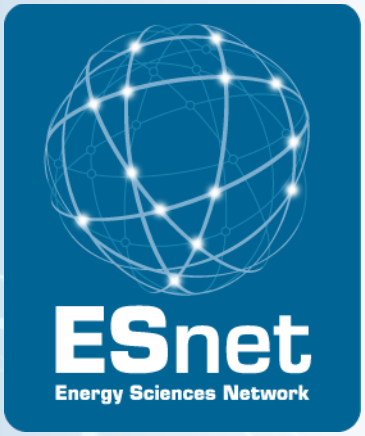


Software Defined Networking for Big-Data Science

Eric Pouyoul

Chin Guok (Presenting)

Inder Monga

DOE NGNS PI Meeting

Mar 18-20, Emeryville, CA

Contents

Network and Service Trends

What is Software Define Networking?

ECSEL: Network Control End-site to End-Site IDC

OneSwitch: A Virtual Distributed (WAN) Programmable Switch

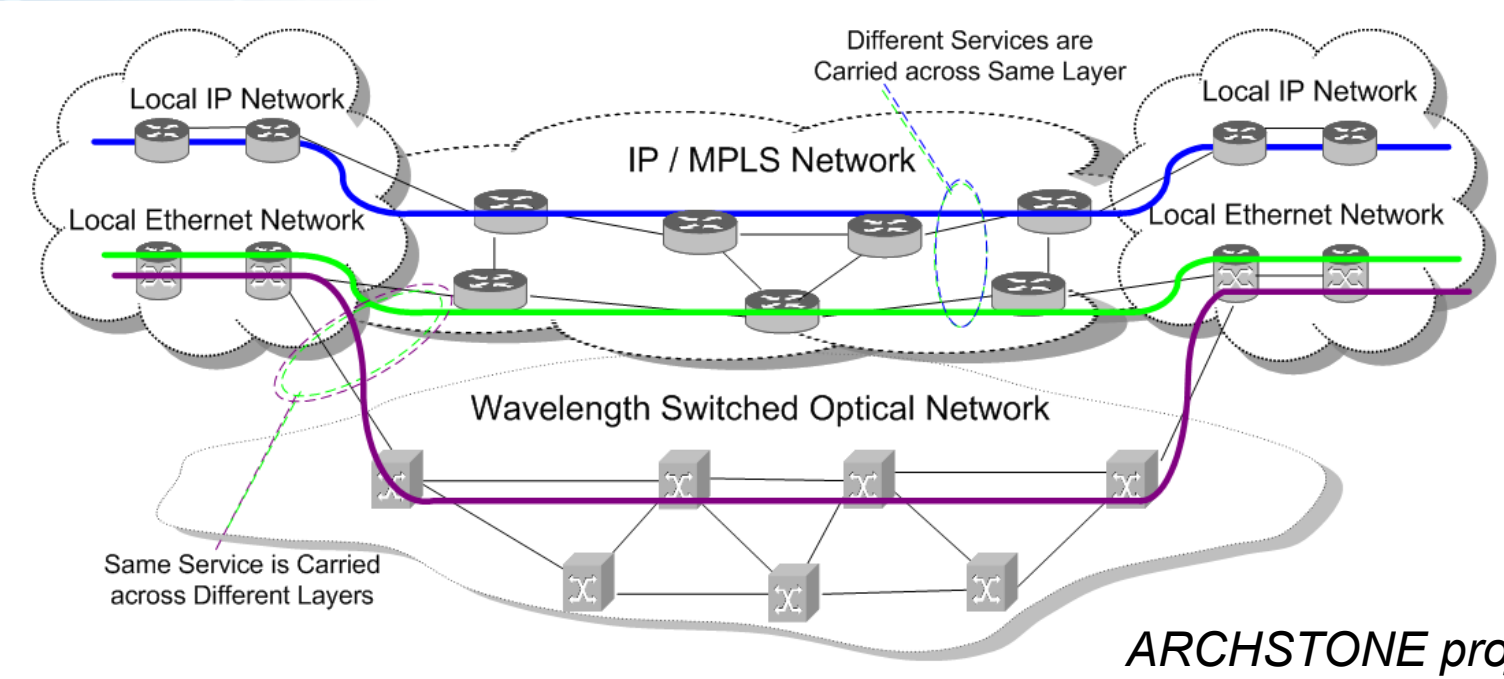
OTS: Optical Transport Switch

OpenFlow Controlled Forwarding Router

Network Trends

Network and Service Trends (1)

- **(Dynamic) Optical Bypass**
  - Separation of large flows, from L2/MPLS circuits to L1 circuits
  - Traditionally has been very complicated, vendor dependent and not-interoperable

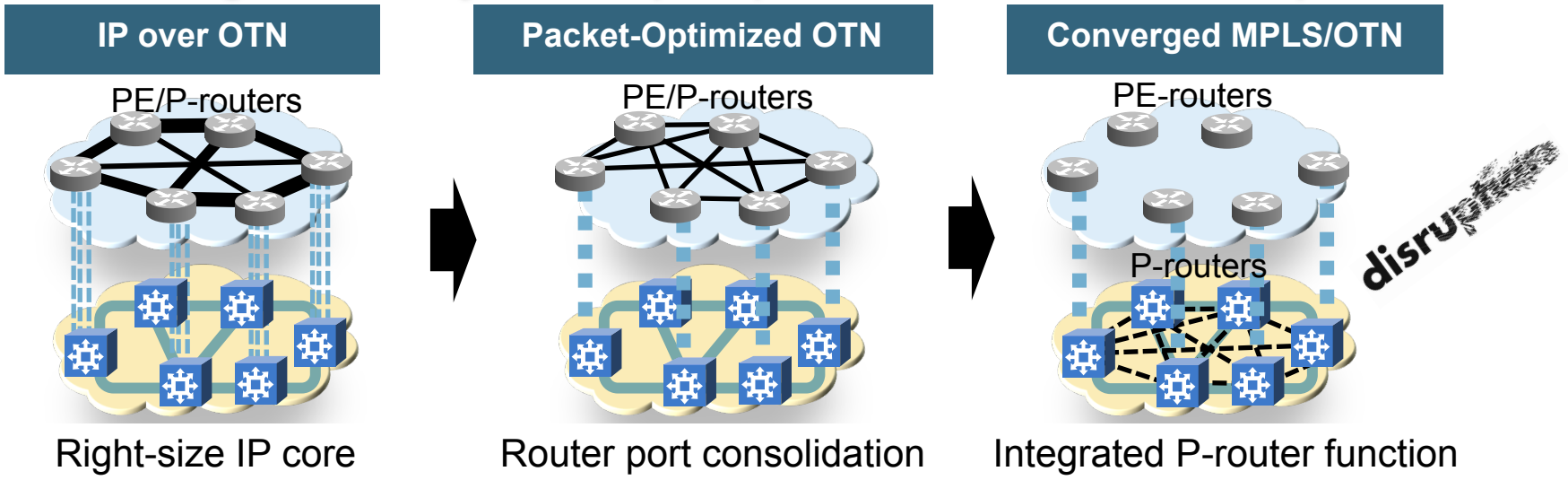


ARCHSTONE project DOE

Network and Service Trends (2)

- **Packet-Optical Integration driving layer collapse**
  - Highly functional packet switching interfaces in Optical transport nodes, thanks to merchant silicon
  - Highly integrated LH DWDM optics in packet routers, thanks to coherent technology and optical integration

Converged Multi-layer Transport (ROADM/WDM/Packet/OTN)



Source: Infinera

Network and Service Trends (3)

- **No uniform way to manage Multi-vendor, Multi-layer networks or multi-domain optical bypass**
  - Packet and Optical devices managed in totally different ways
  - Multi-vendor networks are islands
  - No global topological visibility
    - GMPLS and vendor-specific hacks needed in current approach

Packet World	Transport World
• Connectionless • Enterprise origins • Dynamic flows • Inband control plane • Numerous distributed CP solutions • Monolithic, closed systems	• Connection (circuit) oriented • Service provider origins • Static pipes • EMS/NMS + Cross-connect paradigm • Nascent CP (GMPLS) • Open, programmable systems

Source: Infinera

Network and Service Trends (4)

- **Revolutionize the application use of the network**
  - *Cees*
    - *"Show **Big Bug Bunny** in **4K** on **my Tiled Display** using **Green** Infrastructure"*
    - *"Get **5 PB** of **CDIAC NDP-032** Climate Data from **Livermore**, **Colorado** and **UK** Data centers to **my Amazon cloud storage** by **tomorrow evening**"*
  - Dynamic programmability needs to extend all the way from the application to the physical layer
  - Needs to be practical (dynamic wavelengths still have challenges)

What is Software Defined Networking

What is Software-Defined Networking?

(as defined by Scott Shenker, October 2011)

http://opennetsummit.org/talks/shenker-tue.pdf

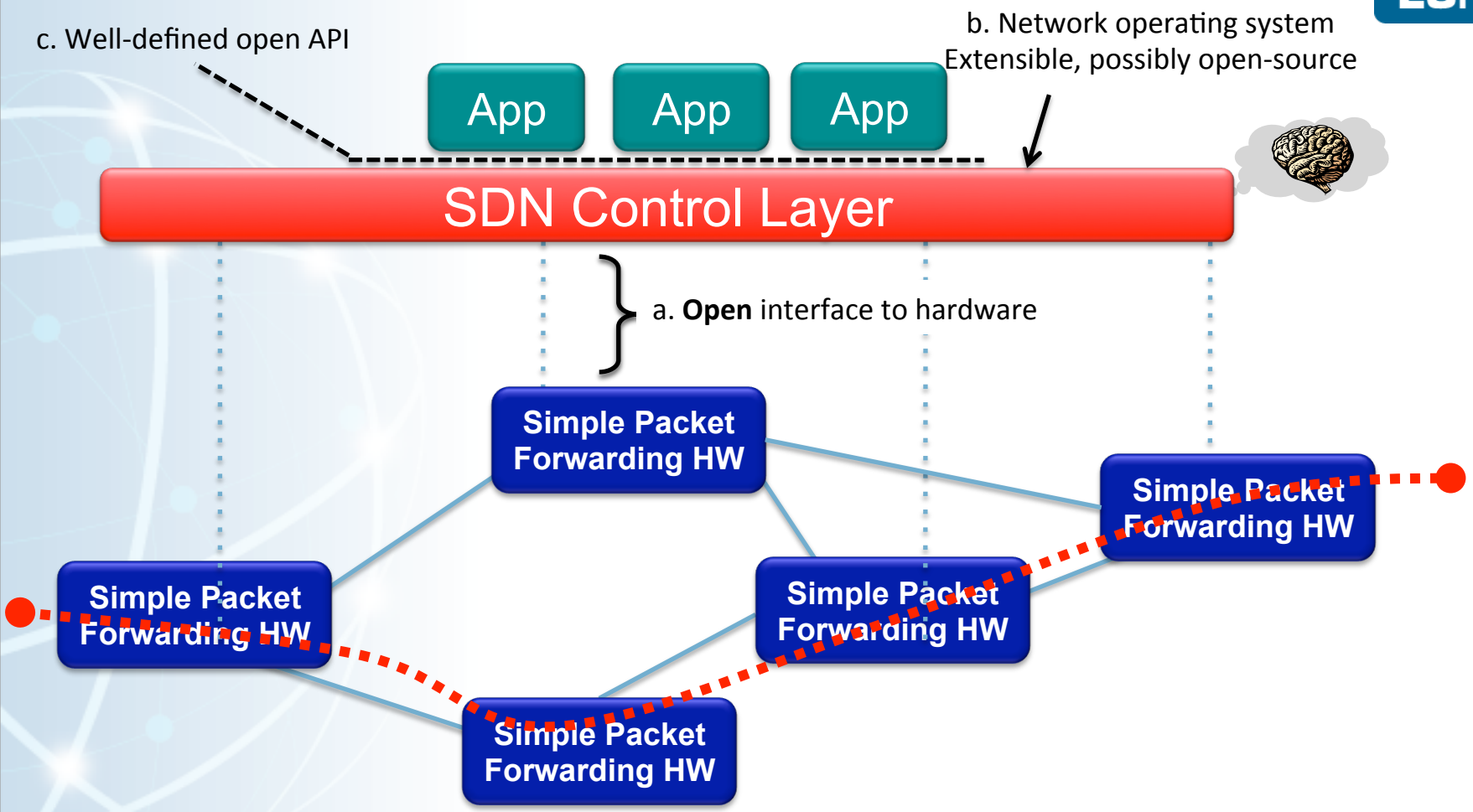
"The ability to master complexity is not the same as the ability to extract simplicity"

"Abstractions key to extracting simplicity"

"SDN is defined precisely by these **three abstractions**

- Distribution, forwarding, configuration "

Software Defined Networking : **Open**, Logically Centralized, Programmable



Source: ONF Tutorial, Open Network Summit 2012

ESCEL: Network Control End-site to End-site IDC

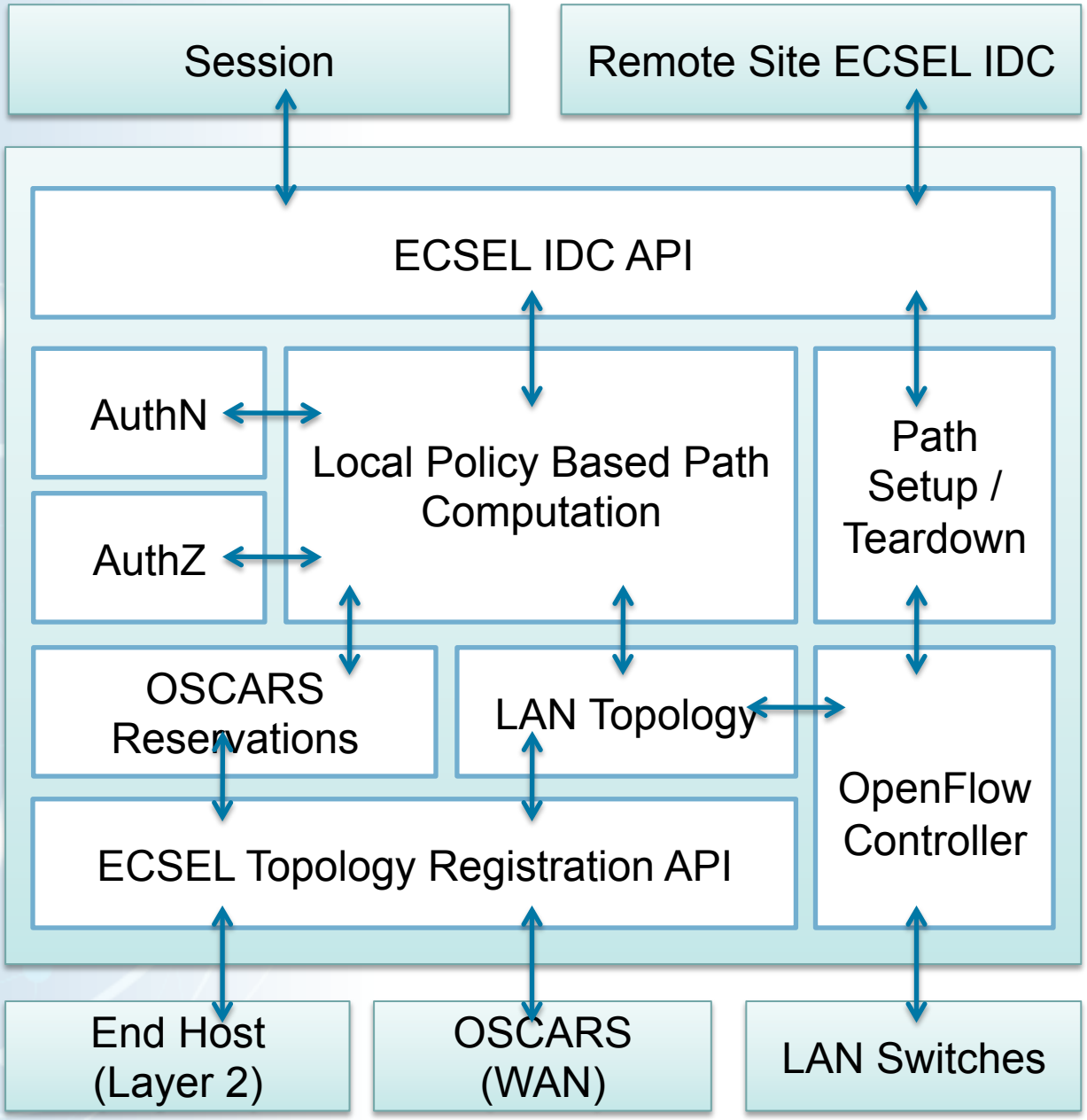
ECSEL: Network Control End-site to End-site IDC

OSCARS already brokers WAN resources across domains and could include LAN resources, however:

- Administrative boundaries limit OSCARS' control.
- LAN topologies are too dynamic for OSCARS path computation.
- Circuits are only point to point.

ECSEL is a modified version of OSCARS supporting end site to end site negotiation of OSCARS resources, and on the fly LAN provisioning.

Each administrative domain runs its own ECSEL instance


Session
Remote Site ECSEL IDC
ECSEL IDC API
AuthN
AuthZ
Local Policy Based Path Computation
Path Setup / Teardown
OSCARS Reservations
LAN Topology
OpenFlow Controller
ECSEL Topology Registration API
End Host (Layer 2)
OSCARS (WAN)
LAN Switches

LAN Auto-Discovery with OpenFlow

OpenFlow switches periodically broadcast LLDP packets onto all the active ports. This allows to maintain dynamic topologies.

The ECSEL OpenFlow controller learns the topology from the switches.

Hosts connected to an OpenFlow switch will also receive LLDP packets discovering where they are in the LAN topology.

Hosts register themselves to the site ECSEL.

WAN Resources Registration

OSCARS circuits are reserved by network engineers (manual) or by a middleware/application (automated).

Network engineers or middleware register OSCARS circuits to ECSEL, adding metadata, such as project names, used later by ECSEL to determine if a user or application has the right to use that resource.

In addition, ECSEL itself can request new OSCARS circuits which are automatically registered.

The OSCARS circuit registration includes in which OpenFlow/port it terminates.

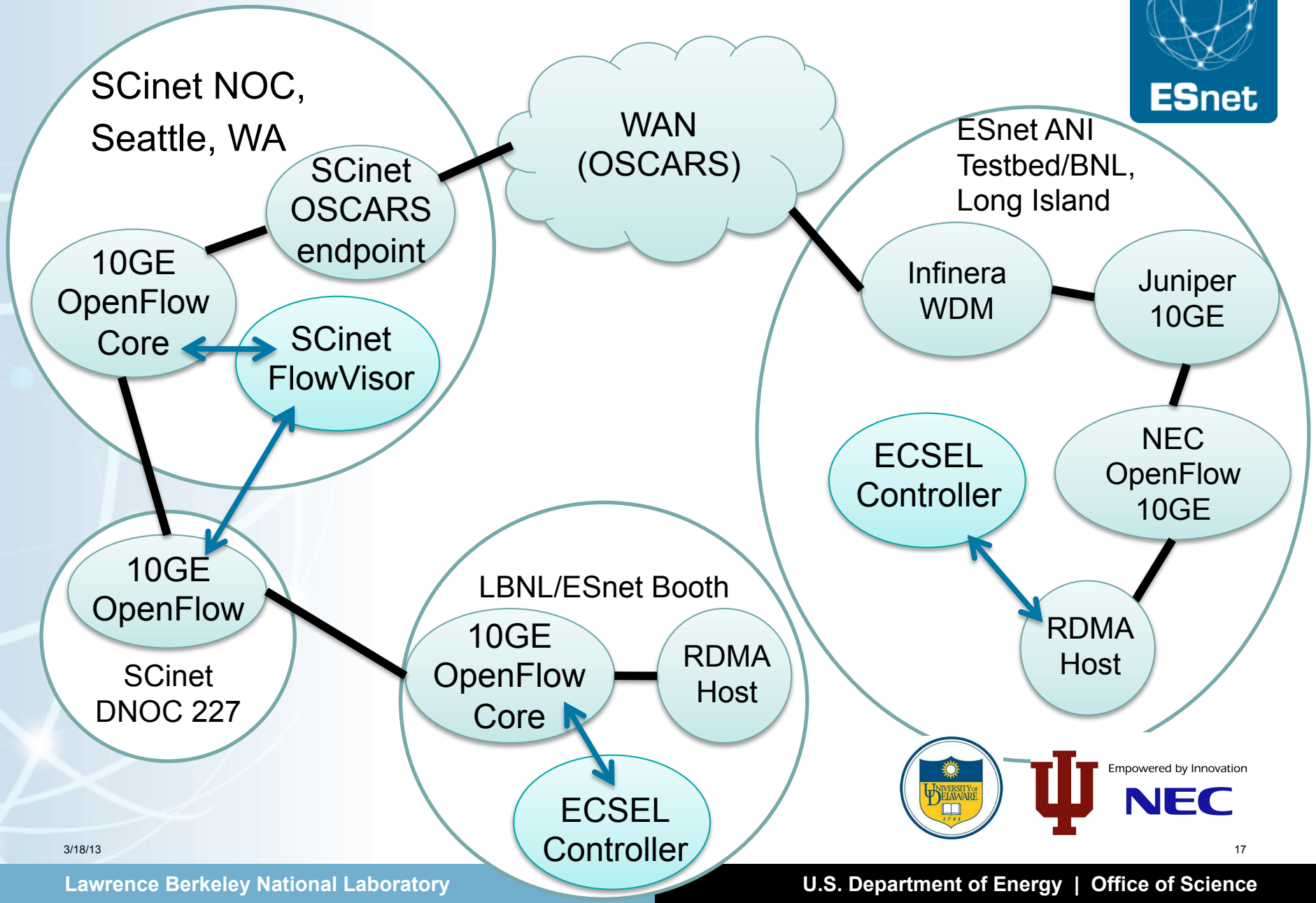
Path Computation / Site Policy

ECSEL's PCE first searches for existing and available OSCARS circuit, matching AuthN and AuthZ criteria, and also matching metadata associated with the OSCARS circuit (project)

Like any IDC, ECSEL asks its peer on the other end of the circuit if it accepts it.

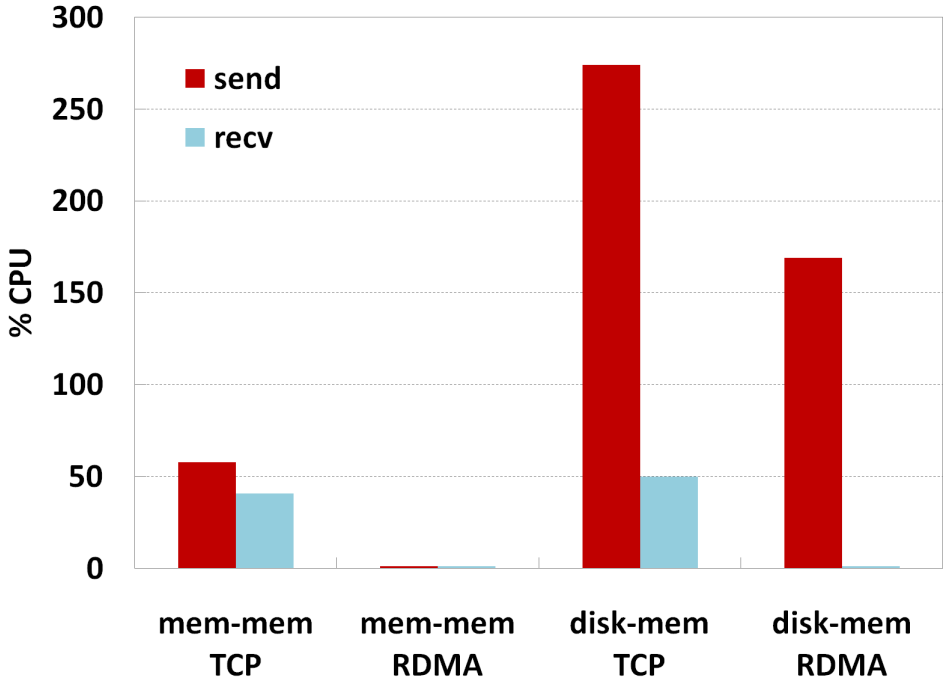
The path in the LAN is not computed at the time the reservation is made, but only during provisioning: this allows for the LAN topology to change.

Super Computing 2011 Demonstration



Results: it works !

Tool	Protocol	Gbps
nuttcp	TCP	9.9
nuttcp	UDP	9.9
xfer_tst	TCP	9.9
xfer_tst	RDMA	9.7
GridFTP	TCP	9.2
GridFTP	UDT	3.3
GridFTP	RDMA	8.1



Almost as fast as TCP, but for only a fraction of the CPU usage

Impact and Future (1)

- Software-defined Networking and OpenFlow
  - OpenFlow easily extends OSCARS virtual-circuits inside campuses
  - End-to-end circuits with OpenFlow enable simpler deployment of non-TCP/IP friendly protocols
  - Middleware like ECSEL provides the architecture to end-site network administrators for automation, policy control and resource management.

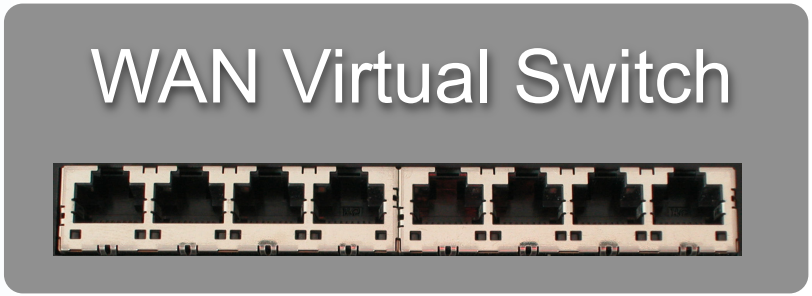
Impact and Future (2)

- RDMA
  - Layer 2 RDMA behaves well over long distance as long as the path is controlled, end host to end host.
  - RDMA eliminates the CPU bottleneck
  - Sites need to deploy services allowing local provisioning
  - Applications need to integrate RDMA to maximize performance.

Due to CPU limitation, host based protocols like TCP will not be able to take advantage of high performance networks and deliver the required data transfer capability.

OneSwitch: A Virtual Distributed (WAN) Programmable Switch

New Network Abstraction: "WAN Virtual Switch"



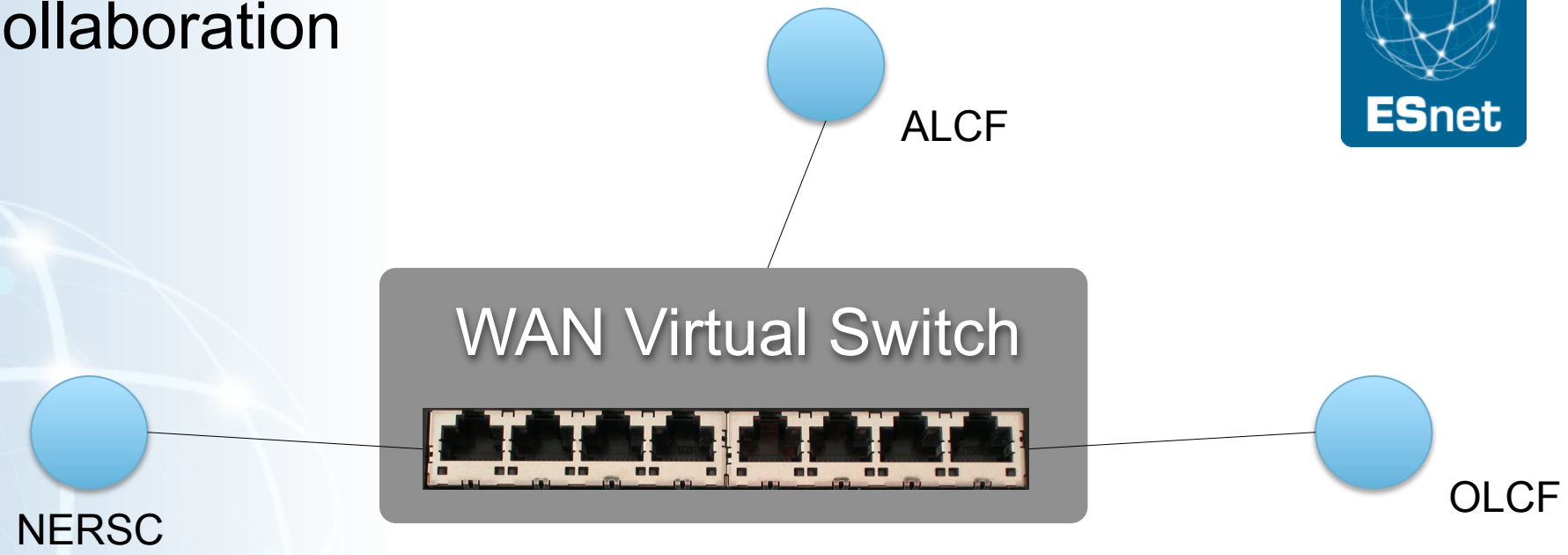
Simple, Multipoint, Programmable

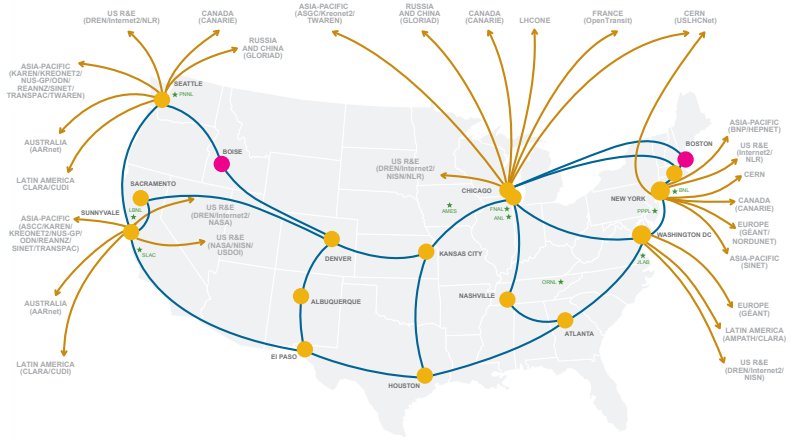
Configuration abstraction:

- Expresses desired behavior
- Hides implementation on physical infrastructure

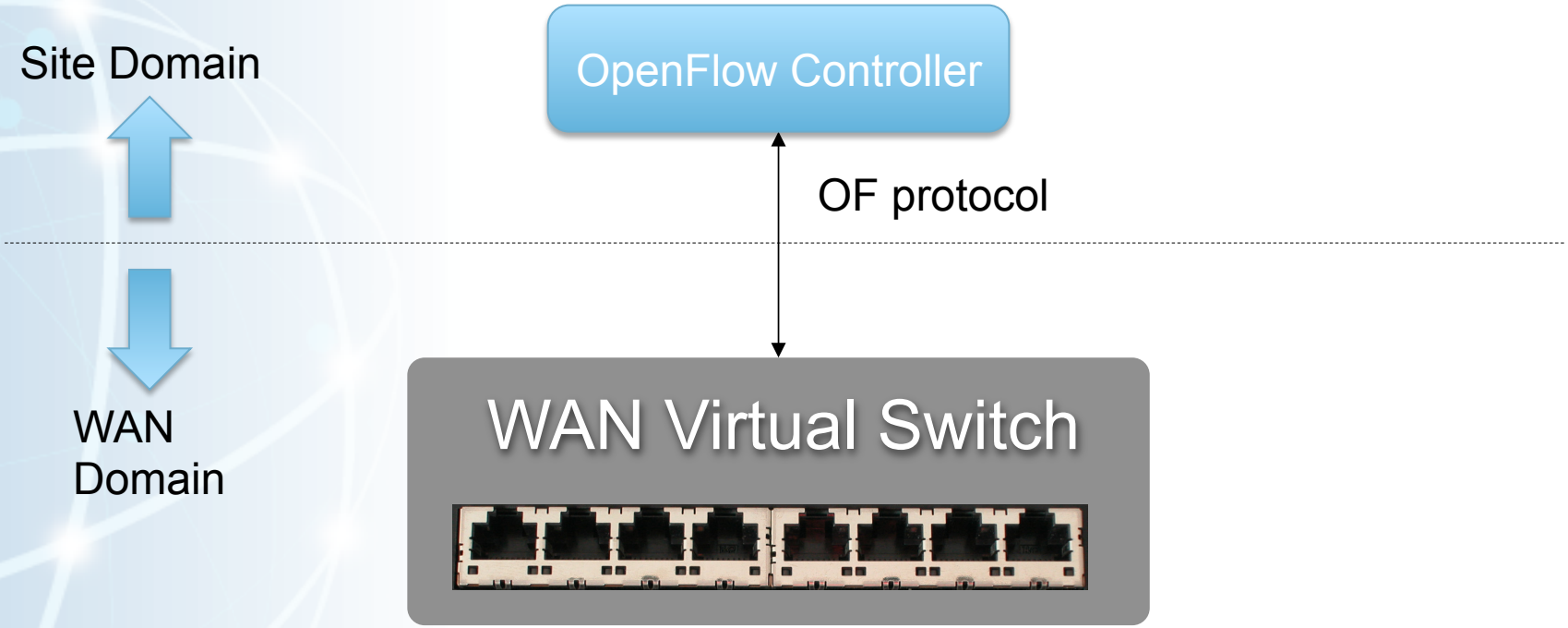
It is not only about the concept, but implementation

Simple Example: One Virtual Switch per Collaboration





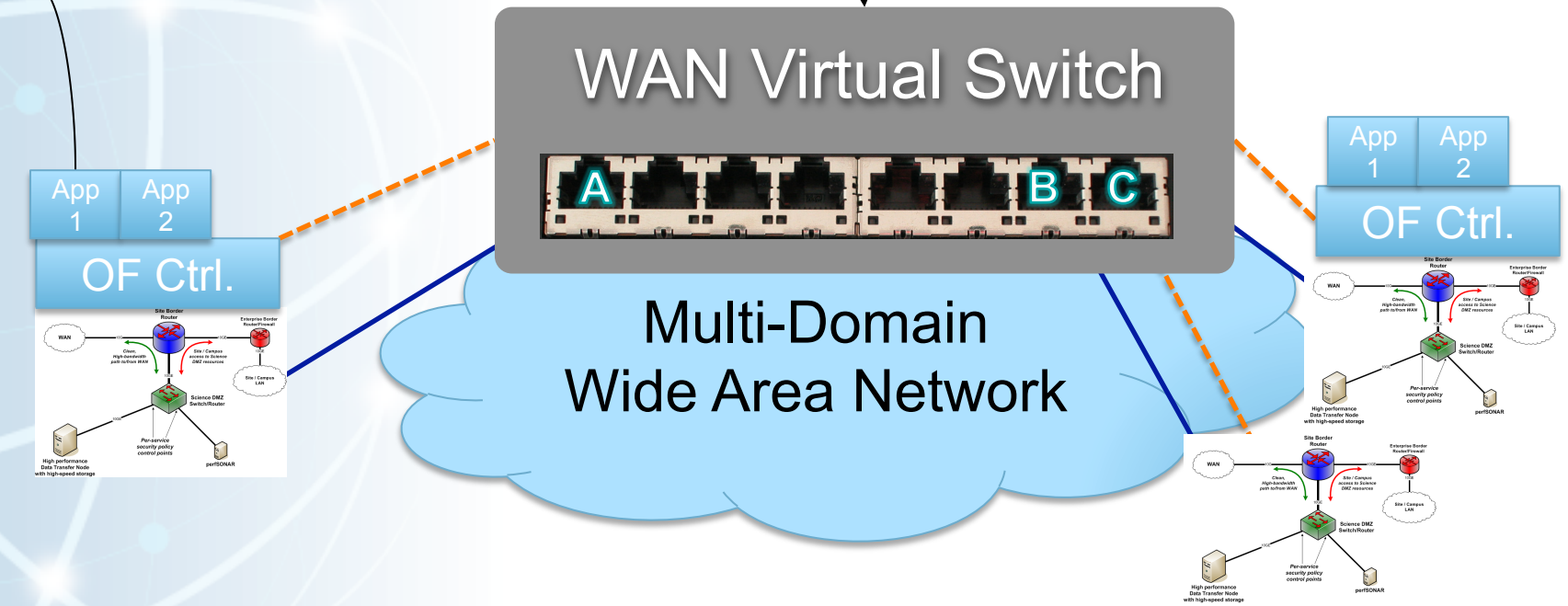
Programmability



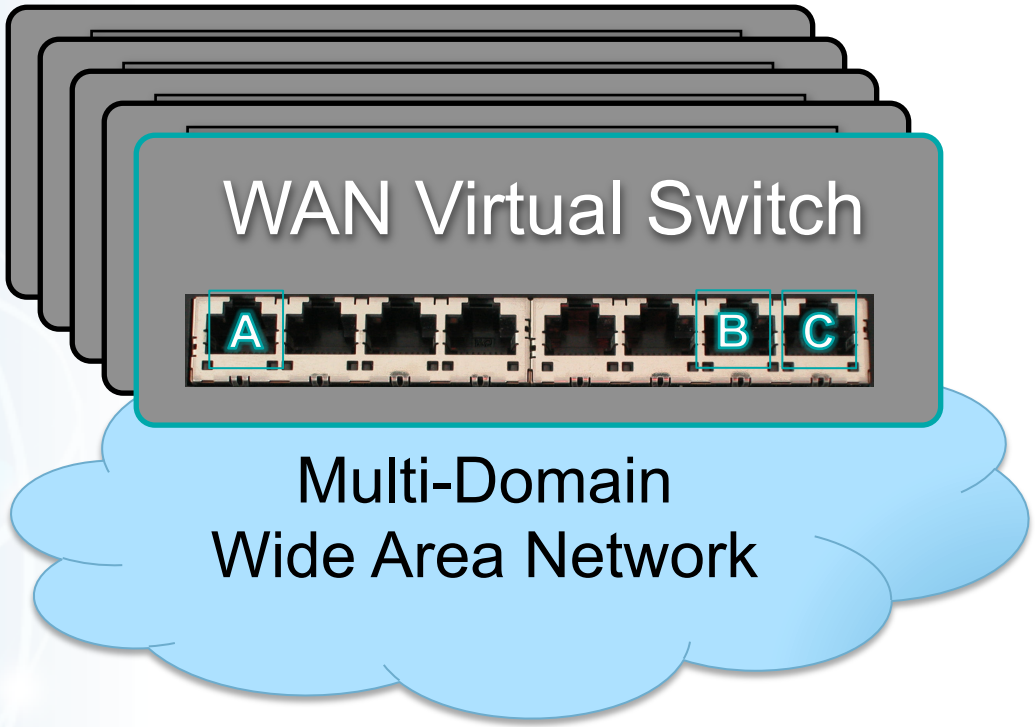
Expose ‘flow’ programming interface leveraging standard OF protocol

“Programmable” by end-sites

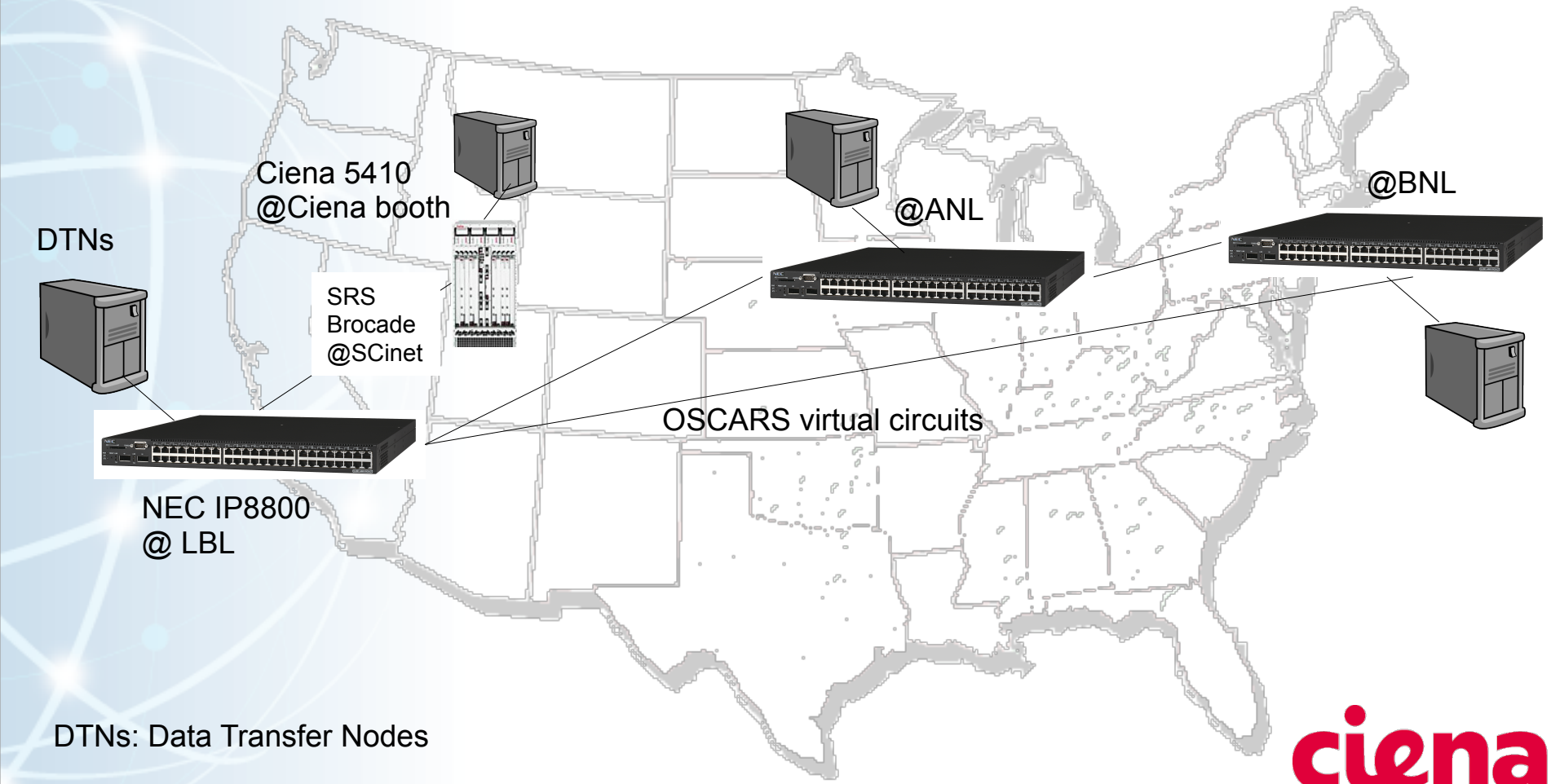
Program flows:
Science Flow1: A → B, QoS, Label
Science Flow2: A → C, VLAN
Science Flow3: A → B,C



Many collaborations, Many Virtual Switches



SRS Demonstration Physical Topology

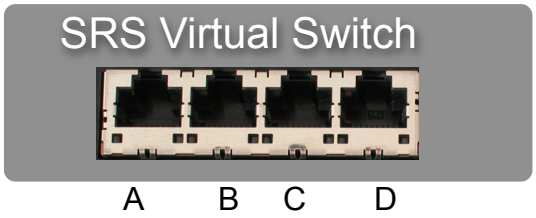


DTNs: Data Transfer Nodes

Virtual Switch Implementation: Mapping abstract model to the physical

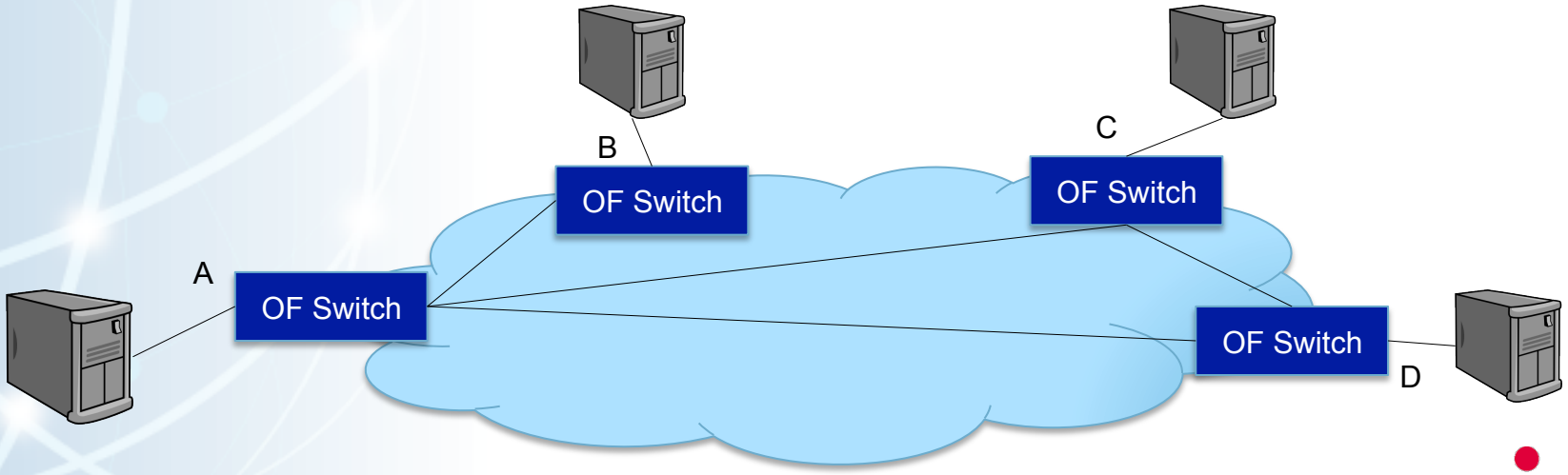
Create Virtual switch:

- Specify edge OF ports
- Specify backplane topology and bandwidth
- Policy constraints like flowspace
- Store the switch into a topology service

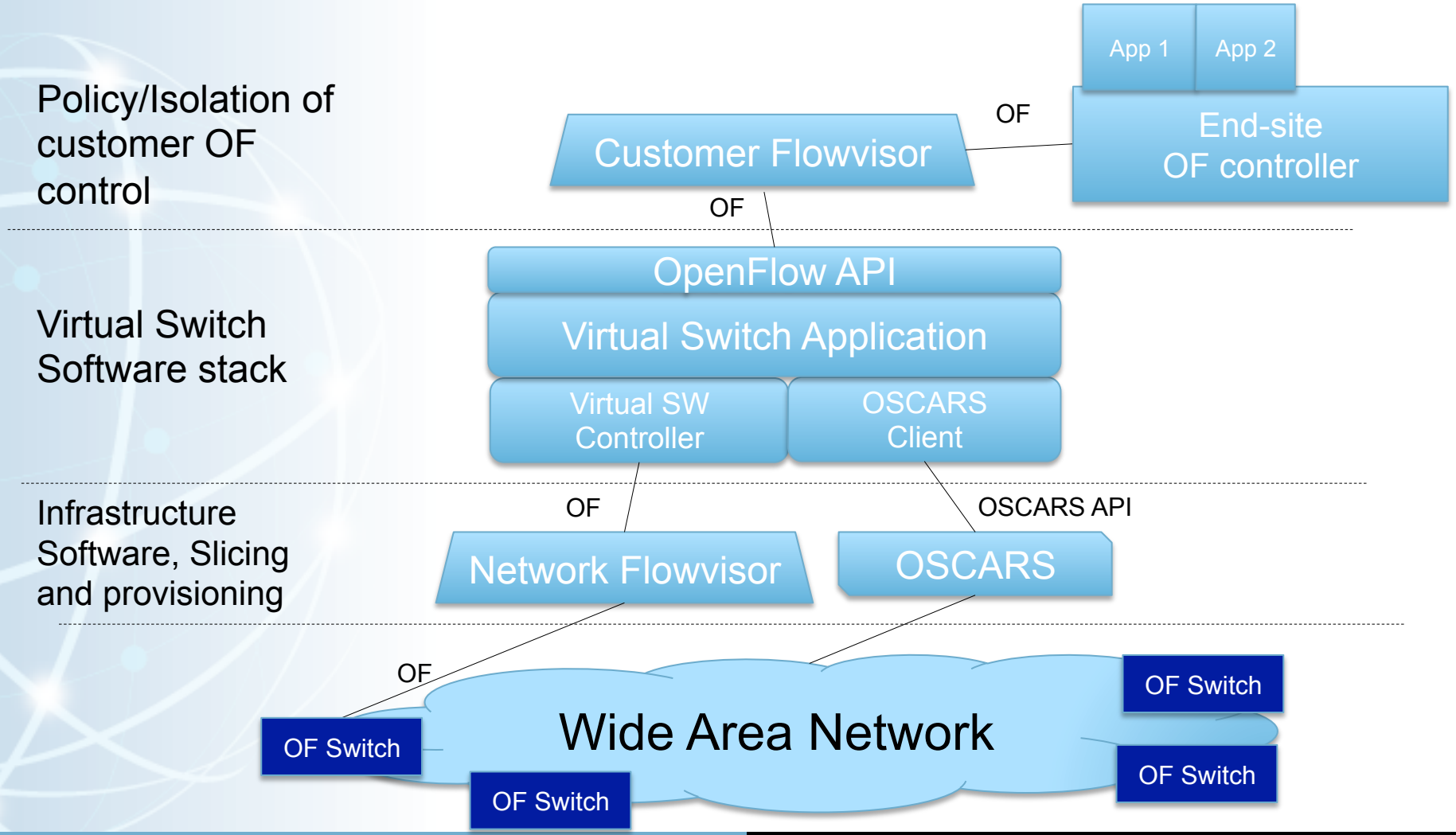


Virtual

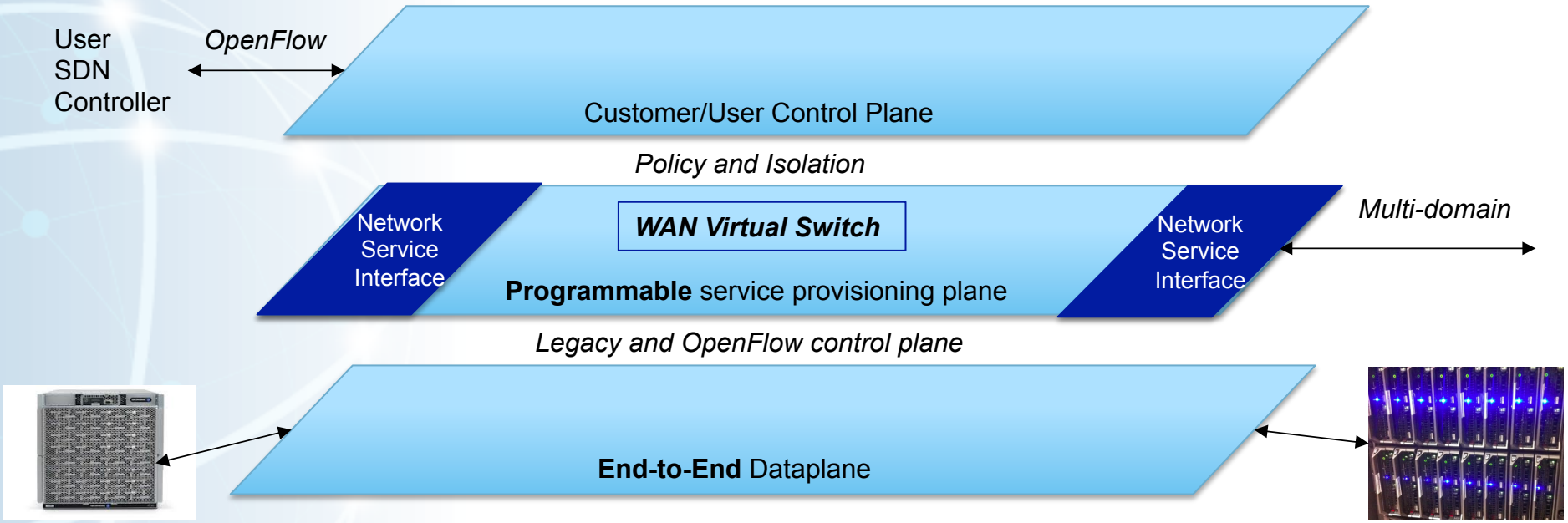
Physical



WAN Virtual Switch: Deploying it as a service



What does this mean for networking?



- Creation of a programmable network provisioning layer
- Sits on top of the “network OS”

Summary

- Powerful network abstraction
  - Files / Storage
- Benefits
  - **Simplicity** for the end-site
    - Works with off-the-shelf, open-source controller
    - Topology simplification
  - **Generic** code for the network provider
    - Virtual switch can be layered over optical, routed or switched network elements
    - OpenFlow support needed on edge devices only, core stays same
  - **Programmability** for applications
    - Allows end-sites to innovate and use the WAN effectively

Acknowledgements

Many folks at ESnet who helped with the deployment and planning

- Sanjay Parab (CMU), Brian Tierney, John Christman, Mark Redman, Patrick Dorn among other ESnet NESG/OCS folks

Ciena Collaborators:

- Rodney Wilson, Marc Lyonnais, Joshua Foster, Bill Webb

SRS Team

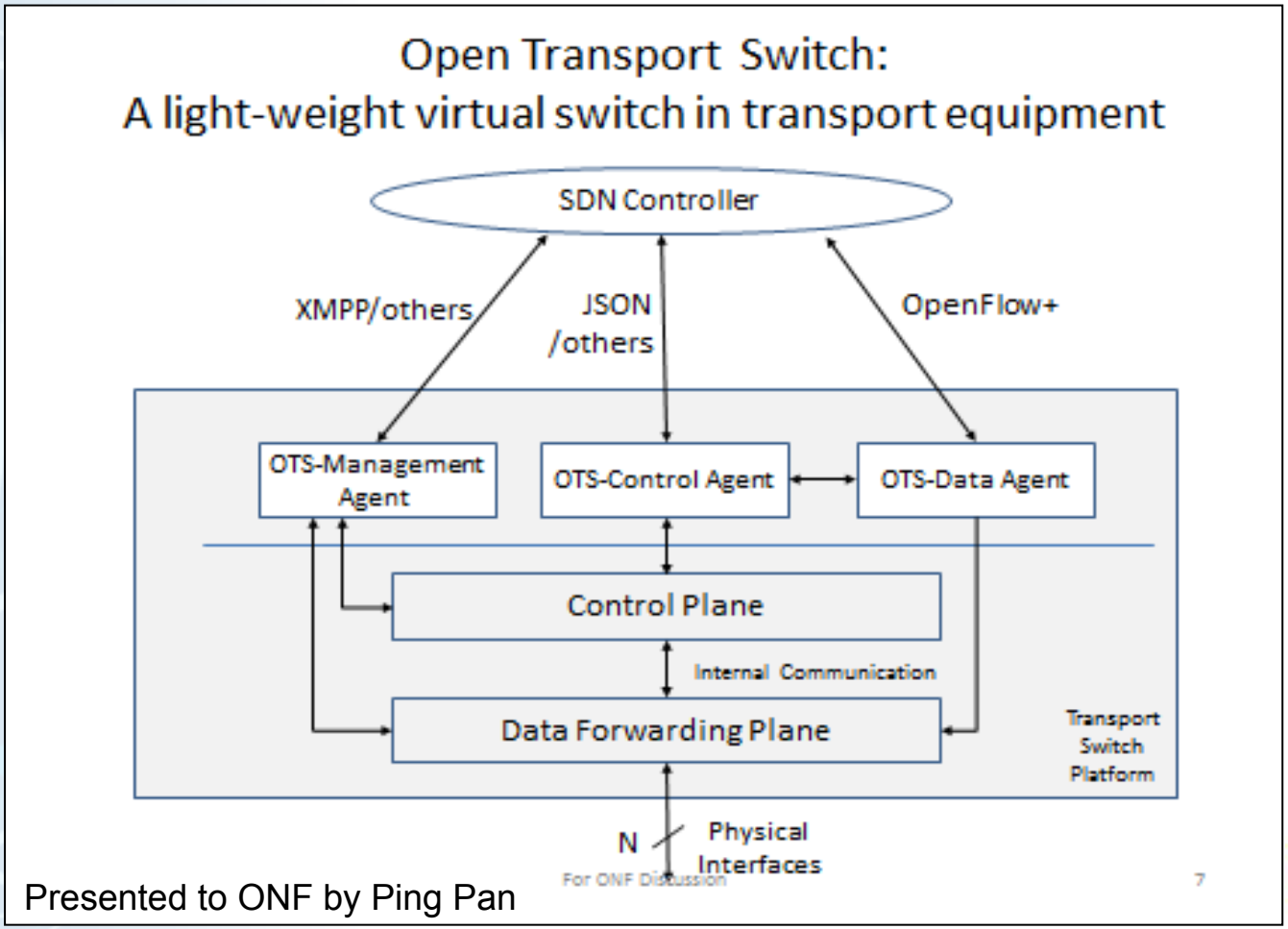
- Andrew Lee, Srini Seetharaman

DOE ASCR research funding that has made this work possible

OTS: Optical Transport Switch

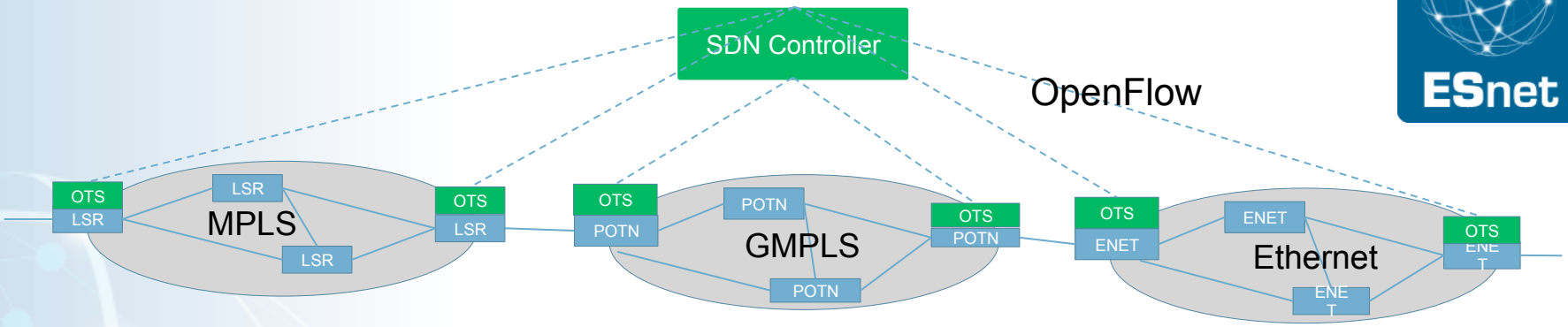
Open Transport Switch Architecture: Uniform APIs for Optical hardware

Expose the switching, cross-connect and flow aggregation capability

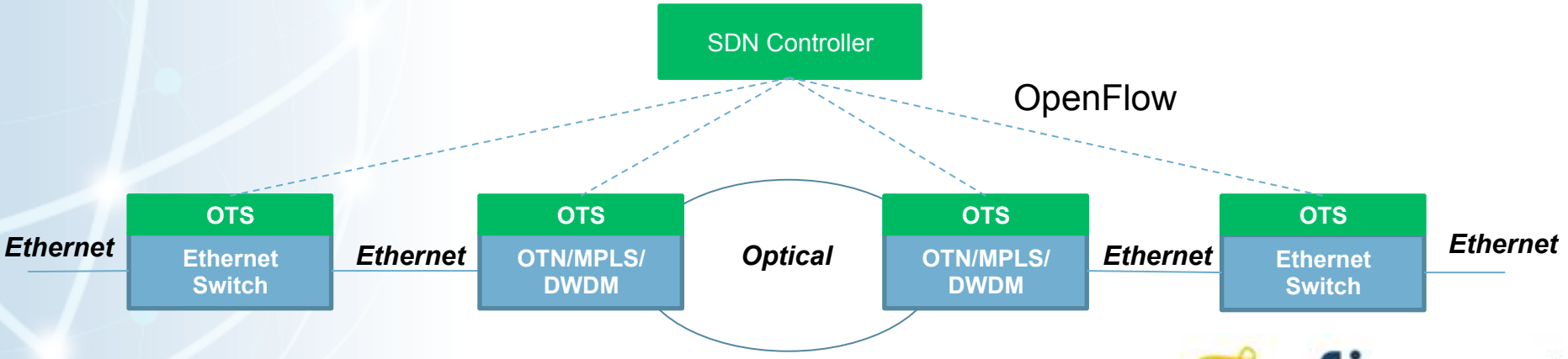


Presented to ONF by Ping Pan

Implicit and Explicit provisioning mechanisms



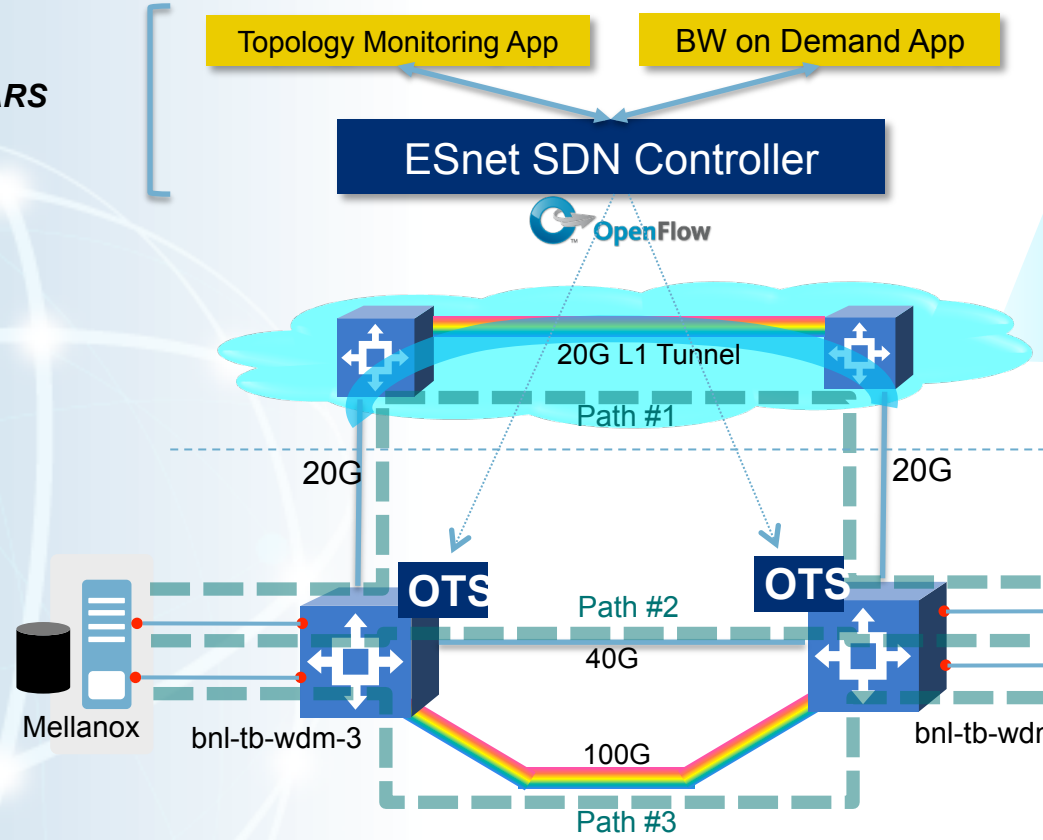
Indirect (Implicit) Path Set Up
(provision edge nodes only,
leverage existing signaling plane)

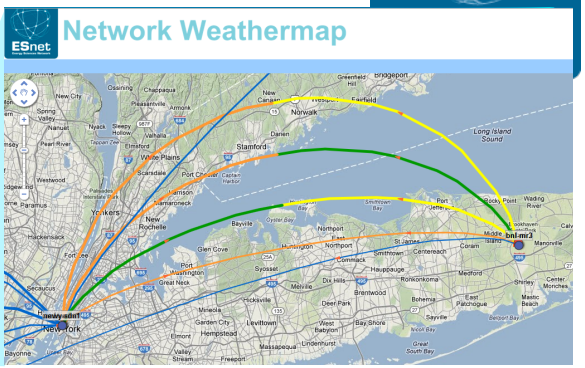


Direct (Explicit) Path Set Up
(provision every node)

infinera
what THE NETWORK will be

ESnet Transport SDN Demo





ESnet LIMAN Production Network

Brookhaven National Laboratory Testbed

SDN Controller communicating with OTS via OpenFlow extensions

Bandwidth on Demand application for Big Data RDMA transport

3 physical transport path options (with varying latencies)

Implicit & explicit provisioning of 10GbE/40GbE services demonstrated

infinera
what THE NETWORK will be

Why Transport SDN?

Automation

- Centralized management applications aka OSCARS

Scaling

- End-to-end service delivery
- Works for IP and non-IP protocols

Virtualization

- Multi-layer abstracted to a single end-to-end connection

What's key about OTS?

OpenFlow: Single provisioning protocol

- SNMP, JunOScript, CLI, TL1, WS, mTOSI….

SDN architecture allows multi-layer topology view

- The reason why GMPLS failed

SLA's require monitoring and management

- Threshold based monitoring
- Different application of interest in the network

Summary

Network flexibility is key to meet the widely divergent needs of the customer-base

Multi-layer networking direction is critical for high-performing applications

SDN enables us a simple, interoperable way to build application intelligence and abstractions.

SDN Demo Collaboration Team

ESnet

- Chin Guok
- Andy Lake
- Inder Monga
- Mike O'Connor
- Eric Pouyoul

Brookhaven National Labs

- Scott Bradley
- Tan Li
- Dantong Yu

Infinera

- Chris Liou
- Dharmendra Naik
- Ping Pan
- Sharfuddin Syed
- Abhinava Sadasivarao

OpenFlow Controlled Forwarding Router

OpenFlow Controlled Forwarding Router

Develop an OF controlled forwarding router (Layer 3)

Investigate FIB compression or other FIB management strategies to reduce number of flows in flow table

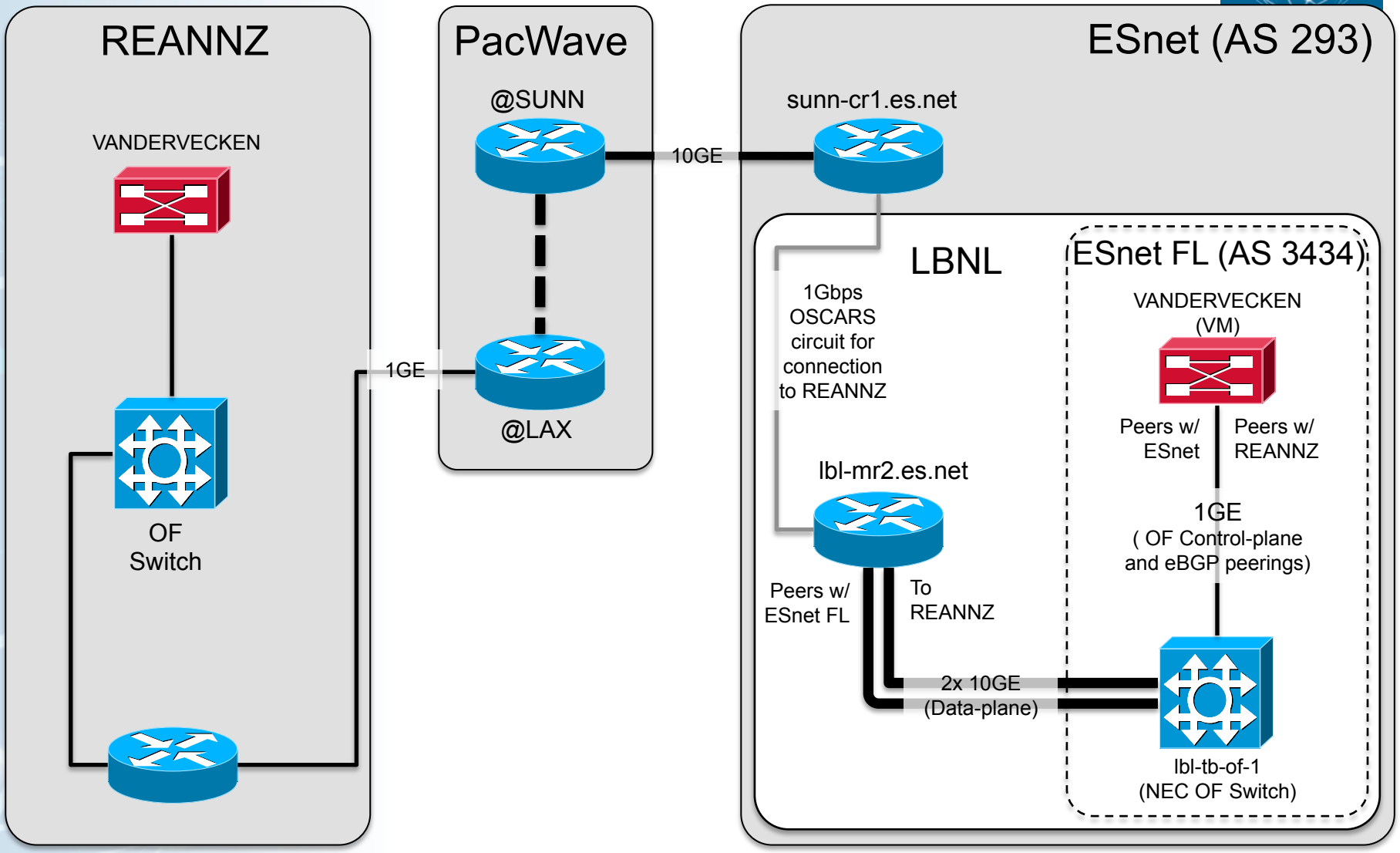
Goal is to demonstrate RIB/FIB scaling for large scale deployments

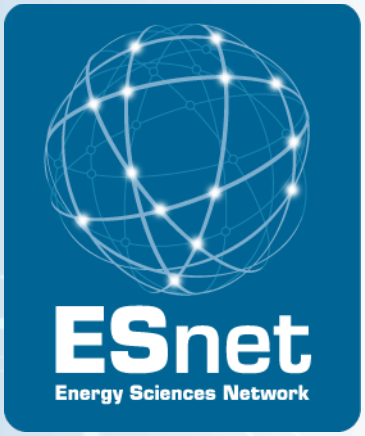
Leverage RouteFlow and Quagga

Project is done in collaboration with:

- ESnet
- Google Network Research
- REANNZ

SDN Routing Demo





Thank you!