Challenges When Designing A Distributed SDX

Sean Donovan, Russ Clark
Georgia Tech

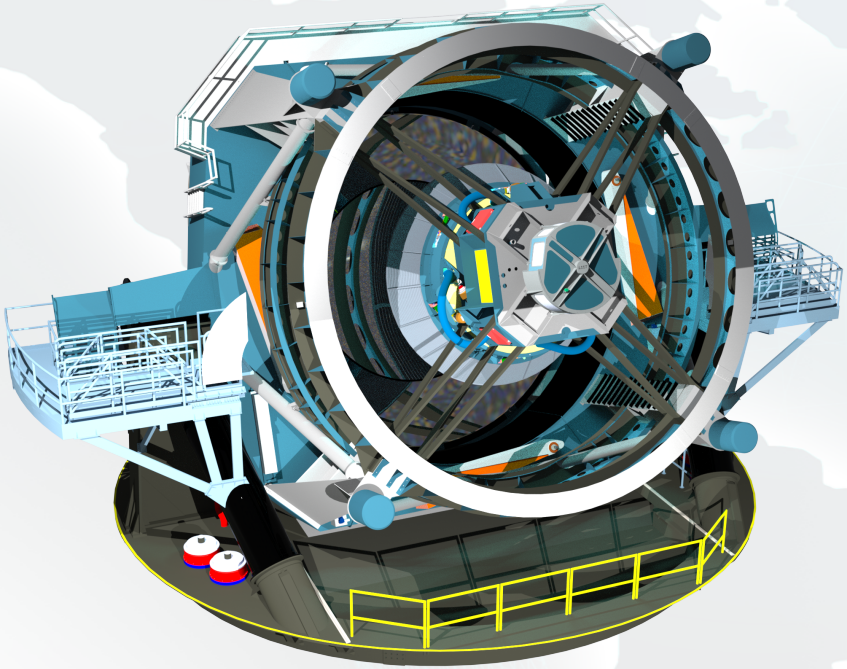
Jeronimo Bezerra
Florida International University

NSF International Research
Network Connections (IRNC)
Grant #ACI-1341024

Julio Ibarra, Heidi Morgan

Joaquin Chung, Cas D'Angelo,
Ankita Lamba, John Skandalakis

Large Synoptic Survey Telescope (LSST)



- High in the mountains in northern Chile
- Engineering First Light in 2019, Science First Light in 2021

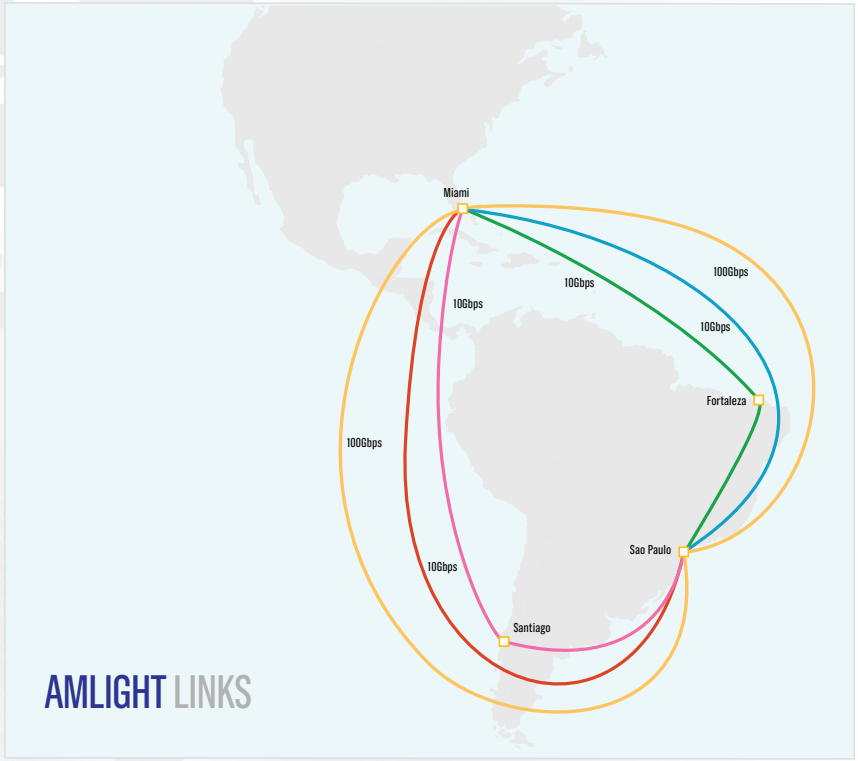
Source: https://www.lsst.org/gallery/telescope-rendering-2013

Huge Bandwidth Requirements

- 8.4 meter primary mirror with 3.2 Gigapixel sensor
- 12.7 GB image taken every 17 seconds
- Needs to be sent from Chile to NCSA/Illinois in 5 seconds
- Peak burst bandwidth of 65 Gbps
- In use all night long

New Connection

- Amlight is installing a new 100Gbps network connection between North and South America
- AtlanticWave/SDX sonnects Atlanta, Miami, and São Paulo over the AMLIGHT network
- Opportunity to innovate with the network



Agenda

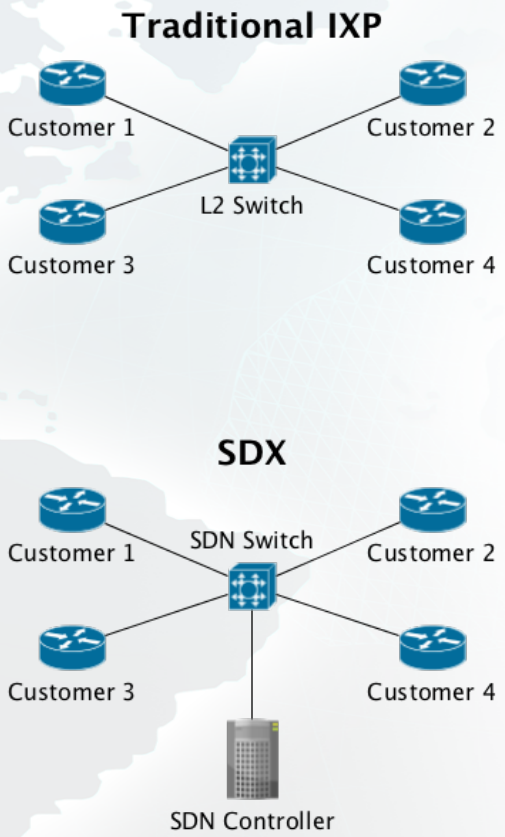
- Introduction
- Design Overview
- Functionality
- Challenges
  - Hardware
  - Abstractions
  - Security
  - Federation
  - Management
  - Sustainability
- Status

Agenda

- **Introduction**
- Design Overview
- Functionality
- Challenges
  - Hardware
  - Abstractions
  - Security
  - Federation
  - Management
  - Sustainability
- Status

Our definition of SDX

- IXP + SDN
  - Not just L2 like an IXP
  - Where participants can write rules
- Multi-site IXP
  - AMS-IX has 10 sites in and around Amsterdam
  - Same administrative domain
- New functionality enabled by SDN at the IXP
  - Not bound by BGP restrictions
  - Application-specific peering



Current SDX Deployments

- Cardigan – Wellington Internet Exchange and REANNZ
  - Very, very early implementation
  - In early 2014, was deployed for 9 months with only 1134 flows
  - Rather traditional IXP
- Maryland/WIX
  - Controller lives “above” Oscars
  - Adding compute to the mix
- PacificWave-SDX
  - This is the most like AtlanticWave/SDX, distributed on the west coast of the US
  - Also a distributed exchange between Seattle, Sunnyvale, CA, and Los Angeles, CA
  - SDX in parallel with their traditional fabric

Current Examples of SDX Research

- Gupta et al., SIGCOMM 2014 – Initial work, where our definition comes from
- Gupta et al., NSDI 2016 – Optimization work, to allow for scalability
- GENI SDX – Early work at deploying an SDX using GENI project infrastructure, still ongoing
- Work at Starlight – Working on evaluating various SDX design
- SDX taxonomy in Chung et al., SoutheastCon 2016.

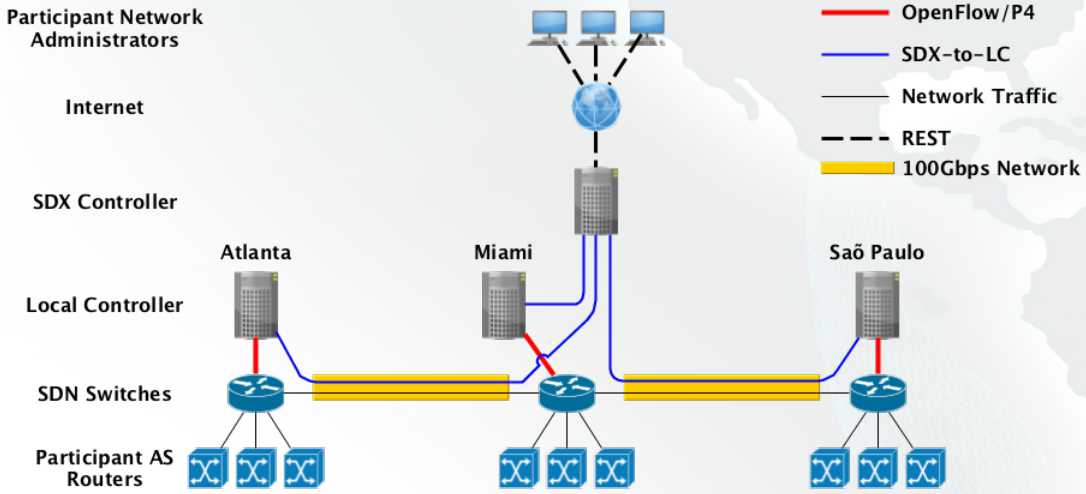
AtlanticWave/SDX

- Another SDX, but with a twist
  - Multiple, international locations
  - Multiple administrative domains
  - REN functionality in addition to SDX functionality
- Lots of telescope data
  - But what about during the day?
  - Have opportunity to do something more interesting

Agenda

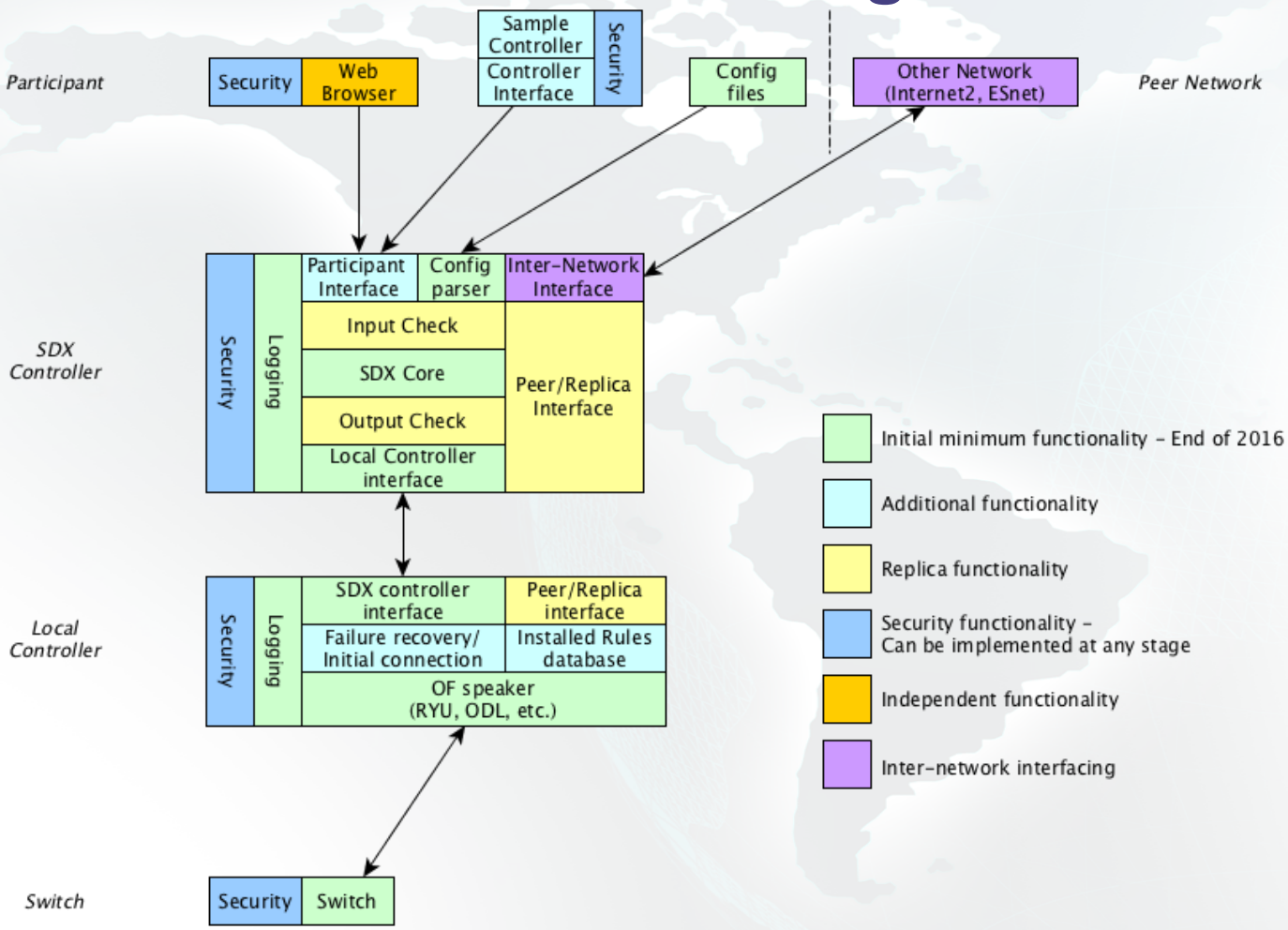
- Introduction
- Design Overview
- Functionality
- Challenges
  - Hardware
  - Abstractions
  - Security
  - Federation
  - Management
  - Sustainability
- Status

Overview



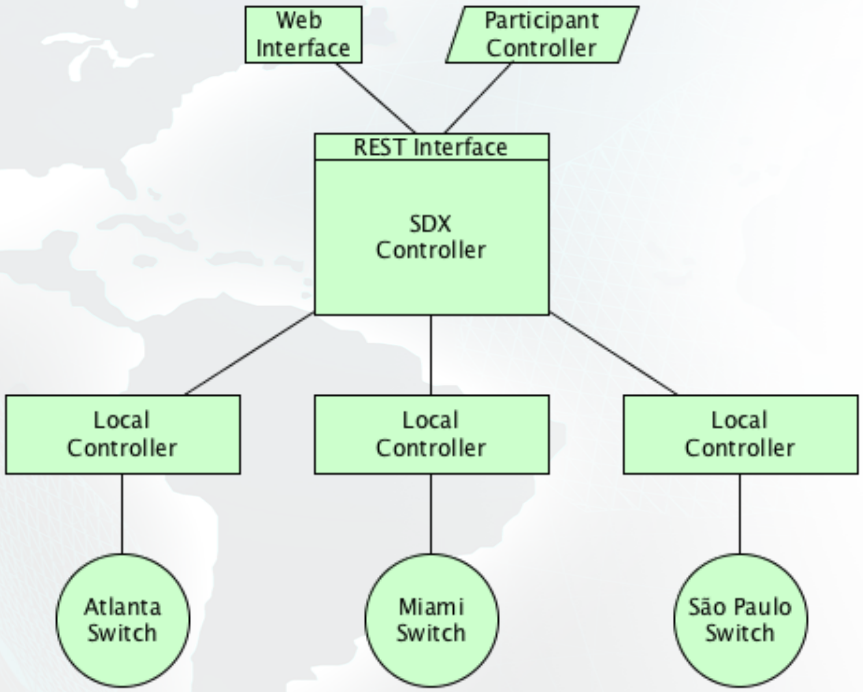
- Initially, three locations to cover
- Thousands of KM of fiber between each location
- Split controller design
  - Central controller for interacting with users
  - Local controllers at each location

Controller Design



Interfaces

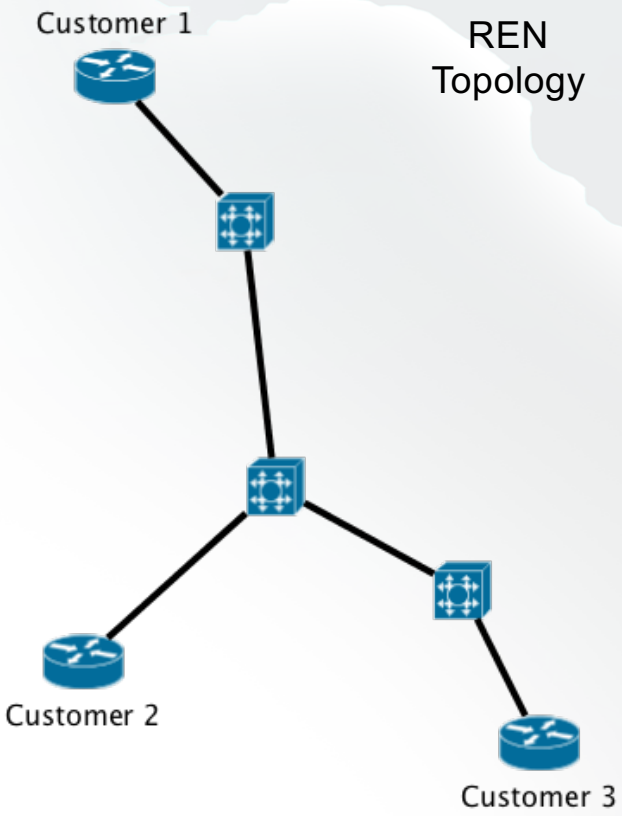
- REST API
- SDX-to-LC
- LC-to-Switch



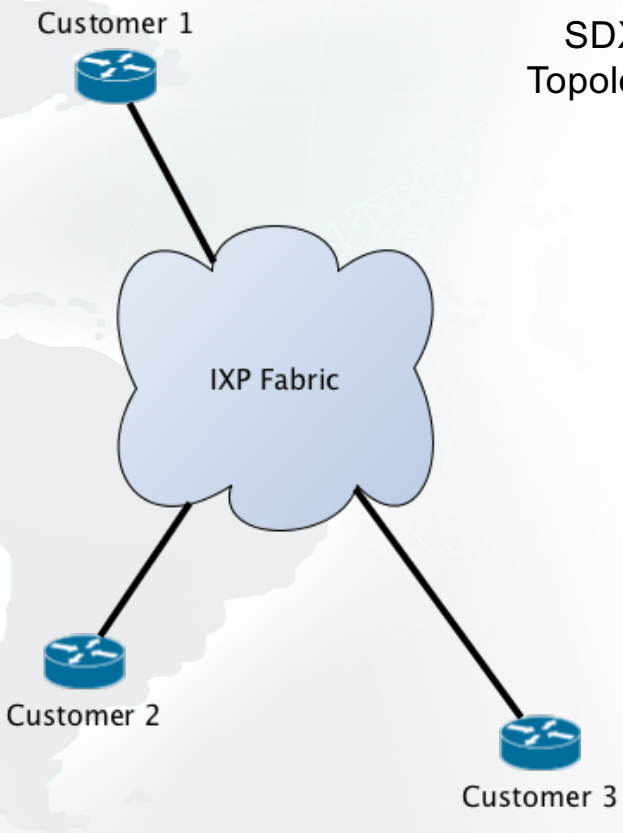
Agenda

- Introduction
- Design Overview
- Functionality
- Challenges
  - Hardware
  - Abstractions
  - Security
  - Federation
  - Management
  - Sustainability
- Status

Functionality



REN
Topology



SDX
Topology

Functionality

- Why not both?
- REN functionality will solve initial use case easily
  - Reserving bandwidth for specific durations
- SDX functionality can be used for unused bandwidth
  - Useful for impromptu transfers

Agenda

- Introduction
- Design Overview
- Functionality
- Challenges
  - Hardware
  - Abstractions
  - Security
  - Federation
  - Management
  - Sustainability
- Status

Challenges

- Like any system, it's complicated
  - But there are some rather unique challenges
- Some solved, but lots of open questions
  - We'd like operator and user help with some of these challenges
- What would *you* want?

Hardware

- We have some specific requirements
  - Multiple Table support
    - To reduce rule sizes dramatically
  - 100Gbps
    - Based on the data rates that we expect
  - Support for most, if not all of OpenFlow 1.3
    - Features in OpenFlow 1.3 that are useful
    - OF Groups, for instance

Need for Multiple Rule Tables

- Each participant has two types of rules
  - Inbound – rules for packets coming into the participant's network
    - 0.0.0.0/24 put on VLAN 3, forward to network
    - 128.0.0.0/24 put on VLAN 4, forward to network
  - Outbound – rules for packets leaving participant's network
    - Strip VLAN tag, forward to neighbor

Cross Multiplication

	A-in	B-in	C-in
A-out			
B-out			
C-out			

Cross Multiplication

	A-in	B-in	C-in
A-out	A-in*A-out	B-in*A-out	C-in*A-out
B-out	A-in*B-out	B-in*B-out	C-in*B-out
C-out	A-in*C-out	B-in*C-out	C-in*C-out

- $O(N^2)$ sets of rules
- Some optimizations are possible
  - The diagonal can be eliminated
  - Gupta, et. al., 2014 discusses other optimizations

Cross Multiplication

	A-in	B-in	C-in
A-out		B-in*A-out	C-in*A-out
B-out	A-in*B-out		C-in*B-out
C-out	A-in*C-out	B-in*C-out	

- $O(N^2)$ sets of rules
- Some optimizations are possible
  - The diagonal can be eliminated
  - Gupta, et. al., 2014 discusses other optimizations

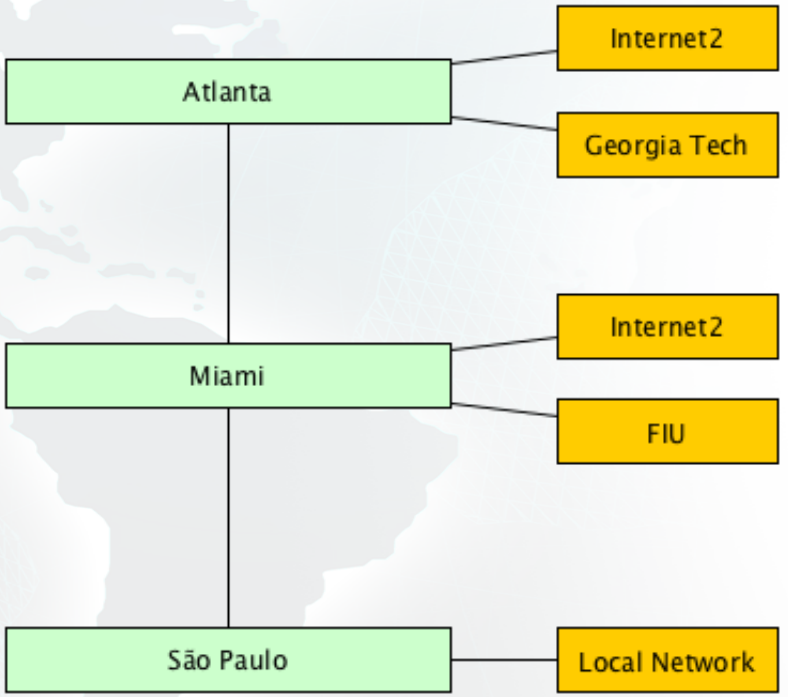
Multiple tables are better

Table 1	Table 2
A-out	A-in
B-out	B-in
C-out	C-in

- With multiple tables, we can pipeline the outbound and inbound rules
- O(2N) sets of rules
  - Much better than $O(N^2)$
- Think of a dozen participants:
  - ~144 sets of rules vs ~24 sets
- Much simpler to implement

100Gbps OpenFlow Equipment is Hard to Find

- Only a few manufacturers have OF 100Gbps gear and big interface buffers
- A lot only have 1 or 2 ports, need 3 or 4, depending on location



OpenFlow 1.3 (non) Support

- Many vendors claim 1.3 support
  - Often single table
  - Only rules X and Y, but not Z
  - Limited number of rules
    - TCAM limitations
- Study about support being overblown
  - Di Lallo et al., IEEE/IFIP NOMS 2016

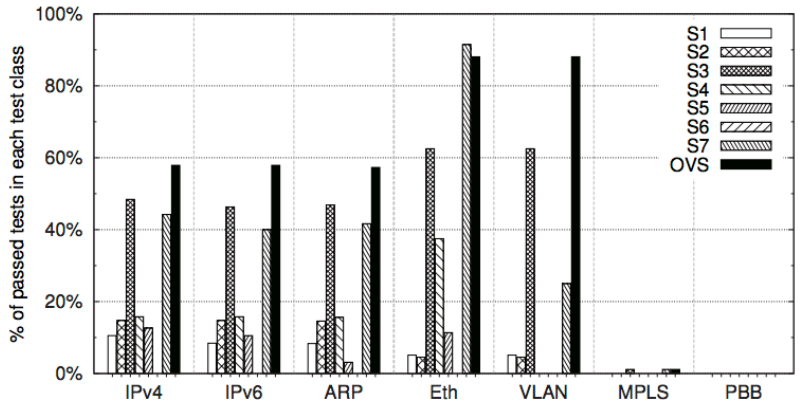
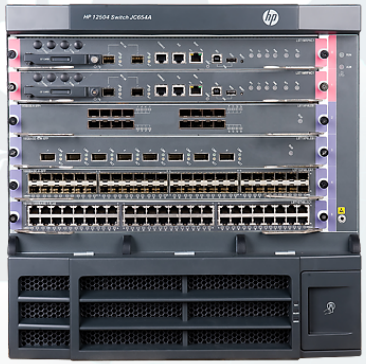


Fig. 5. Percentage of passed OF 1.3 Ryu tests for packets carrying specific protocols.

100Gbps + OpenFlow 1.3

- Rather hard to find!
- Equipment's now trickling out



http://noviflow.com/products/noviswitch/
http://www8.hp.com/us/en/products/networking-switches/product-detail.html?oid=4177453
http://www.corsa.com/products/dp6440/
http://www.brocade.com/en/backend-content/pdf-page.html?/content/dam/common/documents/content-types/datasheet/brocade-mlx-2x100gbe-cfp2-ds.pdf

Abstractions

- What functionality do people need?
  - Point-to-point paths?
  - Point-to-multipoint?
  - Arbitrary routing?
- What should the API look like?
  - REST good enough?
  - Web-based interface?
- Who should it be tailored to?
  - Network admins?
  - Domain scientists?
  - General users?

APIs for Different Audiences

- Administrators

```
{"l2tunnel":{
    "starttime":"2016-10-12T23:20:50",
    "endtime":"2016-10-13T23:20:50",
    "srcswitch":"atl-switch",
    "dstswitch":"mia-switch",
    "srcport":5,
    "dstport":7,
    "srcvlan":1492,
    "dstvlan":1789,
    "bandwidth":1}}
```

- Domain scientists

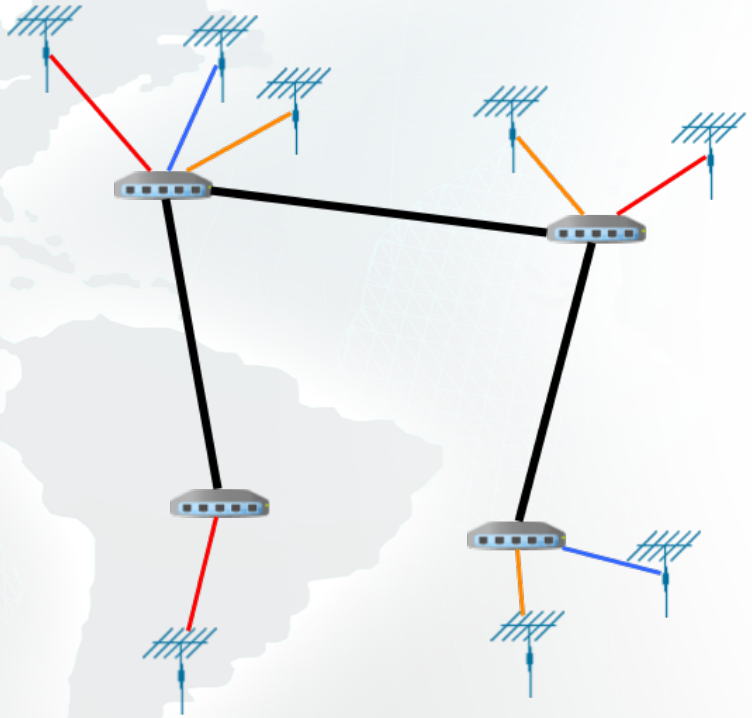
```
{"dtntunnel":{
    "starttime":"2016-10-27T17:00:00",
    "endtime":"2016-10-30T23:59:59",
    "srcdtn":"gt-dtn",
    "dstdtn":"fiu-dtn",
    "bandwidth":1}}
```

What Functionality Would be Useful?

- NSI-like interface planned
  - Partially working now
    - Timers/Bandwidth aren't yet implemented
    - Come see our demo at GLIF!
  - With inter-network NSI integration in the future
- SDX rules based on DNS
  - Based on NetAssay
  - match(domain='example.com')
- Any suggestions?

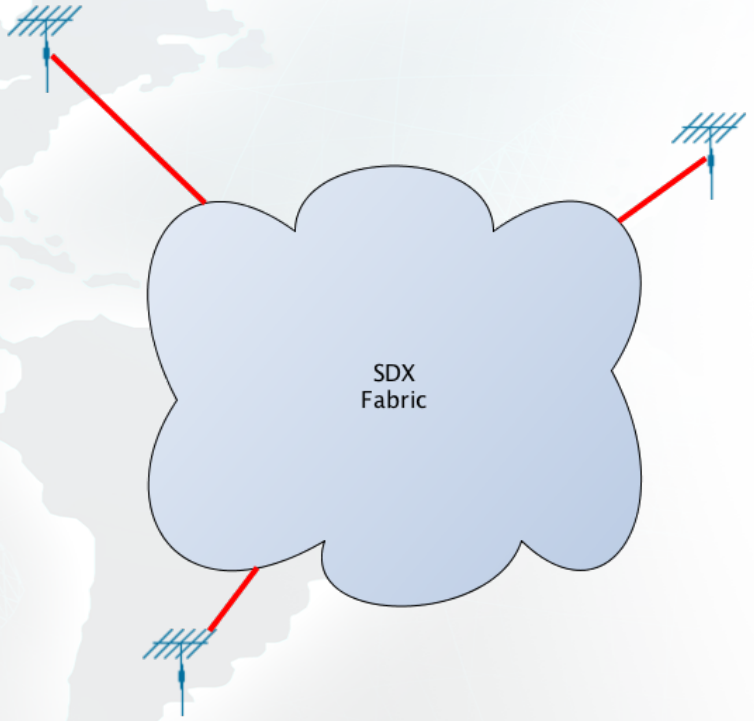
Deployment Outside of AtlanticWave/SDX

- Example deployment
  - In a city with a distributed SDX, like AMS-IX
  - Mobile phone backbone for multiple carriers
- Does this change what sorts of abstractions someone would want?



Deployment Outside of AtlanticWave/SDX

- Example deployment
  - In a city with a distributed SDX, like AMS-IX
  - Mobile phone backbone for multiple carriers
- Does this change what sorts of abstractions someone would want?



Do Administrators Care about Functionality Beyond BGP?

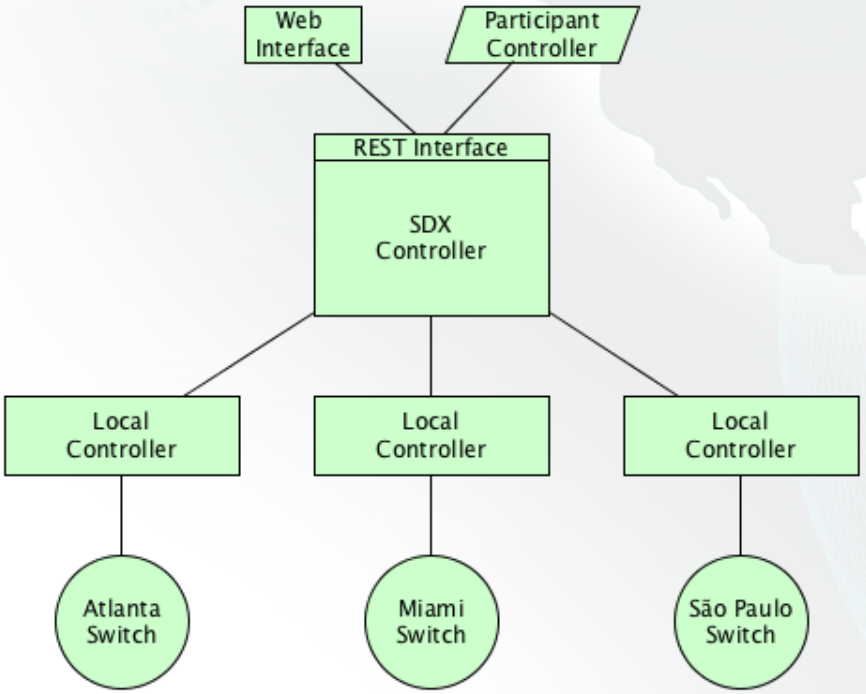
- Application-based peering
  - YouTube through Level3
  - Netflix through Cogent
  - Everything else through AT&T
  - Impossible with BGP
- Shared services at the SDX
  - Shared IDS for small businesses connection to the SDX
  - Web caching at the SDX

- Would administrators be interested in this type of functionality?

Security

- SDN and Security isn't discussed nearly enough
  - Most academic work glosses over security aspects of what they developed
  - New attacks are possible due to the design change over traditional networking
- This is being deployed
  - So we care a lot about security

Security Issues in AtlanticWave/SDX Design



- Information leakage
  - Rules/data leaking to unauthorized users
- DoS attacks
  - REST API is susceptible
  - In-band SDX-to-LC should mitigate
- Policy overlap
  - New user policies must not violate other user's policies

Authentication

- User authentication
  - TLS certificate authentication
  - Would an SSH tunnel with a certificate be enough?
- Local controller and SDX controller
  - Prevent unauthorized rules coming from a fake SDX controller
  - Prevent snooping from a fake local controller
  - Bi-directional TLS authentication with certificates

Authorization

	Admins	Domain Scientists	Data Agent	Research Assistant
GT				
FIU				
NCSA				
UofA				

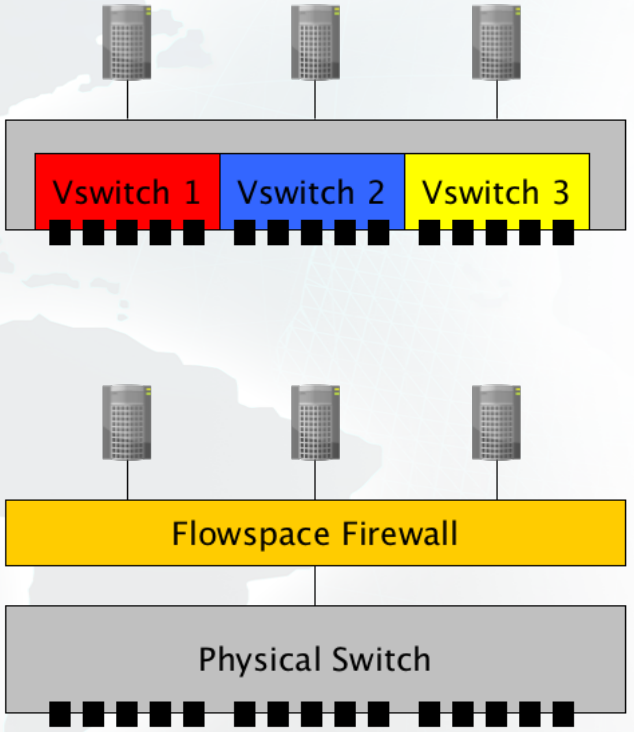
- What's the correct level of granularity in authorization?
  - Roles
  - Organizations
- What Actions should be authorized?
  - At what granularity should actions be authorized?
- Future project

Actions requiring authorization

- Installing rules
  - Per port
  - Per switch
- Removing rules
  - Own rules
  - Same org. rules
- Get Statistics
  - To authorize automated collection methods
- View Rules
  - Per user
  - Per organization
  - Per switch

Federation

- Multiple Controllers with a Single Switch
  - Hardware virtualization
    - Per port, typically
    - New switches allow for per VLAN
  - Software Hypervisor
    - Use something like FlowSpace Firewall
      - Below the LC, for AtlanticWave/SDX
      - FSF does *not* support OF1.3

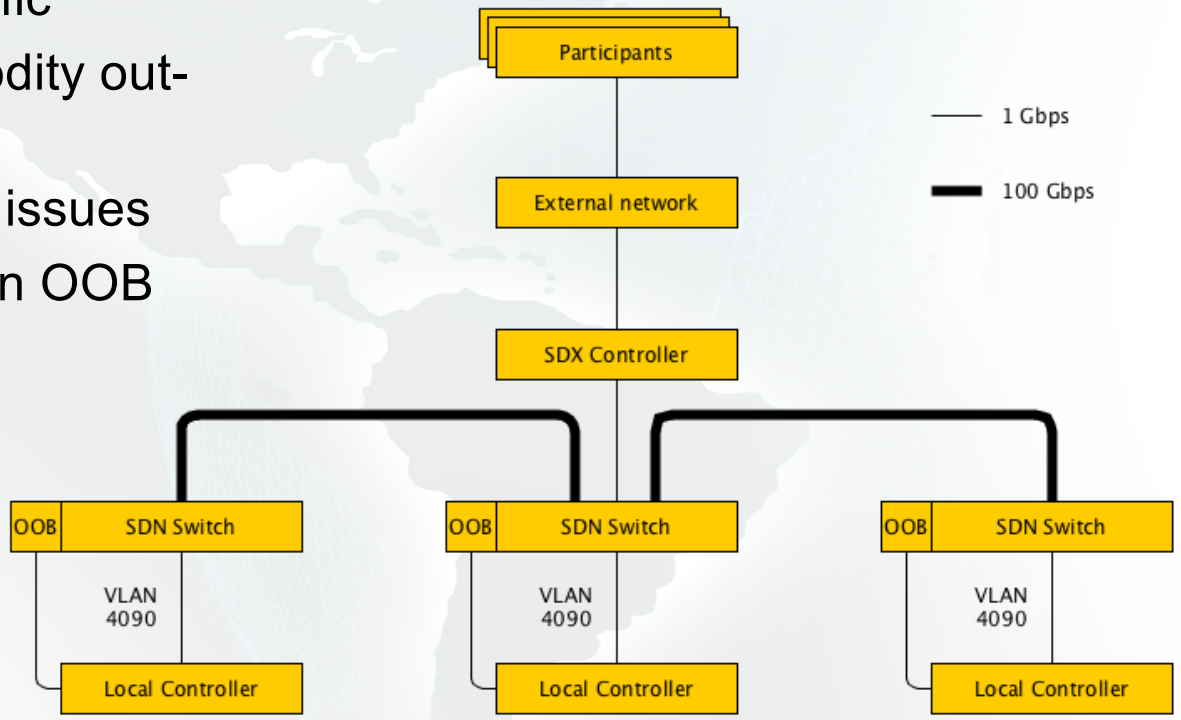


Federation

- Integrating other Networks
  - Integration with NSI
    - There are a number of NSI speakers that could be used to integrate with AtlanticWave/SDX
  - Shibboleth connectivity
    - Difficulty of integration is not yet known
    - Would certificate authentication be better?

Management

- In-band management traffic
- Known delays vs. commodity out-of-band connection
- Helps with some security issues
- Switches still controlled on OOB port
- LC bootstraps switches



Management

- Failover
  - Distance = Latency
  - Latency = Problems
  - AtlanticWave/SDX is not a physically small network
  - Should there be more autonomy at the LC for failover?

	Atlanta	Miami	São Paulo
Atlanta	-	13ms	119ms
Miami	81 MB	-	106ms
São Paulo	743 MB	662 MB	-

https://wondernetwork.com/pings, FIU/AmLight

Sustainability

- Currently supported by NSF Grant #ACI-1341024 2015-2020
- How to make this self sufficient/sustainable?
- What's a good business model?
- Other research networks are facing the same question (e.g., GENI)

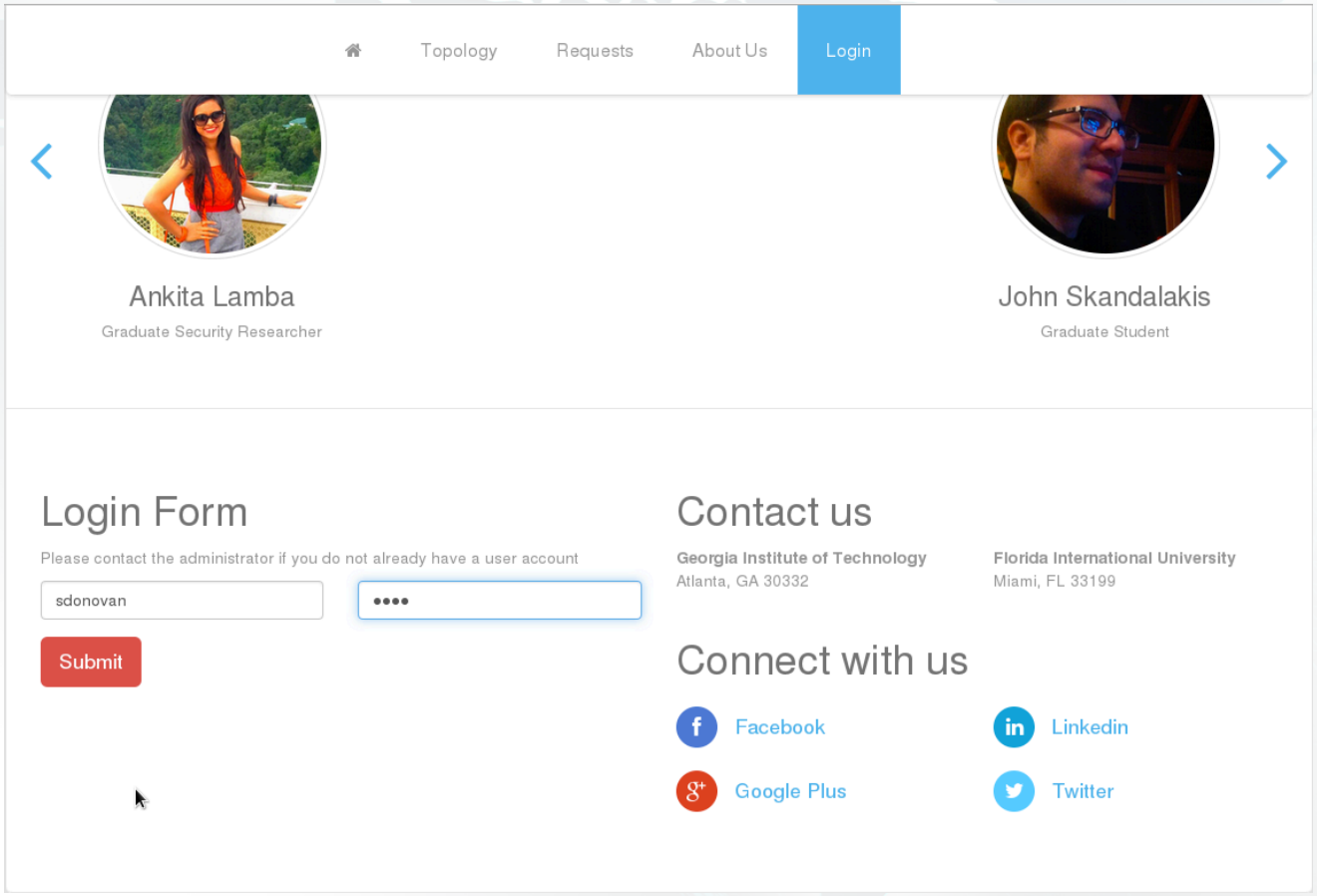
Agenda

- Introduction
- Design Overview
- Functionality
- Challenges
  - Hardware
  - Abstractions
  - Security
  - Federation
  - Management
  - Sustainability
- Status

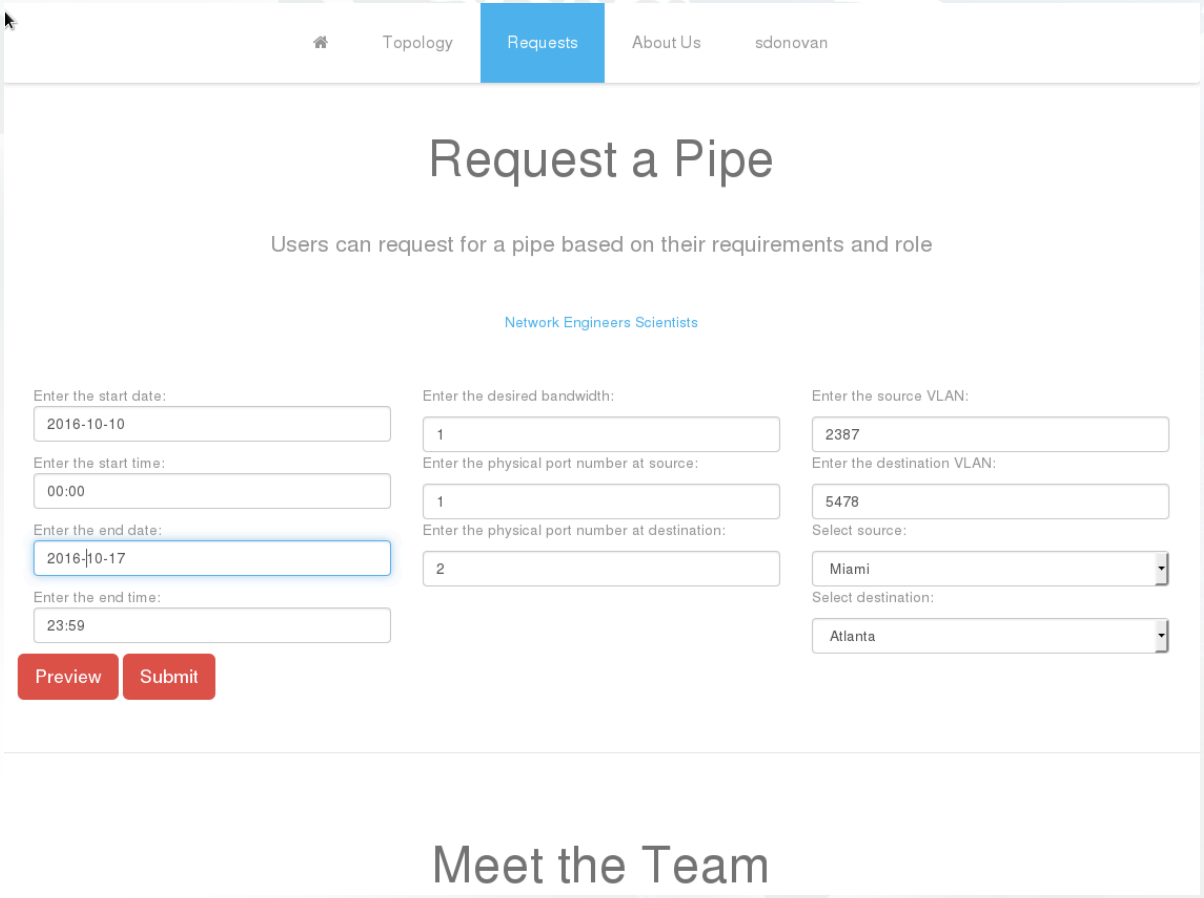
Current Status

- Focusing on NSI-like functionality right now
  - Default IXP behavior will follow
- Initial version of the controller is built
  - Has limitations, but being continuously developed
- Prototype Web Interface
  - Limited to adding rules
- Configuration files for static configurations
  - Users and topology are static at startup

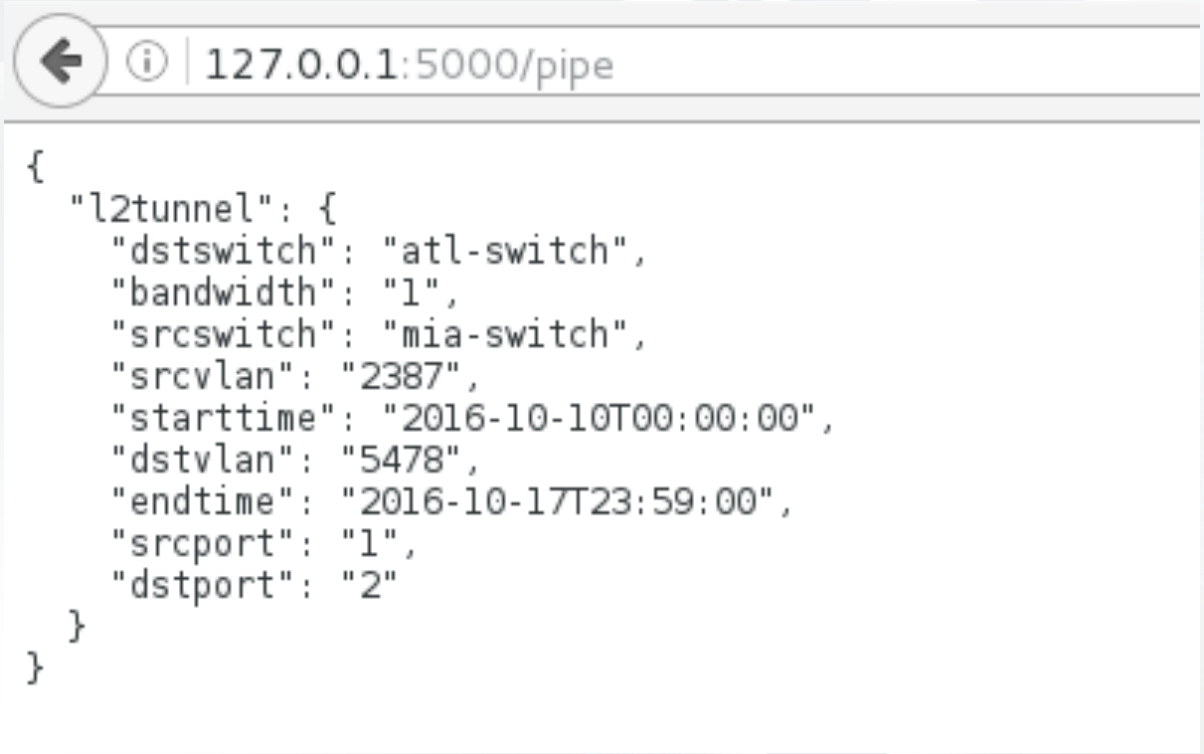
Web Interface



Web Interface



Web Interface



Timeline

- Public Github accessible after this meeting
  - https://github.com/sdonovan1985/atlanticwave-proto
- October for NSI/AL2S-like functionality completed
  - Missing timers and bandwidth reservation as of today
- October for DTN-to-DTN for domain scientists
- November for running on hardware switches
- December for initial SDX functionality

Demo at GLIF

Come to demo night at GLIF

September 29, 6pm

Thanks!

http://www.atlanticwave-sdx.net/

Sean Donovan

sdonovan@gatech.edu

Russ Clark

russ.clark@gatech.edu

Jeronimo Bezerra

jbezerra@fiu.edu

References

- Stringer, Jonathan Philip, et al. "Cardigan: Deploying a distributed routing fabric." *Proceedings of the second ACM SIGCOMM workshop on Hot topics in software defined networking*. ACM, 2013.
- Stringer, Jonathan, et al. "Cardigan: SDN distributed routing fabric going live at an Internet exchange." *2014 IEEE Symposium on Computers and Communications (ISCC)*. IEEE, 2014.
- Gupta, Arpit, et al. "SDX: a software defined internet exchange." *ACM SIGCOMM Computer Communication Review* 44.4 (2015): 551-562.
- Gupta, Arpit, et al. "An industrial-scale software defined internet exchange point." *13th USENIX Symposium on Networked Systems Design and Implementation (NSDI 16)*. 2016.
- Chung, Joaquin, Henry Owen, and Russell Clark. "SDX architectures: A qualitative analysis." *SoutheastCon 2016*. IEEE, 2016.
- di Lallo, Roberto, et al. "On the practical applicability of SDN research." *NOMS 2016-2016 IEEE/IFIP Network Operations and Management Symposium*. IEEE, 2016.