Enhancing Distributed Computing with Programmable and Open Optical Networks

Andrea Fumagalli
The University of Texas at Dallas

7th Annual International Workshop on Innovating the Network for Data-Intensive Science (INDIS 2020)

November 12, 2020

Contributors (UT Dallas)

- **Behzad Mirkhanzadeh**
- **Tianliang Zhang**
- Chencheng Shao
- Ali Shakeri
- Shunmugapriya (Priya) Ramanathan
- Joseph White-Swift
- GI Vania
- Miguel Razo
- Marco Tacca

Outline

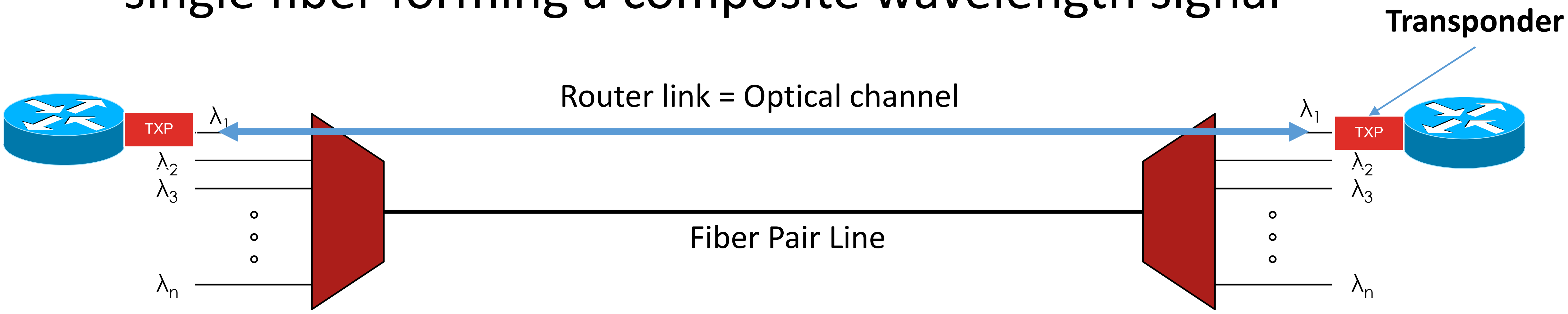
- Optical Networks: What is Unique?
- PROnet I: Using a Proprietary Solution
- Open Optical Network (OON) Efforts
- PROnet II: Embedding GNPy Modeling in Open Line Systems (OLS)
- PROnet III: OpenROADM with Six Optical Vendors
- Enhancing Distributed Computing
- Summary

Outline

- Optical Networks: What is Unique?
- PROnet I: Using a Proprietary Solution
- Open Optical Network (OON) Efforts
- PROnet II: Embedding GNPy Modeling in Open Line Systems (OLS)
- PROnet III: OpenROADM with Six Optical Vendors
- Enhancing Distributed Computing
- Summary

WSON Layer

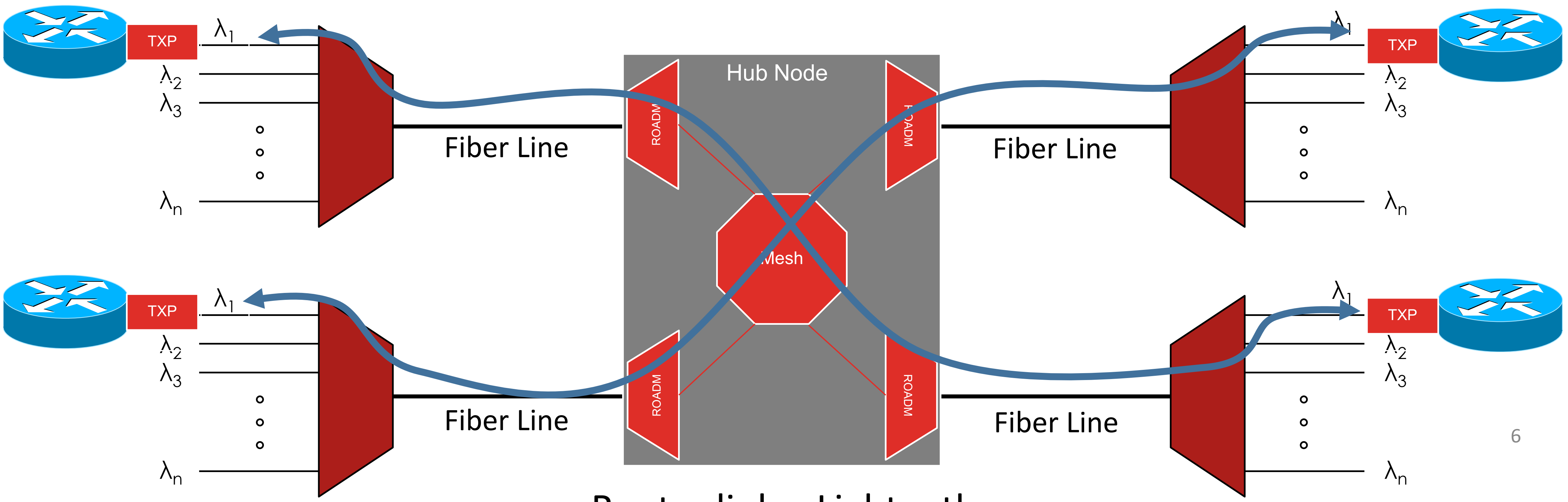
- Wavelength Switched Optical Network (WSON) layer makes use of Dense Wavelength Division Multiplexing (DWDM) technology to create multiple orthogonal channels in a single fiber forming a composite wavelength signal



- Each wavelength channel can be assigned to form a direct link between two client nodes (e.g., routers)

Reconfigurable Optical Add/Drop Multiplexer (ROADM)

- ROADMs can individually route wavelength channels across intermediate optical nodes
- End-to-end optical circuits (lightpaths) can be provisioned



Router link = Lightpath

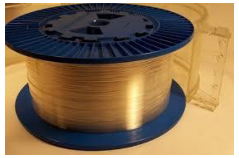
What is Different

- Connecting two Ethernet switches directly: cable length is known and must meet standards (e.g., 300m)
- Radio Link: it is single hop, between two radio heads (the base station and the user equipment)
- WDM circuits: analog signal traversing multiple devices which cause various **transmission impairments** -> **Signal integrity is affected**
  - Difficult to detect what is the cause of high BER at the receiver unless all devices closely work together (network control)
  - FEC are used to reduce BER (e.g., from 10^{-2} to 10^{-15})

Optical Signals and Devices

- Photons (from laser) are modulated
  - Non-coherent signals (legacy)
  - Coherent signals (new generation)

Colorless (Grey) Signal



Fiber Optics

Optical Splitter

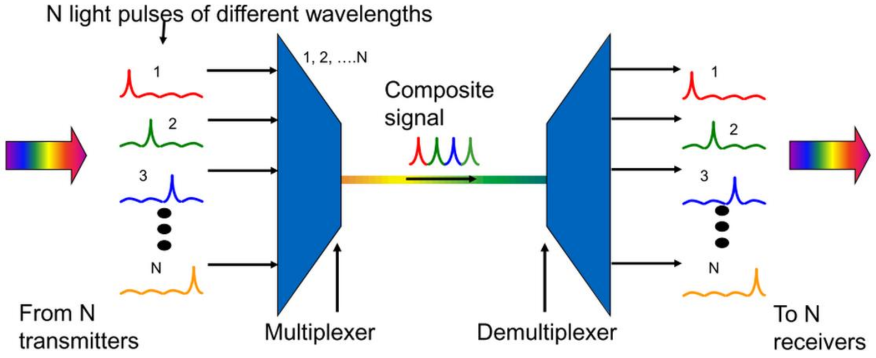
Optical Switch

Loss of Signal Power

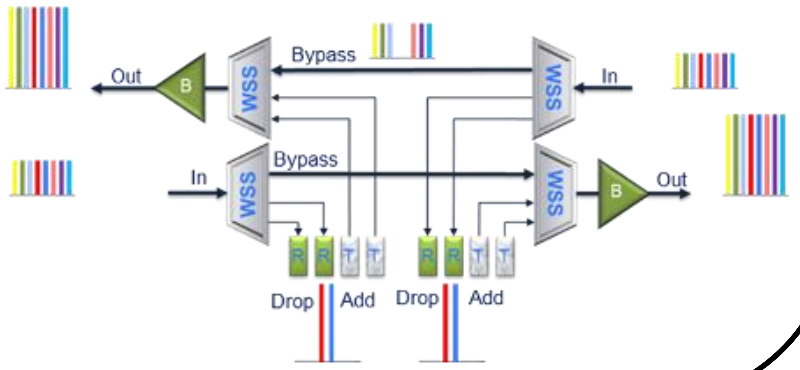
Optical Amplifier

Colored Signal

Mux/Demux



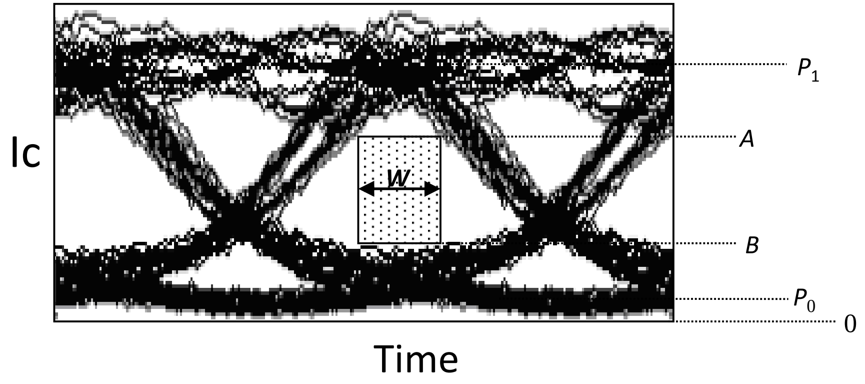
Wavelength Selective Switch (WSS)



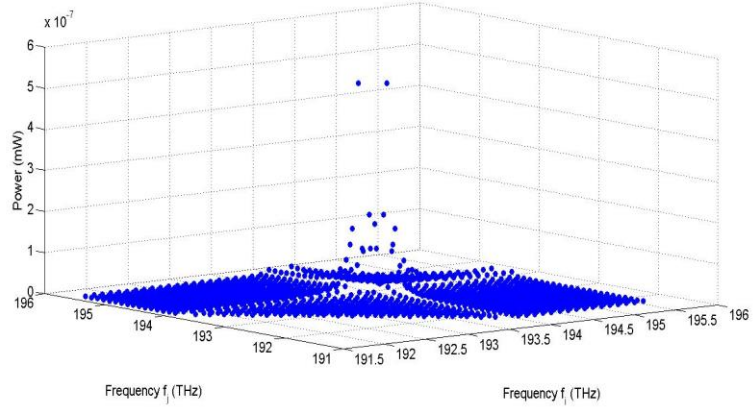
Physical Layer Impairment (PLI) Factors

- Traffic Independent PLIs (TI-PLIs)
  - ✓ power loss
  - ✓ amplified spontaneous emission (ASE) noise
  - ✓ chromatic dispersion (CD)
  - ✓ self-phase modulation (SPM)
  - ✓ polarization mode dispersion (PMD)
- Traffic Dependent PLI (TD-PLI)
  - ✓ four wave mixing (FWM)

Eye diagram at the receiver



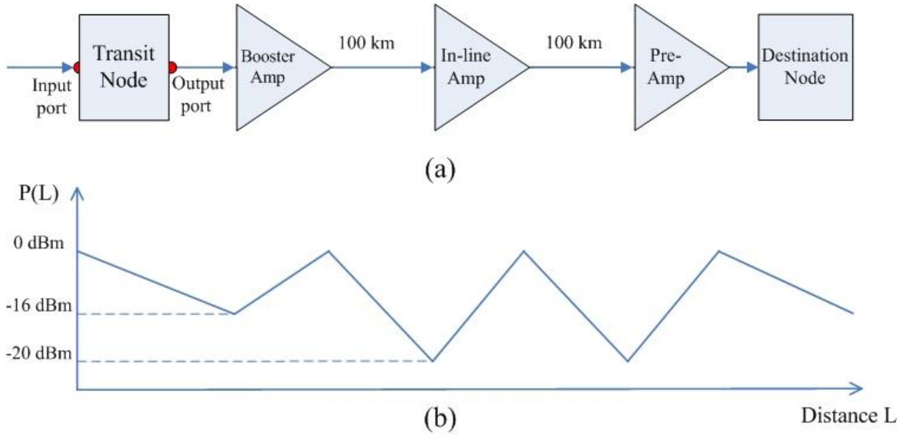
FWM power intensity at 194 THz when only 3 frequency channels are active



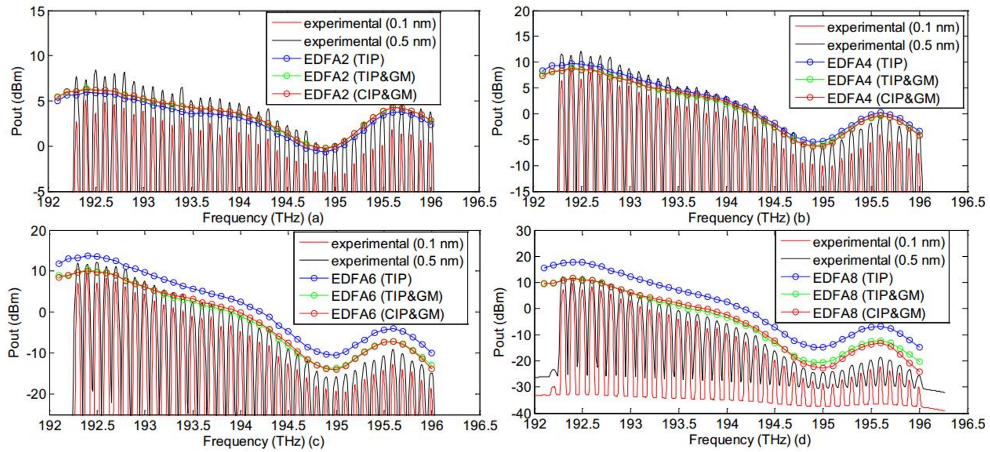
	Bandwidth Efficiency	Using Phase Information	Complexity	Compensation for Linear Impairments	Modulation Format
Non-coherent	Low	No	Low	Hard	OOK, ASK, FSK
Coherent	High	Yes	High	Easy	DPSK, PM-QPSK

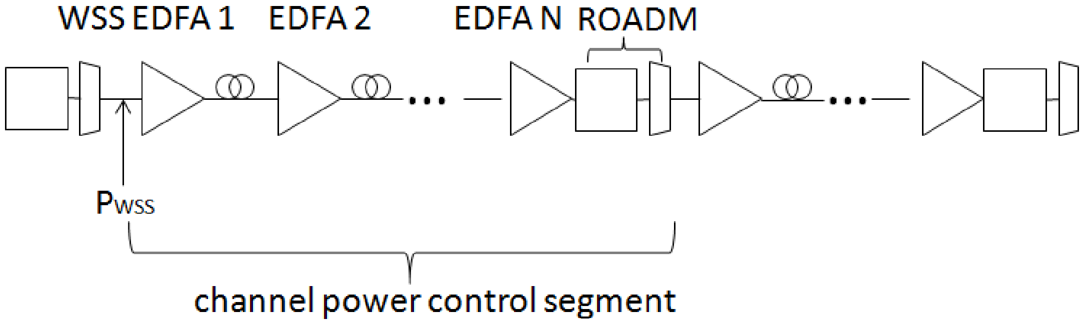
- ➢Intensity of FWM products is higher from close neighbors
- ➢+/-k neighbors are considered (k=3), capturing major interference and reducing the complexity
- ➢XTm threshold to ensure signal quality due to TD-PLI

Optical Signal to Noise Ratio (OSNR) and Automatic Signal Power Control (APC) Strategies



Non-flat gain of amplifier





In order to compensate for the power loss, optical amplifiers are periodically placed along the fiber

EDFA gain

Noise bandwidth

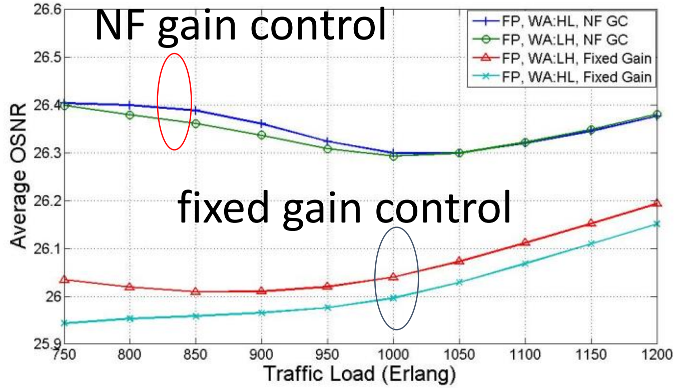
$$\mathrm{OSNR} = \frac{P_{\mathrm{ch}}}{P_{\mathrm{ASE}}} = \frac{P_{\mathrm{ch}}}{NF(G-1)h\upsilon\Delta f}$$

Noise figure

h: Planck constant

v: channel frequency

$$\mathrm{OSNR}_{end-to-end} = \frac{1}{\sum_{i=1}^{N}\frac{1}{OSNR_i}}$$



ASE noise, together with optical signal, is amplified by erbium-doped fiber amplifiers (EDFAs)

Are Optical Networks Becoming (SDN) Programmable?

- “Taking a cue from IT's separation of hardware, operating systems and applications software and, more recently, the separation of compute, storage and networking in data centers, the trend toward disaggregation and open optical networking is starting to impact the broader communications equipment market.”
- “Approaches to optical networking based on disaggregation and software-defined networking control are set to dominate – while key questions still remain.”
- “This impact is already being felt with the shift to SDN, disaggregating the control plane from the forwarding plane, and the shift to network functions virtualization (NFV), disaggregating network hardware from software functions.”

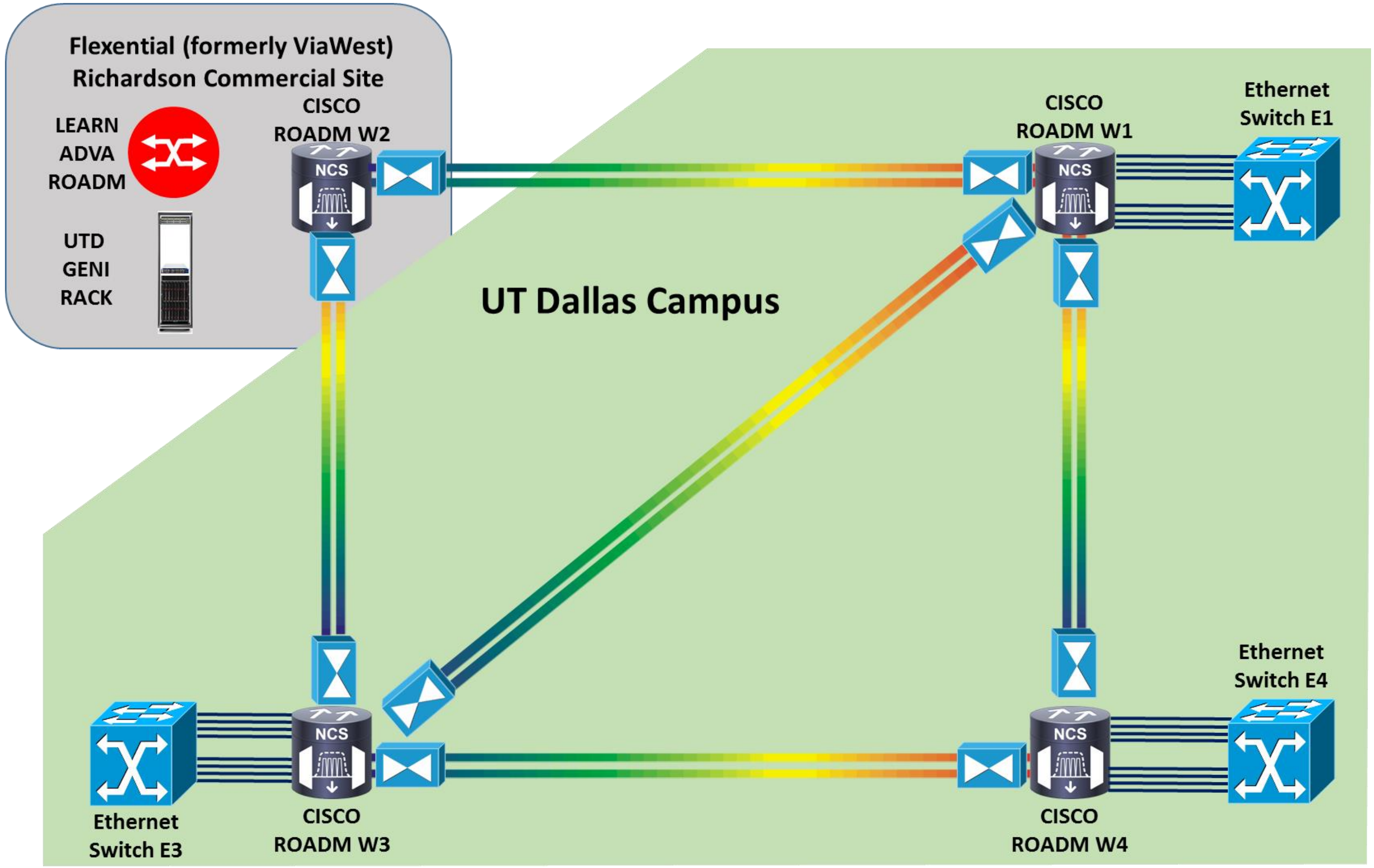
Outline

- Optical Networks: What is Unique?
- PROnet I: Using a Proprietary Solution
- Open Optical Network (OON) Efforts
- PROnet II: Embedding GNPy Modeling in Open Line Systems (OLS)
- PROnet III: OpenROADM with Six Optical Vendors
- Enhancing Distributed Computing
- Summary

History of PROnet

- PROnet stands for Programmable Optical network, invoking the use of Software Defined Networking (SDN) principles applied to the optical network physical layer
- The concept of implementing and deploying PROnet as *both a REN and test-bed* originated during a number of meetings held by the PIs of a former JUNO project titled “ACTION” in June of 2014 (NSF-NICT JUNO workshop at UC Davis)
    - Malathi Veeraraghavan (UVA)
    - Naoaki Yamanaka (Keio University)
    - Eiji Oki (Kyoto University)
    - Andrea Fumagalli (UT Dallas)
- The PROnet concept was driven by the PIs’ interest in experimentally testing technologies and expected advantages to the applications that may result from automatically reconfiguring the optical network on-demand and through well-defined APIs
- Funding for implementing PROnet as a multi-layer optical network was provided by NSF CC*DNI grant #1541461 in September of 2015
- PROnet main collaborative milestones are described in the next slides

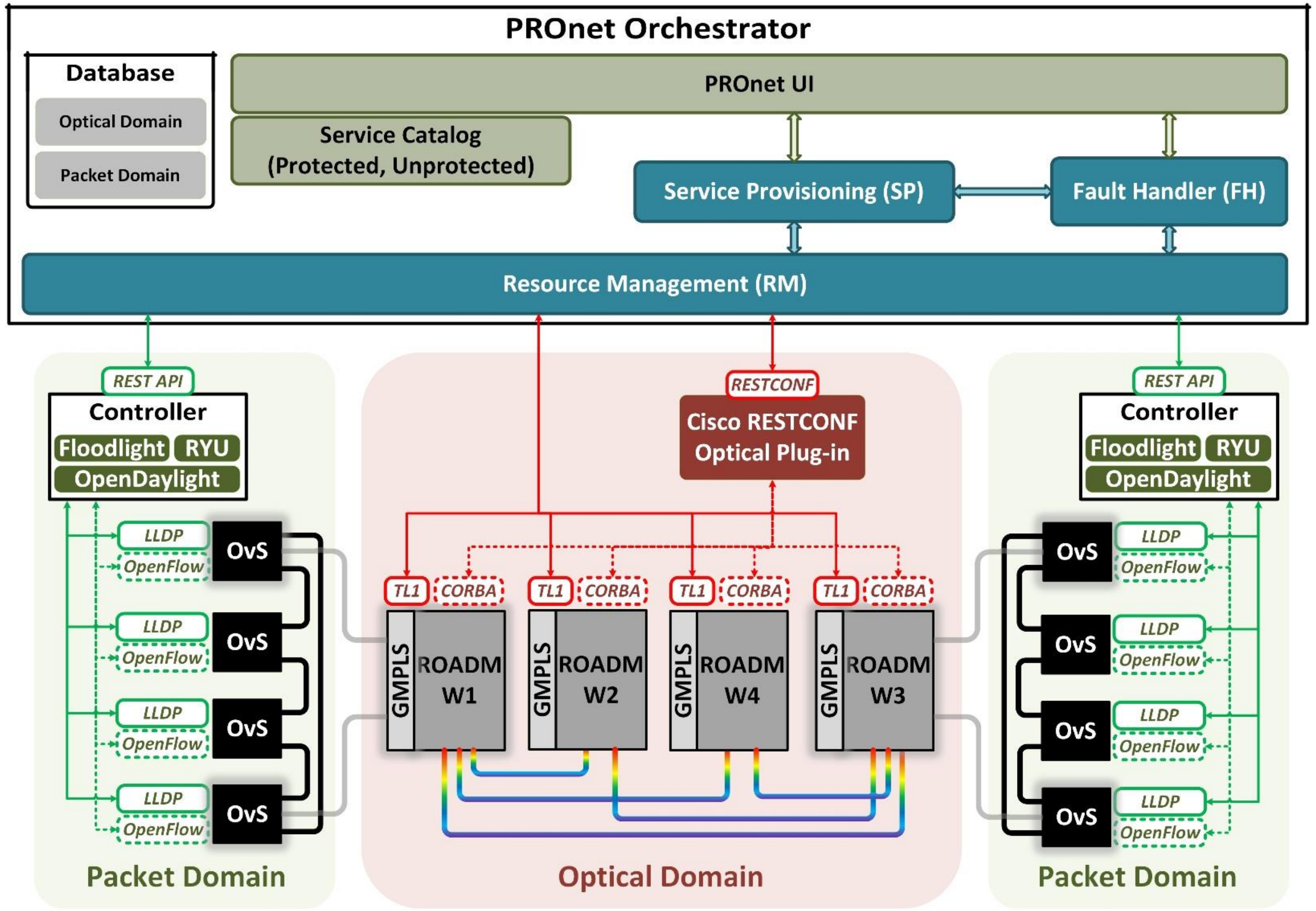
PROnet as a REN at and Around the UT Dallas Campus
(PROnet I: A Single Vendor Solution)



- 4 Cisco NCS 2k ROADM nodes
  - CDC capable
  - Flex Grid capable
  - Transponders supporting 100 and 200 Gbps line transmission rates
- Muxponders can adapt to 10 x 10 Gbps ports of Ethernet switches
- One ROADM is co-located with
  - Adva ROADM of LEARN (Texas REN)
  - GENI rack hosted by UT Dallas
  - Other compute and storage resources

UT Dallas OIT deployed PROnet in Spring/Summer of 2017

The PROnet I: Ethernet-over-WDM SDN Orchestrator



- Interfaces to work with
  - RYU
  - Floodlight
  - OpenDaylight
  - Cisco RESTCONF Optical Plug-in
  - TL1
  - CORBA
- Offered services
  - Unprotected end-to-end flow
  - Protected 1:1 (fiber disjoint) end-to-end flow
  - Highly reliable protection/restoration 1:1+R mechanism, to automatically reconfigure (redesign) the optical network to pre-failure condition

UT Dallas researchers started to develop the PROnet SDN Orchestrator in January of 2016

Lightpath Restoration Procedure and Time

Procedure

- Generalized Multiprotocol Label Switching (GMPLS) is responsible for computing and establishing restoration lightpaths upon link failure
- IEEE RFC 6163 and 6566 describe standards and requirements

How does it work?

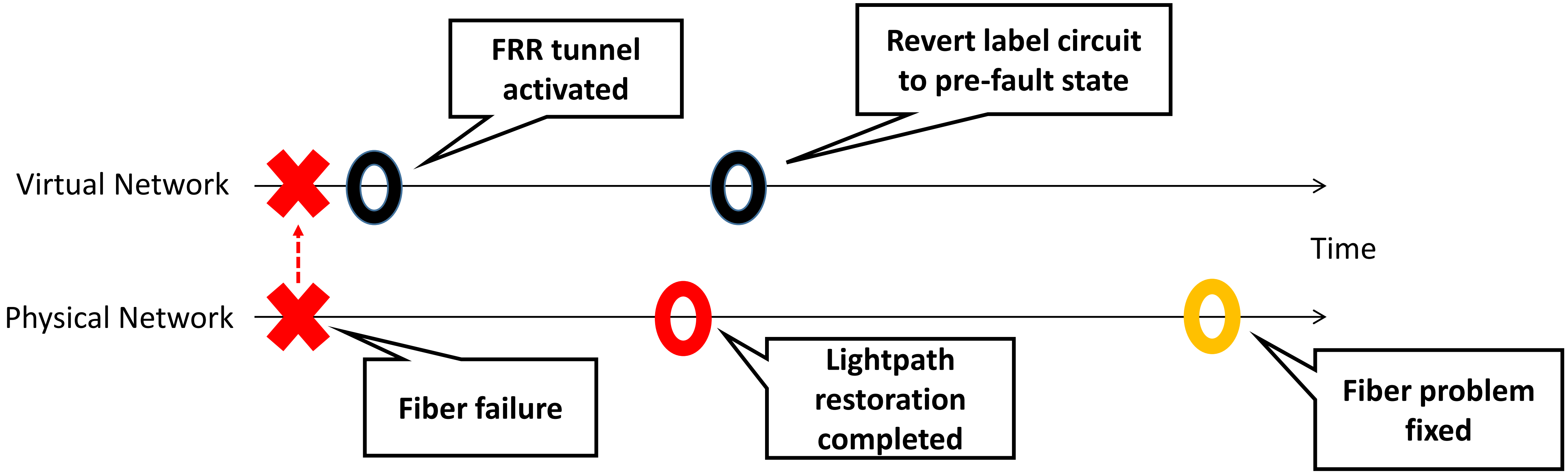
- Automatically switches circuits away from failed or impaired paths
- Can use any available wavelength, but requires ROADMs and tunable TXPs

Completion time depends on

- Failure detection and signaling to inform lightpath end-nodes
- Procedure to compute route and wavelength assignment (RWA)
- Procedure to configure optical devices, provision lightpath, and ensure desirable circuit bit error rate (BER)

Multi-Layer Restoration – 1:1+R

- Virtual Network is back to pre-fault state and FRR tunnel can be used to recover from a subsequent second network failure
- Race conditions between the two recovery mechanisms is avoided thanks to the different time scales
  - FRR protection mechanism responds in milliseconds
  - Lightpath restoration mechanism responds in seconds



Outline

- Optical Networks: What is Unique?
- PROnet I: Using a Proprietary Solution
- Open Optical Network (OON) Efforts
- PROnet II: Embedding GNPy Modeling in Open Line Systems (OLS)
- PROnet III: OpenROADM with Six Optical Vendors
- Enhancing Distributed Computing
- Summary

What is an Open Optical Network (OON)

- “*We define a fully open optical network as one that includes open hardware (transponders and line system equipment) supporting open application programming interfaces (APIs) that can interact with and be managed by open source software*” [*Heidi Adams*]
- ***Disaggregation of optical components*** to avoid/alleviate *proprietary or customer lock-in* problem (high cost to switch to another equipment vendor)
- OON Efforts
  - Open Networking Foundation (ONF)
  - Optical Interworking Forum (OIF)
  - Telecom Infra Project (TIP)’s Open Optical Packet Transport Group
  - Open ROADM Multi-Source Agreement (MSA)

Optical Transport Market Outlook by Telecoms Market Analyst Dell'Oro (Summer 2020)

"Demand for bandwidth – the main driver behind optical market growth – has been given a boost by the recent pandemic."

"This market – largely comprised of DWDM systems – is expected to expand in 2020 and for the next five years reaching nearly US$18 billion."

"Demand for optical transport gears for data center interconnect is expected to take a turn in the near future, with **disaggregated WDN transponder** unit sales annually growing at a double-digit percentage rate."

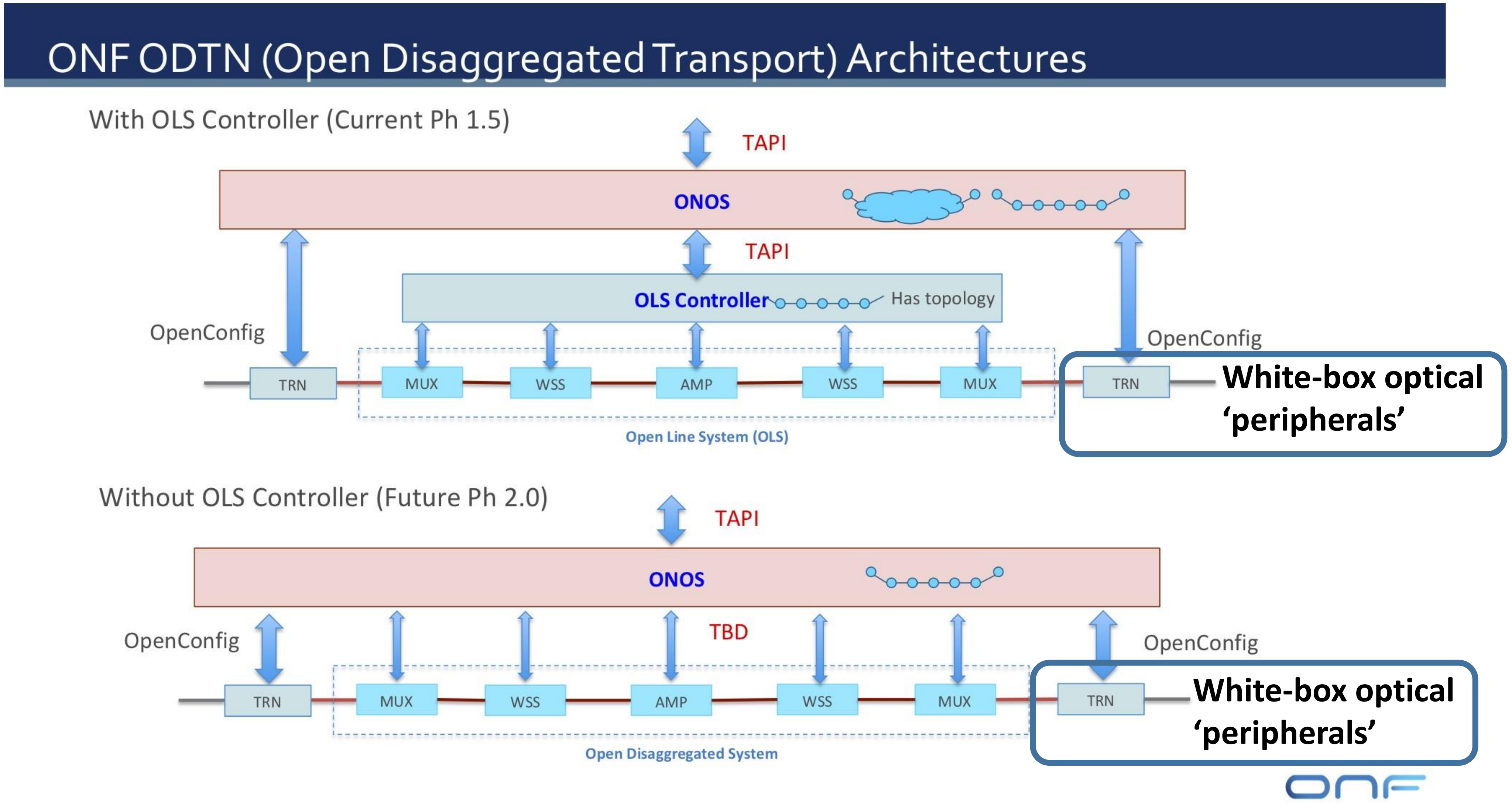
Two Missions

- Make optical equipment use standards procedures and APIs
- Define mechanisms to ensure optical signal integrity in the presence of multiple players in "*a world of analog signals*"

Two Interesting Approaches

- TIP's Open Optical Packet Transport Group
  - Open Disaggregated Transport (ODTN) stressing high performance in WAN deployment
- OpenROADM MSA
  - Stressing full hardware disaggregation in MAN deployment

Open and Disaggregated Transport (ODTN) Architecture



Objective: to disaggregate transponders from (open) line systems

https://www.opennetworking.org/odtn/?utm_referrer=https%3A%2F%2Fwww.google.com%2F

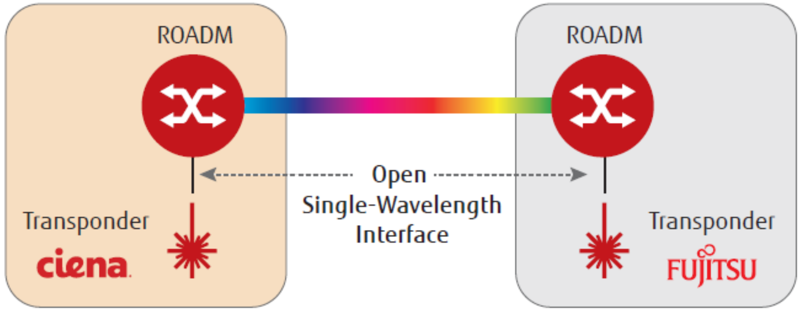
OpenROADM MSA

Open Wavelength Interface (W-Spec) Between Transponders

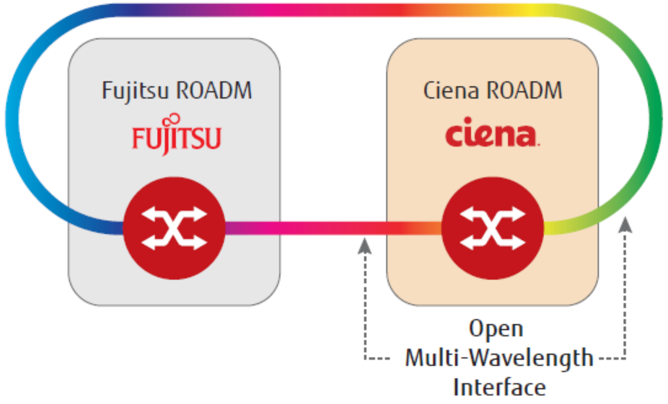
Single-Wavelength Interoperability

Transponder-to-transponder specifications include:

- Electrical framing definition
- Digital signal processor interoperation
- Forward Error Correction definition
- Optical transmission definition
- Optical path definition
- Optical receiver definition



Open Multi-Wavelength Interface (MW-Spec) Between ROADMS



Multi-Wavelength Interoperability

ROADM-to-ROADM specifications include:

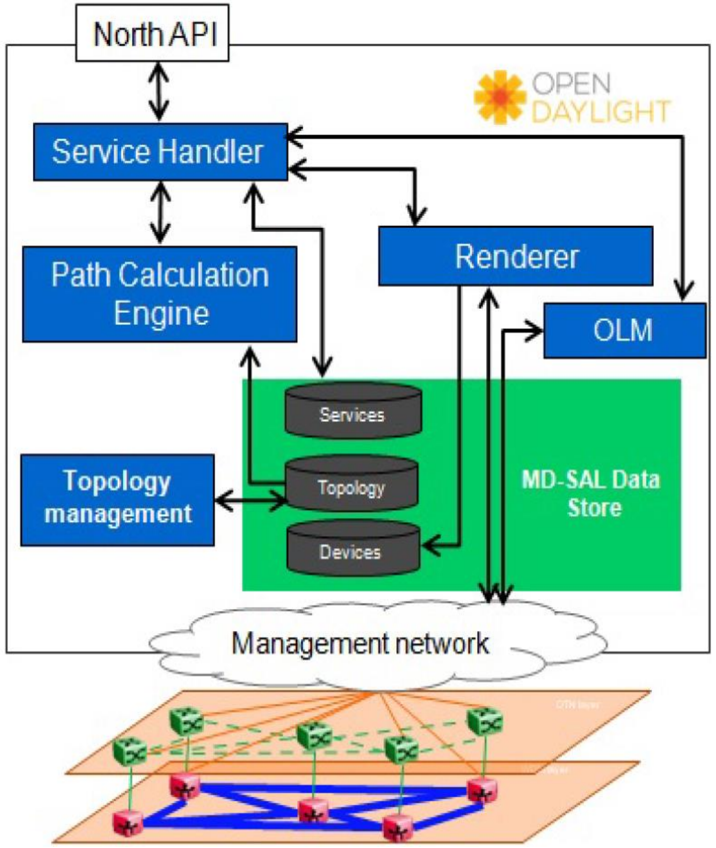
- Operating range characteristics
- Amplifier operational parameters
- ROADM operational parameters
- Optical Service Channel
- Optical power control
- Safety shutdown

<u>Objective: to achieve full disaggregation of optical components</u>

http://OpenROADM.org

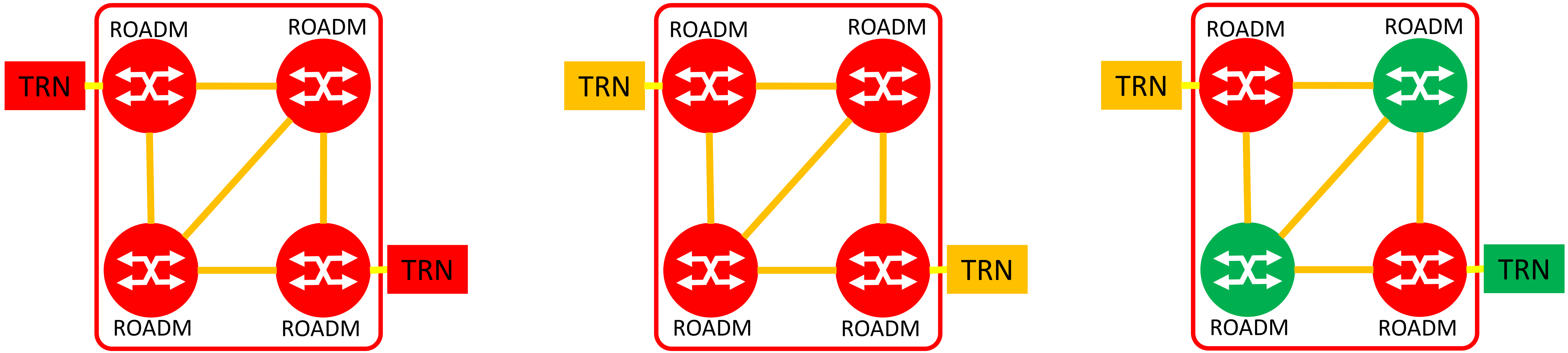
OpenROADM TransportPCE

- Open source OpenROADM control platform
- Implemented by AT&T, Orange, and other groups as a plugin on OpenDaylight controller



OLM = Optical Line Manager

Comparison of Architectures



(a) Proprietary Single Vendor **(b) Open Disaggregated Transport (ODTN)** **(c) Open ROADM**

Single Vendor ⟷ Multi Vendor

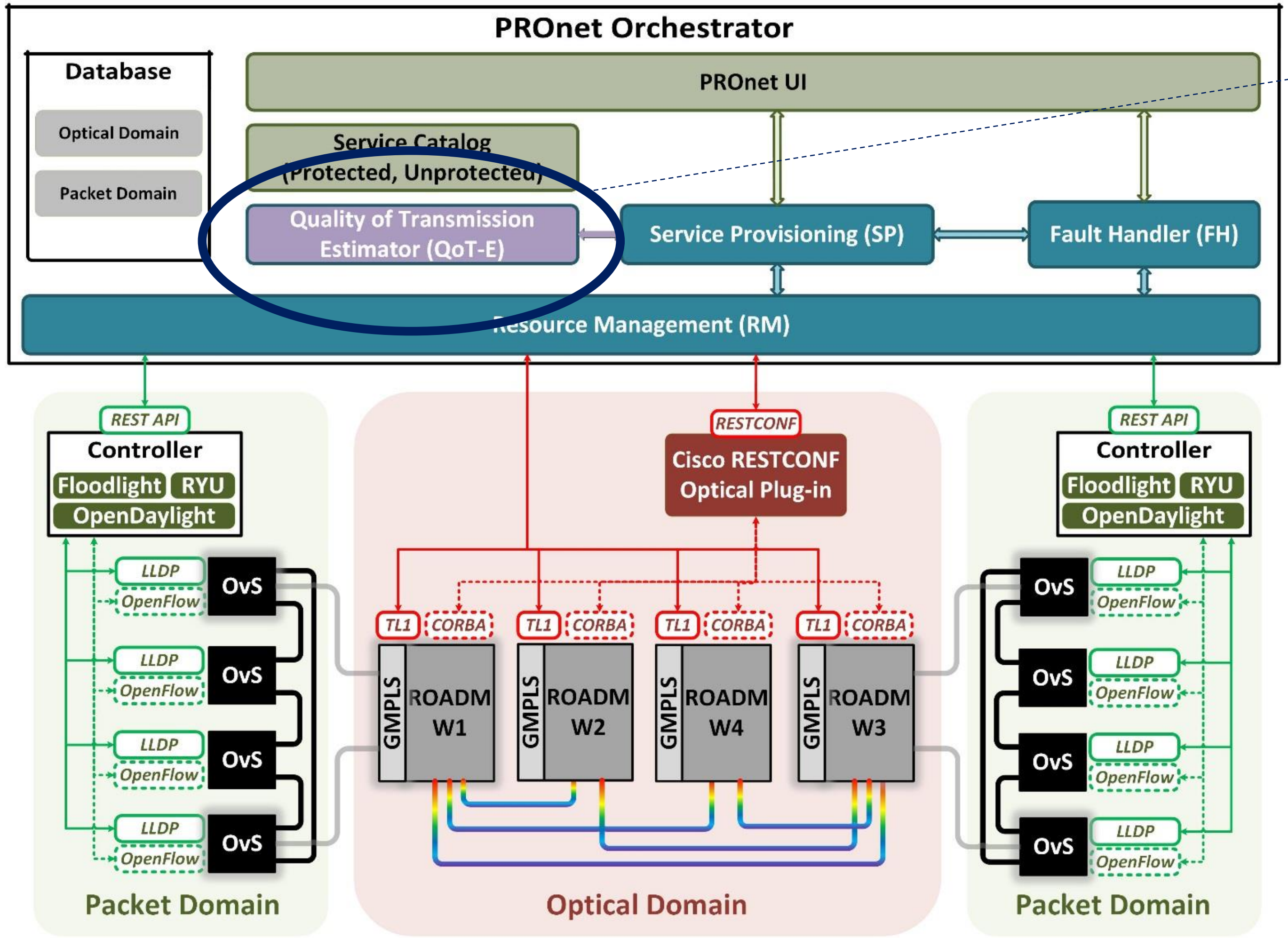
High performance through proprietary coherent DSP capabilities and FEC

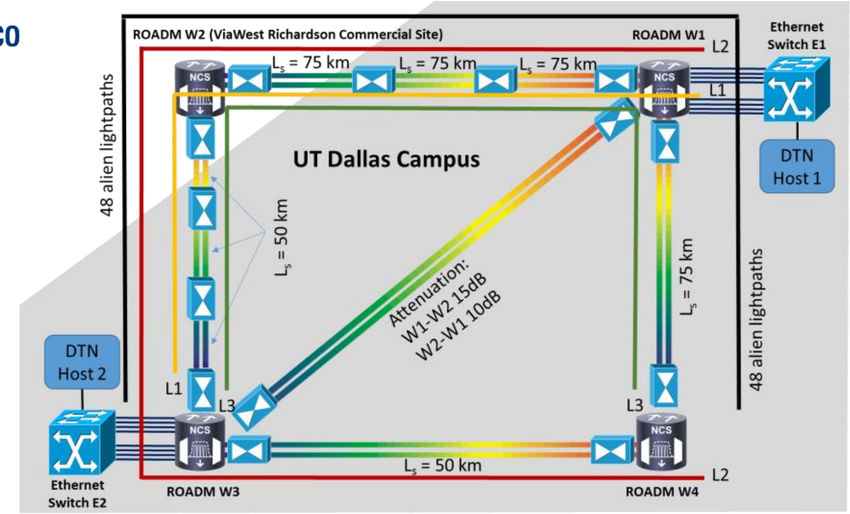
Open and interoperable optical transponders and line system equipment

Outline

- Optical Networks: What is Unique?
- PROnet I: Using a Proprietary Solution
- Open Optical Network (OON) Efforts
- PROnet II: Embedding GNPy Modeling in Open Line Systems (OLS)
- PROnet III: OpenROADM with Six Optical Vendors
- Enhancing Distributed Computing
- Summary

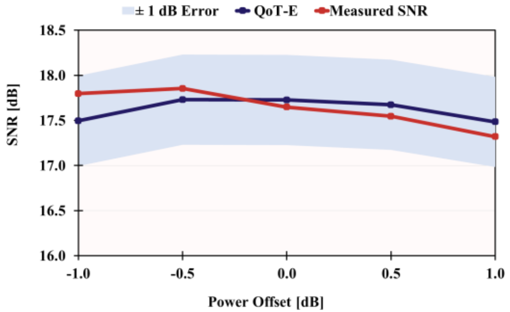
PROnet II: GNPy-based Quality of Transmission Estimator (QoT-E) Module





Two Ph.D. students from PoliTo visited the UT Dallas lab in 2017 to design and develop the QoT-E module, and validated results from the module against experimental results obtained with PROnet equipment

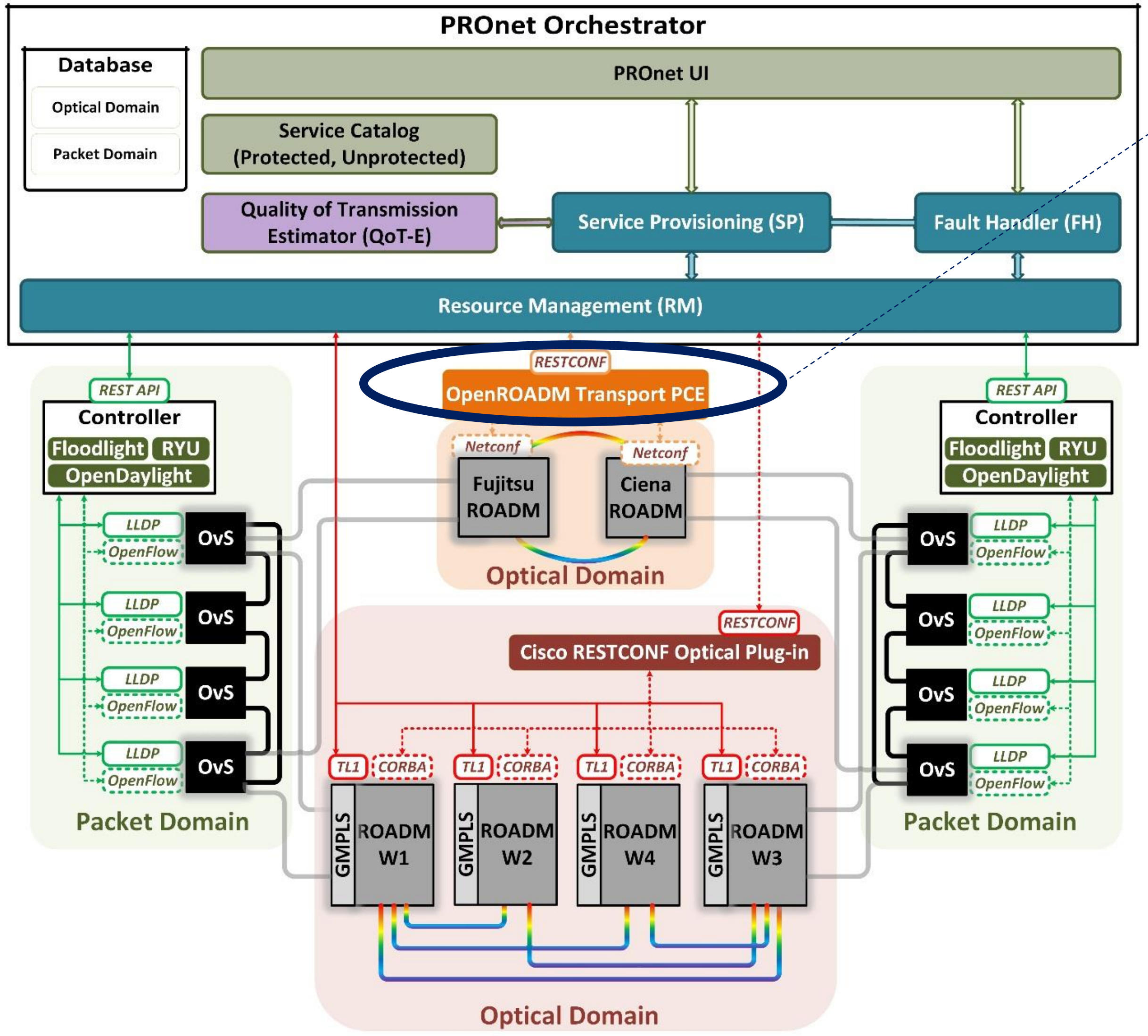
- Cisco provided optical line amplifiers
- PoliTo team provided the QoT-E software module for estimating lightpath OSNR while accounting for both linear and non-linear transmission effects
- QoT-E module makes use of models defined by TIP OOPT-PSE Technical Working Group
- Models were validated using 4 test-
  - Orange Labs
  - Facebook Labs
  - Microsoft Labs
  - UTD Lab



Outline

- Optical Networks: What is Unique?
- PROnet I: Using a Proprietary Solution
- Open Optical Network (OON) Efforts
- PROnet II: Embedding GNPy Modeling in Open Line Systems (OLS)
- PROnet III: OpenROADM with Six Optical Vendors
- Enhancing Distributed Computing
- Summary

PROnet III: OpenROADM Solution with Six Vendors



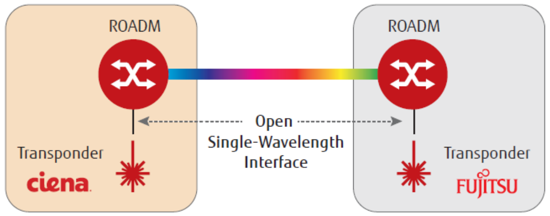
- Interoperate six vendors' ROADMs and transponders (100 Gbps) to demonstrate open interfaces at the wavelength and multi-wavelength layer

Open Wavelength Interface (W-Spec) Between Transponders

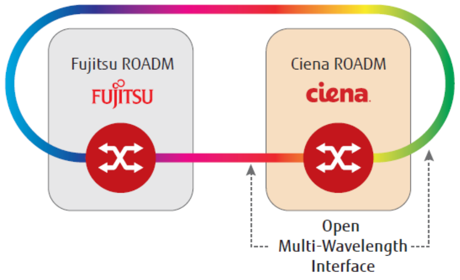
Single-Wavelength Interoperability

Transponder-to-transponder specifications include:

- Electrical framing definition
- Digital signal processor interoperation
- Forward Error Correction definition
- Optical transmission definition
- Optical path definition
- Optical receiver definition



Open Multi-Wavelength Interface (MW-Spec) Between ROADMS



Multi-Wavelength Interoperability

ROADM-to-ROADM specifications include:

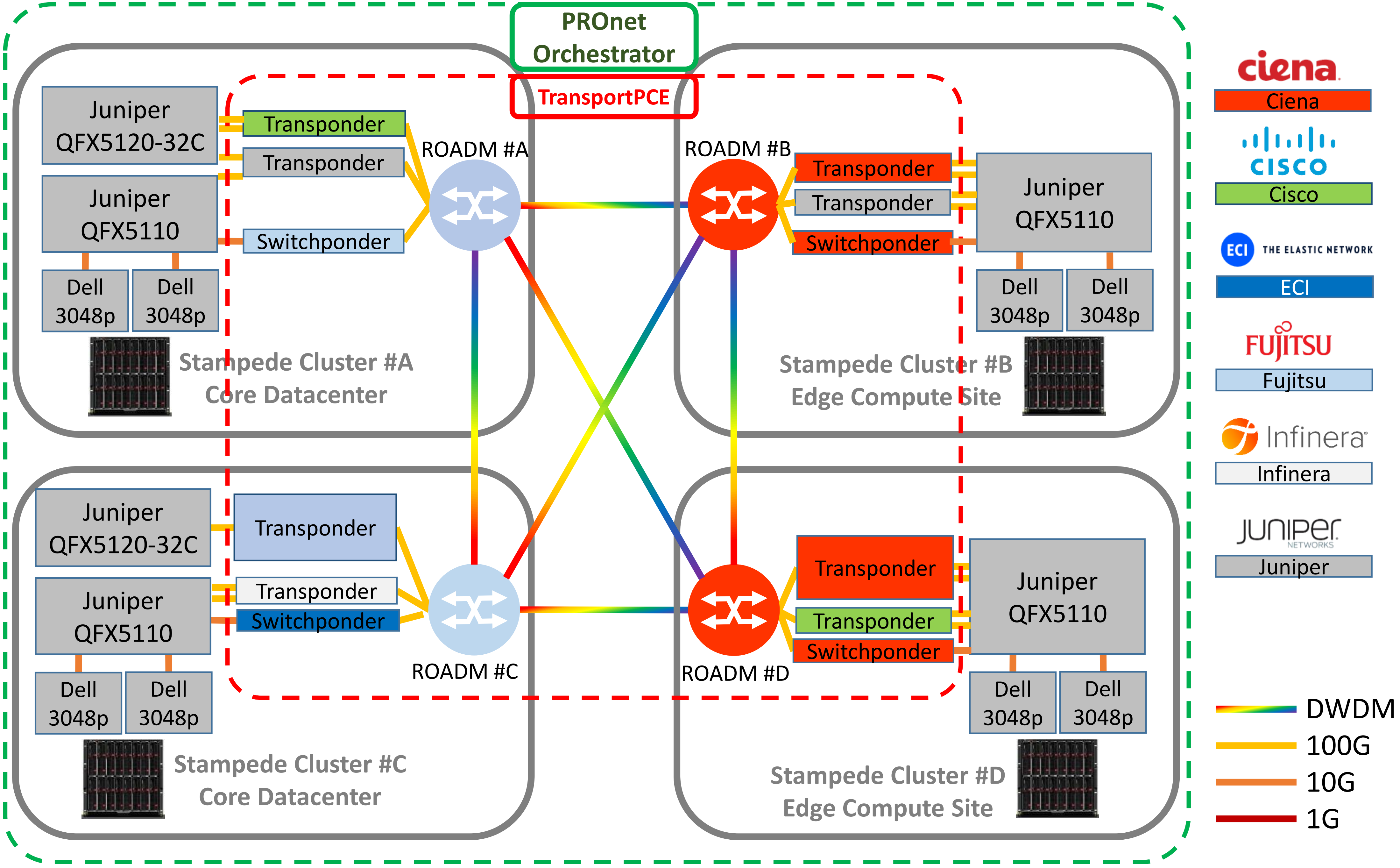
- Operating range characteristics
- Amplifier operational parameters
- ROADM operational parameters
- Optical Service Channel
- Optical power control
- Safety shutdown

UT Dallas researchers added RESTCONF APIs to interface PROnet Orchestrator with OpenROADM Transport PCE module provided by AT&T

http://OpenROADM.org

Live Demonstrations

- **SC 2019**
- **OFC 2020**

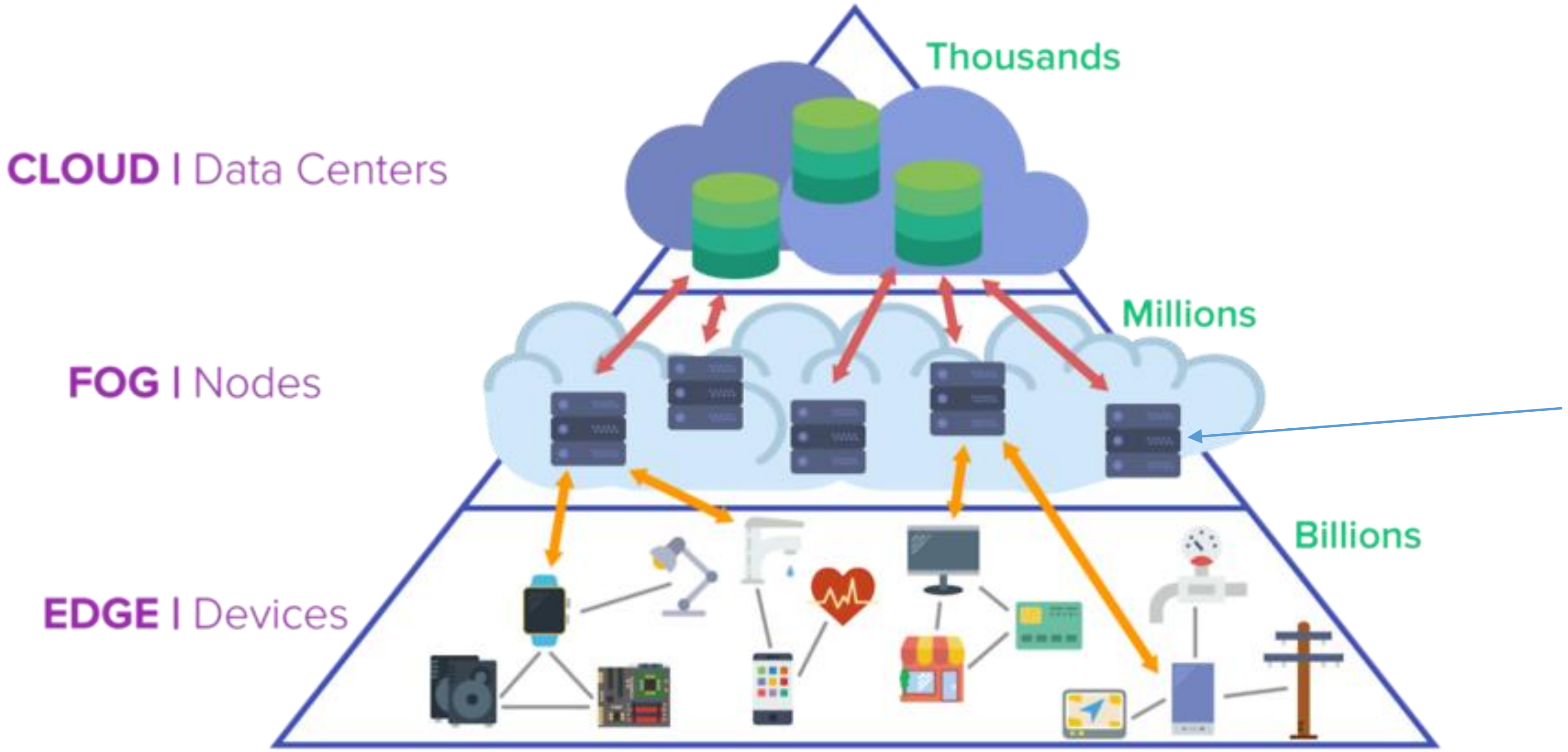


Outline

- Optical Networks: What is Unique?
- PROnet I: Using a Proprietary Solution
- Open Optical Network (OON) Efforts
- PROnet II: Embedding GNPy Modeling in Open Line Systems (OLS)
- PROnet III: OpenROADM with Six Optical Vendors
- Enhancing Distributed Computing
- Summary

Edge Computing

- Efficiently provides compute resources to applications with special QoS requirements
  - Time sensitive applications require tight network latency
- Avoids applications’ data transmission over the WAN to reach a remote cloud site



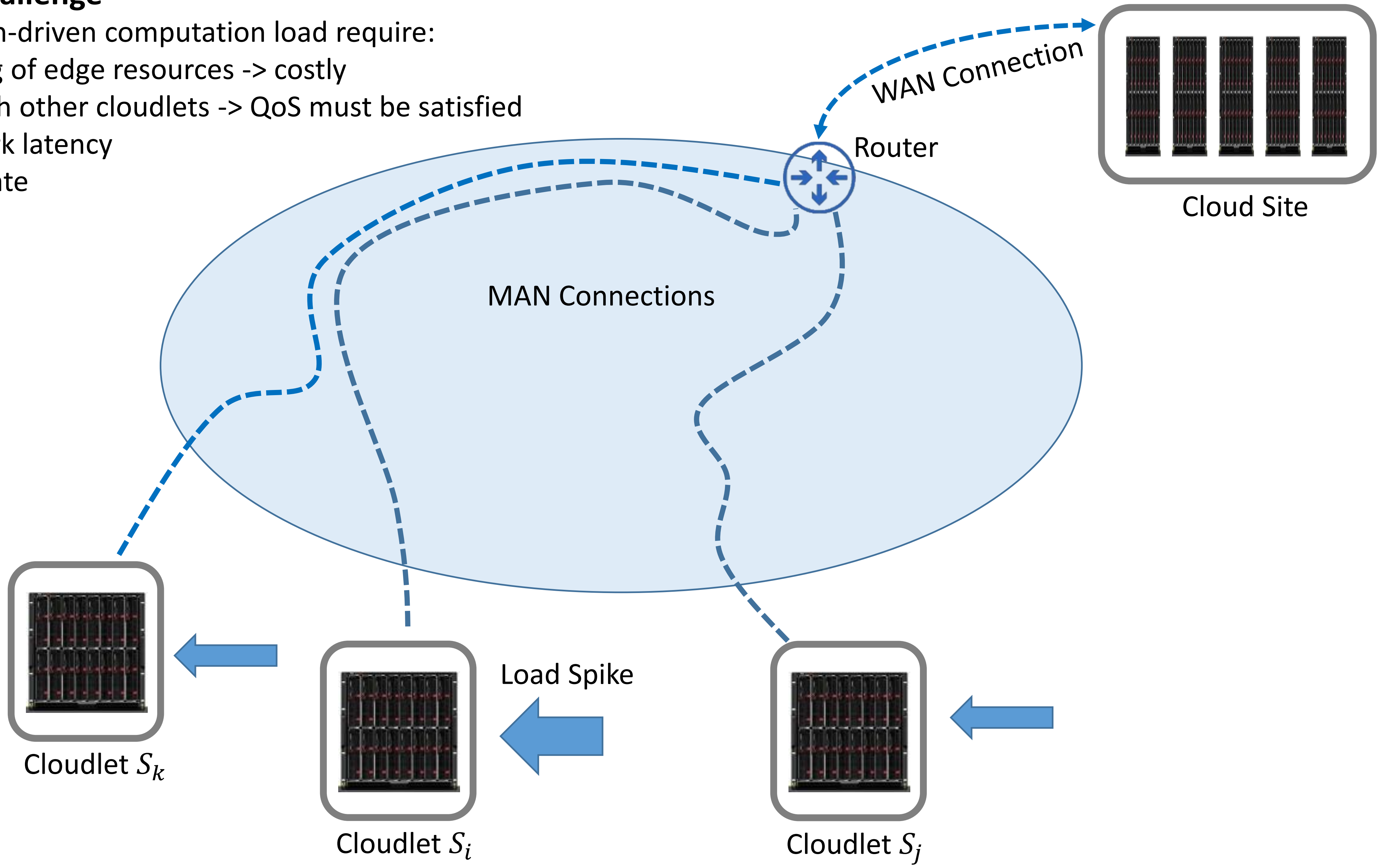
Relatively small cluster of compute resources

Simple Hub and Spokes Architecture

Challenge

Spikes of application-driven computation load require:

- Overprovisioning of edge resources -> costly
- Load sharing with other cloudlets -> QoS must be satisfied
    - Low network latency
    - High data rate

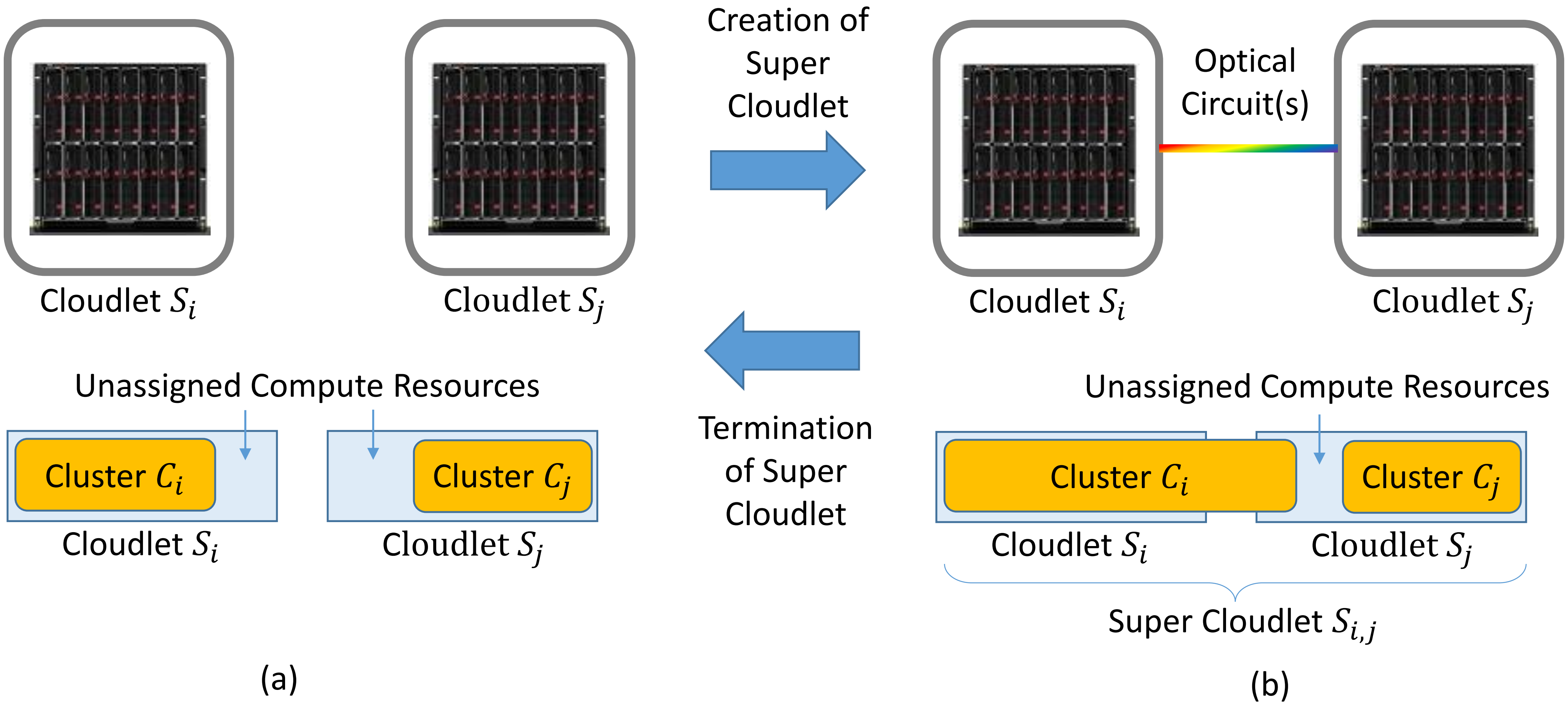


Distributed Super-Cloudlet (Simple Example)

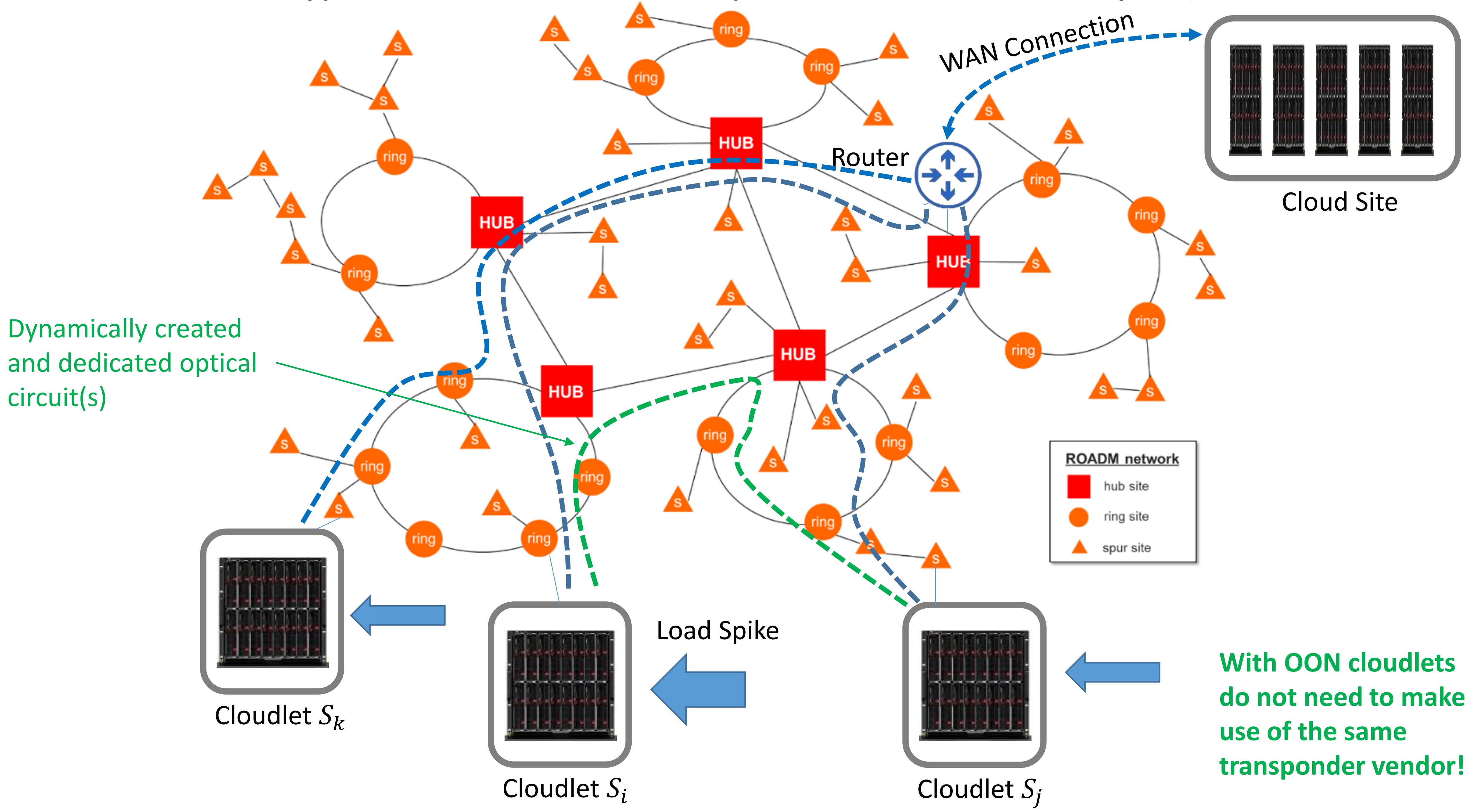
(a) Two cloudlets operating independently, each hosting a single cluster of workers

(b) A super-cloudlet consisting of two cloudlets connected by low-latency and high-data rate optical circuits

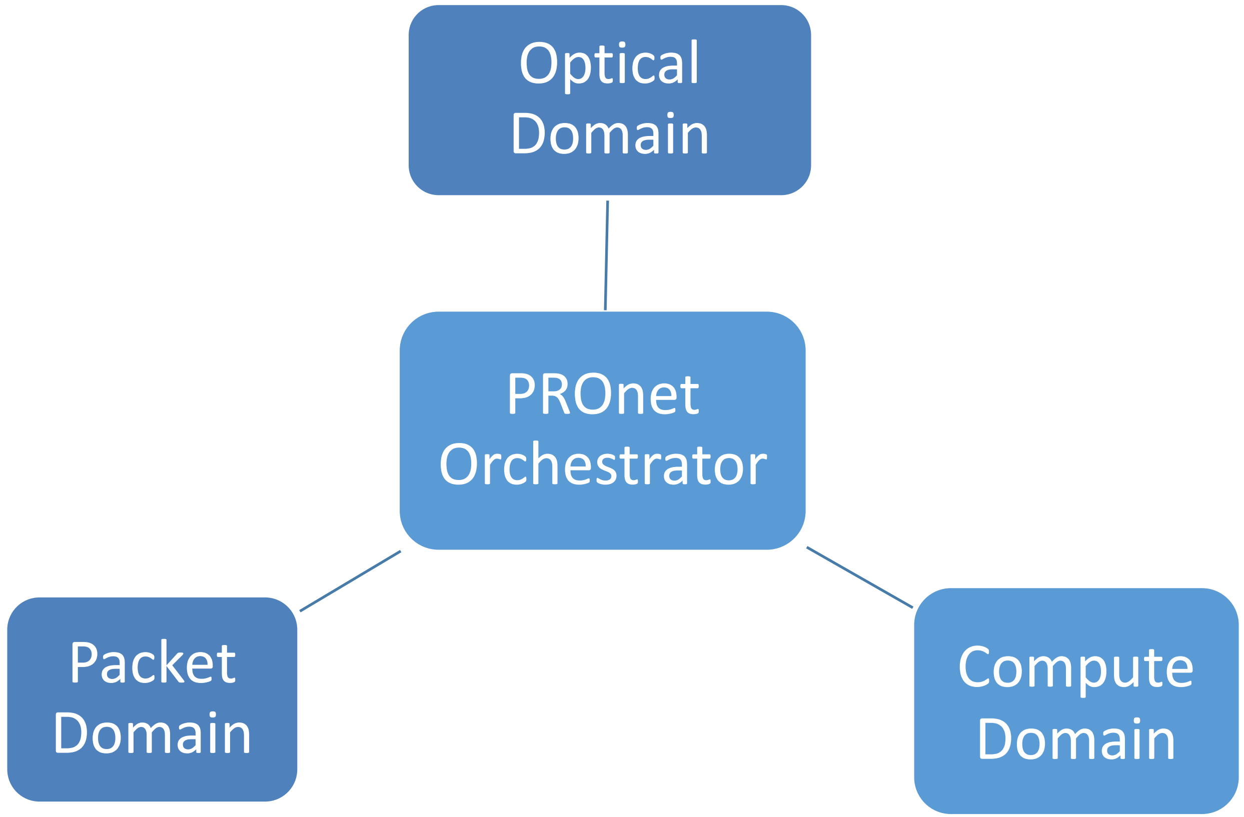
- Worker cluster C_i makes use of compute resources that are provided by cloudlet S_j to cope with a sudden surge in its applications' load



Typical Metro ROADM Transport Network (Source Fujitsu)

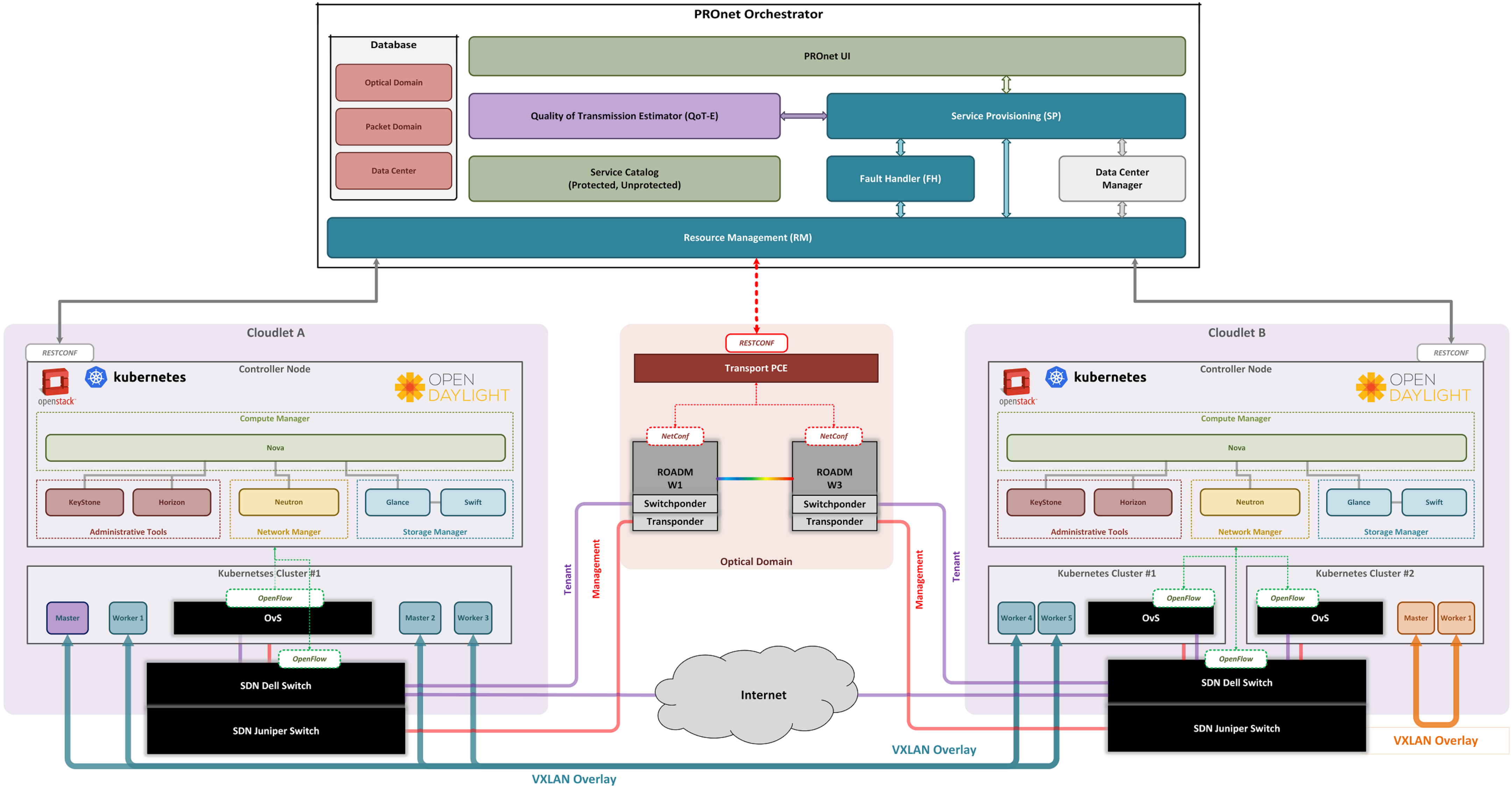


PROnet IV: Orchestrator Updated to Handled Three Domains

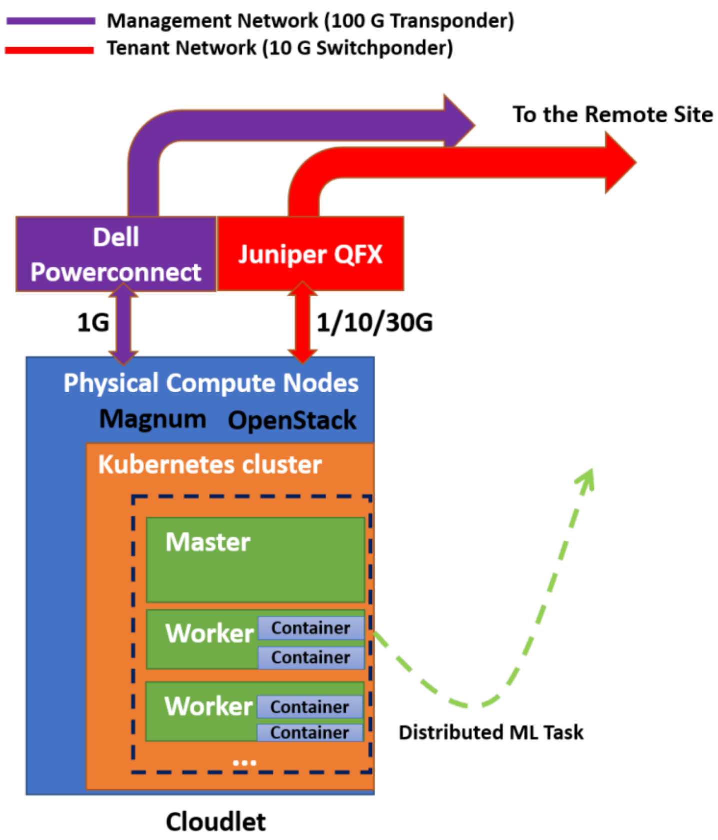


1. Optical Domain
   - TransportPCE
   - Cisco Plugin
2. Ethernet Domain
   - OpenFlow Controllers
3. Compute Domain
   - OpenStack
   - Kubernetes

PROnet Orchestrator Architecture Today



Management and Tenant Networks

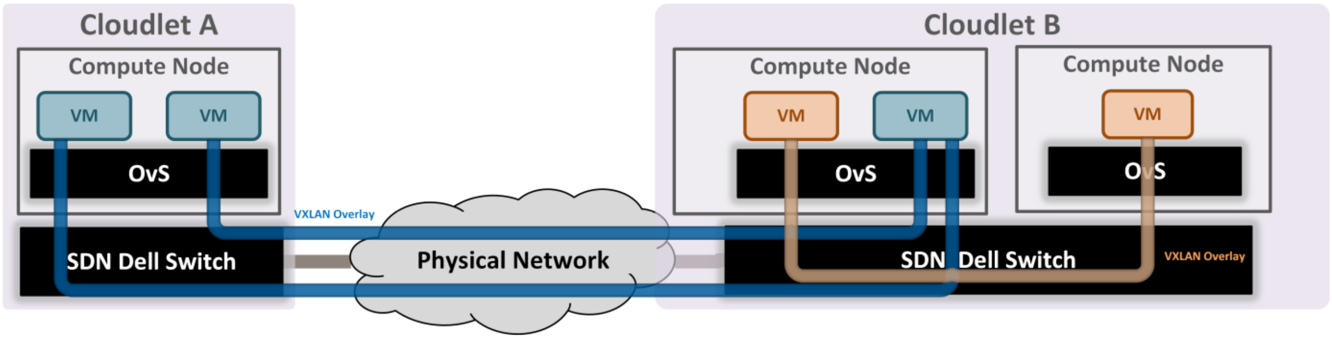


Management Network - Red:

- Migration of VMs (between cloudlets)

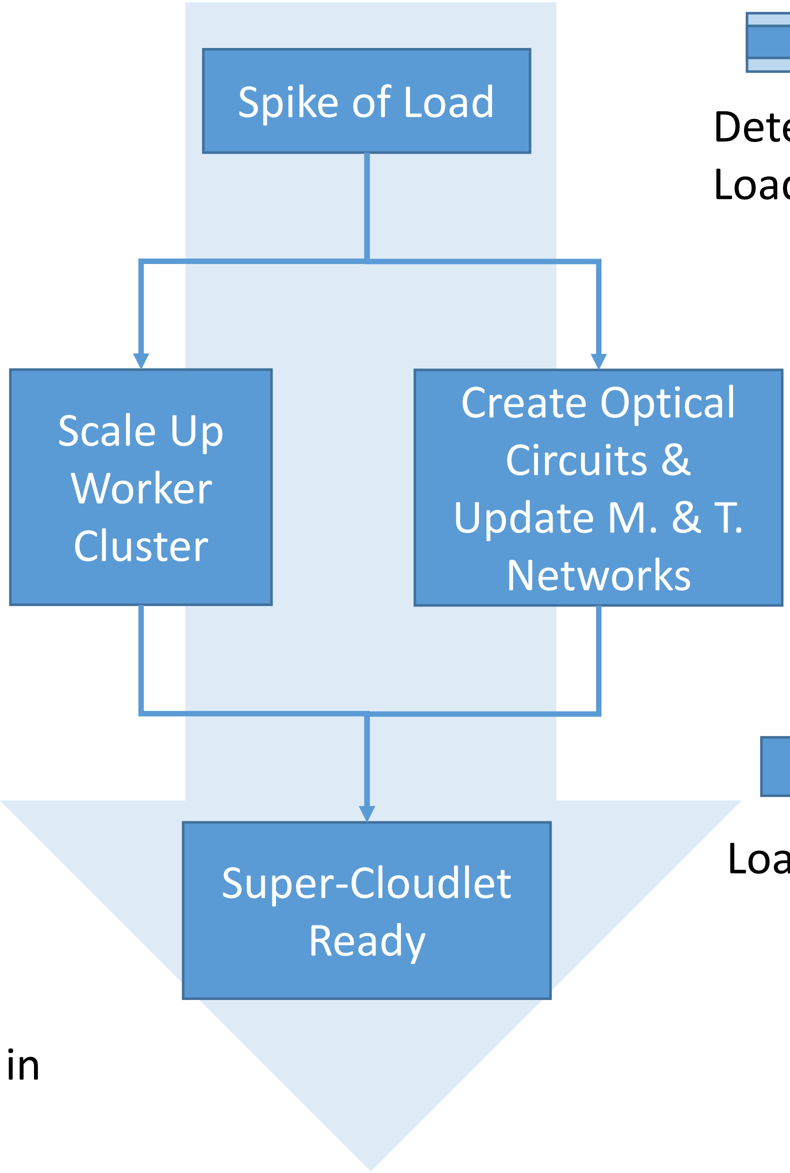
Tenant Network(s) - Purple:

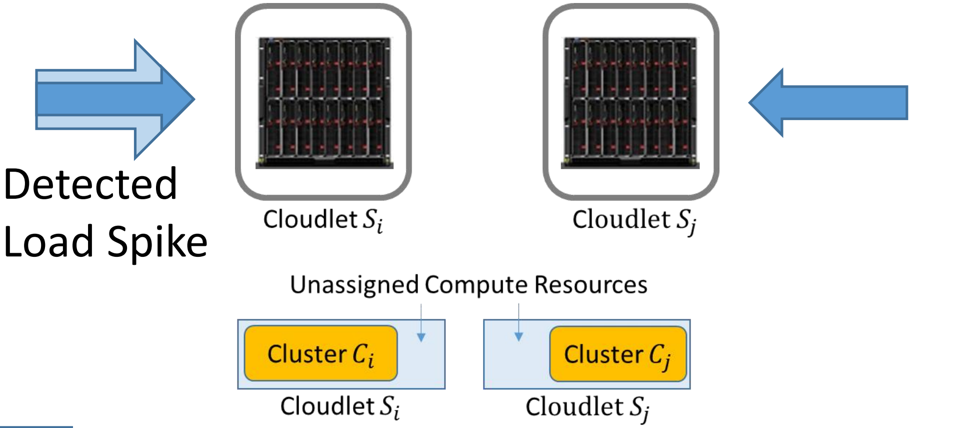
- Supporting application's data traffic (between cloudlets)



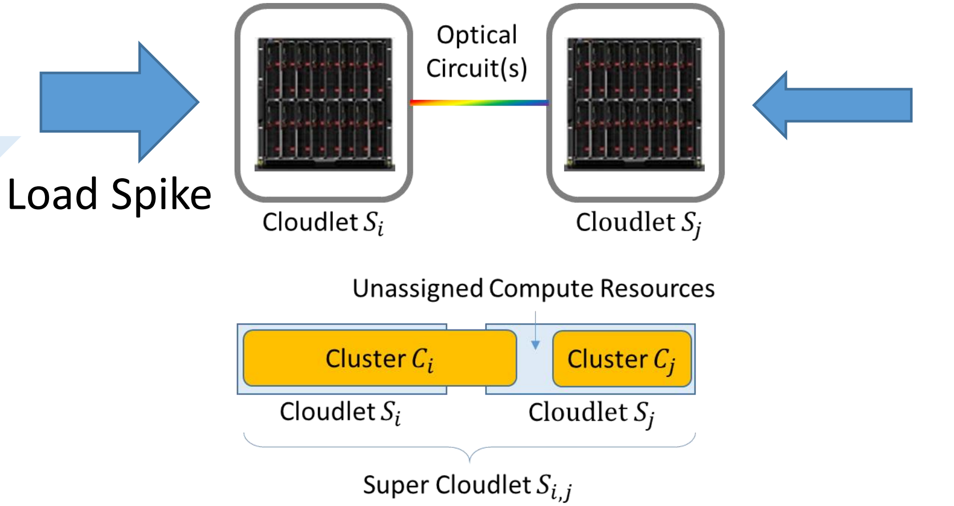
Time Required to Create a Super-Cloudlet

	Mean Scale up	Scale up Variance	Mean Scale Down	Scale Down Variance
1G nodes	420.58s	217.22s	70.90s	13.91s
30G nodes	420.13s	184.13s	68.17s	15.90s





$t_{lightpath\_creation}$	$t_{lightpath\_teardown}$	$t_{L3\_Reachability}$
178s	88s	10s



Note: Time to create a super-cloudlet is comparable to scaling up worker cluster in a single cloudlet

Tenant Application: Distributed Training of ML

Experiment Description:

- Training of RexNet56 (object classification)
- CIFAR10 Dataset (50,000 samples)
- Samples are evenly split over two containers
- Each container splits its own samples to form data batches of size $m = 32, 128$ samples
- During one epoch each container processes one data batch at a time, for a total # of steps:
  - 781 = 50,000/32/2 (when $m = 32$)
  - 195 = 50,000/128/2 (when $m = 128$)
- Data checkpoint: containers must periodically exchange their data with one another at the end of each step

Epoch Completion Time

Batch Size m	# of vCPUs per Container	Single Cloudlet	super-cloudlet Few Meters	super-cloudlet 25km
32	4	1075s	1092s	1095s
32	8	533s	527s	535s
128	4	1058s	1048s	1060s
128	8	501s	502s	504s

Epoch completion time depends on tenant network ability to transfer data quickly between cloudlets

Migration of Containers over Management Network

Experiment Description:

- Training of RexNet56
- CIFAR10 Dataset (50,000 samples)
- Samples are evenly split over three or five containers
- While running the training procedure one of the container is migrated to the other cloudlet
- Migration is performed by using one of these options:
  - Pre-copy with auto-converge
  - Post-copy

For the Management Network Assume:

- 1G connection is available through the "hub and spokes" permanent network
- 30G connection is available through the dynamically created and dedicated optical circuit between cloudlets

Live VM-migration Method	30G node in 3 containers task scenario	30G node in 5 containers task scenario	1G node in 3 containers task scenario	1G node in 5 containers task scenario
Pre-copy with auto-converge	Mean: 75.4s Variance: 3.75s	Mean: 79.37s Variance: 4.2s	Mean: 351.61s Variance: 651.83s	Mean: 362.22s Variance: 583.82s
Post-copy	Mean: 83.49s Variance: 6.86s	Mean: 85.66s Variance: 7.11s	Failed	Failed

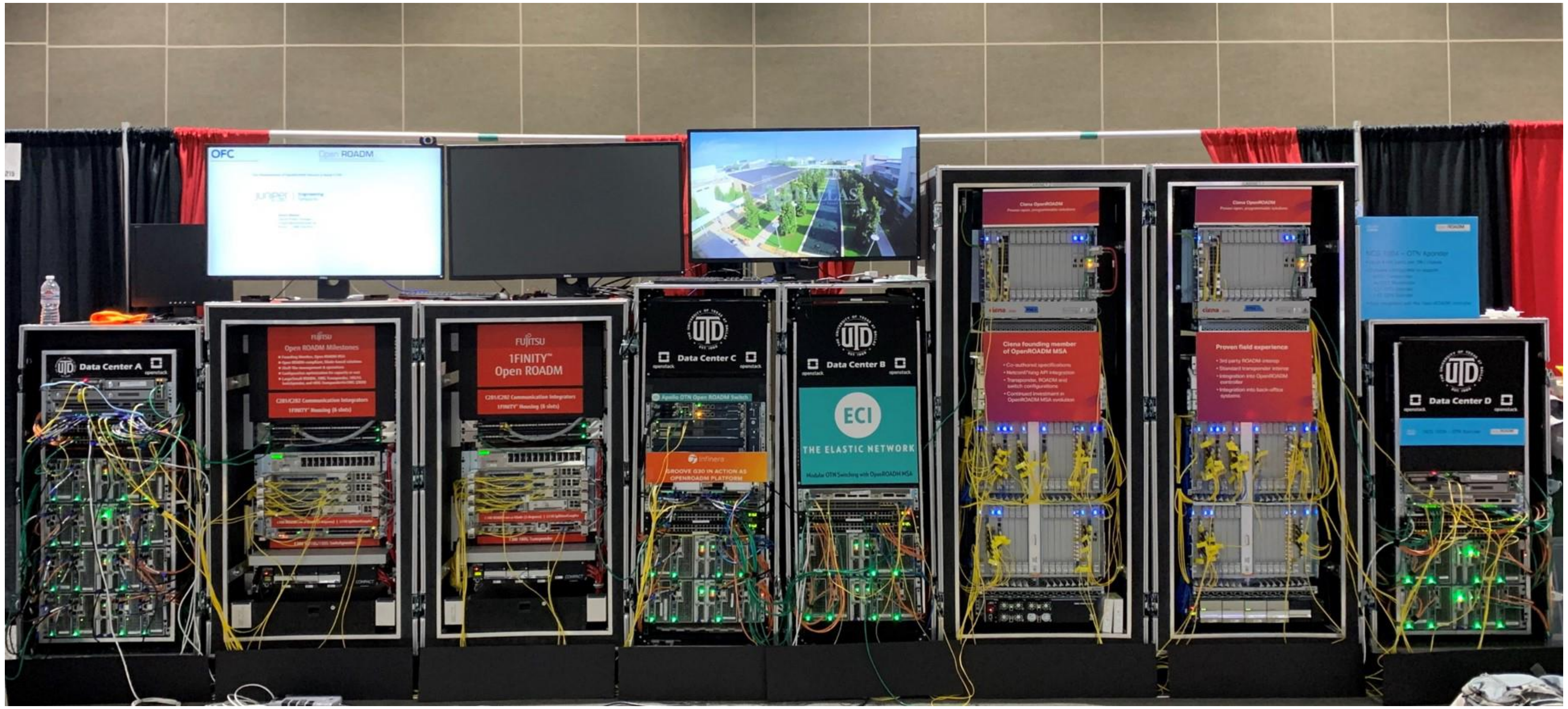
Outline

- Optical Networks: What is Unique?
- PROnet I: Using a Proprietary Solution
- Open Optical Network (OON) Efforts
- PROnet II: Embedding GNPy Modeling in Open Line Systems (OLS)
- PROnet III: Going OpenROADM with Two Optical Vendors
- Next Steps
- Summary

Summary

- Programmable (SDN) optical networks are here to stay!
  - Single vendor and proprietary solutions
  - Open Optical Network (OON) solutions with various degrees of optical component disaggregation
- Edge computing can be enhanced by organizing cloudlets to form *super-cloudlets* thanks to the improved high-data rate and low latency connectivity offered by dedicated optical circuits
- Distributed super-cloudlets and related dedicated optical circuits can be dynamically created in a few minutes to best adapt to changing load conditions originating from the applications with stringent QoS requirements
- Distributed super-cloudlet performance similar to that of a single cloudlet, but the former offers more computation capacity (and dynamically)
- Multi-vendor optical equipment deployment (enabled by OON) offers more flexibility compared to single-vendor one

Live Demonstration of OpenROADM Network at OFC 2020 Booth # 5701



Thank You!

Andrea Fumagalli
andreaf@utdallas.edu

This work is supported in part by NSF grants CNS-1405405, CNS-1409849, ACI-1541461, CNS-1531039, and CNS-1956357

Live Demonstration of OpenROADM Network at OFC 2020 Booth # 5701

Thank You!

Andrea Fumagalli
andreaf@utdallas.edu

This work is supported in part by NSF grants CNS-1405405, CNS-1409849, ACI-1541461, CNS-1531039, and CNS-1956357