Sponsored By

Google intel STORDIS The Open Networking Expert

Silicon One – Multiple roles

Guy Caspary
Engineering Director
Cisco

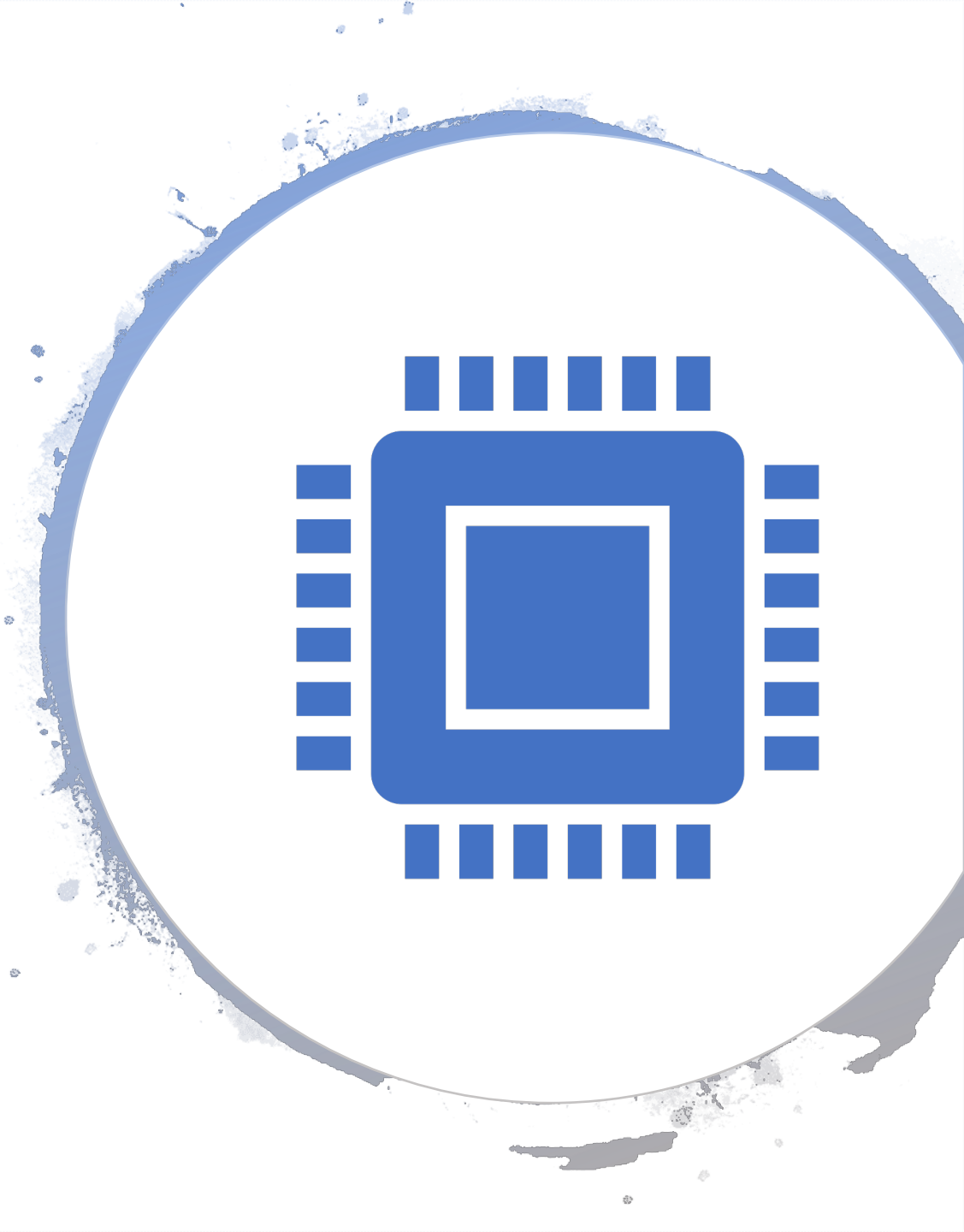
The Vision

A single silicon architecture for all market segments

The problem

- Silicon devices address two different use cases: routers and switches.
- Routers are highly programmable, with deep buffers and large scale.
- Switches are high bandwidth with low power consumption.

The industry diverged into two markets: each defined by unique architecture, systems and software.



CISCO

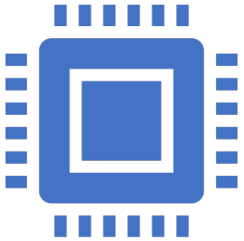
A brief history

- Work on Silicon One started six year ago
- We identified that programmability is a must
  - Features:
    - Faster – SW development is faster than silicon development
    - Longer – features are added pre- and post- silicon
    - Safer – bugs can be fixed post-silicon
  - Broad market applicability
  - A fixed function device must support the super-set of all functions
  - Maintainability

Why P4

- Several tasks were performed in parallel:
  - Coding of high level application to cover all markets
  - Architecture design
  - Development tools (e.g. compiler, debugger)
- Started with C but found out it is too general.
- Application paradigm was naturally match/process.
- Decided to adopt P4 as base.

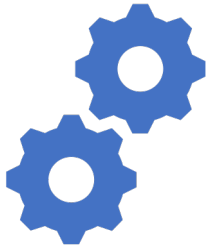
NPU architecture



NPU contains an array of processing engines

Run-to-complete

P4 programmable



One code image covers all markets

Different flows execute different parts of the application

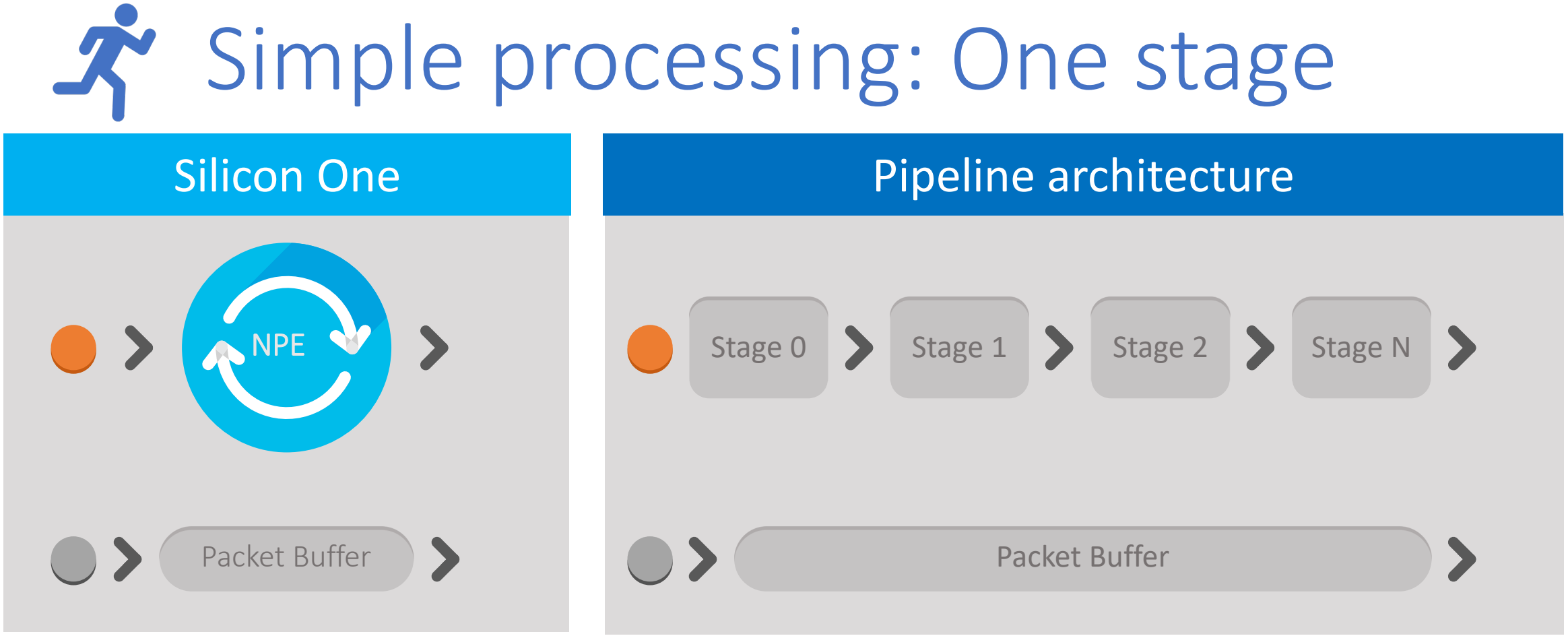
Architecture allows for a different number of processors

Performance and feature scalability

Why run to complete?

- For each packet we spend the “exact” amount of cycles, latency and power required
- Very complex but low bandwidth flows (e.g. control packets) have negligible impact on performance

Simple processing: One stage

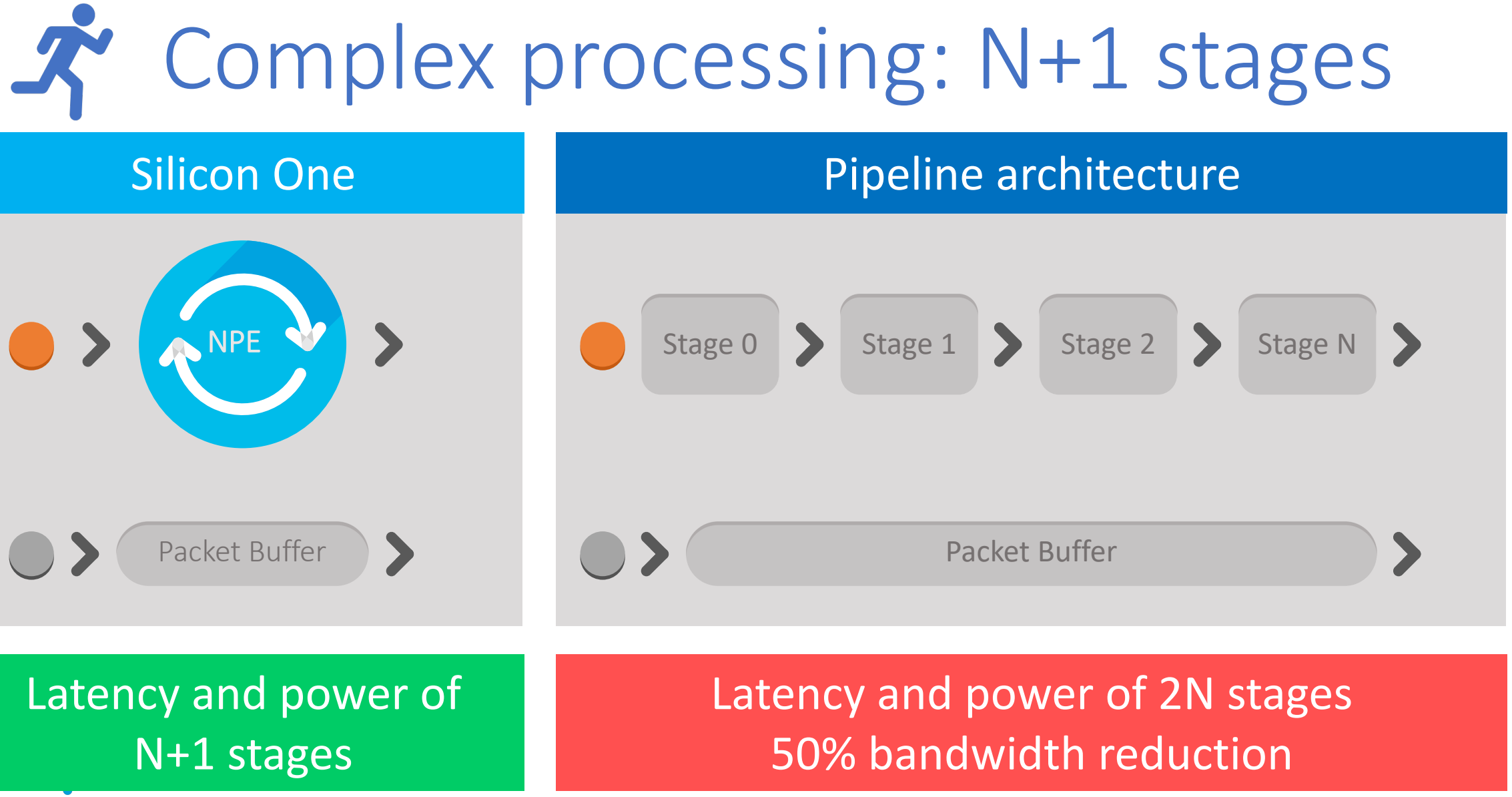


Latency and power of one stage

Latency and power of N stages

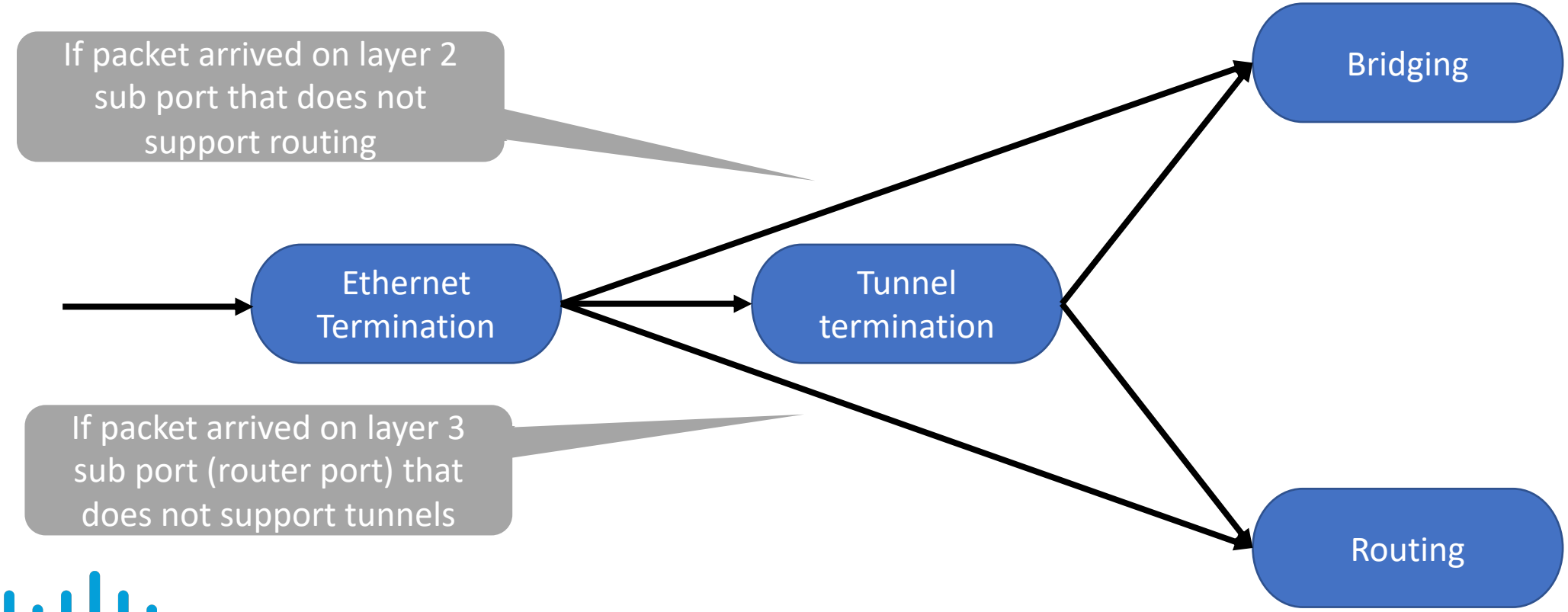
CISCO

Complex processing: N+1 stages



Why run to complete?

- Allows for natural and efficient coding



Cisco Silicon One

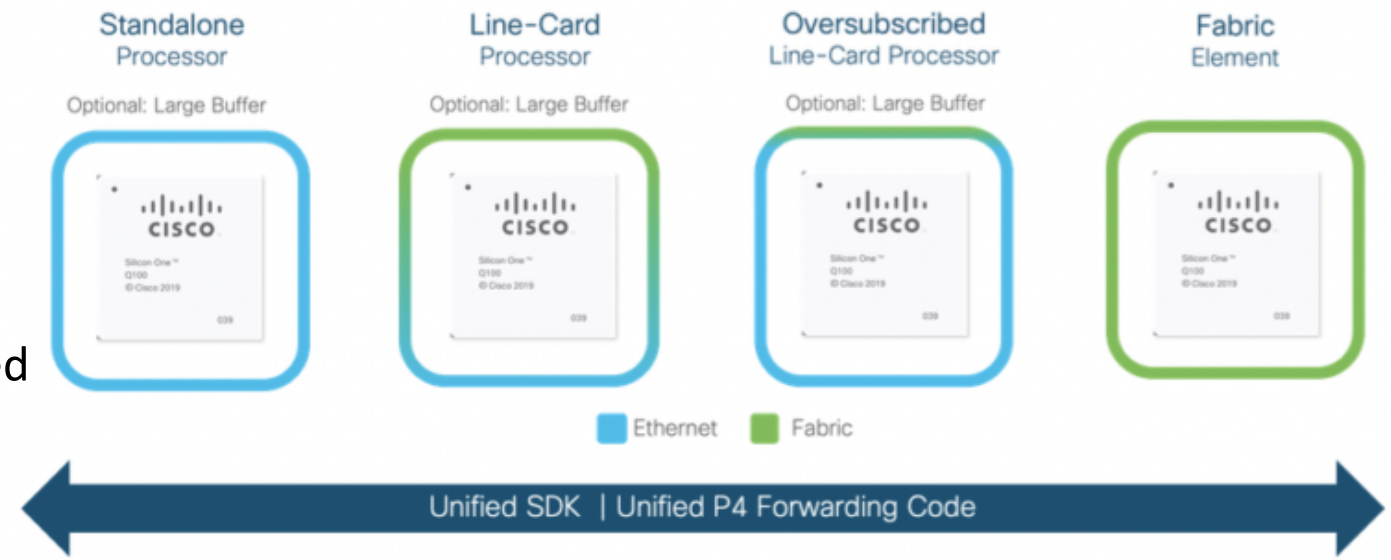
- The first architecture that serves several different market segments – service provider and web-scale.
- Elevating routing silicon performance to the same level as switching silicon performance (bandwidth and power).
- Allows for future product lines to have a consistent silicon architecture.
- ONE experience across the entire network, across all network functions and covering all form factors.
- Significantly reduce OpEx: network engineers save time on testing functionality, qualifying new hardware, and deploying new services with greater consistency and faster time-to-market.

Cisco Silicon One Q100

- The first device in this architecture
- The first routing silicon to break through the 10Tbps benchmark for network bandwidth, without compromising carrier-class capabilities, e.g., feature richness, large queue set, deep buffers, large NPU tables, and advanced programmability.
- Demonstrates many architectural advantages:
  - Supports a fixed switch or router with 10.8T worth of network ports up to large non-blocking distributed routers with Petabit scales.
  - All with non-blocking performance, deep buffering with rich QoS, and programmable forwarding.

Cisco Silicon One Q100 – cont.

- A single device may be used as standalone network processor (optional deep buffers), line card network processor (optional deep buffers) and fabric element in a distributed router.
- All accomplished with a common and unified P4 forwarding code and SDK.



Cisco 8000 series routers

- Powered by Cisco Silicon One ASICs supporting full routing functionality with a single ASIC.
- Platforms scale from 10.8 Tbps to 260 Tbps.
- Common architecture and designs across modular and fixed routers provide topological regularity, scalability, consistent features and operational simplicity.

Cisco 8200 Series

- Uses the Cisco Silicon One Q100 to deliver full routing functionality with a single ASIC per router.
- The architecture supports large forwarding tables, deep buffers, flexible packet operations, and enhanced programmability.
- Two fixed 10.8Tb/s platforms:
  - The Cisco 8201 is a 1RU fixed configuration with 24x400GbE and 12x100GbE ports
  - The Cisco 8202 is a 2RU fixed configuration with 12x400GbE and 60x100GbE ports

Cisco 8800 Series

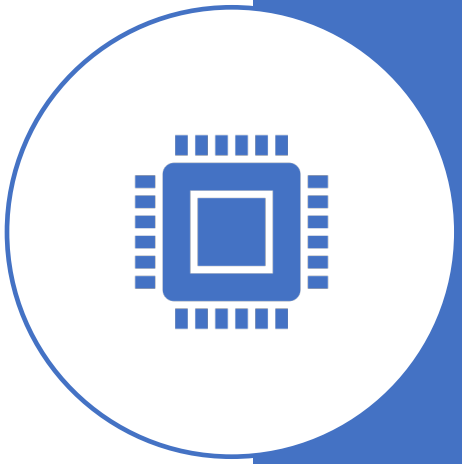
High bandwidth via modular chassis with a redundant control plane and switch fabric:

- The Cisco Router 8808 is an 8-slot, 115.2 Tb/s, 16RU chassis
- The Cisco Router 8812 is a 12-slot, 172.8 Tb/s, 21RU chassis
- The Cisco Router 8818 is an 18-slot, 259.2 Tb/s, 33RU chassis

*Total bandwidth is based on Silicon One Q100; infrastructure provides for more. Stay tuned …

Programmability in action: Pre/Post silicon validation

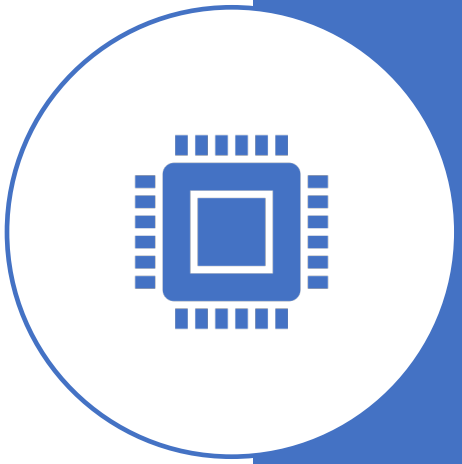
- A dedicated P4 application for silicon verification and validation
- A non-standard injected packet contains commands for the device:
  - Path through device
  - Resources to exercise:
    - Lookups
    - Counter increments
    - more



CISCO

Programmability in action: power measurement

- Fast bring-up of power testing application defined by customer:
  - Ethernet packets
  - Specific flow through ports
  - Specific database lookups
- Prototype in 2 hours
- Debug and testing in 2 days
- Delivered a week after request …
- Shipping and customs clearance 7 days



CISCO

Silicon One

Sponsored By

Google intel STORDIS The Open Networking Expert

Thank You

Contact: Guy Caspary <gcaspary@cisco.com>

Sponsored By

Google intel STORDIS The Open Networking Expert

Security Stability and Transparency: is P4 the answer?

Dr. Paola Grosso
Associate Professor “Multiscale Networked Systems”
University of Amsterdam – The Netherlands

mns-research.nl

Science usecases

Photo source: CERN

High energy physics (LHC)

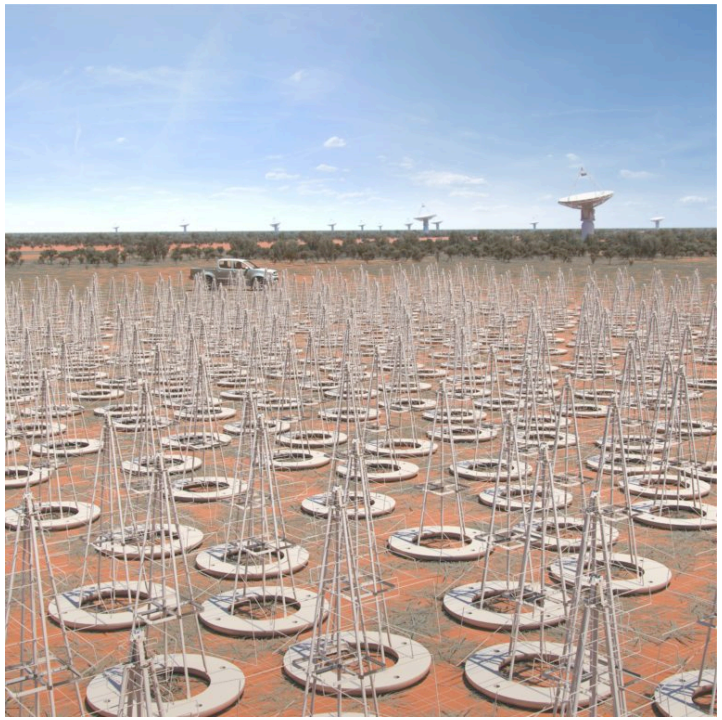


Photo source: SKA organisation

Radio astronomy (SKA)

Science data is moving around the world

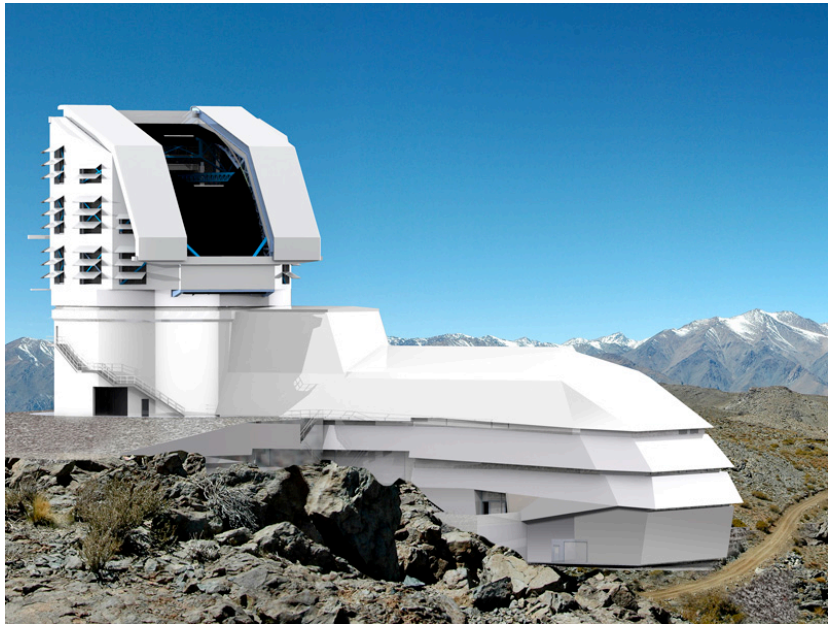


Photo source: LSST/NSF/AURA

Radio astronomy (LSST)

Societal usecases

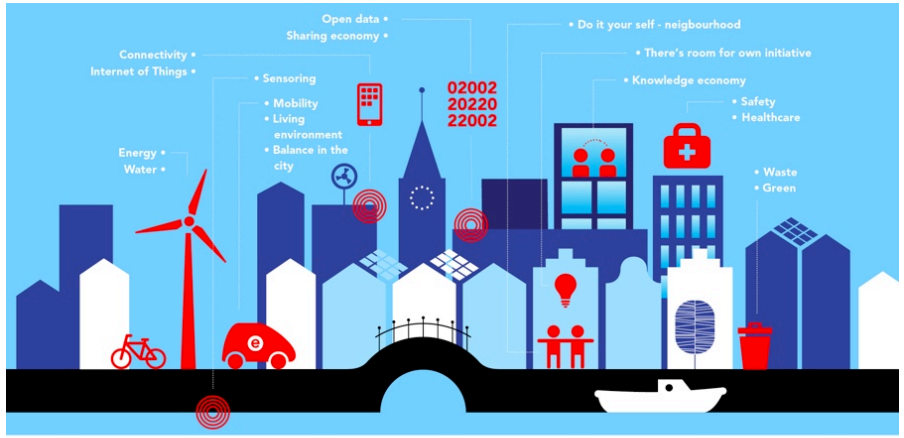
Photo source: Genetic Literacy Project

Personalized medicine (EPI)

Photo source: DL4LD project

Logistics (DL4LD)

Personal data is shared by many parties.



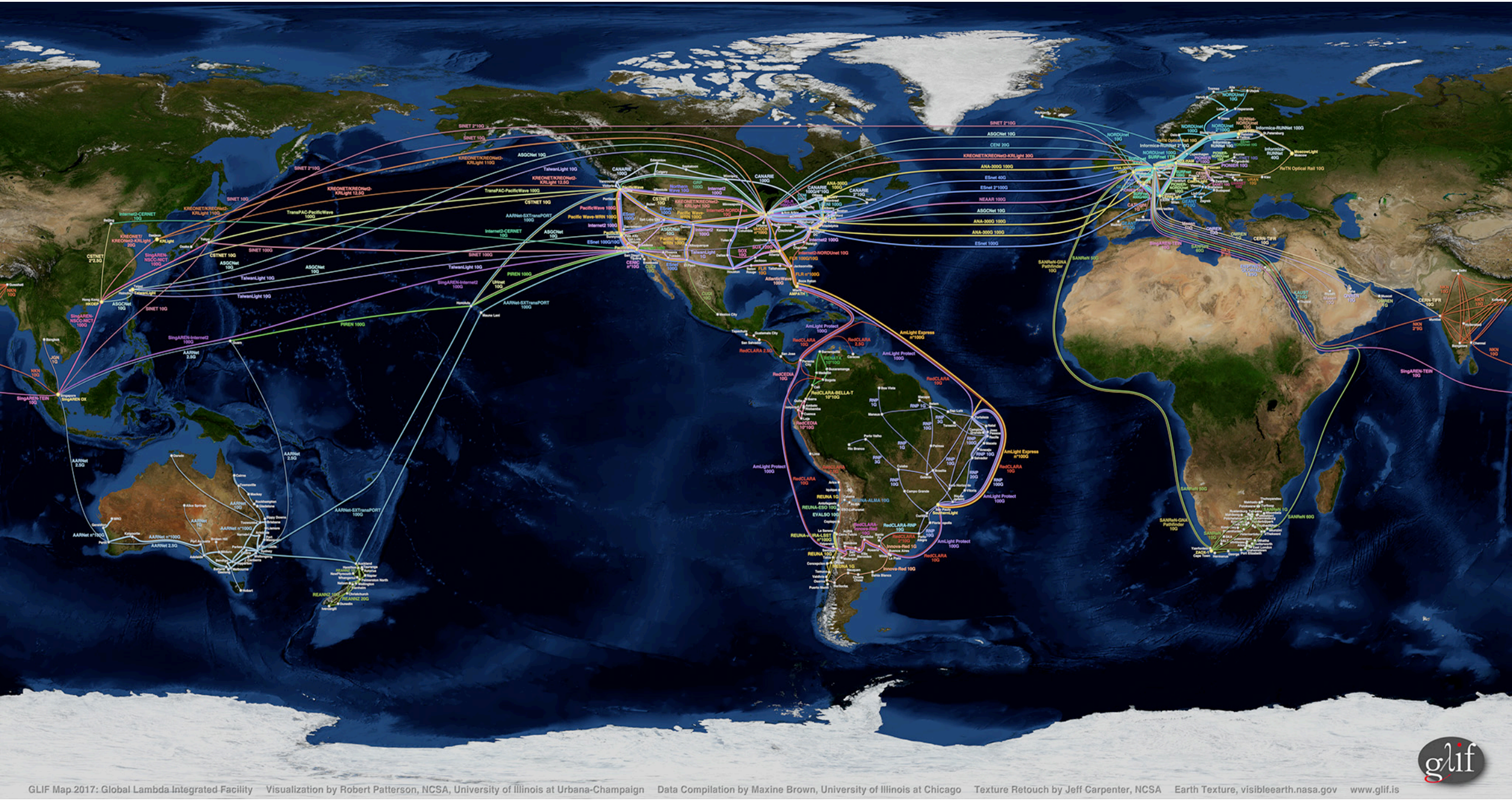
Smart cities (AMDeX)

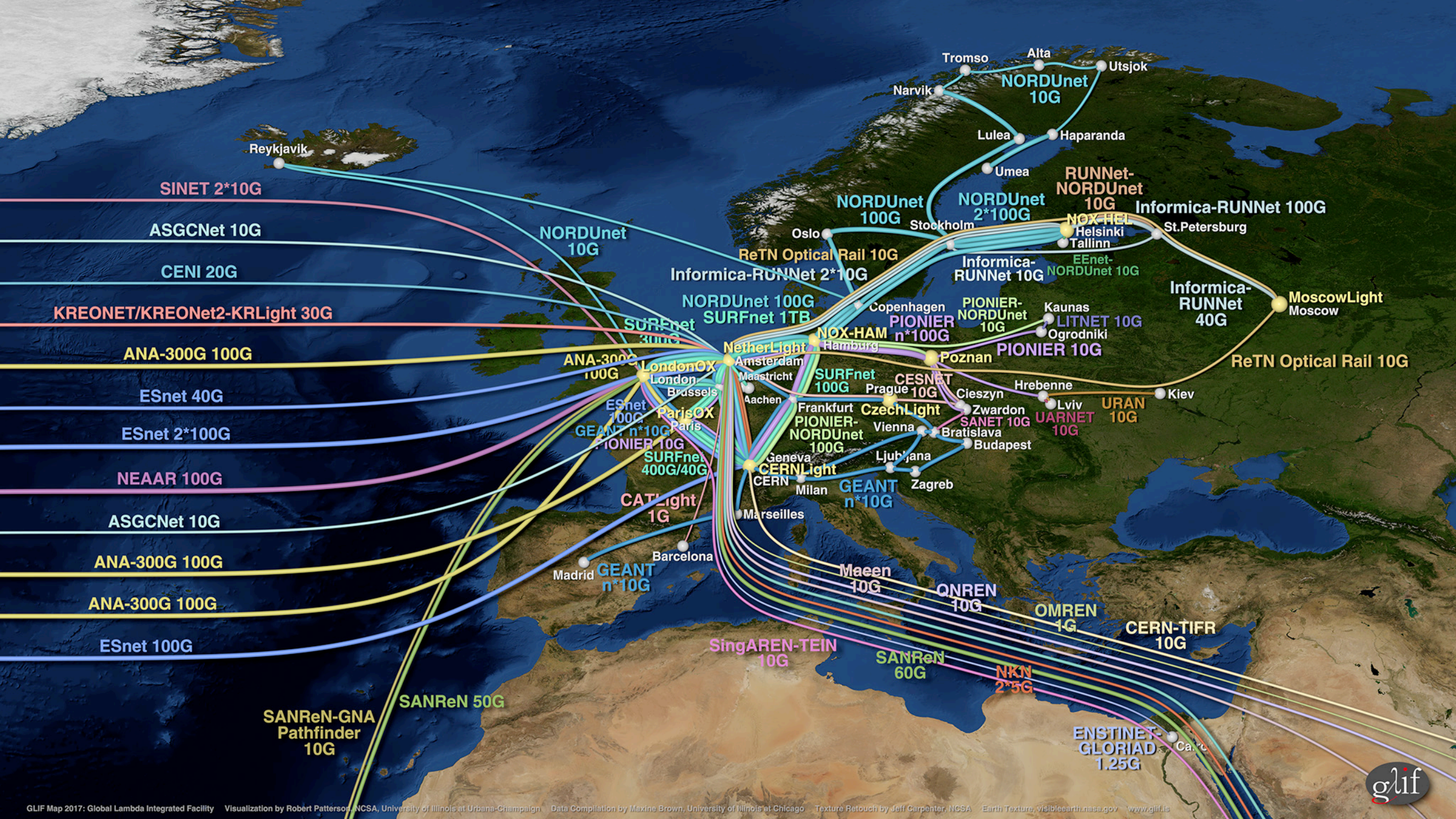
Photo source: AMS Economic Board

Multiscale Networked Systems

How can we provide security, stability and transparency in the networks of the Future?

Not just from the perspective of network operators but also for end users!


SINET 2*10G
SINET 10G
ASGCNet 10G
KREONET/KREONet2-KRLight 11.5G
KREONET/KREONet2-KRLight 12.5G
TransPAC-PacificWave 100G
CSTNET 10G
Internet2-CERNET 10G
KREONET/KREONet2-KRLight 20G
SingAREN-NSCC-NICT 100G
CSTNET 2*2.5G
ASGCNet 10G
TaiwanLight 10G
SingAREN-Internet2 100G
AARNet 2.5G
PIREN 100G
AARNet-SXTransPORT 100G
CANARIE 100G
PacificWave 100G
Pacific Wave-WRN 100G
Internet2 100G
ESnet 100G/10G
CSTNET 10G
AARNet-SXTransPORT 100G
Internet2-CERNET 10G
ASGCNet 10G
SINET 100G
TaiwanLight 10G
PIREN 100G
UHnet 10G
AARNet-SXTransPORT 100G
SingAREN-Internet2 100G
AARNet 2.5G
SingAREN-TEIN 10G
AARNet 2.5G
AARNet n*100G
AARNet 10G
AARNet-SXTransPORT 100G
REANNZ 10G
REANNZ 20G
SINET 2*10G
ASGCNet 10G
CENI 20G
KREONET/KREONet2-KRLight 30G
ANA-300G 100G
ESnet 40G
ESnet 2*100G
NEAAR 100G
ASGCNet 10G
ANA-300G 100G
ANA-300G 100G
ESnet 100G
NORDUnet 10G
NORDUnet 100G
NORDUnet 2*100G
RUNNet-NORDUnet 10G
Informica-RUNNet 100G
Informica-RUNNet 2*10G
NORDUnet 100G
SURFnet 1*10G
PIONIER 10G
GÉANT 10G
ReTN Optical Rail 10G
MoscowLight Moscow
Informika-RUNNet 40G
CERN-TIFR 10G
SANReN 50G
SANReN-GNA Pathfinder 10G
SANReN 60G
SANReN 1G
SingAREN-TEIN 10G
AmLight Protect 100G
AmLight Express n*100G
RedCLARA 10G
RedCLARA 2.5G
RedCLARA-BELLA-T 10*10G
RedCEDIA 10G
RedCEDIA 10*10G
RENATA 10*10G
RNP 1G
RNP 3G
RNP 10G
RNP 20G
RNP 100G
REUNA 1G
REUNA-ALMA 10G
REUNA-ESO 10G
EVALSO 10G
RedCLARA-Innova-Red 10G
RedCLARA-RNP 10G
RedCLARA 2*10G
Innova-Red 1G
Innova-Red 10G
REUNA-AURA-LSST n*10G
GLIF Map 2017: Global Lambda Integrated Facility
Visualization by Robert Patterson, NCSA, University of Illinois at Urbana-Champaign
Data Compilation by Maxine Brown, University of Illinois at Chicago
Texture Retouch by Jeff Carpenter, NCSA
Earth Texture, visibleearth.nasa.gov
www.glif.is


Tromso
Alta
Utsjok
NORDUnet 10G
Narvik
Lulea
Haparanda
Umea
Reykjavik
SINET 2*10G
ASGCNet 10G
CENI 20G
KREONET/KREONet2-KRLight 30G
ANA-300G 100G
ESnet 40G
ESnet 2*100G
NEAAR 100G
ASGCNet 10G
ANA-300G 100G
ANA-300G 100G
ESnet 100G
NORDUnet 10G
NORDUnet 100G
NORDUnet 2*100G
RUNNet-NORDUnet 10G
Informica-RUNNet 100G
Stockholm
NOX-HEL
Helsinki
Tallinn
St.Petersburg
Oslo
ReTN Optical Rail 10G
Informica-RUNNet 2*10G
Informica-RUNNet 10G
EEnet-NORDUnet 10G
Informica-RUNNet 40G
MoscowLight
Moscow
NORDUnet 100G
SURFnet 1TB
Copenhagen
PIONIER-NORDUnet 10G
Kaunas
LITNET 10G
Ogrodniki
PIONIER n*100G
NOX-HAM
Hamburg
NetherLight
Amsterdam
Poznan
PIONIER 10G
ReTN Optical Rail 10G
SURFnet 100G
LondonOX
London
Maastricht
Brussels
Aachen
Prague
CESNET 10G
Cieszyn
Hrebenne
Kiev
Frankfurt
CzechLight
Zwardon
Lviv
URAN 10G
ESnet 100G
ParisOX
Paris
GEANT n*10G
PIONIER 10G
PIONIER-NORDUnet 100G
Vienna
SANET 10G
UARNET 10G
Bratislava
Budapest
SURFnet 400G/40G
Geneva
CERNLight
CERN
Milan
Ljubljana
Zagreb
GEANT n*10G
CATLight 1G
Marseilles
Barcelona
Madrid
GEANT n*10G
Maeen 10G
QNREN 10G
OMREN 1G
CERN-TIFR 10G
SingAREN-TEIN 10G
SANReN 60G
NKN 2*5G
SANReN 50G
SANReN-GNA Pathfinder 10G
ENSTINET-GLORIAD 1.25G
GLIF Map 2017: Global Lambda Integrated Facility
Visualization by Robert Patterson, NCSA, University of Illinois at Urbana-Champaign
Data Compilation by Maxine Brown, University of Illinois at Chicago
Texture Retouch by Jeff Carpenter, NCSA
Earth Texture, visibleearth.nasa.gov
www.glif.is
glif

Multiscale Networked Systems

Why P4?

Per packet processing in the dataplane provides advantages compared to out-of-band approaches for fine grained telemetry. No need for summarization and a wealth of information that is usable in many contexts.

- Transparency goal:
  - From telemetry we acquire insights in what is happening in the network, eg the path taken by flows.
- Security goal:
  - From telemetry follows the possibility to identify attacks and feed intrusion detection systems (see SARnet project).
- Stability goal:
  - From telemetry follows you can identify bottlenecks and buffers filling up along the path, eg the amount of data in queues leveraging time stamping.

Multiscale Networked Systems

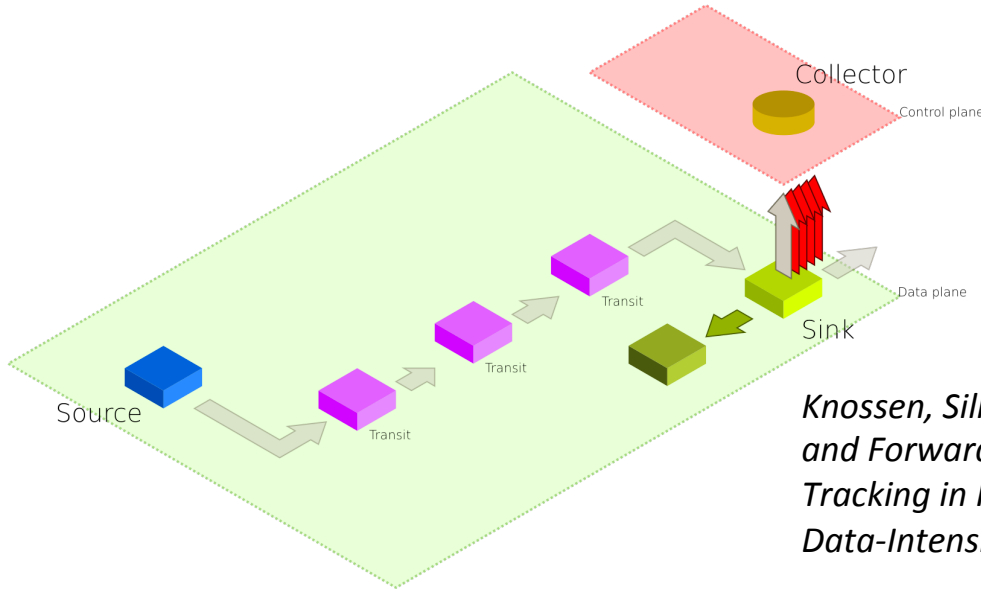
How we use P4?

Use of the software switch and compilers from p4.org for our initial research and then moved to hardware (*).

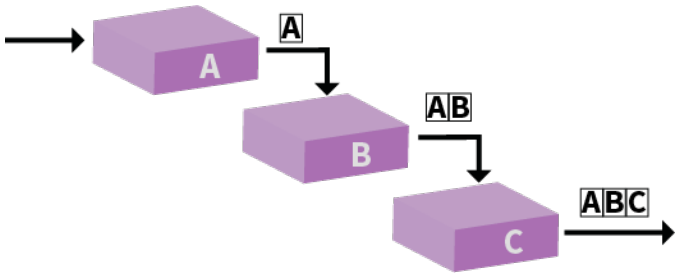
- Barefoot switches give us better performance switching capabilities
- Smart NICS allow us to bring the P4 capabilities all the way to host

(*) Still very useful in educational settings for students labs.

Telemetry

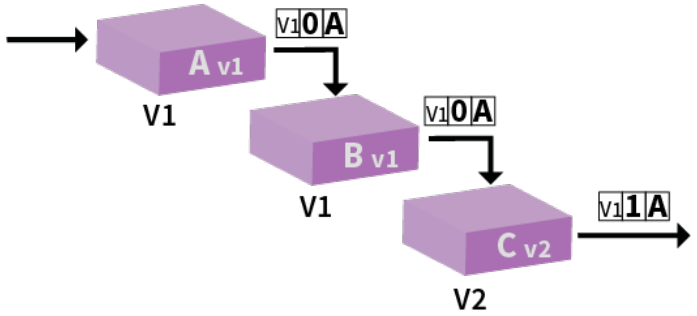


Knossen, Silke, Joseph Hill, and Paola Grosso. "Hop Recording and Forwarding State Logging: Two Implementations for Path Tracking in P4." 2019 IEEE/ACM Innovating the Network for Data-Intensive Science (INDIS). IEEE, 2019.



Next Header : next header type	Header Extension Length: 0x02	Padding: 0x00 * 6
Option Type: 0x3F	Option Data Length: 0x06	Option Data: Node Identifier
Option Type: 0x3F	Option Data Length: 0x06	Option Data: Node Identifier

8 bytes

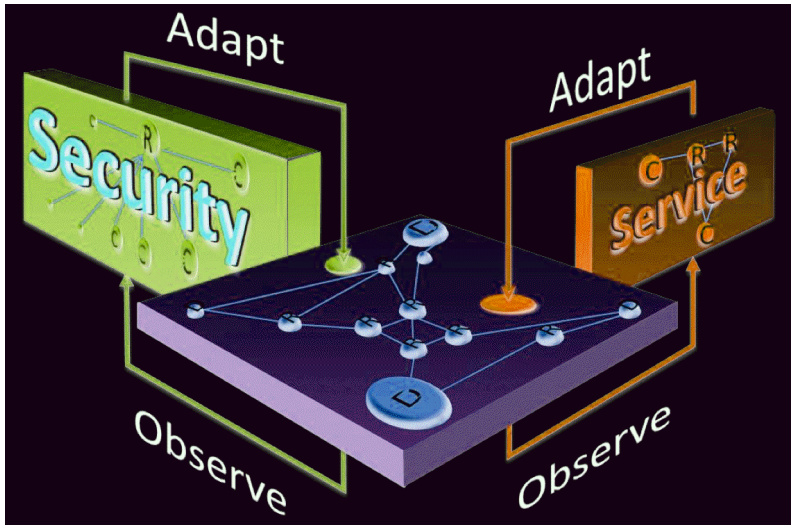


Next Header : next header type	Header Extension Length: 0x01	Padding: 0x00 * 6
Option Type: 0x3F	Option Data Length: 0x06	Option Data

version	trackable	Node Identifier

8 bytes

Security

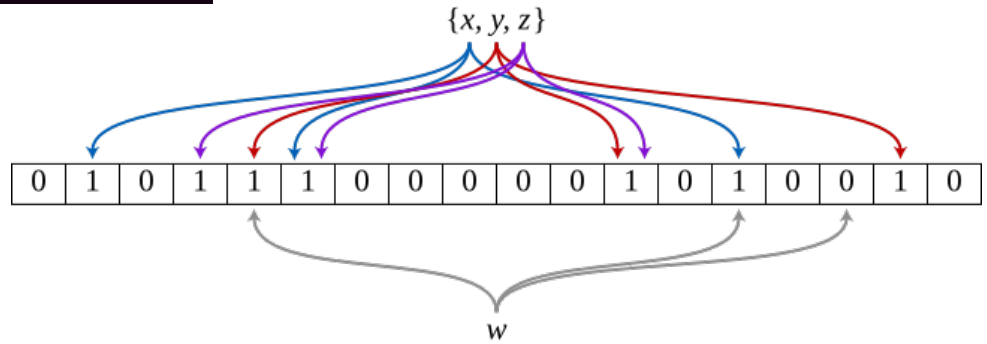


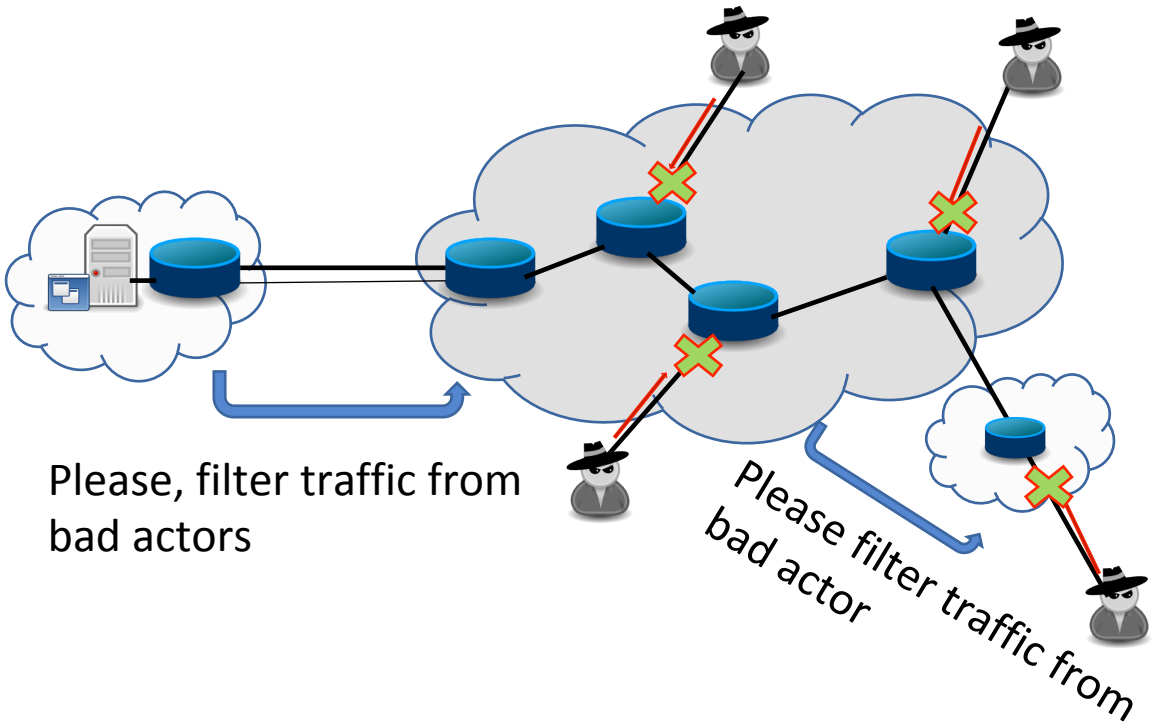
Multiscale Networked Systems

Adapting for autonomous response

(ML learning)

Bloom filters in P4

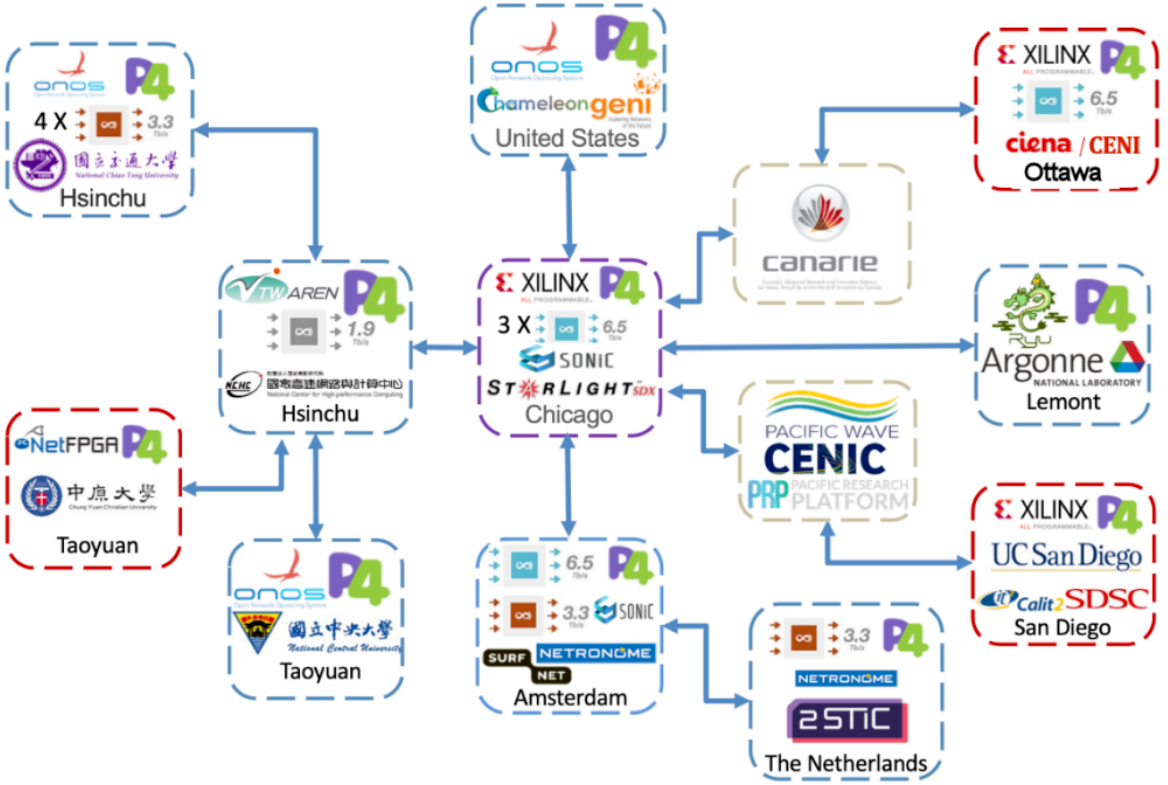




Hill, Joseph, Mitchel Aloserij, and Paola Grosso. "Tracking network flows with P4."
2018 IEEE/ACM Innovating the Network for Data-Intensive Science (INDIS). IEEE, 2018.

2STiC

SECURITY, STABILITY AND TRANSPARENCY OF INTER-NETWORK COMMUNICATIONS



2STiC testbed

High performance and large scale deployment to test the protocols

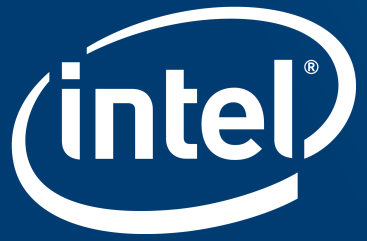
Sponsored By

Google intel STORDIS The Open Networking Expert

Thank You…
… and thanks to Silke Knossen, Mitchell Aloserij and Joseph Hill (from UvA)

p.grosso@uva.nl

2stic.nl
mns-research.nl



P4-Programmable Data Plane Use-cases

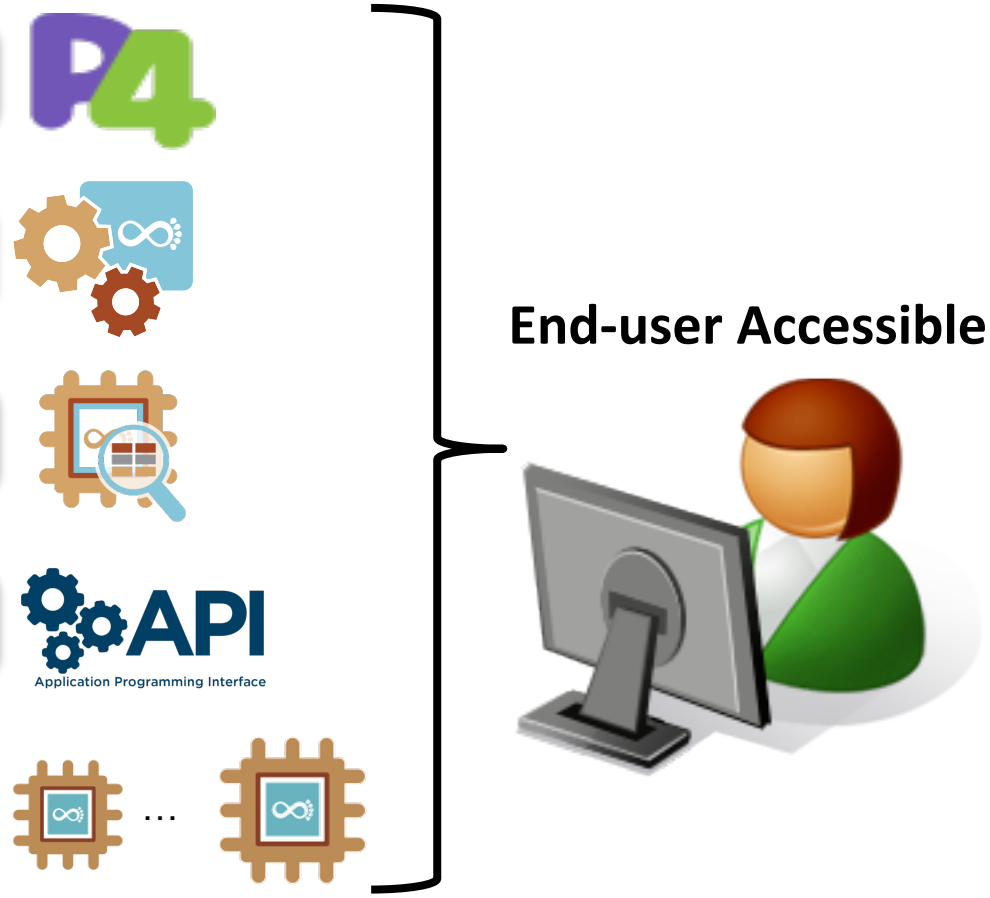
Arkadiy Shapiro, Product Line Manager, Software

Barefoot Division, Connectivity Group, Intel

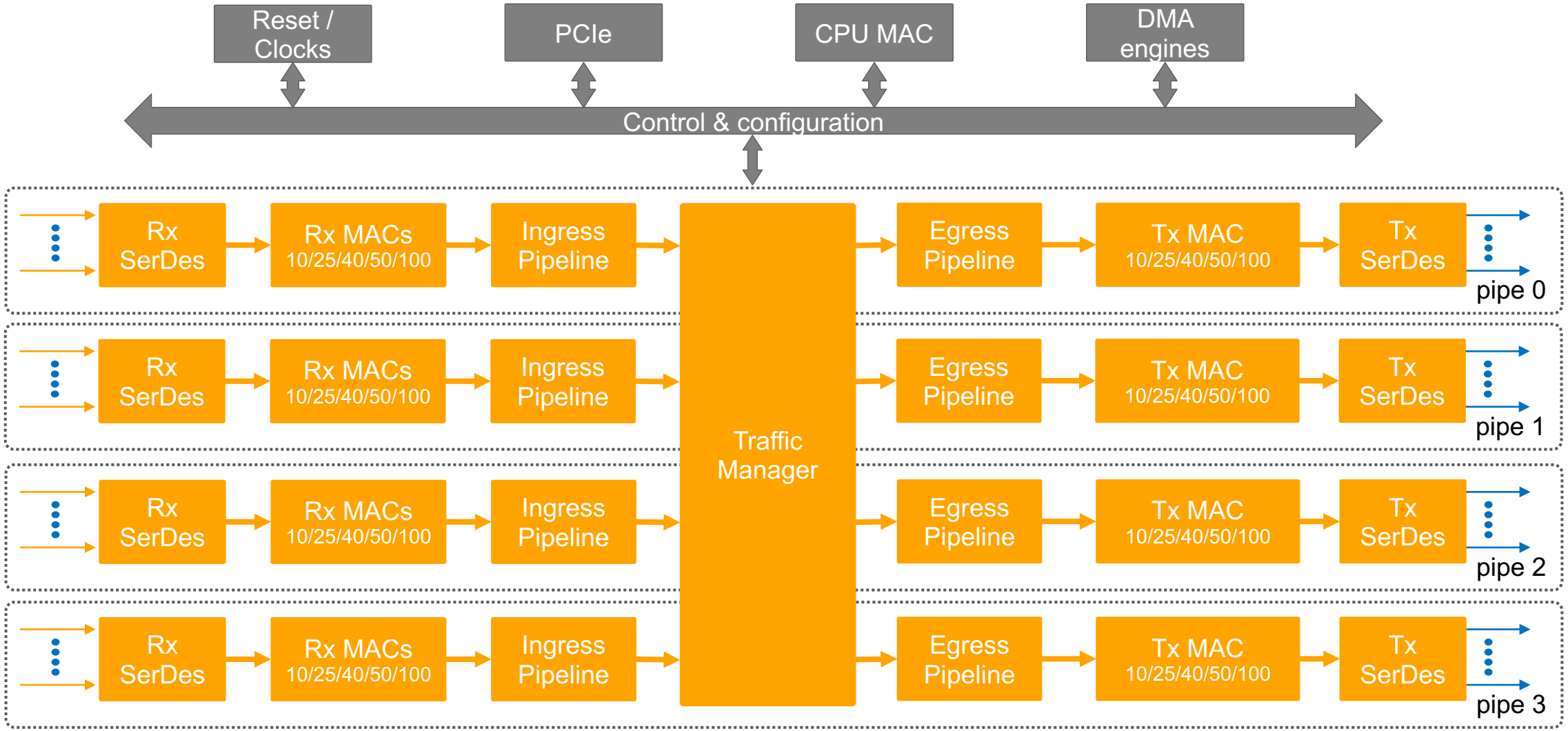
Arkadiy.Shapiro@intel.com

Programmable Data Plane Building Blocks

1. Open programming language & Community
2. Compiler & Chip Simulation Model
3. Visualization Tools
4. Flexible / Auto-generated APIs
5. Fully Programmable ASIC



Barefoot Tofino – Block Diagram



P4 vs Alternatives

P4	Alternatives
Large open community / ecosystem	Little to zero community
Target / architecture independent	Tied to specific ASIC
High level language / generic constructs	Constructs not re-usable
Designed for network pipeline programming	Designed for x86 applications and adapted to networking
Proven	Not proven

Barefoot Baremetal Switch Ecosystem

Switch OS / Fabric Solutions

ipinfusion™
SONiC
STRATUM
NoviFlow
stateless
kaloom

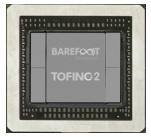
White Box Hardware (ODMs)

Edge-corE® NETWORKS
STORDIS
英業達 Inventec
恒为科技 EmbedWay
ufiSpace
DNI
Celestica™
WNC Wistron NeWeb Corp.

Barefoot ASICs

6.4 Tb/s

3.2 Tb/s, 2.4 Tb/s, 2.0 Tb/s, 1.8 Tb/s



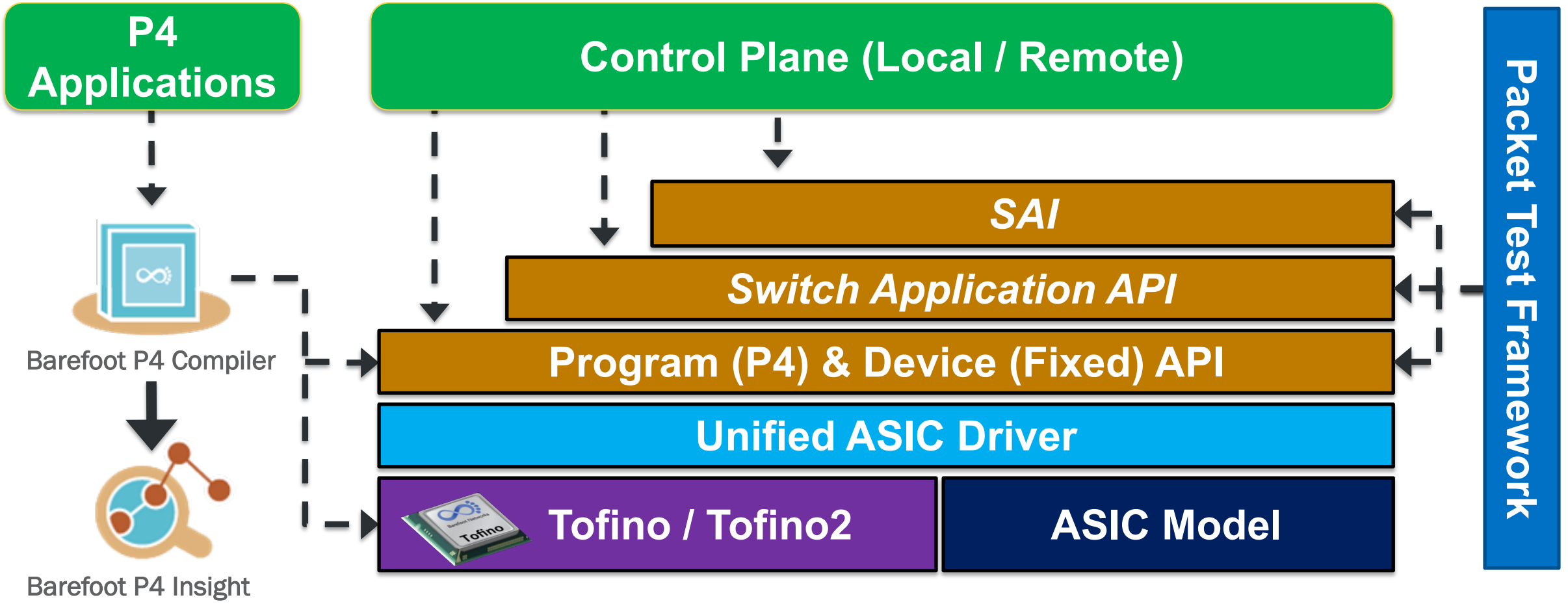
12.8 Tb/s

8.0 Tb/s, 6.4 Tb/s

Disaggregated Programmable Solution Building Blocks

ASIC	Tofino	Tofino2				
P4 Data Plane Provider	ODM	Barefoot	End Customer	3rd Party Consultant		
Control Plane / NOS	Stratum	SONiC	Commercial	Other / custom		
System Vendor (OEM / ODM)	Stordis	Edge-core	Inventec	Ufi Space	DNI	Embedway

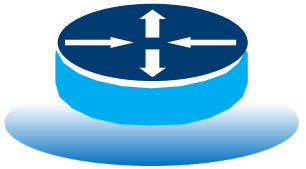
Control Plane Integration with Programmable Data Plane



Barefoot P4* Applications – Many More Use Cases

L4
Load-balancing

Real-time
Telemetry



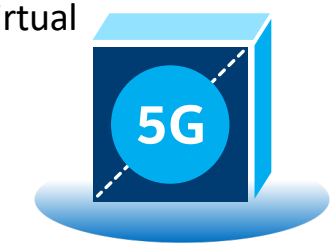
Enhanced
Routing

Enhanced
Switching

Security, DDoS
Detection

Physical
to Virtual

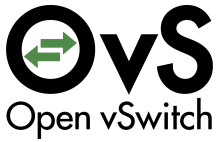
DNS Caching

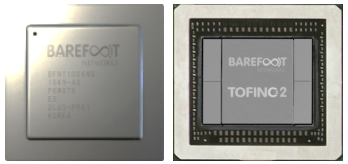


5G Gateway


```
table int_table {
  reads {
    ip.protocol;
  }
  actions {
    export_queue_latency;
  }
}
```

```
actionadd_header(int_header);
  modify_field(int_header.kind, TCP_OPTION_INT);
  modify_field(int_header.len, TCP_OPTION_INT_LEN);
  modify_field(int_header.sw_id, sw_id);
  modify_field(int_header.q_latency,
               intrinsic_metadata.deq_timedelta);
  add_to_field(tcp.dataOffset, 2);
  add_to_field(ipv4.totalLen, 8);
  subtract_from_field(ingress_metadata.tcpLength,
                      12);
}
 export_queue_latency (sw_id) {
```



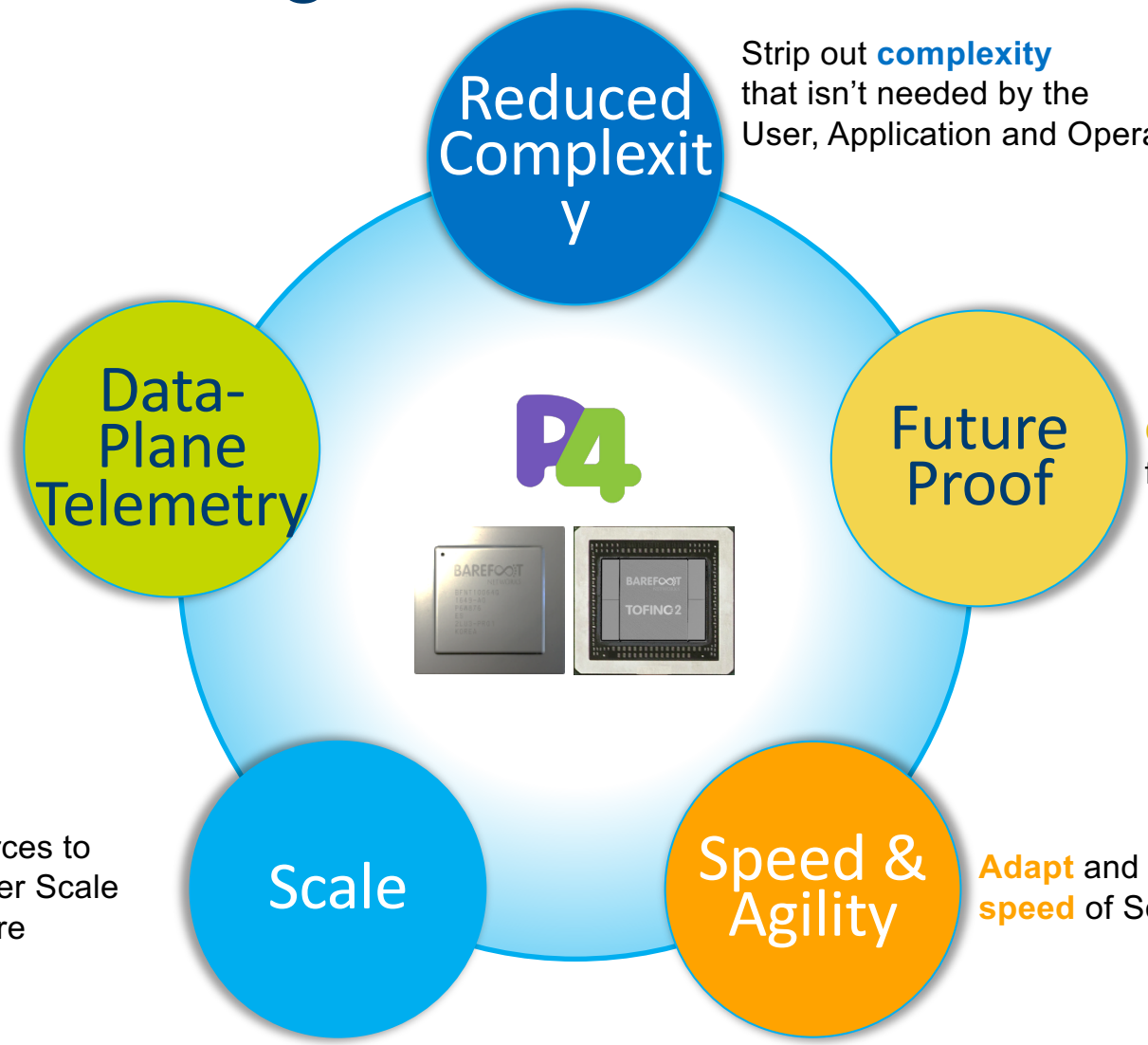


Tofino™ & Tofino™ 2

Bare Metal

Smart NIC

Benefits of P4* Programmable Switches



Strip out **complexity** that isn't needed by the User, Application and Operator

Continuously deliver and **evolve** features on the same Hardware

Adapt and innovate at **speed** of Software

Scale data-plane resources to match the needs of Hyper Scale and Service Infrastructure

Instrument the Data Plane with Barefoot **SPRINT**™ Smart, Programmable, Real-time, In-band Network Telemetry (INT)

Use-case: Flexible Scaling

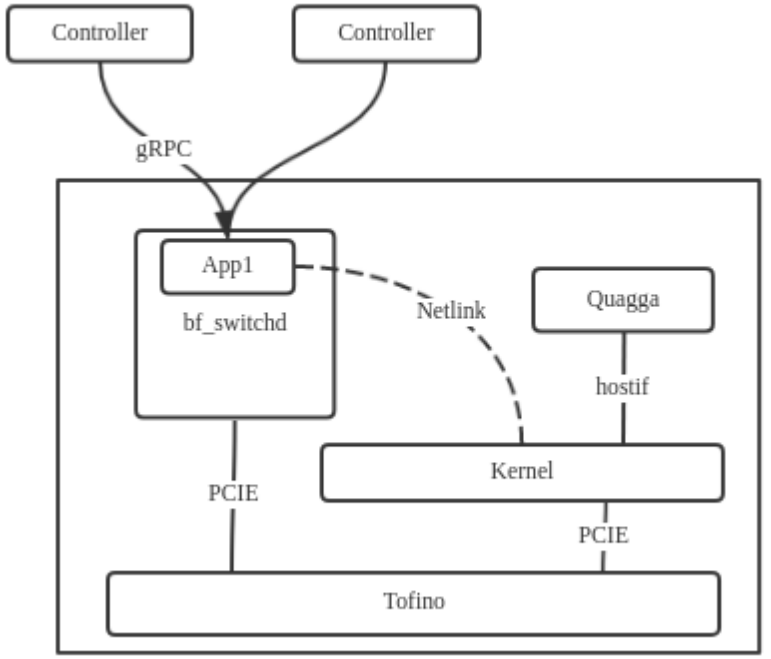
- Different table sizes for leaf and spine
- Different table sizes for different deployments
- Update with changing network conditions
- Example: IPv4 vs IPv6 heavy fabric

Parameter	**Scenario 1**	**Scenario 2**
IPv4 Host Local	Heavy	Light
IPv4 LPM	Light	Heavy
IPv6 Host	Heavy	Light
IPv6 LPM	Light	Heavy

Use-case: Cloud Gateway

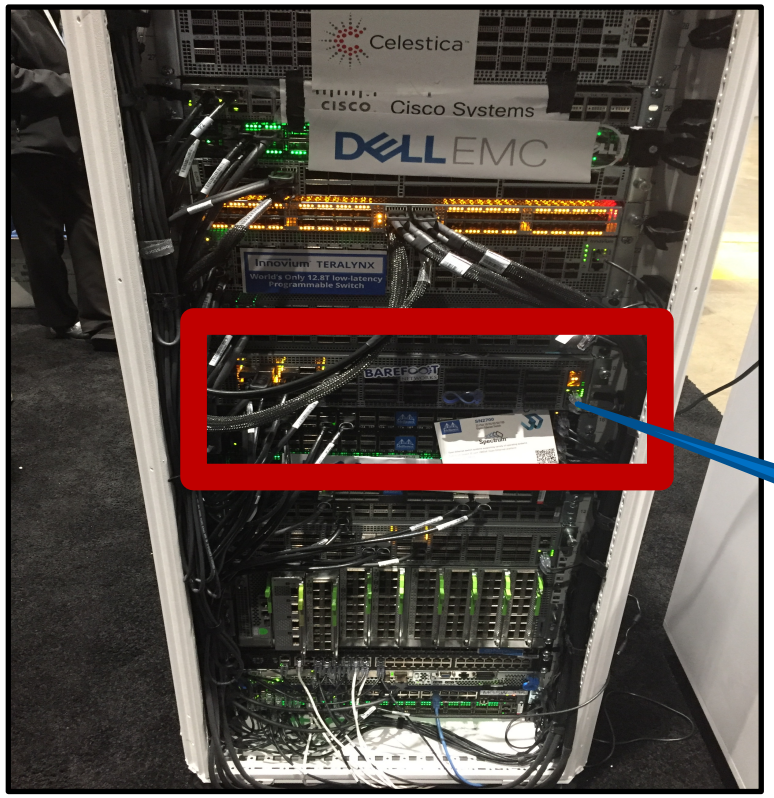
- **Details**
  - IPv6 GRE gateway solution
  - Migration from DPDK / OpenFlow
  - Whitebox switch + ONL
  - Custom P4-16 data plane, controlled via program APIs using GRPC
  - Custom SDN controller
- **Benefits**
  - Higher performance
  - Flexibility and control

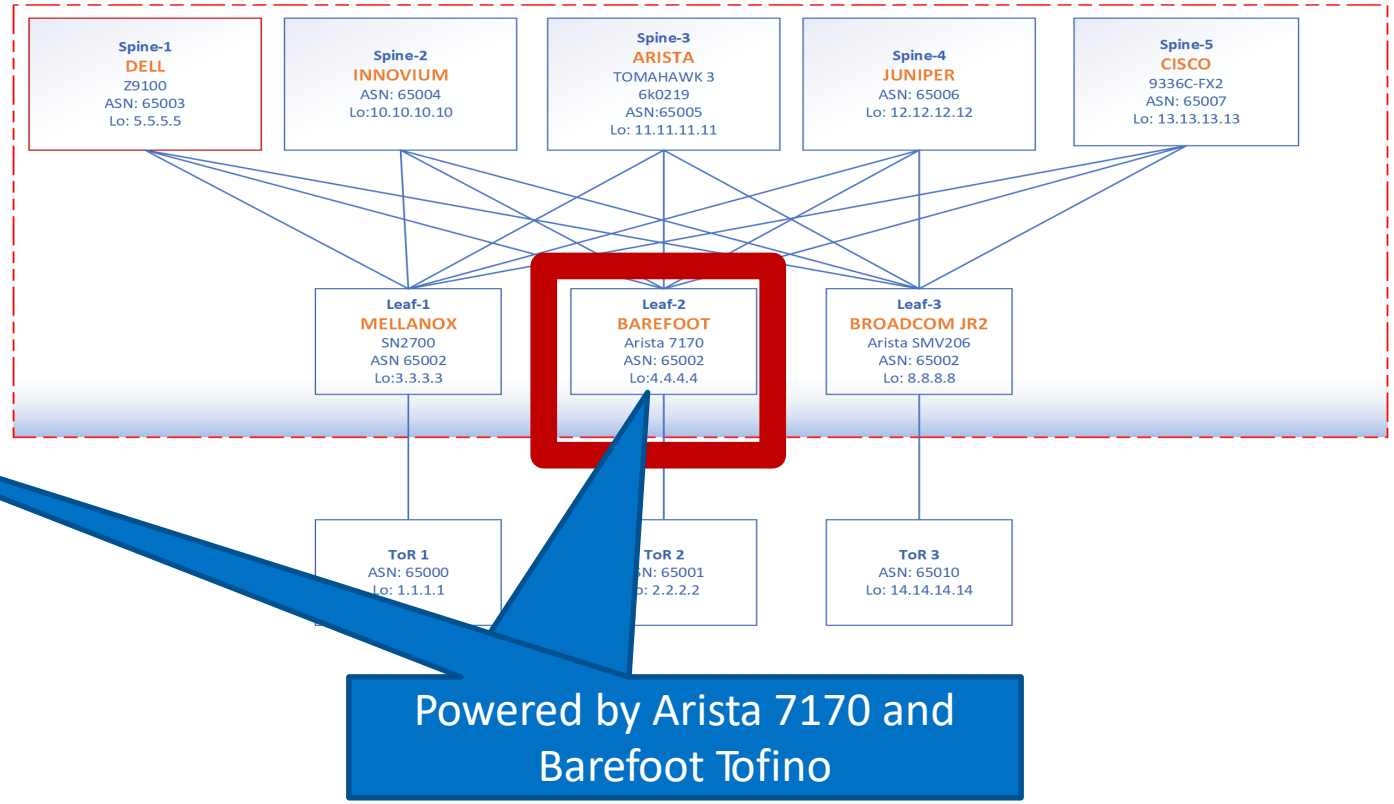
UCLOUD



Use-case: Baremetal Server Hosting with SONiC

Joint demo with Microsoft and Arista

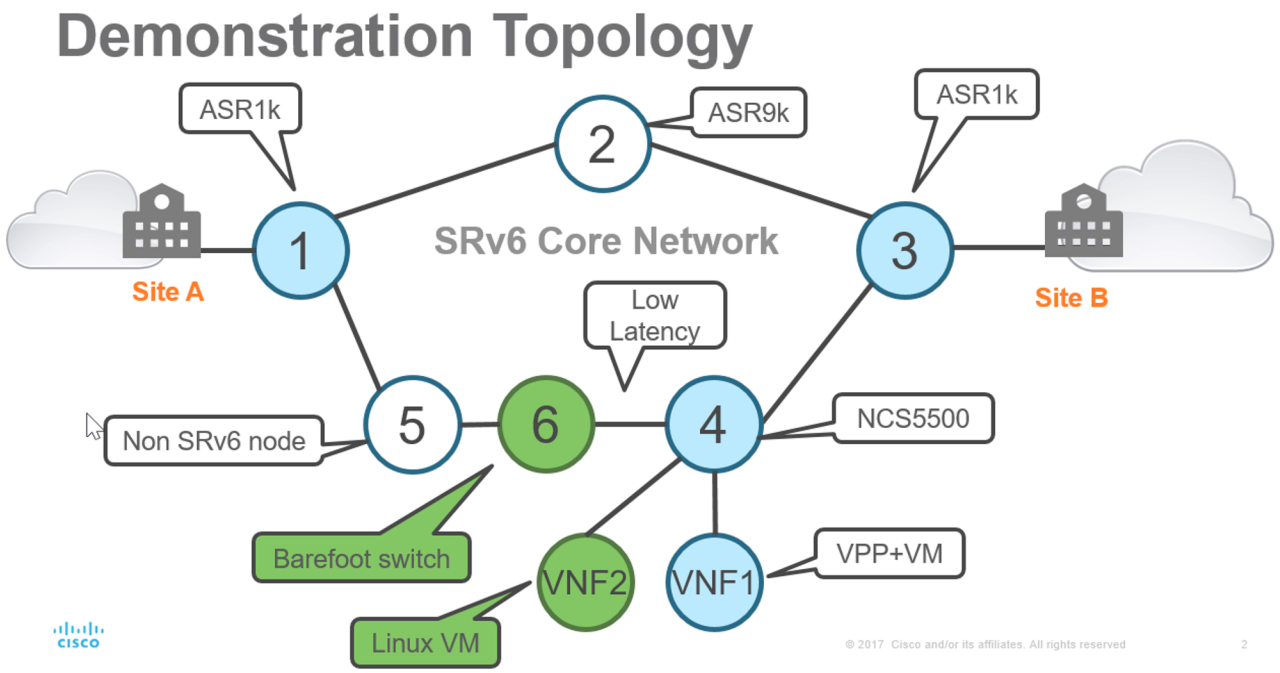




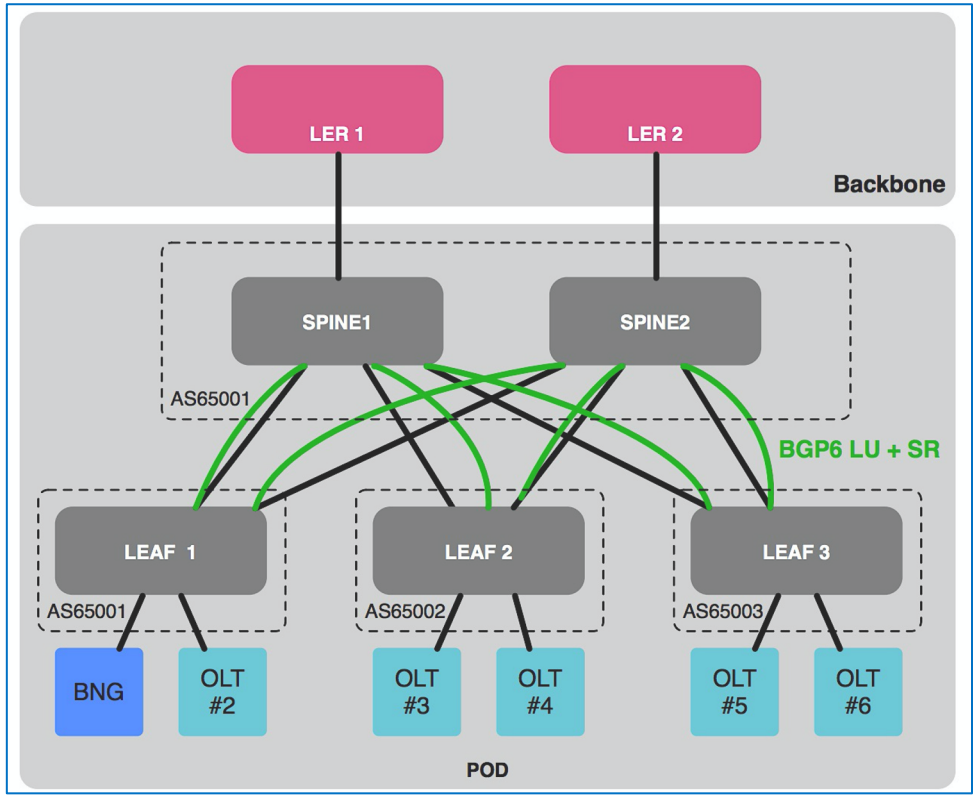
Use-case: Network Packet Brokers

- **Data plane & API Simplification**
- Header Stripping
- Packet Slicing
- Flexible Tunnel Encapsulation / Decapsulation
- Enhanced load-sharing (weighted, GTP correlation)
- Flexible TCAM table scaling for policies

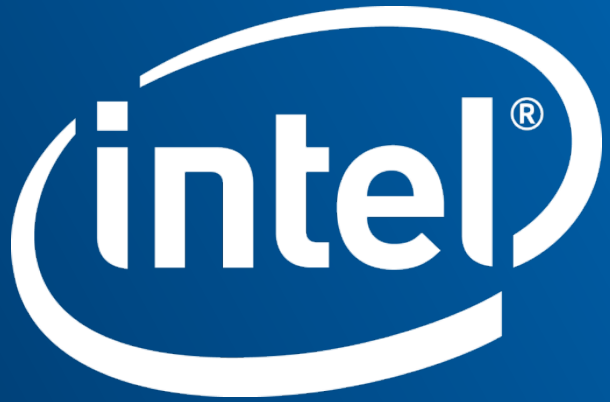
Use-case: Emerging Technologies



SRv6 endpoint and transit node functionality



Tofino performs large scale **PPPoE** session termination


intel®