ_{v_1}	in_1	$\bot$	(v_1, v_2)	$push(10)$
	in_2	$\bot$	(v_1, v_2)	$push(20)$
τ_{v_2}	(v_1, v_2)	10	(v_2, v_3)	$swap(11)$
	(v_1, v_2)	20	(v_2, v_3)	$swap(21)$
τ_{v_3}	(v_2, v_3)	11	(v_3, v_4)	$swap(12)$
	(v_2, v_3)	21	(v_3, v_8)	$swap(22)$
	(v_7, v_3)	11	(v_3, v_4)	$swap(12)$
	(v_7, v_3)	21	(v_3, v_8)	$swap(22)$
τ_{v_4}	(v_3, v_4)	12	out_1	pop
τ_{v_5}	(v_2, v_5)	40	(v_5, v_6)	pop
τ_{v_6}	(v_2, v_6)	30	(v_6, v_7)	$swap(31)$
	(v_5, v_6)	30	(v_6, v_7)	$swap(31)$
	(v_5, v_6)	61	(v_6, v_7)	$swap(62)$
	(v_5, v_6)	71	(v_6, v_7)	$swap(72)$
τ_{v_7}	(v_6, v_7)	31	(v_7, v_3)	pop
	(v_6, v_7)	62	(v_7, v_3)	$swap(11)$
	(v_6, v_7)	72	(v_7, v_8)	$swap(22)$
τ_{v_8}	(v_3, v_8)	22	out_2	pop
	(v_7, v_8)	22	out_2	pop



local FFT	Out-I	In-Label	Out-I	op
τ_{v_2}	(v_2, v_3)	11	(v_2, v_6)	$push(30)$
	(v_2, v_3)	21	(v_2, v_6)	$push(30)$
	(v_2, v_6)	30	(v_2, v_5)	$push(40)$
global FFT	**Out-I**	**In-Label**	**Out-I**	**op**
τ'_{v_2}	(v_2, v_3)	11	(v_2, v_6)	$swap(61)$
	(v_2, v_3)	21	(v_2, v_6)	$swap(71)$
	(v_2, v_6)	61	(v_2, v_5)	$push(40)$
	(v_2, v_6)	71	(v_2, v_5)	$push(40)$

Router configurations
(Cisco, Juniper, etc.)

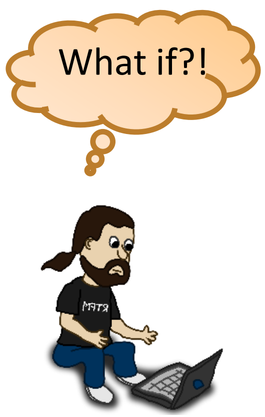


Many alternatives: ***automata*** theory, binary decision diagrams (***BDDs***), ***games*** (e.g., Stackelberg, Petri nets), ***SMTs***, ***ILPs*** …

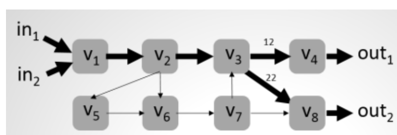
Even more automation:

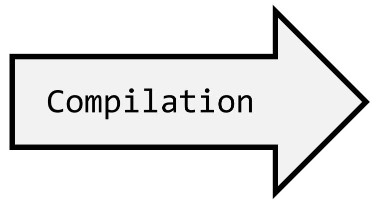
Synthesis

⇢ Formal methods good for verifying networks! E.g., P-Rex for MPLS (Jensen et al. CoNEXT’19)



FT	In-I	In-Label	Out-I	op
τ_{v_1}	in_1	$\bot$	(v_1, v_2)	$push(10)$
	in_2	$\bot$	(v_1, v_2)	$push(20)$
τ_{v_2}	(v_1, v_2)	10	(v_2, v_3)	$swap(11)$
	(v_1, v_2)	20	(v_2, v_3)	$swap(21)$
τ_{v_3}	(v_2, v_3)	11	(v_3, v_4)	$swap(12)$
	(v_2, v_3)	21	(v_3, v_8)	$swap(22)$
	(v_7, v_3)	11	(v_3, v_4)	$swap(12)$
	(v_7, v_3)	21	(v_3, v_8)	$swap(22)$
τ_{v_4}	(v_3, v_4)	12	out_1	pop
τ_{v_5}	(v_2, v_5)	40	(v_5, v_6)	pop
τ_{v_6}	(v_2, v_6)	30	(v_6, v_7)	$swap(31)$
	(v_5, v_6)	30	(v_6, v_7)	$swap(31)$
	(v_5, v_6)	61	(v_6, v_7)	$swap(62)$
	(v_5, v_6)	71	(v_6, v_7)	$swap(72)$
τ_{v_7}	(v_6, v_7)	31	(v_7, v_3)	pop
	(v_6, v_7)	62	(v_7, v_3)	$swap(11)$
	(v_6, v_7)	72	(v_7, v_8)	$swap(22)$
τ_{v_8}	(v_3, v_8)	22	out_2	pop
	(v_7, v_8)	22	out_2	pop





Where configuration not compliant?

local FFT	Out-I	In-Label	Out-I	op
τ_{v_2}	(v_2, v_3)	11	(v_2, v_6)	$push(30)$
	(v_2, v_3)	21	(v_2, v_6)	$push(30)$
	(v_2, v_6)	30	(v_2, v_5)	$push(40)$

global FFT	Out-I	In-Label	Out-I	op
τ'_{v_2}	(v_2, v_3)	11	(v_2, v_6)	$swap(61)$
	(v_2, v_3)	21	(v_2, v_6)	$swap(71)$
	(v_2, v_6)	61	(v_2, v_5)	$push(40)$
	(v_2, v_6)	71	(v_2, v_5)	$push(40)$

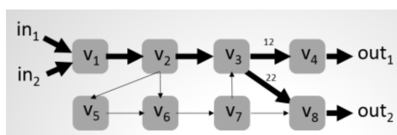
Router configurations
(Cisco, Juniper, etc.)

Even more automation:

Synthesis

- Formal methods good for verifying networks! E.g., P-Rex for MPLS (Jensen et al. CoNEXT’19)

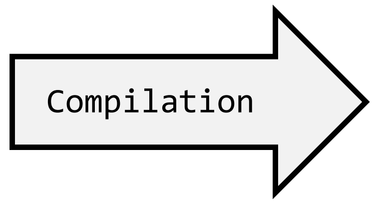
FT	In-I	In-Label	Out-I	op
τ_{v_1}	in_1	$\bot$	(v_1, v_2)	$push(10)$
	in_2	$\bot$	(v_1, v_2)	$push(20)$
τ_{v_2}	(v_1, v_2)	10	(v_2, v_3)	$swap(11)$
	(v_1, v_2)	20	(v_2, v_3)	$swap(21)$
τ_{v_3}	(v_2, v_3)	11	(v_3, v_4)	$swap(12)$
	(v_2, v_3)	21	(v_3, v_8)	$swap(22)$
	(v_7, v_3)	11	(v_3, v_4)	$swap(12)$
	(v_7, v_3)	21	(v_3, v_8)	$swap(22)$
τ_{v_4}	(v_3, v_4)	12	out_1	pop
τ_{v_5}	(v_2, v_5)	40	(v_5, v_6)	pop
τ_{v_6}	(v_2, v_6)	30	(v_6, v_7)	$swap(31)$
	(v_5, v_6)	30	(v_6, v_7)	$swap(31)$
	(v_5, v_6)	61	(v_6, v_7)	$swap(62)$
	(v_5, v_6)	71	(v_6, v_7)	$swap(72)$
τ_{v_7}	(v_6, v_7)	31	(v_7, v_3)	pop
	(v_6, v_7)	62	(v_7, v_3)	$swap(11)$
	(v_6, v_7)	72	(v_7, v_8)	$swap(22)$
τ_{v_8}	(v_3, v_8)	22	out_2	pop
	(v_7, v_8)	22	out_2	pop



local FFT	Out-I	In-Label	Out-I	op
τ_{v_2}	(v_2, v_3)	11	(v_2, v_6)	$push(30)$
	(v_2, v_3)	21	(v_2, v_6)	$push(30)$
	(v_2, v_6)	30	(v_2, v_5)	$push(40)$

global FFT	Out-I	In-Label	Out-I	op
τ'_{v_2}	(v_2, v_3)	11	(v_2, v_6)	$swap(61)$
	(v_2, v_3)	21	(v_2, v_6)	$swap(71)$
	(v_2, v_6)	61	(v_2, v_5)	$push(40)$
	(v_2, v_6)	71	(v_2, v_5)	$push(40)$

Router configurations
(Cisco, Juniper, etc.)



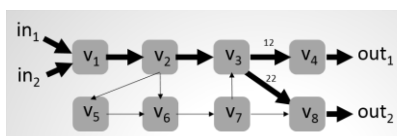
Where configuration not compliant?

Even more automation:

Synthesis

- Formal methods good for verifying networks! E.g., P-Rex for MPLS (Jensen et al. CoNEXT’19)

FT	In-I	In-Label	Out-I	op
τ_{v_1}	in_1	$\bot$	(v_1, v_2)	$push(10)$
	in_2	$\bot$	(v_1, v_2)	$push(20)$
τ_{v_2}	(v_1, v_2)	10	(v_2, v_3)	$swap(11)$
	(v_1, v_2)	20	(v_2, v_3)	$swap(21)$
τ_{v_3}	(v_2, v_3)	11	(v_3, v_4)	$swap(12)$
	(v_2, v_3)	21	(v_3, v_8)	$swap(22)$
	(v_7, v_3)	11	(v_3, v_4)	$swap(12)$
	(v_7, v_3)	21	(v_3, v_8)	$swap(22)$
τ_{v_4}	(v_3, v_4)	12	out_1	pop
τ_{v_5}	(v_2, v_5)	40	(v_5, v_6)	pop
τ_{v_6}	(v_2, v_6)	30	(v_6, v_7)	$swap(31)$
	(v_5, v_6)	30	(v_6, v_7)	$swap(31)$
	(v_5, v_6)	61	(v_6, v_7)	$swap(62)$
	(v_5, v_6)	71	(v_6, v_7)	$swap(72)$
τ_{v_7}	(v_6, v_7)	31	(v_7, v_3)	pop
	(v_6, v_7)	62	(v_7, v_3)	$swap(11)$
	(v_6, v_7)	72	(v_7, v_8)	$swap(22)$
τ_{v_8}	(v_3, v_8)	22	out_2	pop
	(v_7, v_8)	22	out_2	pop



local FFT	Out-I	In-Label	Out-I	op
τ_{v_2}	(v_2, v_3)	11	(v_2, v_6)	$push(30)$
	(v_2, v_3)	21	(v_2, v_6)	$push(30)$
	(v_2, v_6)	30	(v_2, v_5)	$push(40)$

global FFT	Out-I	In-Label	Out-I	op
τ'_{v_2}	(v_2, v_3)	11	(v_2, v_6)	$swap(61)$
	(v_2, v_3)	21	(v_2, v_6)	$swap(71)$
	(v_2, v_6)	61	(v_2, v_5)	$push(40)$
	(v_2, v_6)	71	(v_2, v_5)	$push(40)$

Router configurations
(Cisco, Juniper, etc.)

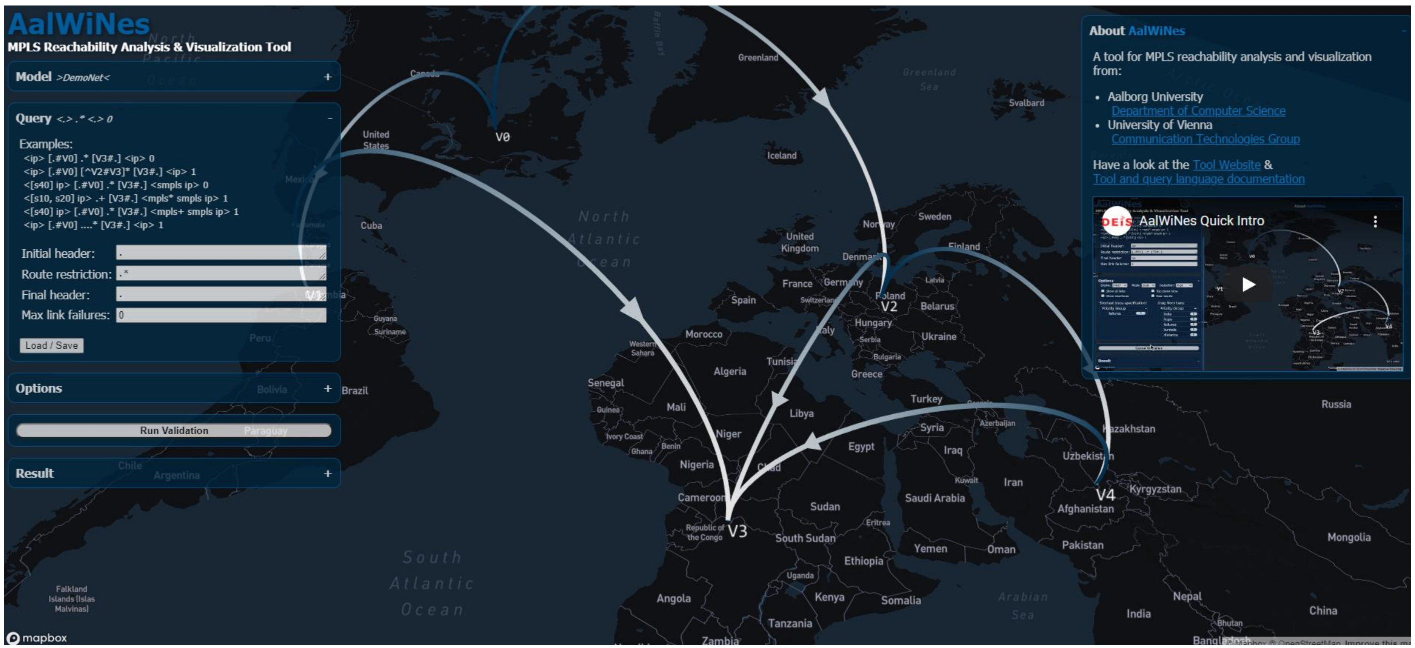
Compilation

Where configuration not compliant?

Synthesis!

Literature: P-Rex: Fast Verification of MPLS Networks with Multiple Link Failures. Jensen et al. ACM CoNEXT, 2018.

P-Rex / AalWiNes Tool



Tool: https://demo.aalwines.cs.aau.dk/
Youtube: https://www.youtube.com/watch?v=mvXAn9i7_Q0

Efficient Synthesis?

ML+FM!

- Formal ***synthesis slower*** than verificatio
- An opportunity for using ML!
- ***Ideally ML+FM***: guarantees from formal methods, performance from ML
- For example: synthesize with ML then verify with formal methods
- Examples: DeepMPLS, DeepBGP, …

Roadmap

- Performance: Self-adjusting datacenter networks
- Modelling: How to model workloads, such as ML workloads?
- Dependability: Self-correcting MPLS networks
- More Use cases for self-driving networks

Great Opportunities

- ⇢ Self-driving switches
- ⇢ Self-driving congestion control
- ⇢ Let's discuss! ☺

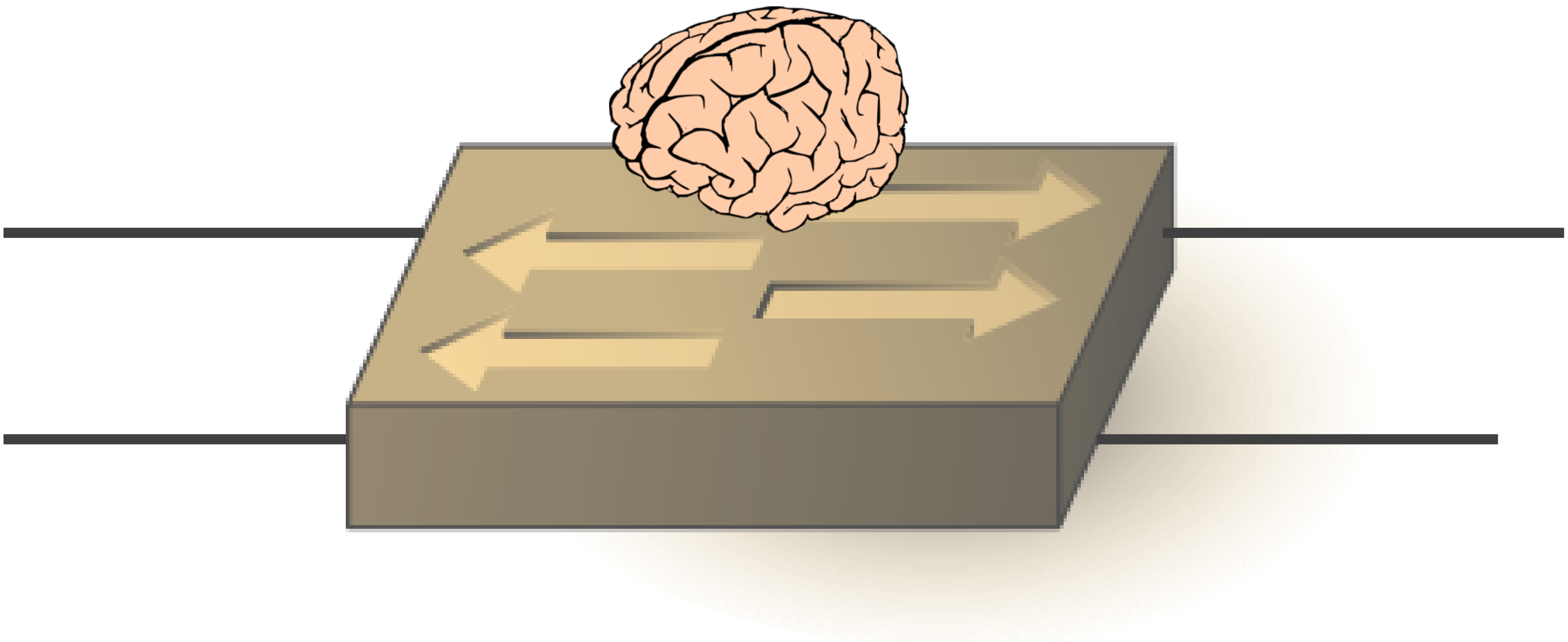
Addanki et al. (NSDI 2024)

Smart Switches

Addanki et al. (NSDI 2024)

Smart Switches

⇢ What if switches become smart?



Scenario 1

⇢ What if switches become smart? Assume: shared memory size 3.

Scenario 1

⇢ What if switches become smart? Assume: shared memory size 3.

Scenario 1

- What if switches become smart? Assume: shared memory size 3.

Scenario 1

⇢ What if switches become smart? Assume: shared memory size 3.

⇢ Suboptimal: green packets could be transmitted in parallel, but there is no more space! (Output rate 1 vs 2!)

Scenario 1

⇢ What if switches become smart? Assume: shared memory size 3.

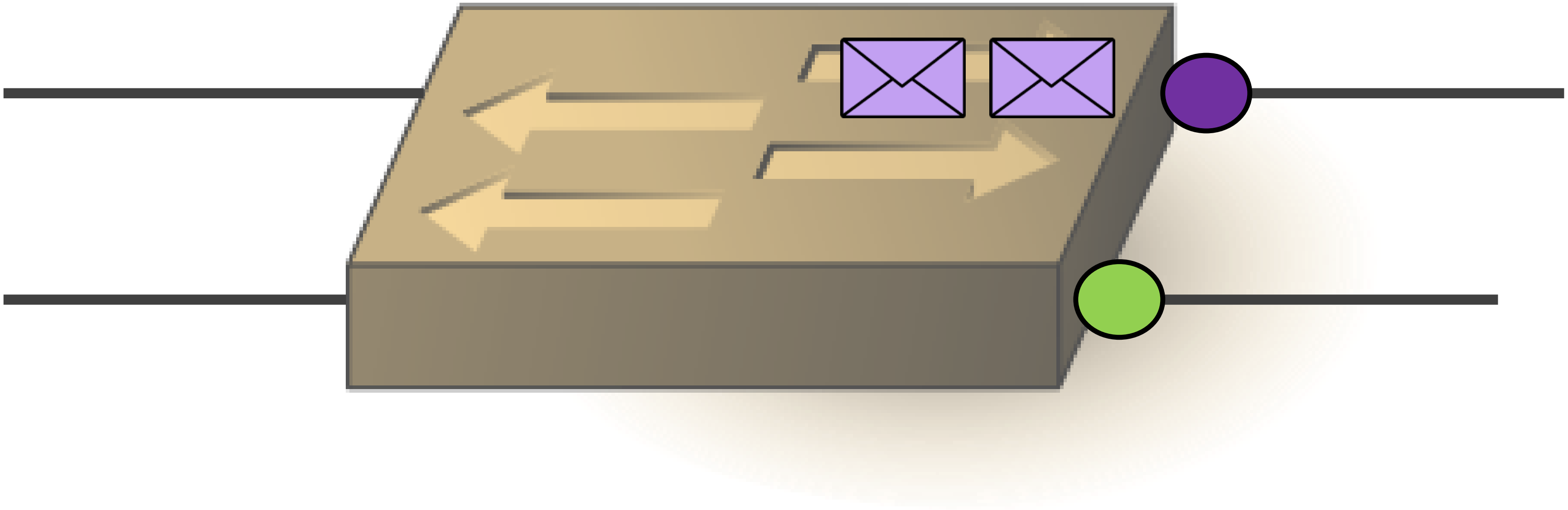
⇢ Suboptimal: green packets could be transmitted in parallel, but there is no more space! (Output rate 1 vs 2!)

Scenario 2

- What if switches become smart? Assume: shared memory size 3.

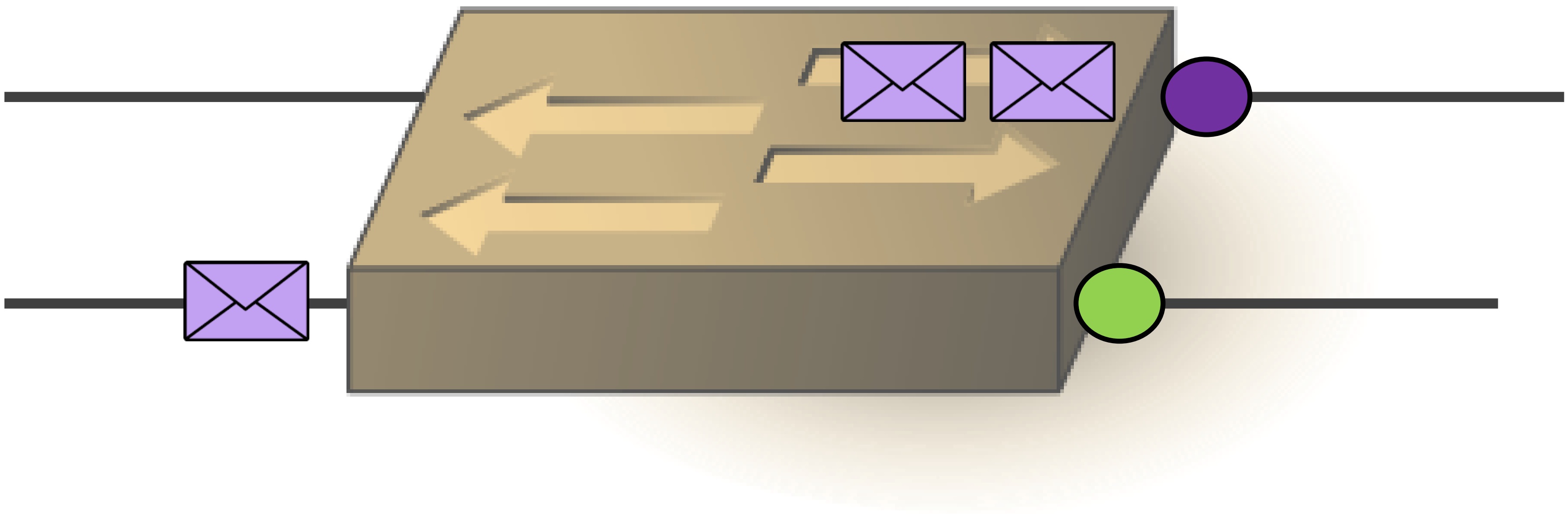
Scenario 2

⇢ What if switches become smart? Assume: shared memory size 3.



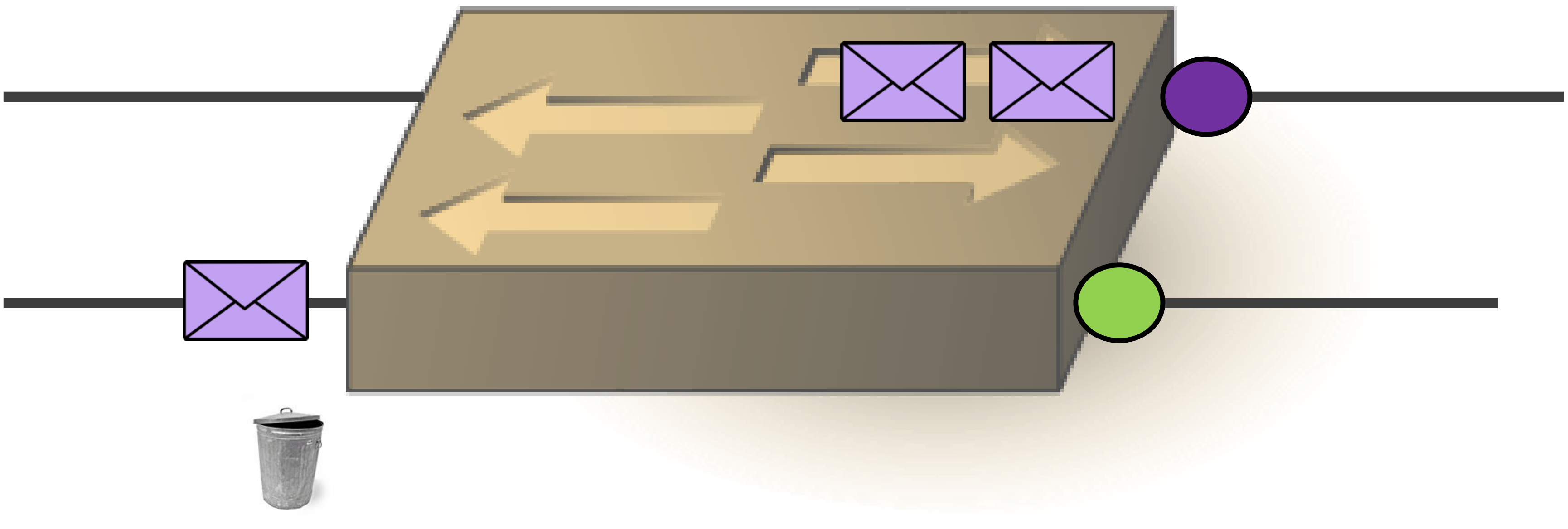
Scenario 2

⇢ What if switches become smart? Assume: shared memory size 3.



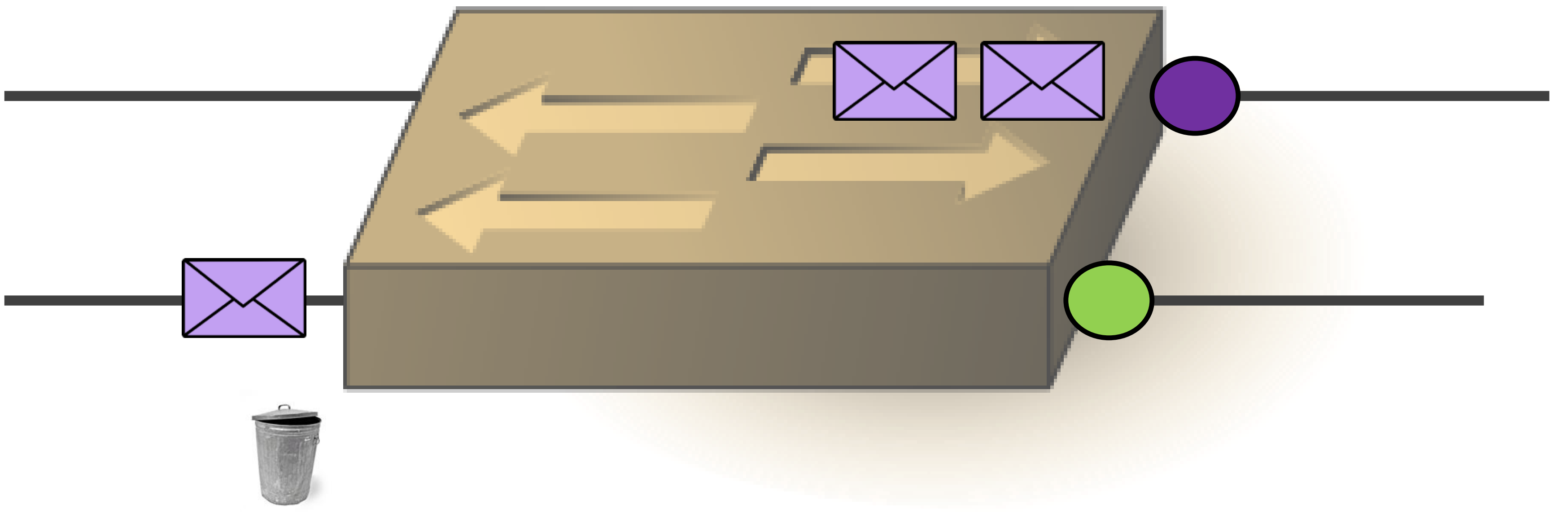
Scenario 2

⇢ What if switches become smart? Assume: shared memory size 3.



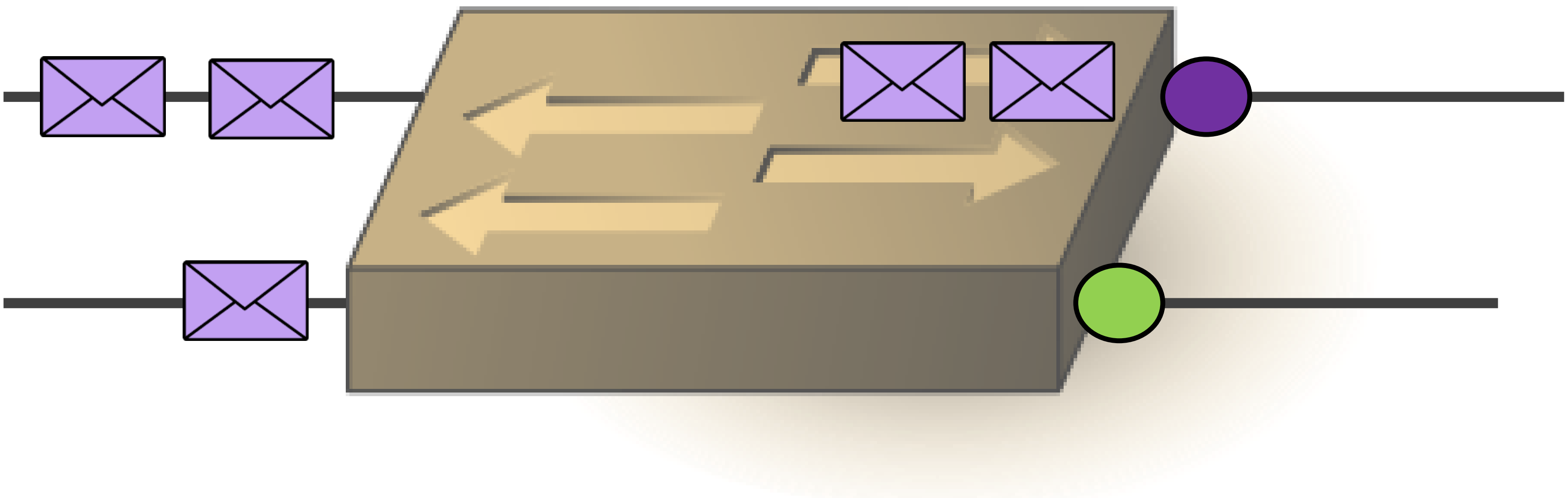
Scenario 2

⇢ What if switches become smart? Assume: shared memory size 3.



Scenario 2

- What if switches become smart? Assume: shared memory size 3.



- Suboptimal: drop to leave space but no space needed!

Addanki et al. (NSDI 2024)

Credence

⇢ Traffic at switch can be ***predicted*** fairly well

⇢ AI/ML could significantly ***improve buffer management***…

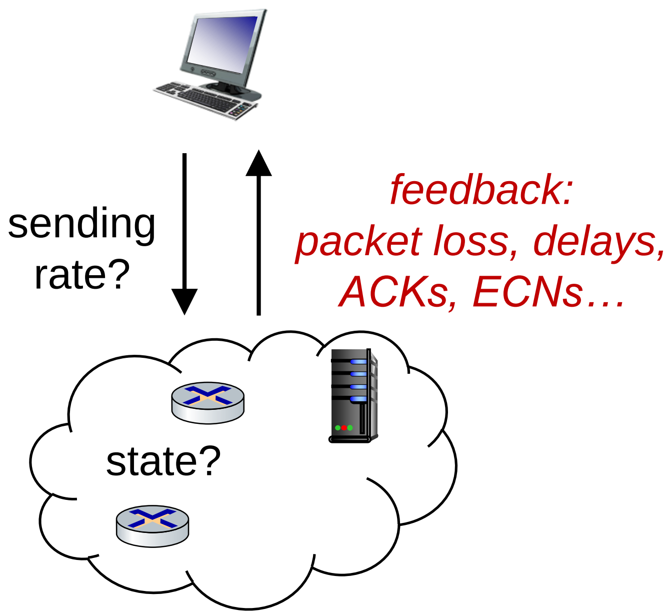
⇢ … and hence ***admission control and throughput***!

Further reading:

Credence: Augmenting Datacenter Switch Buffer Sharing with ML Predictions
Vamsi Addanki, Maciej Pacut, and Stefan Schmid.
21st USENIX Symposium on Networked Systems Design and Implementation (**NSDI**), 2024.

Congestion Control

- One of the big success stories stories of the Internet!
- Thanks to Internet protocol TCP: no congestion collapse since 1990s
- Same mechanism since 30+ years, while traffic increased by factor 1 billion!
- Still much innovation (and research, e.g., on fairness) Google's BBR, QUIC, ECN, etc.



Modeling BBR

Model-Based Insights on the Performance, Fairness, and Stability of BBR

Simon Scherrer
simon.scherrer@inf.ethz.ch
ETH Zurich
Switzerland

Markus Legner
markus.legner@inf.ethz.ch
ETH Zurich
Switzerland

Adrian Perrig
adrian.perrig@inf.ethz.ch
ETH Zurich
Switzerland

Stefan Schmid
stefan_schmid@univie.ac.at
TU Berlin/Uni. Vienna
Germany/Austria

ABSTRACT

Google's BBR is the most prominent result of the recently revived quest for efficient, fair, and flexible congestion-control algorithms (CCAs). While the performance of BBR has been investigated by numerous studies, previous work still leaves gaps in the understanding of BBR performance: Experiment-based studies generally only consider network settings that researchers can set up with manageable effort, and model-based studies neglect important issues like convergence.

To complement previous BBR analyses, this paper presents a fluid model of BBRv1 and BBRv2, allowing both efficient simulation under a wide variety of network settings and an-

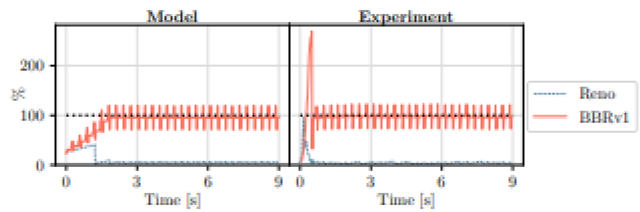


Figure 1: Competition of sending rates (in % of link bandwidth) between a Reno flow and a BBRv1 flow, according to our fluid model and experiment data.

however, a deep theoretical understanding of BBR also requires a model that is valid for general settings and allows

Summary

- Opportunity: *adaptable networks* and *structure* in demand
- Opportunity: *AI/ML* for performance and *formal methods* for dependability
- Enables *self-driving networks*
- Requires: models and automated, computer-driven designs
- Great research opportunities ahead!

Some References

On the Complexity of Traffic Traces and Implications
Chen Avin, Manya Ghobadi, Chen Griner, and Stefan Schmid.
ACM **SIGMETRICS** and ACM Performance Evaluation Review (**PER**), Boston, Massachusetts, USA, June 2020.

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

Cerberus: The Power of Choices in Datacenter Topology Design (A Throughput Perspective)
Chen Griner, Johannes Zerwas, Andreas Blenk, Manya Ghobadi, Stefan Schmid, and Chen Avin.
ACM **SIGMETRICS** and ACM Performance Evaluation Review (**PER**), Mumbai, India, June 2022.

Cerberus: The Power of Choices in Datacenter Topology Design (A Throughput Perspective)
Chen Griner, Johannes Zerwas, Andreas Blenk, Manya Ghobadi, Stefan Schmid, and Chen Avin.
ACM **SIGMETRICS** and ACM Performance Evaluation Review (**PER**), Mumbai, India, June 2022.

AalWiNes: A Fast and Quantitative What-If Analysis Tool for MPLS Networks
Peter Gjøl Jensen, Morten Konggaard, Dan Kristiansen, Stefan Schmid, Bernhard Clemens Schrenk, and Jiri Srba.
16th ACM International Conference on emerging Networking EXperiments and Technologies (**CoNEXT**), Barcelona, Spain, December 2020.

Latte: Improving the Latency of Transiently Consistent Network Update Schedules
Mark Glavind, Niels Christensen, Jiri Srba, and Stefan Schmid.
38th International Symposium on Computer Performance, Modeling, Measurements and Evaluation (**PERFORMANCE**) and ACM Performance Evaluation Review (**PER**), Milan, Italy, November 2020.

Model-Based Insights on the Performance, Fairness, and Stability of BBR (IRTF Applied Networking Research Prize)
Simon Scherrer, Markus Legner, Adrian Perrig, and Stefan Schmid.
ACM Internet Measurement Conference (**IMC**), Nice, France, October 2022.

Credence: Augmenting Datacenter Switch Buffer Sharing with ML Predictions
Vamsi Addanki, Maciej Pacut, and Stefan Schmid.
21st USENIX Symposium on Networked Systems Design and Implementation (**NSDI**), Santa Clara, California, USA, April 2024.