Revolutionizing Datacenter Networks via Reconfigurable Topologies

Stefan Schmid (TU Berlin)

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

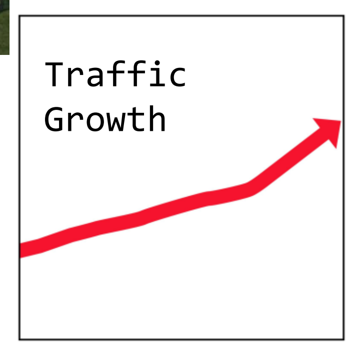
Acknowledgements:

Trend

Data-Centric Applications

Datacenters ("hyper-scale")

+network



Source: Facebook

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

Trend

Data-Centric Applications

Datacenters (“hyper-scale”)

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

Credits: Marco Chiesa

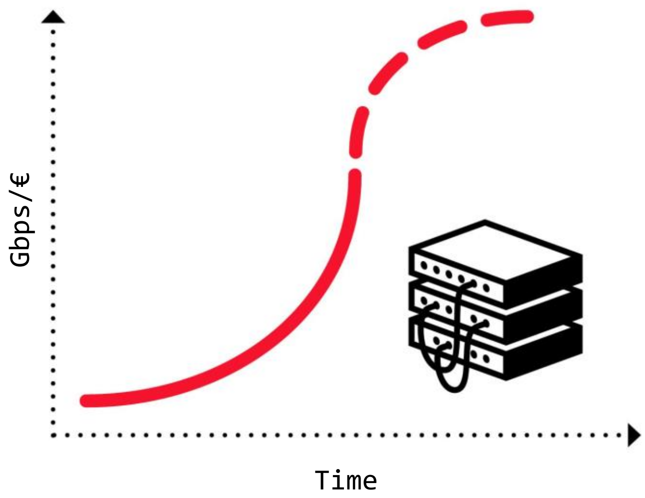
The Problem

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**



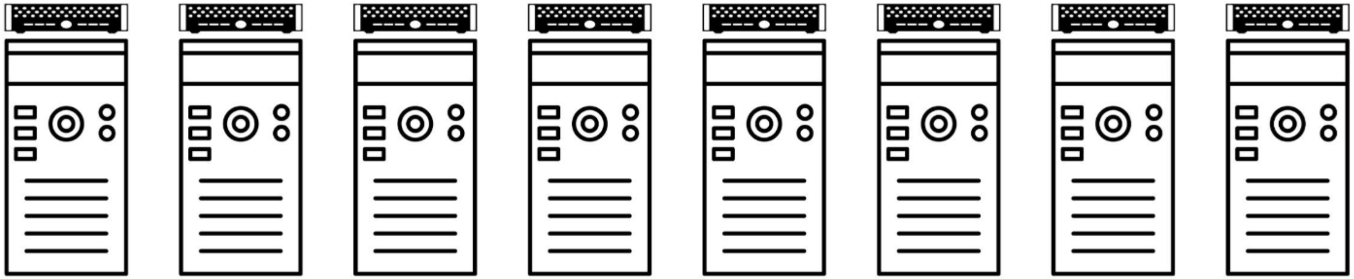
[1] Source: Microsoft, 2019

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

Root Cause

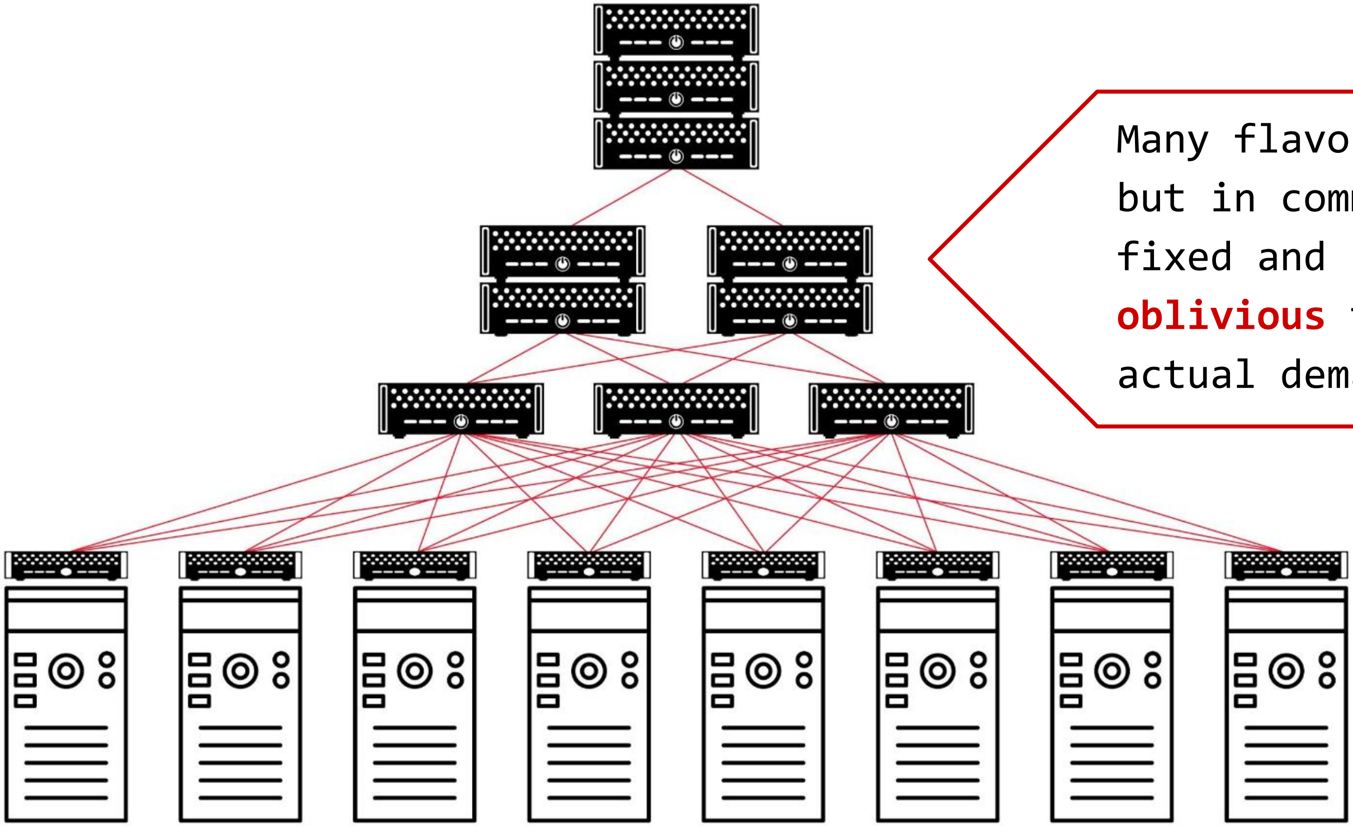
Fixed and Demand-Oblivious Topology

How to interconnect?



Root Cause

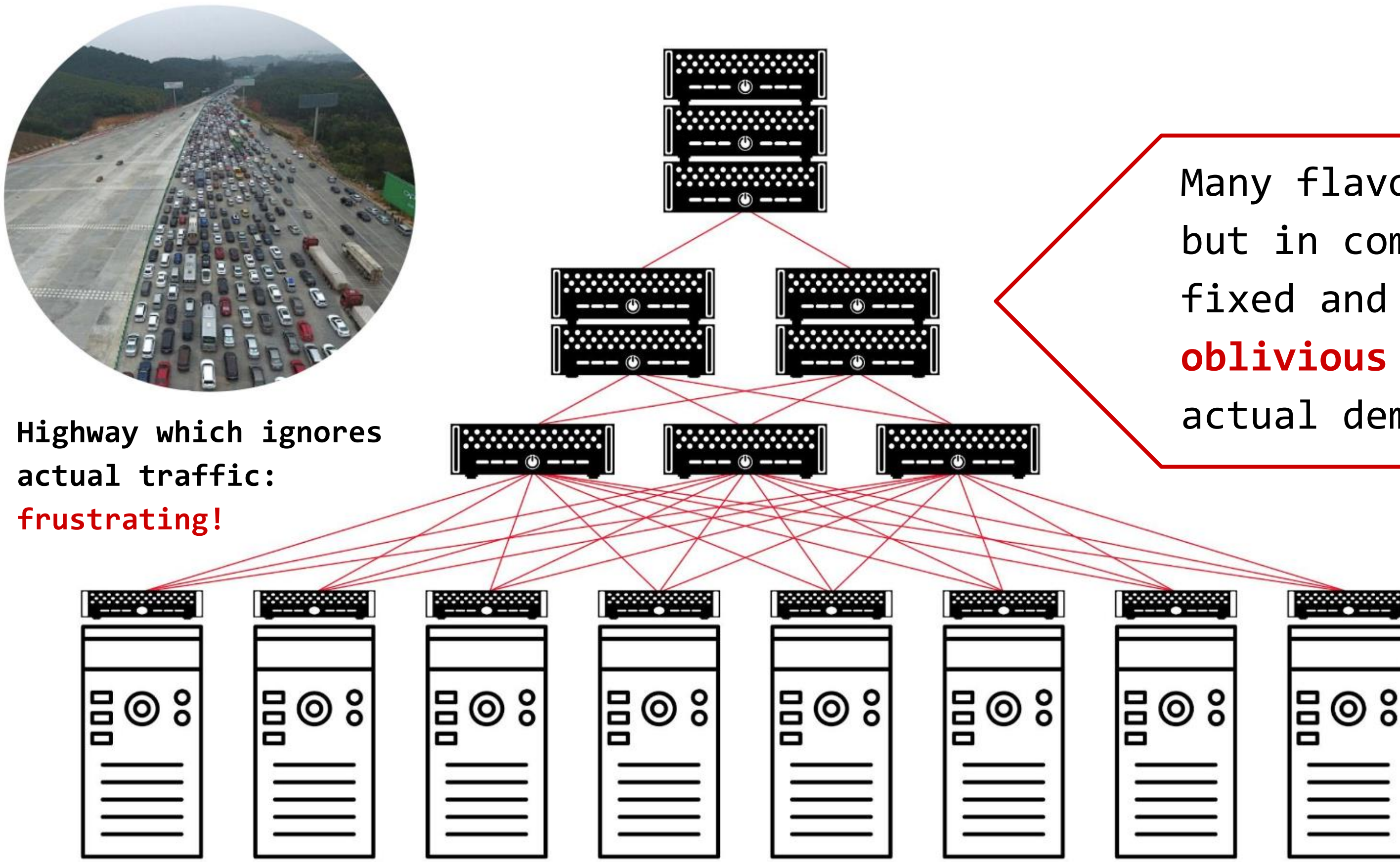
Fixed and Demand-Oblivious Topology



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

Root Cause

Fixed and Demand-Oblivious Topology

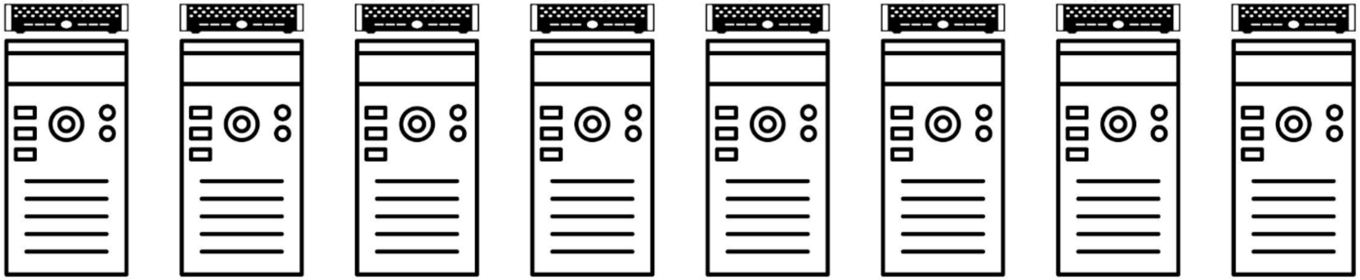


Highway which ignores actual traffic: frustrating!

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

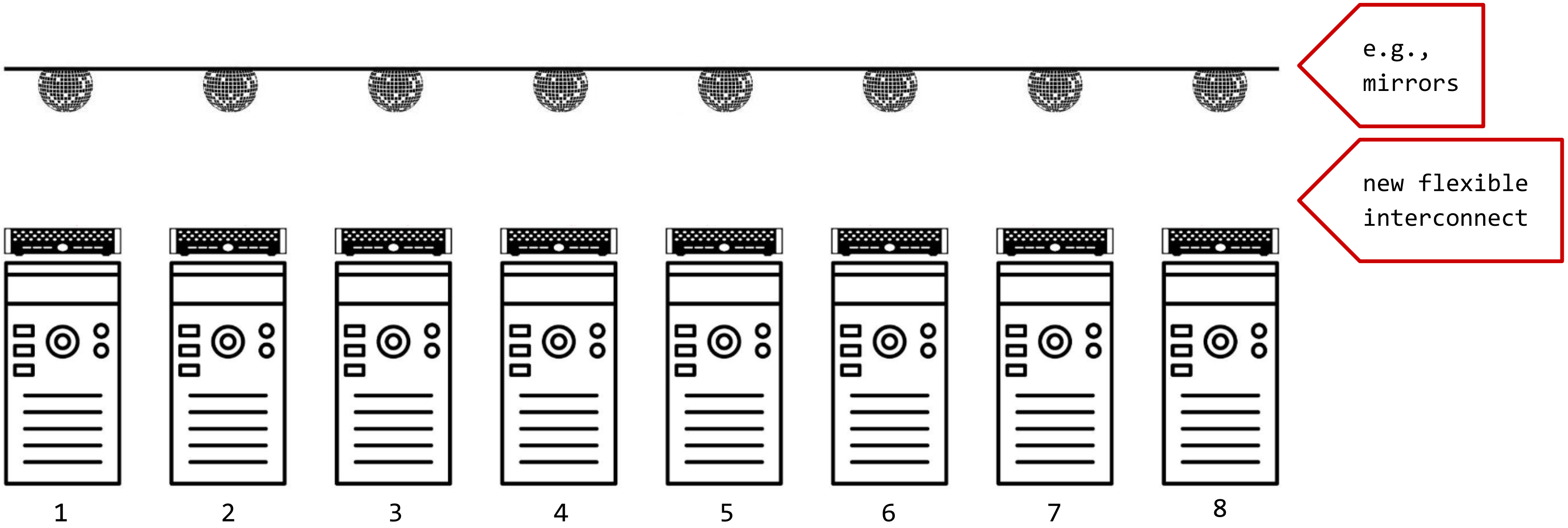
A Vision

Flexible and Demand-Aware Topologies



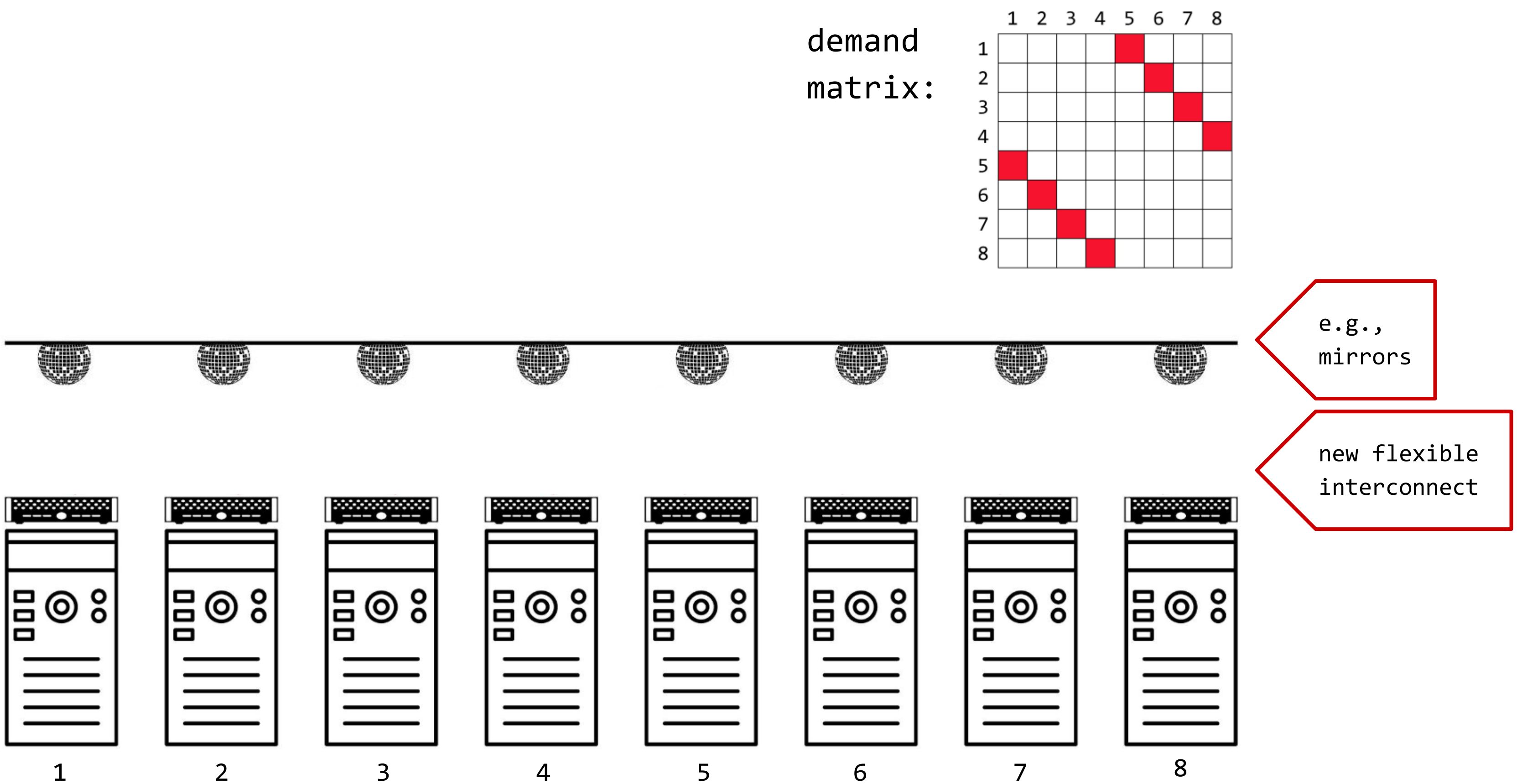
A Vision

Flexible and Demand-Aware Topologies



A Vision

Flexible and Demand-Aware Topologies

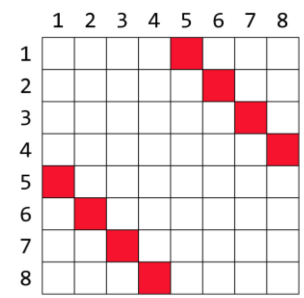


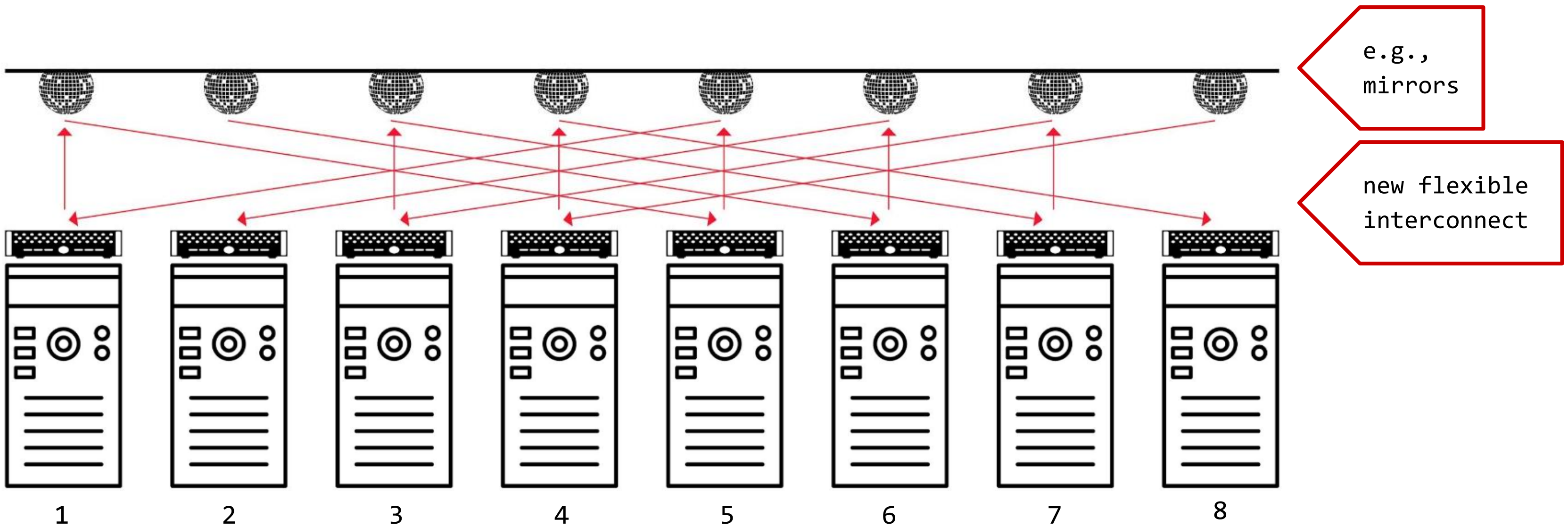
A Vision

Flexible and Demand-Aware Topologies

Matches demand

demand matrix:

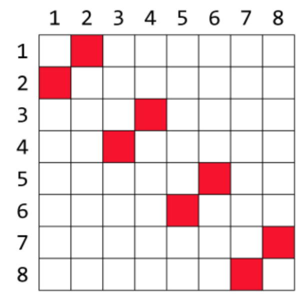


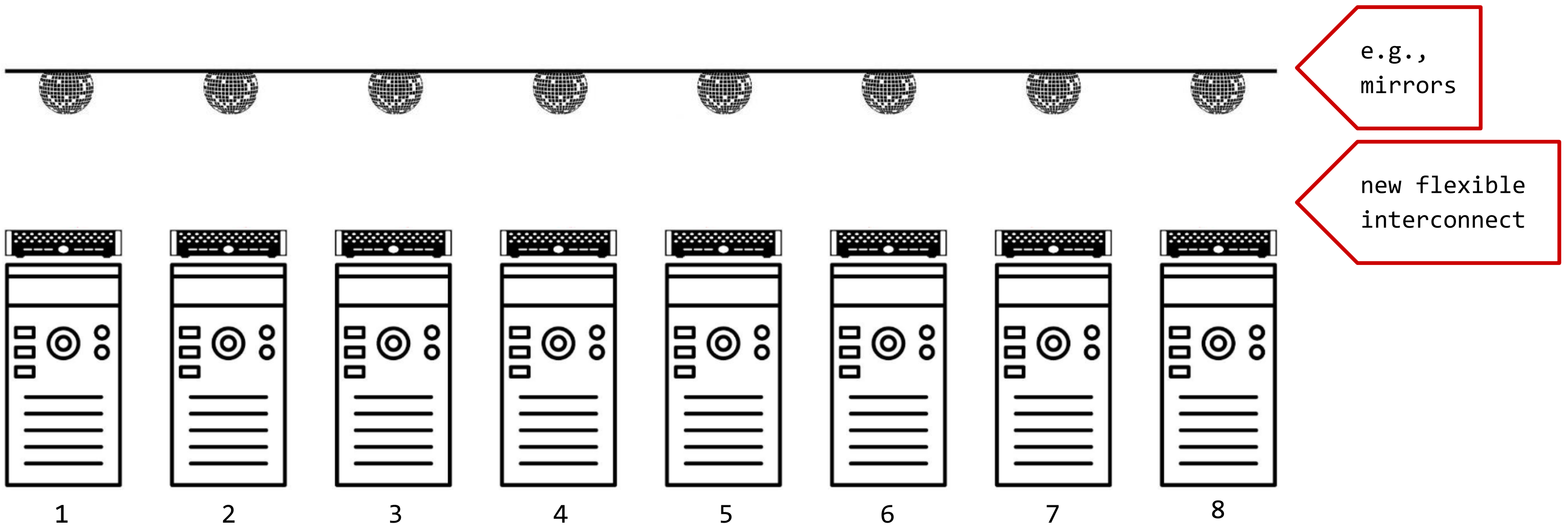


A Vision

Flexible and Demand-Aware Topologies

new demand:



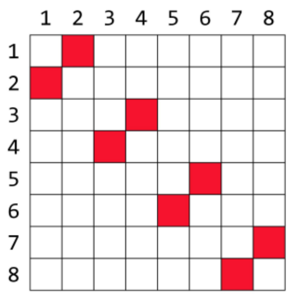


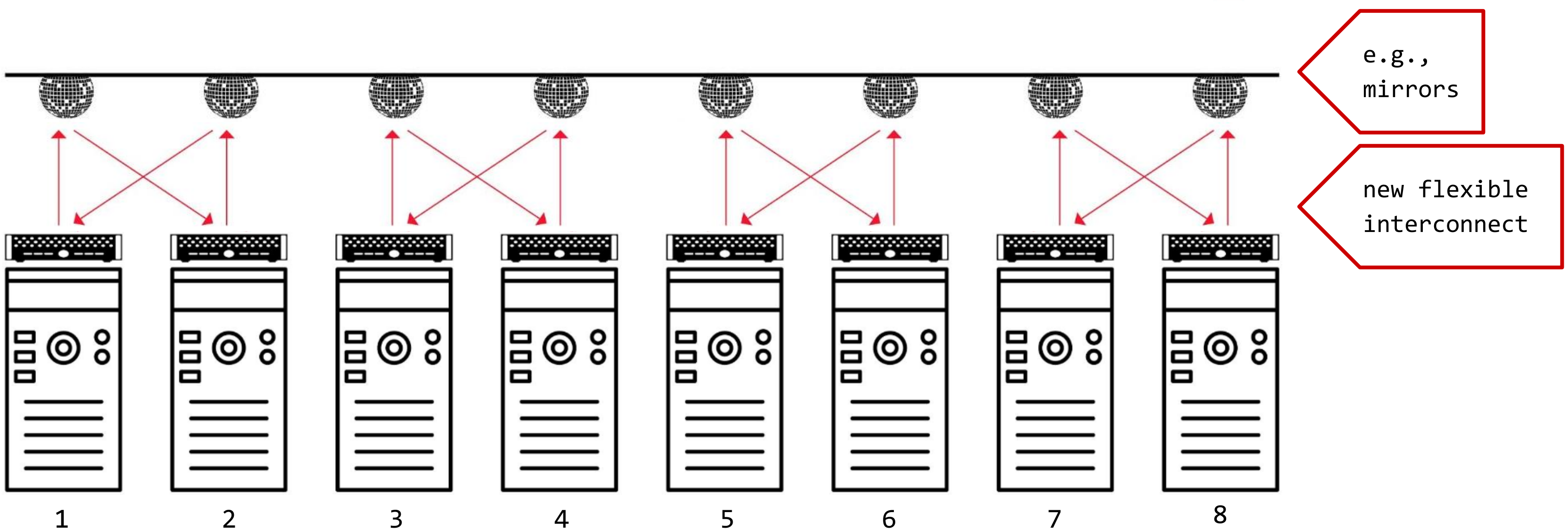
A Vision

Flexible and Demand-Aware Topologies

Matches demand

new demand:



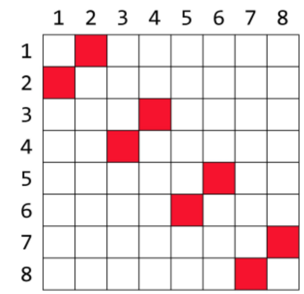


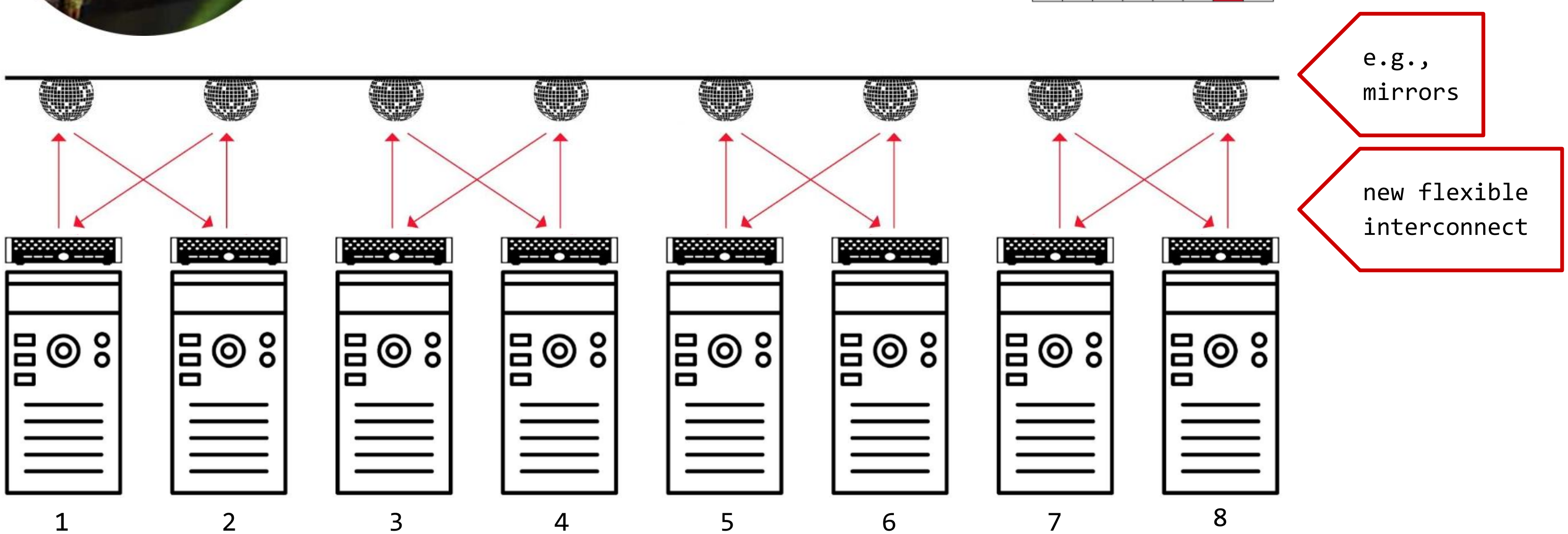
A Vision

Flexible and Demand-Aware Topologies

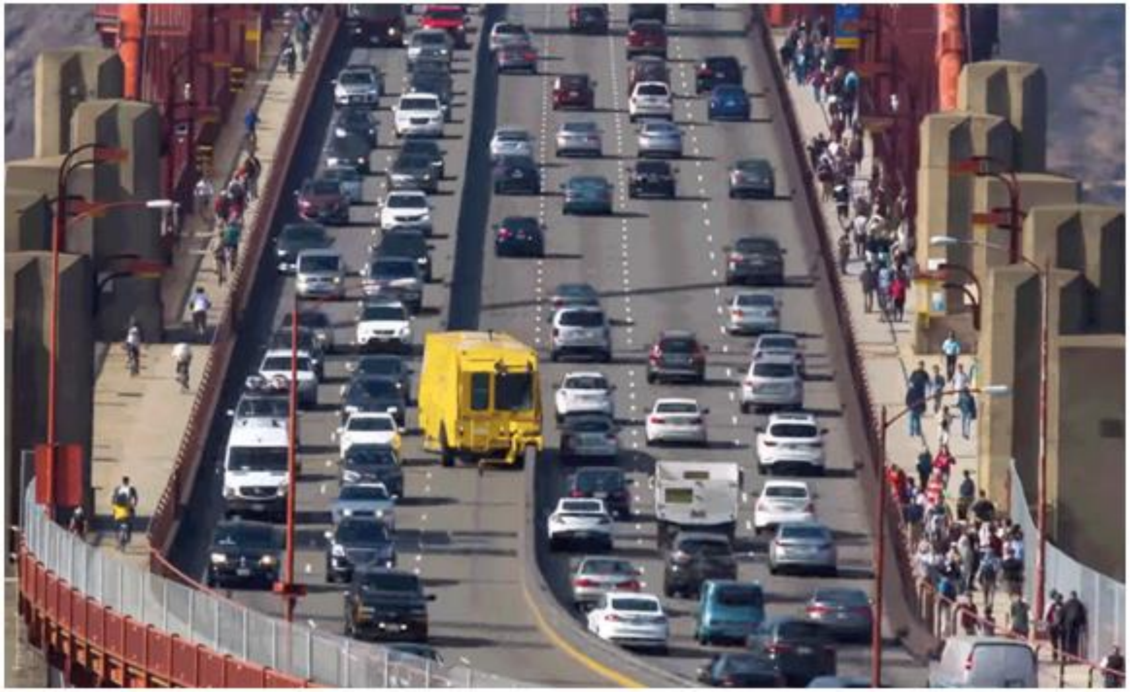
Self-Adjusting Networks

new demand:



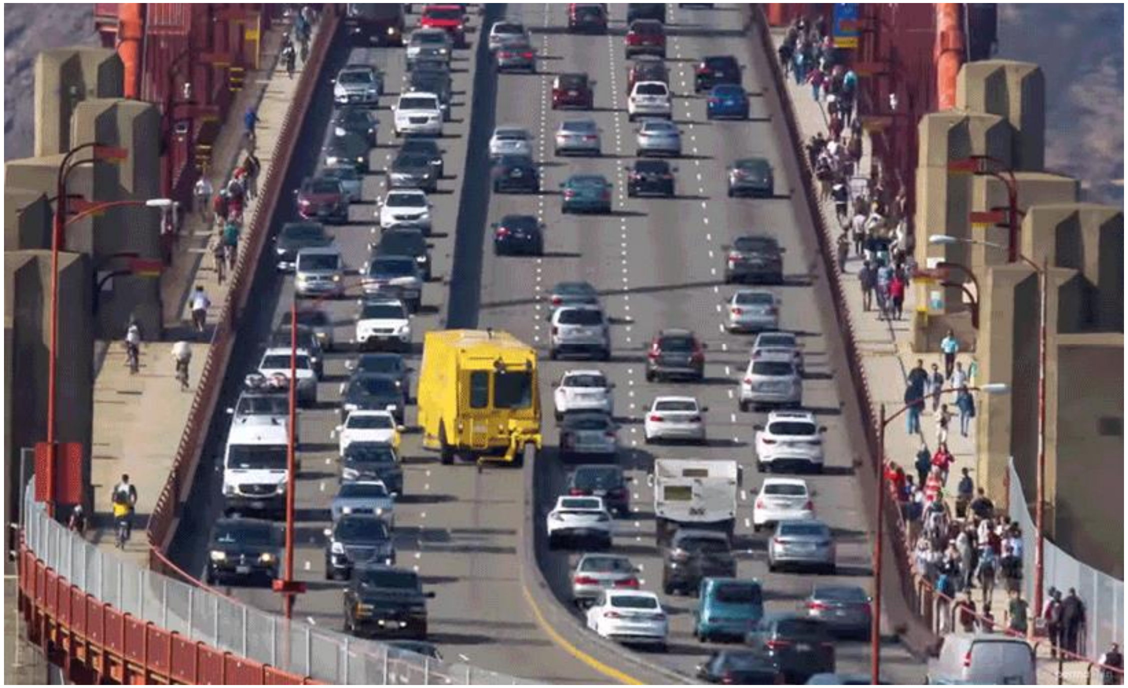


Analogy



Golden Gate Zipper

Analogy



Golden Gate Zipper

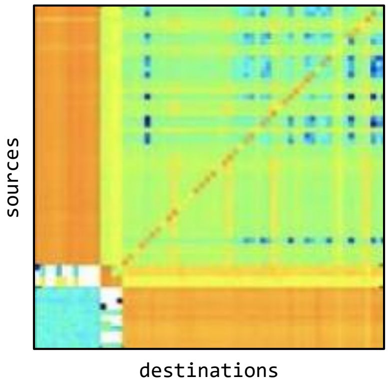
The Motivation

Much Structure in the Demand

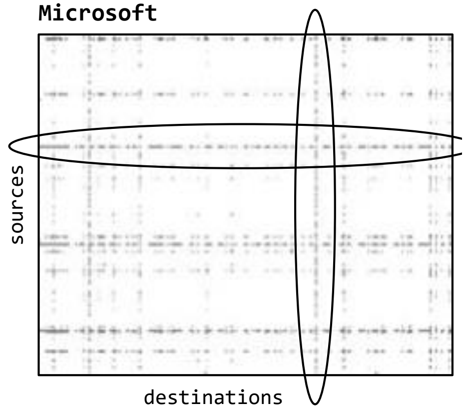
Empirical studies:

traffic matrices sparse and skewed

Facebook

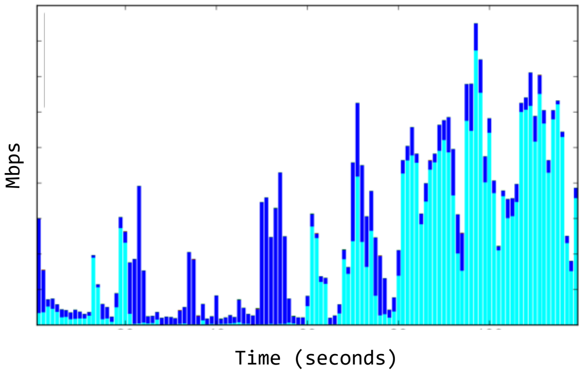


Microsoft



traffic bursty over time

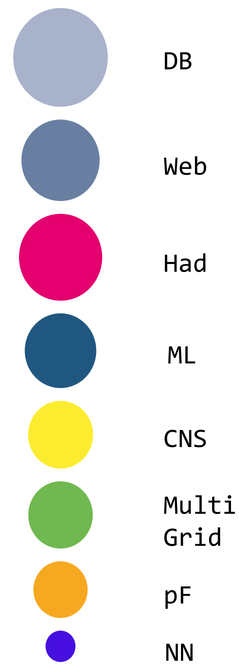
Facebook



The hypothesis: can be exploited.

Recent Representation of Trace Structure:

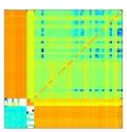
Complexity Map

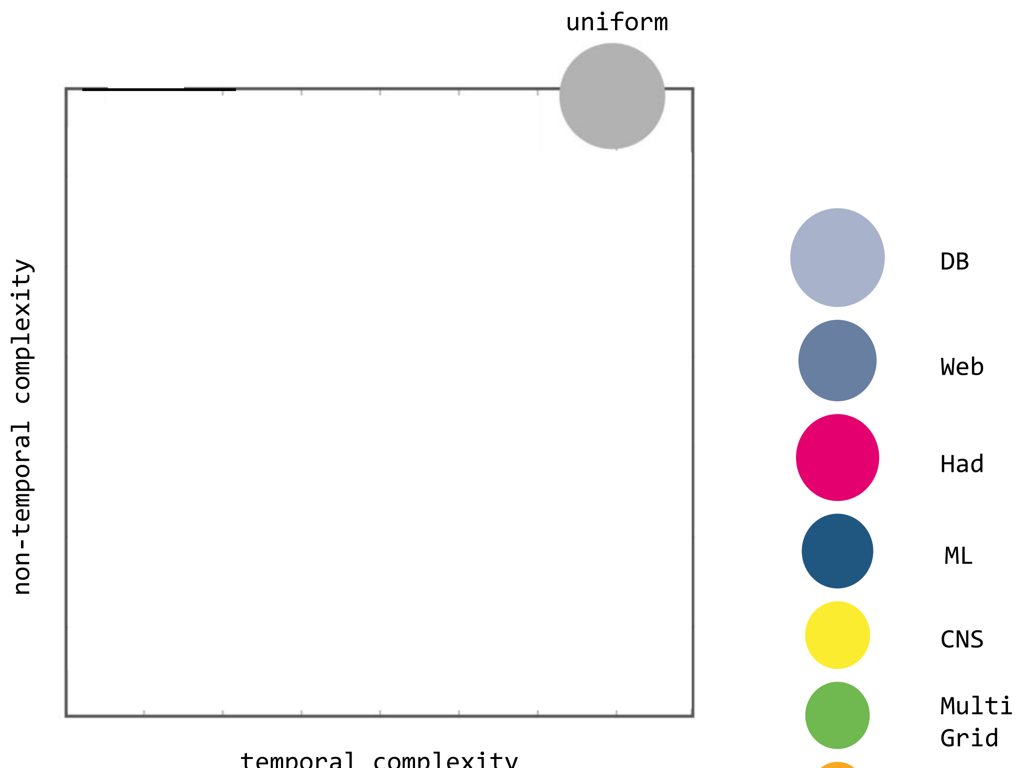


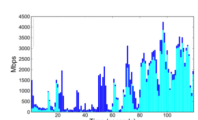
Griner et al., SIGMETRICS 2020

Recent Representation of Trace Structure:

Complexity Map



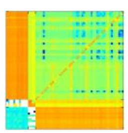


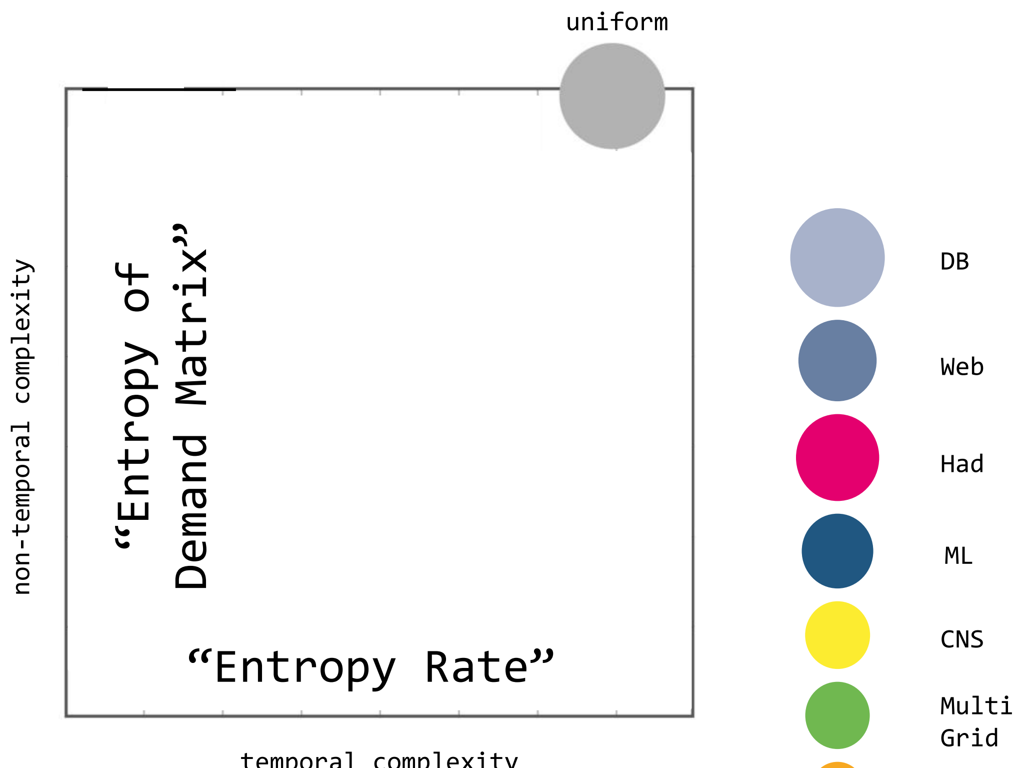


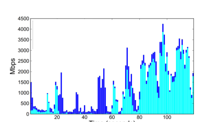
Griner et al., SIGMETRICS 2020

Recent Representation of Trace Structure:

Complexity Map



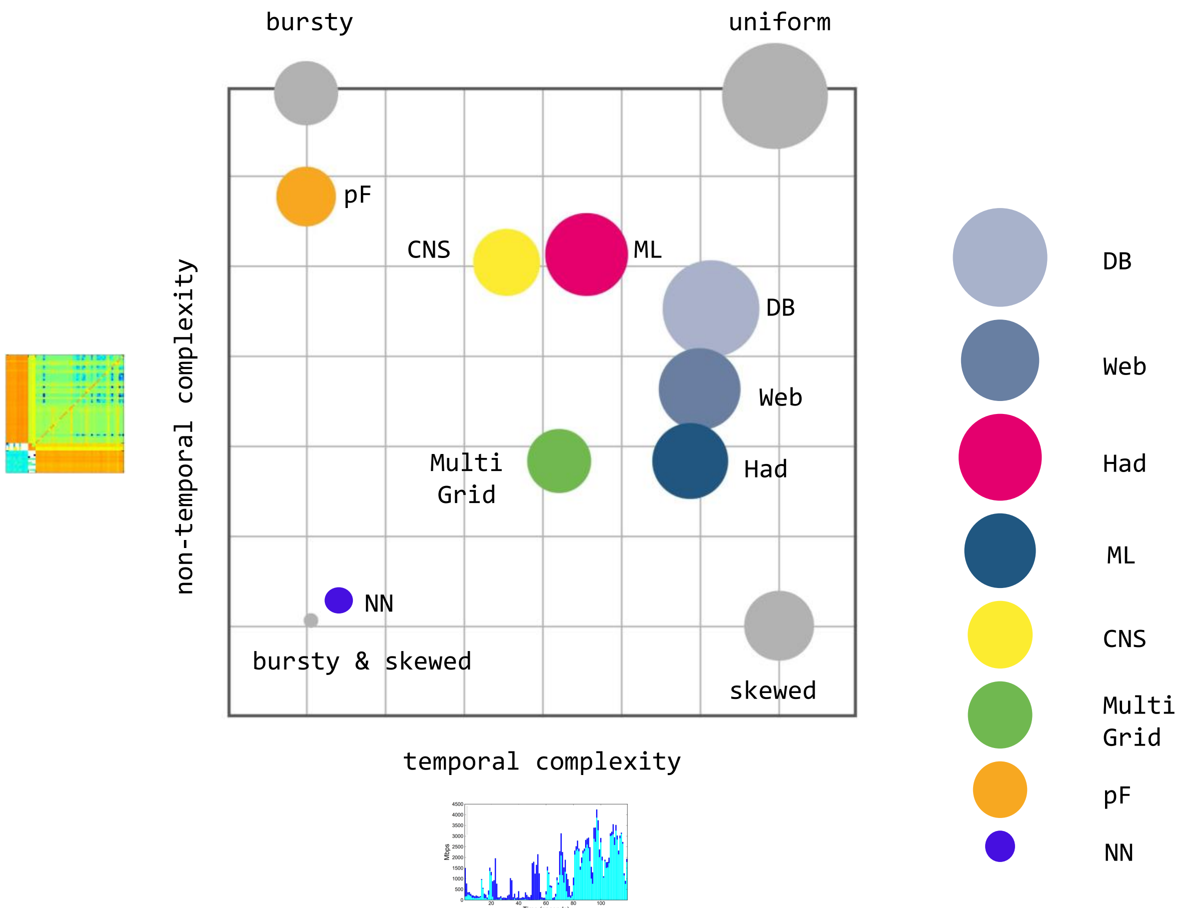




Griner et al., SIGMETRICS 2020

Recent Representation of Trace Structure:

Complexity Map

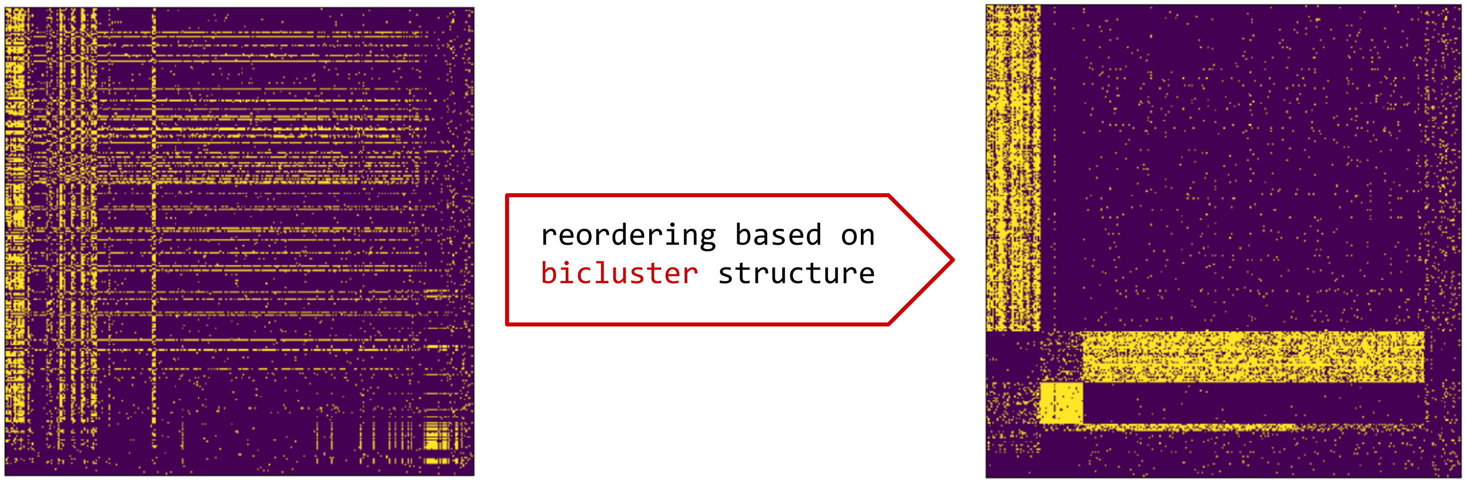


Different structures!

Griner et al., SIGMETRICS 2020

Traffic is also clustered:

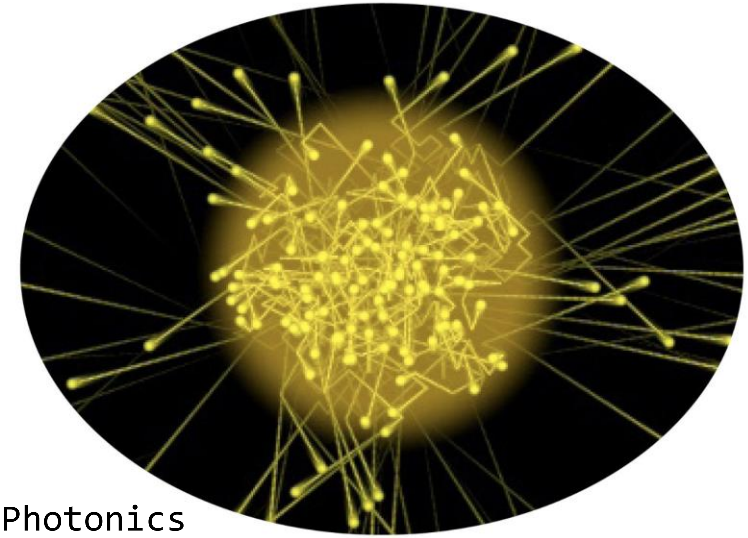
Small Stable Clusters



Opportunity: *exploit* with little reconfigurations!

Förster et al., Analyzing the Communication Clusters in Datacenters. WWW 2023

Sounds Crazy? Emerging Enabling Technology.



Photonics

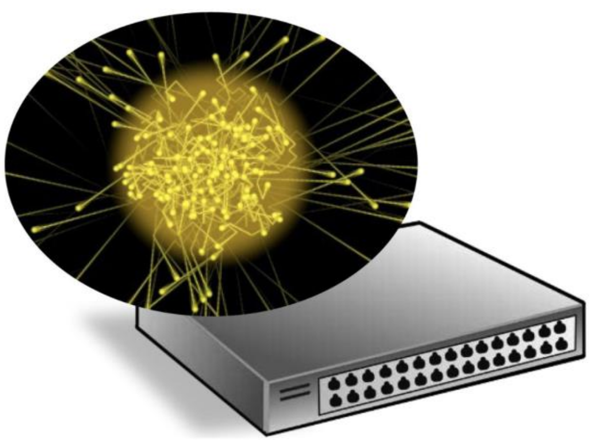
H2020:
"Photonics one of only five key enabling technologies for future prosperity."

US National Research Council:
"Photons are the new Electrons."

Enabler

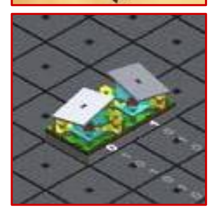
Novel Reconfigurable Optical Switches

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



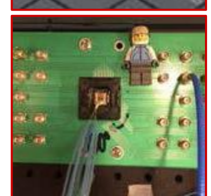
Prototype 1

Moving antenna (ms)



Prototype 2

Moving mirrors (mus)



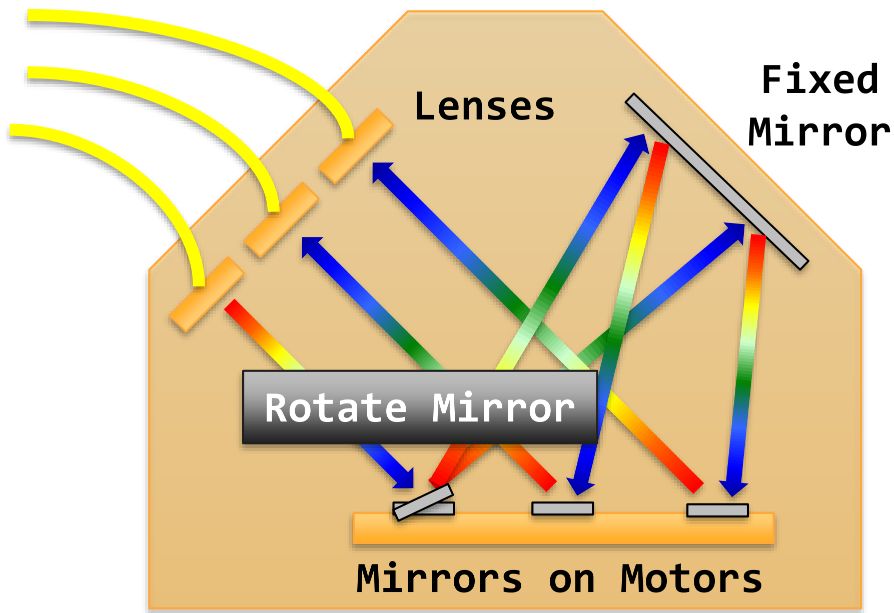
Prototype 3

Changing lambdas (ns)

Example

Optical Circuit Switch

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



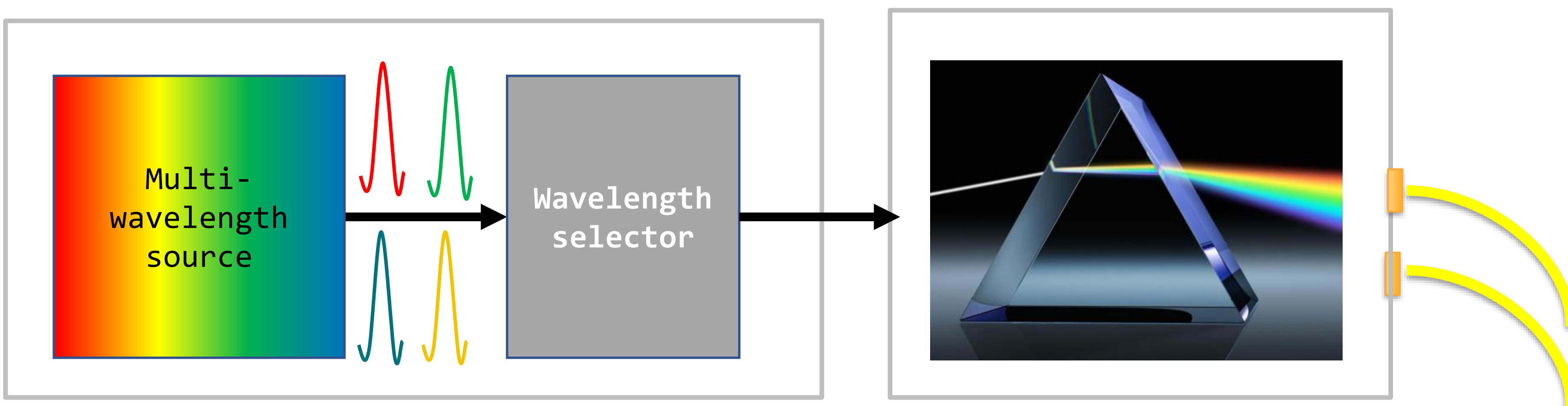
Optical Circuit Switch

By Nathan Farrington, SIGCOMM 2010

Another Example

Tunable Lasers

- ⇢ Depending on wavelength, forwarded differently
- ⇢ Optical switch is passive



Electrical switch with tunable laser

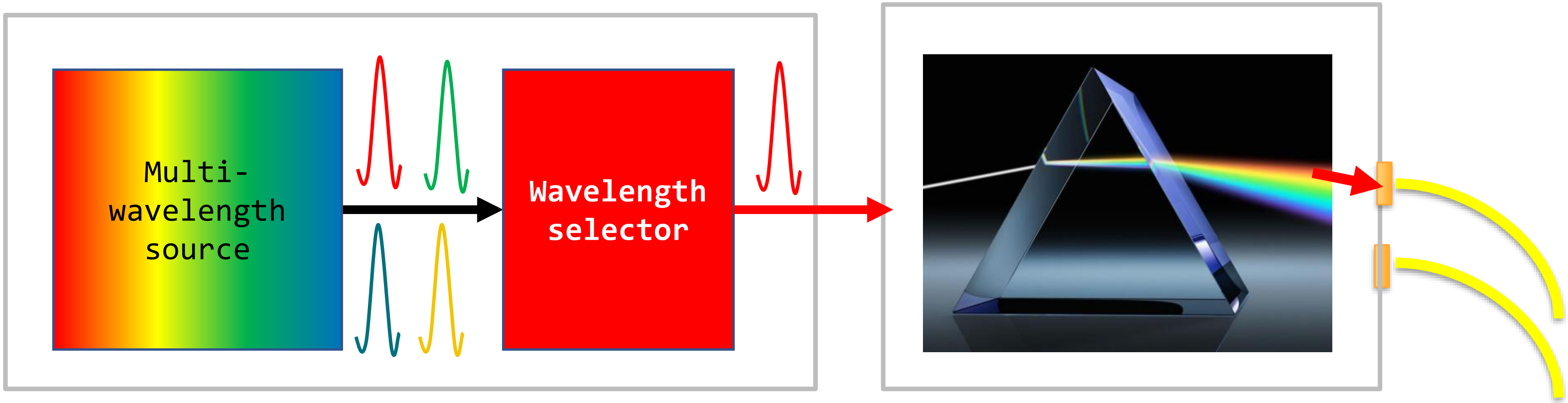
Optical switch
Passive

Ballani et al., Sirius, ACM SIGCOMM 2020.

Another Example

Tunable Lasers

- Depending on wavelength, forwarded differently
- Optical switch is passive



Electrical switch with tunable laser

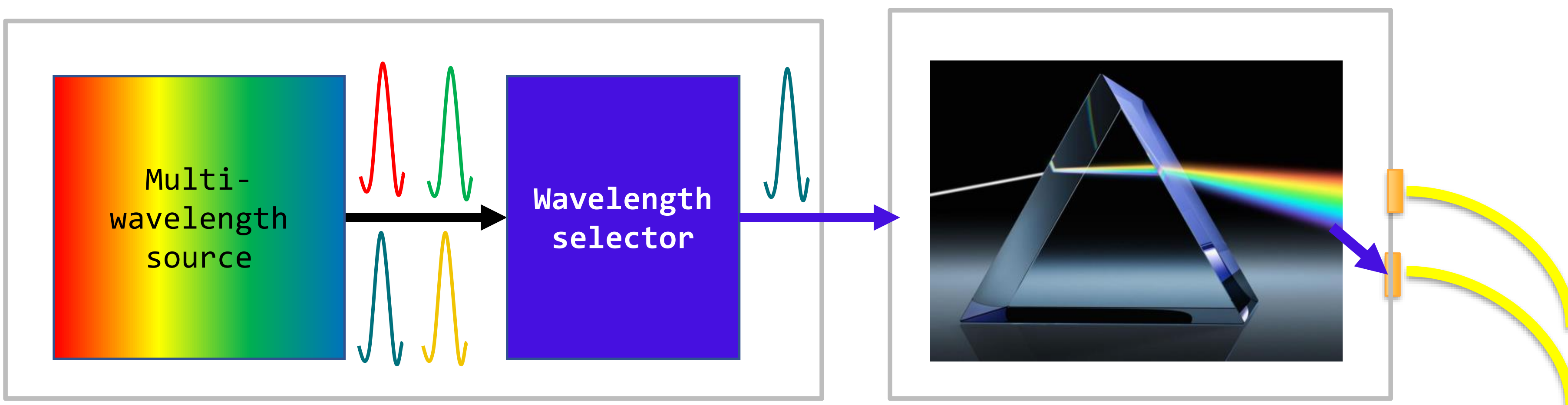
Optical switch ***Passive***

Ballani et al., Sirius, ACM SIGCOMM 2020.

Another Example

Tunable Lasers

- Depending on wavelength, forwarded differently
- Optical switch is passive



Electrical switch with tunable laser

Optical switch *Passive*

Ballani et al., Sirius, ACM SIGCOMM 2020.

First Deployments

E.g., Google’s Datacenter Jupiter

Systems

Jupiter evolving: Reflecting on Google’s data center network transformation

August 24, 2022

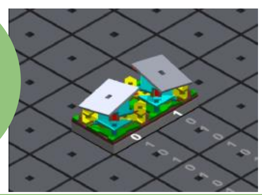
Amin Vahdat

VP & GM, Systems and Services Infrastructure

The Big Picture

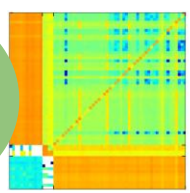
New!

Flexibility



More!

Structure

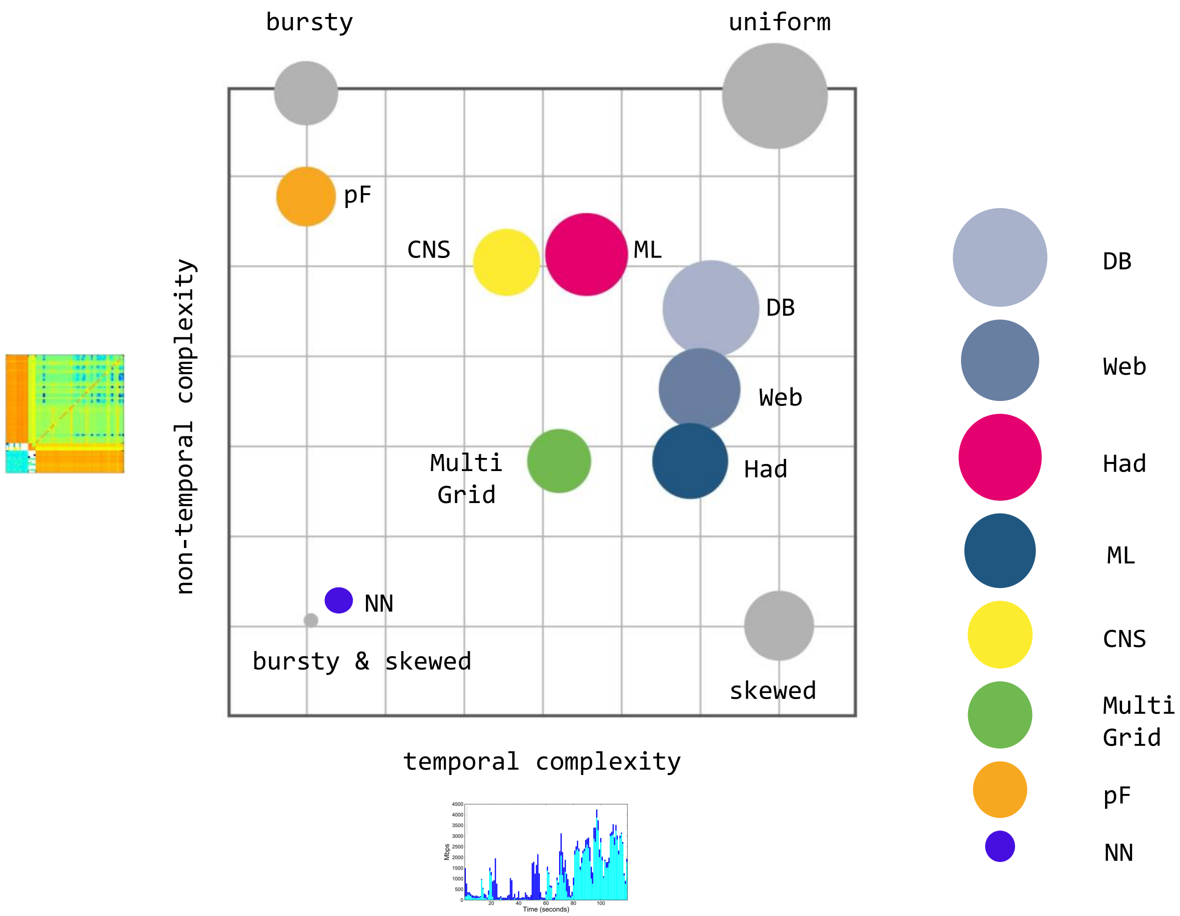


Self-Adjusting Networks

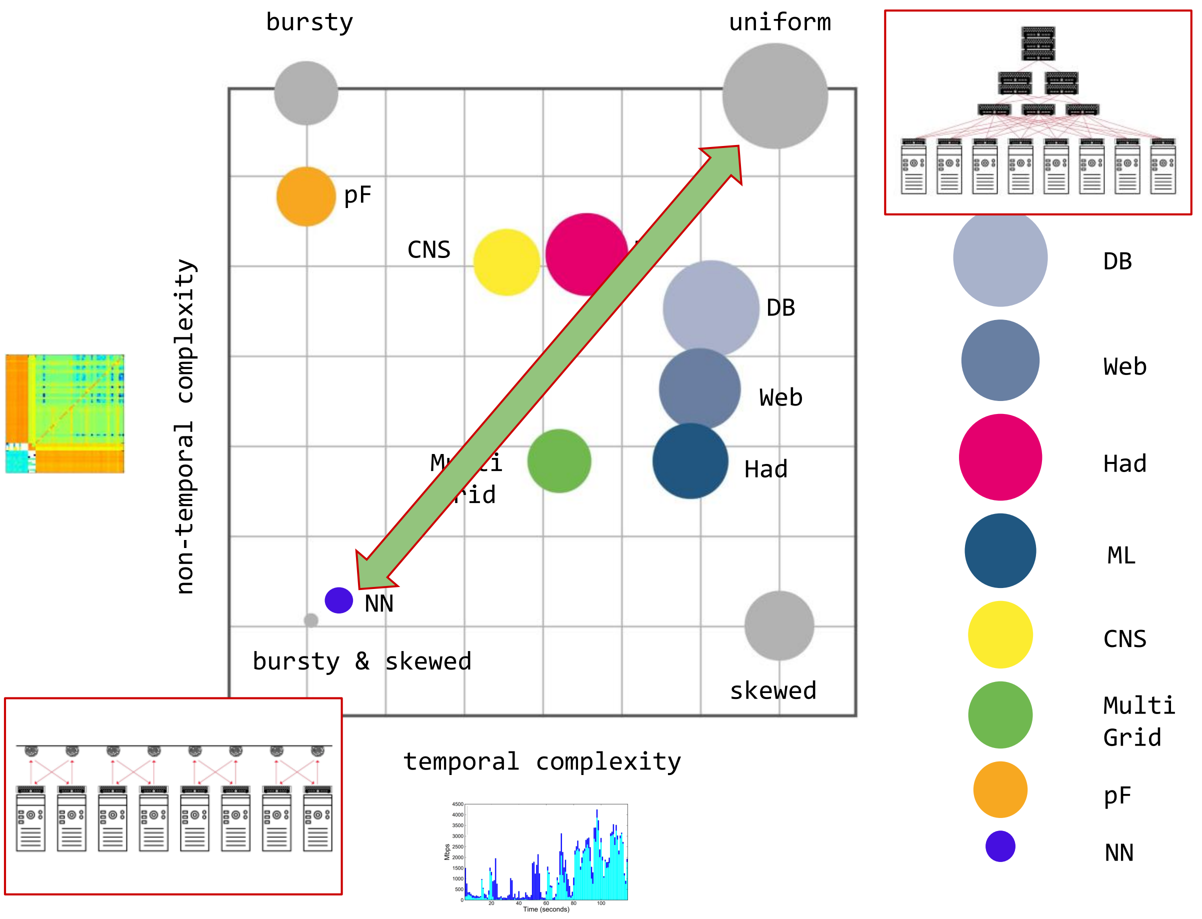
Efficiency

Now is the time!

Potential Gain



Potential Gain



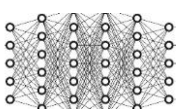
Unique Position

Demand-Aware, Self-Adjusting Systems

Everywhere, but mainly in software

Algorithmic trading

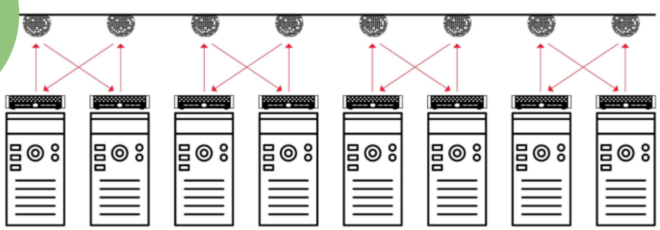
Recommender systems



Neural networks

VS

Our focus in this talk: in hardware



Design Choices

Diverse topology components:

→ demand-oblivious and

demand-aware

Demand-oblivious

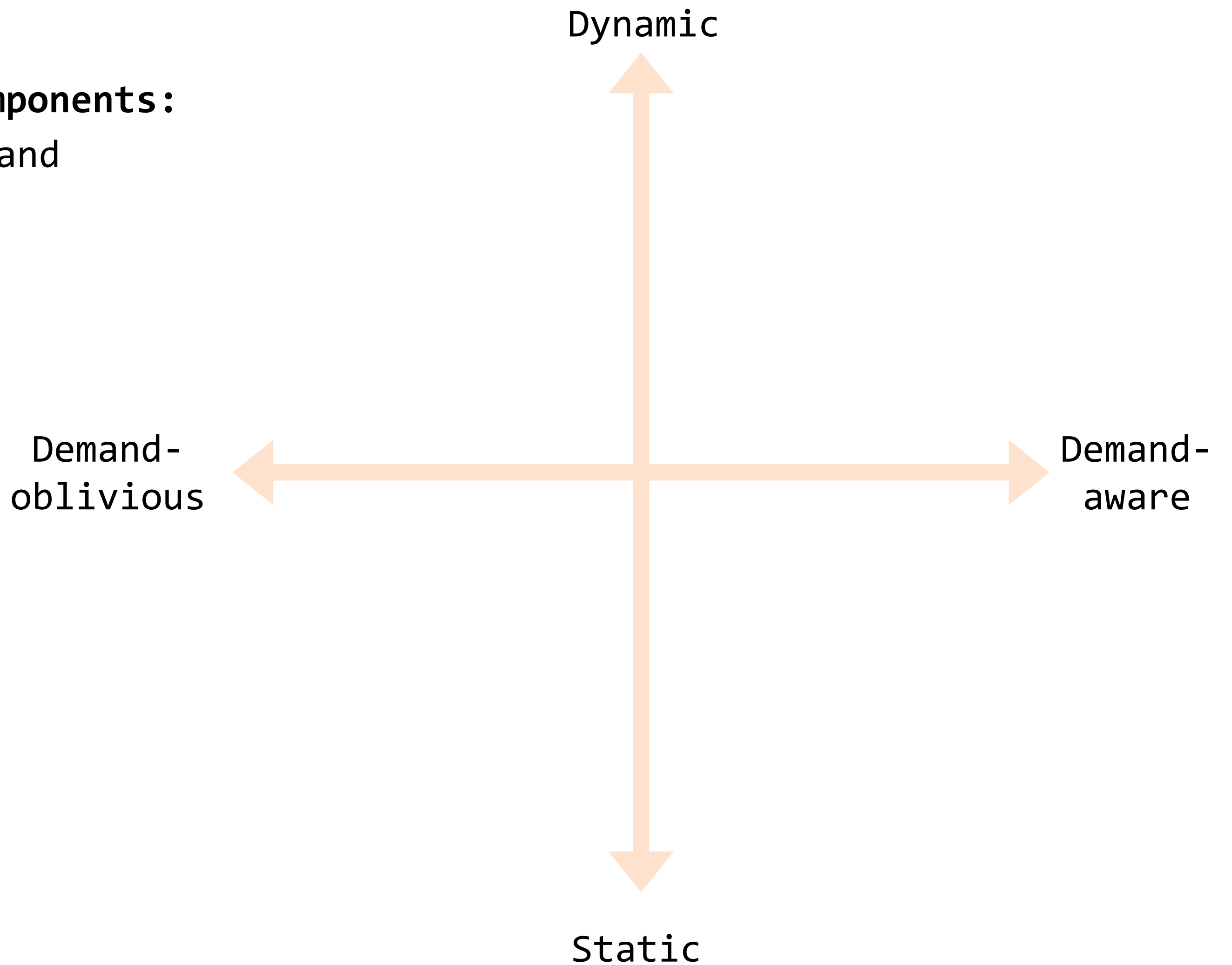
Demand-aware

Design Choices

Diverse topology components:

⇀ demand-oblivious and
demand-aware

⇀ static vs dynamic

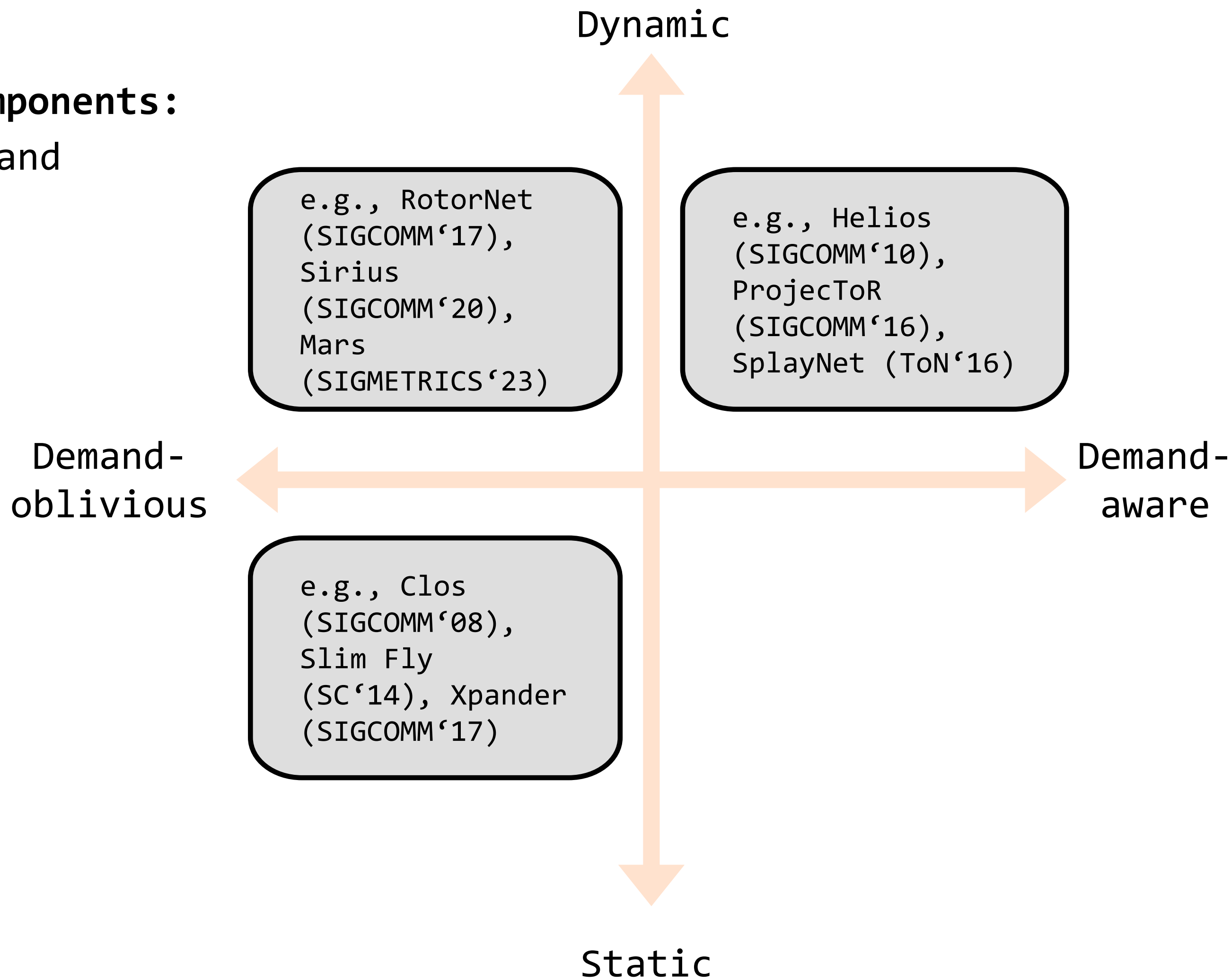


Design Choices

Diverse topology components:

→ demand-oblivious and demand-aware

→ static vs dynamic

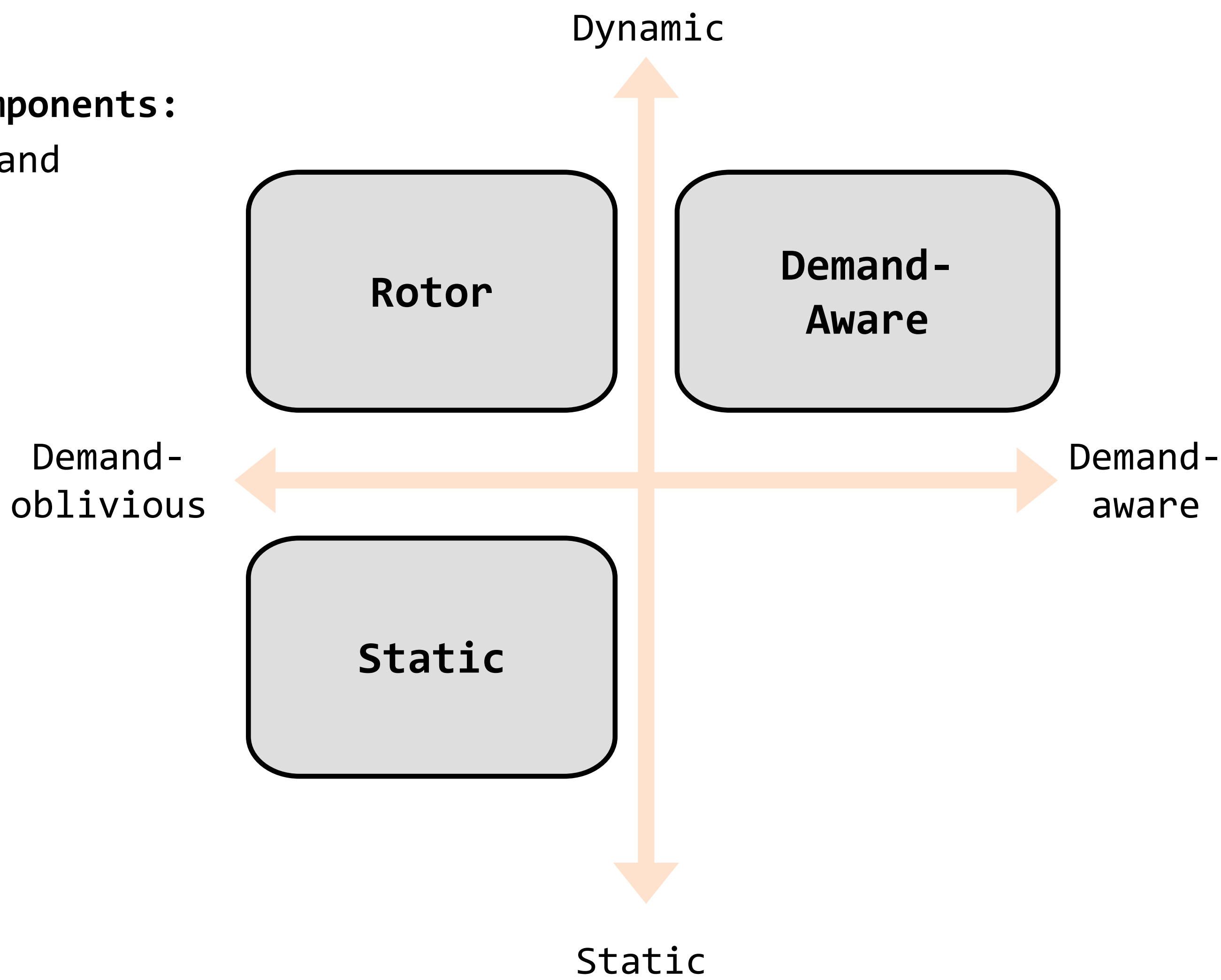


Design Choices

Diverse topology components:

→ demand-oblivious and demand-aware

→ static vs dynamic

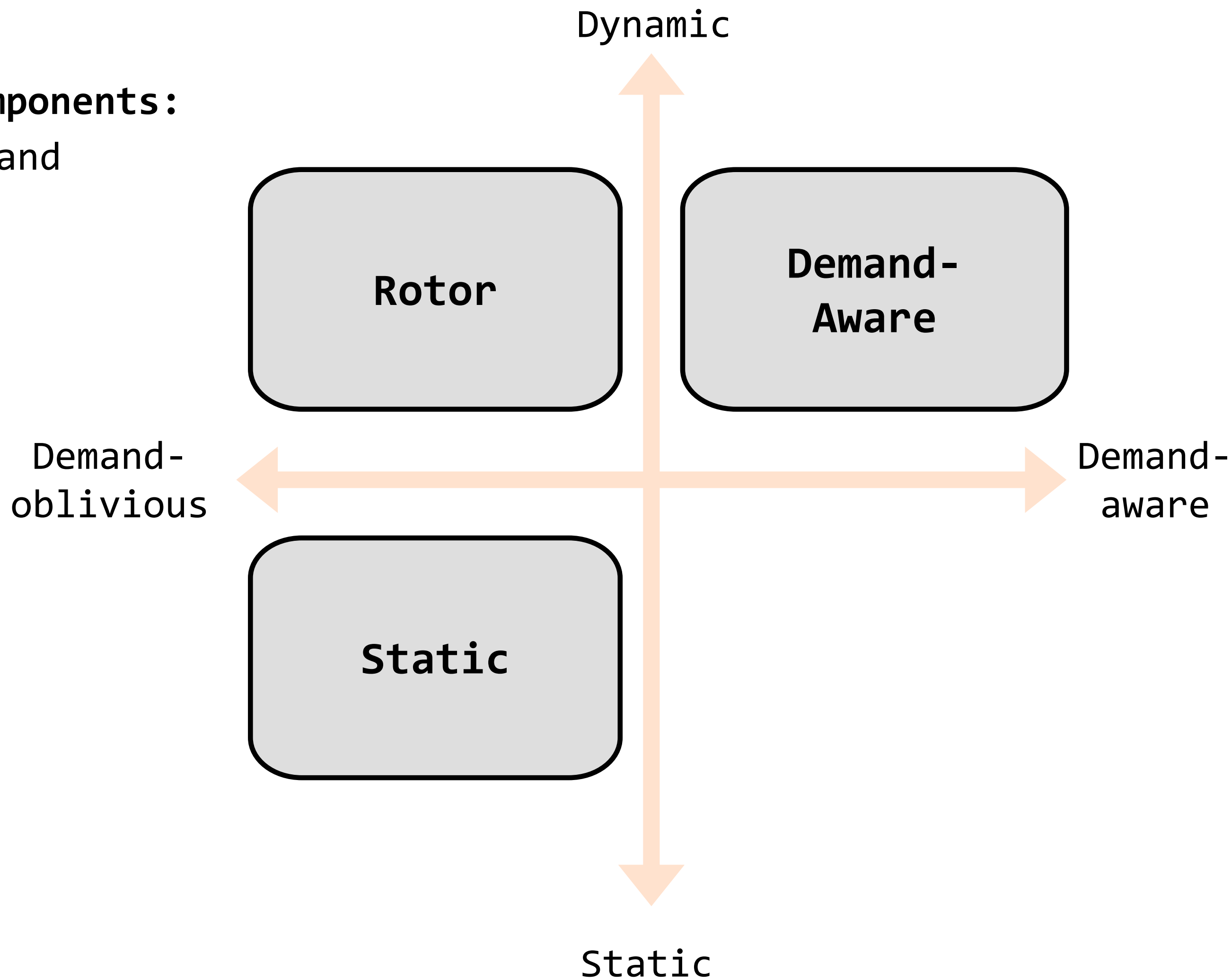


Design Choices

Diverse topology components:

→ demand-oblivious and
  demand-aware

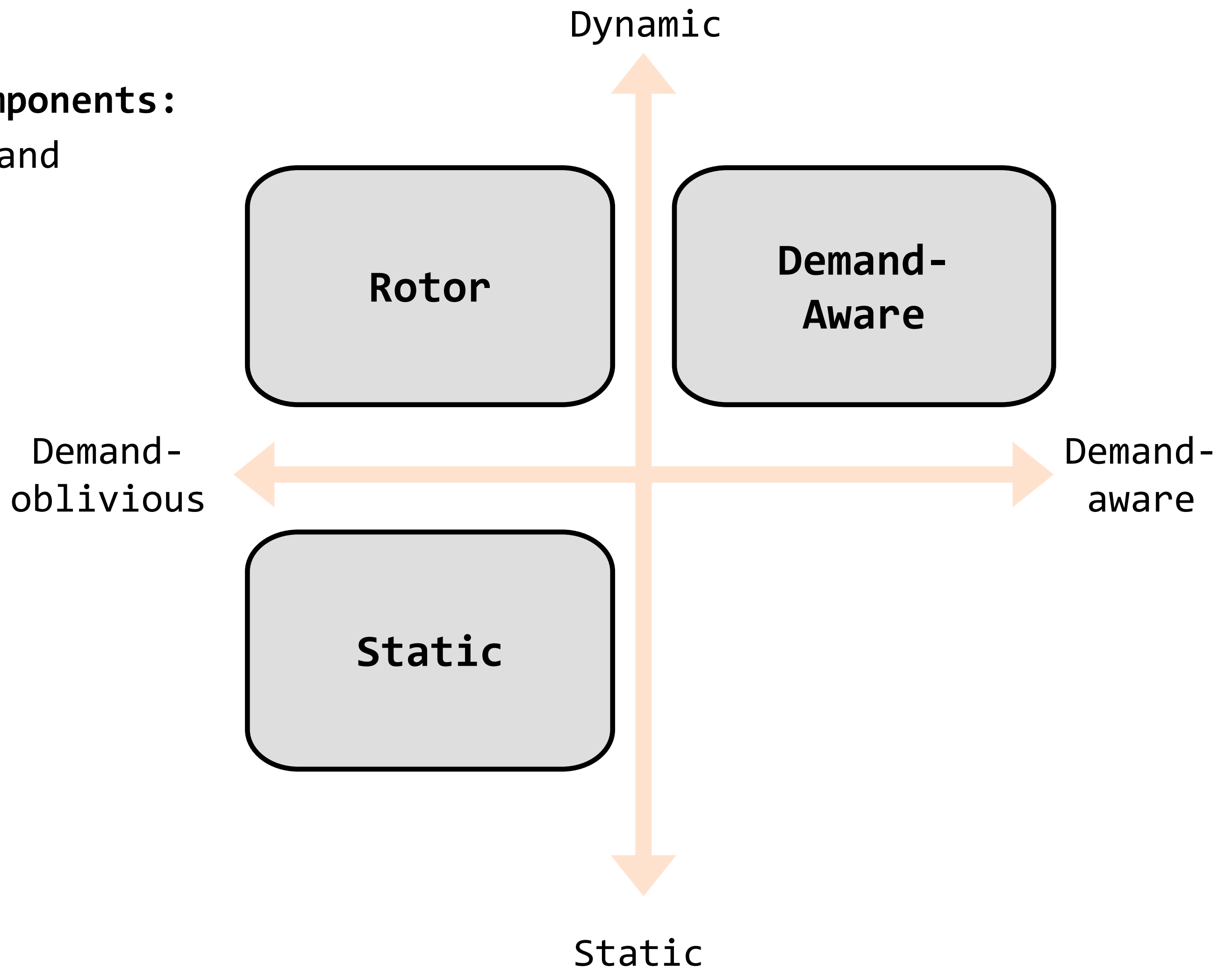
→ static vs dynamic



Which approach is best?

Design Choices

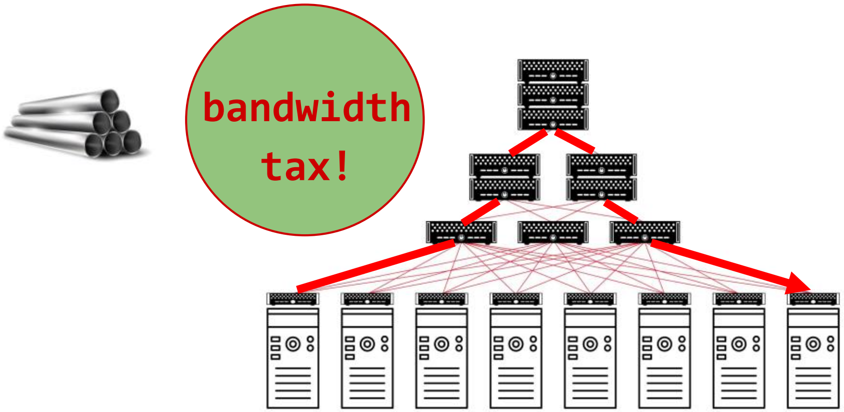
Diverse topology components:
⇀ demand-oblivious and
demand-aware
⇀ static vs dynamic



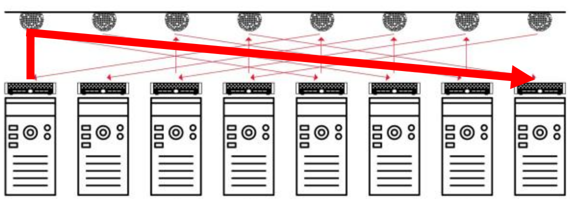
As always in CS:
It depends…

Costs and Tradeoffs

→ **Good:** Demand-aware networks may be really useful to serve large flows (elephant flows): avoiding multi-hop routing



vs



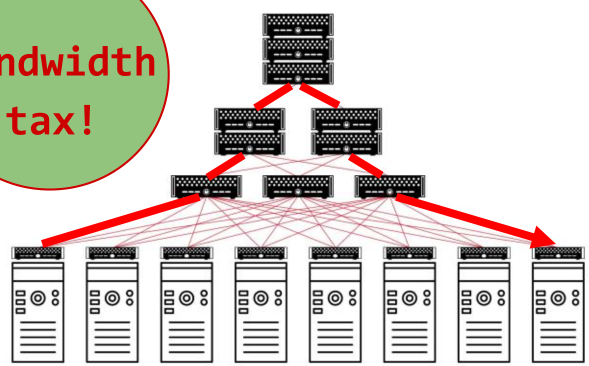
6 hops

1 hop

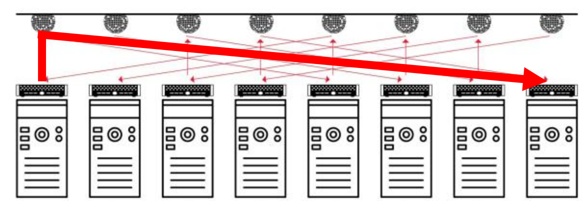
Costs and Tradeoffs

→ **Good:** Demand-aware networks may be really useful to serve large flows (elephant flows): avoiding multi-hop routing

bandwidth tax!



vs

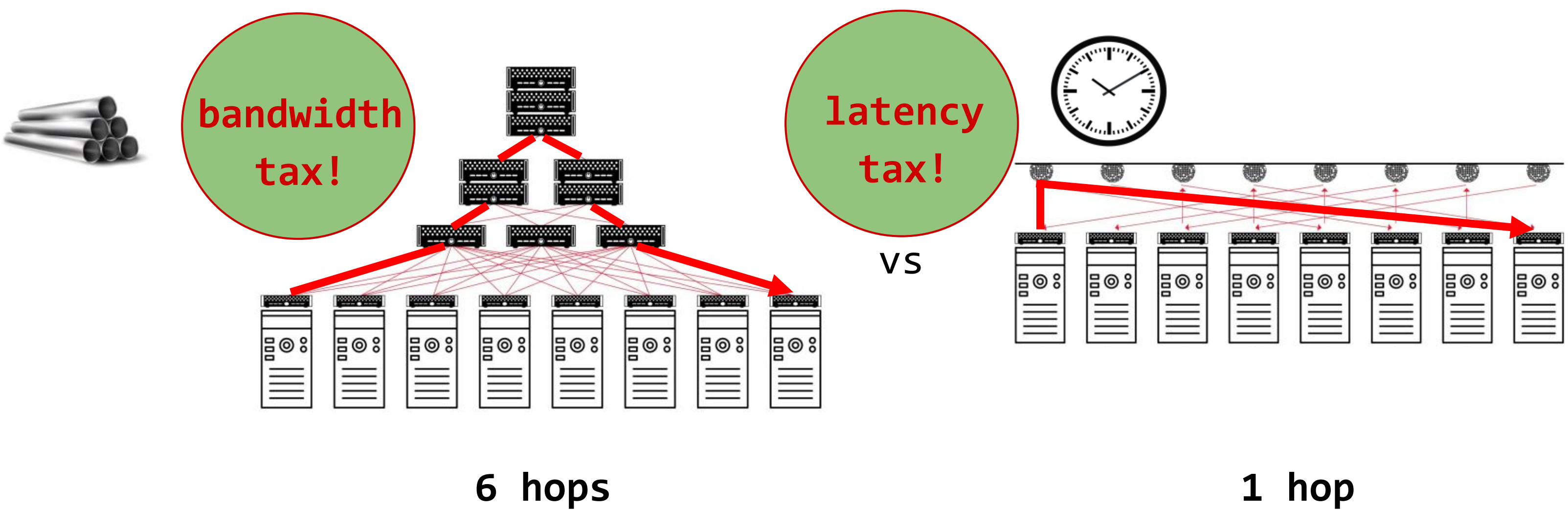


6 hops **1 hop**

→ **However:** requires optimization and adaption, which takes time

Costs and Tradeoffs

→ **Good:** Demand-aware networks may be really useful to serve large flows (elephant flows): avoiding multi-hop routing



→ **However:** requires optimization and adaption, which takes time

Optimal Design Depends on

Traffic Types

Diverse patterns:

⇀ Shuffling/Hadoop:
all-to-all

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

⇀ Elephant flows

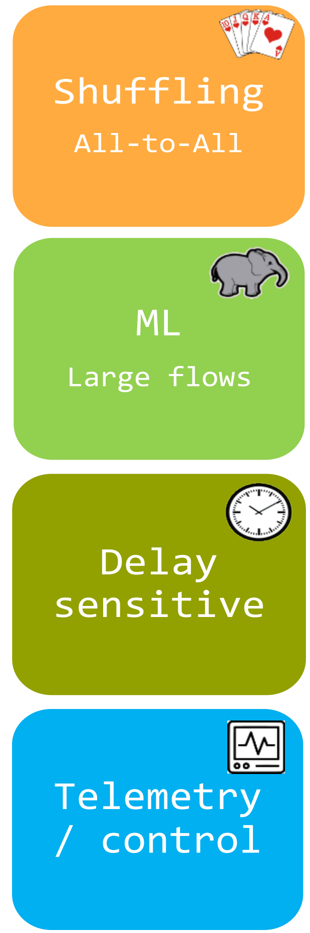
⇀ Query traffic: skewed

⇀ Mice flows

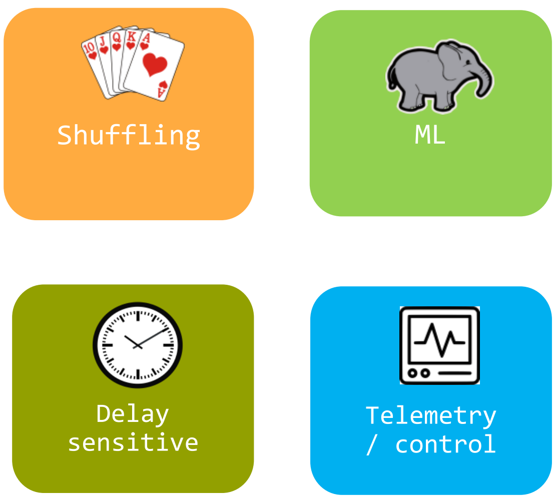
⇀ Control traffic: does not evolve
but has non-temporal structure

Diverse requirements:

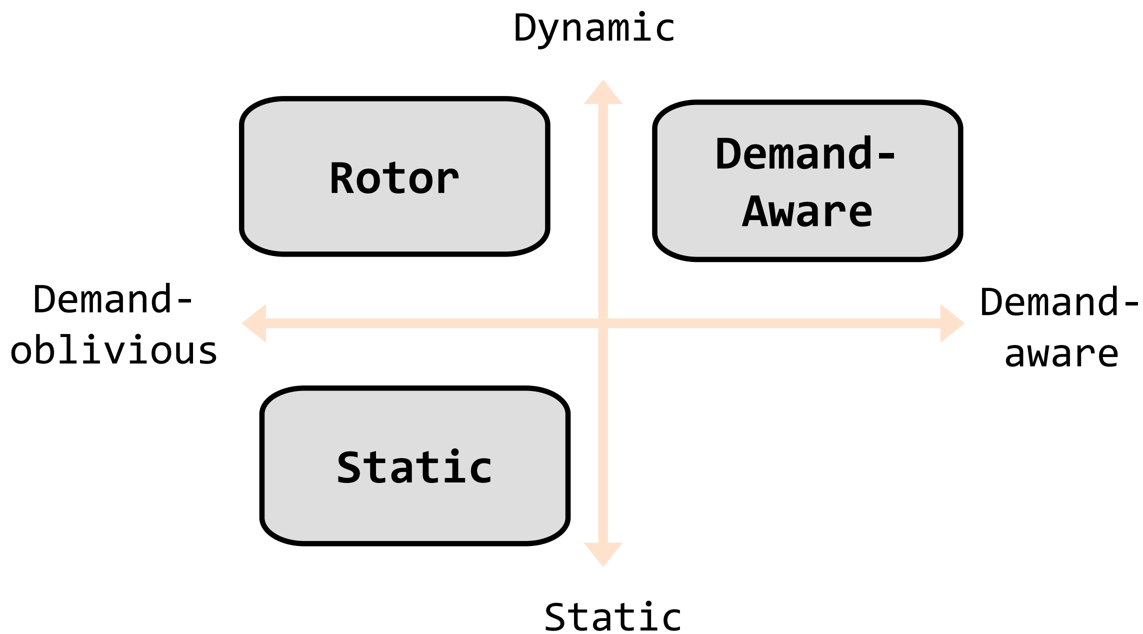
⇀ ML is bandwidth hungry,
small flows are latency-
sensitive



Examples:
Match or Mismatch?

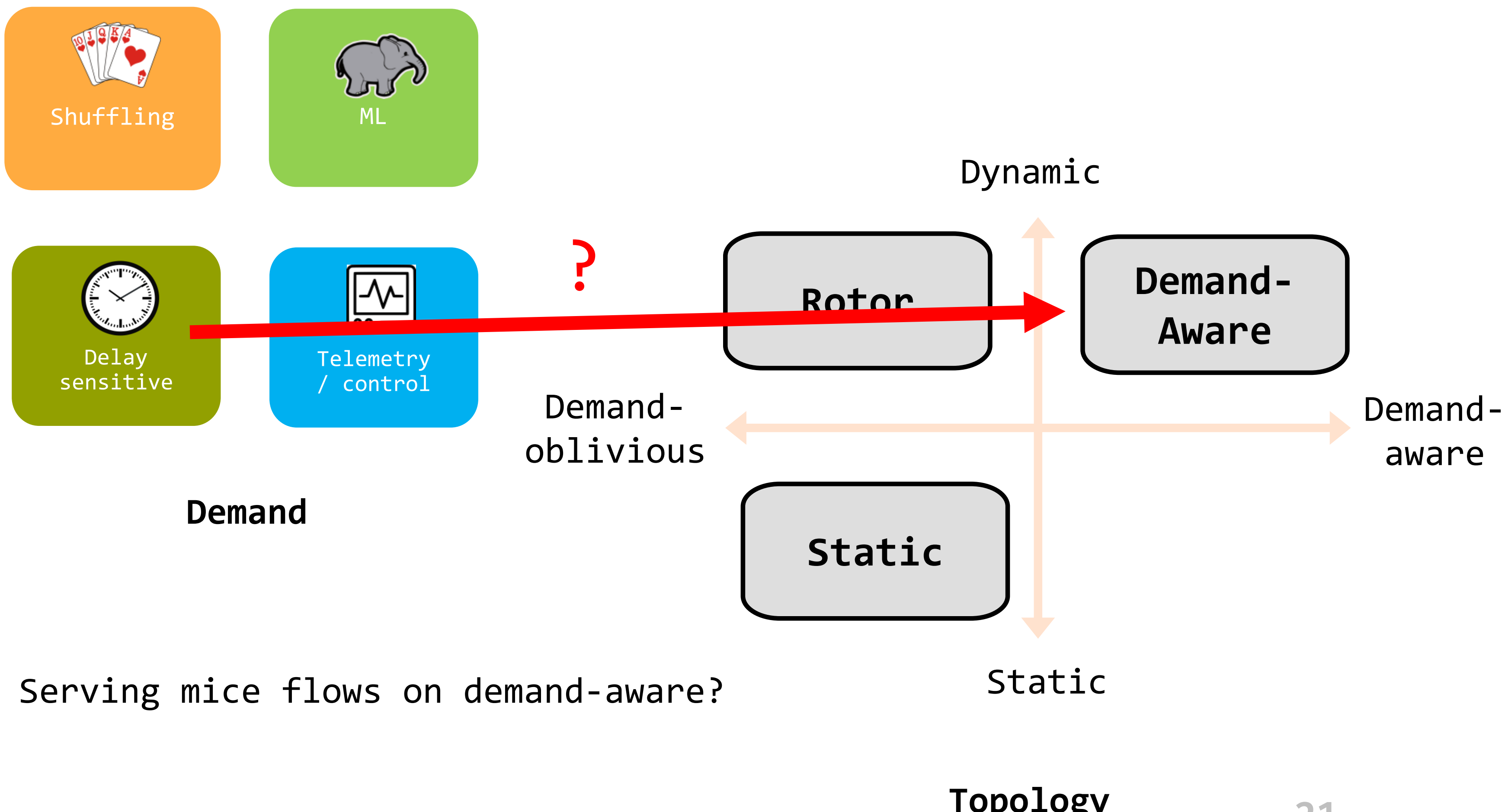


Demand



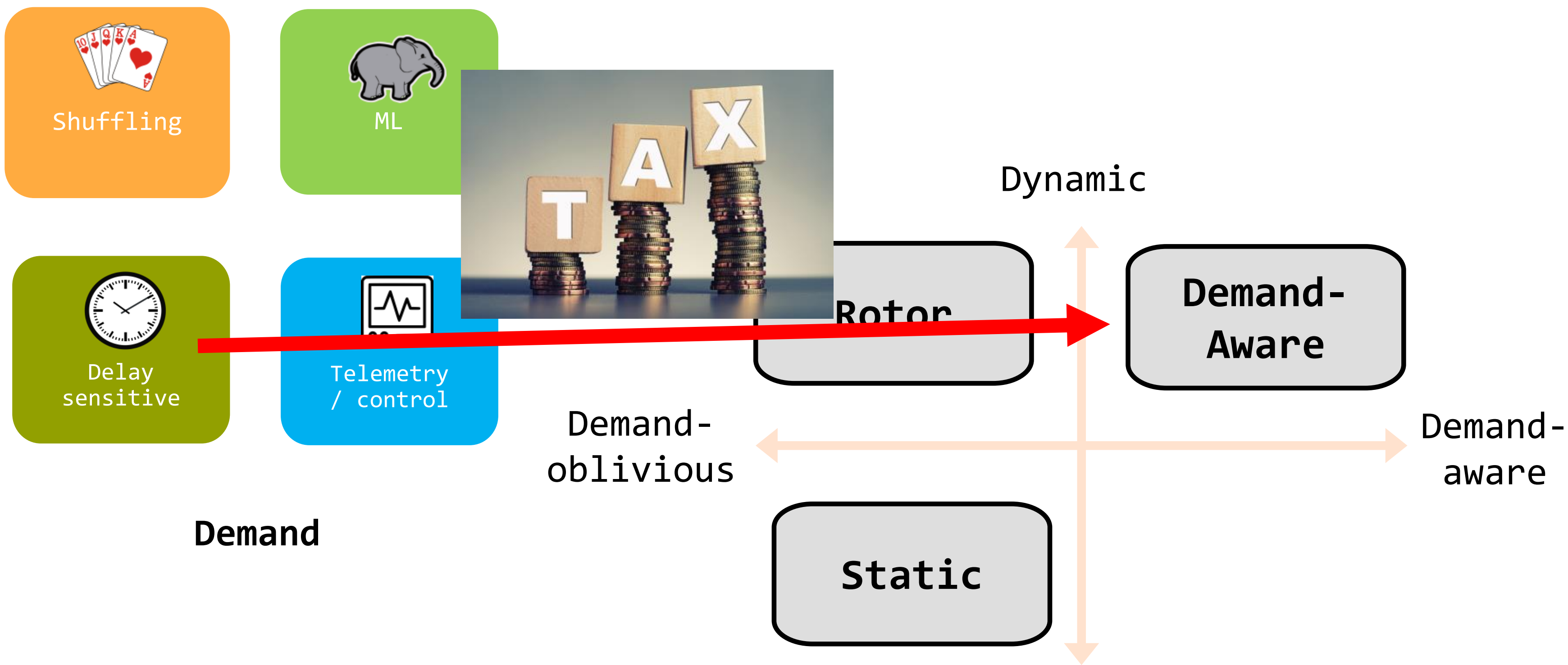
Topology

Examples:
Match or Mismatch?



Serving mice flows on demand-aware?

Examples:
Match or Mismatch?



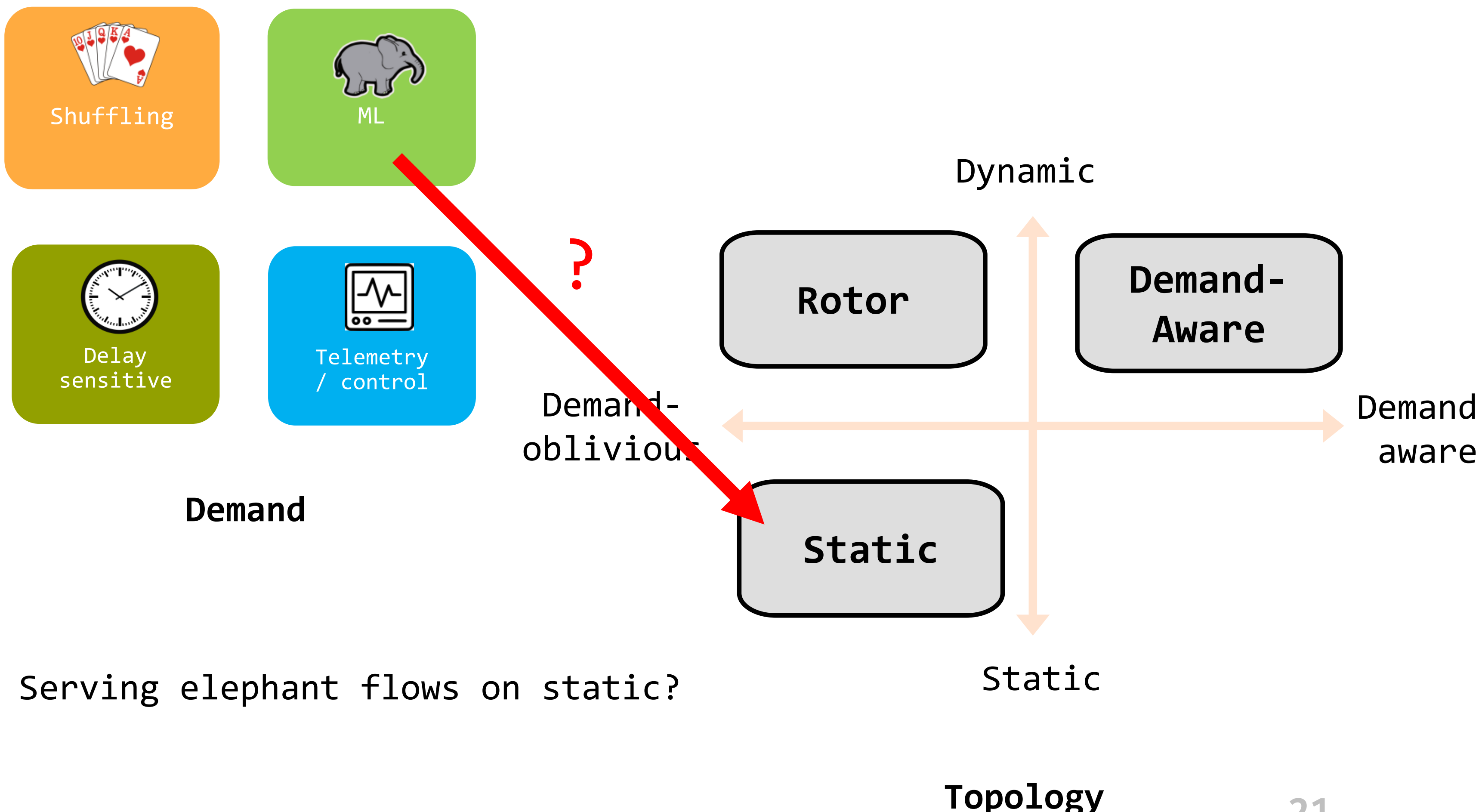
Serving mice flows on demand-aware?

Bad idea! Latency tax.

Static

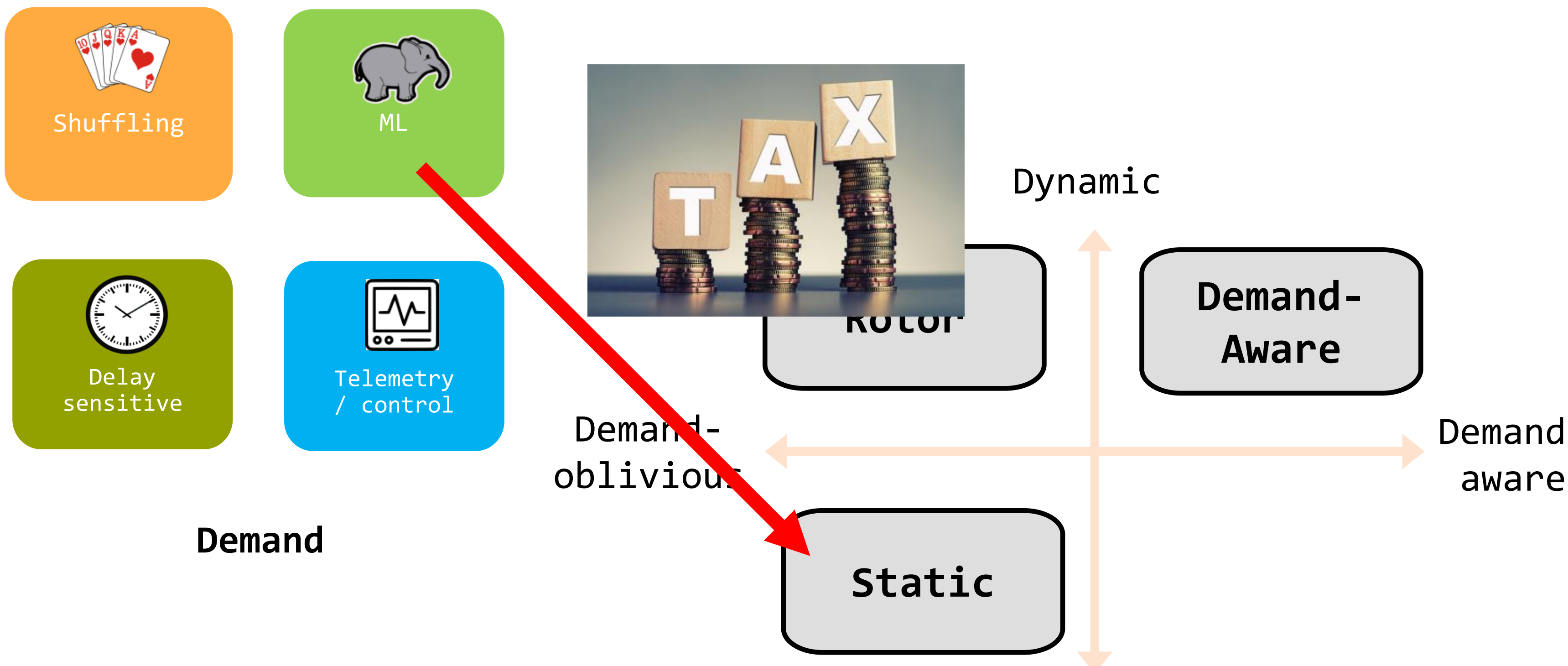
Topology

Examples:
Match or Mismatch?



Serving elephant flows on static?

Examples:
Match or Mismatch?



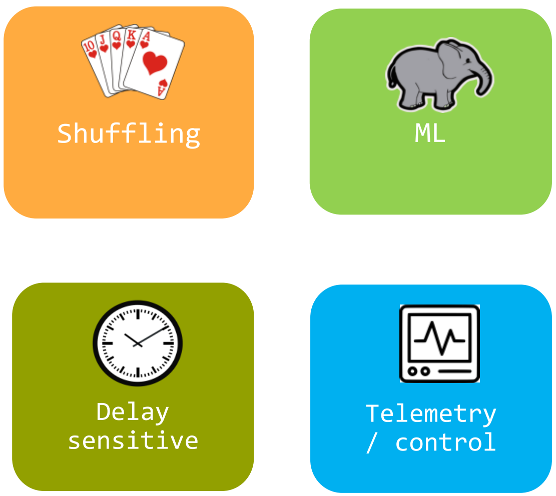
Static

Topology

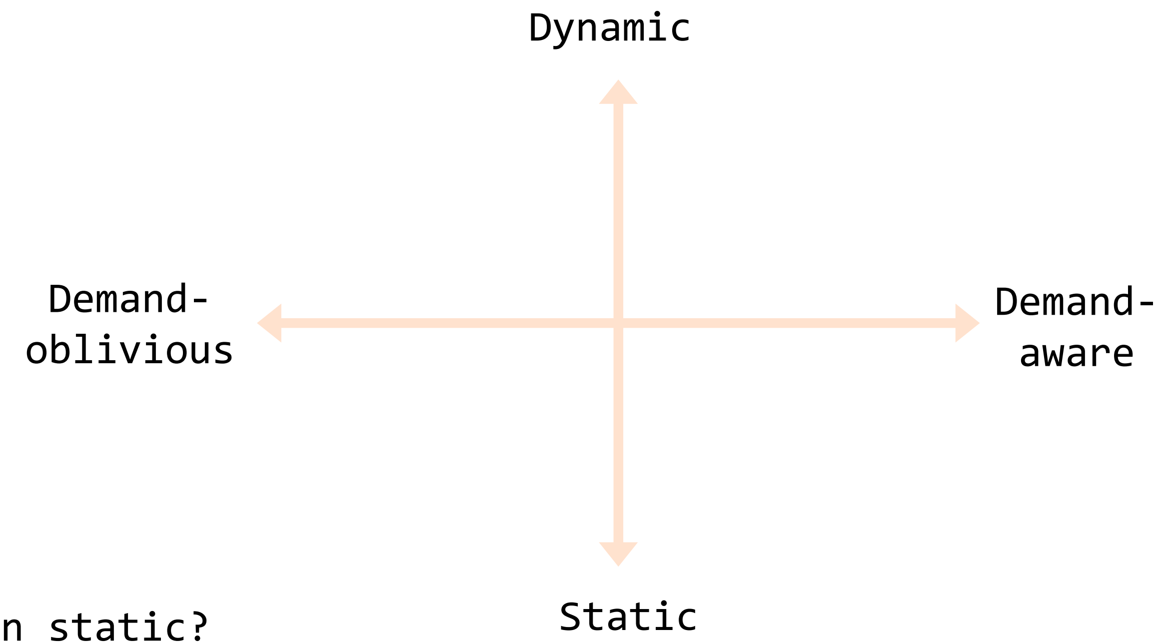
Serving elephant flows on static?

Bad idea! Bandwidth tax.

Examples:
Match or Mismatch?



Demand

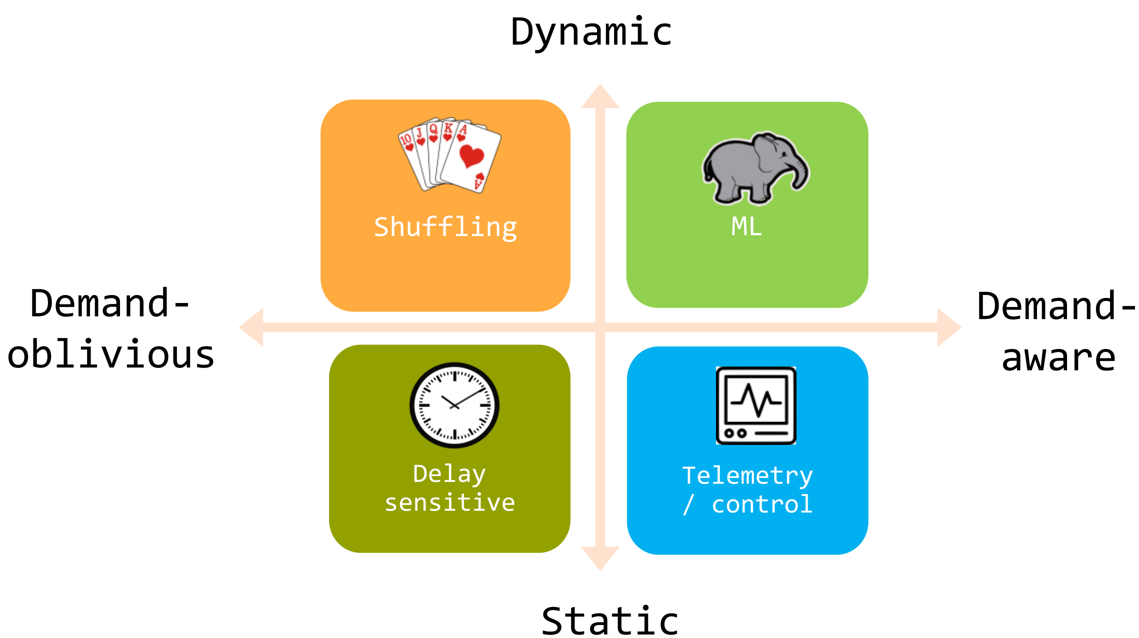


Serving elephant flows on static?

Bad idea! Bandwidth tax.

Topology

A First Guess



A first approach:

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

* Griner et al., ACM SIGMETRICS 2022

Conclusion

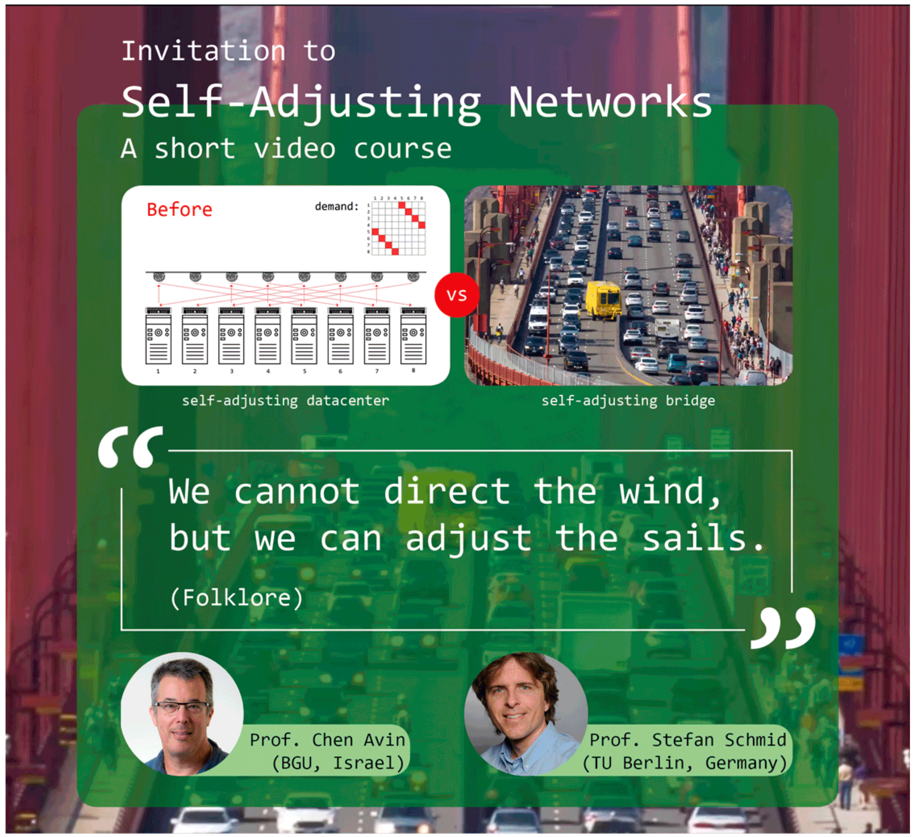
- Opportunity: *structure* in demand and *reconfigurable* networks
- So far: tip of the iceberg
- Many challenges
  - Optimal design depends on traffic pattern
  - How to *measure/predict* traffic?
  - Impact on other *layers*?
  - Routing and congestion control?
  - *Scalable control* plane
  - *Application-specific* self-adjusting networks?
- Many more *opportunities* for optical networks

YouTube Interview & CACM

Check out our **YouTube interviews**
on Reconfigurable Datacenter Networks:

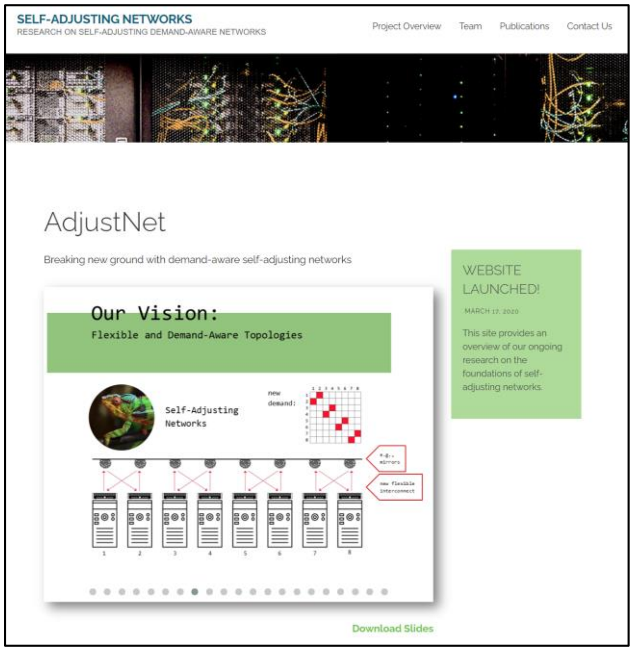
[Revolutionizing Datacenter Networks via Reconfigurable Topologies](https://www.youtube.com/@self-adjusting-networks-course)
Chen Avin and Stefan Schmid.
Communications of the ACM (**CACM**), 2025.
Watch here: https://www.youtube.com/@self-adjusting-networks-course

Online Video Course

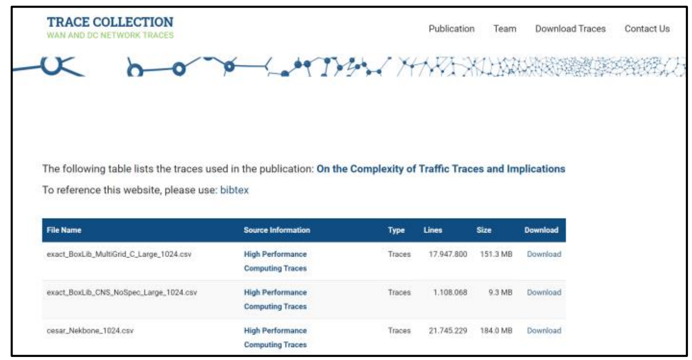


https://self-adjusting.net/course

Websites



http://self-adjusting.net/
Project website



https://trace-collection.net/
Trace collection website

Upcoming CACM Article

Revolutionizing Datacenter Networks via Reconfigurable Topologies

CHEN AVIN, is a Professor at Ben-Gurion University of the Negev, Beersheva, Israel
STEFAN SCHMID, is a Professor at TU Berlin, Berlin, Germany

With the popularity of cloud computing and data-intensive applications such as machine learning, datacenter networks have become a critical infrastructure for our digital society. Given the explosive growth of datacenter traffic and the slowdown of Moore's law, significant efforts have been made to improve datacenter network performance over the last decade. A particularly innovative solution is reconfigurable datacenter networks (RDCNs): datacenter networks whose topologies dynamically change over time, in either a demand-oblivious or a demand-aware manner. Such dynamic topologies are enabled by recent optical switching technologies and stand in stark contrast to state-of-the-art datacenter network topologies, which are fixed and oblivious to the actual traffic demand. In particular, reconfigurable demand-aware and "self-adjusting" datacenter networks are motivated empirically by the significant spatial and temporal structures observed in datacenter communication traffic. This paper presents an overview of reconfigurable datacenter networks. In particular, we discuss the motivation for such reconfigurable architectures, review the technological enablers, and present a taxonomy that classifies the design space into two dimensions: static vs. dynamic and demand-oblivious vs. demand-aware. We further present a formal model and discuss related research challenges. Our article comes with complementary video interviews in which three leading experts, Manya Ghobadi, Amin Vahdat, and George Papen, share with us their perspectives on reconfigurable datacenter networks.

KEY INSIGHTS

- **Datacenter networks have become a critical infrastructure** for our digital society, serving explosively growing communication traffic.
- **Reconfigurable datacenter networks (RDCNs)** which can adapt their topology dynamically, based on innovative **optical switching technologies**, bear the potential to improve datacenter network performance, and to simplify datacenter planning and operations.
- Demand-aware dynamic topologies are particularly interesting because of the **significant spatial and temporal structures** observed in real-world traffic, e.g., related to distributed machine learning.
- The study of RDCNs and self-adjusting networks raises many **novel technological and research challenges** related to their design, control, and performance.

References (1)

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.

Revolutionizing Datacenter Networks via Reconfigurable Topologies
Chen Avin and Stefan Schmid.
Communications of the ACM (**CACM**), 2025.

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.

References (2)

Mars: Near-Optimal Throughput with Shallow Buffers in Reconfigurable Datacenter Networks
Vamsi Addanki, Chen Avin, and Stefan Schmid.
ACM **SIGMETRICS** and ACM Performance Evaluation Review (**PER**), Orlando, Florida, USA, June 2023.

Duo: A High-Throughput Reconfigurable Datacenter Network Using Local Routing and Control
Johannes Zerwas, Csaba Györgyi, Andreas Blenk, Stefan Schmid, and Chen Avin.
ACM **SIGMETRICS** and ACM Performance Evaluation Review (**PER**), Orlando, Florida, USA, June 2023.

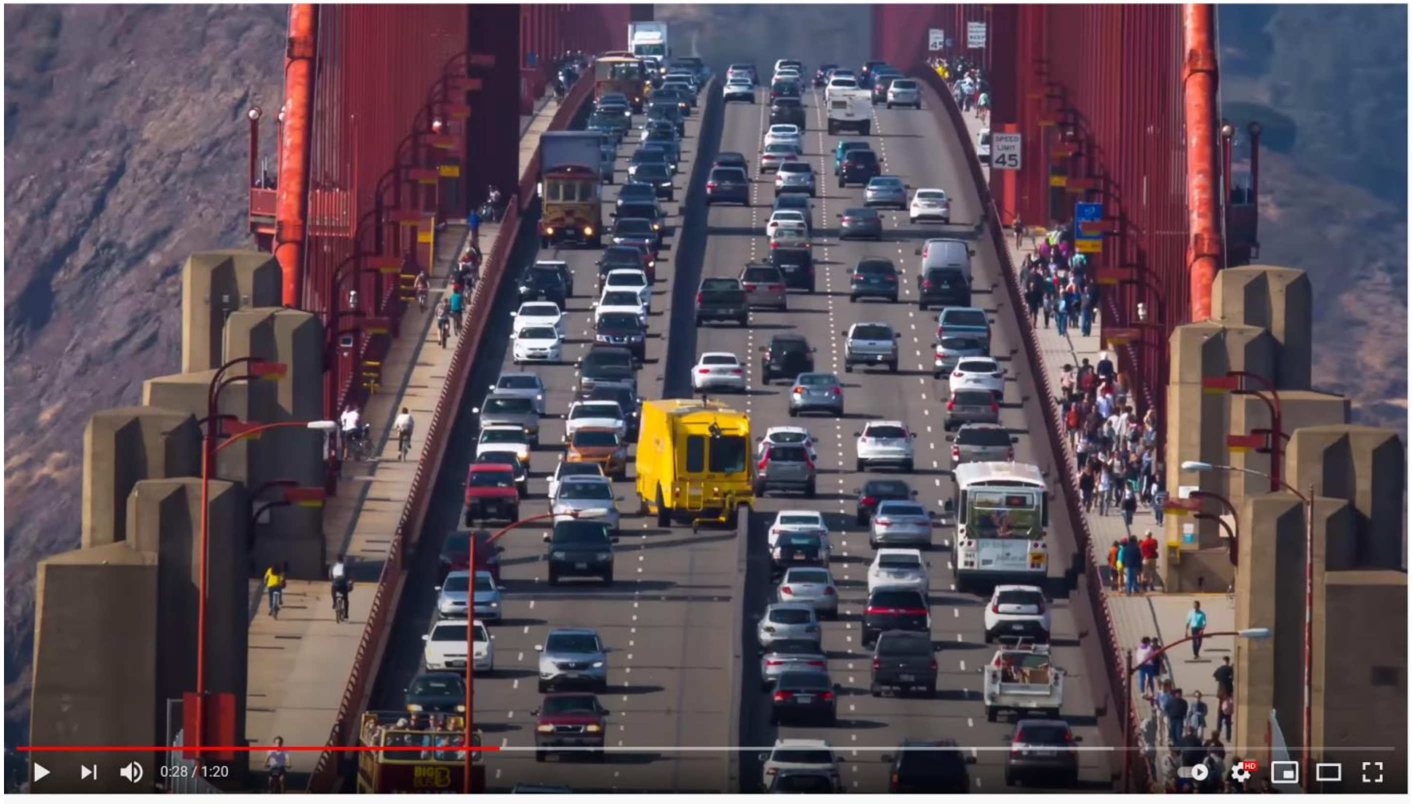
SyPer: Synthesis of Perfectly Resilient Local Fast Rerouting Rules for Highly Dependable Networks
Csaba Györgyi, Kim G. Larsen, Stefan Schmid, and Jiri Srba.
IEEE Conference on Computer Communications (**INFOCOM**), Vancouver, Canada, May 2024.

Demand-Aware Network Design with Minimal Congestion and Route Lengths
Chen Avin, Kaushik Mondal, and Stefan Schmid.
IEEE/ACM Transactions on Networking (**TON**), 2022.

A Survey of Reconfigurable Optical Networks
Matthew Nance Hall, Klaus-Tycho Foerster, Stefan Schmid, and Ramakrishnan Durairajan.
Optical Switching and Networking (**OSN**), Elsevier, 2021.

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.
.

Questions?



Slides available here:

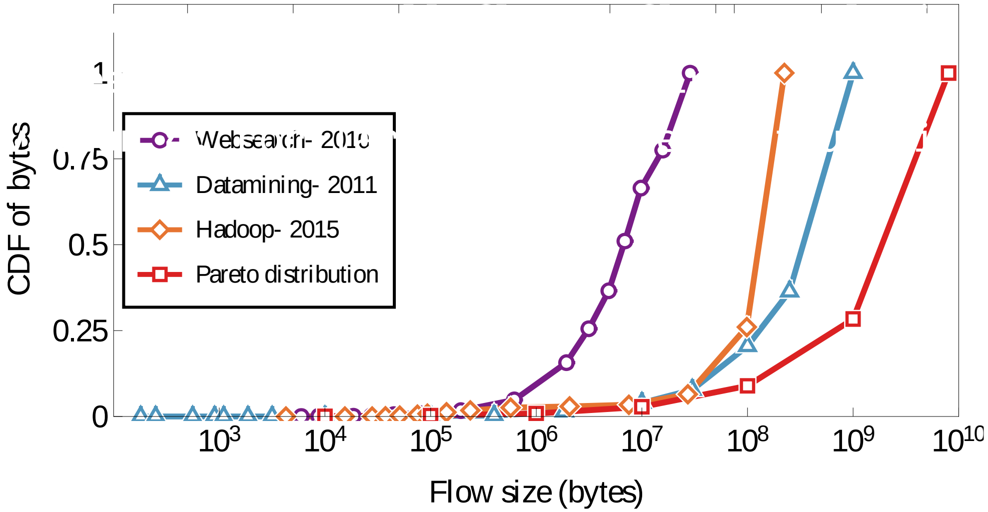
Backup

Flow Size Matters

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

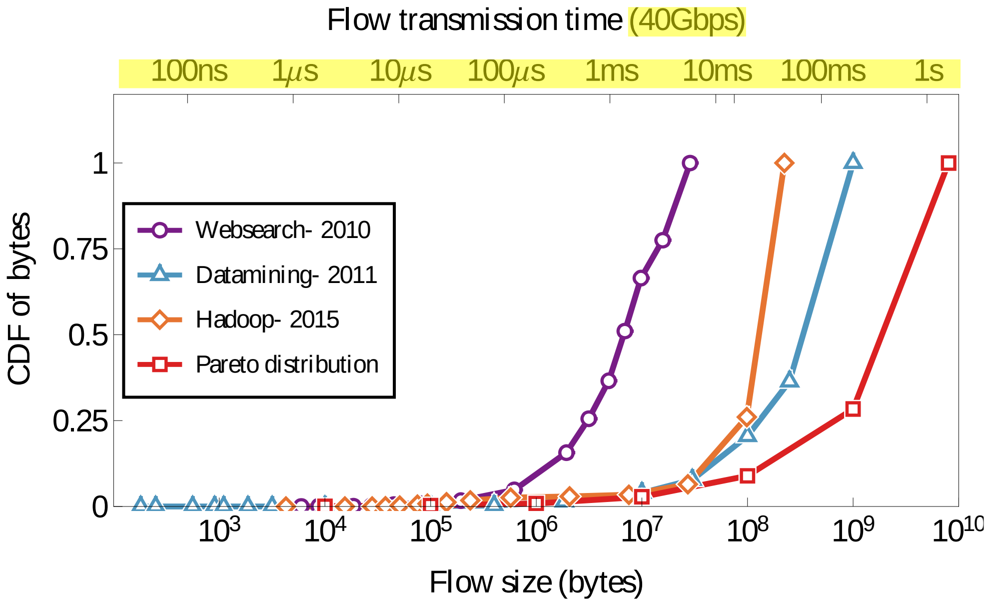
Flow Size Matters

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



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

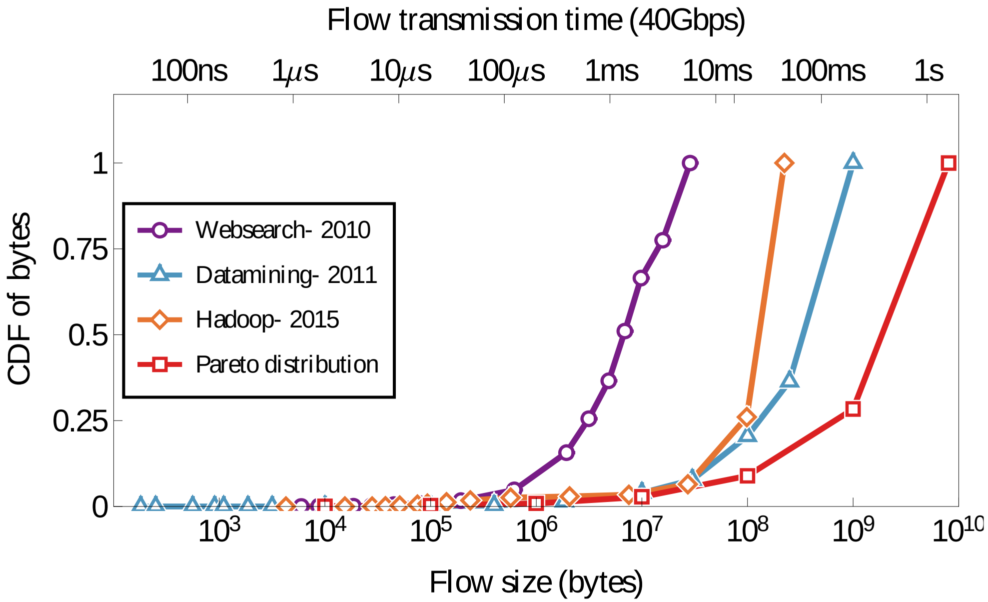
Flow Size Matters



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

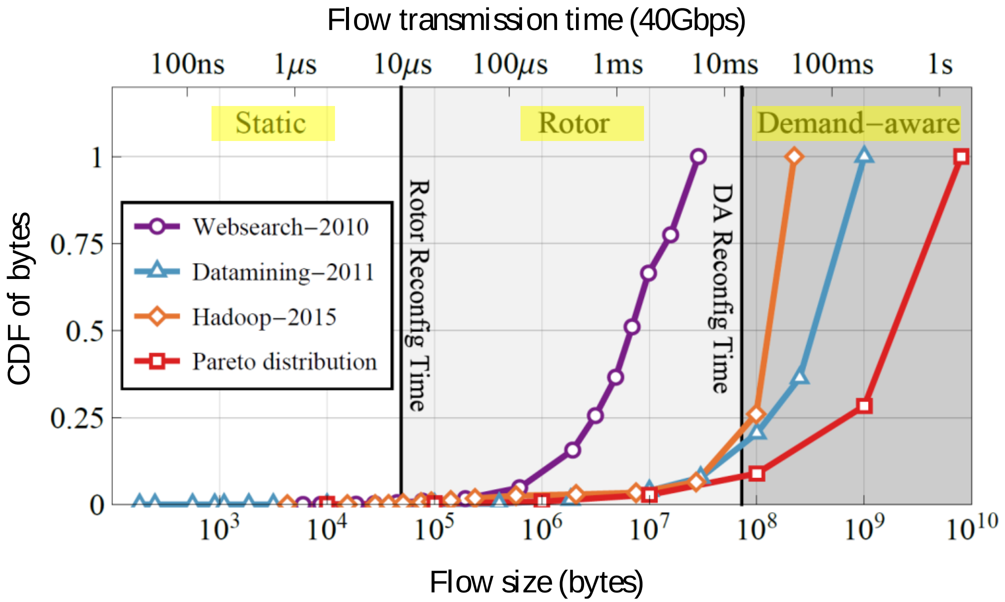
⇢ **Observation 2:** The transmission time of a flow depends on its size.

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



⇢ **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.

Excursion

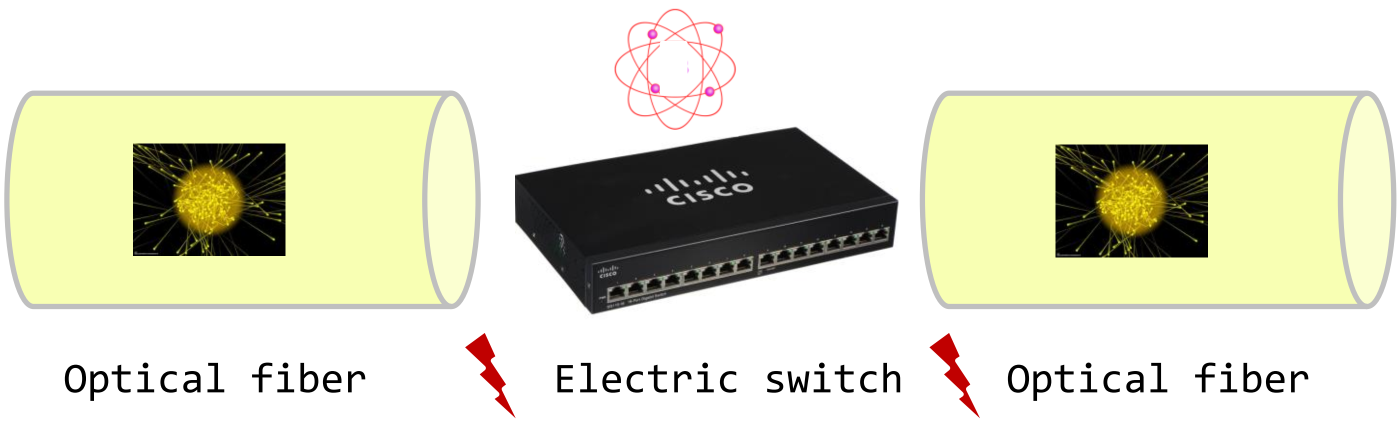
More benefits of optical & reconfigurable switching

So far: focus on throughput performance.

Benefit 1:

Energy and Latency

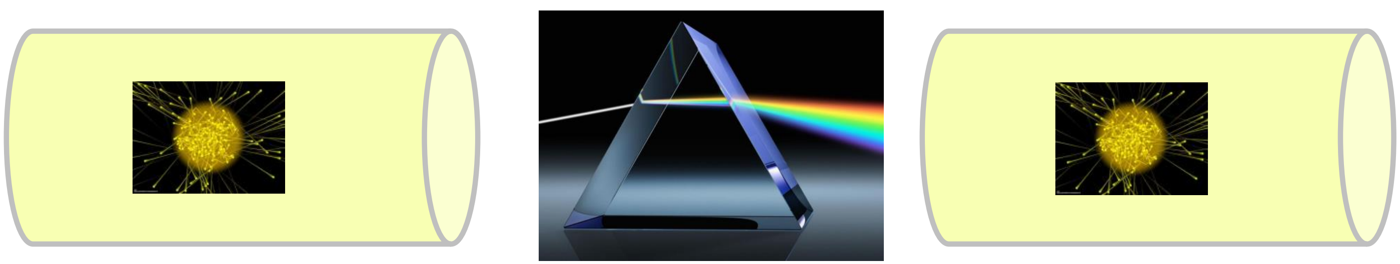
- No need to *convert* photons in fiber to electrons in switch (and back)
- Can safe *energy* and reduce *latency* (in addition to enabling almost unlimited throughput)



Benefit 1:

Energy and Latency

- No need to *convert* photons in fiber to electrons in switch (and back)
- Can safe *energy* and reduce *latency* (in addition to enabling almost unlimited throughput)

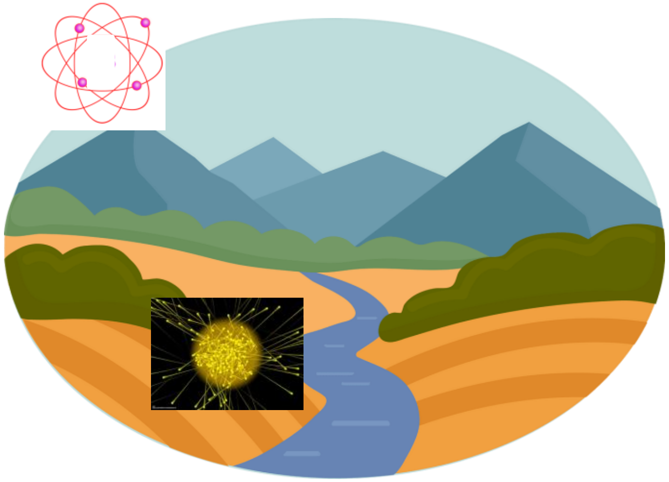


Optical fiber —— Optical switch —— Optical fiber

Benefit 2:

Resilience

Floodings in South Germany destroyed much electrical network infrastructure



Solution: deploy optical infrastructure (in valleys) and electrical *on hills* where safe?

Benefit 3:

Evolving Datacenters

- Reconfigurable datacenter networks naturally support *heterogeneous* network elements
- And therefore also *incremental* hardware upgrades

Amin Vahdat
Google