Network Technology Watch update

Edoardo Martelli, Rolf Seuster

HOW2019 – March 2019

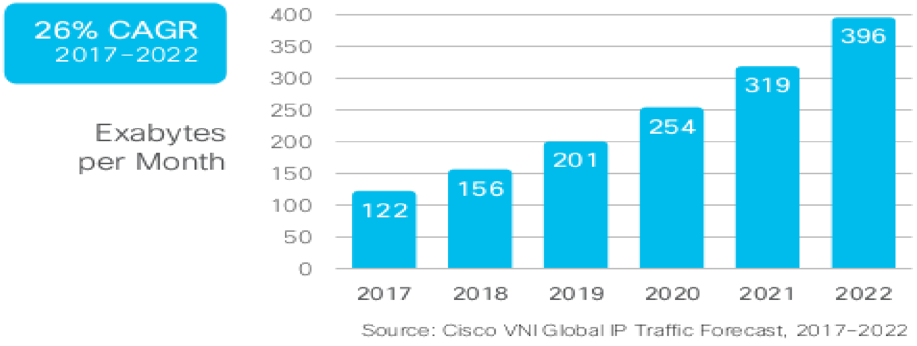
Content:

- HEPiX TechWatch Network sub-wg
    - Network technology watch

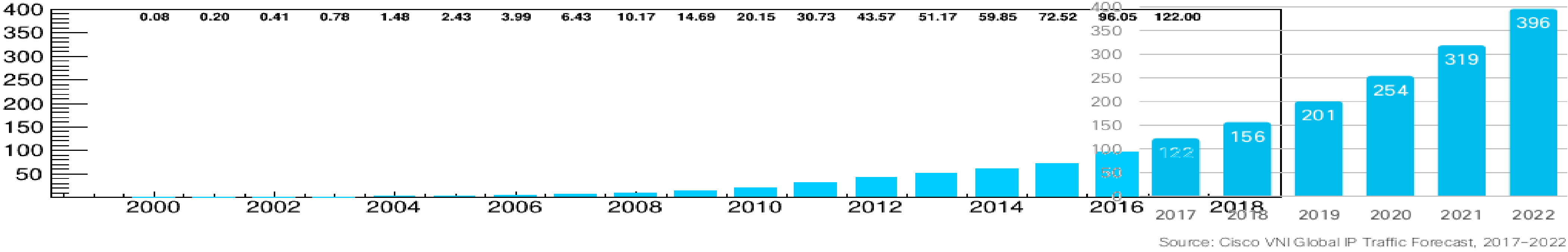
- HEPiX Network Function Virtualization sub-WG
    - Current activities

Introduction

- Internet traffic in next years expected to increase >3x until 2022 w.r.t. 2017
- one main driver 5G, together with its related computing needs (streaming movies etc.)



Introduction



- continues large increase year-by-year in last ~5years
  - down from 1.6x or 2x in early years ...
- technology so far has delivered (as has Moore's law) but the really free lunch is over
  - getting harder and harder to increase bandwidth, but there's still a few low hanging fruits …

→ Techwatch group !

HEPiX TechWatch Network sub-wg

Mission:

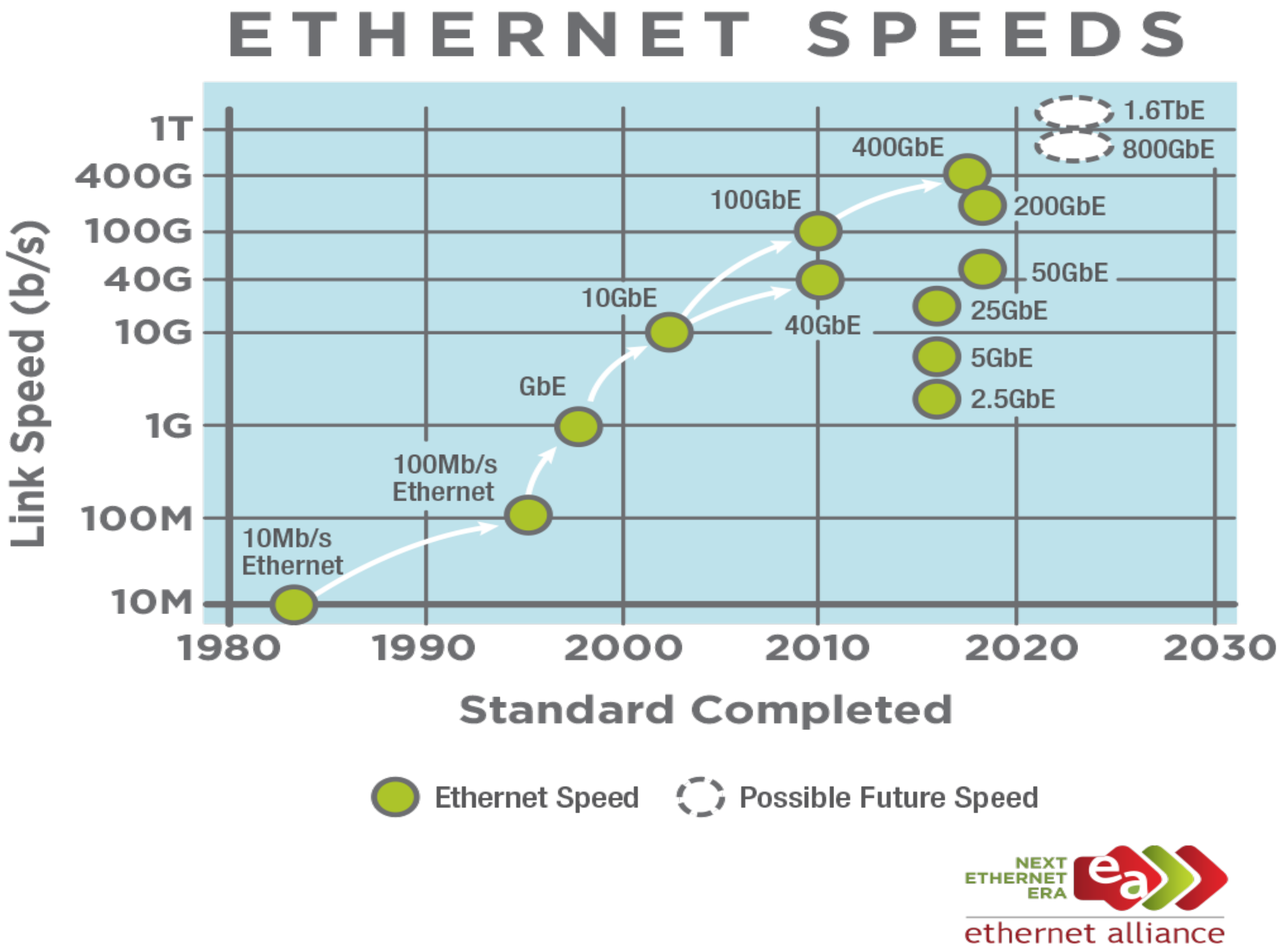
- follow technology trends
- analyse implications for HEP
- foreseen possible shortcoming

This sub-WG is part of the HEPiX Technology Watch activity

Links:

- Web site: http://w3.hepix.org/techwatch/network.html
- Twiki: https://twiki.cern.ch/twiki/bin/view/HEPIX/TechwatchNetwork/WebHome

Ethernet evolution



Ethernet in 2019

IEEE P802.3cn 50 Gb/s, 200 Gb/s, and 400 Gb/s over Single-Mode Fiber:
this fast-tracked project will leverage the PAM4 technology developed to support links at these rates currently up to 10km and build upon them to expand their reach to 40km.

IEEE P802.3cp Bidirectional 10 Gb/s, 25 Gb/s, and 50 Gb/s Optical Access PHYs:
this effort will develop bidirectional optical access PHYs for 10GbE, 25GbE, and 50GbE for point-to-point applications where the availability of fibers is limited. Wireless infrastructure is one of the key application spaces that this effort targets.

IEEE P802.3ct 100 Gb/s and 400 Gb/s over DWDM Systems:
this effort will see Ethernet evolve to support reaches up to 80km over a DWDM system. While the main drivers for this effort have been Multi-Service Operators (MSO) and Data Center Interconnect (DCI), it is easy to see how these solutions could be utilized for future mobile network aggregation and core backhaul.

100-200-400Gbps

The 100G Lambda Multisource Agreement (MSA) Group has released draft 2.0 of three specifications targeted to support 100-Gbps per wavelength transmission using PAM4 modulation.

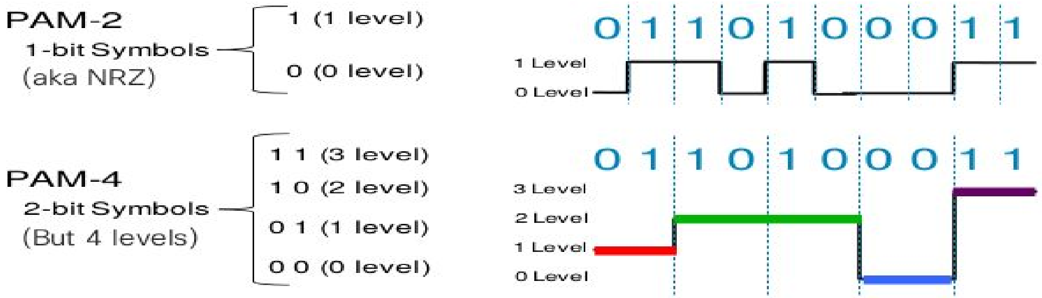
The draft specifications include **100G-FR and 100G-LR for 100 Gigabit Ethernet (GbE)** duplex single-mode fiber links over 2 km and 10 km, respectively, as well as the 400G-FR4 specification for 400GbE duplex single-mode fiber links in a 4x100G wavelength design.

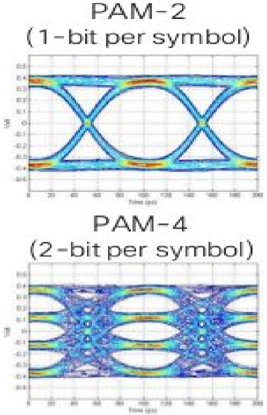
The MSA says it has begun work on a 400G-LR4 specification for a 10 km reach at 400GbE as well.

NRZ(PEM-2) vs PEM-4

- doubling information per transportation unit

Can fit into existing networks, but requires new endpoints

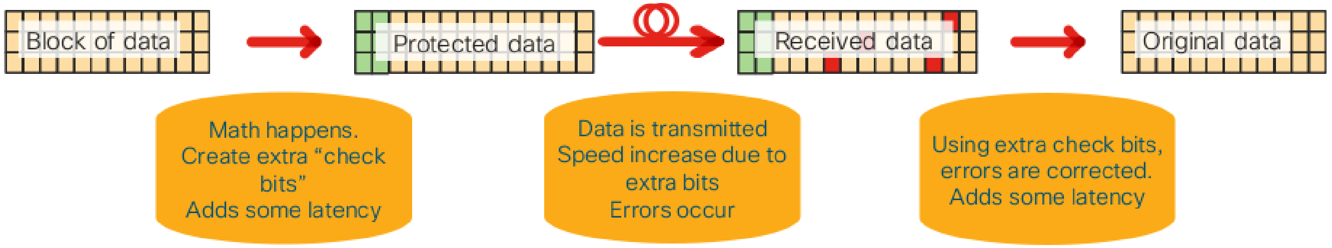




- doubles data rate on same hardware w.r.t NRZ
- reduces cost of optical equipment w.r.t. NRZ
- transmitters become more complex (de-mux)

Forward Error Correction

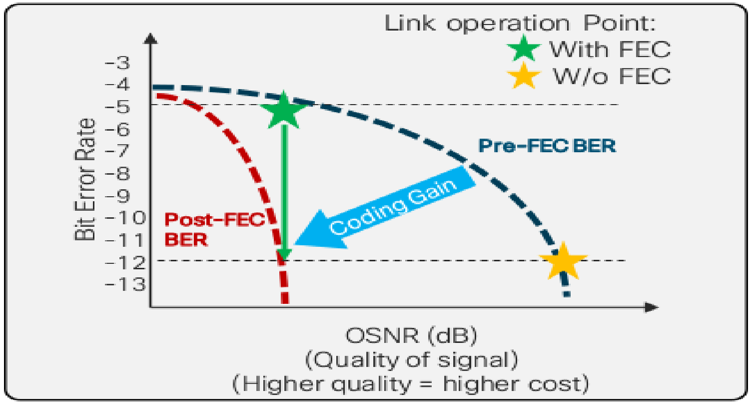
Adopted by clients optics starting @ 25 Gb/s and above



- inflates payload by N bits to protects M bits
  - e.g. 300 / 5140 [RS(544,514,10)] or less for other standards
- can use lower quality optical specs to reduce cost significantly
- requires less retransmits as long as all errors can be corrected
  → overall higher bandwidth
- can show 'cliff' in bandwidth if corrections fail due to too many transmission errors

FEC - Benefits

- added latency due to FEC is just O(100ns)



Usage of lower quality optical specifications **significantly reduces** cost and power of solutions

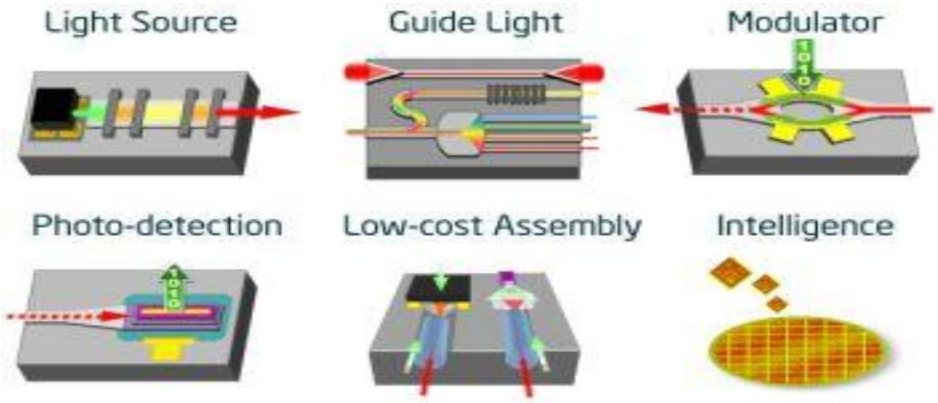
Different FEC algorithms can be used all with different performance properties

- Reed-Solomon: most common in Ethernet
- Higher performance FECs (e.g. used in Coherent optics)

- Incremental latency impact is dependent on implementation and data rate.
- For common Ethernet interfaces latency increase in range of ~50 to 100 ns (equivalent to time of flight over 5-10m of fiber)

- other FEC algorithms available that reduce overhead in size and/or latency
  - e.g. RS(272,257+1,10,7) in new 25G specs
    vs RS(544,514,10) as in 400G specs

Silicon Photonics



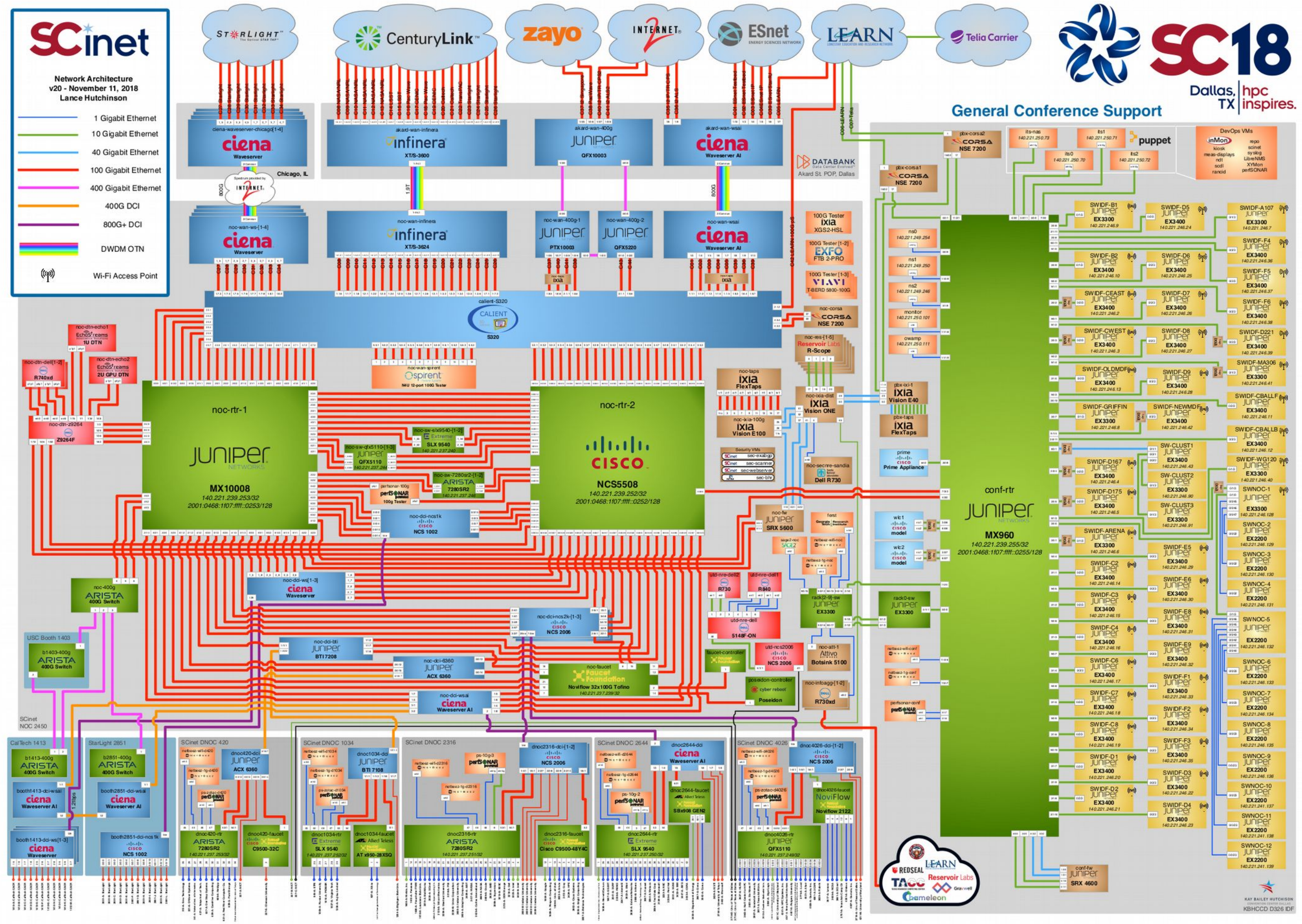
- advantages:
  - cheaper to produce
    - lower price will eventually trickle down to end users
    - almost all part on same die
      - allows also for more integration testing
    - manufacturing in existing CMOS production processes
  - lower power consumption than conventional transceivers
  - available since ~2 years:
    other manufacturers: Cisco(Luxtera), Intel, GlobalFoundries, IBM (research only?), …

Luxtera

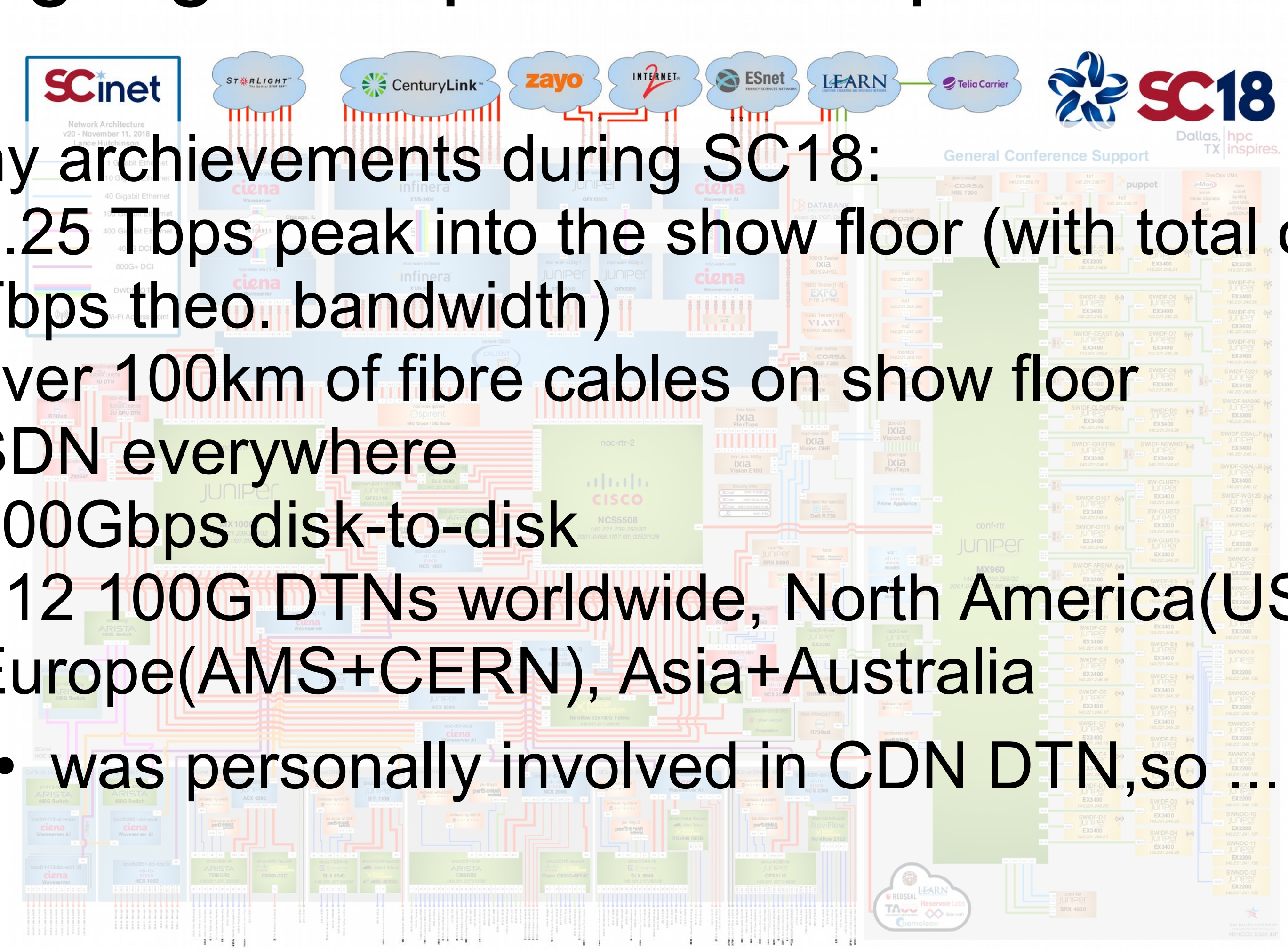
Intel

- see also e.g. https://community.mellanox.com/s/article/inside-the-silicon-photonics-transceiver

SC18 highlights + personal experience



SC18 highlights + personal experience



- many archievements during SC18:
  - 1.25 Tbps peak into the show floor (with total of 4.02 Tbps theo. bandwidth)
  - over 100km of fibre cables on show floor
  - SDN everywhere
  - 400Gbps disk-to-disk
  - ~12 100G DTNs worldwide, North America(US+CDN), Europe(AMS+CERN), Asia+Australia
    - was personally involved in CDN DTN,so ...

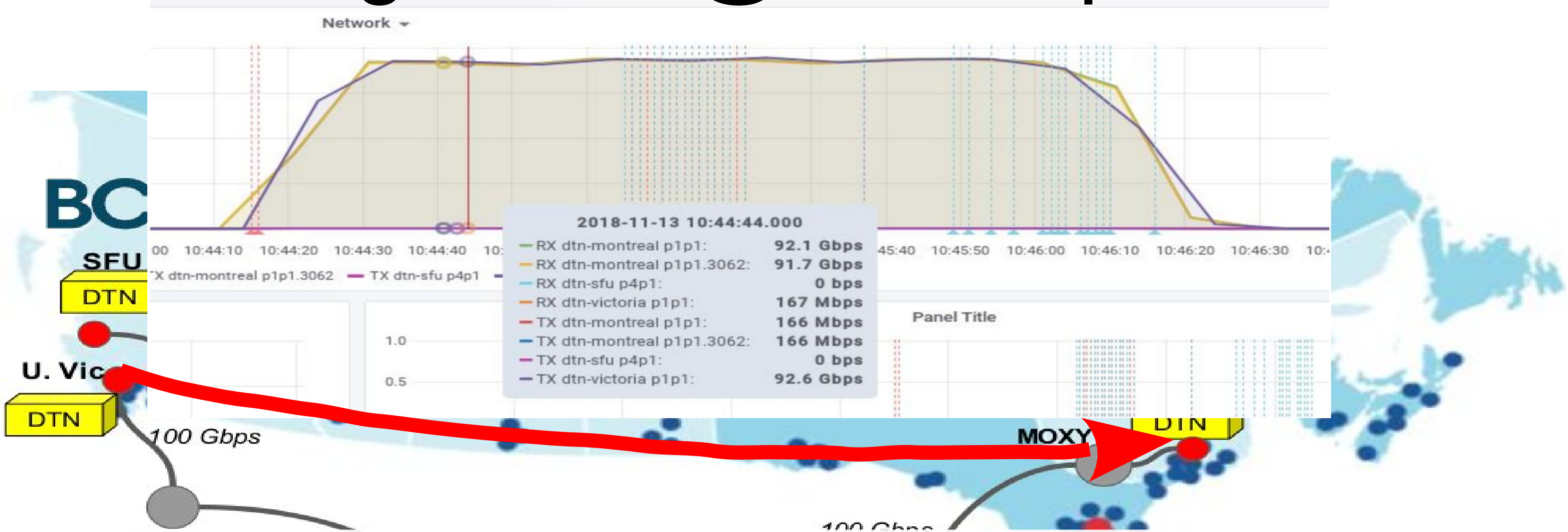
Canadian part in testing DTNs at SC18

attempt to do disk-to-disk at 100Gpbs with 'commodity' HW

results presented here are a combined collaborative effort between ComputeCanada and HepNetCanada, with lots of help from others !!

Network topology

SC18-testing DTNs @ 100Gbps inCanada



- reached >92Gbps over ~2 minutes, repeated transfers showed similar results
- my personal conclusion: on modern OS, apply manufacturer's drivers and tunes, apply moderate OS tuning will give already very good results, doing better is hard work (and often yields worse results !)
- Nice resource: from RedHat https://access.redhat.com/sites/default/files/attachments/20150325_network_performance_tuning.pdf
  from perfSonar/ESNet https://fasterdata.es.net/
  other's had more success, e.g. https://indico.cern.ch/event/676324/contributions/2967991/

Network Evolution Areas

The following are some of the key areas for HEP Networking R&D:

- **Improving efficiency of data transfers**
  - TCP BBR - version 2 is in the works with promising improvements
  - Exploring alternative protocols for transfers (UDP)
- **Caching**
  - Data caches co-located with network hubs in a similar way as on commercial CDNs
- **Federations/Clouds**
  - Overlay networks spanning multiple domains
  - Multi-clouds - expanding DC networking via L3VPNs
- **Technology**
  - SDN/NFV approaches - currently looked at by HEPiX NFV WG
    - Compute - Agile service delivery on Cloud Infrastructures (OpenStack, Kubernetes)
    - Data Transfers - Network resource optimisation - dynamically optimising the network based on its load and state (more in Shawn/Ilija)
  - SD-WAN approaches - https://www.mode.net/

Other things to look out

HEPiX

- What to expect from Mellanox/NVidia deal?
  Nvidia paid $6.9b for Mellanox (~$1b revenue)
  - 3 CPU/GPU(FPGA) interconnects now:
    - NVlink / OpenCAPI (Coherent Accelerator Processor Interface)
      AMD, IBM, Google, Micron, Mellanox (Nvidia as ‘contributor’)
    - CCIX (Cache Coherent Interconnect for Accelerators)
      AMD, ARM, IBM, Qualcomm, Xilinx, Huawei, Mellanox
    - CXL (Compute Express Link)
      Cisco, Dell EMC, Facebook, Google, HPE, Huawei, Intel, Microsoft
- Remote DMA (RDMA) with all its flavours
  - RoCE RDMA over Converged Ethernet, iWARP, ...

Ethernet interfaces

EMERGING INTERFACES AND NOMENCLATURE

	Electrical Interface	Backplane	Twinax Cable	Twisted Pair (1 Pair)	Twisted Pair (4 Pair)	MMF	500m PSM4	2km SMF	10km SMF	20km SMF	40km SMF	80km SMF
10BASE-		T1S		T1S/T1L								
100BASE-				T1								
1000BASE-				T1	T							
2.5GBASE-		KX		T1	T							
5GBASE-		KR		T1	T							
10GBASE-				T1	T				BIDI Access	BIDI Access	BIDI Access	
25GBASE-	25GAUI	KR	CR/CR-S		T	SR			LR/ EPON/ BIDI Access	EPON/ BIDI Access	ER/ BIDI Access	
40GBASE-	XLAUI	KR4	CR4		T	SR4/eSR4	PSM4	FR	LR4			
50GBASE-	LAUI-2/50GAUI-2 50GAUI-1	KR	CR			SR		FR	EPON/ BIDI Access LR	EPON/ BIDI Access	BIDI Access ER	
100GBASE-	CAUI-10 CAUI-4/100GAUI-4 100GAUI-2 100GAUI-1	KR4 KR2 KR1	CR10 CR4 CR2 CR1			SR10 SR4 SR2	PSM4 DR	10X10 CWDM4/ CLR4 100G-FR	LR4/ 4WDM-10 100G-LR	4WDM-20	ER4/ 4WDM-40	ZR
200GBASE-	200GAUI-4 200GAUI-2	KR4 KR2	CR4 CR2			SR4	DR4	FR4	LR4		ER4	
400GBASE-	400GAUI-16 400GAUI- 8 400GAUI-4	KR4	CR4			SR16 SR8/SR4.2	DR4	FR8 400G-FR4	LR8 400G-LR4		ER8	ZR

Gray Text = IEEE Standard Red Text = In Standardization Green Text = In Study Group
Blue Text = Non-IEEE standard but complies to IEEE electrical interfaces

ethernet alliance

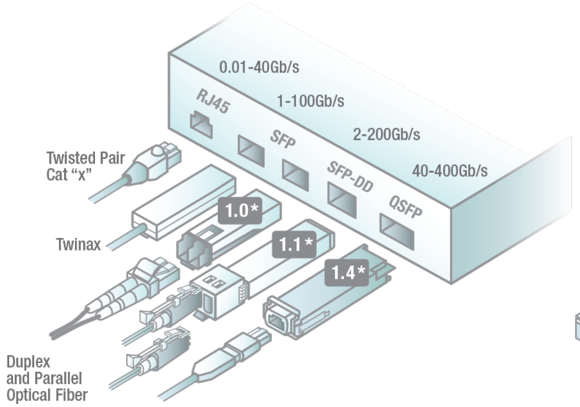
Form factors

FORM FACTORS

This diagram shows the most common form factors used in Ethernet ports. Hundreds of millions of RJ45 ports are sold a year while tens of millions of SFP and millions of QSFP ports ship a year.

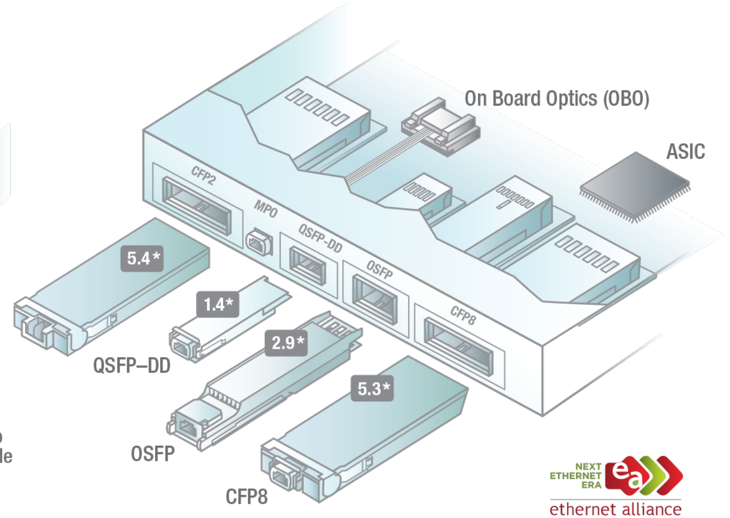
This diagram shows new form factors initially designed for 100GbE and 400GbE Ethernet ports. All have 4 or 8 lanes and the OBO has up to 16 lanes. The power consumption of the modules is proportional to the surface area of the module.

1–4 Lane Interfaces



*Square inches of top surface of the module

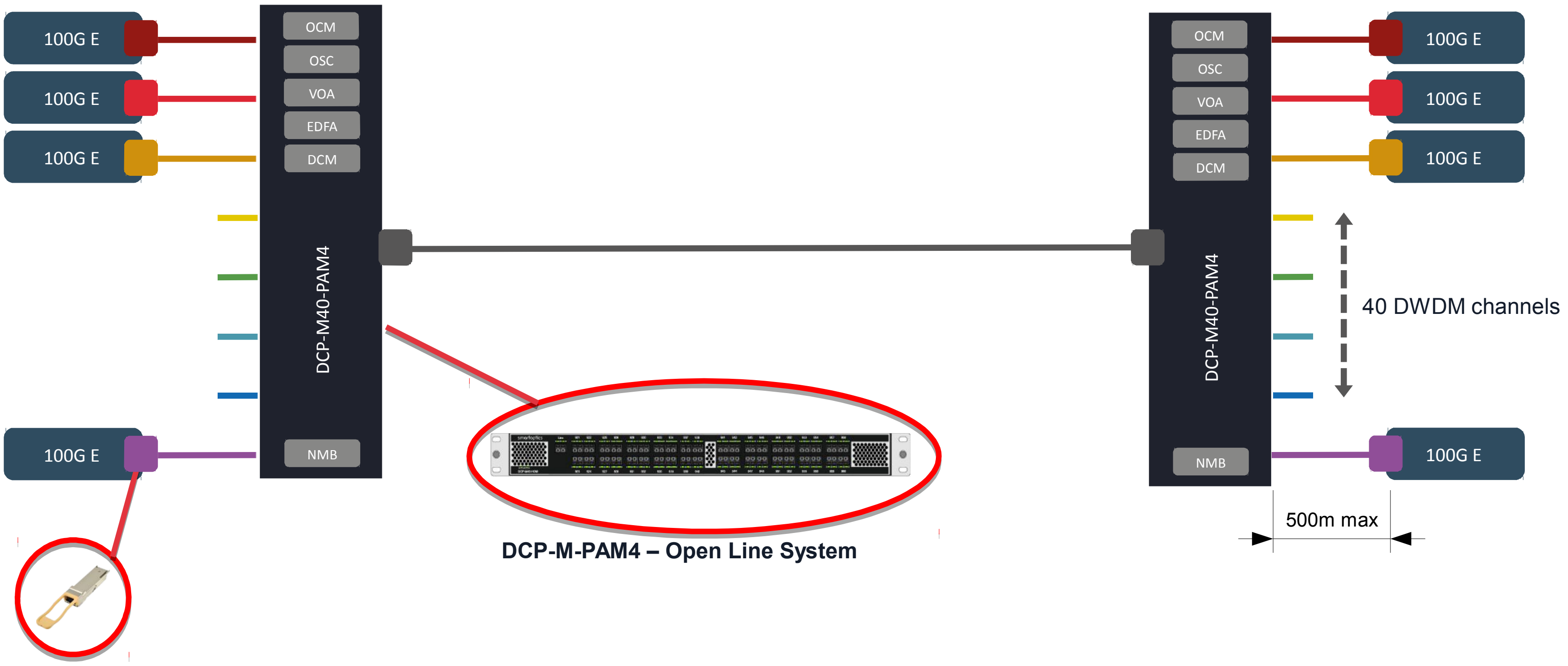
4-16 Lane Interfaces



NEXT ETHERNET ERA ethernet alliance

DCI with DWDM PAM4

100G p-t-p embedded over <80km distance based on a **cost effective solution**



DCP-M-PAM4 – Open Line System

PAM4 QSFP28 DWDM TRX
Embedded in to 100G switch

[Source: Smartoptics]

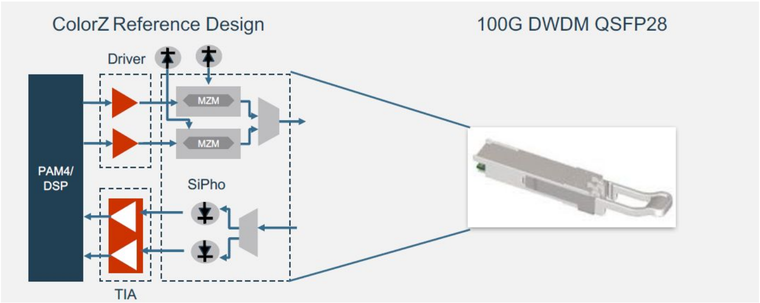
100G PAM4 DWDM QSFP28

100G DWDM Transceiver

- 2 wavelengths (lanes) on a 50GHz grid
- Transceiver output power -11 dBm
- Minimum required input power -2 dBm
- Dispersion tolerance +/-6 km on G.652 fiber
- High OSNR required (>31 dB)

Requires an active line system to address these parameters

Can be plugged directly into standard switches and NICs



[Source: Smartoptics]

HEPiX Network Function Virtualization sub-wg update

On behalf of Marian Babik and Shawn McKee

Network Function Virtualization WG

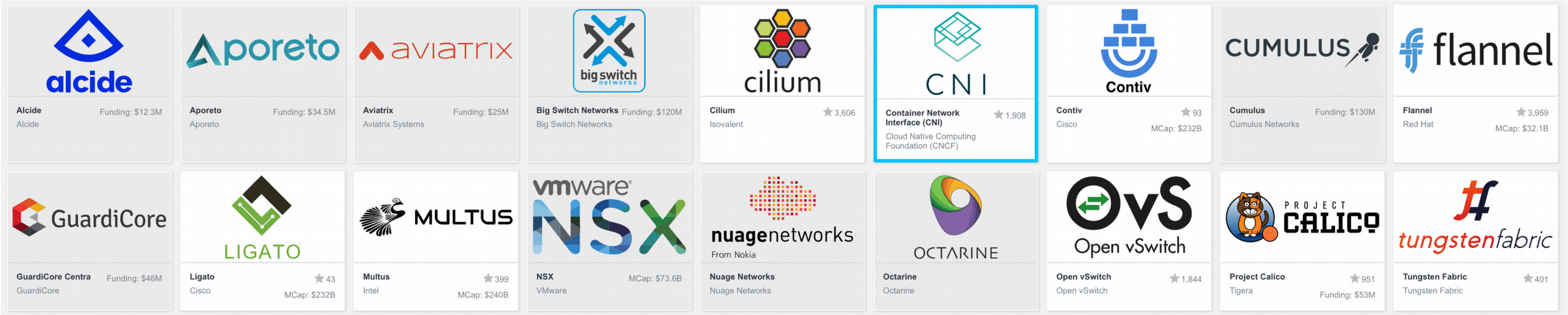
Mandate: Identify use cases, survey existing approaches and evaluate whether and how SDN/NFV should be deployed in HEP.

Team: 60 members including R&Es (GEANT, ESNet, Internet2, AARNet, Canarie, SURFNet, GARR, JISC, RENATER, NORDUnet), sites (ASGC, PIC, BNL, CNAF, CERN, KISTI, KIAE, FIU, AGLT2, Caltech, DESY, IHEP, Nikhef, IN2P3)

Monthly meetings (https://indico.cern.ch/category/10031/)
Mailing list: https://listserv.in2p3.fr/cgi-bin/wa?SUBED1=hepix-nfv-wg

Cloud Native Networking

- One of the WG objectives is to explore SDN/NFV approaches for **Cloud Compute -** SDN solutions for **data centres** supporting OpenStack/Kubernetes/OpenShift, etc.
- A range of approaches, both open-source (white) and commercial (grey) exist, tracked by the **Cloud Native Computing Foundation** (http://l.cncf.io)



- Current focus is on **Open vSwitch/Open Virtual Networking** and **Tungsten Fabric -** both open source, best match to our use cases – for comparison see OVS talk by Y. Yang
- Existing **Tungsten Fabric** pilot deployments at CERN and Nikhef
- Experiences with commercial deployments are reported on regular basis at HEPiX by IHEP

LF Networking and SmartNICs

- Additional projects that improve SDN/NFV performance, provide alternative SDN controllers, offer programmable off-loading capabilities are hosted by **Linux Foundation**

- **SmartNICs** - now offered from multiple vendors - make it possible to **maximise capacity while providing full programmability** for virtual switching and routing, tunnelling (VXLAN, MPLS), ACLs and security groups, congestion and tail latency reduction, etc.
  - FPGA-based SmartNICs broadly deployed in Microsoft Azure
  - Good overview provided in this ACM SIGARCH article

Programmable Networks

- Another objective of the WG is to explore SDN/NFV approaches for distributed storage, end-to-end transfers (e.g. data lakes) and HPCs.
  We envision an operational model that coordinates use of network, CPU and storage resources, as it also seeks to improve workflow
- Number of interesting projects exist in this area:
  - **SDN for End-to-End Networked Science at Exascale (SENSE)**
    - Pioneering new paradigm in application to network interactions
  - **BigData Express** - high performance data transfer engine with on-demand provisioning of network paths
  - Existing project also in **ATLAS** (SDN btw AGLT2/MWT2/KIT), SDN aspects also in NSF-funded **SLATE**, **OSIRIS** and CERN's **NOTED** project

- WG also exploring **R&E activities and plans** : AMLight, ESNet6, GEANT
  - Important as it'll impact both future capacity and network capabilities !

Plans

- Different projects currently ongoing related to SDN/NFV
    - Not only within HEP, but also in other domains
- WLCG DOMA has network area listed as one its priorities
    - There are several use cases being discussed within DOMA activities that would fit well with SDN/NFV (service chaining, VPN provisioning, traffic engineering, etc.)
- Bridging the gap between network and experiments groups important to ensure that the new technologies will be adopted
    - Surveying existing approaches/activities important to help them gain visibility
    - Identifying existing testbeds and how they can be accessed would help plan/test prototypes
    - Help to gain visibility into what is currently possible with SDN/NFV and how it can impact existing workflows
- We're currently working on a white paper that should help us engage the experiments, agree on terminology and have a focused discussion on what we'd like to do in the future.
- Always looking for feedback and additional volunteers/sites for help.

Conclusion

Summary

Network Techwatch:
- 100Gbps becoming commodity, but too many variants
- 400Gbps is out, QSFP-DD seems preferred
- PAM4 and FEC will drive higher speeds and port density
- SC18 demonstrated maturity of 100Gbps servers

HEPiX NFV WG:
- Exploring applications of SDN and NFV to improve datacentre efficiency
- Preparing white paper to help HEP community understand and adopt SDN/NFV

References

TechWatch Networking sub-WG

- Web site: http://w3.hepix.org/techwatch/network.html
- Twiki: https://twiki.cern.ch/twiki/bin/view/HEPIX/TechwatchNetwork/WebHome
- Ethernet Roadmap: https://ethernetalliance.org/the-2019-ethernet-roadmap/

NFV working group:

- WG meetings and notes: https://indico.cern.ch/category/10031/
- F2F meeting: https://indico.cern.ch/event/725706/
- WP: https://docs.google.com/document/d/1w7XUPxE23DJXn--j-M3KvXIfXHUnYgsVUhBpKFjyjUQ/edit?usp=sharing