Self-Adjusting Networks

Stefan Schmid (TU Berlin)

“We cannot direct the wind,
but we can adjust the sails.”

(Folklore)

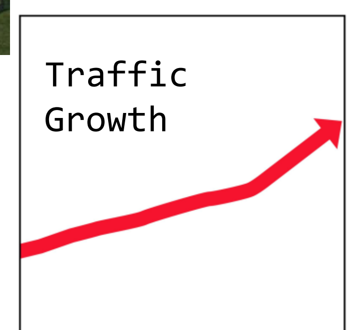
Trend

Data-Centric Applications

Datacenters ("hyper-scale")

+network

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



Source: Facebook

Trend

Data-Centric Applications

NETFLIX

Datacenters ("hyper-scale")

+network

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

Credits: Marco Chiesa

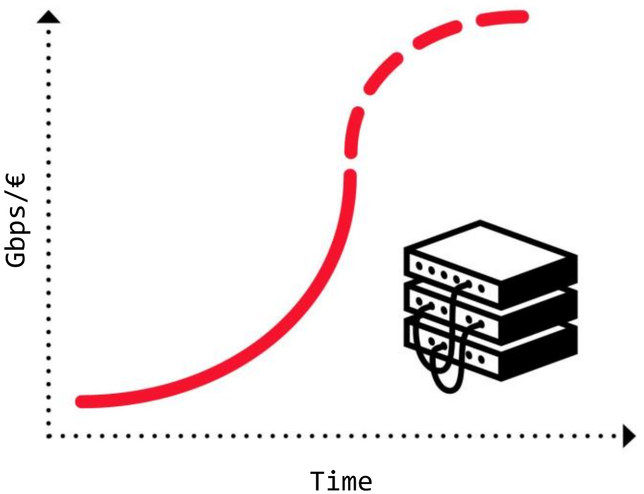
The Problem

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**



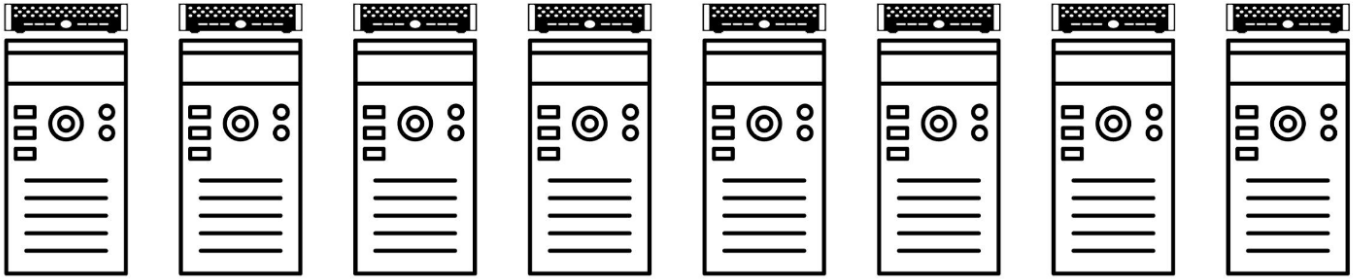
[1] Source: Microsoft, 2019

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

Root Cause

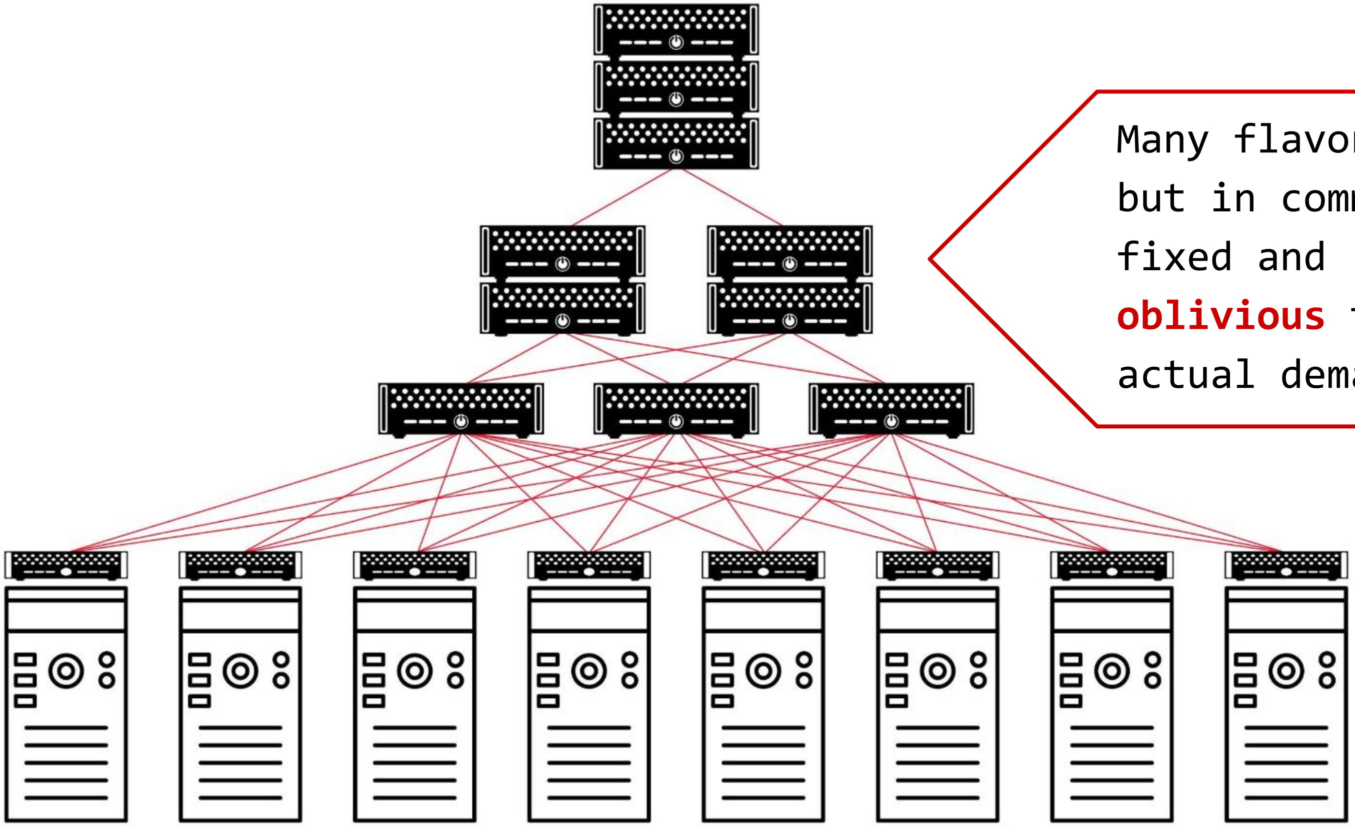
Fixed and Demand-Oblivious Topology

How to interconnect?



Root Cause

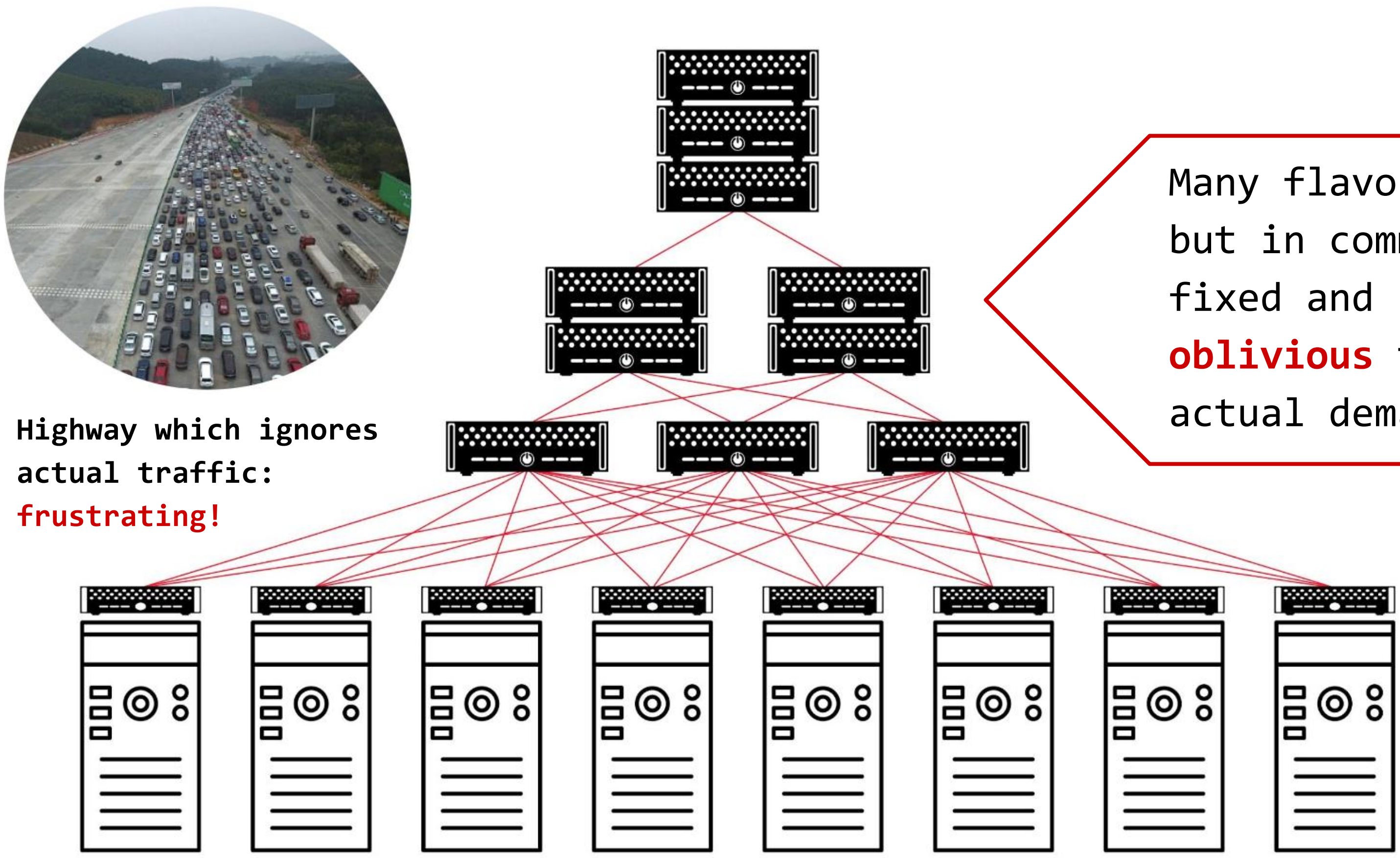
Fixed and Demand-Oblivious Topology



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

Root Cause

Fixed and Demand-Oblivious Topology

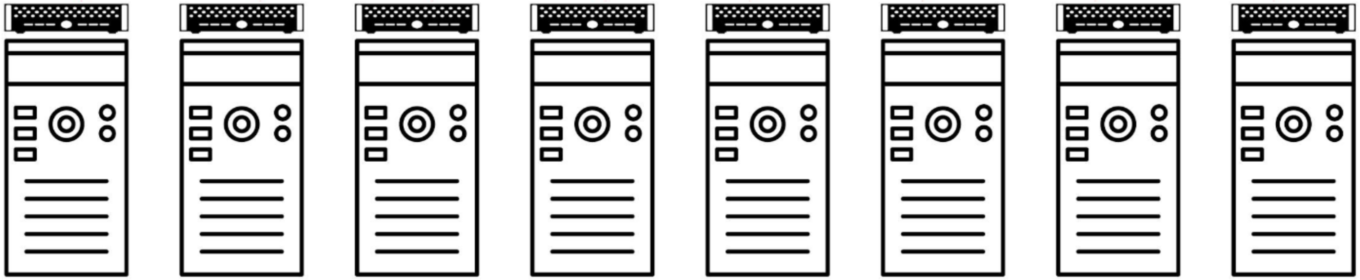


Highway which ignores actual traffic: **frustrating!**

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

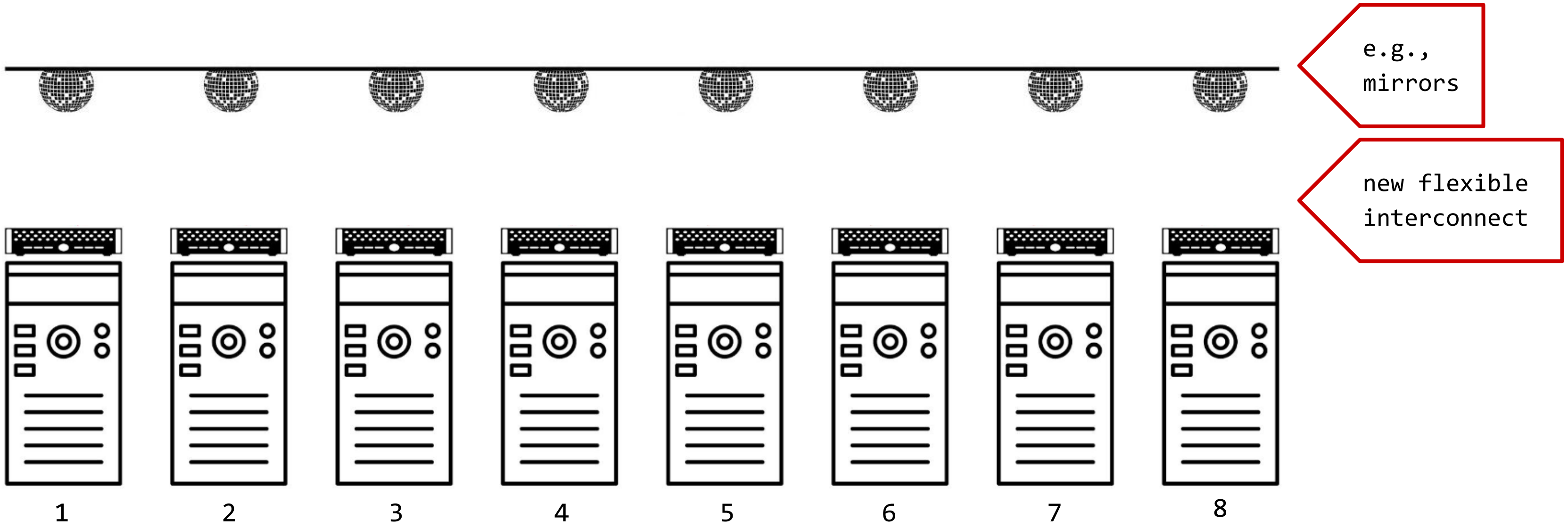
A Vision

Flexible and Demand-Aware Topologies



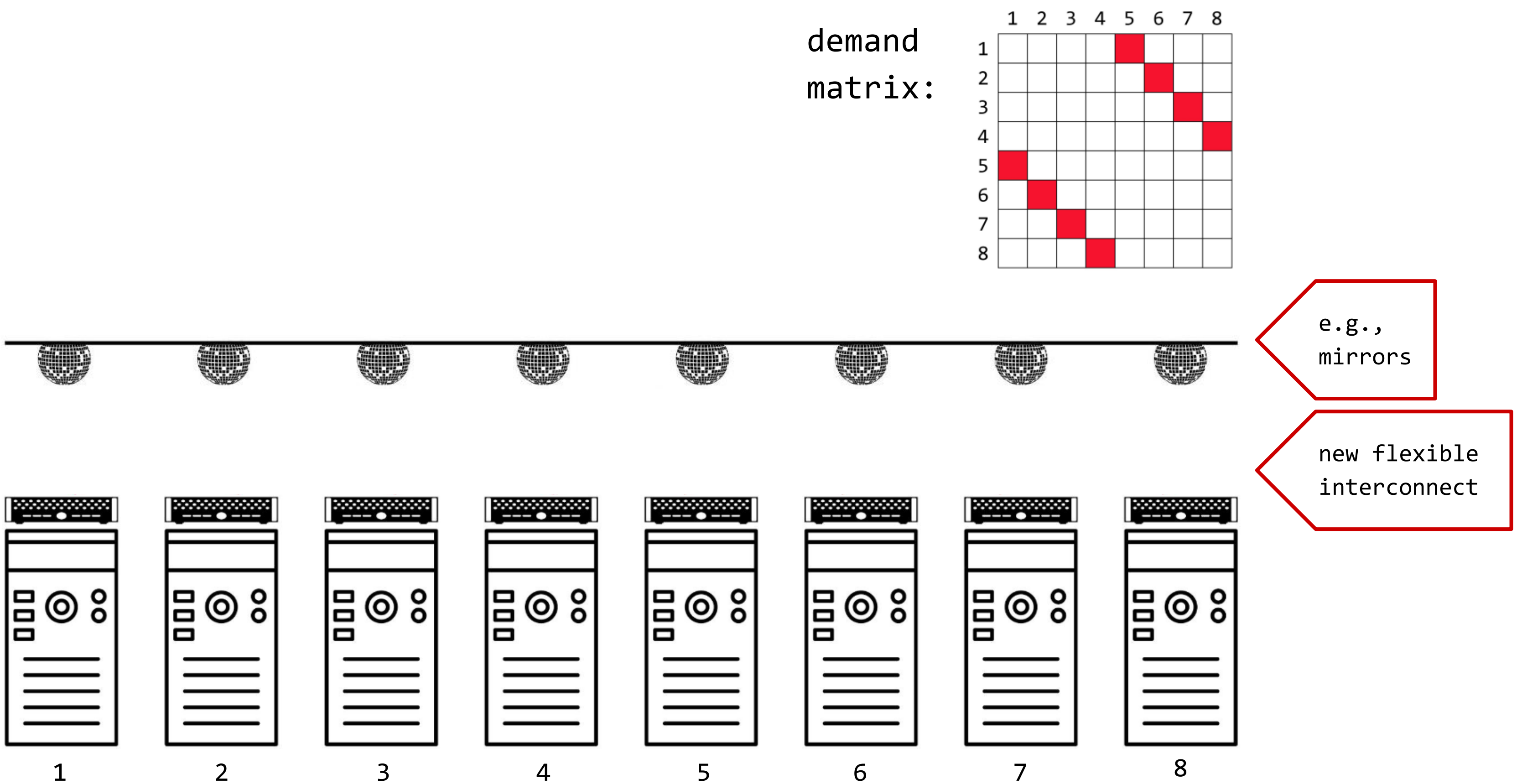
A Vision

Flexible and Demand-Aware Topologies



A Vision

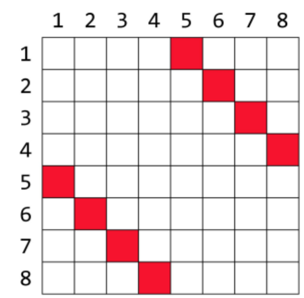
Flexible and Demand-Aware Topologies



A Vision

Flexible and Demand-Aware Topologies

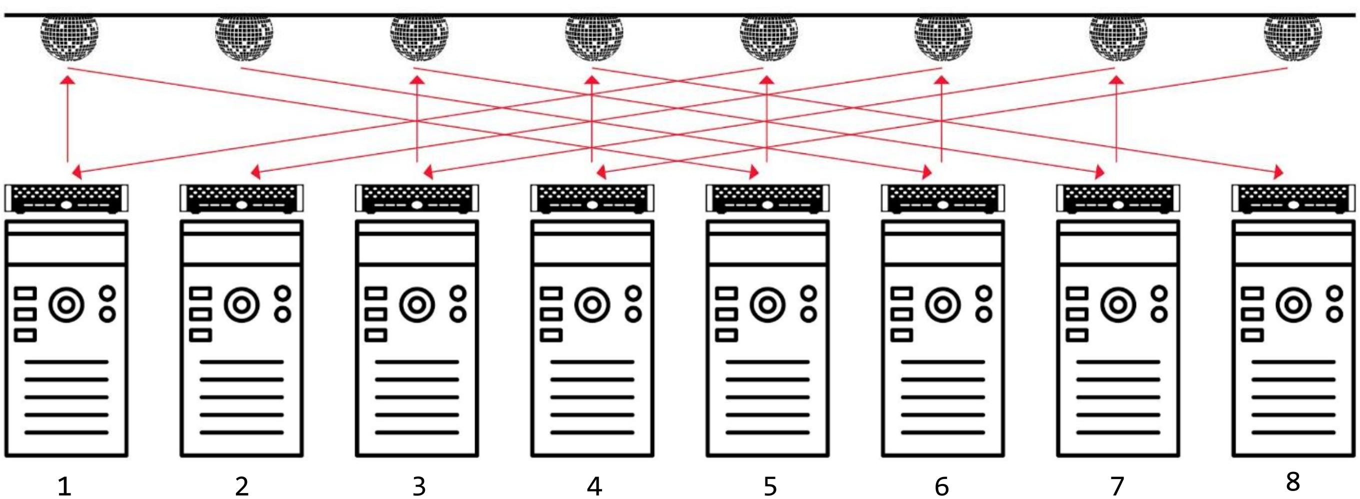
demand matrix:



Matches demand

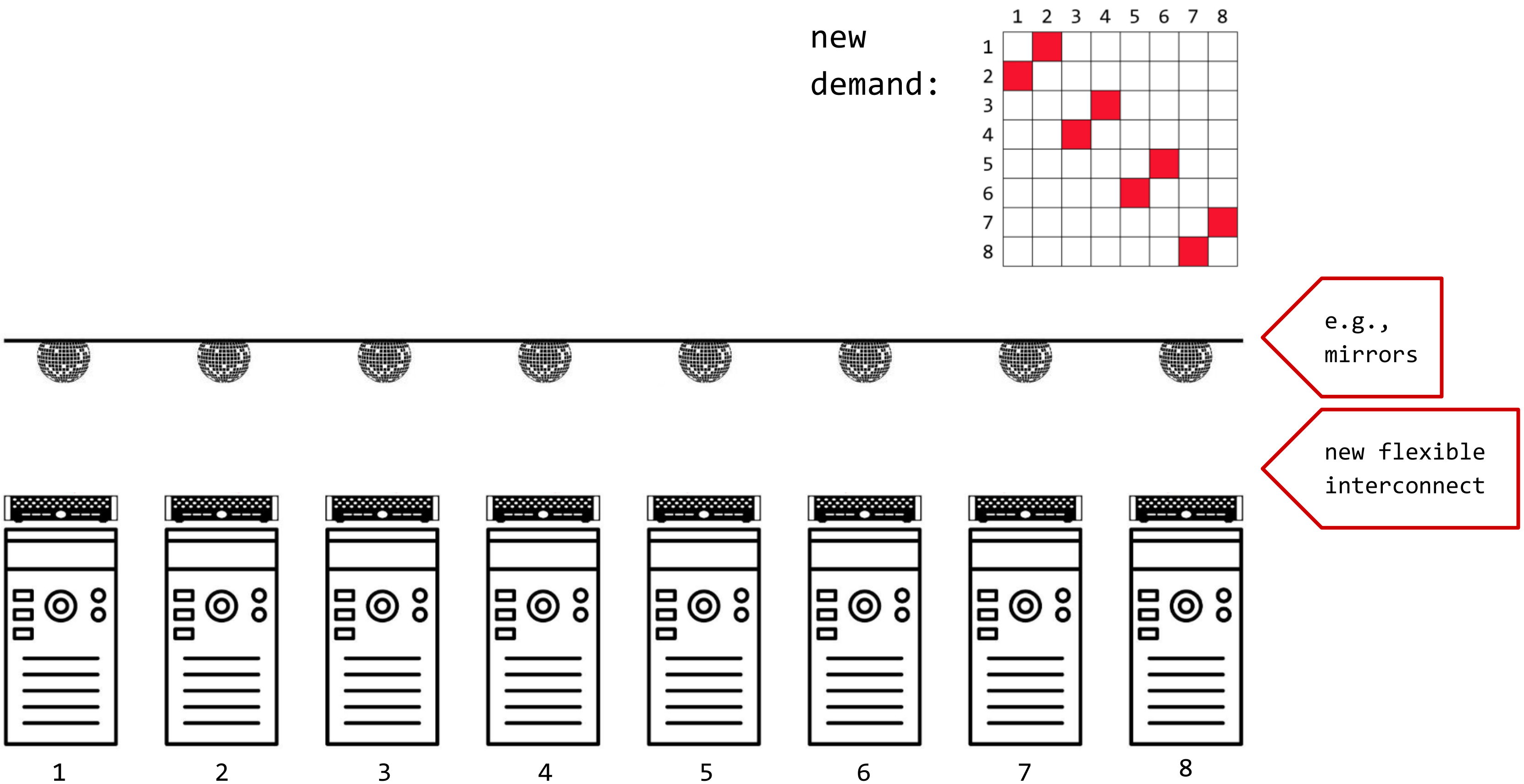
e.g., mirrors

new flexible interconnect



A Vision

Flexible and Demand-Aware Topologies

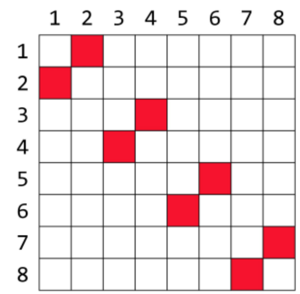


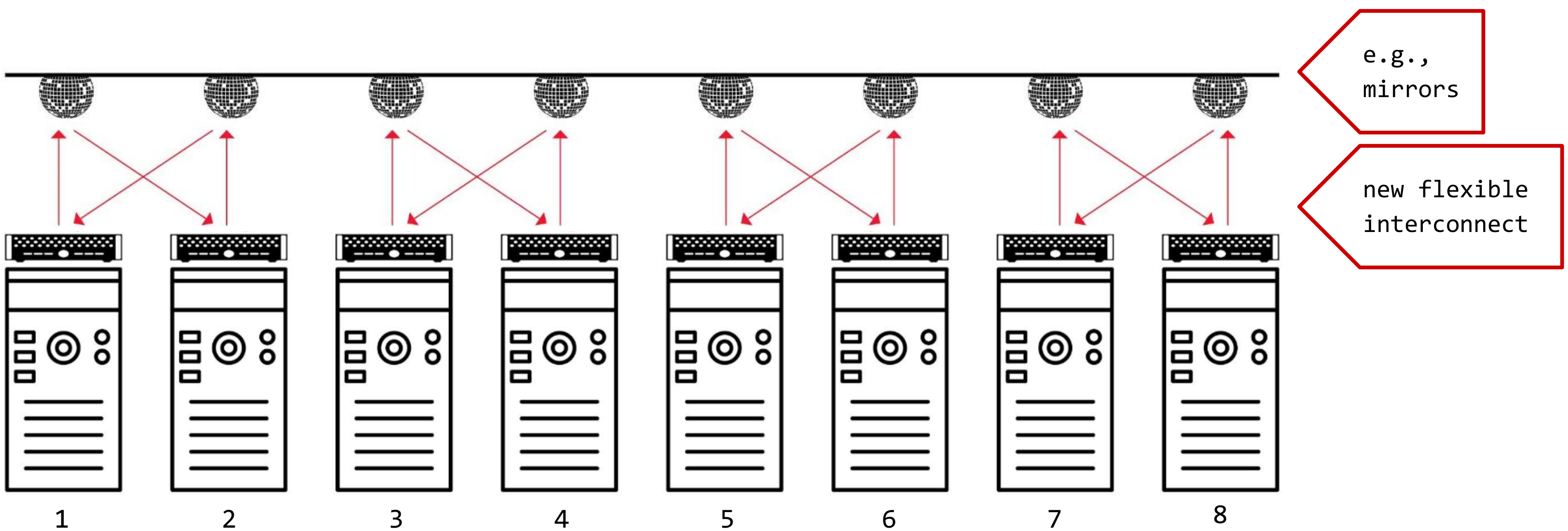
A Vision

Flexible and Demand-Aware Topologies

Matches demand

new demand:



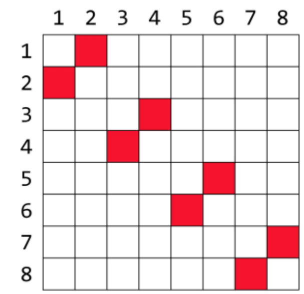


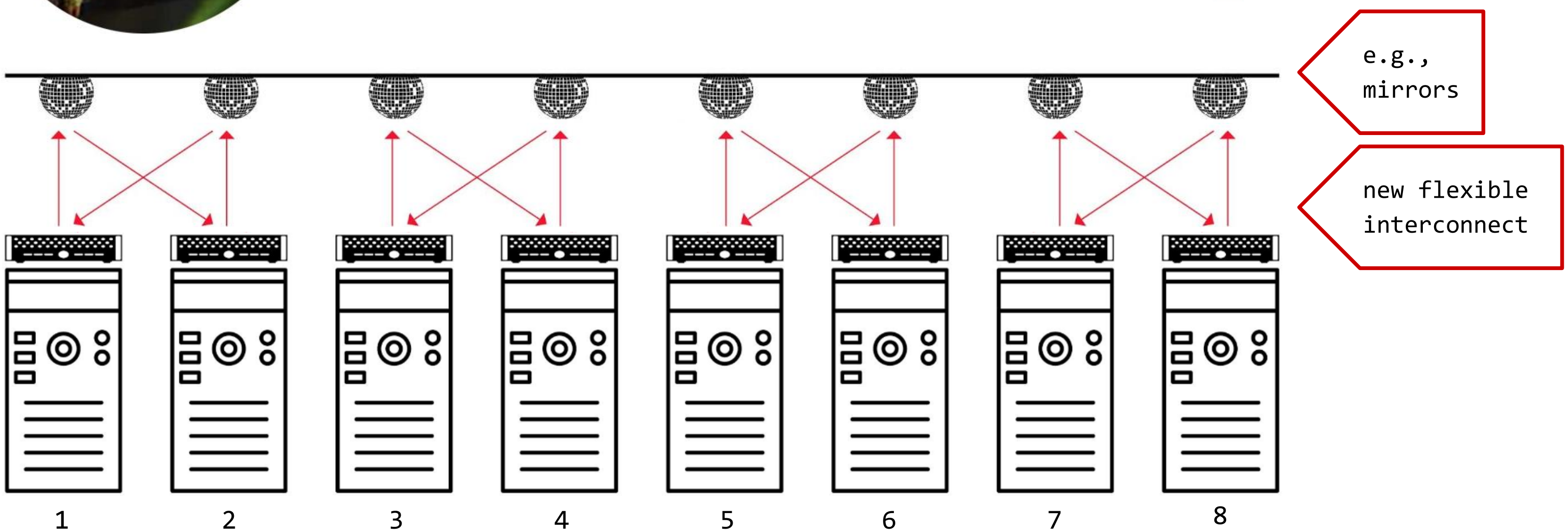
A Vision

Flexible and Demand-Aware Topologies

Self-Adjusting
Networks

new
demand:



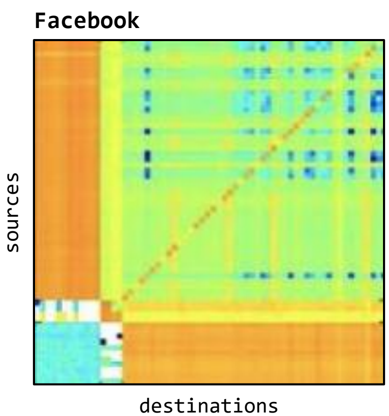


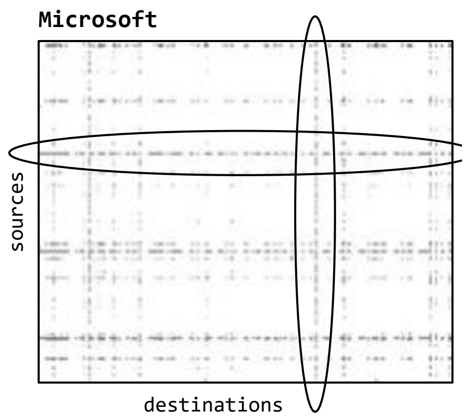
The Motivation

Much Structure in the Demand

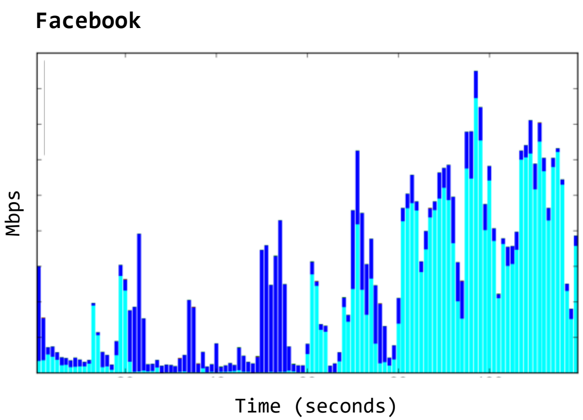
Empirical studies:

traffic matrices sparse and skewed





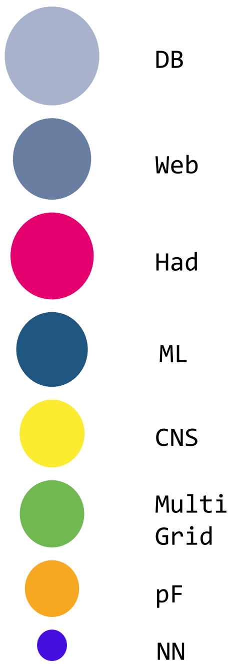
traffic bursty over time



The hypothesis: can be exploited.

Recent Representation of Trace Structure:

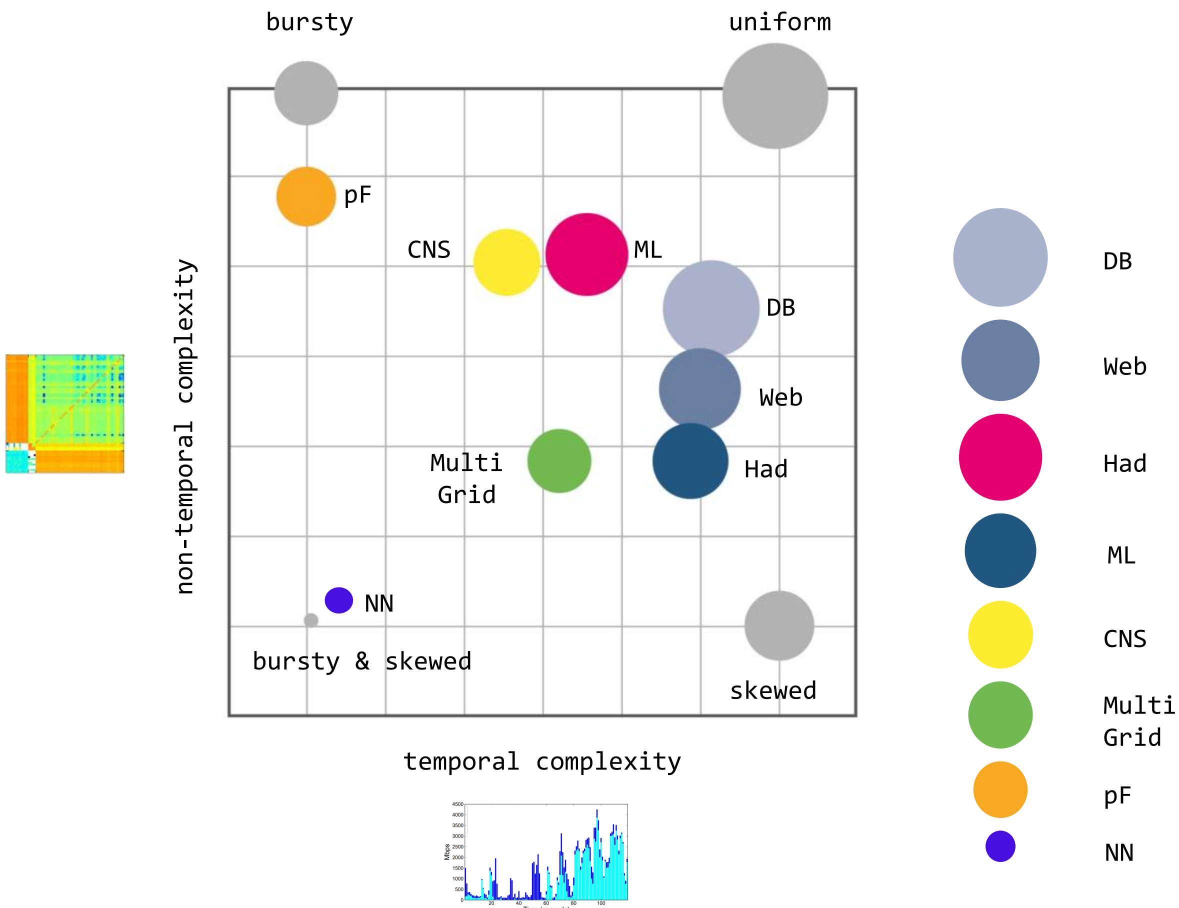
Complexity Map



Griner et al., SIGMETRICS 2020

Recent Representation of Trace Structure:

Complexity Map

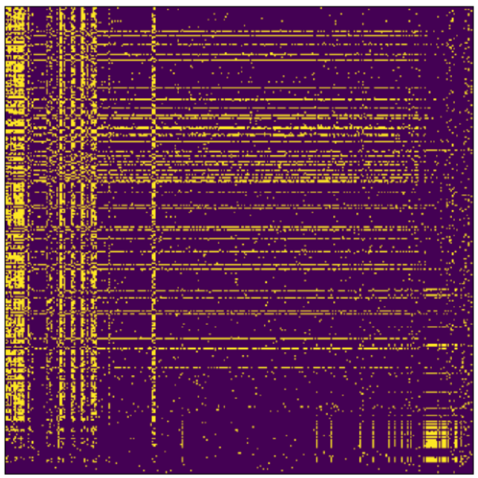


Different structures!

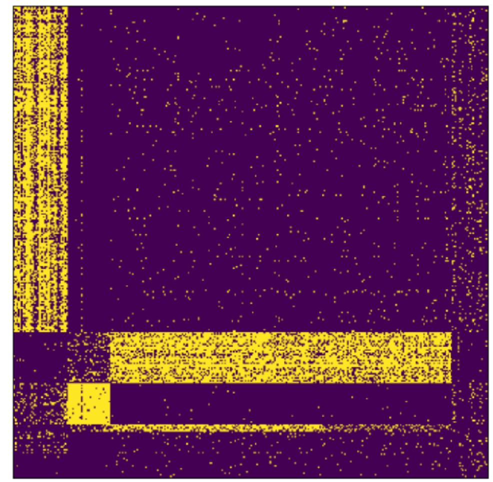
Griner et al., SIGMETRICS 2020

Traffic is also clustered:

Small Stable Clusters



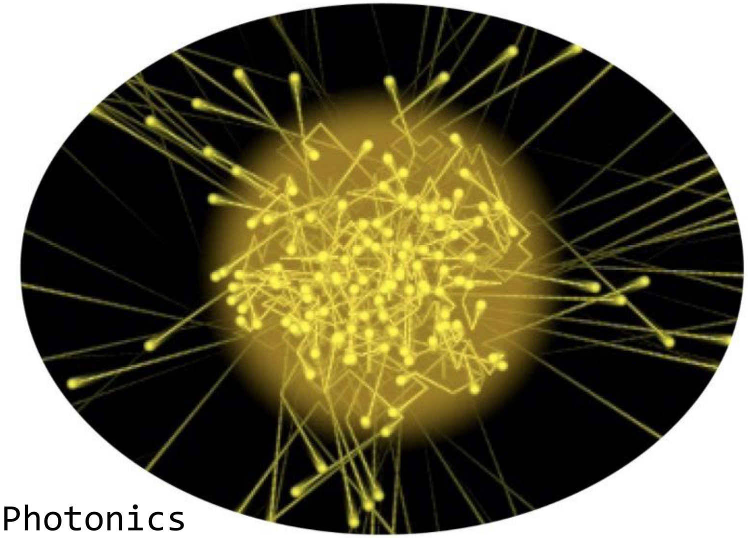
reordering based on
bicluster structure



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.”

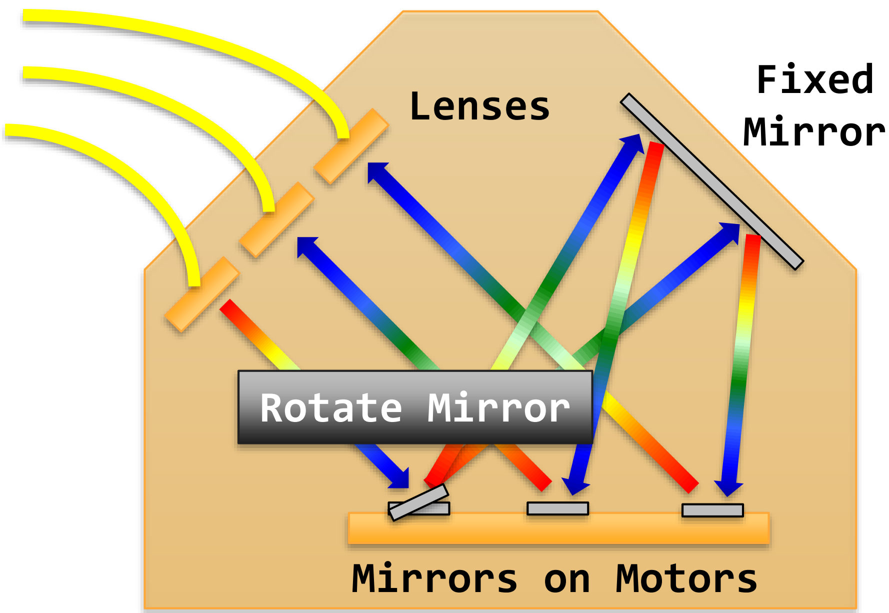
US National Research Council:

“Photons are the new Electrons.”

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

Systems

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

August 24, 2022

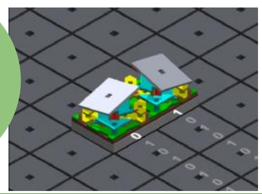
Amin Vahdat

VP & GM, Systems and Services Infrastructure

The Big Picture

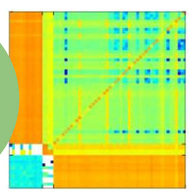
New!

Flexibility



More!

Structure



Self-Adjusting Networks

Efficiency

Now is the time!

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

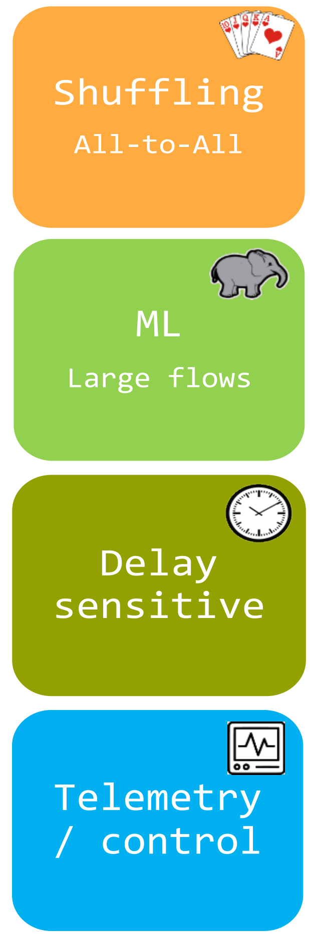
⇀ Query traffic: skewed

⇀ 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

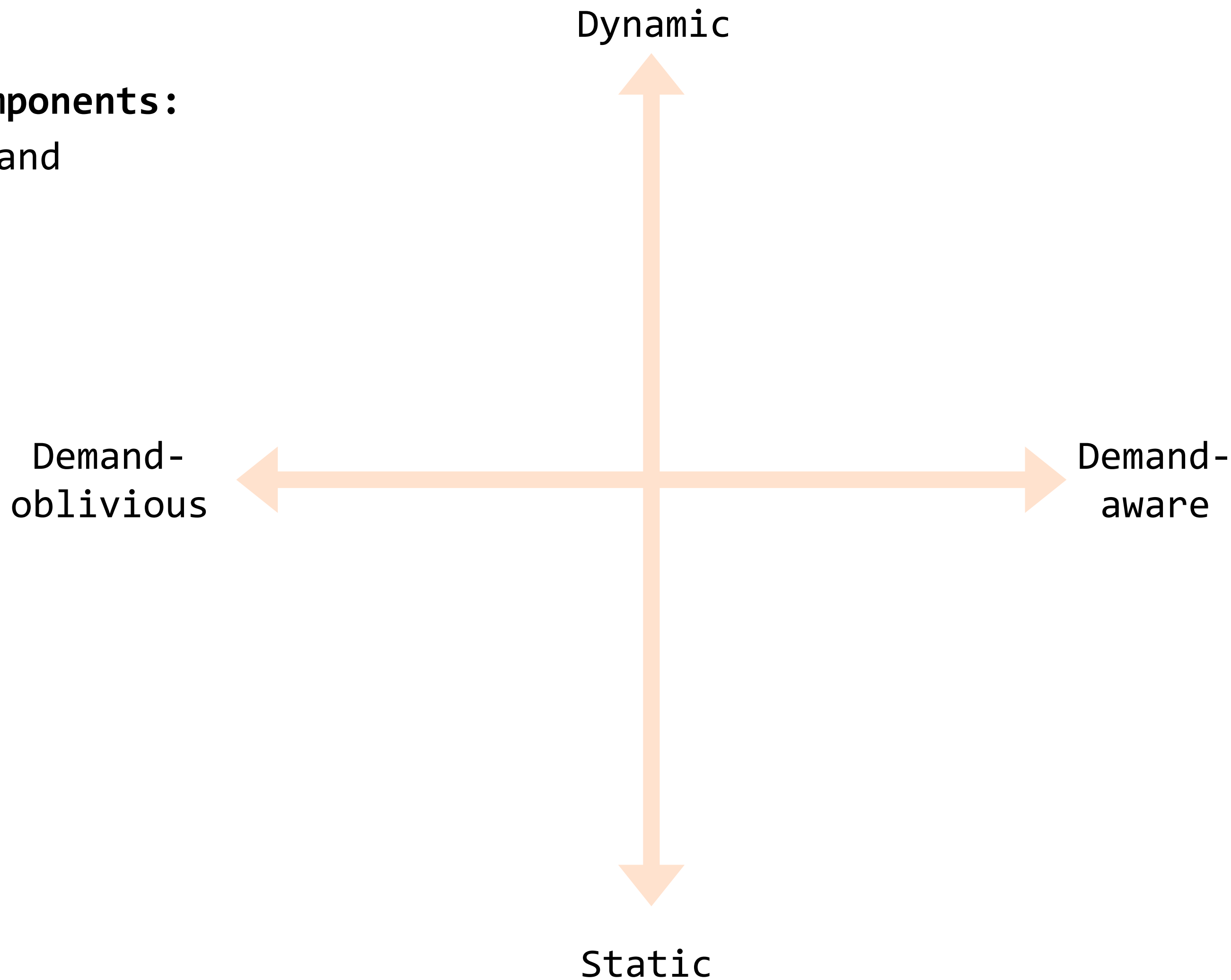
Demand-oblivious ⟷ Demand-aware

Opportunity: Tech Diversity

Diverse topology components:

⇀ demand-oblivious and
  demand-aware

⇀ static vs dynamic

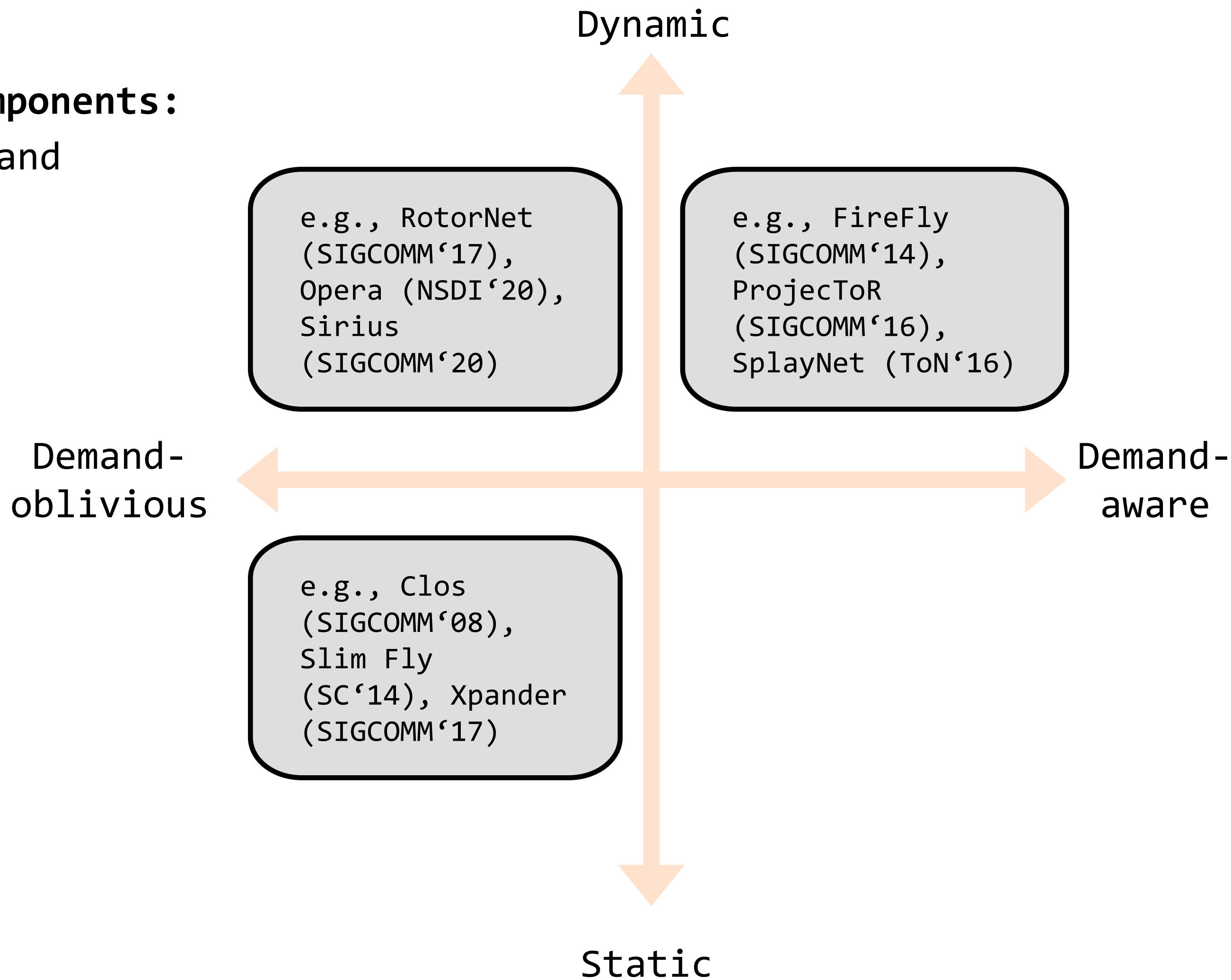


Opportunity: Tech Diversity

Diverse topology components:

⇀ demand-oblivious and
demand-aware

⇀ static vs dynamic

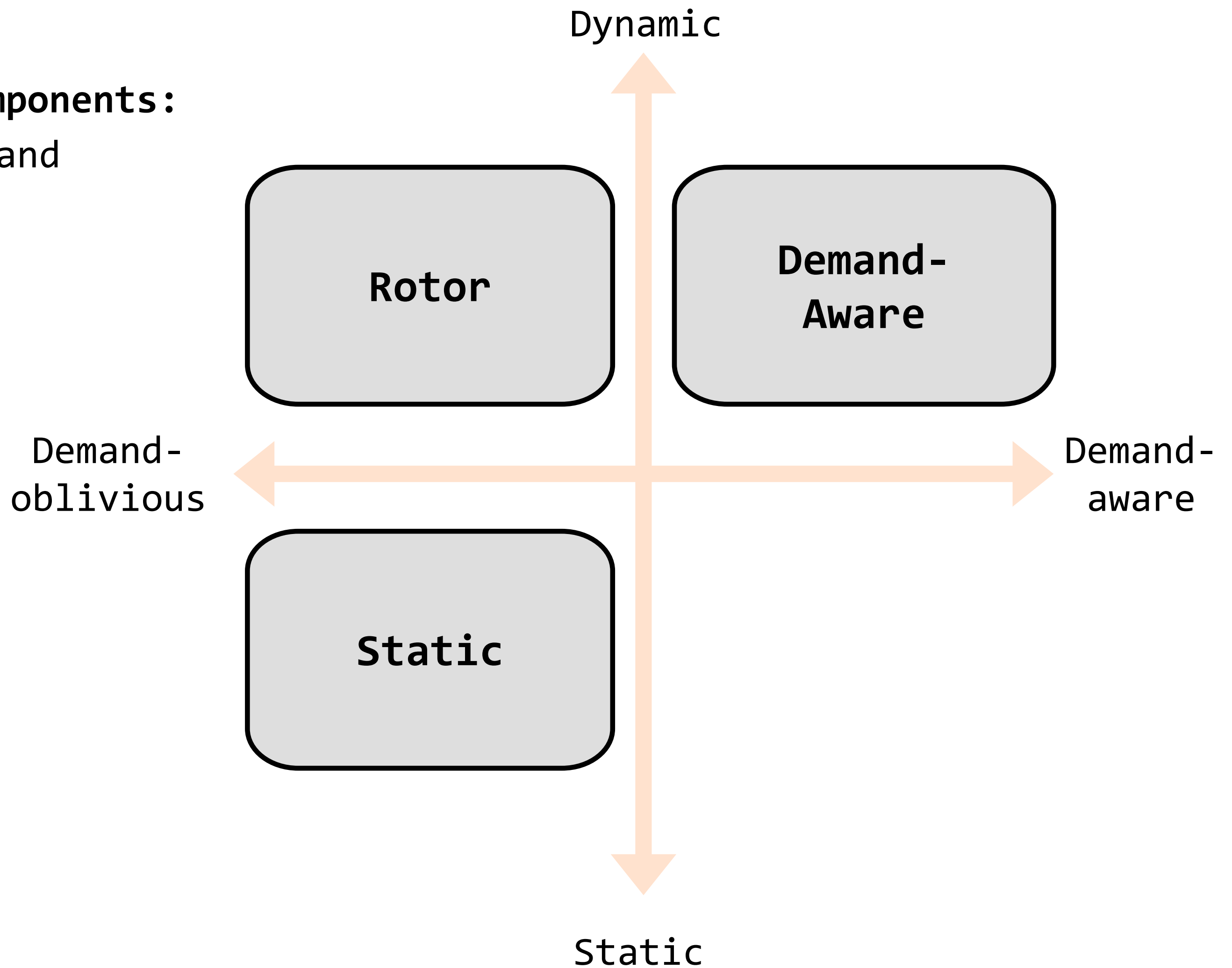


Opportunity: Tech Diversity

Diverse topology components:

⇀ demand-oblivious and demand-aware

⇀ static vs dynamic

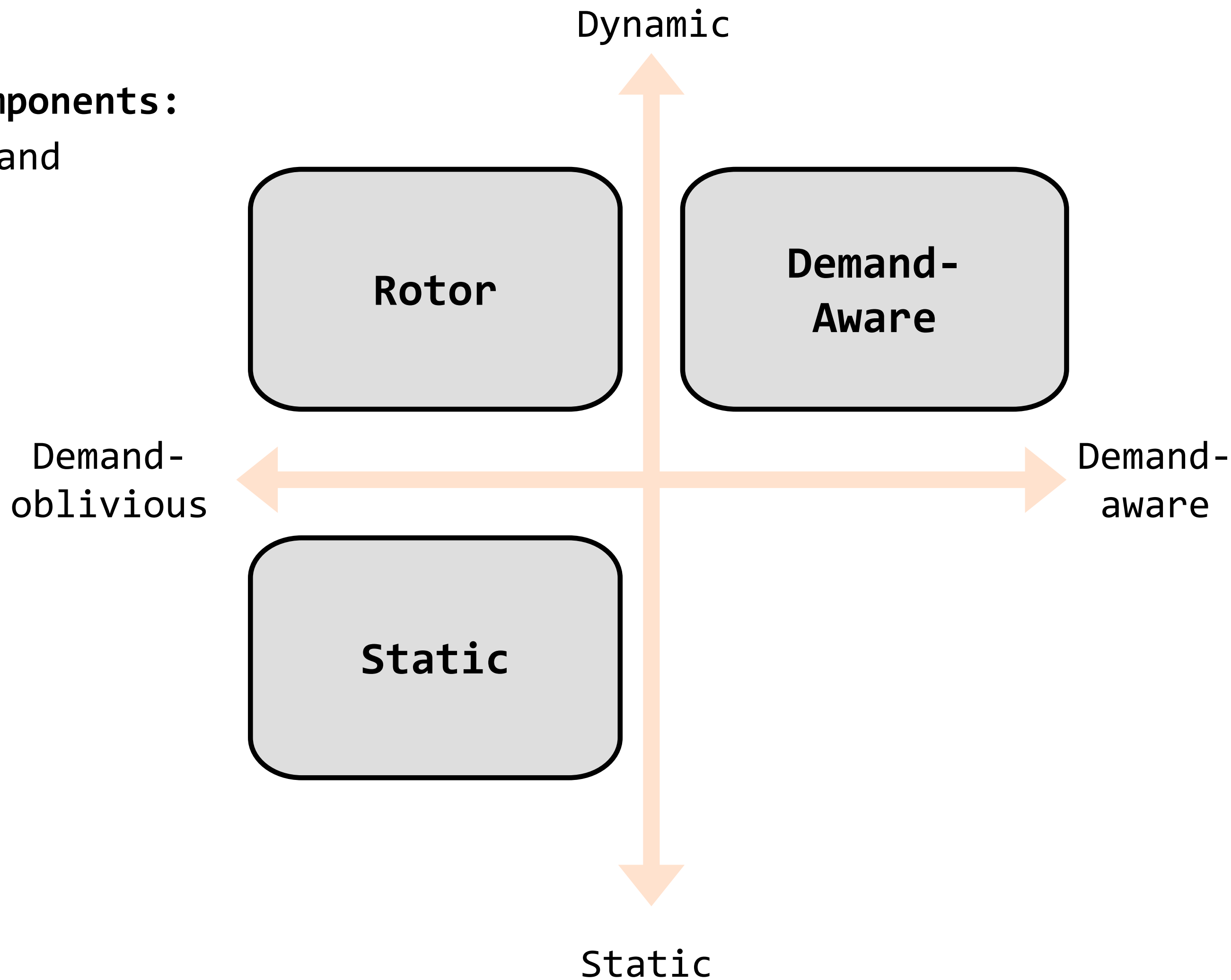


Opportunity: Tech Diversity

Diverse topology components:

⇀ demand-oblivious and demand-aware

⇀ static vs dynamic



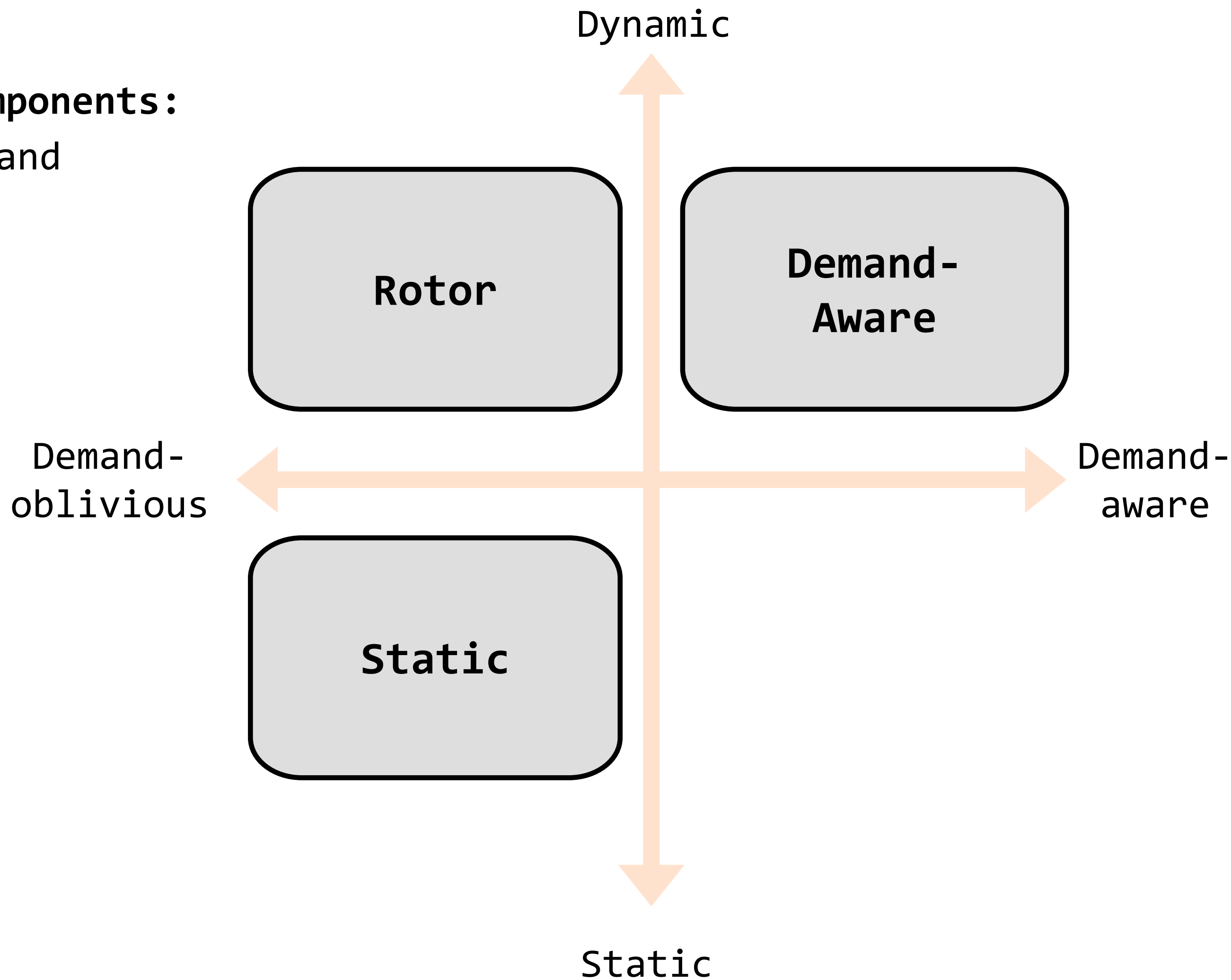
Which approach is best?

Opportunity: Tech Diversity

Diverse topology components:

⇀ demand-oblivious and demand-aware

⇀ static vs dynamic



Which approach is best?

As always in CS: It depends…

Opportunity: Tech Diversity

Diverse topology components:

⇀ demand-oblivious and demand-aware

⇀ static vs dynamic

Dynamic

Demand-oblivious

Demand-aware

Multihop forwarding:

bandwidth tax

Static

Which approach is best?

As always in CS: It depends…

Opportunity: Tech Diversity

Diverse topology components:

⇀ demand-oblivious and
demand-aware

⇀ static vs dynamic

Dynamic

Reconfiguration times:
delay tax

Demand-oblivious

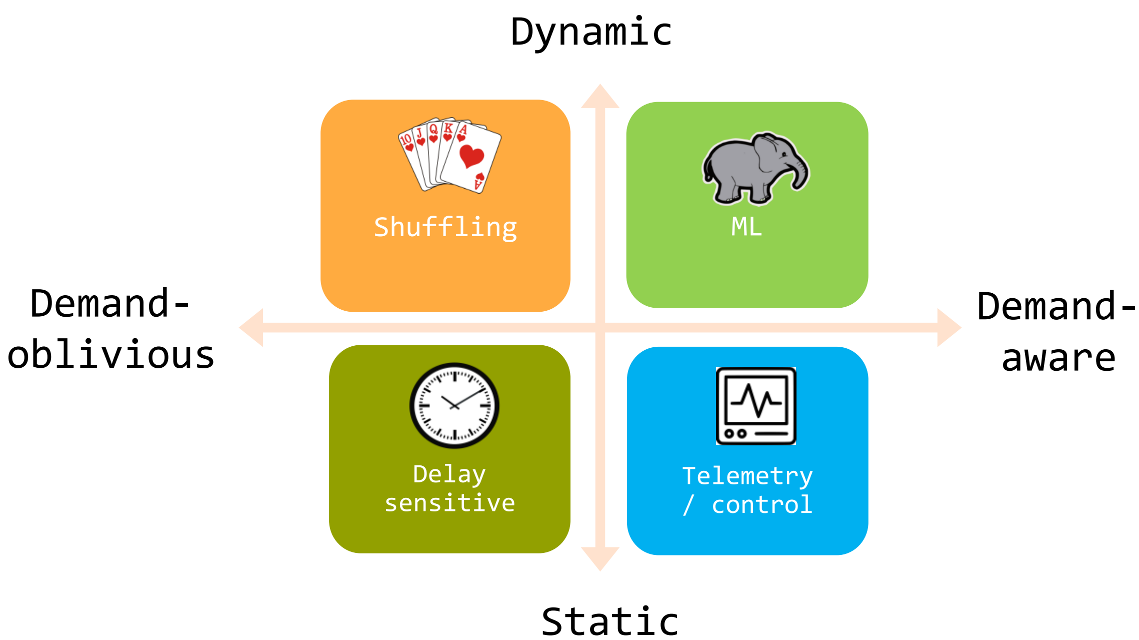
Demand-aware

Static

Which approach is best?

As always in CS:
It depends…

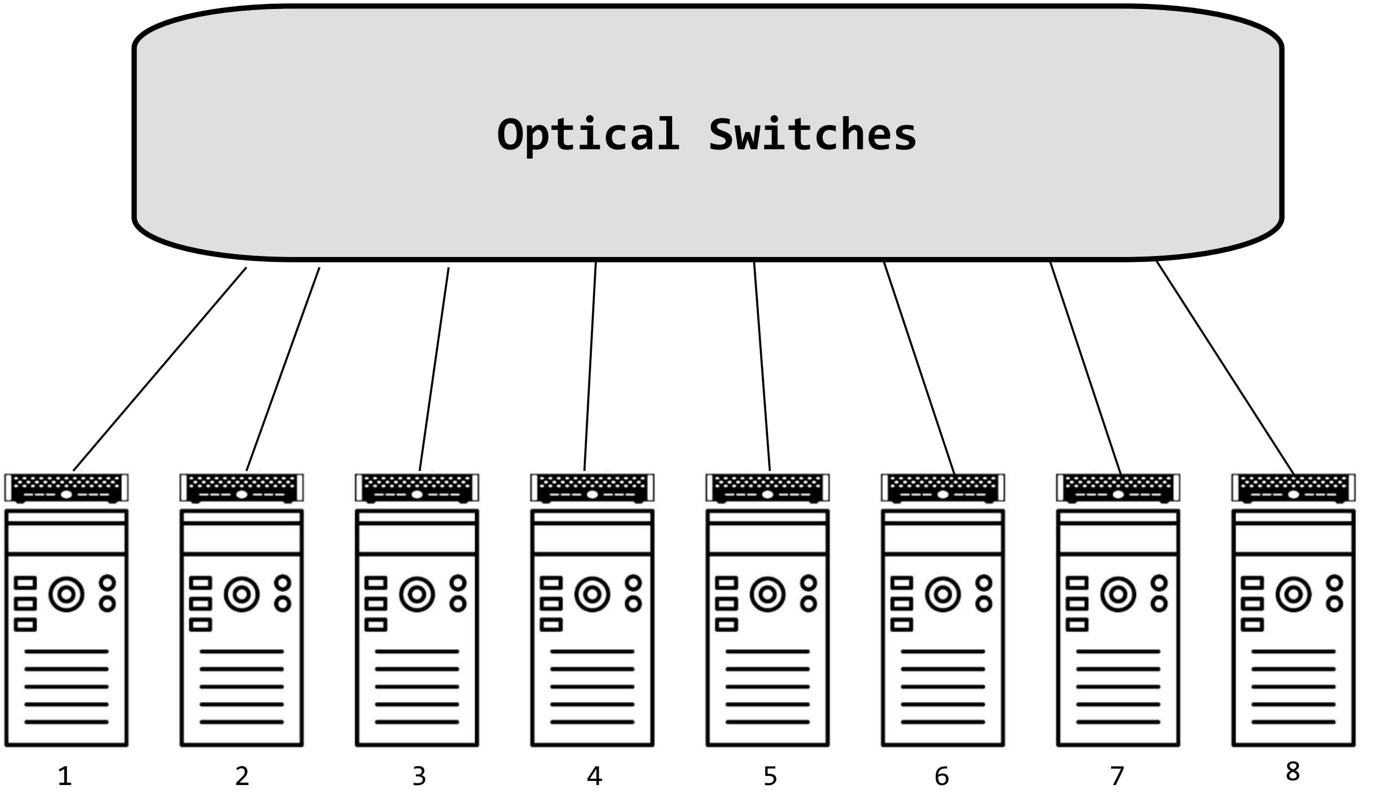
Cerberus: It's a Match!



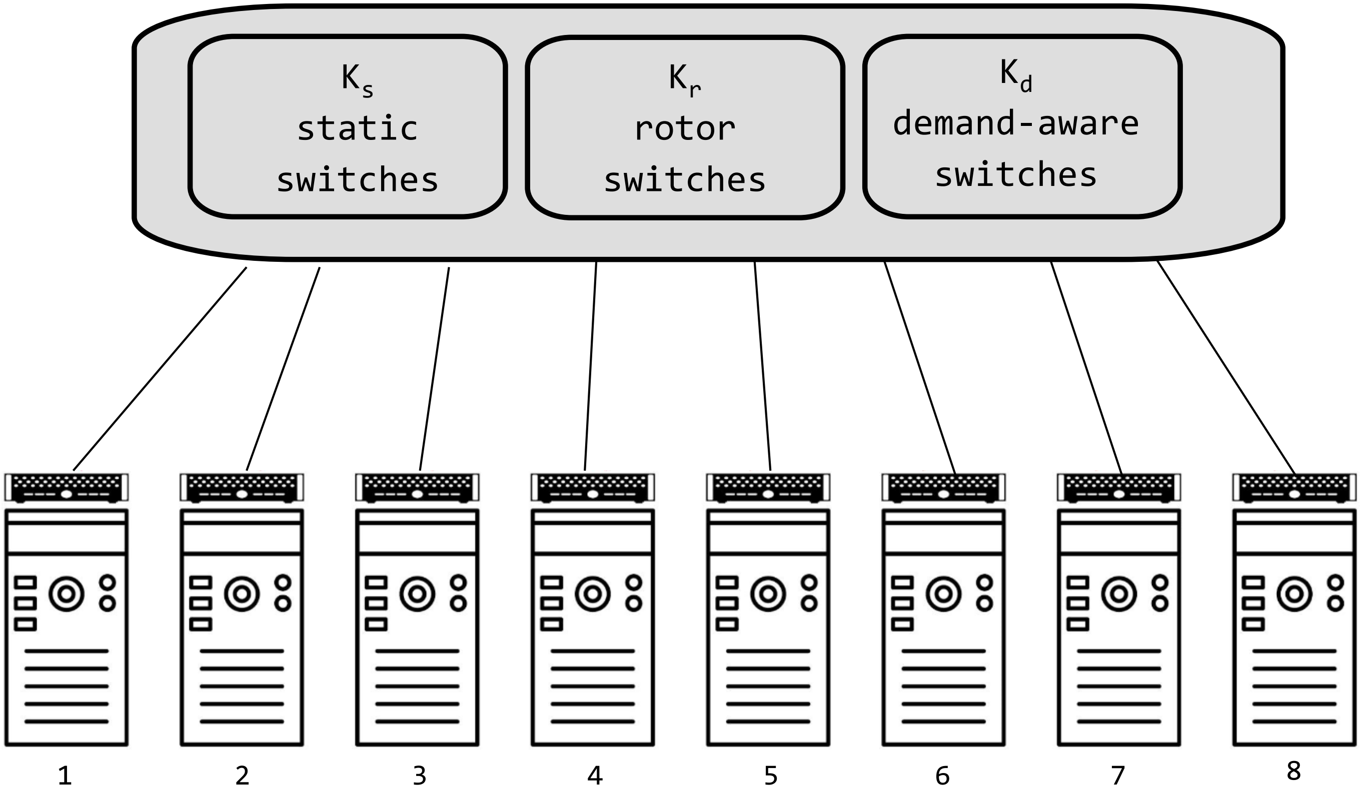
We have a first approach:
Cerberus* serves traffic on the “best topology”! (Optimality open)

* Griner et al., ACM SIGMETRICS 2022

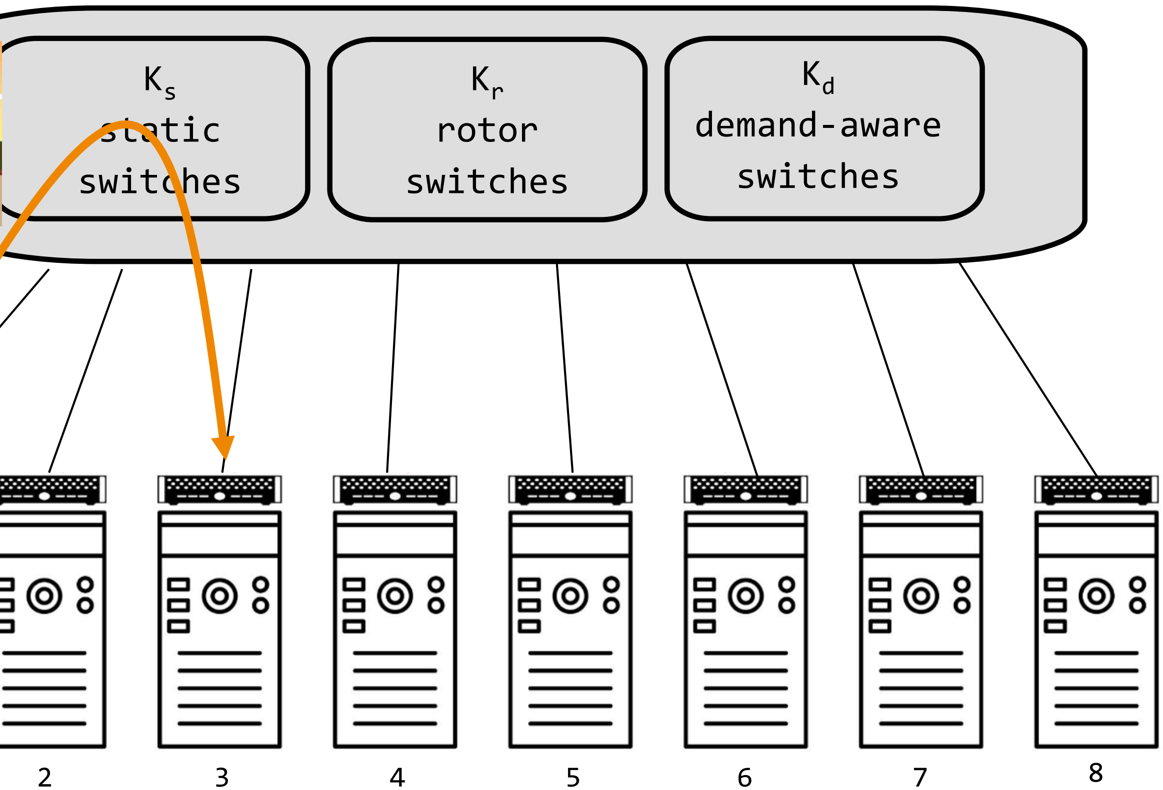
Cerberus



Cerberus

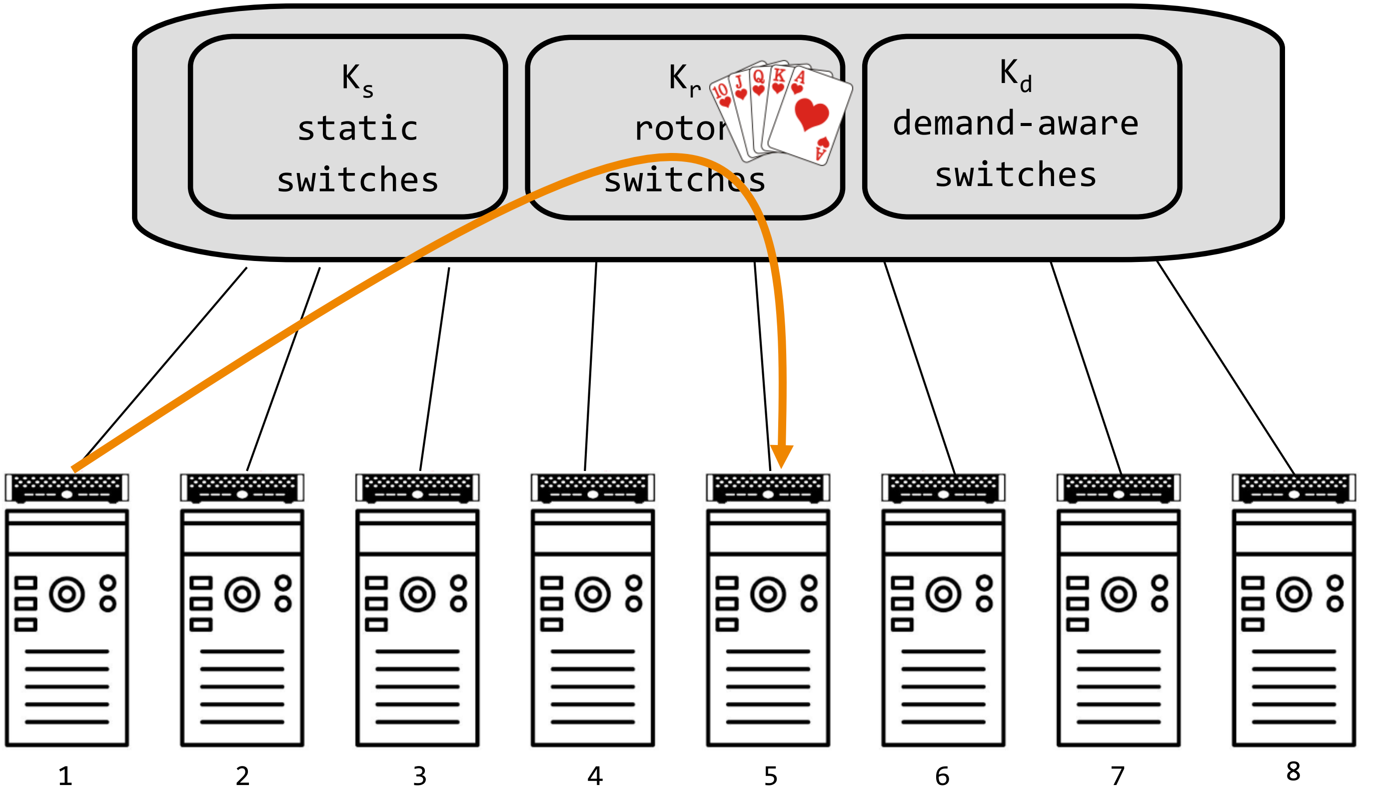

K_s
static
switches
K_r
rotor
switches
K_d
demand-aware
switches
1
2
3
4
5
6
7
8

Cerberus



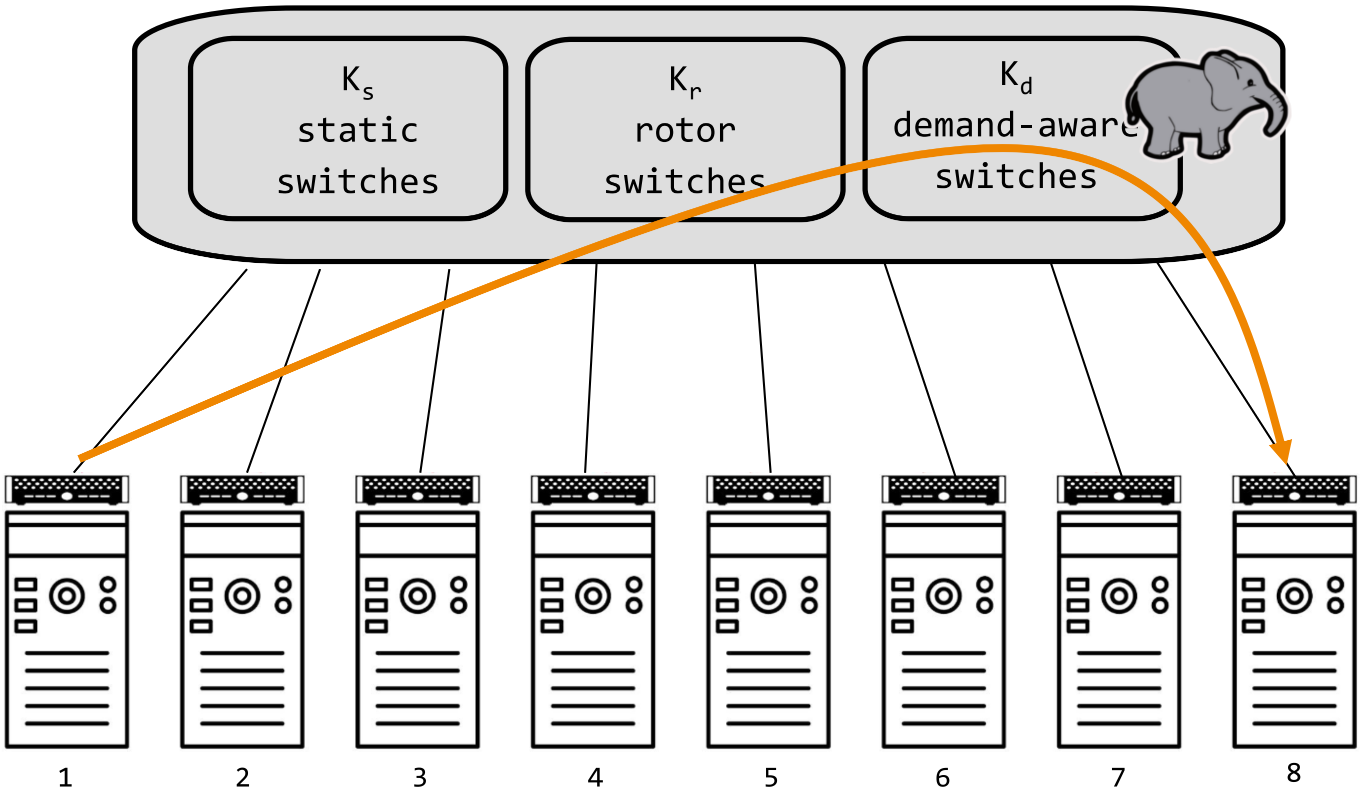
Scheduling: Small flows go via static switches…

Cerberus



Scheduling: … medium flows via rotor switches…

Cerberus



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

Summary

- Opportunity: *structure* in demand and *reconfigurable* networks
- Cerberus aims to assign traffic to its best topology
  - Depending on flow size
  - ***Open questions:*** Analysis of throughput? Optimality?

“Zukunftsmusik”

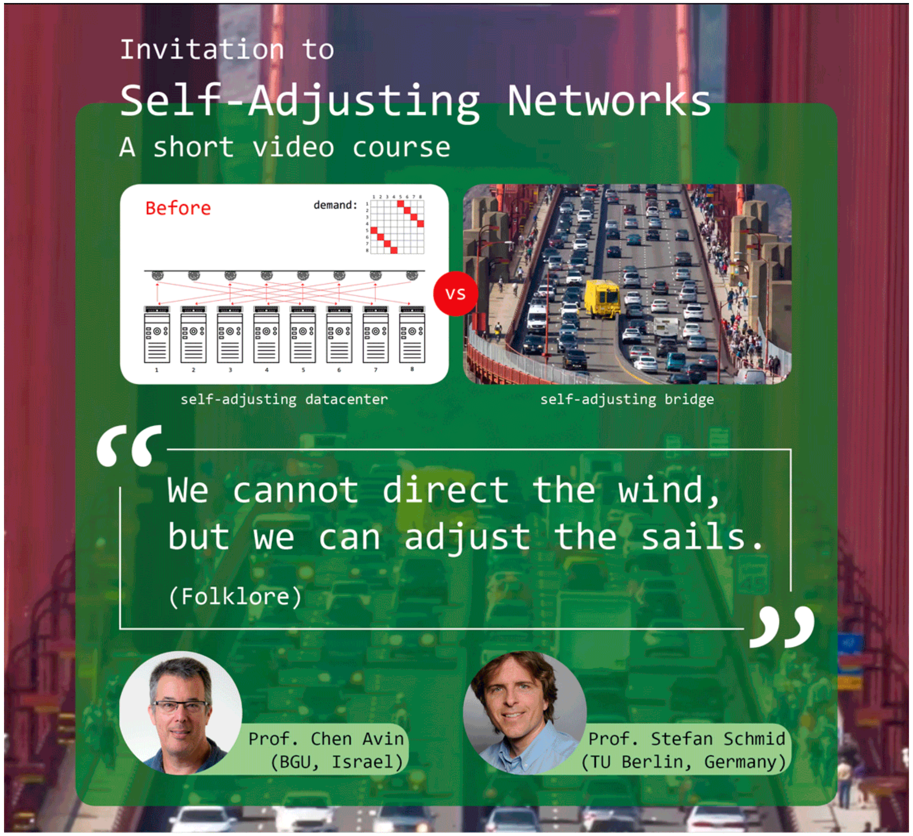
⇢ So far: tip of the iceberg

⇢ Many more challenges

- ⇀ Shock wave through *layers*:
  impact on routing and congestion control?
- ⇀ *Scalability* of control in dynamic graphs:
  local algorithms? Greedy routing?
- ⇢ Complexity of demand-aware graphs
  (*pure vs hybrid*, e.g., SplayNet)
- ⇀ *Application-specific* self-adjusting networks:
  e.g., for AI, or similar to *active dynamic networks* (independent sets, consensus, …)
- ⇀ etc.

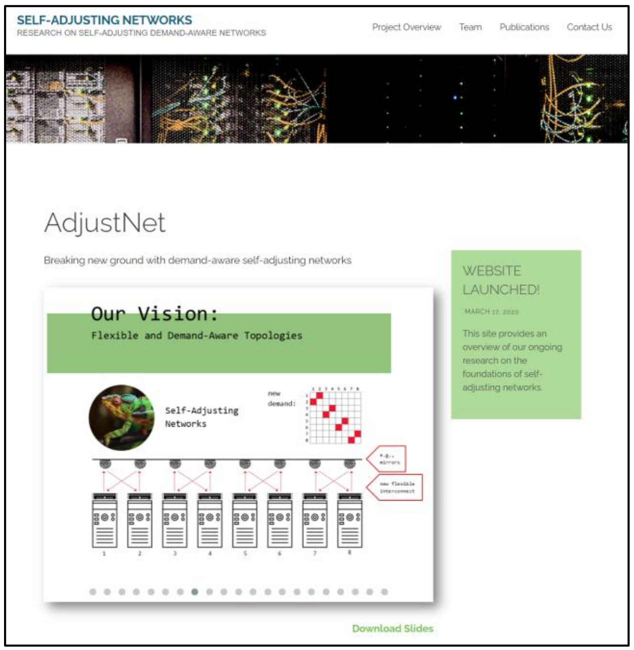
Thank you!

Online Video Course

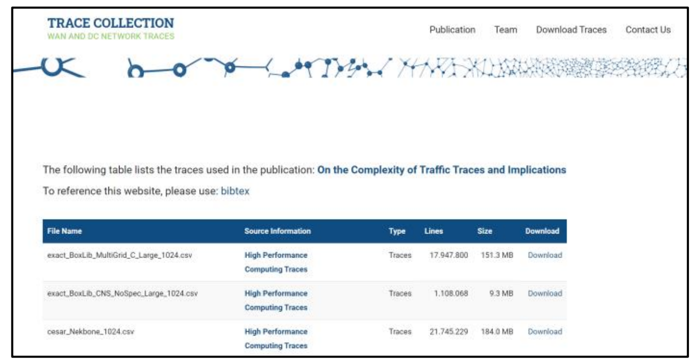


https://self-adjusting.net/course

Websites

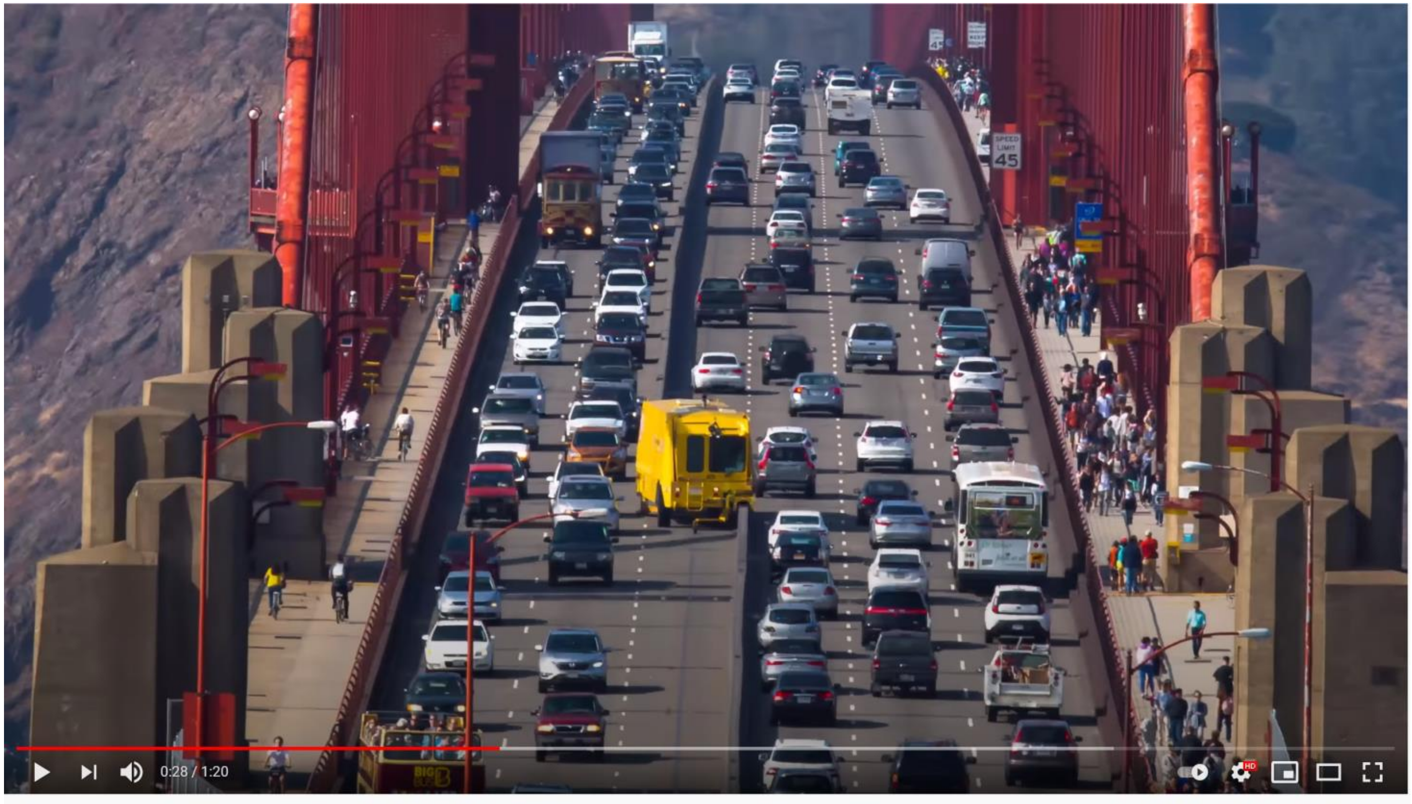


http://self-adjusting.net/
Project website



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

Questions?



Golden Gate Zipper

Further Reading

Overview: Models

Toward Demand-Aware Networking: A Theory for Self-Adjusting Networks

Chen Avin
Ben Gurion University, Israel
avin@cse.bgu.ac.il

Stefan Schmid
University of Vienna, Austria
stefan_schmid@univie.ac.at

This article is an editorial note submitted to CCR. It has NOT been peer reviewed.
The authors take full responsibility for this article's technical content. Comments can be posted through CCR Online.

ABSTRACT

The physical topology is emerging as the next frontier in an ongoing effort to render communication networks more flexible. While first empirical results indicate that these flexibilities can be exploited to reconfigure and optimize the network toward the workload it serves and, e.g., providing the same bandwidth at lower infrastructure cost, only little is known today about the fundamental algorithmic problems underlying the design of reconfigurable networks. This paper initiates the study of the theory of demand-aware, self-adjusting networks. Our main position is that self-adjusting networks should be seen through the lense of self-adjusting datastructures. Accordingly, we present a taxonomy classifying the different algorithmic models of demand-oblivious, fixed demand-aware, and reconfigurable demand-aware networks, introduce a formal model, and identify objectives and evaluation metrics. We also demonstrate, by examples, the inherent

Figure 1: Taxonomy of topology optimization

design of efficient datacenter networks has received much attention over the last years. The topologies underlying modern datacenter networks range from trees [7, 8] over hypercubes [9, 10] to expander networks [11] and provide high connectivity at low cost [1].

Until now, these networks also have in common that their topology is *fixed* and ***oblivious*** to the actual demand (i.e.,

Trace Complexity

On the Complexity of Traffic Traces and Implications

CHEN AVIN, School of Electrical and Computer Engineering, Ben Gurion University of the Negev, Israel
MANYA GHOBADI, Computer Science and Artificial Intelligence Laboratory, MIT, USA
CHEN GRINER, School of Electrical and Computer Engineering, Ben Gurion University of the Negev, Israel
STEFAN SCHMID, Faculty of Computer Science, University of Vienna, Austria

This paper presents a systematic approach to identify and quantify the types of structures featured by packet traces in communication networks. Our approach leverages an information-theoretic methodology, based on iterative randomization and compression of the packet trace, which allows us to systematically remove and measure dimensions of structure in the trace. In particular, we introduce the notion of *trace complexity* which approximates the entropy rate of a packet trace. Considering several real-world traces, we show that trace complexity can provide unique insights into the characteristics of various applications. Based on our approach, we also propose a traffic generator model able to produce a synthetic trace that matches the complexity levels of its corresponding real-world trace. Using a case study in the context of datacenters, we show that insights into the structure of packet traces can lead to improved demand-aware network designs: datacenter topologies that are optimized for specific traffic patterns.

Dynamic DAN

SplayNet: Towards Locally Self-Adjusting Networks

Stefan Schmid*, Chen Avin*, Christian Scheideler, Michael Borokhovich, Bernhard Haeupler, Zvi Lotker

***Abstract*—This paper initiates the study of locally self-adjusting networks: networks whose topology adapts dynamically and in a decentralized manner, to the communication pattern σ. Our vision can be seen as a distributed generalization of the self-adjusting datastructures introduced by Sleator and Tarjan [22]: In contrast to their splay trees which dynamically optimize the lookup costs from a *single node* (namely the tree root), we seek to minimize the routing cost between arbitrary *communication pairs* in the network.**

As a first step, we study distributed binary search trees (BSTs), which are attractive for their support of greedy routing. We introduce a simple model which captures the fundamental tradeoff between the benefits and costs of self-adjusting networks. We present the *SplayNet* algorithm and formally analyze its performance, and prove its optimality in specific case studies. We also introduce lower bound techniques based on interval cuts and edge expansion, to study the limitations of any demand-optimized network. Finally, we extend our study to multi-tree networks, and highlight an intriguing difference between classic and distributed splay trees.

I. Introduction

In the 1980s, Sleator and Tarjan [22] proposed an appealing new paradigm to design efficient Binary Search Tree (BST) datastructures: rather than optimizing traditional metrics such toward static metrics, such as the diameter or the length of the longest route: the self-adjusting paradigm has not spilled over to distributed networks yet.

We, in this paper, initiate the study of a distributed generalization of self-optimizing datastructures. This is a non-trivial generalization of the classic splay tree concept: While in classic BSTs, a *lookup request* always originates from the same node, the tree root, distributed datastructures and networks such as skip graphs [2], [13] have to support *routing requests* between arbitrary pairs (or *peers*) of communicating nodes; in other words, both the source as well as the destination of the requests become variable. Figure 1 illustrates the difference between classic and distributed binary search trees.

In this paper, we ask: Can we reap similar benefits from self-adjusting *entire networks*, by adaptively reducing the distance between frequently communicating nodes?

As a first step, we explore fully decentralized and self-adjusting Binary Search Tree networks: in these networks, nodes are arranged in a binary tree which respects node identifiers. A BST topology is attractive as it supports greedy routing: a node can decide locally to which port to forward a request given its destination address.

Cerberus

Cerberus: The Power of Choices in Datacenter Topology Design*

A Throughput Perspective

CHEN GRINER, School of Electrical and Computer Engineering, Ben Gurion University of the Negev, Israel
JOHANNES ZERWAS, Technical University of Munich, Germany
ANDREAS BLENK, Technical University of Munich, Germany
MANYA GHOBADI, Computer Science and Artificial Intelligence Laboratory, MIT, USA
STEFAN SCHMID, Faculty of Computer Science, University of Vienna, Austria
CHEN AVIN, School of Electrical and Computer Engineering, Ben Gurion University of the Negev, Israel

The bandwidth and latency requirements of modern datacenter applications have led researchers to propose various topology designs using static, dynamic demand-oblivious (rotor), and/or dynamic demand-aware switches. However, given the diverse nature of datacenter traffic, there is little consensus about how these designs would fare against each other. In this work, we analyze the throughput of existing topology designs under different traffic patterns and study their unique advantages and potential costs in terms of bandwidth and latency "tax". To overcome the identified inefficiencies, we propose Cerberus, a unified, two-layer leaf-spine optical datacenter design with three topology types. Cerberus systematically matches different traffic patterns with their most suitable topology type: e.g., latency-sensitive flows are transmitted via a static topology,

Selected References

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.
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.
On the Complexity of Traffic Traces and Implications
Chen Avin, Manya Ghobadi, Chen Griner, and Stefan Schmid.
ACM SIGMETRICS, Boston, Massachusetts, USA, June 2020.
Survey of Reconfigurable Data Center Networks: Enablers, Algorithms, Complexity
Klaus-Tycho Foerster and Stefan Schmid.
SIGACT News, June 2019.
Toward Demand-Aware Networking: A Theory for Self-Adjusting Networks (Editorial)
Chen Avin and Stefan Schmid.
ACM SIGCOMM Computer Communication Review (**CCR**), October 2018.
Demand-Aware Network Design with Minimal Congestion and Route Lengths
Chen Avin, Kaushik Mondal, and Stefan Schmid.
38th IEEE Conference on Computer Communications (**INFOCOM**), Paris, France, April 2019.
Distributed Self-Adjusting Tree Networks
Bruna Peres, Otavio Augusto de Oliveira Souza, Olga Goussevskaia, Chen Avin, and Stefan Schmid.
38th IEEE Conference on Computer Communications (**INFOCOM**), Paris, France, April 2019.
Demand-Aware Network Designs of Bounded Degree
Chen Avin, Kaushik Mondal, and Stefan Schmid.
31st International Symposium on Distributed Computing (**DISC**), Vienna, Austria, October 2017.
SplayNet: Towards Locally Self-Adjusting Networks
Stefan Schmid, Chen Avin, Christian Scheideler, Michael Borokhovich, Bernhard Haeupler, and Zvi Lotker.
IEEE/ACM Transactions on Networking (**TON**), Volume 24, Issue 3, 2016. Early version: IEEE **IPDPS** 2013.
Characterizing the Algorithmic Complexity of Reconfigurable Data Center Architectures
Klaus-Tycho Foerster, Monia Ghobadi, and Stefan Schmid.
ACM/IEEE Symposium on Architectures for Networking and Communications Systems (**ANCS**), Ithaca, New York, USA, July 2018.