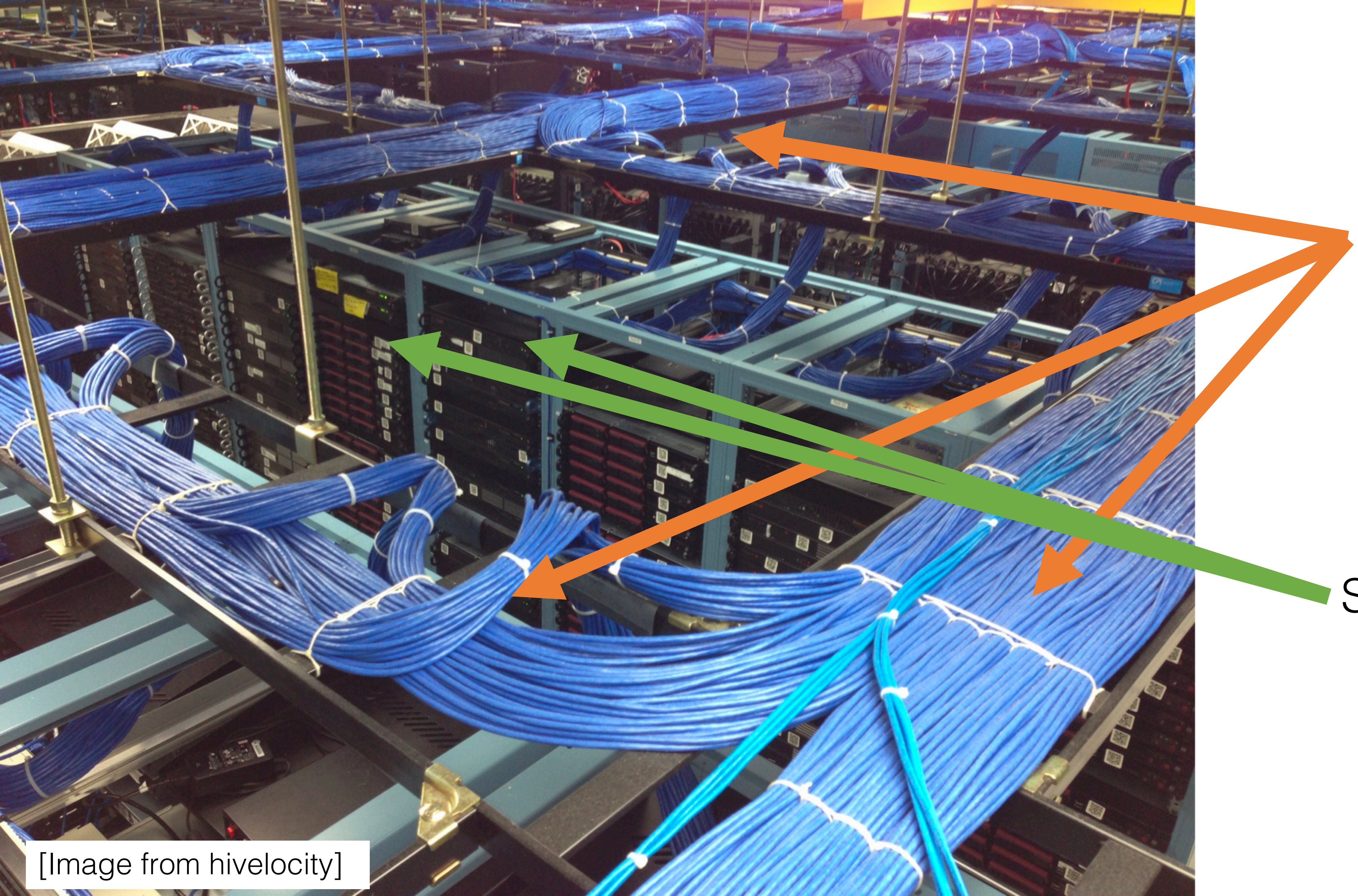
RAIL: A Case for Redundant Arrays of Inexpensive Links in Data Center Networks

Danyang Zhuo, Monia Ghobadi, Ratul Mahajan, Amar Phanishayee,
Xuan Kelvin Zou, Hang Guan, Arvind Krishnamurthy, Thomas Anderson

Microsoft Research

University of Washington

Columbia University



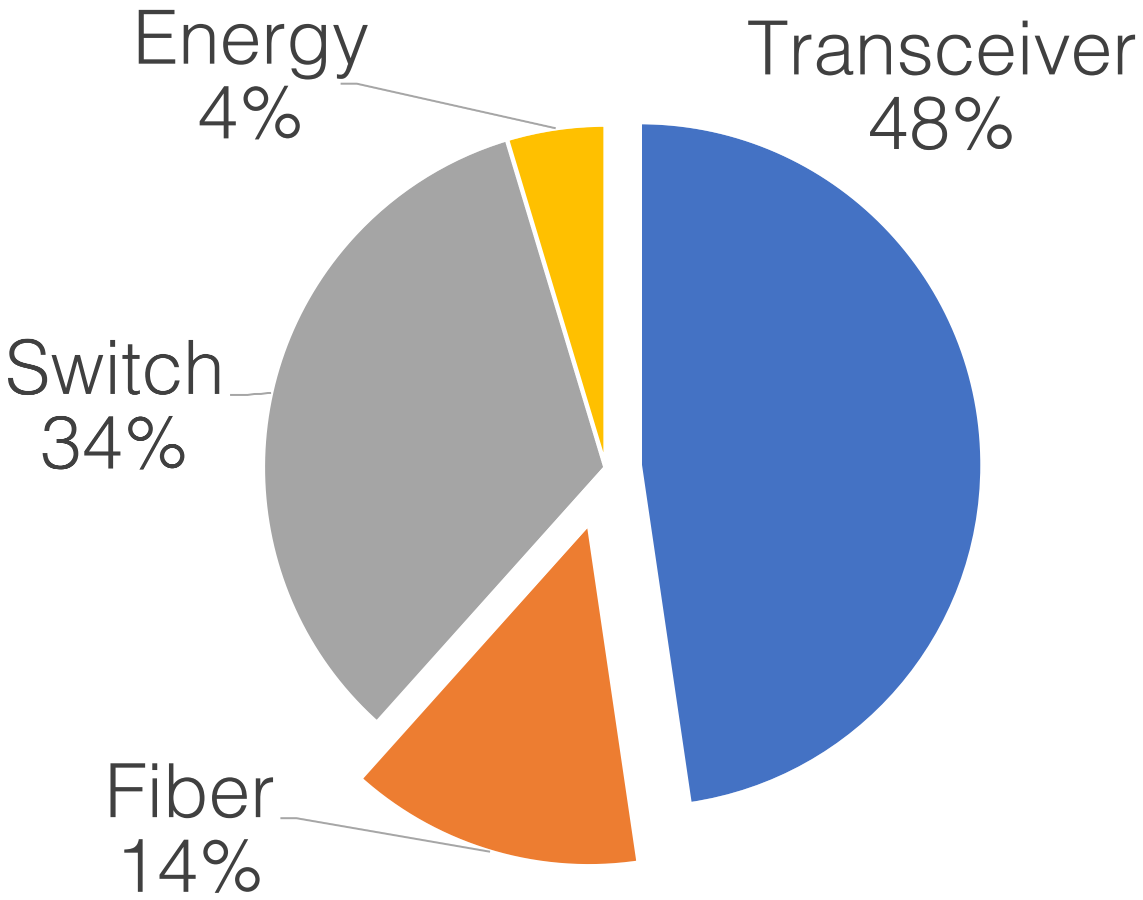
Fiber

Switch

[Image from hivelocity]

Optical Links are Expensive

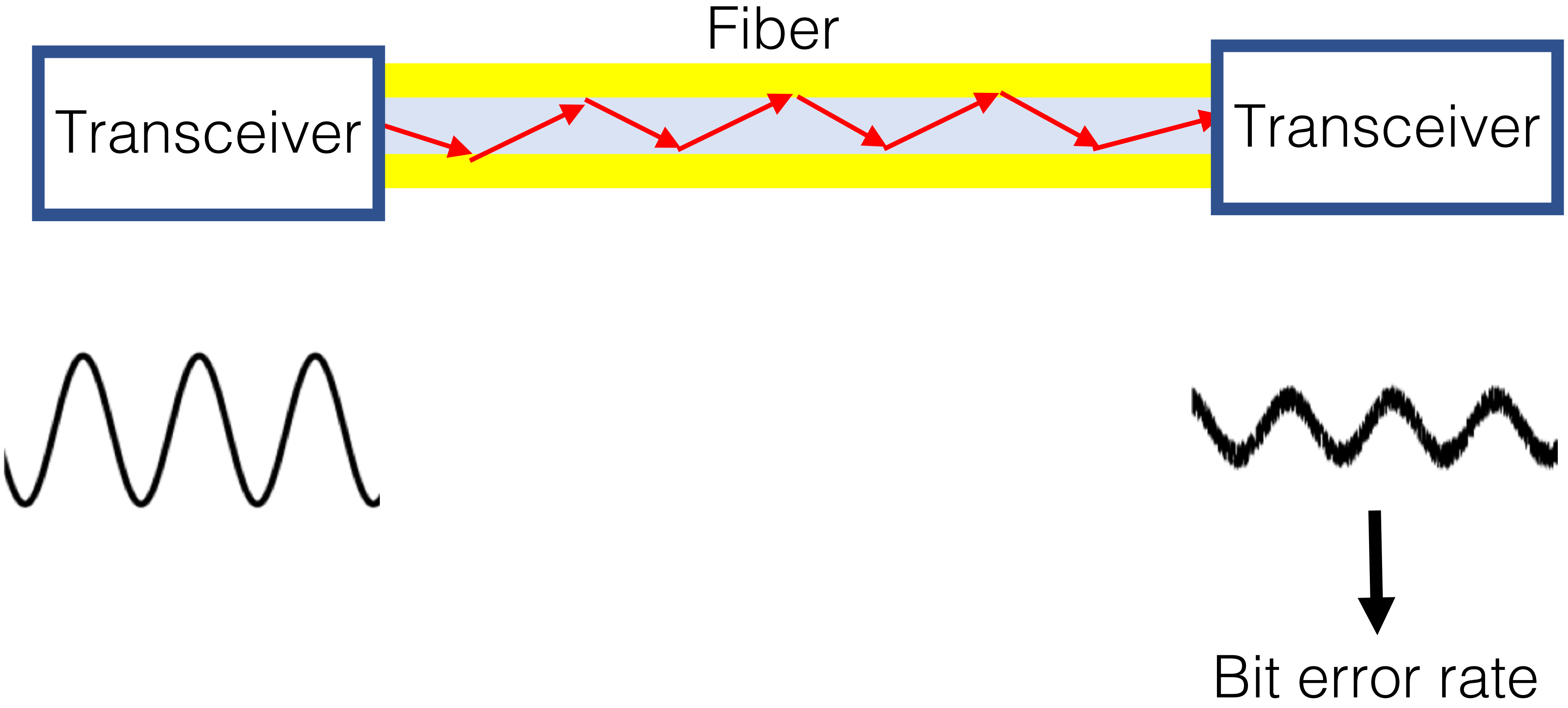
Transceiver



In this talk..

- First large-scale measurement of optical links in DCN
  - Significant over-engineering in optics
- Reduce over-engineering by using transceivers beyond design reach limit
  - Cost saving up to 40% of DCN
- Challenge: Packet loss on a small fraction of links
  - RAIL protects loss sensitive applications from this packet loss

Signal Strength and Bit Error Rate



Fiber

Transceiver

Transceiver

Bit error rate

Transceiver Classification

→	10G-SR	10G-LR	40G-SR4	40G-LR4
→	300m	10km	100m	10km
→	$45	$111	$165	$1249

SR 300m multi-mode 10G rate SFP+ 850nm

$45.00

LR 10Km single-mode 10G rate SFP+ 1310nm

$111.00

SR4 100m multi-mode 40G rate QSFP+ MPO connector 850nm

$165.00

LR4 10Km single-mode 40G rate QSFP+ LC connector CWDM

$1,249.00

[Image from robofiber]

Transceiver Classification

	10G-SR		10G-LR	40G-SR4		40G-LR4
	300m		10km	100m		10km
→	$45	<	$111	$165	<	$1249

SR 300m multi-mode 10G rate SFP+ 850nm

$45.00

LR 10Km single-mode 10G rate SFP+ 1310nm

$111.00

SR4 100m multi-mode 40G rate QSFP+ MPO connector 850nm

$165.00

LR4 10Km single-mode 40G rate QSFP+ LC connector CWDM

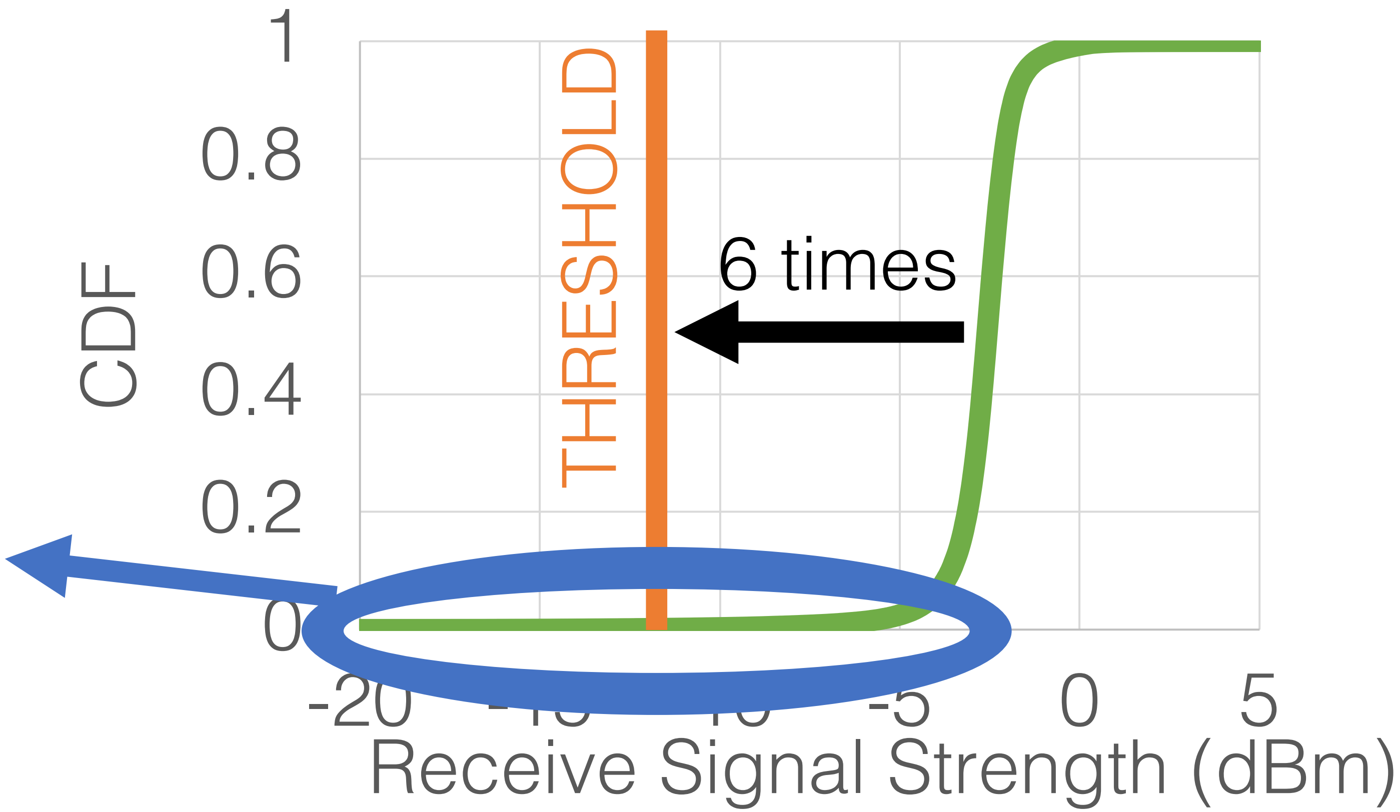
$1,249.00

[Image from robofiber]

Over-engineering in Optics

300K links

10G-SR



Uniform randomly distributed

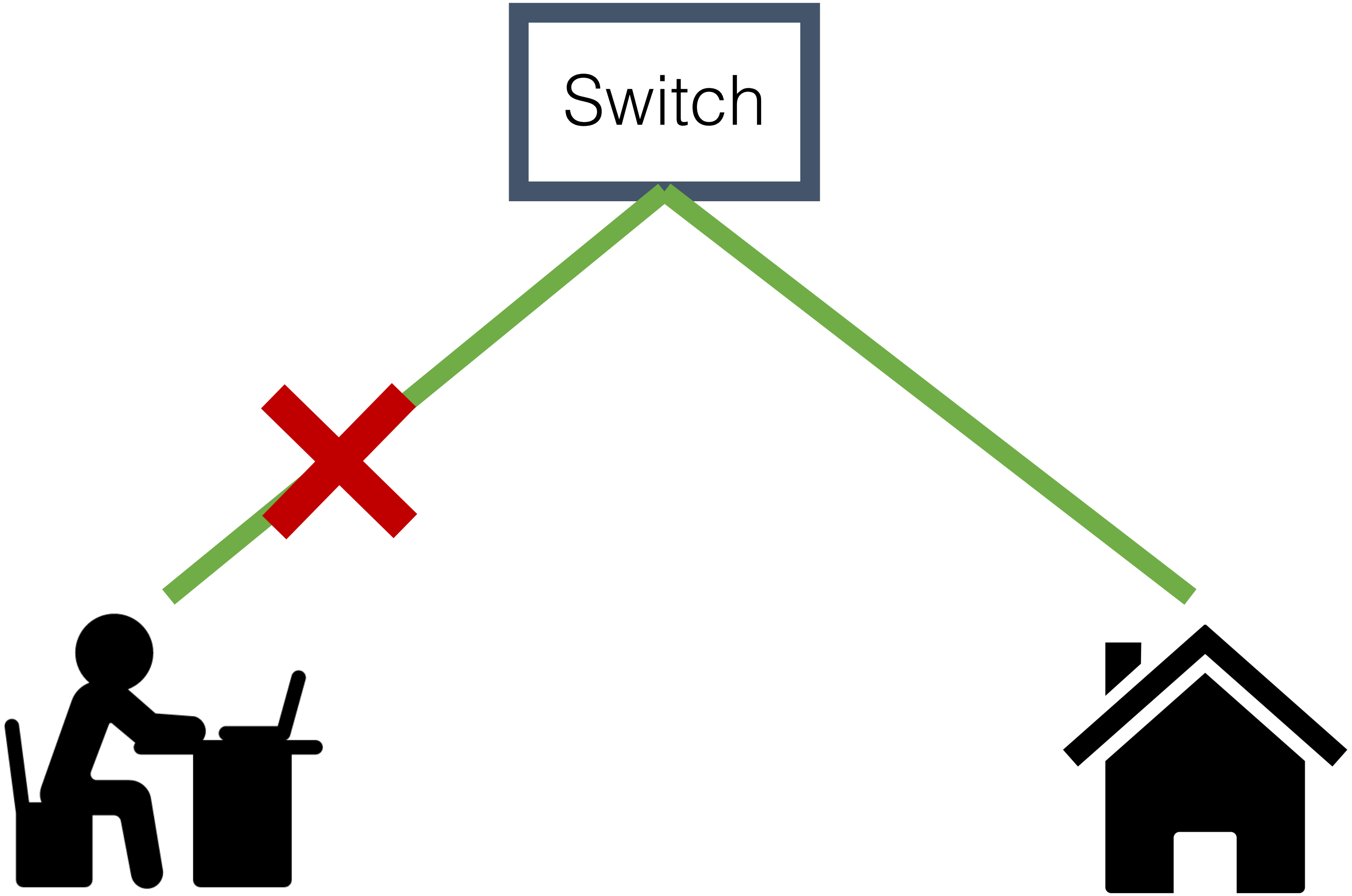
Pervasive across: technology types (10G, 40G, 100G), 20 data centers, 5 major transceiver manufacturers, 10 months

Why so much over-engineering?

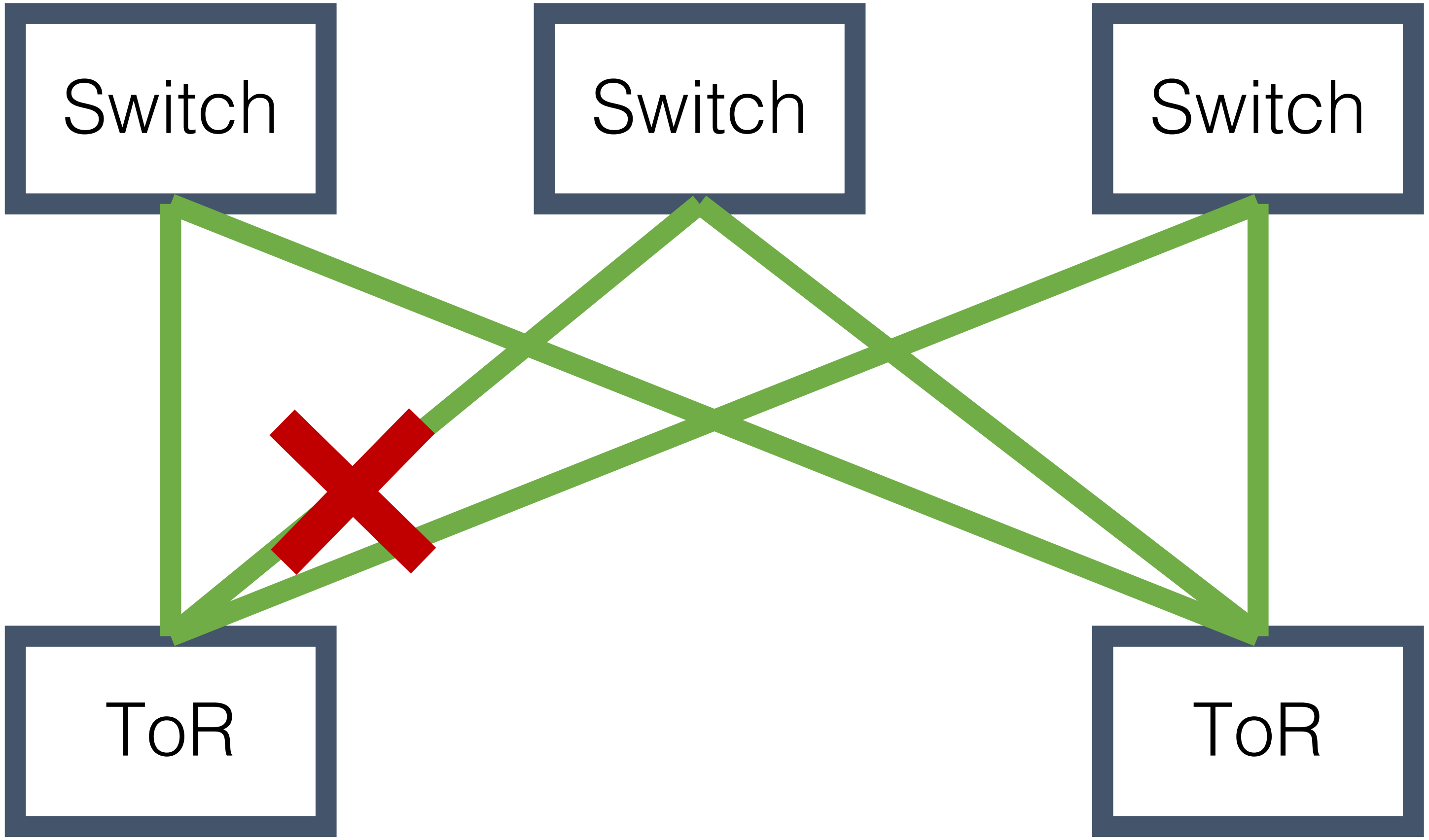
IEEE 802.3 Standards

- Ensure every link is reliable under worst-case assumption
  - Fiber quality
  - Connector loss
  - Dispersion
- Derive reach limit based on worst-case assumption

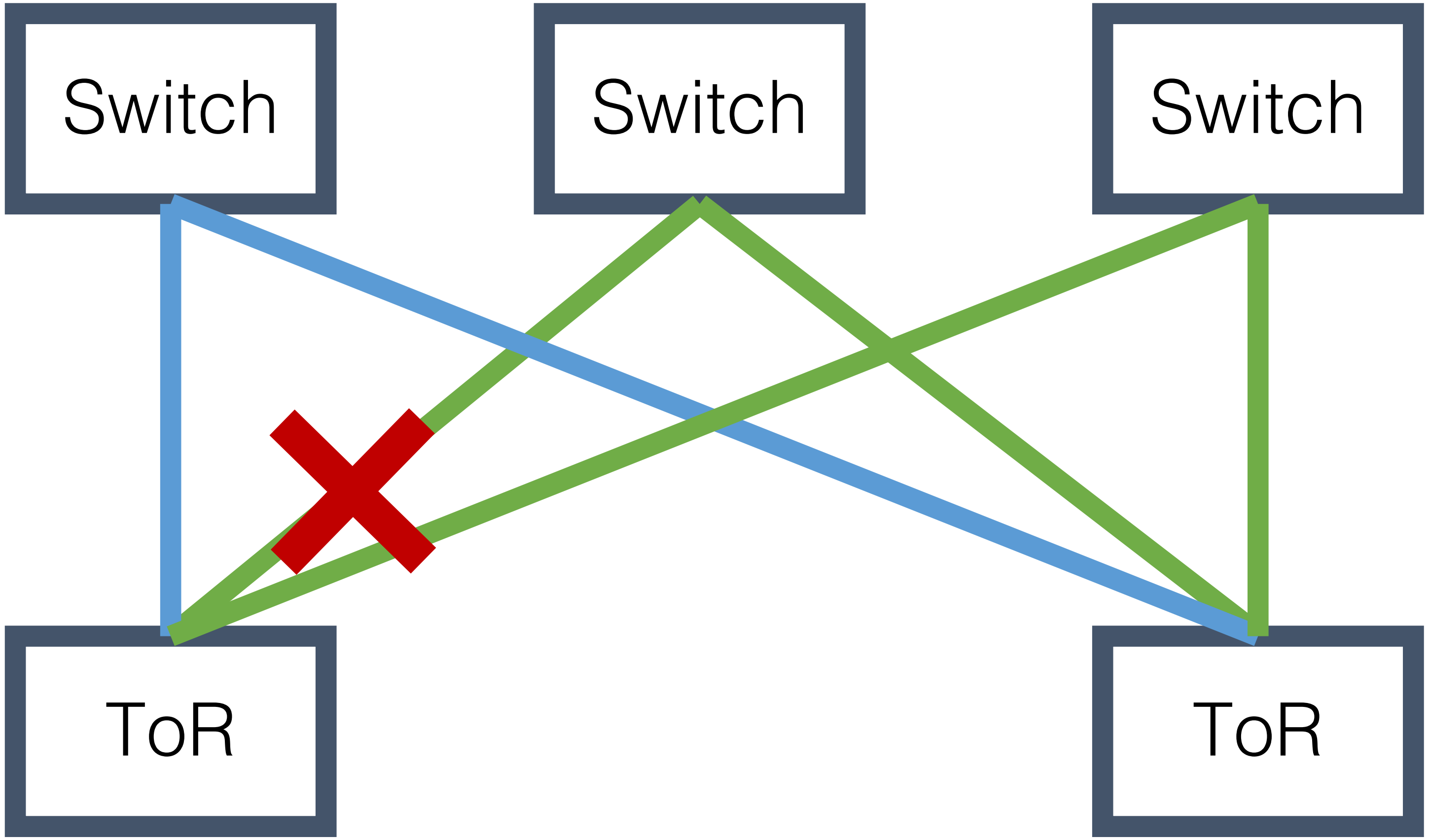
IEEE 802.3 Standards



IEEE 802.3 Standards

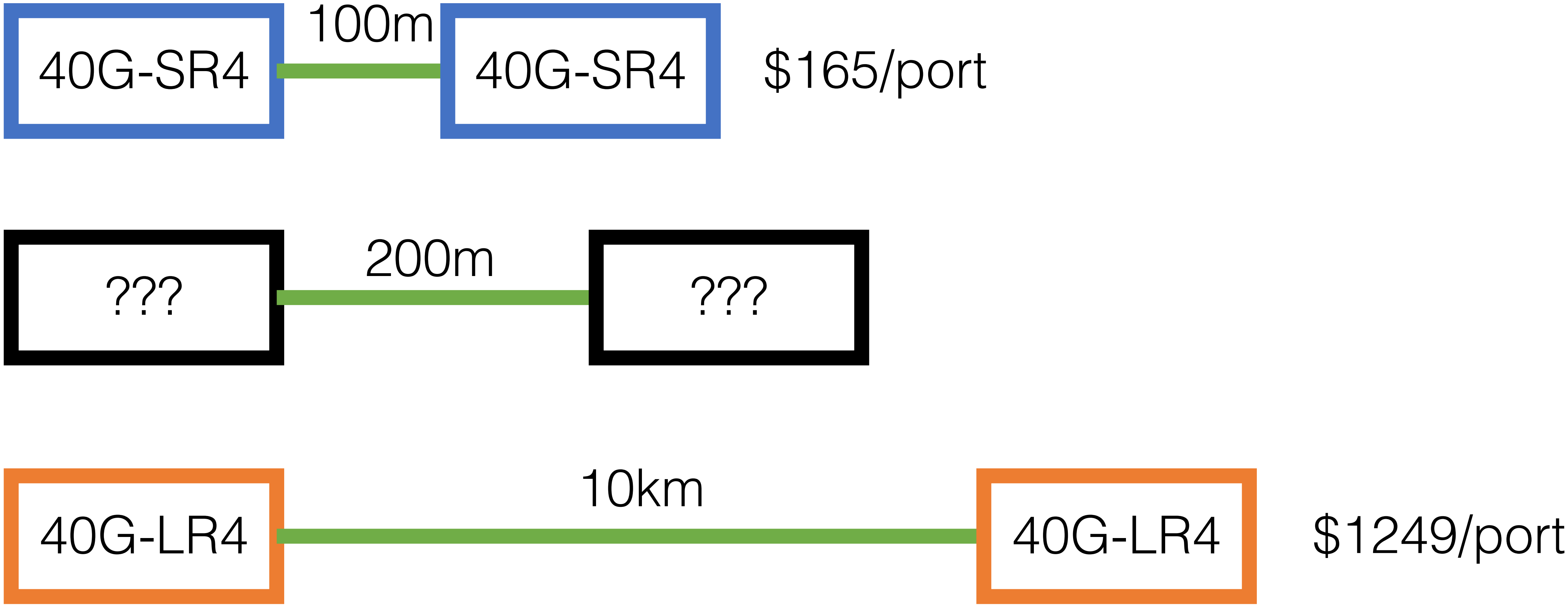


IEEE 802.3 Standards



How to reduce over-engineering?

Use Transceivers beyond Design Reach Limit

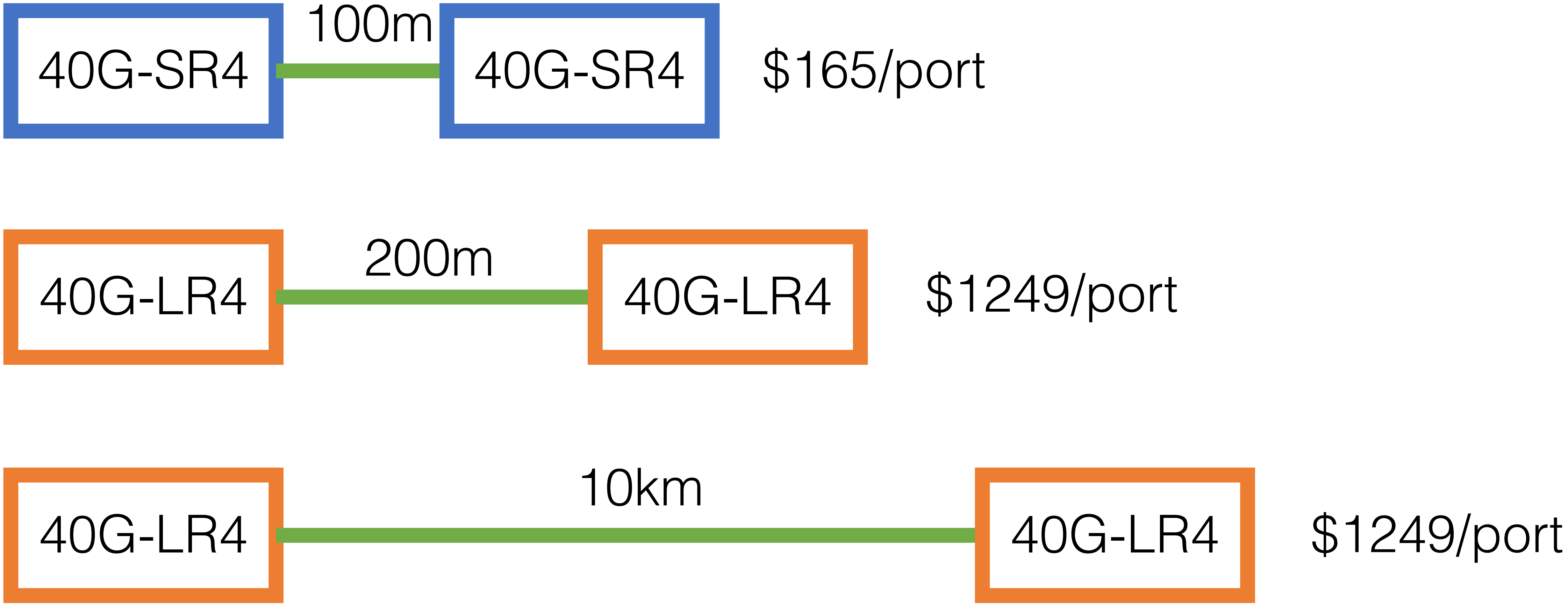


40G-SR4 100m 40G-SR4 $165/port

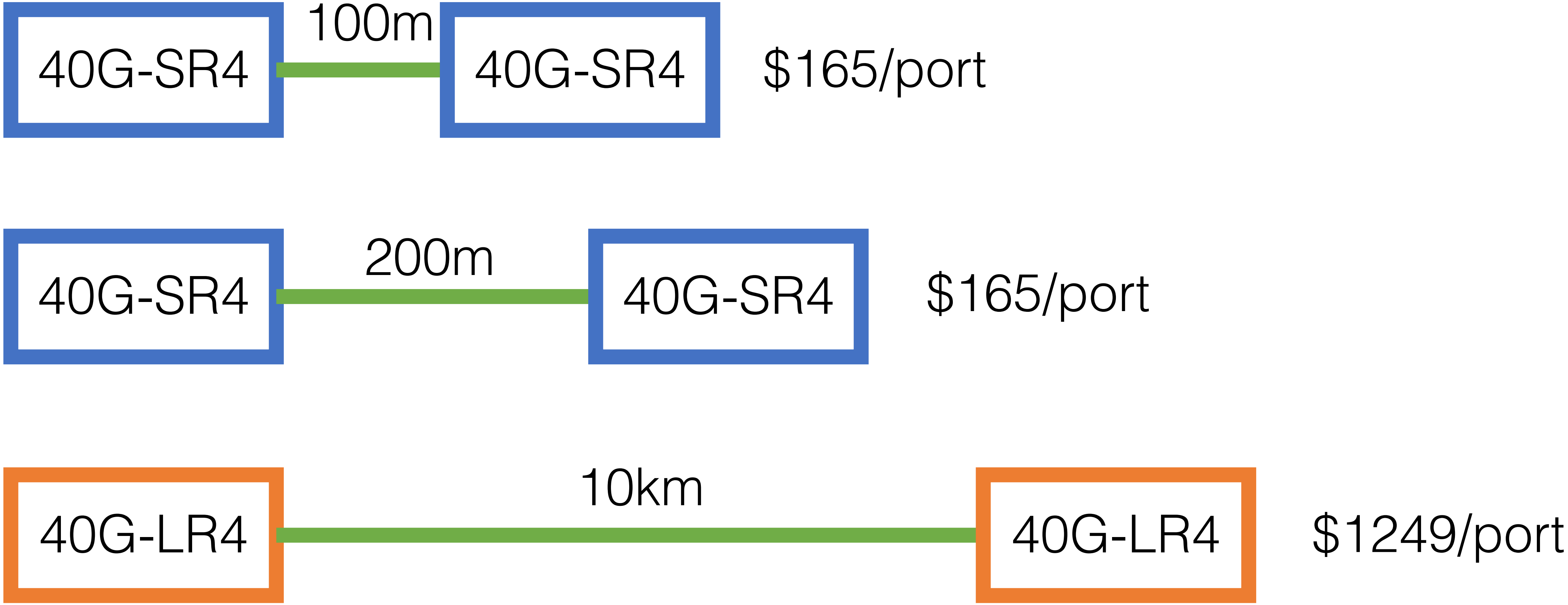
??? 200m ???

40G-LR4 10km 40G-LR4 $1249/port

Use Transceivers beyond Design Reach Limit



Use Transceivers beyond Design Reach Limit

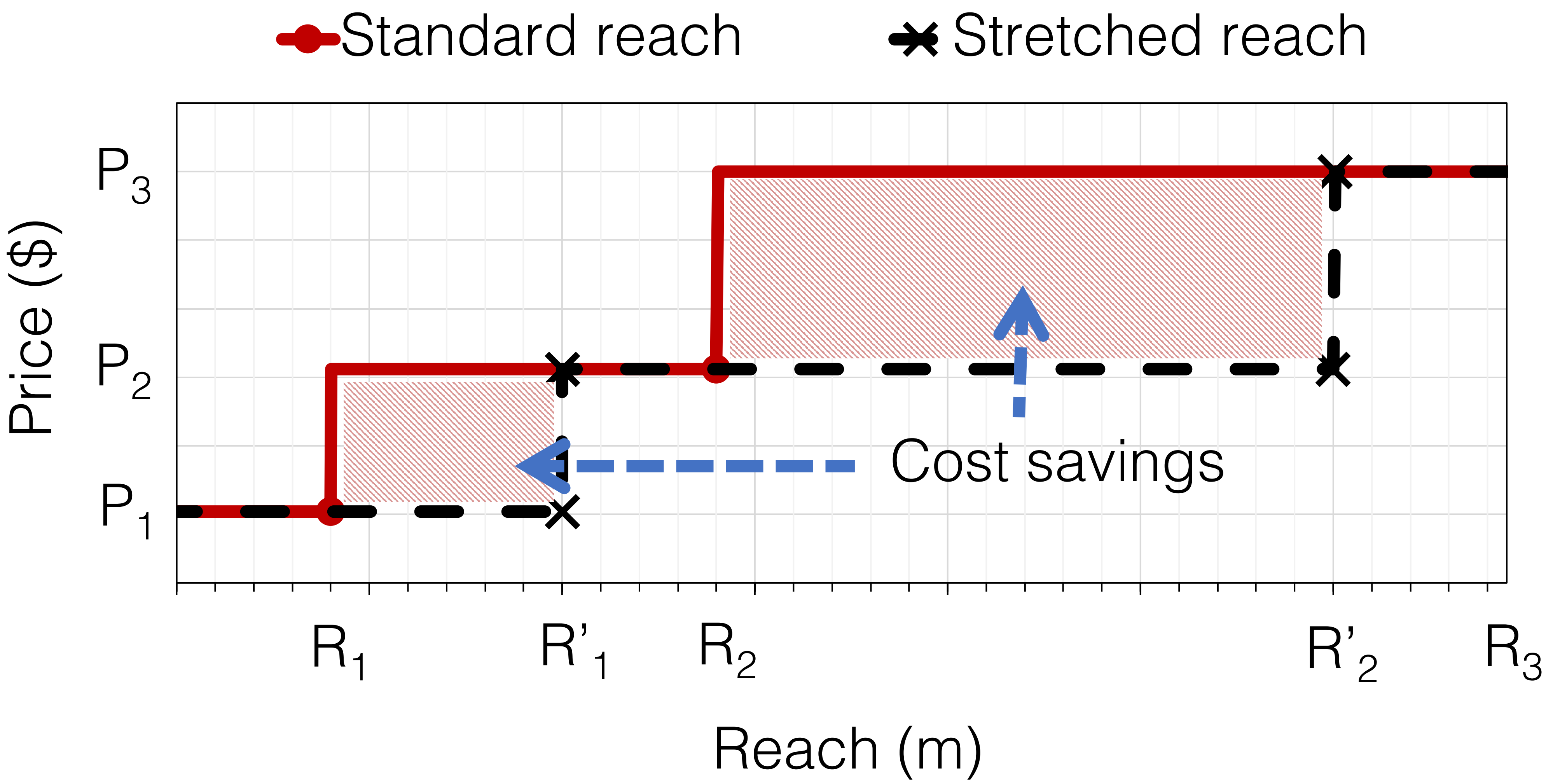


40G-SR4 100m 40G-SR4 $165/port

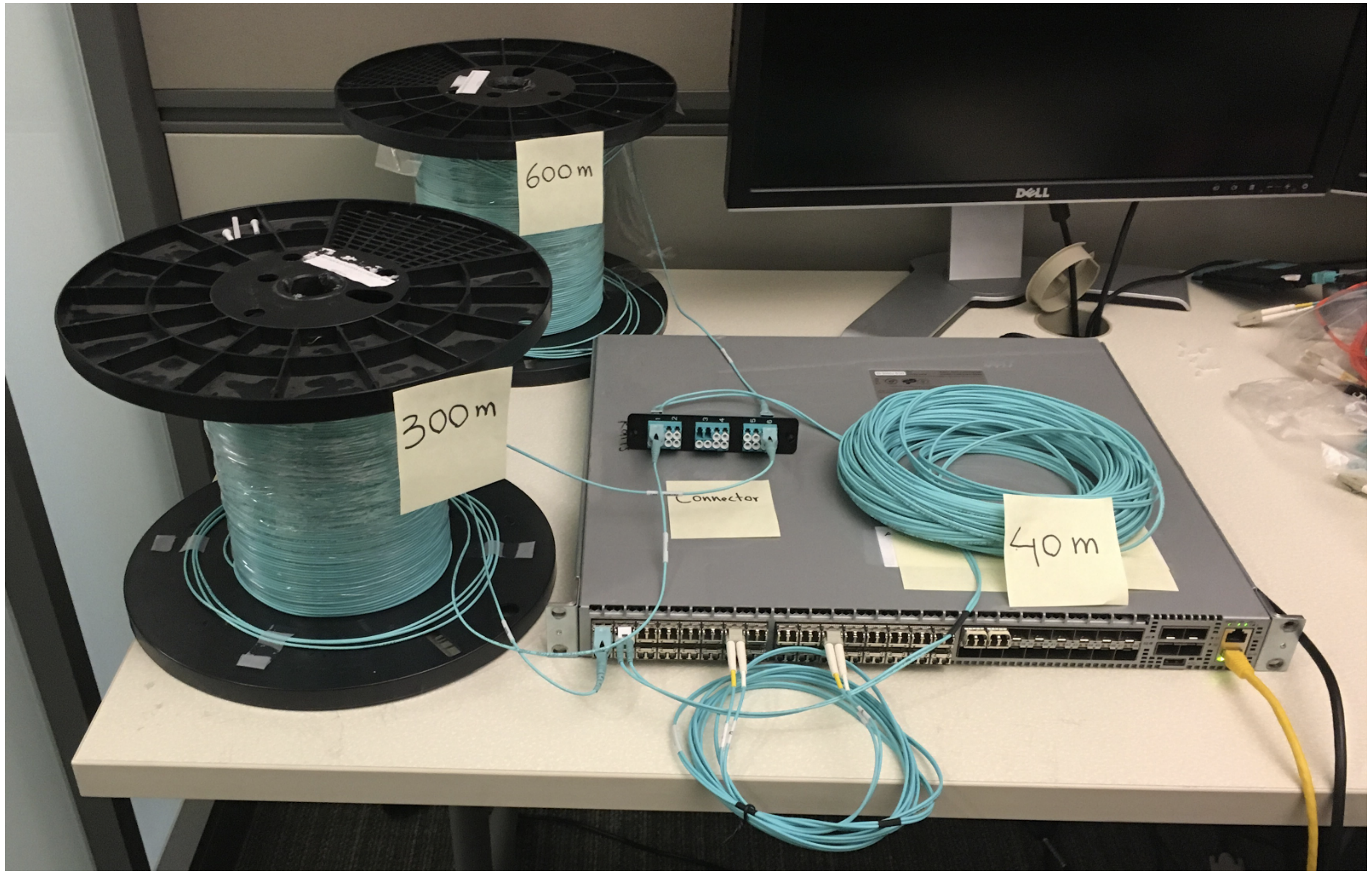
40G-SR4 200m 40G-SR4 $165/port

40G-LR4 10km 40G-LR4 $1249/port

Reducing DCN Cost



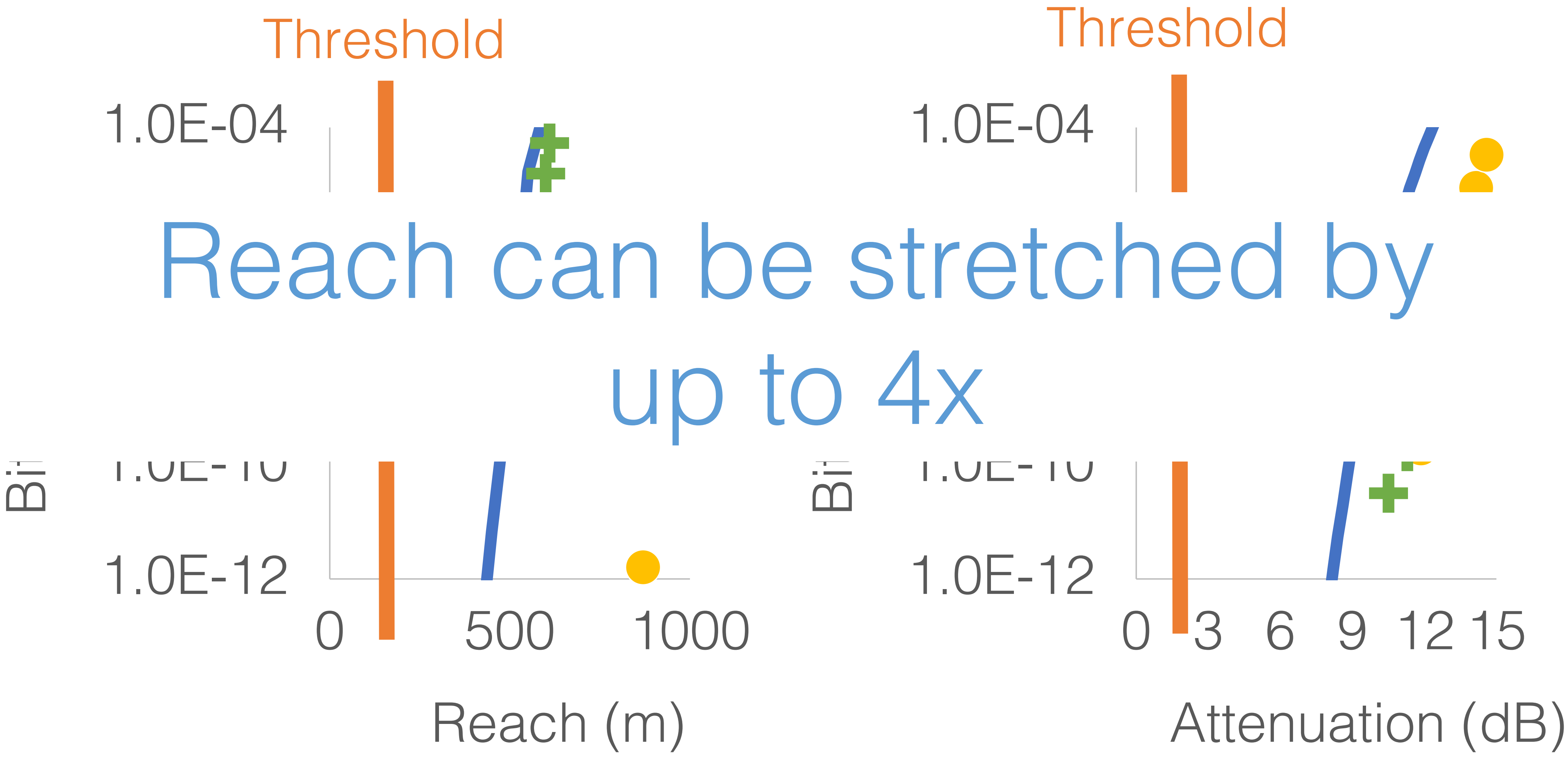
Quantifying Stretch Limit



- Concatenate short fibers to emulate a long fiber.
- Use optical attenuator to emulate dirty on optical connectors.
- Test packet error rate and convert it to bit-error rate (BER).
- Modeling effect of stretching on BER.

Quantifying Stretch Limit

40G-SR4



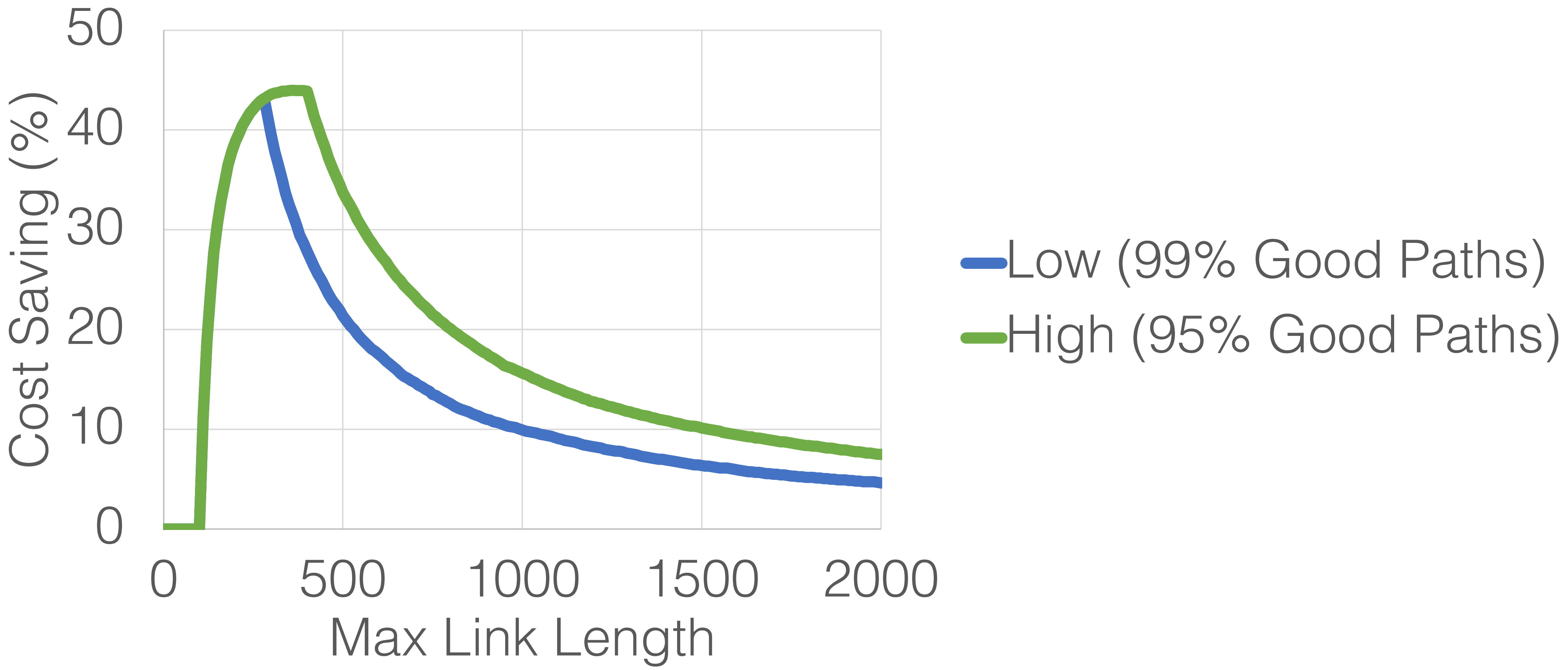
Reach can be stretched by up to 4x

Modeling Network Cost

- 3-stage Clos network (512 ToRs, 512 Aggs, 256 Cores)
- Uniform link length distribution (max length = 10m – 2 km)
    - Pick cheapest optical technology for each link ← Stretch
- Calculate the total DCN cost by summing up:
    - Fiber cost
    - Transceiver cost
    - Switch cost
    - Energy cost

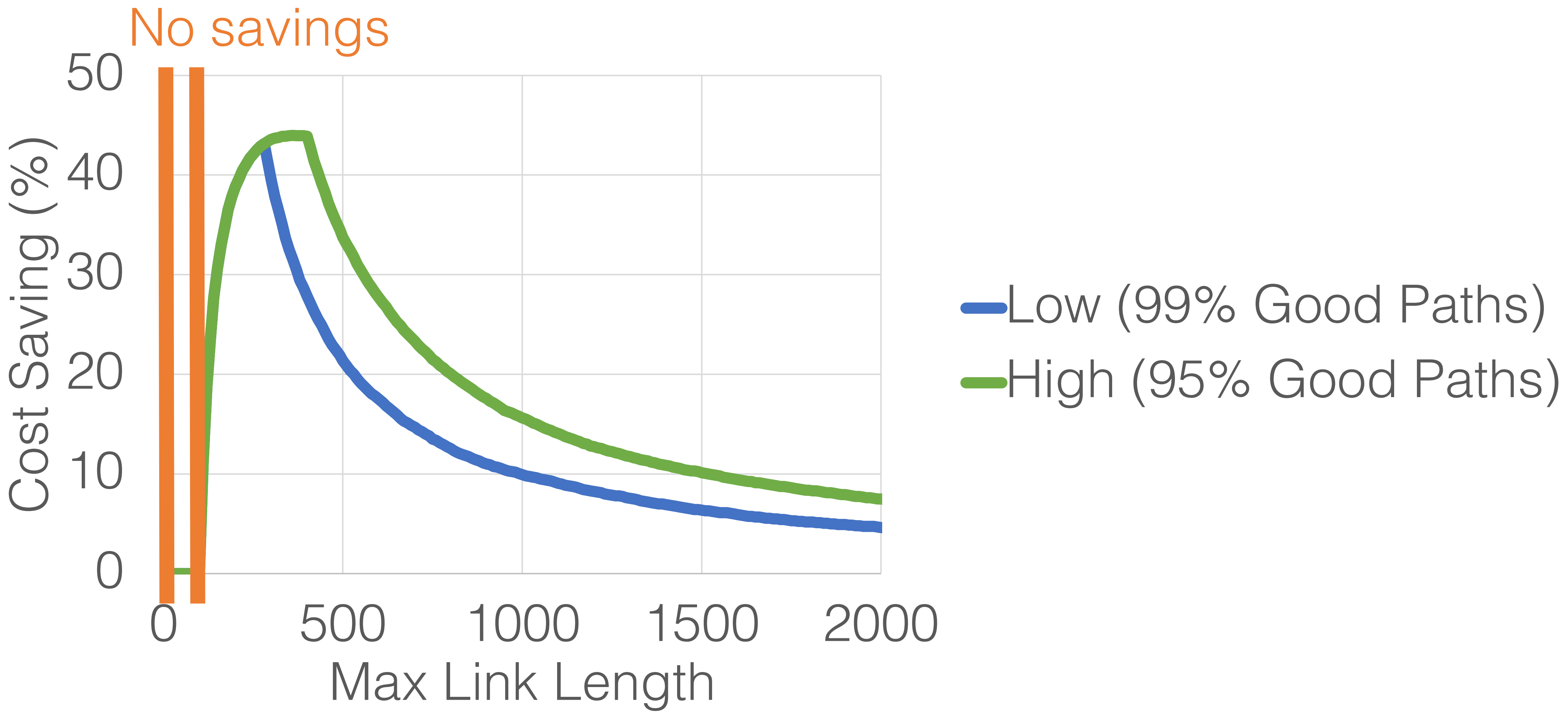
Cost Saving after Stretch

100m 40G-SR4
10km 40G-LR4



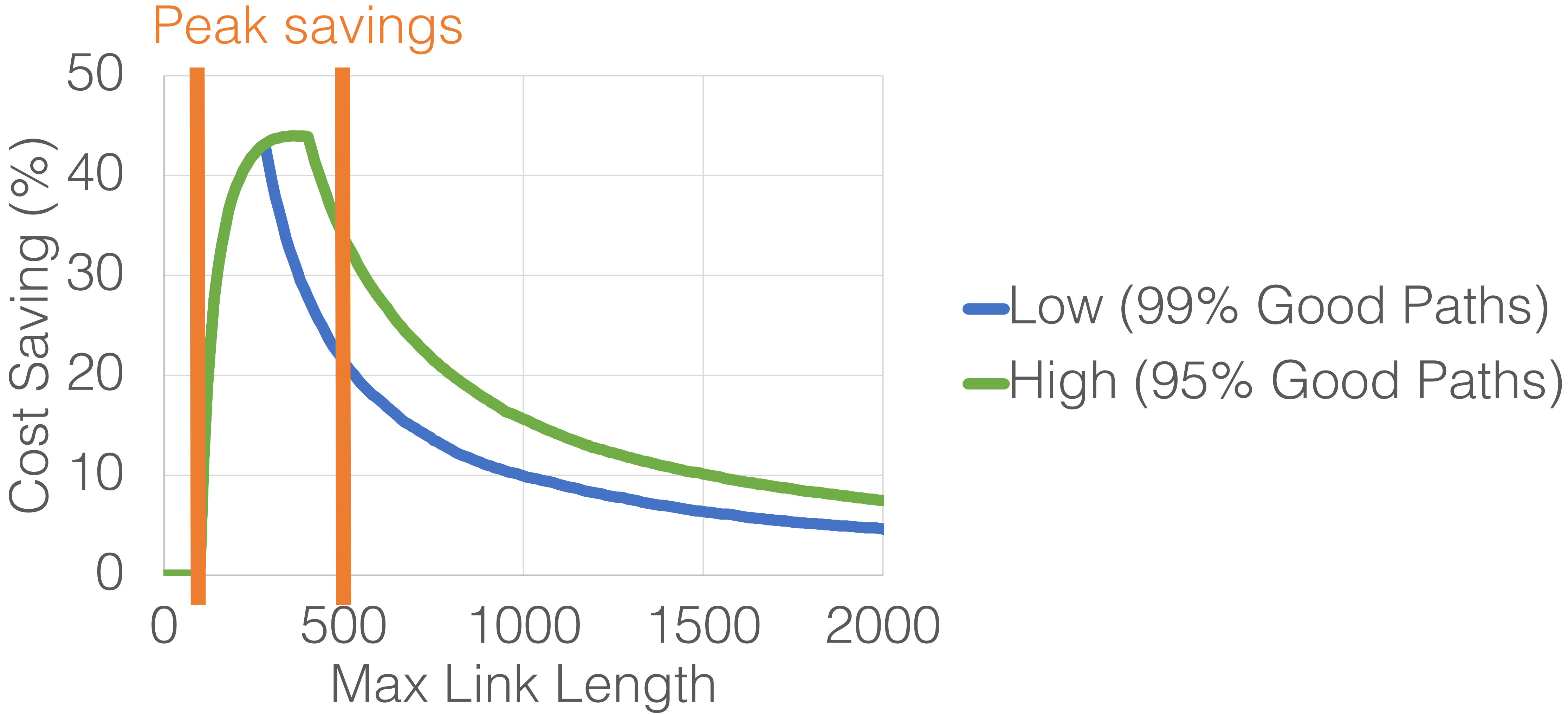
Cost Saving after Stretch

100m 40G-SR4
10km 40G-LR4



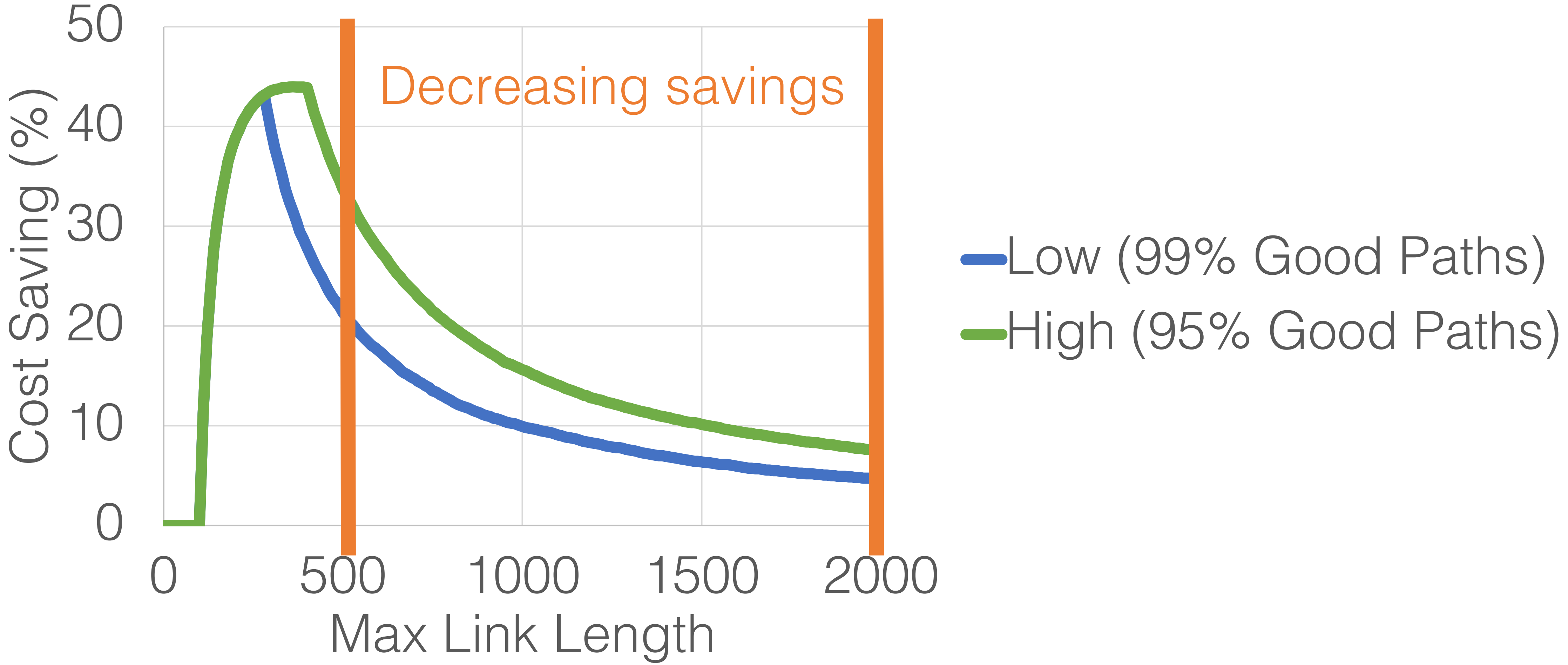
Cost Saving after Stretch

100m 40G-SR4
10km 40G-LR4



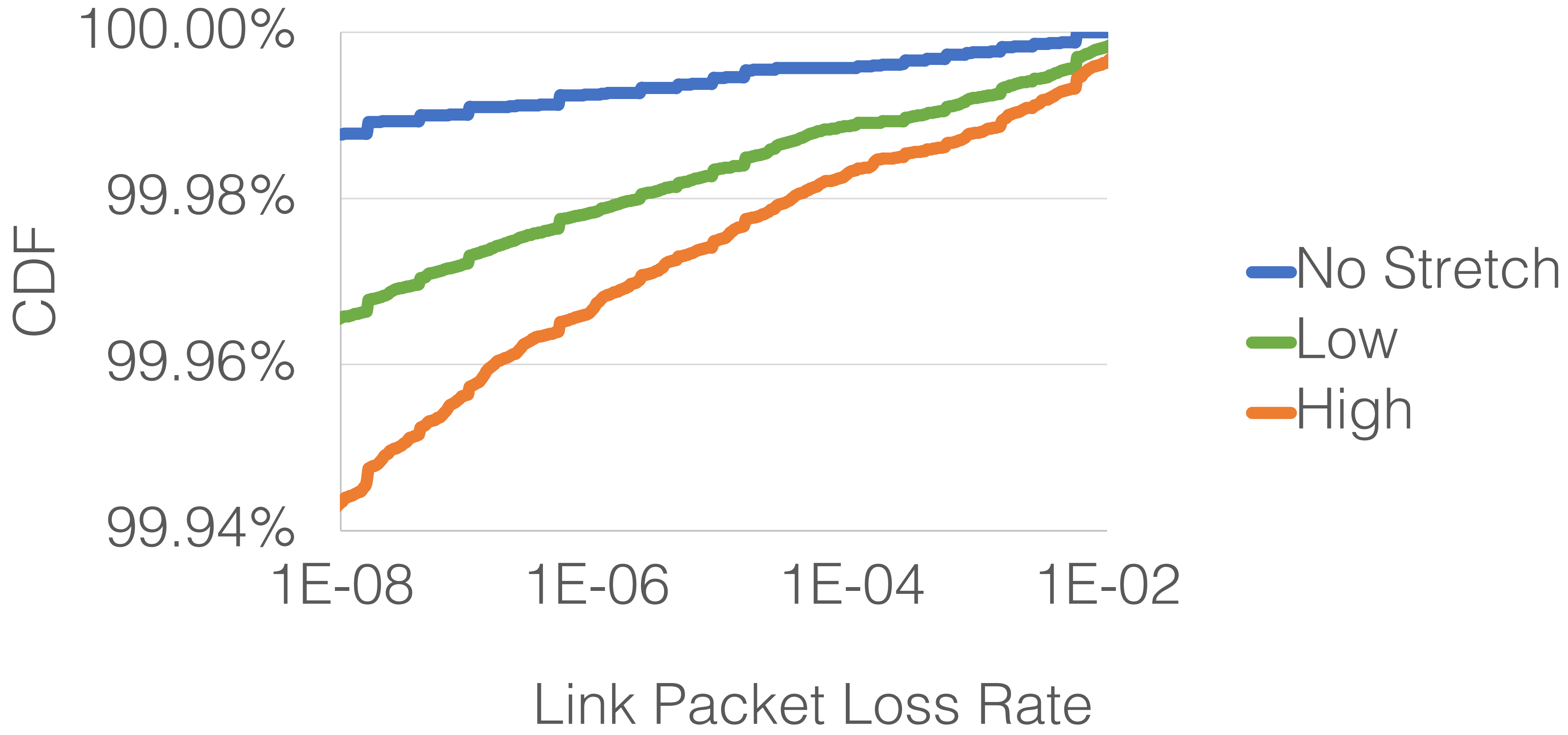
Cost Saving after Stretch

100m 40G-SR4
10km 40G-LR4



Impact on Packet Loss

0-500m, 40G



How to protect loss-sensitive applications from a small number of low loss links?

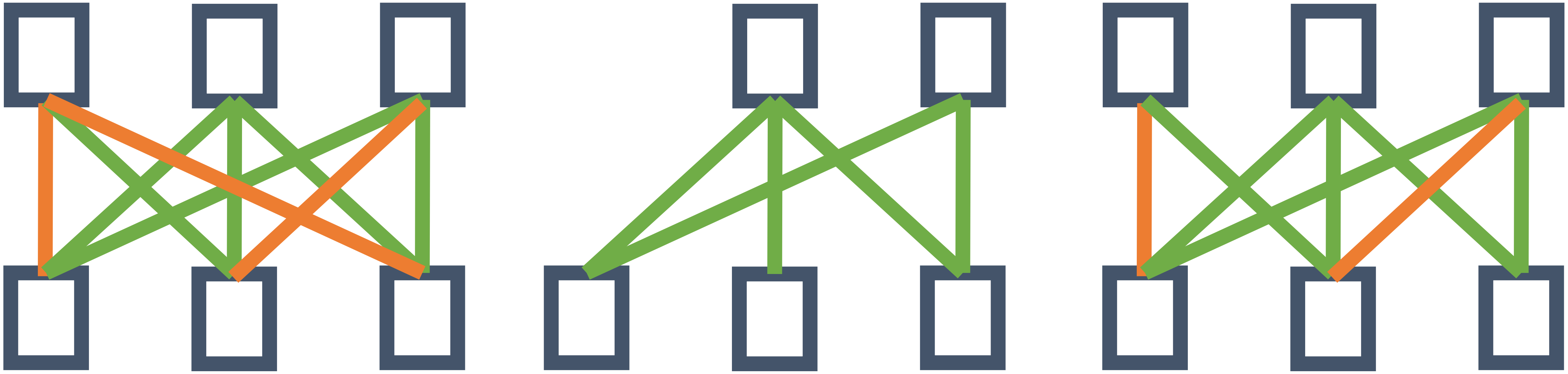
Possible Solutions

- Strawman: Source-routing
  - Source server picks a path that meets application requirement
  - Hard to scale
- Strawman: Error-correction code
  - Need to encode with per-path error rate to avoid bandwidth overhead
  - Latency of error correction for short flows

RAIL’s Approach

- Virtual Topology
    - Ensures a maximum end-to-end path packet error rate
    - Higher class virtual topology has higher loss rate
    - Applications choose virtual topology
- Error correction for higher class
    - Route with error-correction appropriate for path

Virtual Topologies



Physical Topology

0% path loss rate

0.1% path loss rate

🟩 0% link loss rate

🟧 0.1% link loss rate

Support Unmodified Applications

- Server exposes multiple virtual NICs where each virtual NIC corresponds to a path loss rate guarantee
- Applications simply choose a virtual NIC to bind
  - Flow is transparently error-encoded when traversing high loss path
  - Loss rate is queried from centralized controller

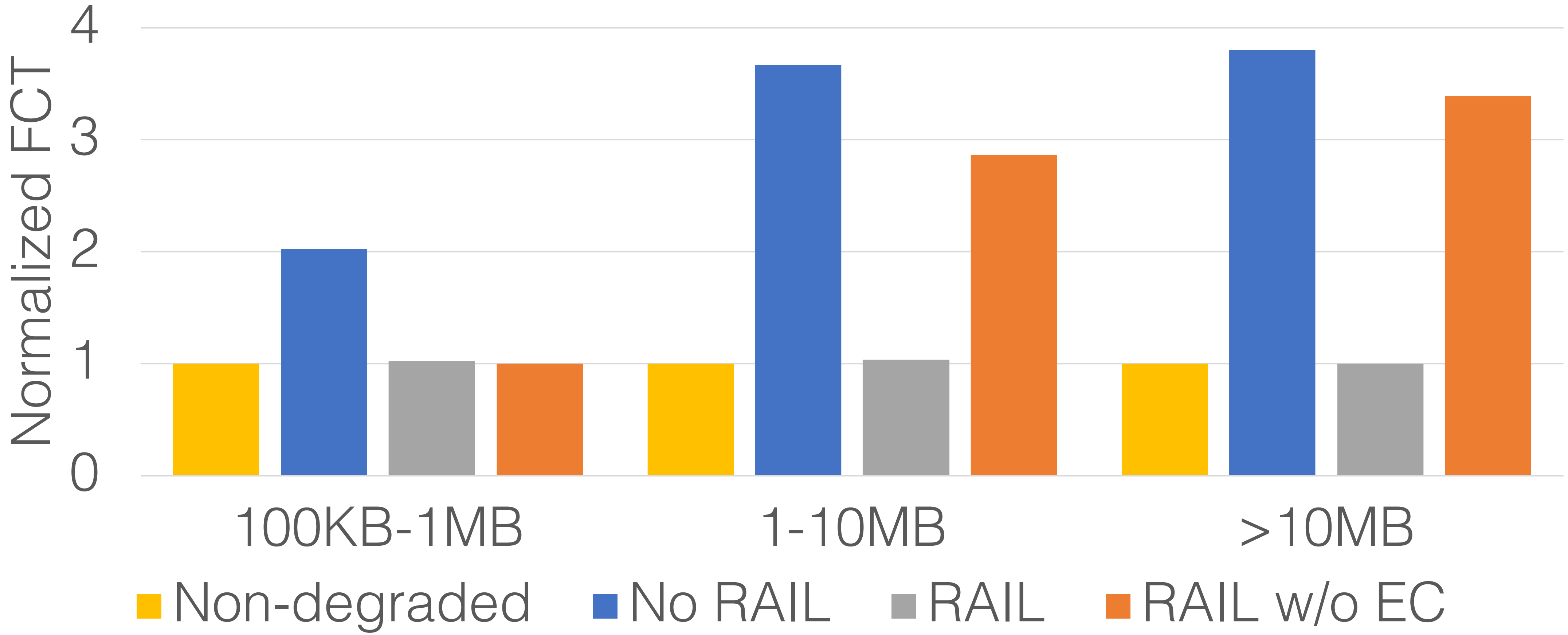
Testbed Experiments

- 3-stage Clos network (4 ToRs, 4 Aggs, 2 Cores)
  - 10G-SR optical technology
  - TCP CUBIC on Linux 3.19
- Use an optical attenuator to degrade the quality of a single link
- Two virtual topologies
  - Virtual topology #1: Without the degraded link
  - Virtual topology #2: With the degraded link

Evaluation Methodology

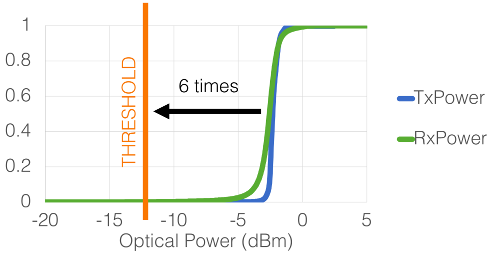
- Comparison
  - No RAIL
  - RAIL
  - RAIL w/o EC, use virtual topology #1 to protect flows less than 1MB
- Compute flow completion time normalized by performance on non-degraded network; Flow length distribution from pFabric.
  - Binned by flow sizes

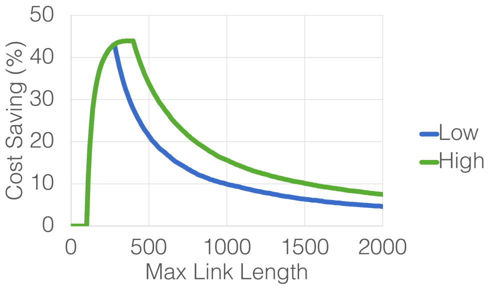
Evaluations

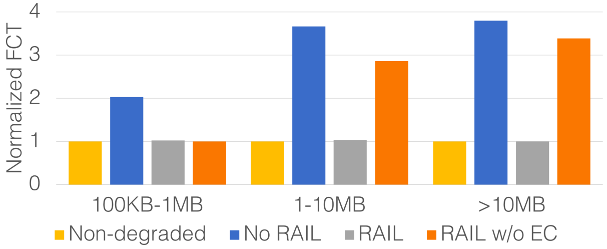


Summary

- Room for cost saving in optics used in DCN
- Reducing over-engineering
  - Stretching design reach limit for optical links can save up to 40%
- RAIL protects loss sensitive applications from packet loss due to reduction in over-engineering.


Optical Power (dBm)
THRESHOLD
6 times
TxPower
RxPower


Cost Saving (%)
Max Link Length
Low
High


Normalized FCT
100KB-1MB
1-10MB
>10MB
Non-degraded
No RAIL
RAIL
RAIL w/o EC