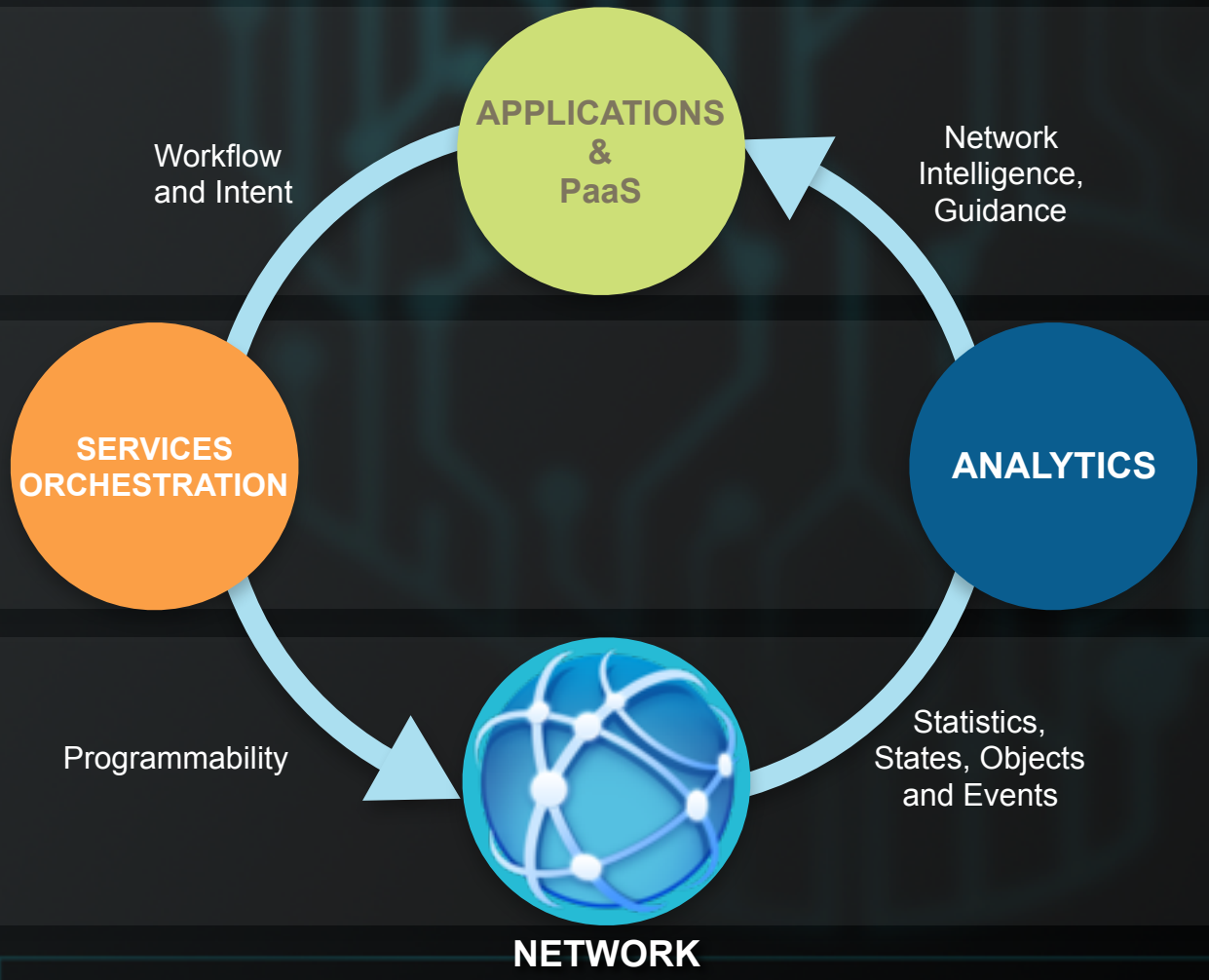
REACTIVE NETWORKING

All reactions require entropy = OSS

David Ward | Chief Architect & CTO of Engineering
dward@cisco.com

SDN + DATA DRIVEN NETWORKING = REACTIVE NETWORKING

| Turning data into information to drive intent

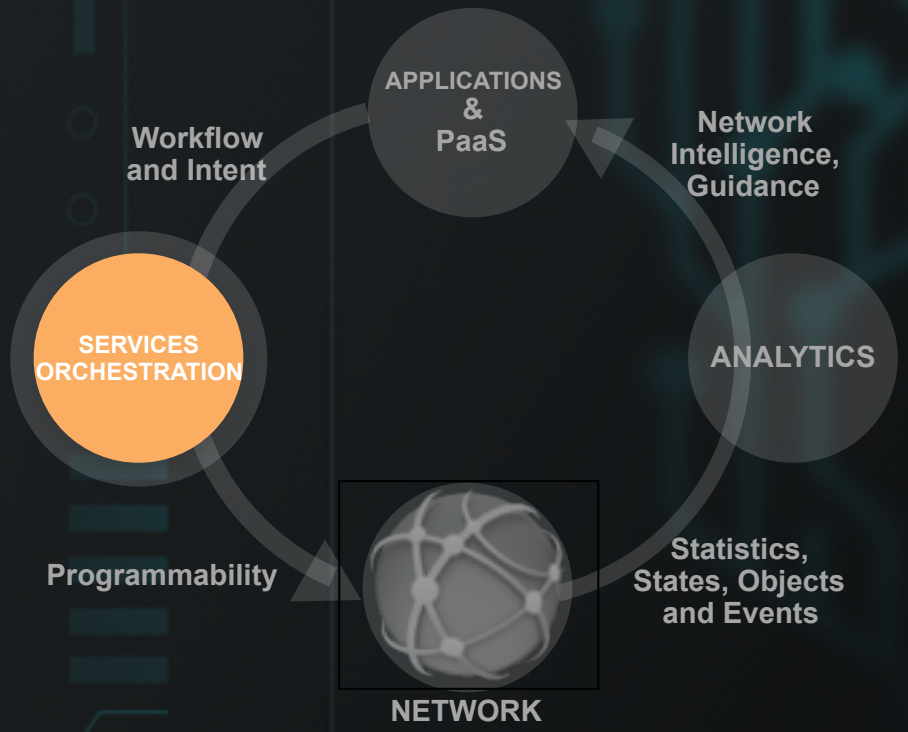


Application Software
New Businesses
SaaS + Integration
Operations BI

Infrastructure Software
Management
Orchestration
Analytics, Controllers

Embedded Software
Route, Switch, Appliance
IOS, XR, NXOS, others…

© 2016 Cisco Systems, Inc. All rights reserved.



Services Orchestration Platform

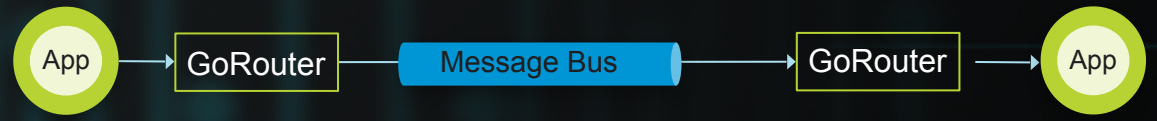
Incorporating model and driven orchestration along with SDN controller functionality to realize policy intent.

Providing complete network services life-cycle management and VNF life cycle management to implement both simple and complex service topologies that span the network and data center domains.

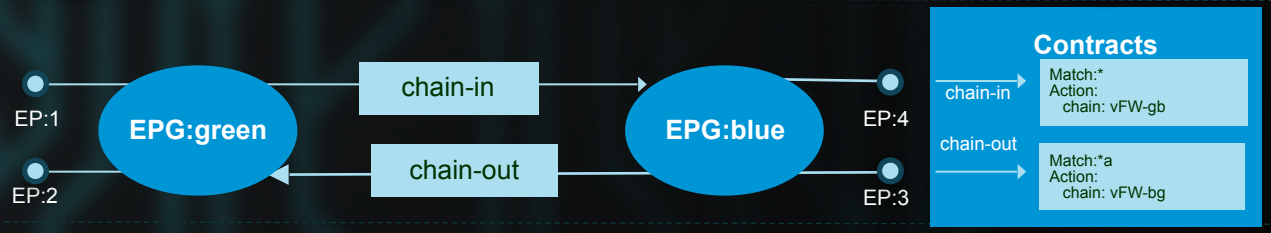
© 2016 Cisco Systems, Inc. All rights reserved.

ORCHESTRATION LAYERS VIEW

Application View

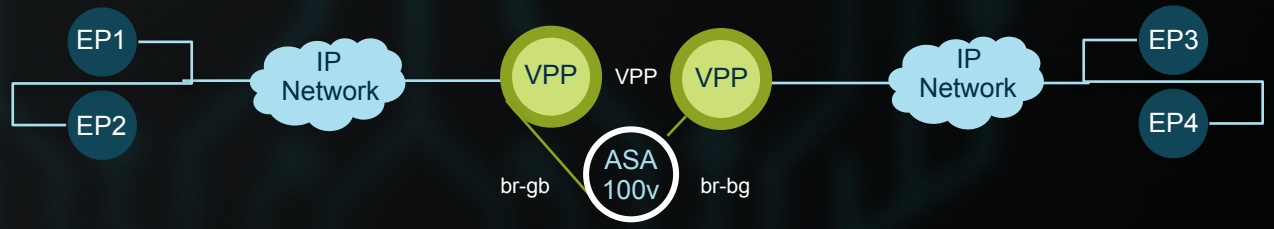


Policy View

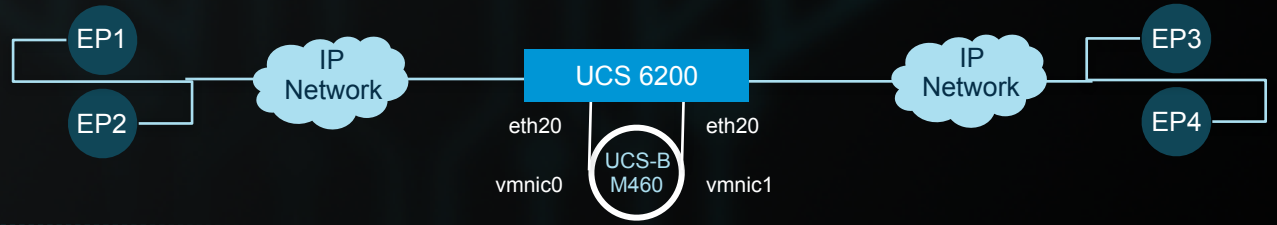


Service View

Virtual Topology



Physical Topology



Resource View

© 2016 Cisco Systems, Inc. All rights reserved.

SERVICES ORCHESTRATION PLATFORM

Service Provisioning
Workload Placement
Service Configuration

VM Policy
Service Monitoring
Auto Healing
Elastic Scaling

Network Controller
Group Policy Service
Chaining Physical and
Virtual Network control

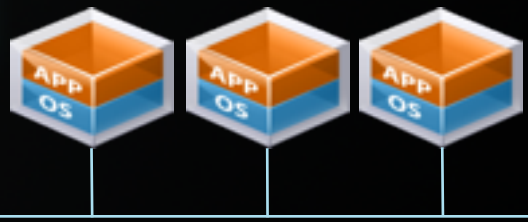
Service Model | Service Model | Service Model

Service Life Cycle Manager
Network Services Orchestrator

Virtual Machine Life Cycle Manager
Elastic Services Controller

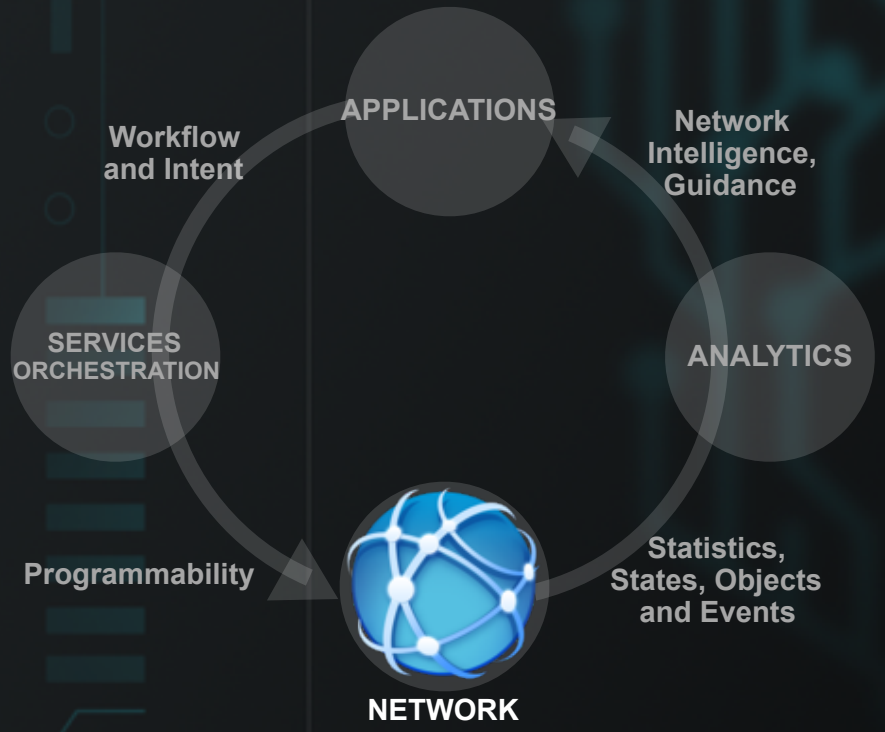
OPEN DAYLIGHT [f d.io] Fast Data forwarder

VM Policy
Network Policy
Service Provisioning
Service Configuration
Service Chaining
Service Monitoring
Auto Recovery
Elastic Scaling
Workload Placement
Service Assurance



Hypervisor

Compute | Network | Storage



NETWORK

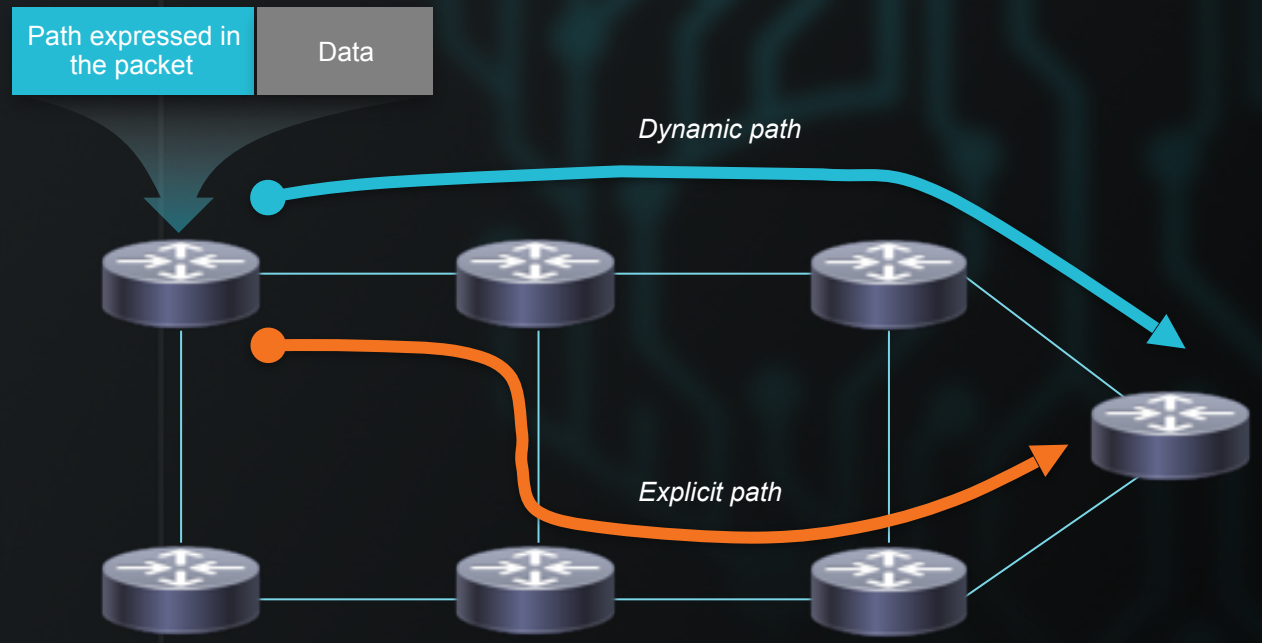
Focus on simplification, scale, improving management and operations, performance and cost optimization

Whilst creating the scalable foundation for services across Edge, WAN and Data Center

© 2016 Cisco Systems, Inc. All rights reserved.

SEGMENT ROUTING: REMOVING PROTOCOLS

Simplicity



Control Plane

Routing protocols with extensions (IS-IS,OSPF, BGP)	SDN controller

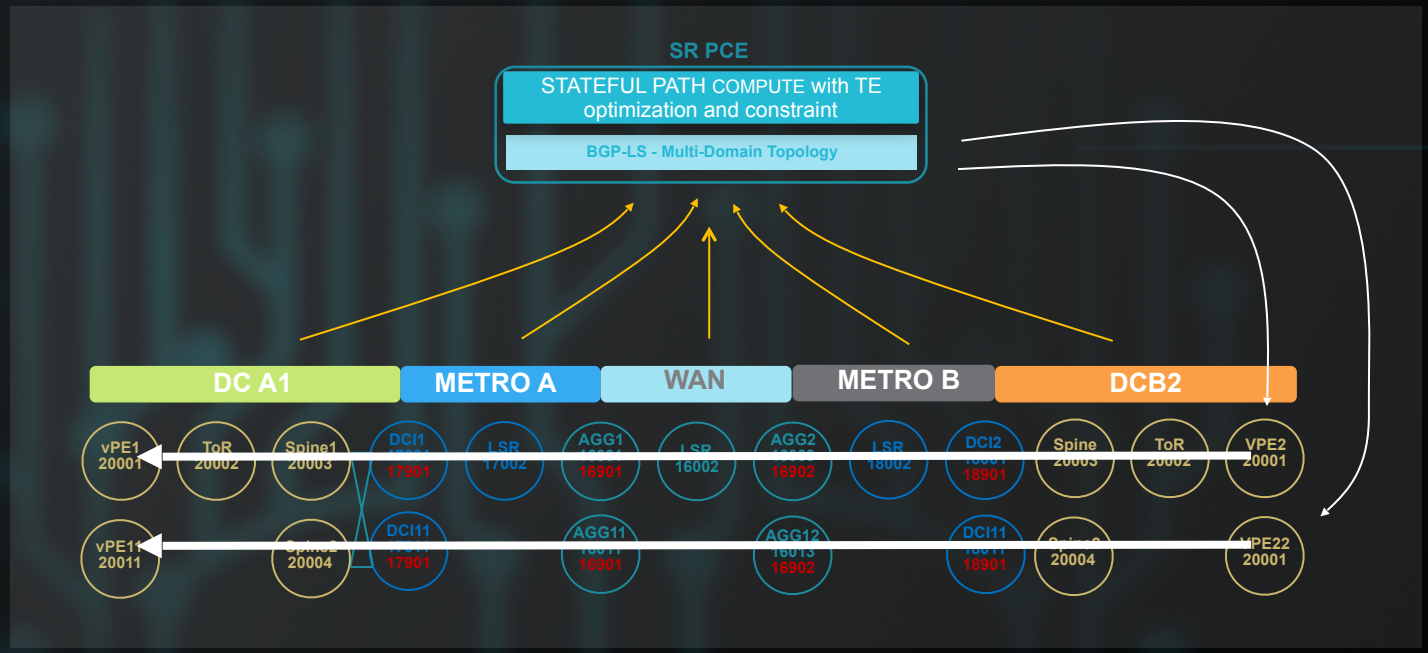
Data Plane

MPLS (segment labels)	IPv6 (+SR header)

Paths Options

Dynamic (STP computation)	Explicit (expressed in the packet)

SR PCE



- Multi-domain topology
  - Realtime reactive feed via BGP-LS/ISIS/OSPF from multiple domains
  - Including IP address and SID
- Multi-domain path compute with TE optimization and constraint
  - Latency, Disjoint paths, Inclusion/Exclusion (affinity, SRLG, address)
  - Bandwidth

© 2016 Cisco Systems, Inc. All rights reserved.

BIER: REMOVING REPLICATION STATE

- BIER – Bit Indexed Explicit Replication
  - Next Gen Stateless Multicast
  - No multicast control protocols
    - No PIM, MLDP, P2MP RSVP
  - Deterministic fast convergence
  - Simple to deploy and operate
  - OPEX inline with Unicast

- BIERMan
  - ODL BIER Application
  - Central control of multicast flows
  - Map multicast channels to app-defined paths.
  - Open-source

Discover Topology

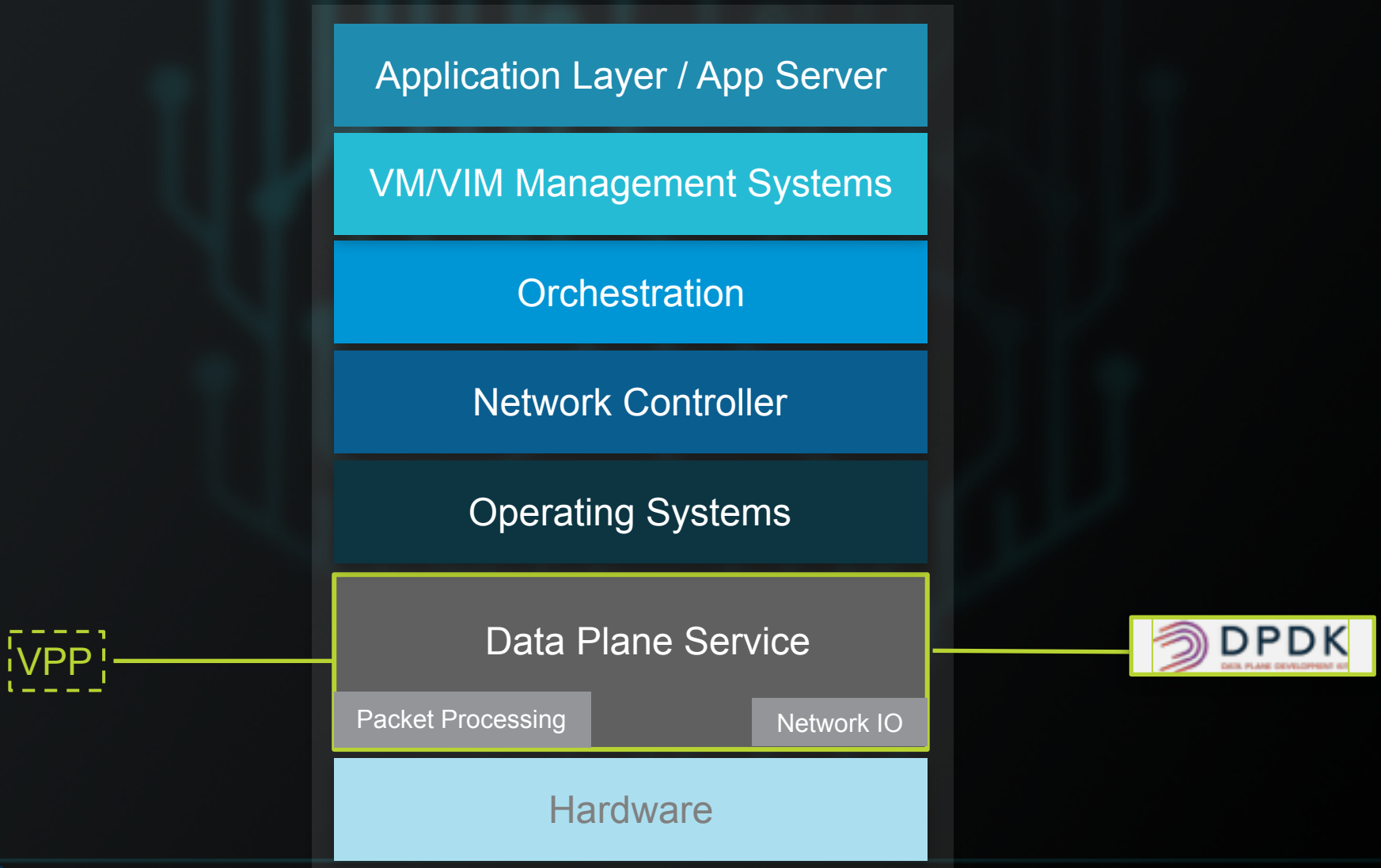
Define MP Path

Assign Channel

Deploy!

© 2016 Cisco Systems, Inc. All rights reserved.

VPP IN THE OVERALL STACK = SCALE



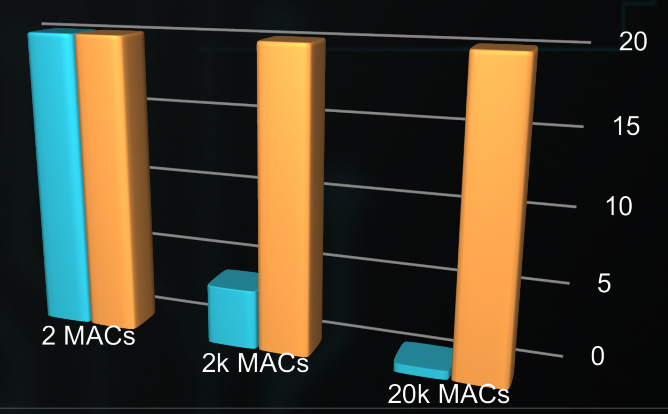
© 2016 Cisco Systems, Inc. All rights reserved.

VPP technology in a nutshell

- Throughput NDR = “No Drop Rate”
- “No Drop Rate” = Read zero frame loss
  - And not “ITU 10^-7 packet loss ratio”
- Simply put NDR here is better than ”ITU NDR” ☺

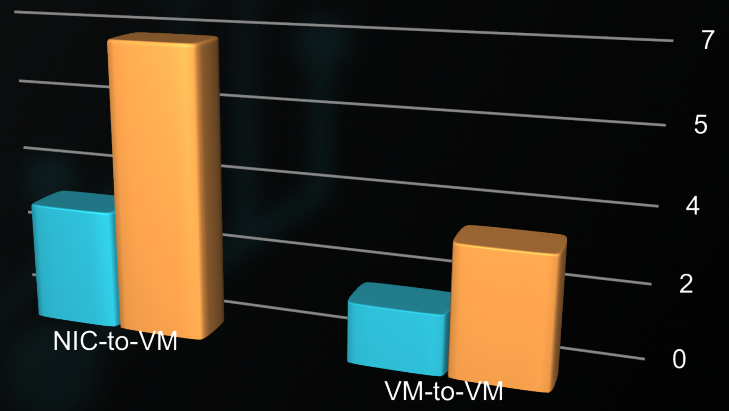
| NDR rates for 2p10GE, 1 core, L2 NIC-to-NIC

[IMIX Gbps]



| NDR rates for 2p10GE, 1 core, L2 NIC-to-VM/VM-to-VM

[IMIX Gbps]



- Virtual network infrastructure – benchmarking the efficiency
- All tests with minimal connections only, single CPU core
- Higher performance with more connections, more CPU cores
- Latest SW at the time: OVSDPDK 2.4.0, VPP 09/2015

© 2016 Cisco and/or its affiliates. All rights reserved.

VNET-SLA BENCHMARKING AT SCALE: IPV6

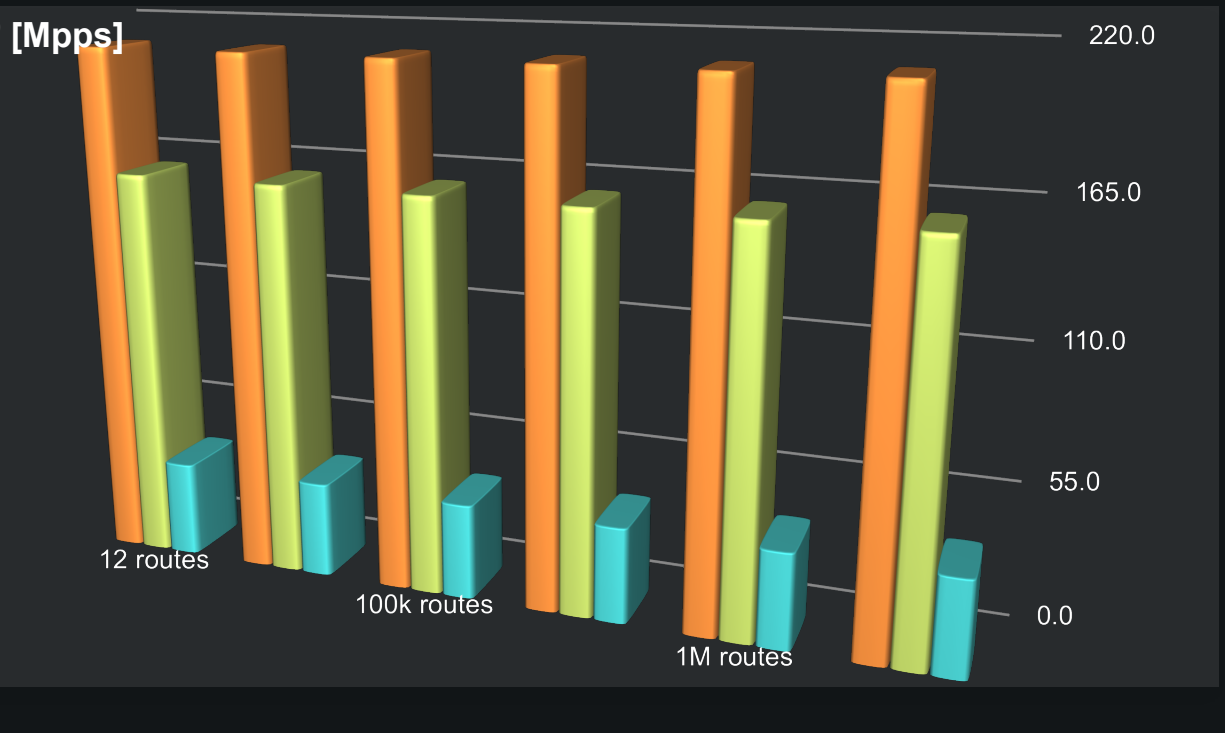
VPP-based vSwitch

Phy-VS-Phy

Zero-packet-loss Throughput for 12 port 40GE, 24 cores, IPv6

- FD.io VPP data plane throughput not impacted by large size of IPv6 FIB
- VPP tested on UCS 4-CPU-socket server with 4 of Intel “Haswell" x86-64 processors E7-8890v3 18C 2.5GHz
- 24 Cores used – Another 48 cores can be used for other network services!

Zero-packet-loss Throughput for 12 port 40GE, 24 cores, IPv6



VPP vSwitch IPv4 routed forwarding
FIB with 2 milion IPv6 entries
12x40GE (480GE) 64B frames

200Mpps zero frame loss

NIC and PCIe is the limit not VPP

VPP vSwitch IPv4 routed forwarding
FIB with 2 milion IPv6 entries

12x40GE (480GE) IMIX frames

480Gbps zero frame loss

“Sky” is the limit not VPP

VNET-SLA BENCHMARKING AT SCALE: L2 MAC

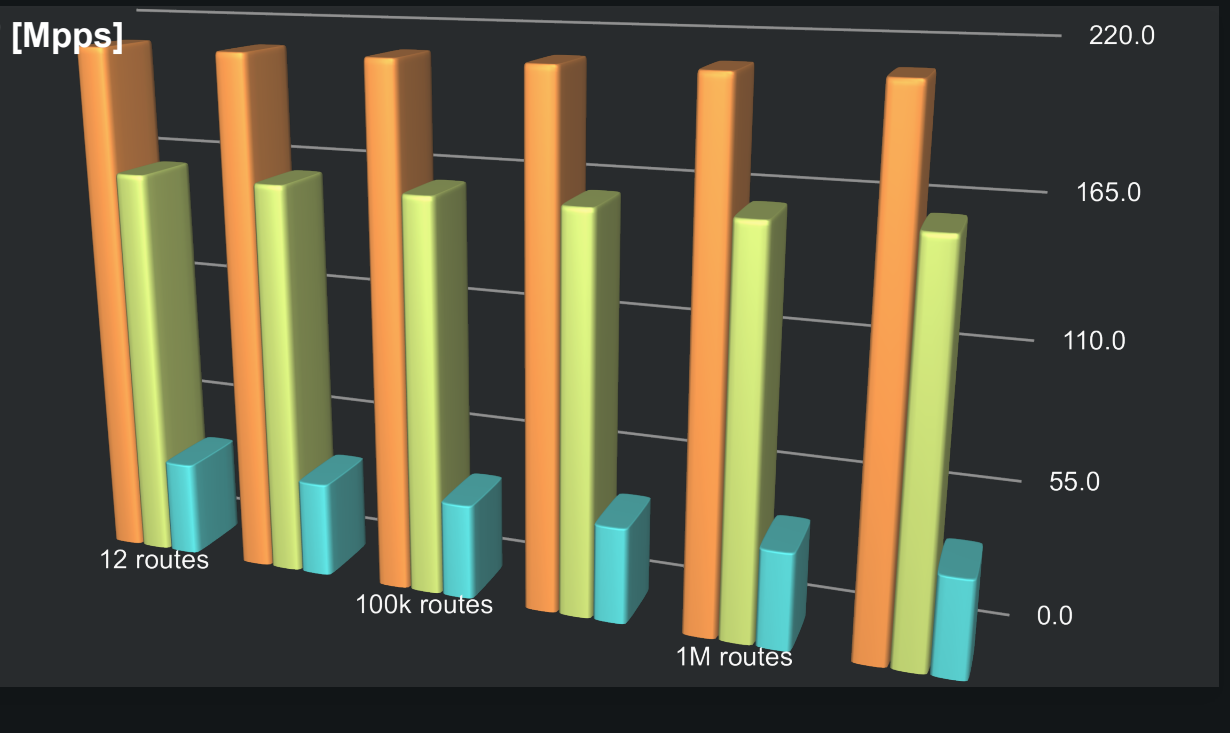
VPP-based vSwitch

Phy-VS-Phy

Zero-packet-loss Throughput for 12 port 40GE, 24 cores L2 MAC

- FD.io VPP data plane throughput not impacted by large L2 MAC FIB size
- VPP tested on UCS 4-CPU-socket server with 4 of Intel “Haswell" x86-64 processors E7-8890v3 18C 2.5GHz
- 24 Cores used – Another 48 cores can be used for other network services!

Zero-packet-loss Throughput for 12 port 40GE, 24 cores, L2 MAC



VPP vSwitch IPv4 routed forwarding FIB with 2 million L2 MAC entries 12x40GE (480GE) 64B frames

200Mpps zero frame loss

NIC and PCIe is the limit not VPP

VPP vSwitch IPv4 routed forwarding FIB with 2 million L2 MAC entries

12x40GE (480GE) IMIX frames

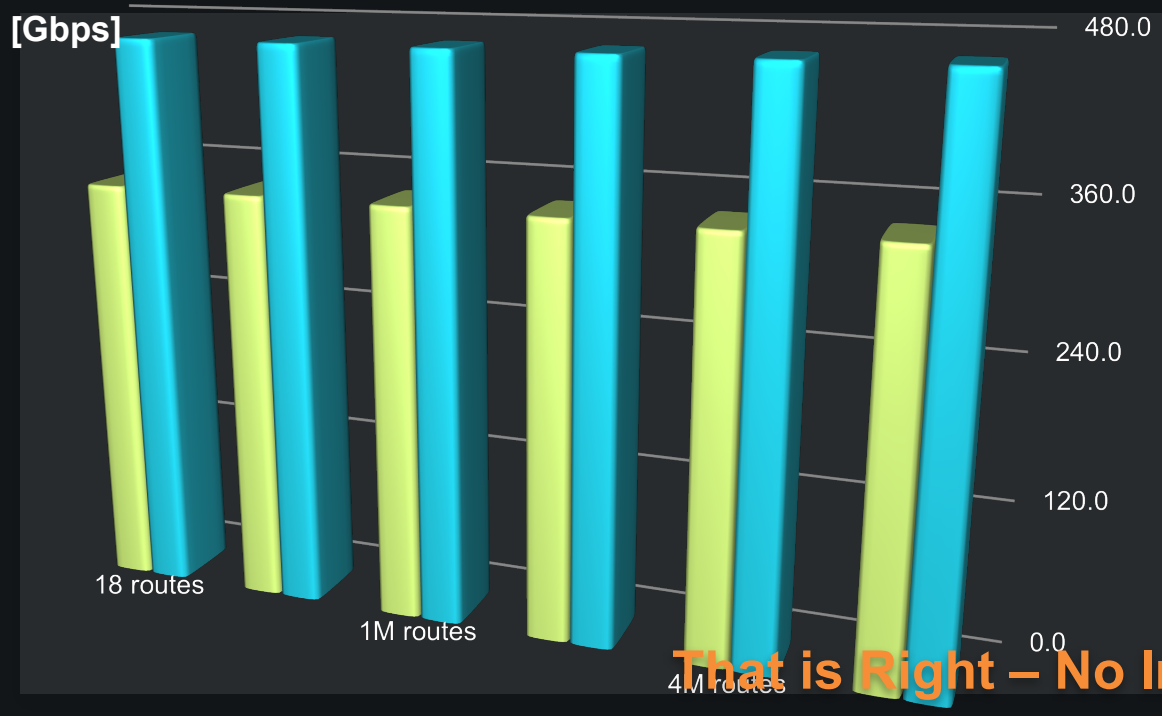
480Gbps zero frame loss

“Sky” is the limit not VPP

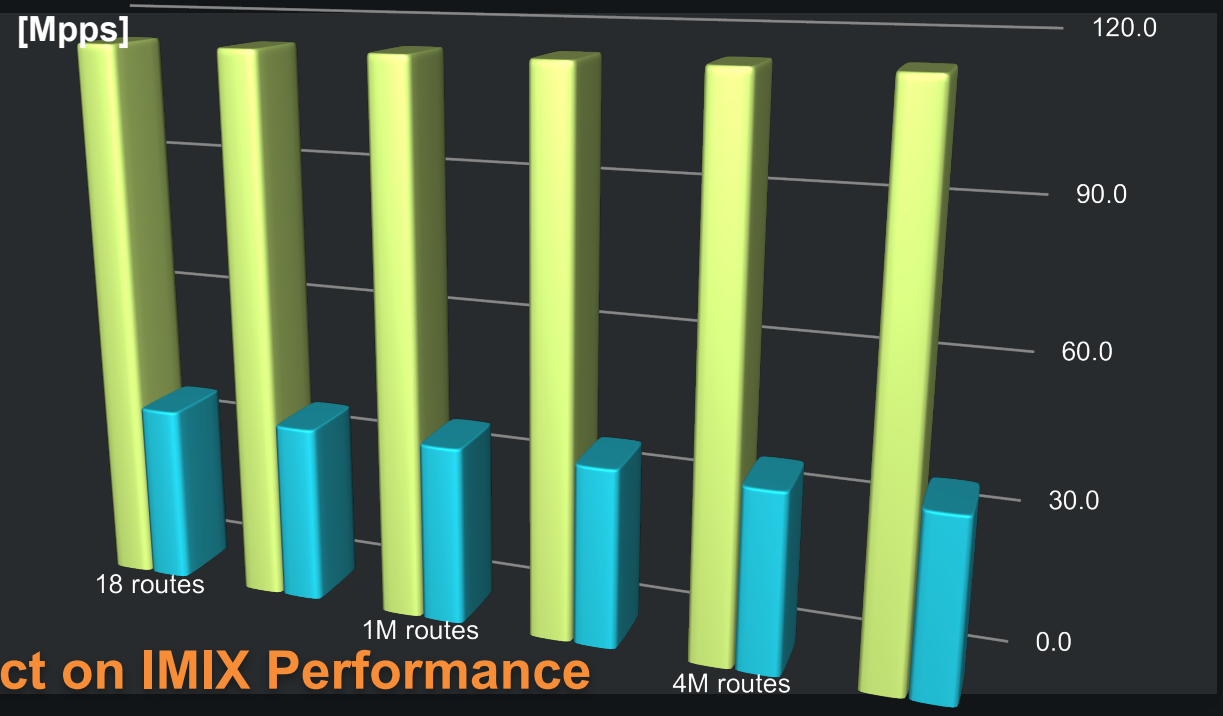
VNET-SLA BENCHMARKING AT SCALE: IPV4 + SECURITY

VPP-based vSwitch

Zero-packet-loss Throughput for 18 port 40GE, 36 cores, IPv4



Zero-packet-loss Throughput for 18 port 40GE, 36 cores, IPv4



That is Right – No Impact on IMIX Performance

- FD.io VPP data plane throughput not impacted by large IPv4 FIB size, and stateless security input access lists and white-lists & black-lists*
- VPP tested on UCS 4-CPU-socket server with 4 of Intel “Haswell" x86-64 processors E7-8890v3 18C 2.5GHz
- **36 Cores used – Another 36cores can be used for other network services!**

VPP Cores Not Completely Busy
VPP Vectors Have Space For More Services and More Packets!!
PCIe 3.0 and NICs Are The Limit

VPP vSwitch IPv4 routed forwarding FIB with 2M IPv4 entries, 2k white-list* 18x40GE IMIX frame sequence

342Gbps zero frame loss

Sky” is the limit not VPP

VPP vSwitch IPv4 routed forwarding FIB with 2M IPv4 entries, 2k white-list* 18x40GE 1518B frames

462Gbps zero frame loss

Pcle 3.0 is the limit not VPP

* Test results extrapolated from 120GE 2-Socket machine. See next slides for why.

VPP Cores Not Completely Busy
VPP Vectors Have Space For More Services and More Packets!!
PCIe 3.0 and NICs Are The Limit

And How Do We Know This?
Simples – A Well Engineered Telemetry
In Linux and VPP Tells Us So

```
========
TC5     120ge.vpp.24t24pc.ip4.cop
TC5.0     120ge.2pnic.6nic.rss2.vpp.24t24pc.ip4.cop
d. testcase-vpp-ip4-cop-scale
         120ge.2pnic.6nic.rss2.vpp.24t24pc.ip4.2m.cop.2.copip4dst.2k.match.100
            64B, 138.000Mpps, 92,736Gbps
            IMIX, 40.124832Mpps, 120.000Gbps
            1518, 9.752925Mpps, 120.000Gbps
            ---------------
            Thread 1 vpp_wk_0 (lcore 2)
            Time 45.1, average vectors/node 23.44, last 128 main loops 1.44 per node 23.00
              vector rates in 4.6791e6, out 4.6791e6, drop 0.0000e0, punt 0.0000e0
                         Name                 State         Calls          Vectors        Suspends         Clocks       Vectors/Call
            TenGigabitEtherneta/0/1-output   active           9003498       211054648               0          1.63e1          23.44
            TenGigabitEtherneta/0/1-tx       active           9003498       211054648               0          7.94e1          23.44
            cop-input                        active           9003498       211054648               0          2.23e1          23.44
            dpdk-input                       polling         45658750       211054648               0          1.52e2           4.62
            ip4-cop-whitelist                active           9003498       211054648               0          4.34e1          23.44
            ip4-input                        active           9003498       211054648               0          4.98e1          23.44
            ip4-lookup                       active           9003498       211054648               0          6.25e1          23.44
            ip4-rewrite-transit              active           9003498       211054648               0          3.43e1          23.44
            ---------------
            Thread 24 vpp_wk_23 (lcore 29)
            Time 45.1, average vectors/node 27.04, last 128 main loops 1.75 per node 28.00
              vector rates in 4.6791e6, out 4.6791e6, drop 0.0000e0, punt 0.0000e0
                         Name                 State         Calls          Vectors        Suspends         Clocks       Vectors/Call
            TenGigabitEthernet88/0/0-outpu   active           7805705       211055503               0          1.54e1          27.04
            TenGigabitEthernet88/0/0-tx      active           7805705       211055503               0          7.75e1          27.04
            cop-input                        active           7805705       211055503               0          2.12e1          27.04
            dpdk-input                       polling         46628961       211055503               0          1.60e2           4.53
            ip4-cop-whitelist                active           7805705       211055503               0          4.35e1          27.04
            ip4-input                        active           7805705       211055503               0          4.86e1          27.04
            ip4-lookup                       active           7805705       211055503               0          6.02e1          27.04
            ip4-rewrite-transit              active           7805705       211055503               0          3.36e1          27.04
```

VPP Cores Not Completely Busy
VPP Vectors Have Space For More Services and More Packets!!
PCIe 3.0 and NICs Are The Limit

And How Do We Know This?
Simples – A Well Engineered Telemetry
In Linux and VPP Tells Us So

VPP average vector size below shows 23-to-27
This indicates VPP program worker threads are not busy
Busy VPP worker threads should be showing 255
This means that VPP worker threads operate at 10% capacity

```
========
TC5     120ge.vpp.24t24pc.ip4.cop
TC5.0     120ge.2pnic.6nic.rss2.vpp
d. testcase-vpp-ip4-cop-scale
         120ge.2pnic.6nic.rss2.vpp.
            64B, 138.000Mpps, 92,7
            IMIX, 40.124832Mpps, 1
            1518, 9.752925Mpps, 12
            ---------------
            Thread 1 vpp_wk_0 (lcore 2)
            Time 45.1, average vectors/node 23.44, last 128 main loops 1.44 per node 23.00
              vector rates in 4.6791e6, out 4.6791e6, drop 0.0000e0, punt 0.0000e0
                         Name                 State         Calls          Vectors        Suspends         Clocks       Vectors/Call
            TenGigabitEtherneta/0/1-output   active           9003498       211054648               0          1.63e1           23.44
            TenGigabitEtherneta/0/1-tx       act
            cop-input                        act
            dpdk-input                       pol
            ip4-cop-whitelist                act
            ip4-input                        act
            ip4-lookup                       act
            ip4-rewrite-transit              act
            ---------------
            Thread 24 vpp_wk_23 (lcore 29)
            Time 45.1, average vectors/node 27.0
              vector rates in 4.6791e6, out 4.67
                         Name                 State         Calls          Vectors        Suspends         Clocks       Vectors/Call
            TenGigabitEthernet88/0/0-outpu   active           7805705       211055503               0          1.54e1           27.04
            TenGigabitEthernet88/0/0-tx      active           7805705       211055503               0          7.75e1           27.04
            cop-input                        active           7805705       211055503               0          2.12e1           27.04
            dpdk-input                       polling         46628961       211055503               0          1.60e2            4.53
            ip4-cop-whitelist                active           7805705       211055503               0          4.35e1           27.04
            ip4-input                        active           7805705       211055503               0          4.86e1           27.04
            ip4-lookup                       active           7805705       211055503               0          6.02e1           27.04
            ip4-rewrite-transit              active           7805705       211055503               0          3.36e1           27.04
```

VPP is also counting the cycles-per-packet (CPP)
We know exactly what feature, service, packet processing activity is using the CPU cores
We can engineer, we can capacity plan, we can automate service placement

We can scale across many many CPU cores and computers
And AUTOMATE it easily – as it is after all just SOFTWARE

ONE MORE THING – THE LONG TERM MAX DELAY

- Low long-term max packet delay with FD.io VPP
- >>120 msec long-term max packet delay measured by others for other vSwitches
- But it is just not not there with VPP and stock Ubuntu 14.04 (no Linux tuning!)

The Fast Data Project (FD.io)

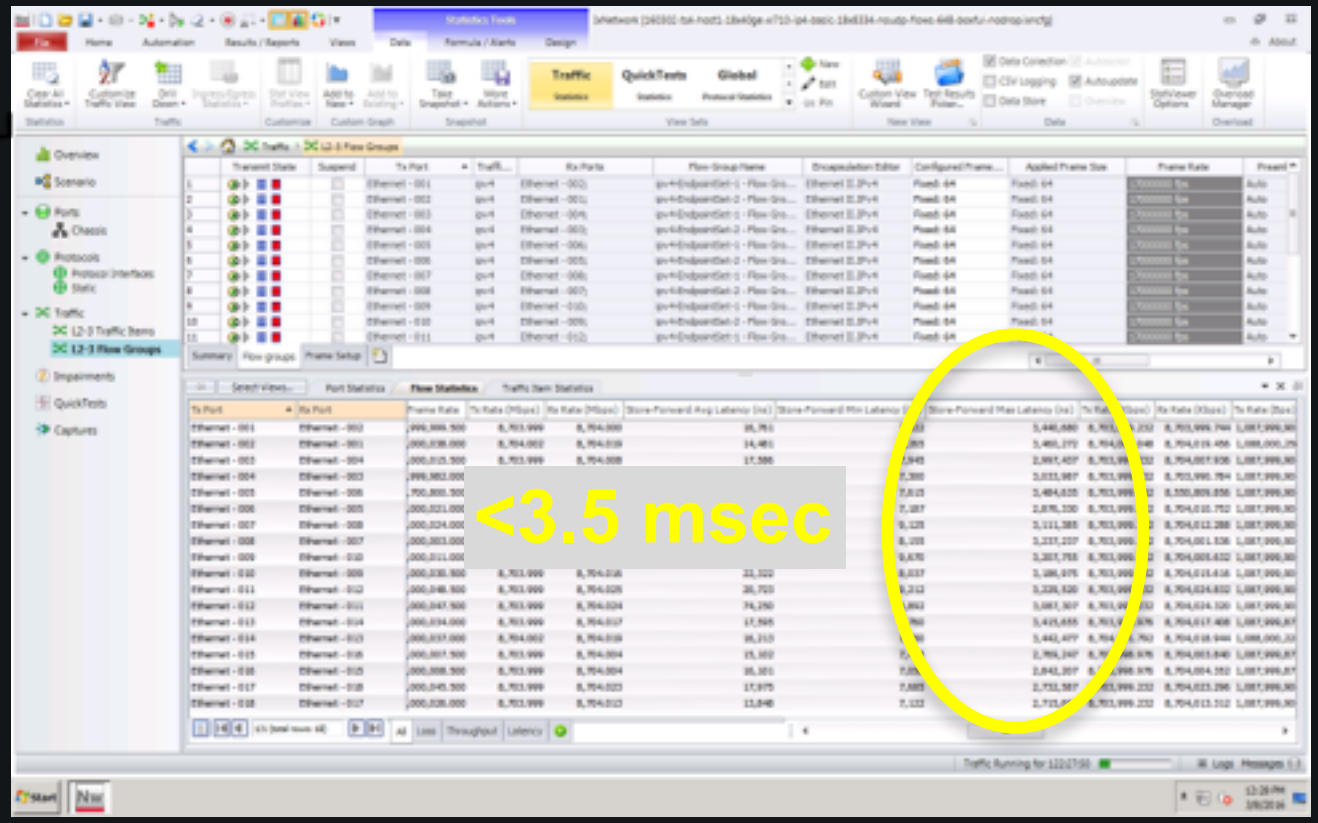
Compute Node Software	Version
Host Operating System	Ubuntu 14.04.3 LTS Kernel version: 3.13.0-63-generic
DPDK	DPDK 2.2.0
FD.io VPP	vpp v1.0.0-174~g57a90e5

Compute Node Hardware	Cisco UCS C460 M4
Chipset	Intel® C610 series chipset
CPU	4 x Intel® Xeon® Processor E7-8890 v3 (18 cores, 2.5GHz, 45MB Cache)
Memory	2133 MHz, 512 GB Total
NICs	9 x 2p40GE Intel XL710 18 x 40GE = 720GE !!

The Soak Test Proof:

18 x 7.7trillion packets forwarded.

VPP: 3.5msec Max Packet Delay incl. the outliers!!



VPP FEATURES

IPv4/IPv6

14+ MPPS, single core
Multimillion entry FIBs
Source RPF
Thousands of VRFs
Controlled cross-VRF lookups
Multipath – ECMP and Unequal Cost
Multiple million Classifiers – Arbitrary N-tuple
VLAN Support – Single/Double tag

Counters for everything
Mandatory Input Checks:
•TTL expiration
•header checksum
•L2 length < IP length
•ARP resolution/snooping
•ARP proxy

IPv4

GRE, MPLS-GRE, NSH-GRE, VXLAN
IPSEC
DHCP client/proxy
CG NAT

IPv6

Neighbor discovery
Router Advertisement
DHCPv6 Proxy
L2TPv3
Segment Routing
MAP/LW46 – IPv4aas
iOAM

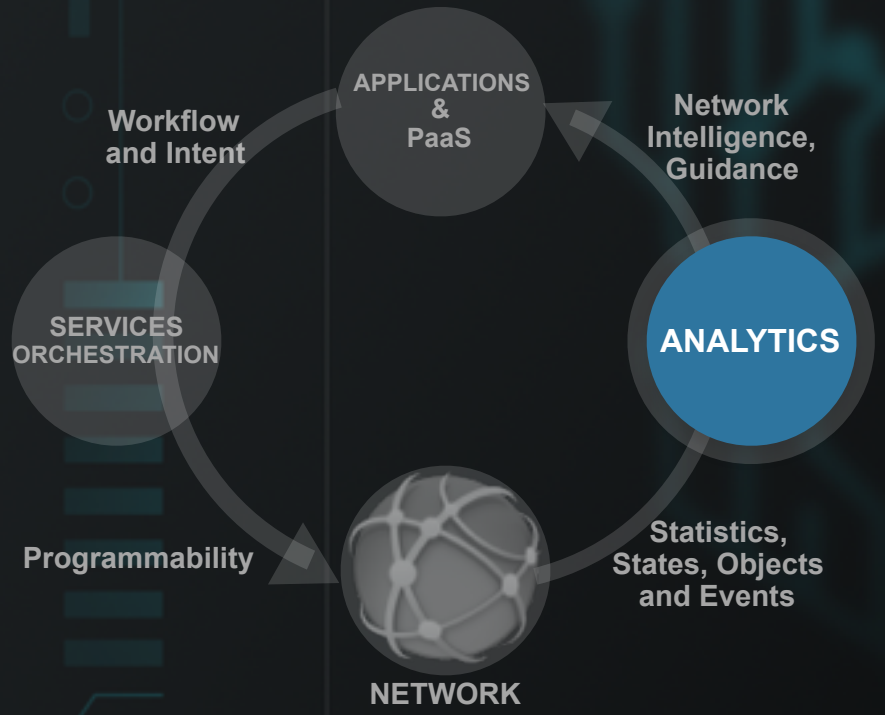
L2

VLAN Support
Single/ Double tag
L2 forwarding with EFP/Bridge Domain concepts
VTR – push/pop/Translate (1:1,1:2, 2:1,2:2)
Mac Learning – default limit of 50k addresses
Bridging – Split-horizon group support/EFP Filtering
Proxy Arp
Arp termination
IRB – BVI Support with Router Mac assignment
Flooding
Input ACLs
Interface cross-connect

MPLS

MPLS-o-Ethernet – Deep label stacks supported

© 2016 Cisco Systems, Inc. All rights reserved.



Operational and Business Intelligence

Platform and tools to aggregate and process the data from the Orchestrated networks and services to gain actionable insight in real-time and enable Data Driven Network and Services

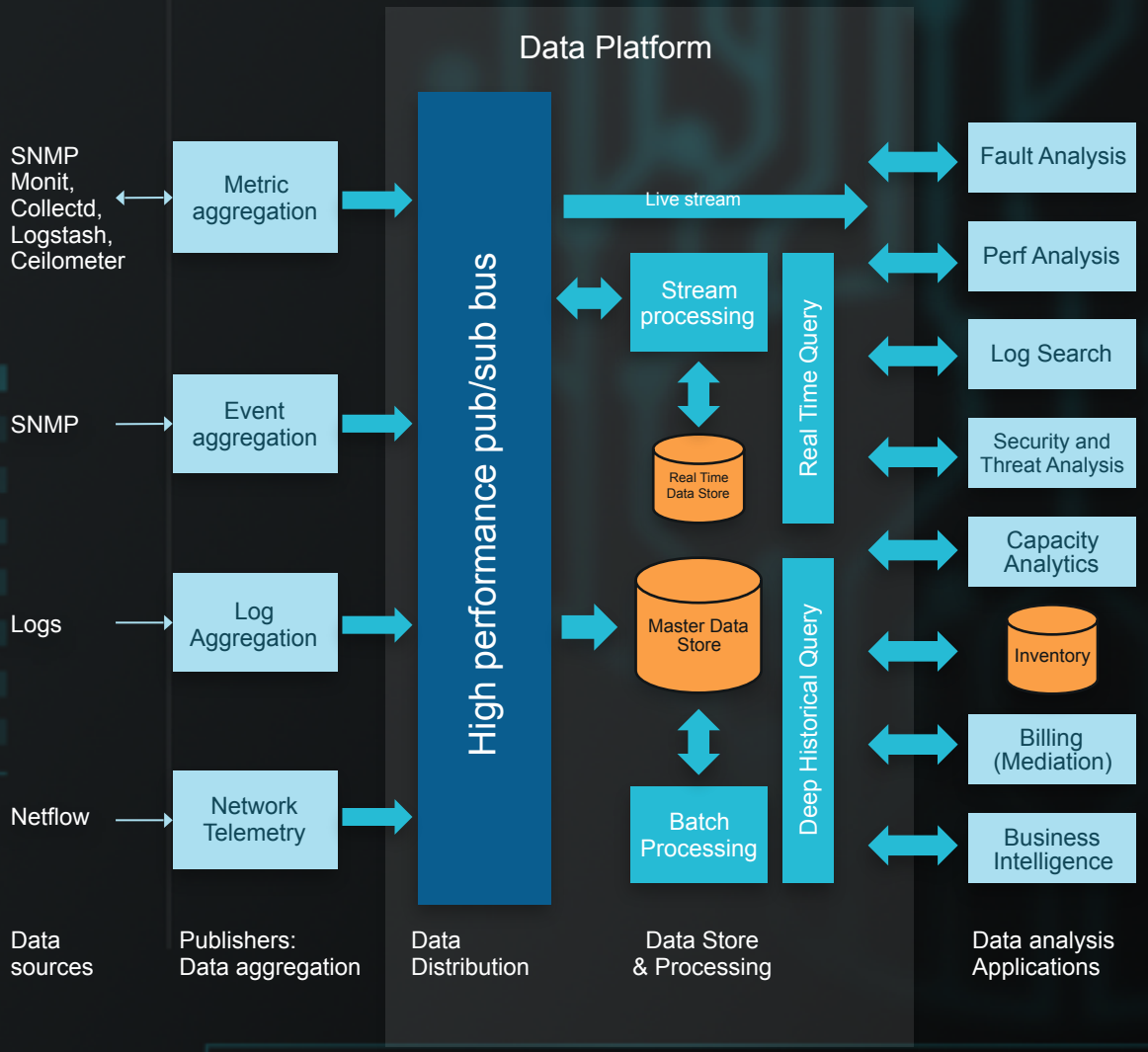
© 2016 Cisco Systems, Inc. All rights reserved.

OSS ANALYTICS IS A BIG DATA PROBLEM

- OSS and BSS analytics applications can be addressed by performing a query function against the entire OSS data set, e.g.
  - Fault management = f(event data)
  - Performance management = f(metric data)
  - Billing mediation = f(event data, metric data)
  - Capacity management = f(metric data)
  - Security analytics = f(metric data, route data

© 2016 Cisco Systems, Inc. All rights reserved.

PLATFORM FOR NETWORK DATA ANALYTICS



| Principles

- Decouple data aggregation (publishers) from data analysis (consumers) – allow any OSS app the potential to access any data source
- Simple, scalable, open data distribution platform
  - Scale-out architecture with support for horizontal scale in all core components
  - Very highly available core platform
  - Low and predictable latency
- Immutable Dataset
  - All data stored raw
  - Minimal filtering/ procesing on ingress
- Minimal filtering/processing on ingress
- Analytics based approach to analysis functions
- Support for streaming apps, real-time queries and batch processing

© 2016 Cisco Systems, Inc. All rights reserved.

PREDICTIVE SERVICES MANAGEMENT

INSTRUMENTATION

OPEN DAYLIGHT

RESTfulAPI

Northbound ODL-Kafka Plugin

Aggregated Events

Compile yang Model and register

NB APIs from Models

Data models (YANG)

Event Handling

Subscribe to a topic (pattern matching)

SAL Service Registry

OpenDaylight

odl-message-bus

SB APIs from Models

Data models (YANG)

Southbound SNMP Plugin

DATA

POLICY

Service Info

Geographical Info

CONTEXT

PaNDA

High performance pub/sub bus

NOTIFICATION

EVENT

DATA

POLICY

Derived Metrics

Trend Analysis

Policy Adaptation

QUERY

CLIENT

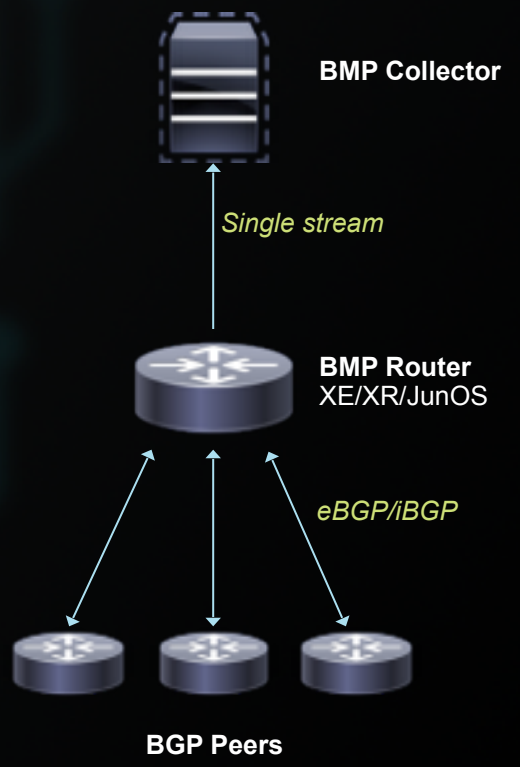
Data Distribution

Data Store & Processing

Consu

ROUTING ANALYTICS LEVERAGING BMP

- BGP Monitoring Protocol (BMP) encapsulates BGP messages from one or more BGP peers into a single TCP stream to one or more collectors
- Efficient, *[near]* real-time, low memory/CPU on router, little to no service impact with peering
- Simplified configuration (one-time setup) with granular controls per peer
- All address families supported
- https://tools.ietf.org/html/draft-ietf-grow-bmp-07



© 2016 Cisco Systems, Inc. All rights reserved.

ROUTING ANALYTICS APPLICATION

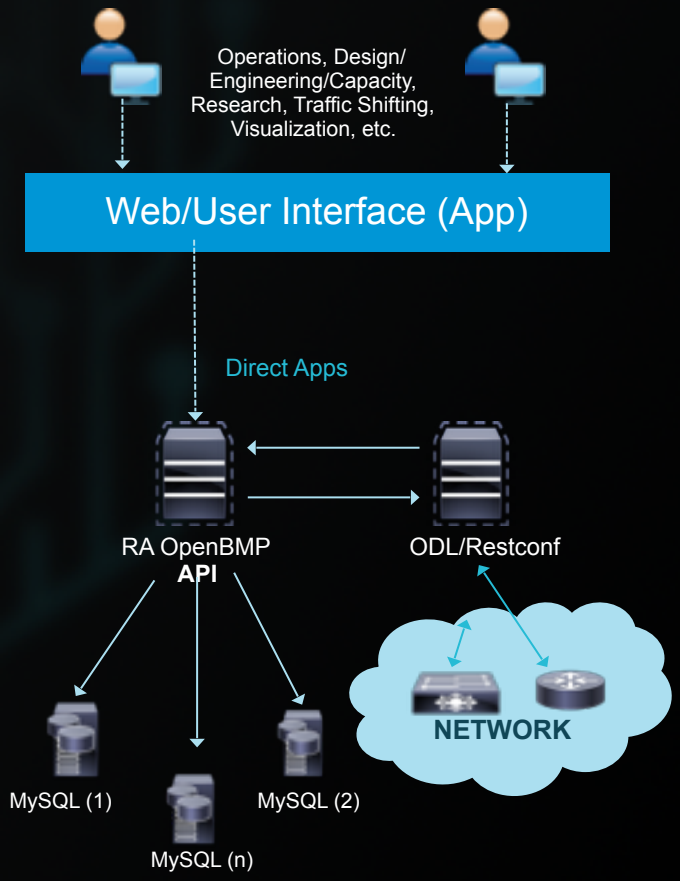
- RA OpenBMP User Interface is a Web based GUI interface that implements the RA OpenBMP API

Dashboards (*includes geo-location/maps*)

- Global router view with peer browser
- Peer view with rib and AS browser
- Link state view with on-demand topology and SPF
- Looking glass

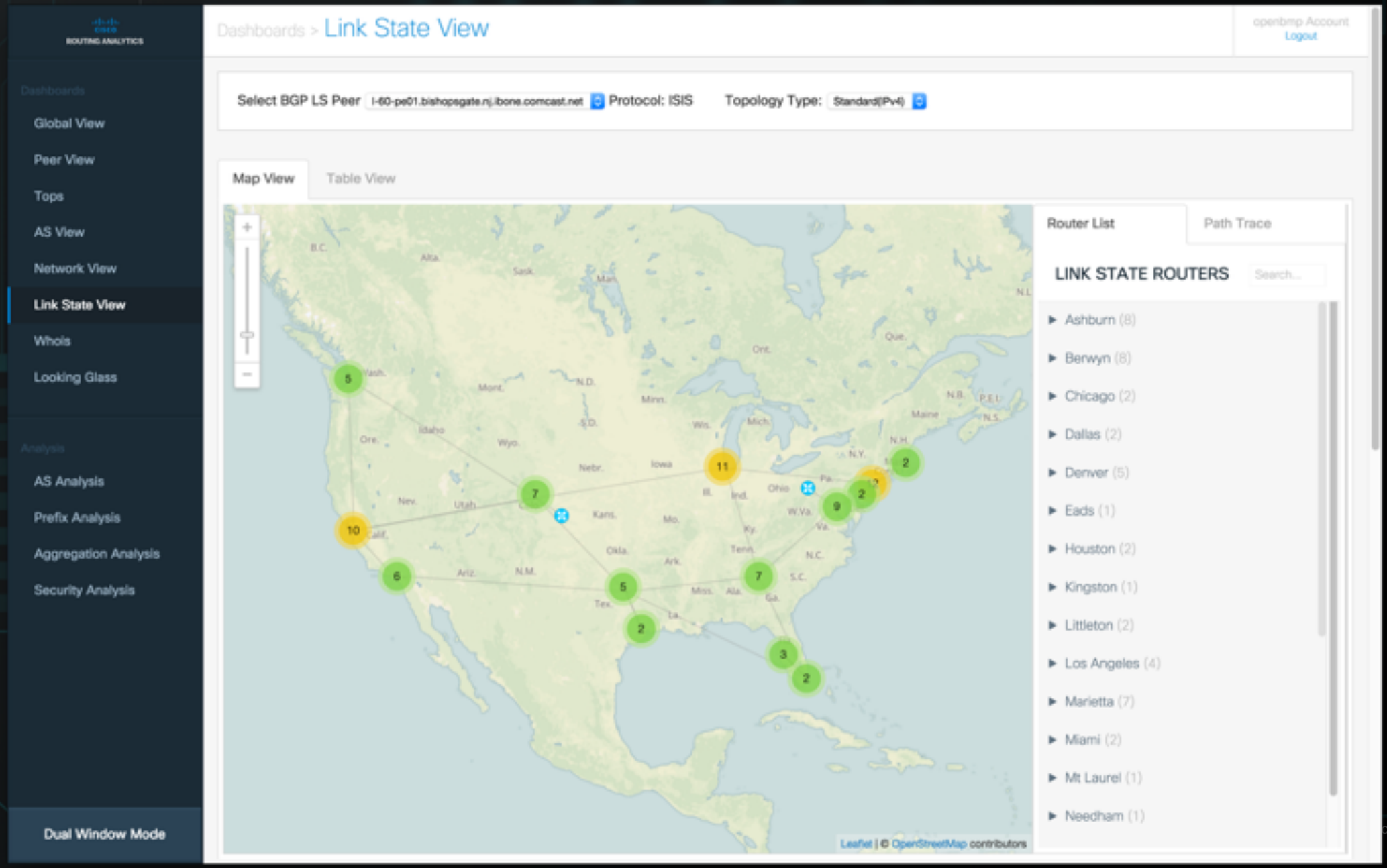
Analysis/Reporting

- Peers
- ASNs
- Prefixes



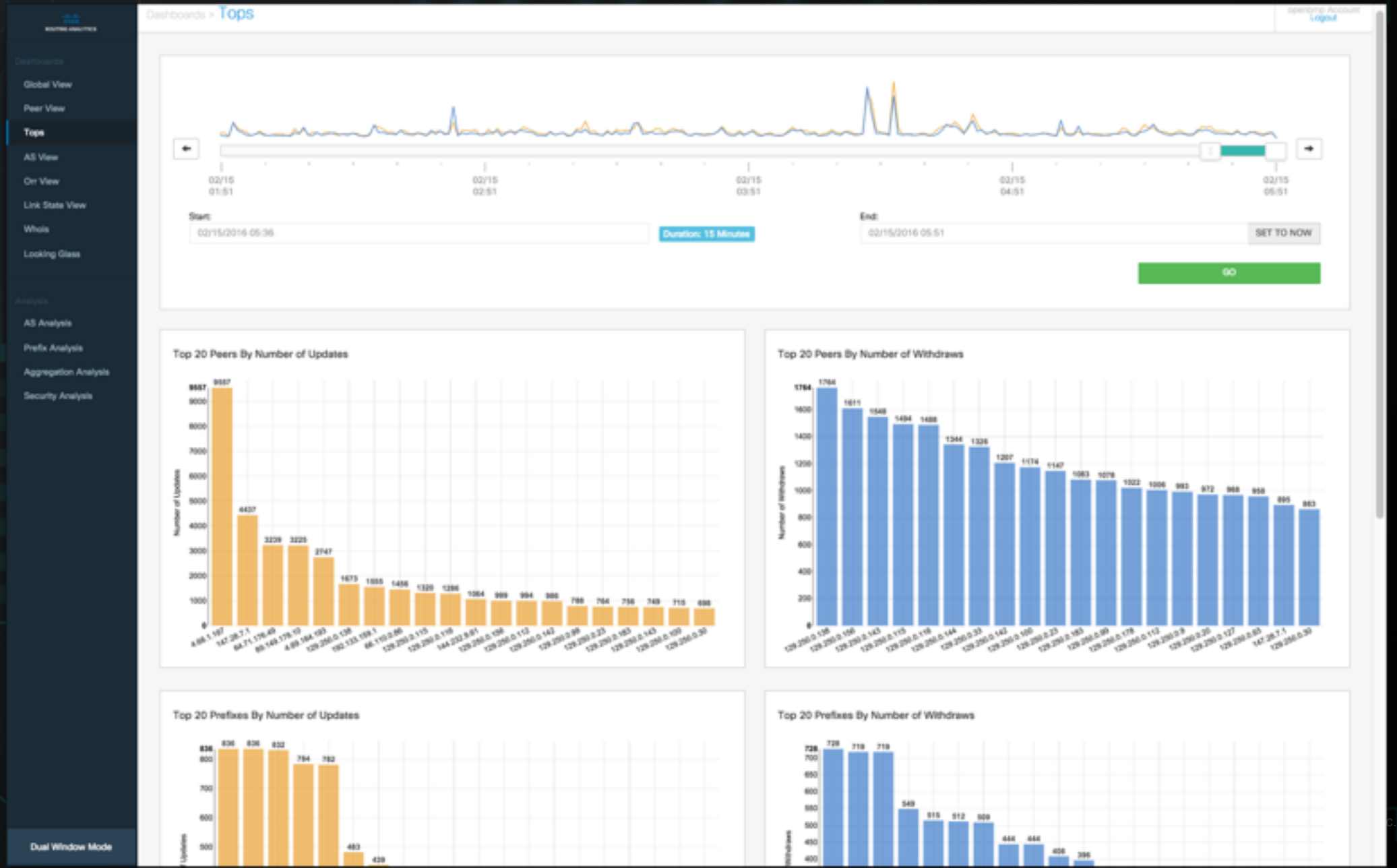
© 2016 Cisco Systems, Inc. All rights reserved.

ROUTING ANALYTICS



c. All rights reserved.

ROUTING ANALYTICS



c. All rights reserved.

Putting the Pieces together

Applications/
Business Outcomes

Model Driven
Orchestration

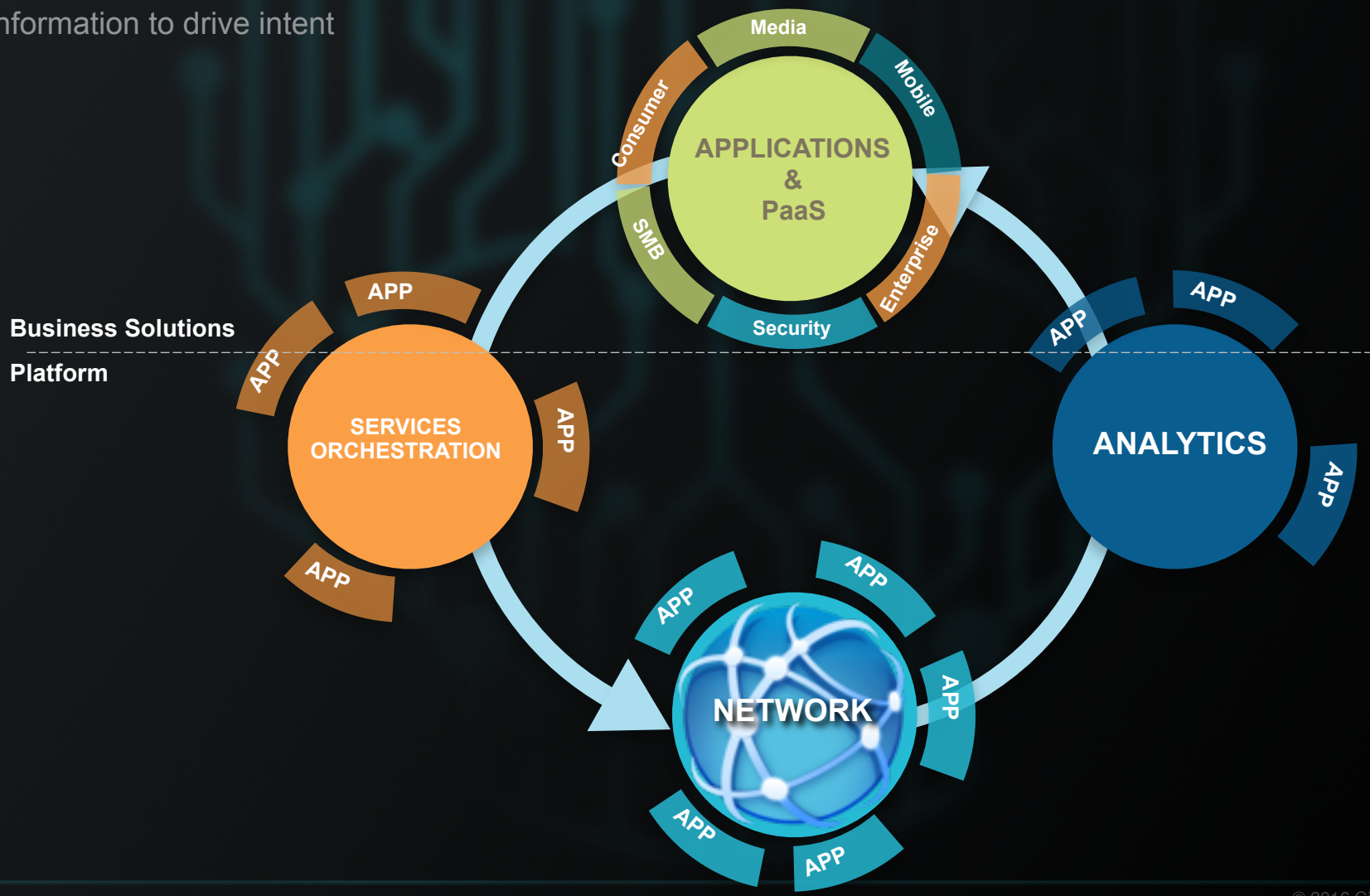
Cloud
Platform

Compute, Network,
Storage Infrastructure

© 2016 Cisco Systems, Inc. All rights reserved.

SDN + DATA DRIVEN NETWORKING = REACTIVE NETWORKING

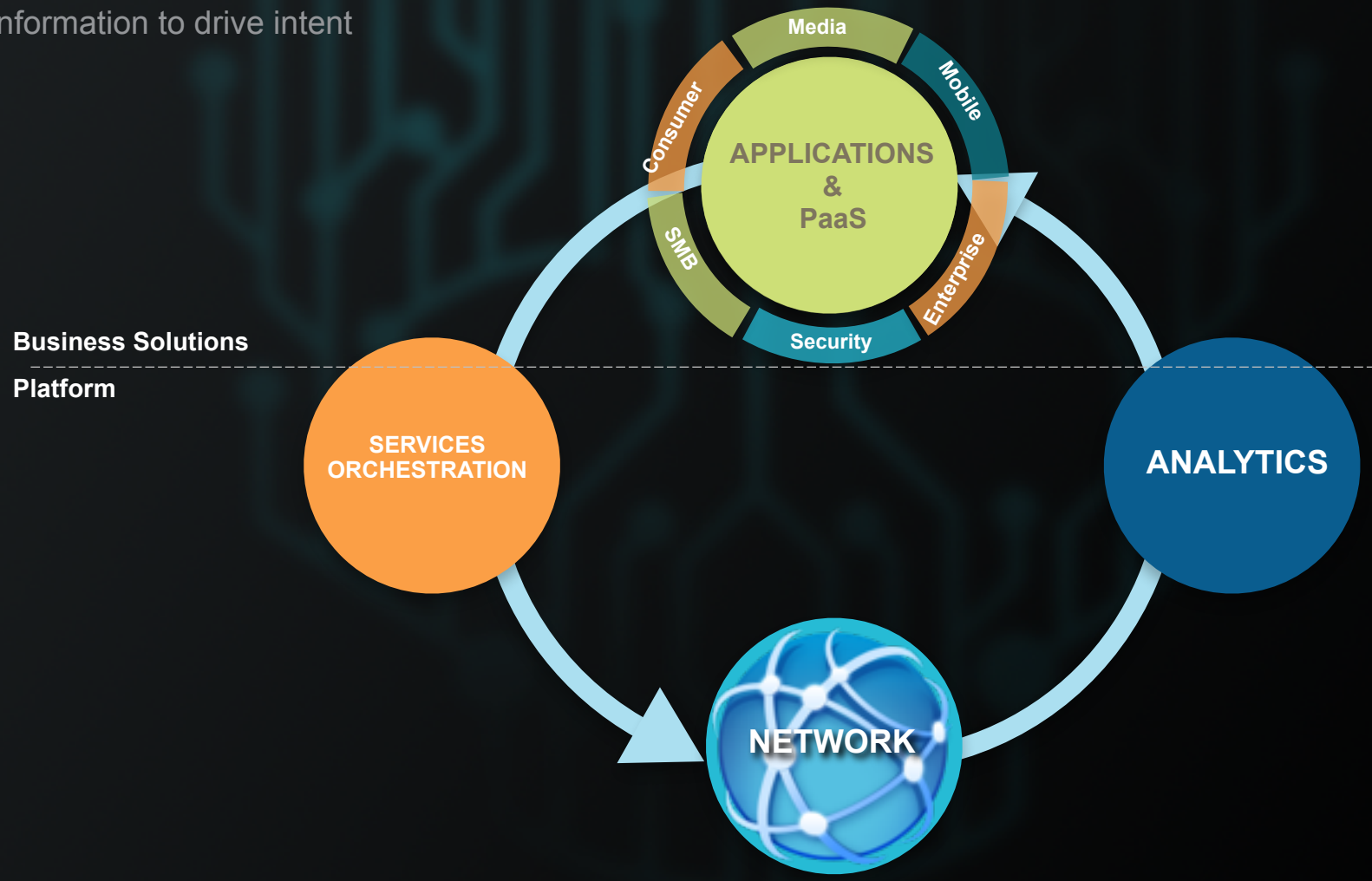
| Turning data into information to drive intent



© 2016 Cisco Systems, Inc. All rights reserved.

SDN + DATA DRIVEN NETWORKING = REACTIVE NETWORKING

| Turning data into information to drive intent



Business Centric

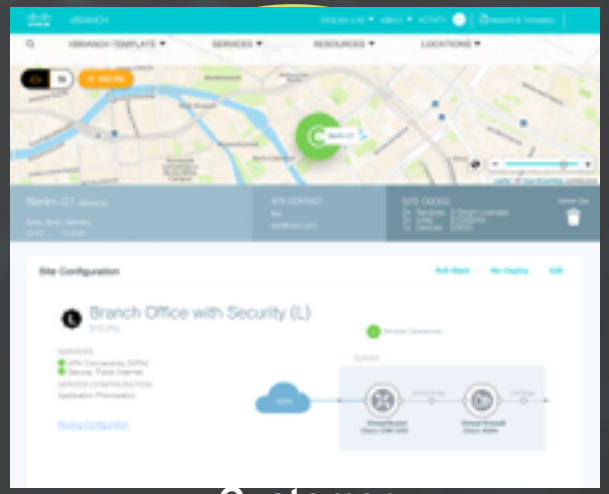
Technology Centric

© 2016 Cisco Systems, Inc. All rights reserved.

SDN + DATA DRIVEN NETWORKING = REACTIVE NETWORKING

| Turning data into information to drive intent

Service Order App



Customer Marketplace

Service Catalog App

Service Assurance App

© 2016 Cisco Systems, Inc. All rights reserved.

CISCO