Dis-Aggregation as a Vehicle for Hyper-Scalability in Optical Networks

Dan Kilper

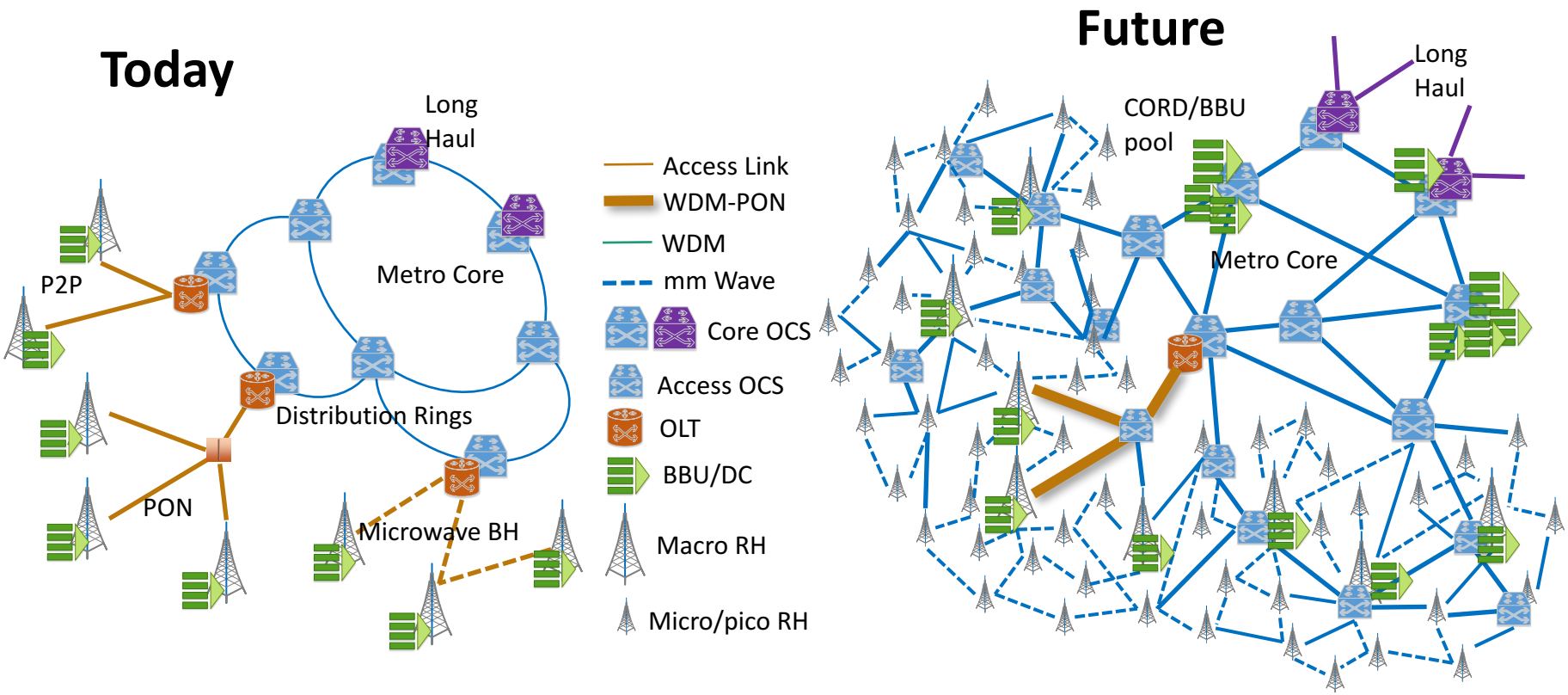
May 14, 2018

Hyper Scale Computing

- Method to scale data centers to ‘warehouse’ sizes
  - 100k’s servers
  - Entire data center becomes the system
    - Hardware/software separation enabled DC-wide control
    - Trade off server performance for cost & DC performance
- Merchant silicon opened door for data center operators to design their own servers
  - Enabled holistic DC architectures
  - Computer ‘integrators’ bounced back by designing whole rack and pod solutions

ARIZONA CIAN

Densification of Wireless Access



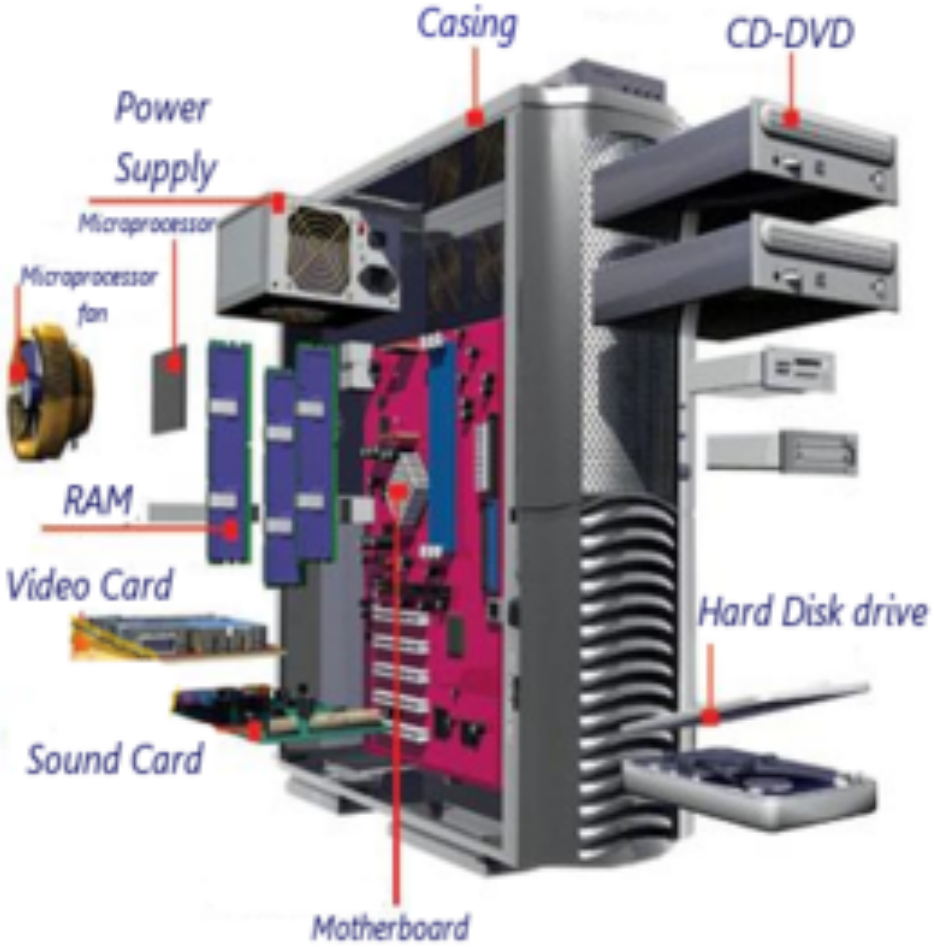
- Network operators requesting 10k's of access points in each US city
- Each access point > 10 Gb/s backhaul/fronthaul
- Operators offering whole wavelength access (e.g. Pilot)

ARIZONA CIAN

What is Dis-Aggregation?

- Dis-aggregation is economic concept
  - Different vendors provide parts that make up a system
  - Whether to disaggregate is usually driven by market and supply chain considerations
- Dis-aggregation is an architecture concept
  - Physical or control integration is separated
  - Often determined by performance requirements

Market Driven Computer Dis-Aggregation Enabled Hyperscale DC Architecture



ARIZONA CIAN

Two Main Drivers for Dis-Aggregation

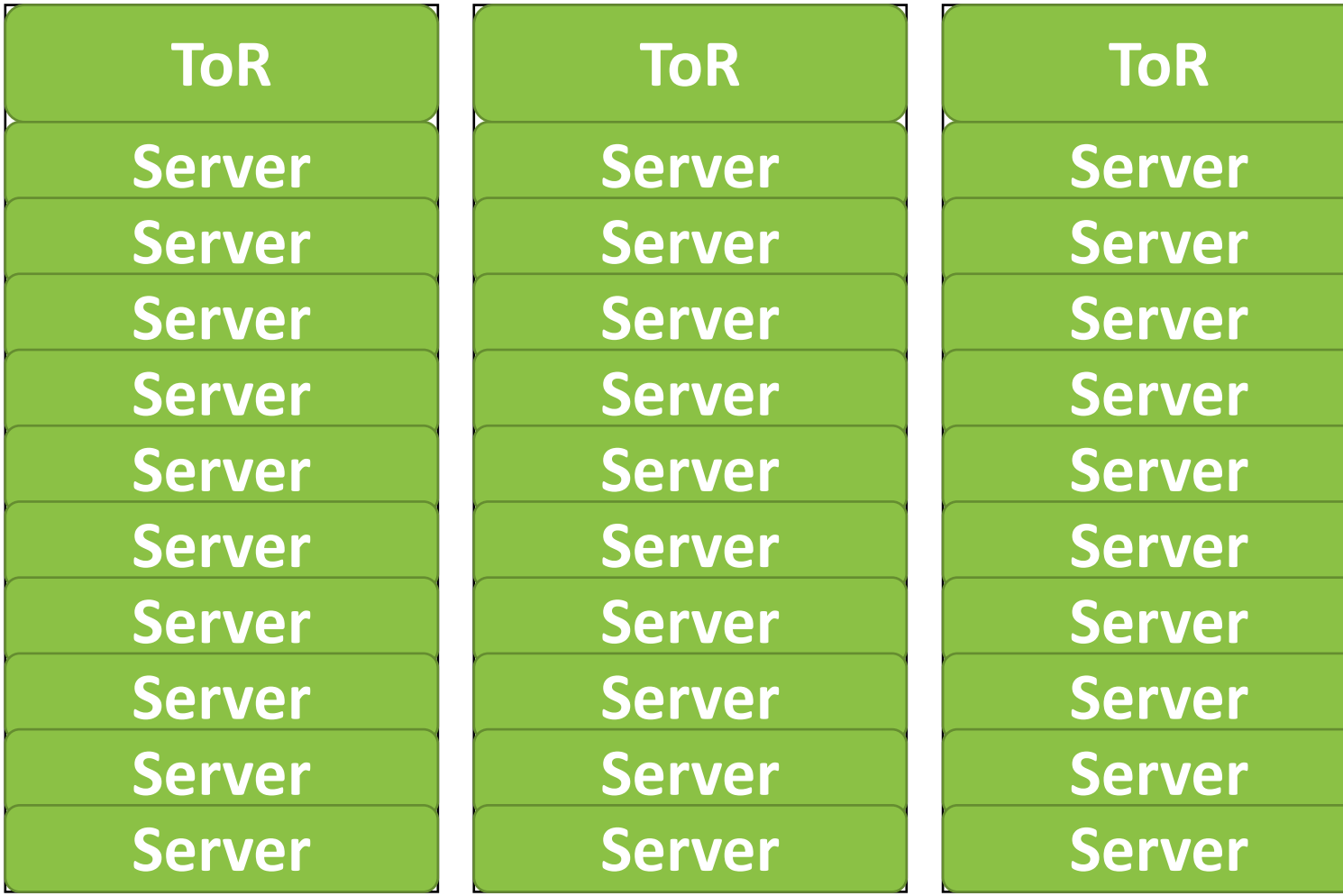
- Market
  - When performance is less important
  - When scalability is needed
  - Use market competition to drive down cost

- Performance
  - When component performance is more important than system performance
  - When technologies reach new performance levels enabling disaggregation
  - Use architecture enhancements to drive down cost

ARIZONA CIAN

Conventional Data Center

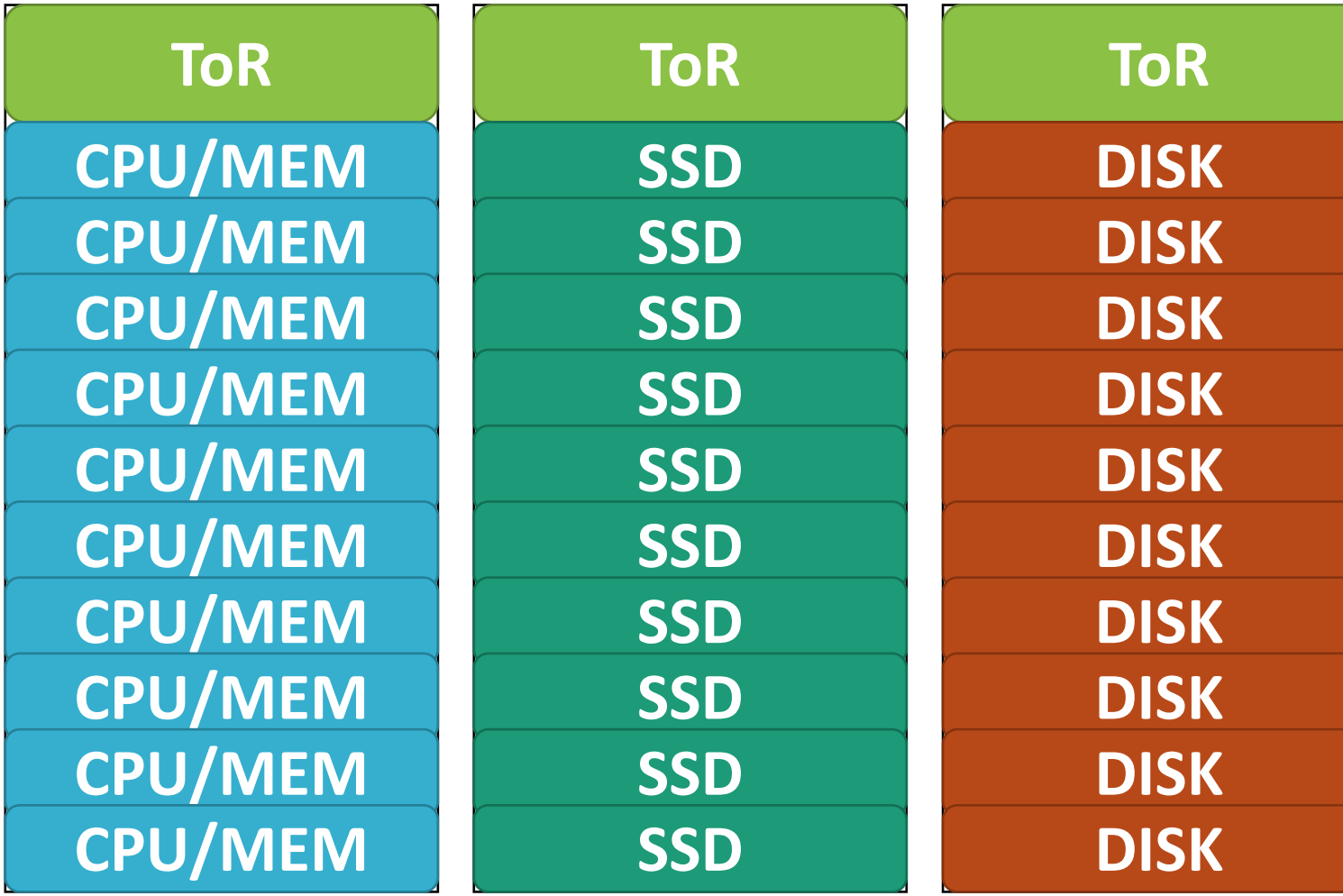
Pack Servers into Racks



ARIZONA CIAN

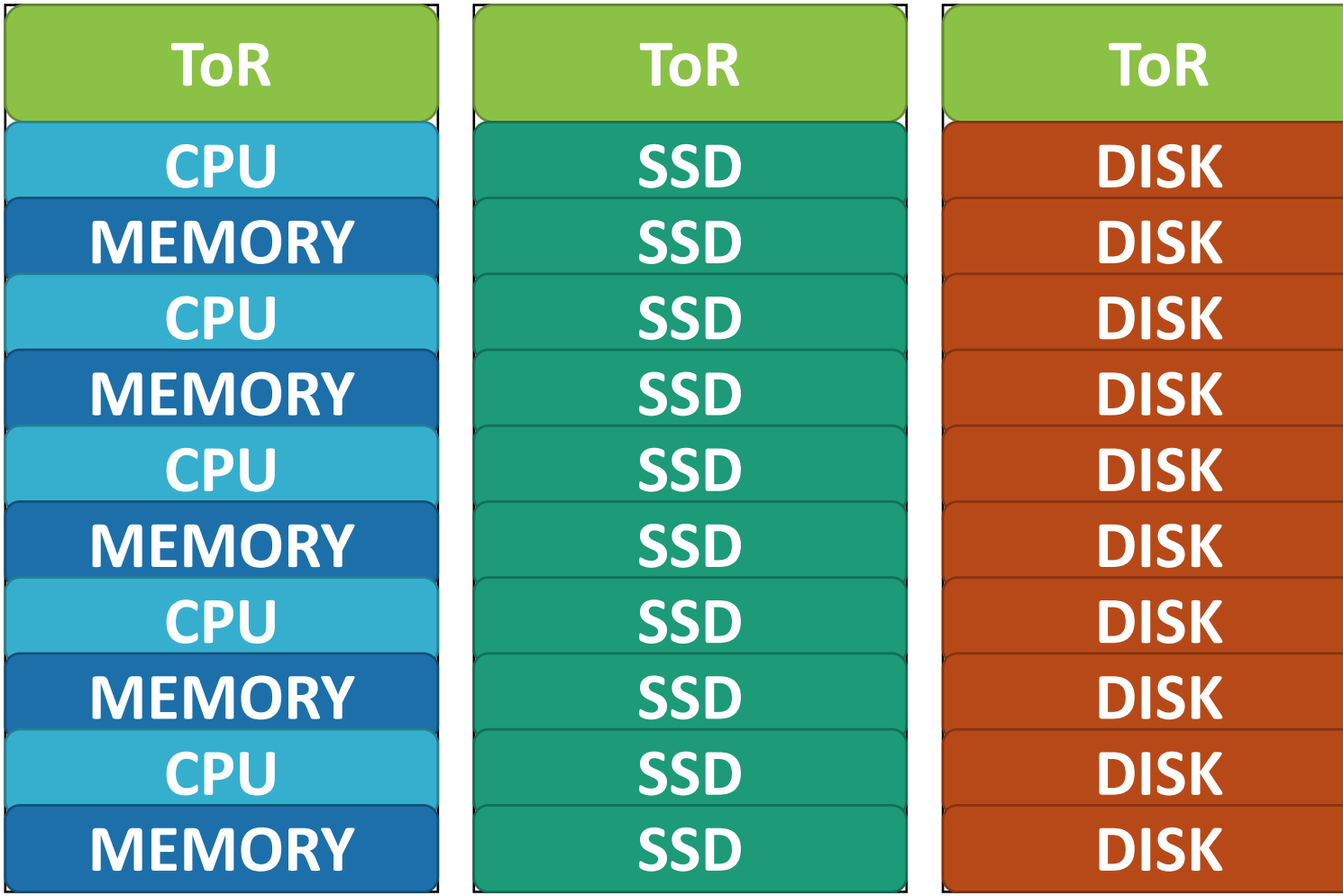
Dis-aggregated Data Center

Resource per Shelf



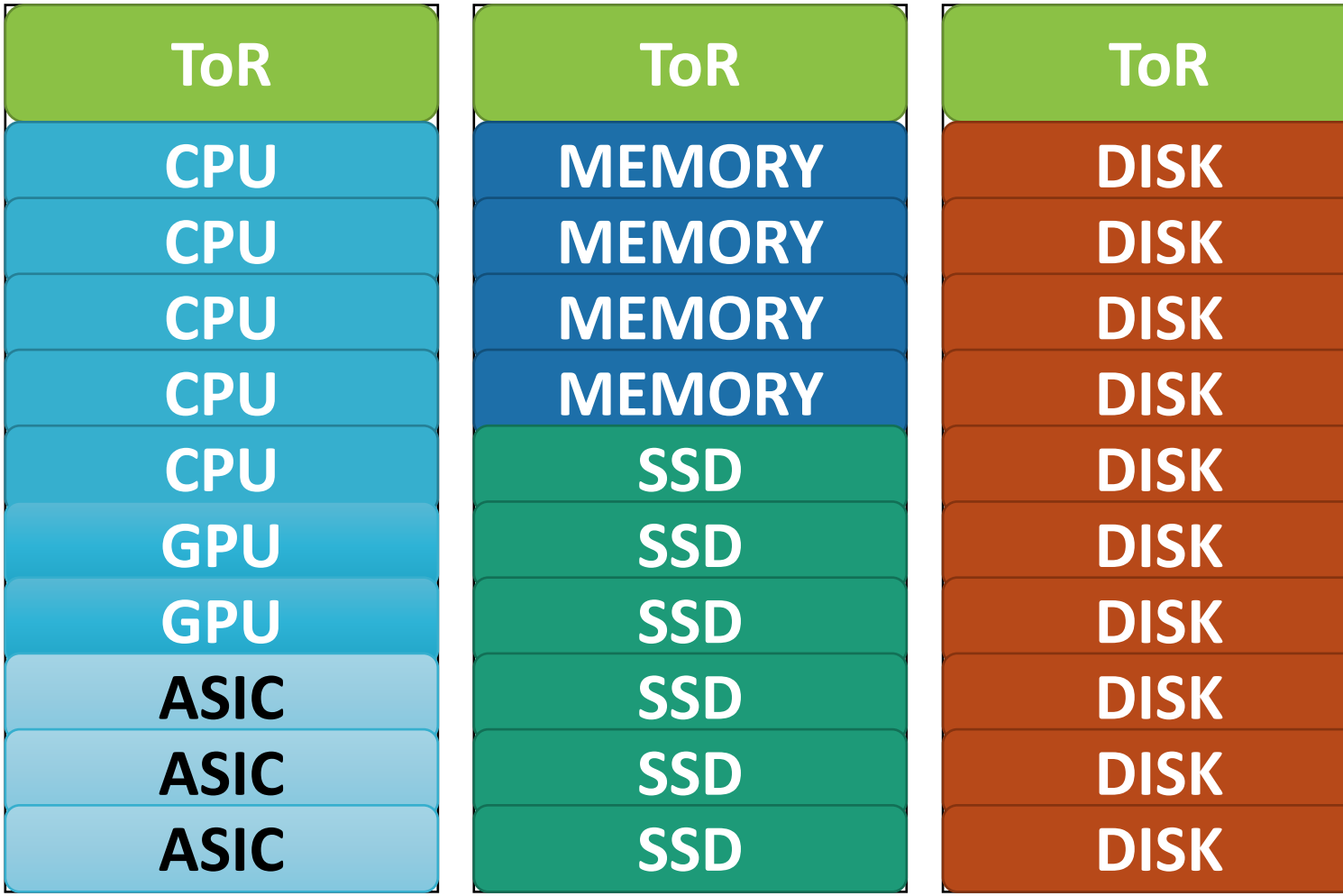
ARIZONA CIAN

Dis-aggregated Data Center



ARIZONA CIAN

Dis-aggregated Data Center



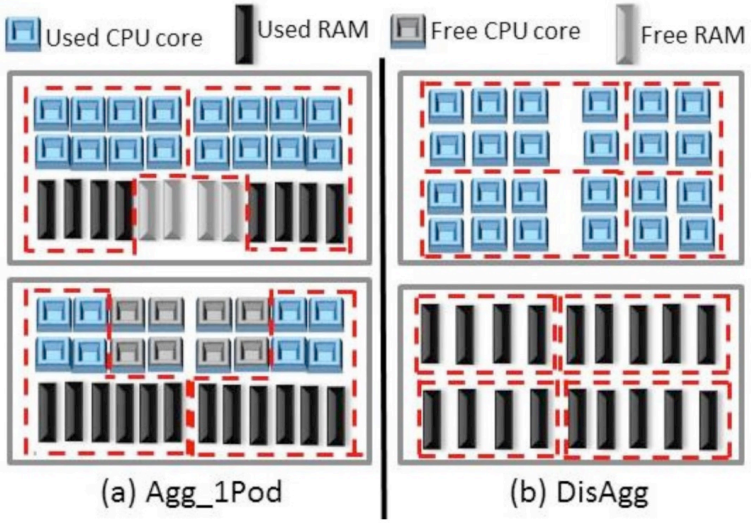
ARIZONA CIAN

Why Dis-Aggregate Again?

- If you have optics to the components then increase interconnect distances to ~100m
  - Latency requirement becomes the limitation
- Is server optimum combination of cpu/memory/disk/ storage/NIC?
  - Can virtualization be more efficient if remove artificial boundaries created by server architecture?
  - Server memory locked to CPUs
- Does server allow for best network architecture?
- Optimize thermal management to device requirements
  - At shelf and rack level

ARIZONA CIAN

Architecture Dis-Aggregation Benefits



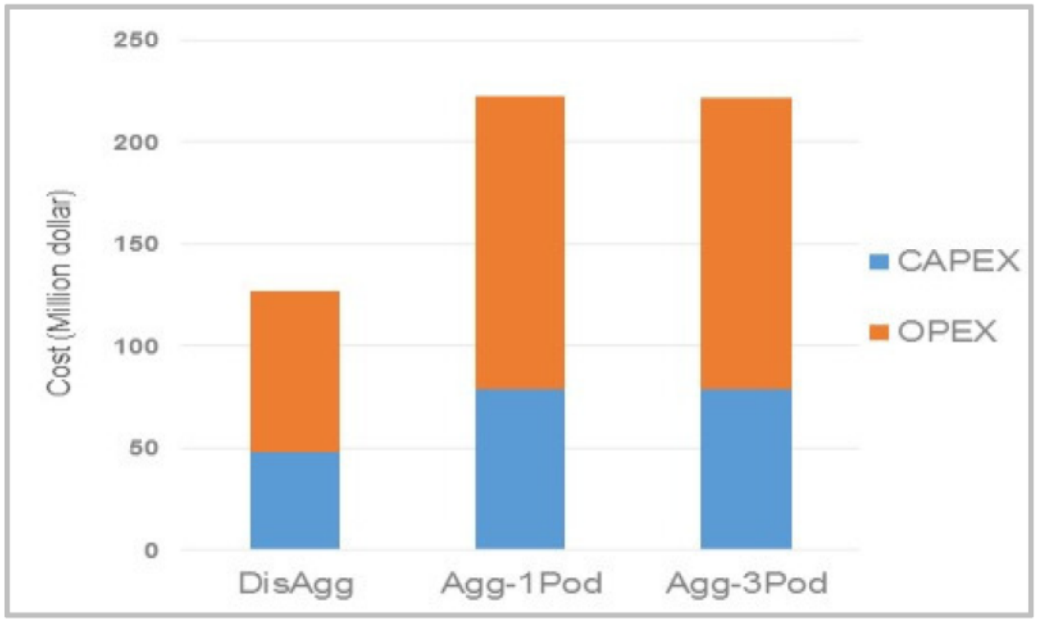


Fig. 4. Total cost for 10 years

Bringing Optics Inside the Computer

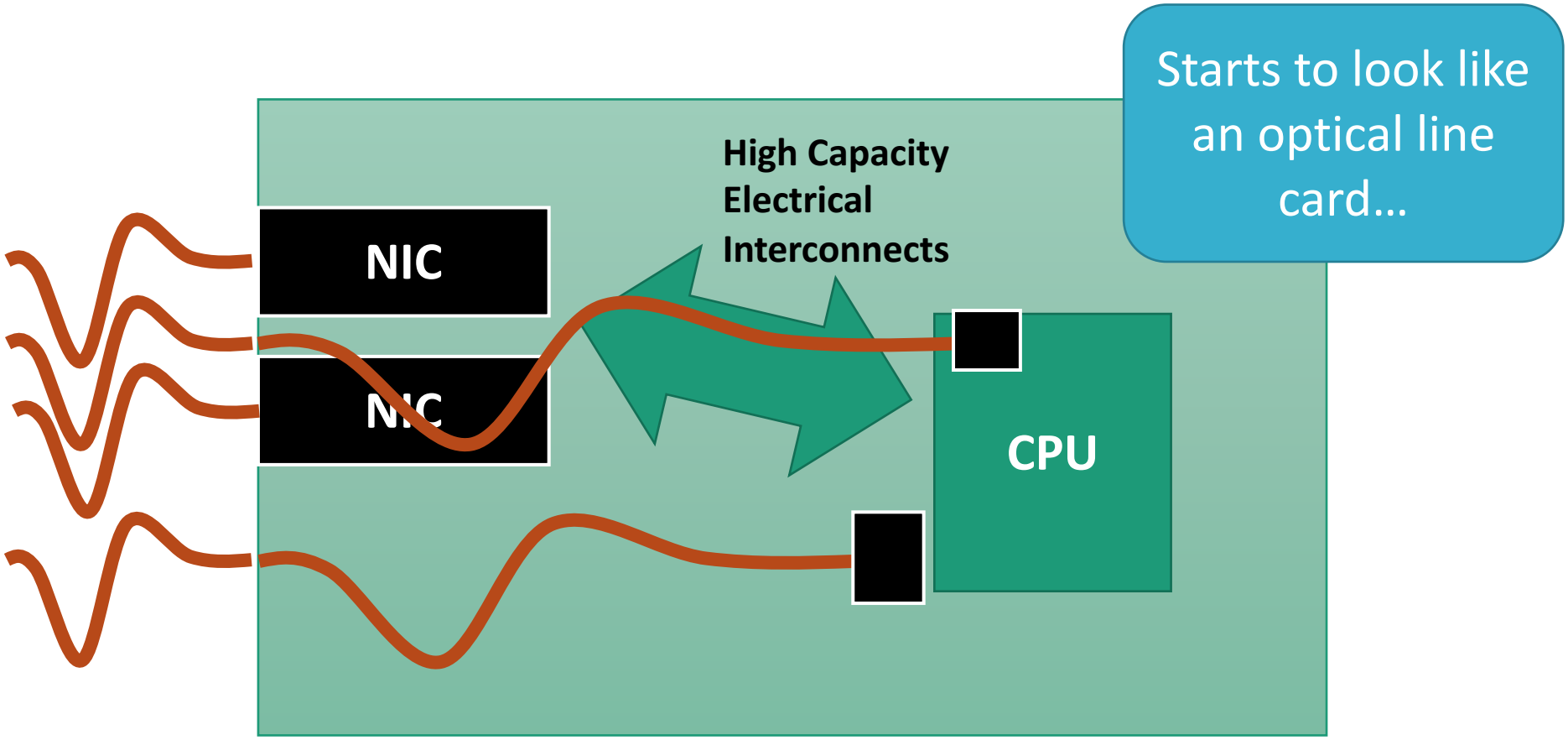
- CPU IO Bottleneck:
  - Need optics for CPU to memory interconnects
    - Its going to be there no matter what
  - What are the prospects for scaling this to 10-100m?
  - D.A.B. Miller Proc. IEEE 2009

- Embedded optics: moving the NIC onto the board
  - Expanding the NIC and integrating it on board

- Data Center Optical Networks
  - If you have a network, why not dis-aggregate?

ARIZONA CIAN

Embedded Optics



ARIZONA CIAN

Dis-Aggregating Optical Systems

Some History

- Late 90's: MCI/Globecom tried to build their own systems from components
- ~2000: Unified control plane attempt to merge control of optical systems into L3 control
  - GMPLS/MPLS was result
- Mid 00's: JDSU/Nortel introduce 'generic' ROADM building block systems
- Late 00's: Coherent transceivers change system engineering (no dispersion maps, PMD)
- Early 10's: Enterprises/DC operators build their own optical networks
- 2020: 5G is coming!

FAIL FAIL FAIL

Performance Change

Market Changes

ARIZONA CIAN

Optical System Vendors

- Historically optical system vendors NOT ‘system integrators’
  - Optical systems are engineered products
    - Components and sub-systems highly specific to system design
    - Tightly coupled hardware and software design
    - Long R&D and test cycles to develop product

- Key question: Can optical system vendors move to system integrator model?
  - Similar to Dell or HP
  - Or operating system model? e.g. Microsoft

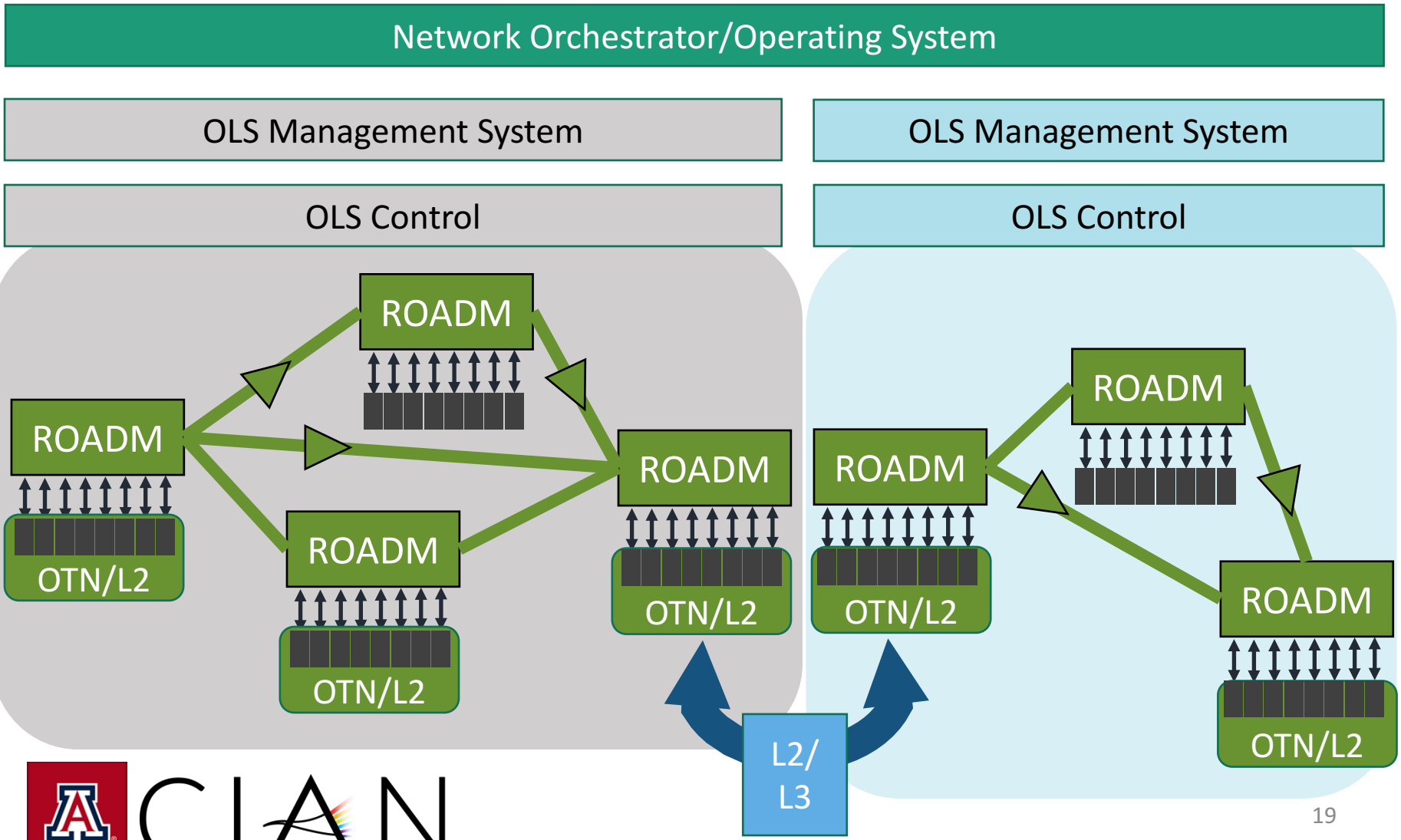
ARIZONA CIAN

Hyperscale Attributes

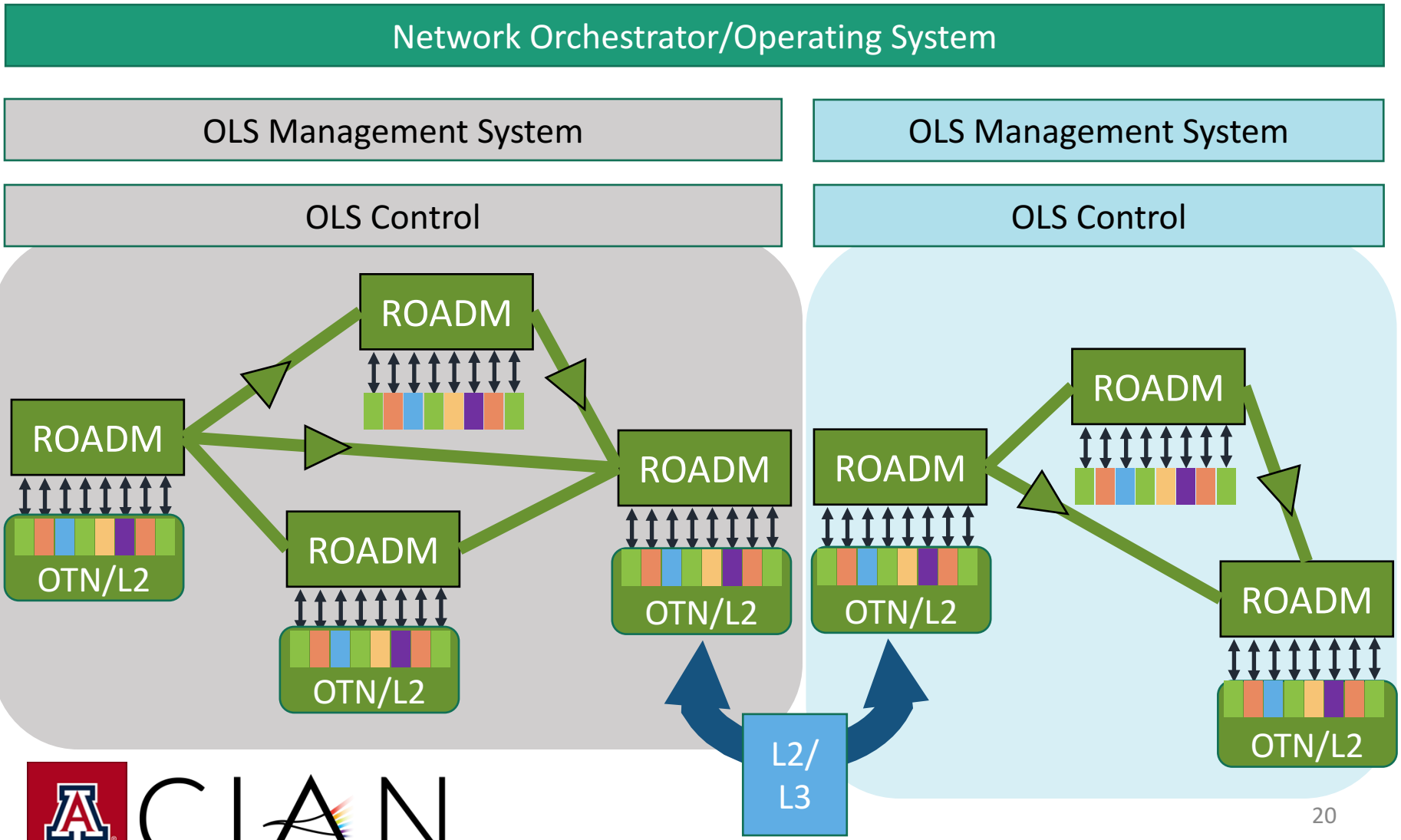
- Large numbers of access points (ROADM nodes)
  - Go from 100's per city to 10k-100k per city
  - Designed at the network level to achieve scalability

- Unified and scalable software control
  - Remove 'siloing' – hardware tied to software (operating system)

Proprietary Optical Systems

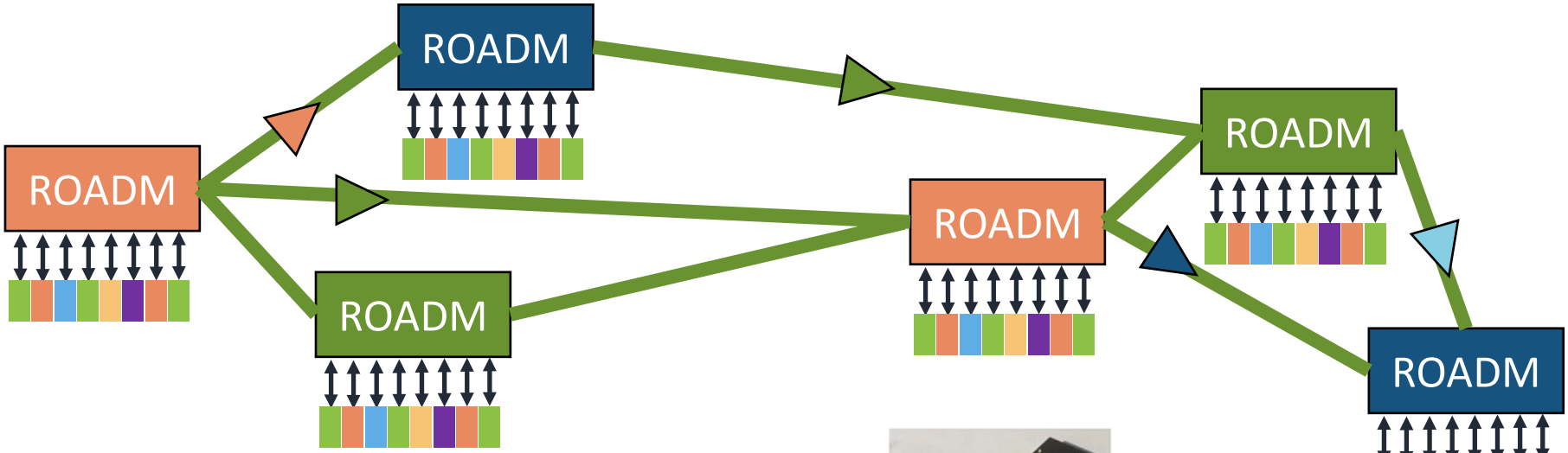


Transceiver Disaggregation (Alien λs)



Whitebox/openROADM Systems

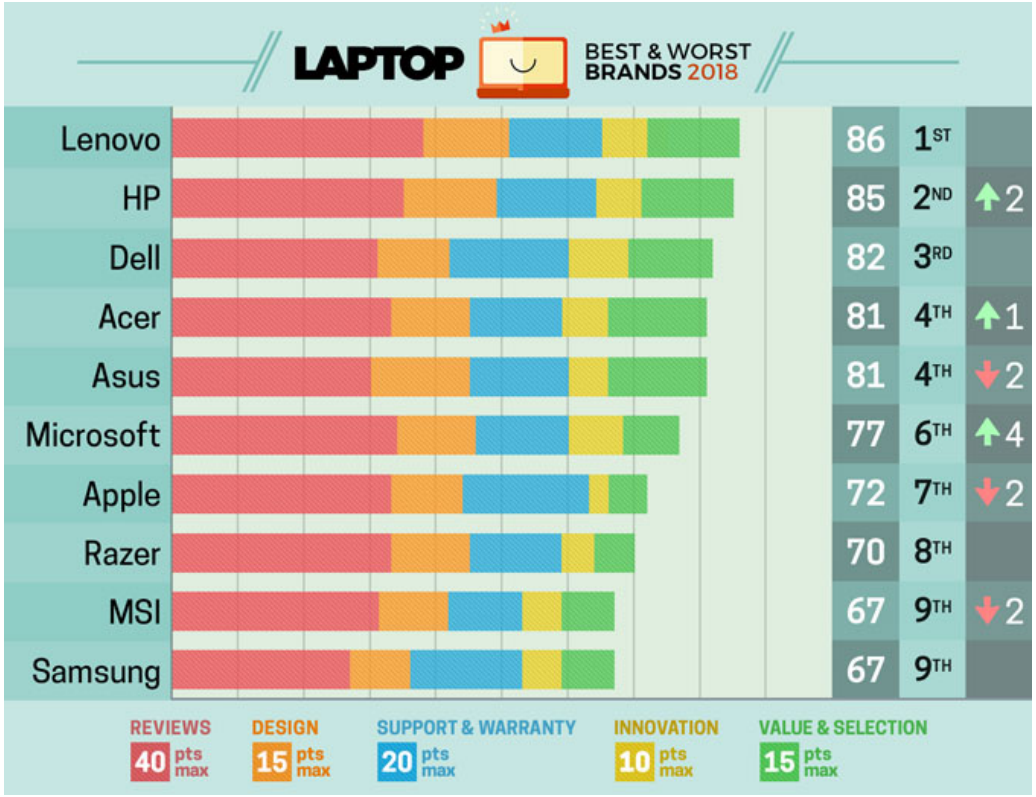
Network Orchestrator/Operating System



Menara:
Built in OTN

Computer System Integration

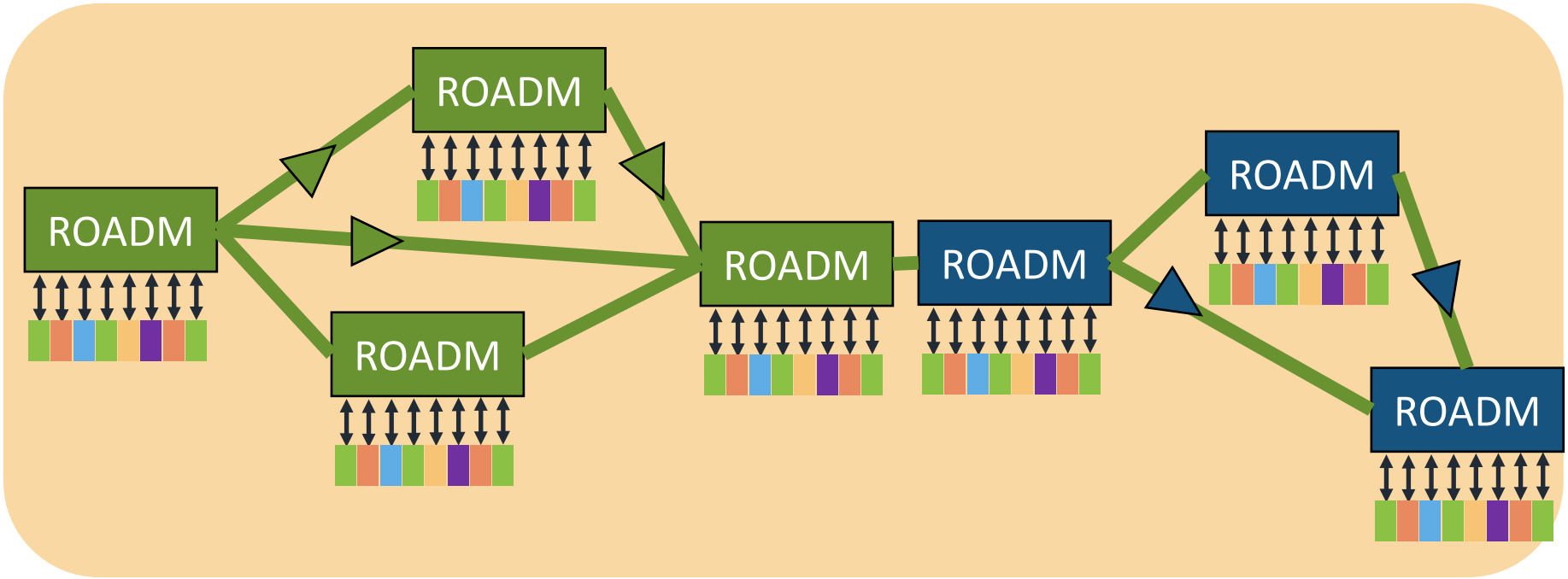
- Still value in matching components to motherboard and good system design principles



Whitebox/Open Optical Networks

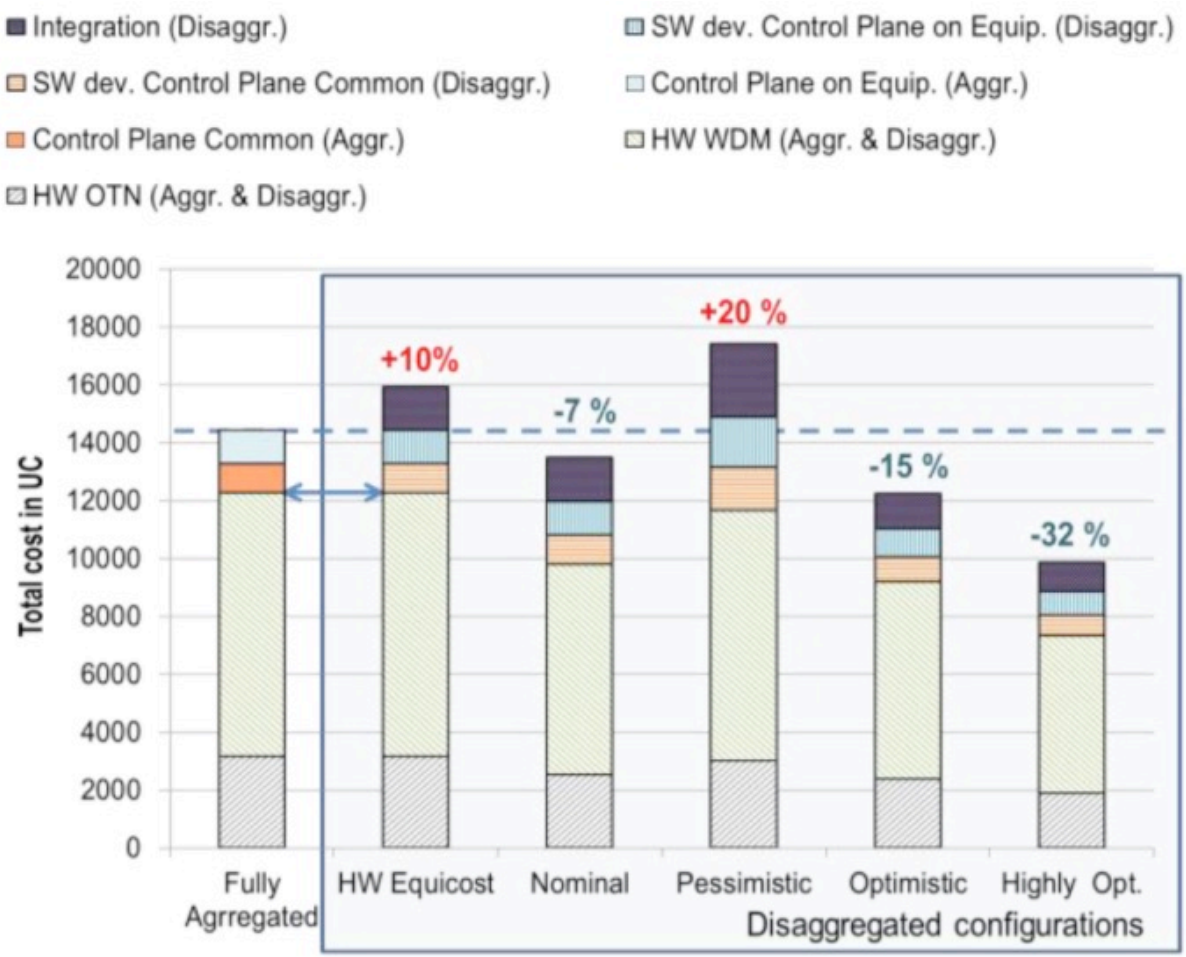
Network Orchestrator/Operating System

OLS Control & Management System



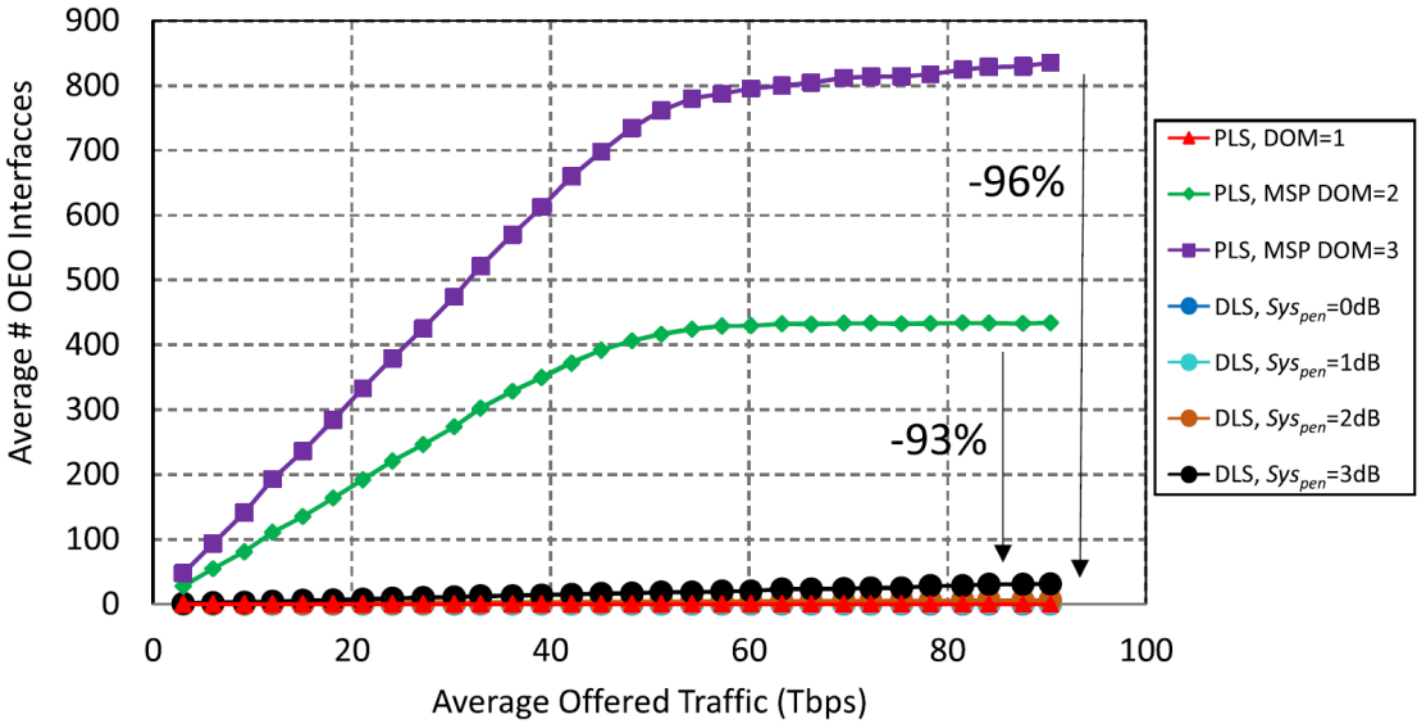
ARIZONA CIAN

Cost Models: Where's the Savings?



Riccardi, et. al. JLT 2018

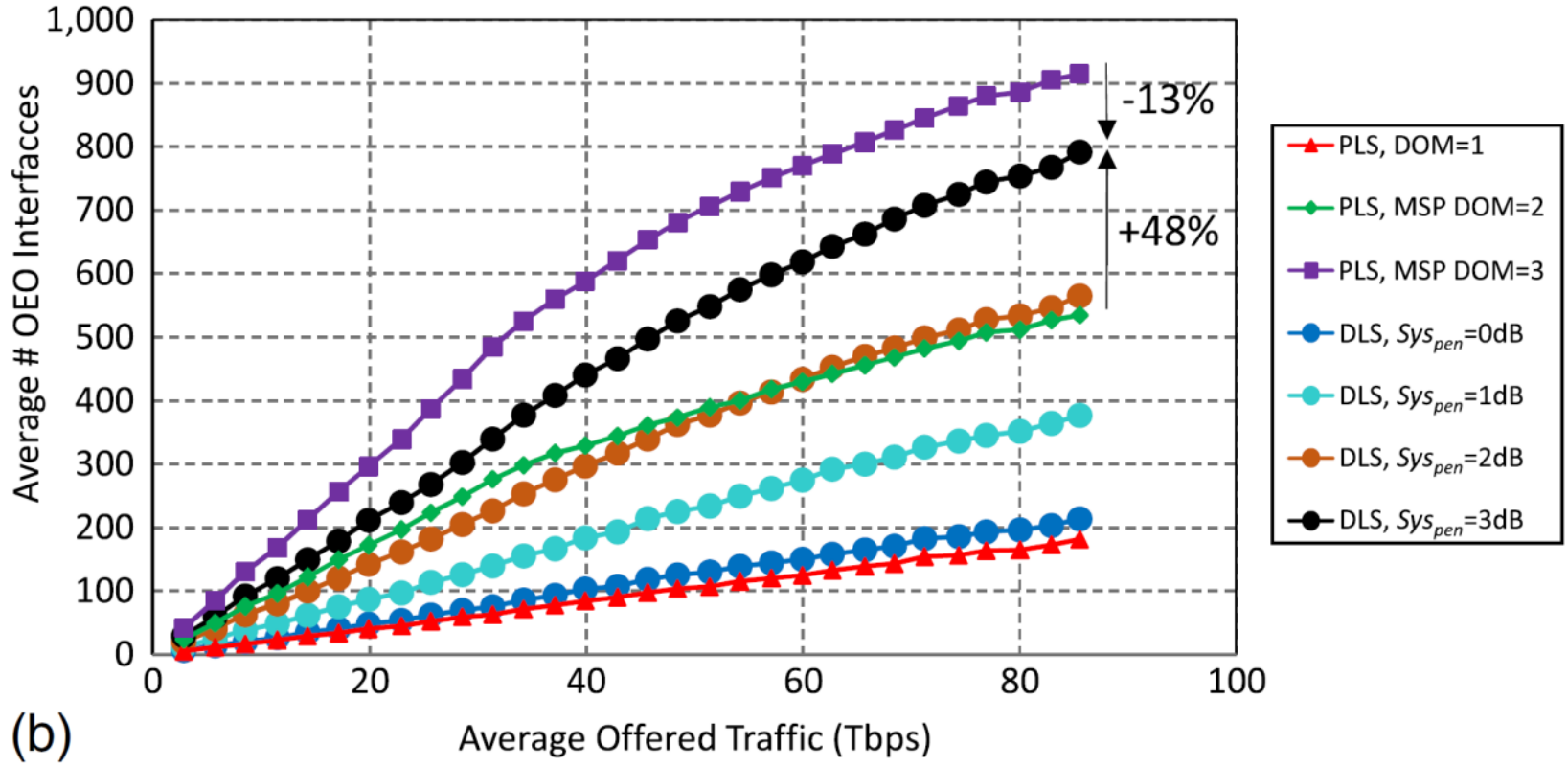
Transceiver Savings: Avoid Regens



- With almost no regeneration

J. Santos et. al. JOCN 2018

With Regeneration

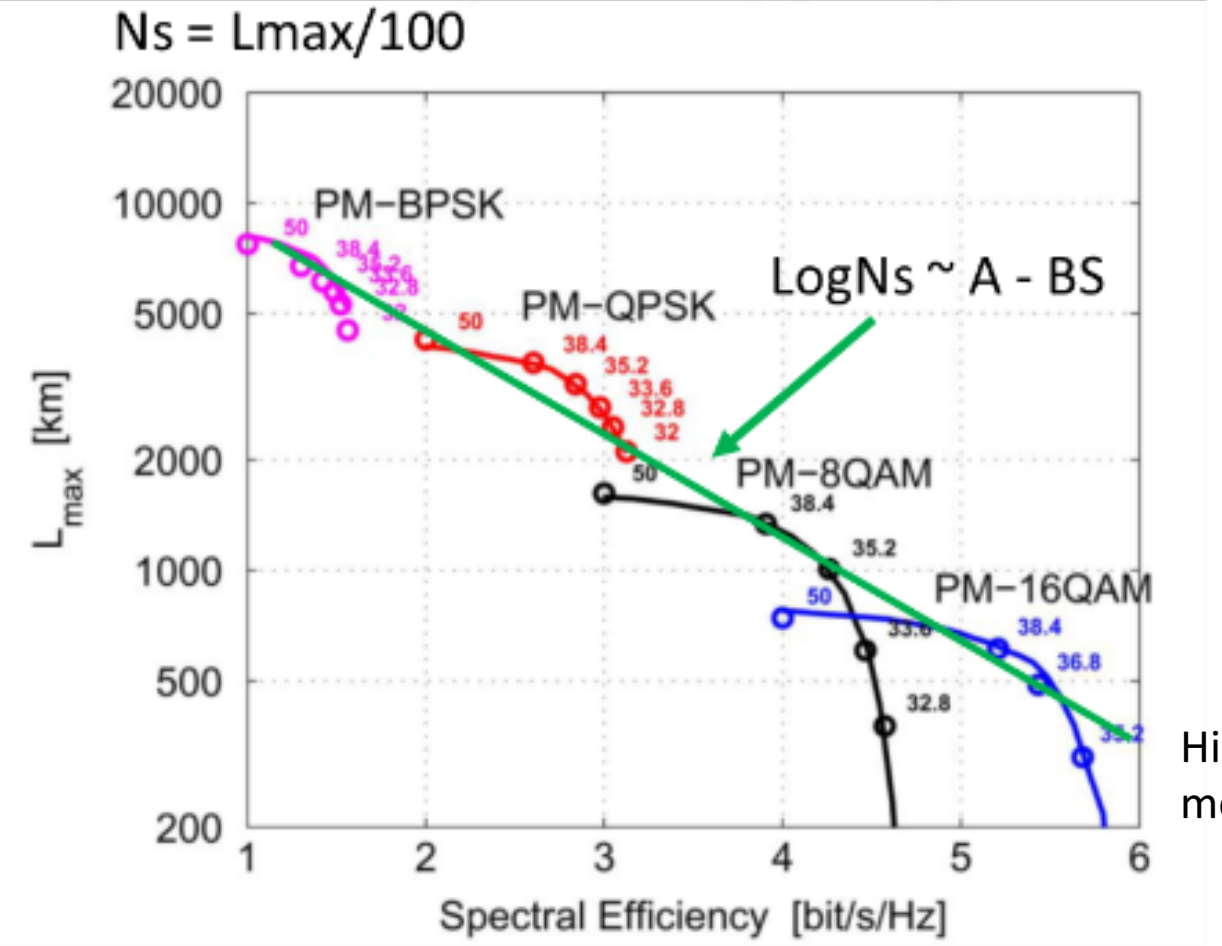


- Disaggregation penalty & network domains make a difference

ARIZONA CIAN

J. Santos et. al. JOCN 2018

Transmission Reach



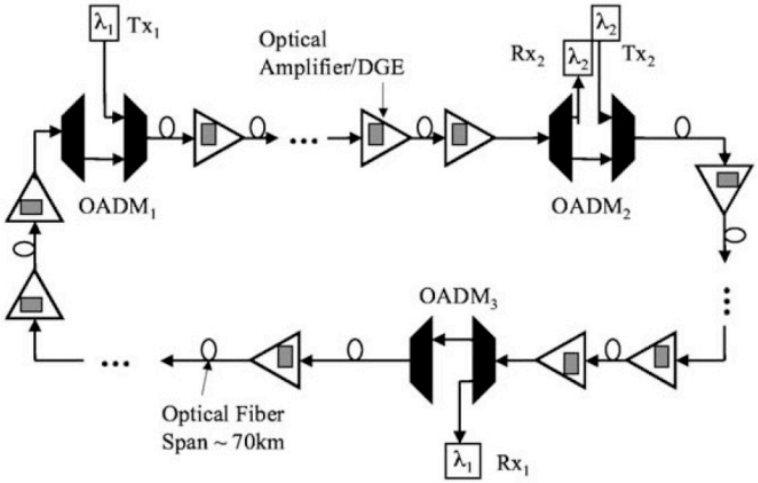
In Metro & data center networks:
Distance = # Hops
2000 km ~ 20 hops

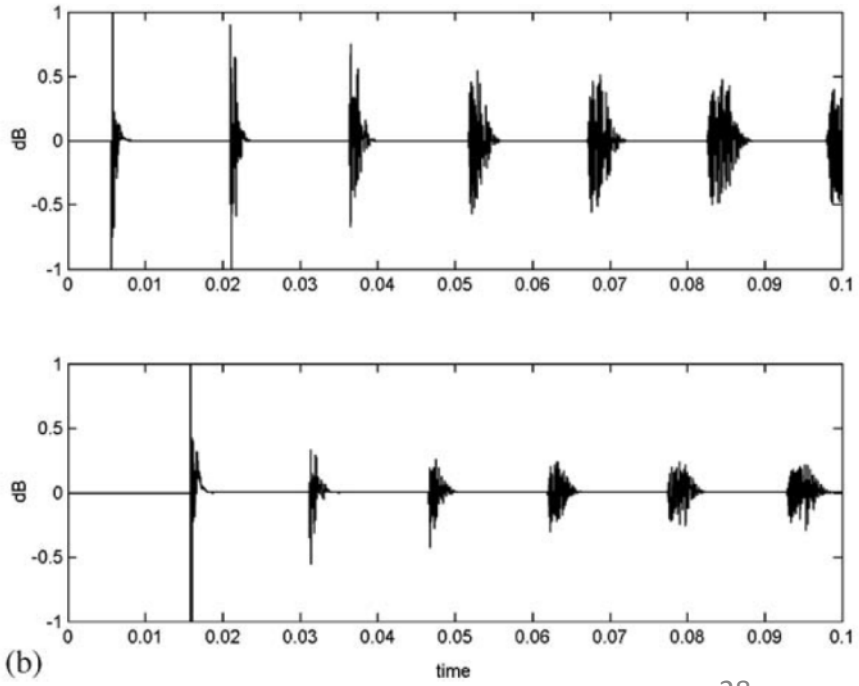
Higher order modulation

Bosco, et. al. JLT 2011

Optical Power Dynamics

- Optical power dynamics in OADM ring network
- Simulations & modeling of channel power oscillations and instability
  - L. Pavel Automatica 2004
  - Gorinevsky & Farber JLT 2004

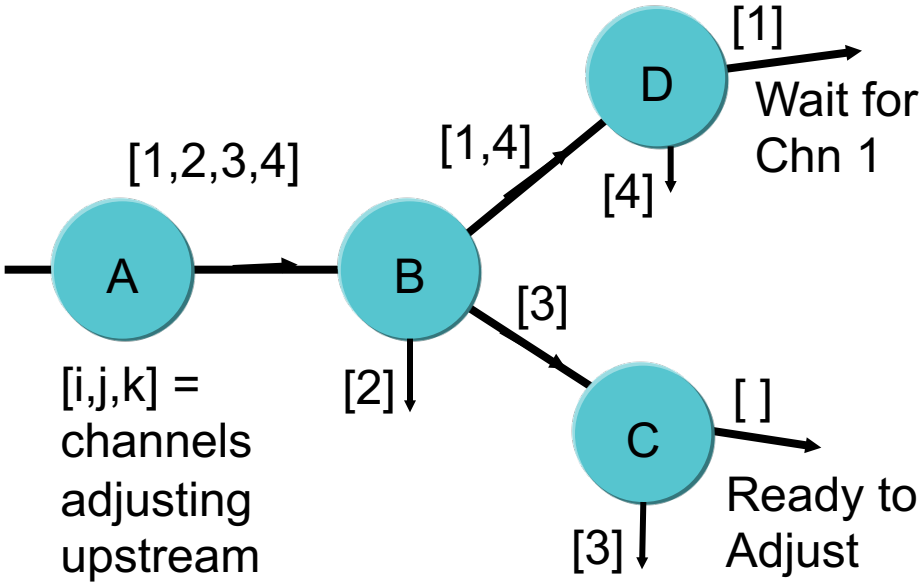


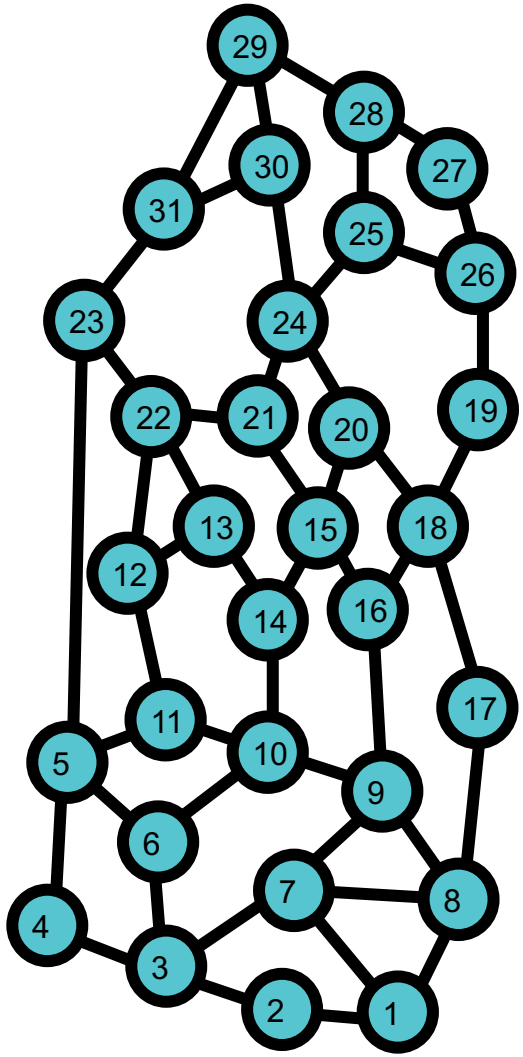


Sustained Oscillations over Long Periods

Dynamic Domain Power Control Algorithm

- Power drifts over time and new channels are provisioned: need periodic power control to stay within margins
- Adjust nodes in parallel within 'optically' isolated domains
  - Node ordering based on channel routes





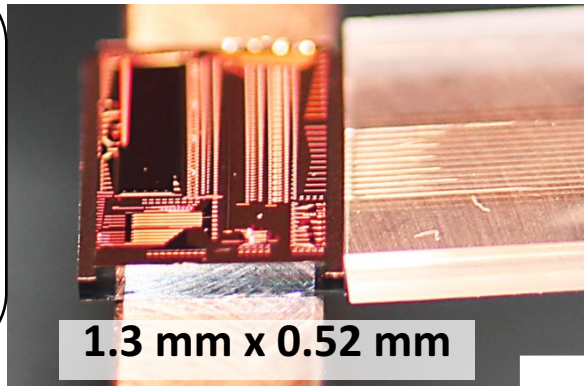
Kilper & White OFC 2007

WDM Network Node on-a-Chip: Lower performance, but much lower cost

R. Aguinaldo, H. Grant, S. Mookherjea (UCSD) + Sandia

Objectives:

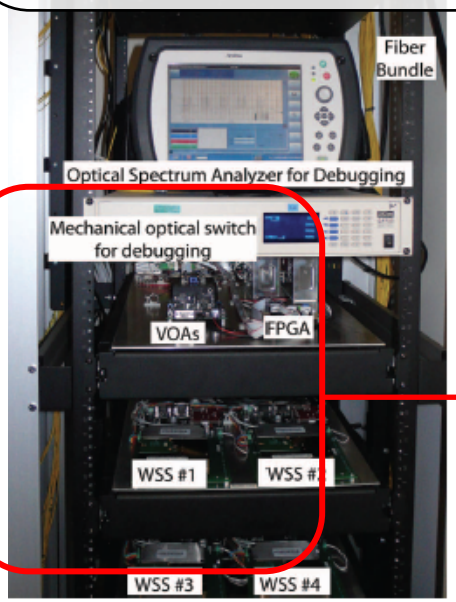
- tunable drop (reject) 4-channel tunable add 4+1 channel VOA
- 100,000 times smaller
- approx. 250 mW
- no moving parts

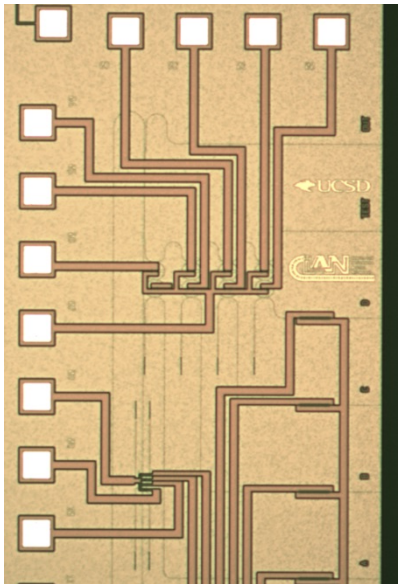


8 fiber V-groove array

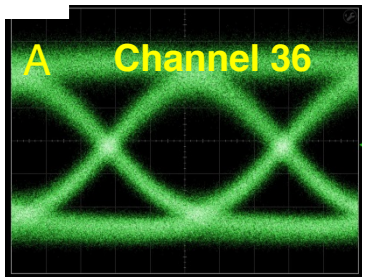
In/Out

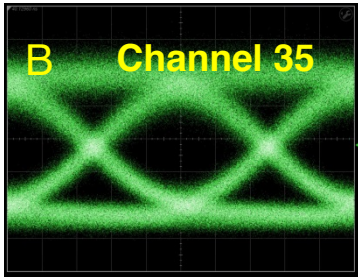
- Common (23 ch) ←
- OUT (23 ch) →
- Diag ("Test") →
- A B, C, D (in) ←

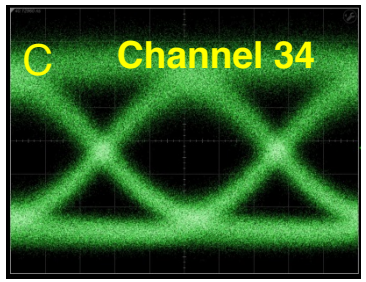


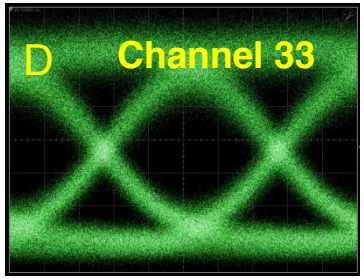


4 x 10 Gbps added from individual IN fibers to common OUT









ARIZONA CIAN

all channels on ITU-T 100 GHz grid → 100 ps ←

System Level Issues

- Transceiver & system performance interactions
  - Bigger problem for bleeding edge performance
  - Transceivers complex systems on their own
- Blocking bad corner cases
- Handling the wide range of system functions
- System testing pulls in margins
  - Too many uncertainties
- Control dynamics
  - Optical power dynamics

ARIZONA CIAN

Research Questions

- At what metro reach (number of node hops) do the different disaggregation models become problematic? For which transceiver types?
- How does physical layer software control scale with number of nodes?
  - DICONET and other examples for long haul need to be adapted here
  - Need tools to develop and test control at scale (see next talk)
- What components can be scaled to very large numbers?
  - Need integrated photonics

ARIZONA CIAN

Conclusions

- Computing systems are going through multiple rounds of disaggregation in order to continue hyperscale growth
  - Market and/or performance driven architectural change
- 5G creates potential for optical systems to jump to hyperscale models
- Not just about opening competition for transceivers, need full network design for hyperscale growth
- Transmission engineering remains an obstacle
  - Hardware & Software
  - Need new tools tackle problem (machine learning?)
- Savings need to come from high volumes: need to think hyperscale

ARIZONA CIAN

www.cian-erc.org

Center for Dis-Integrated and Dis-Aggregated Networks

Thank You

Our Group:
https://wp.optics.arizona.edu/dkilper/

CIAN:
www.cian-erc.org