TOMORROW starts here.

Evolution of Core Routing Hardware and Software

BRKSPG-2640

LJ Wobker, CCIE 5020

Technical Marketing & Systems Architecture

High-End Routing & Optical Group

Expectations, Level-setting, and Plea for Forgiveness...

- Apologies for the accent.
  - Throw something when I go to fast.
- Primarily a discussion of problems and technology
  - but I will talk a little about product.
  - biased towards the big routing stuff (NCS-6000, ASR-9000)
  - no configs, no troubleshooting, no specific, actionable data
  - but I hope you walk out knowing a lot more than when you walked in.
- Too much to cover in 90 minutes!
  - World of solutions, meet the engineer, get my email...
  - I will try for ~15 min of Q&A at the end – but INTERRUPT ME!

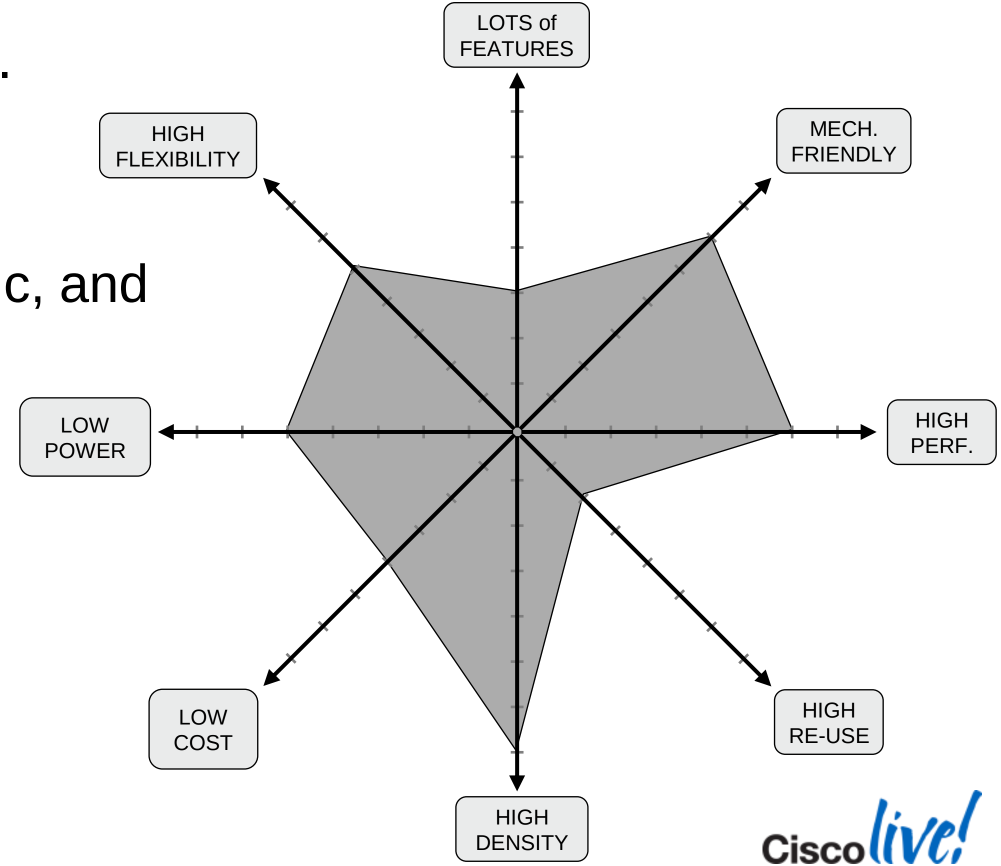
Agenda

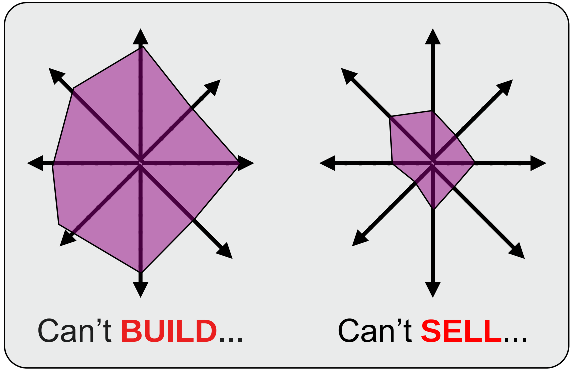
- **High End Routing Challenges**
- **Software**
  - (virtualised IOS XR)
- **Forwarding Silicon / Network Processors**
  - (Cisco nPower)
- **Mechanical / Thermal / Power**
  - (Cisco NCS-6008)
- **Optical density & flexibility**
  - (Cisco CPAK)
- **Optical reach, manageability, and granularity**
  - (Cisco nLight)

Challenges in High End Routing Systems

Router Engineering : The Greatest Game...

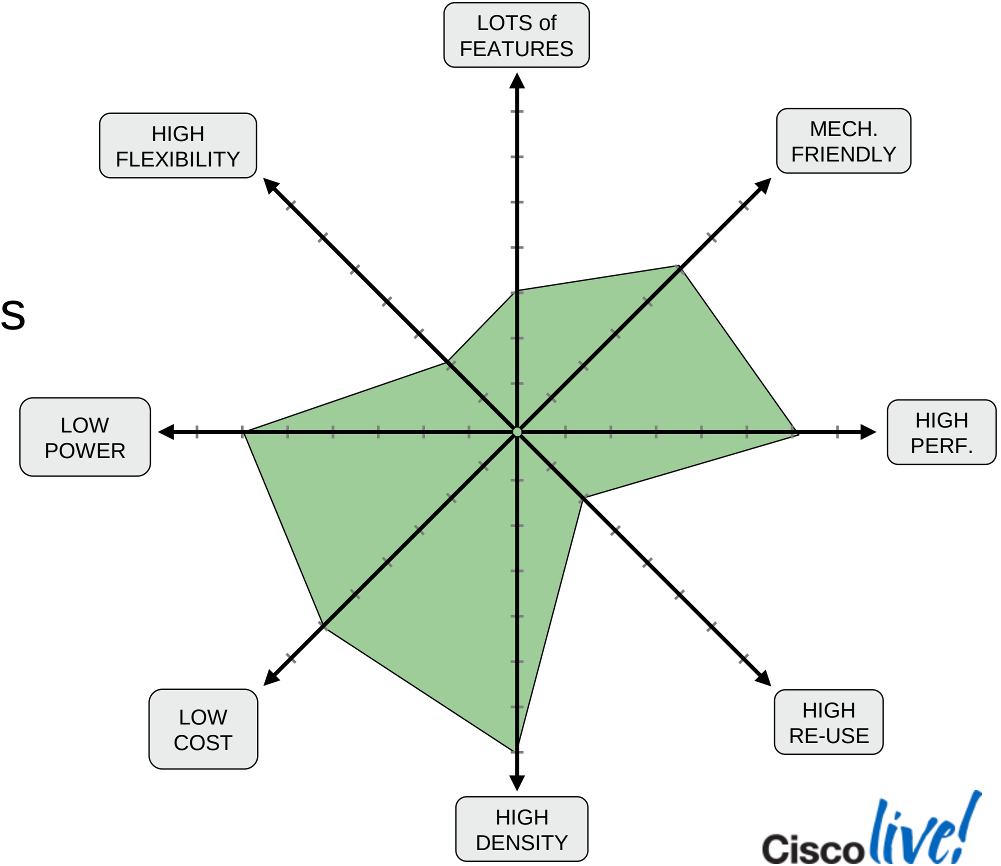
- So, you want to design a router...
- Ask: "Optimise for everything"
- Reality: lots of physical, economic, and temporal constraints





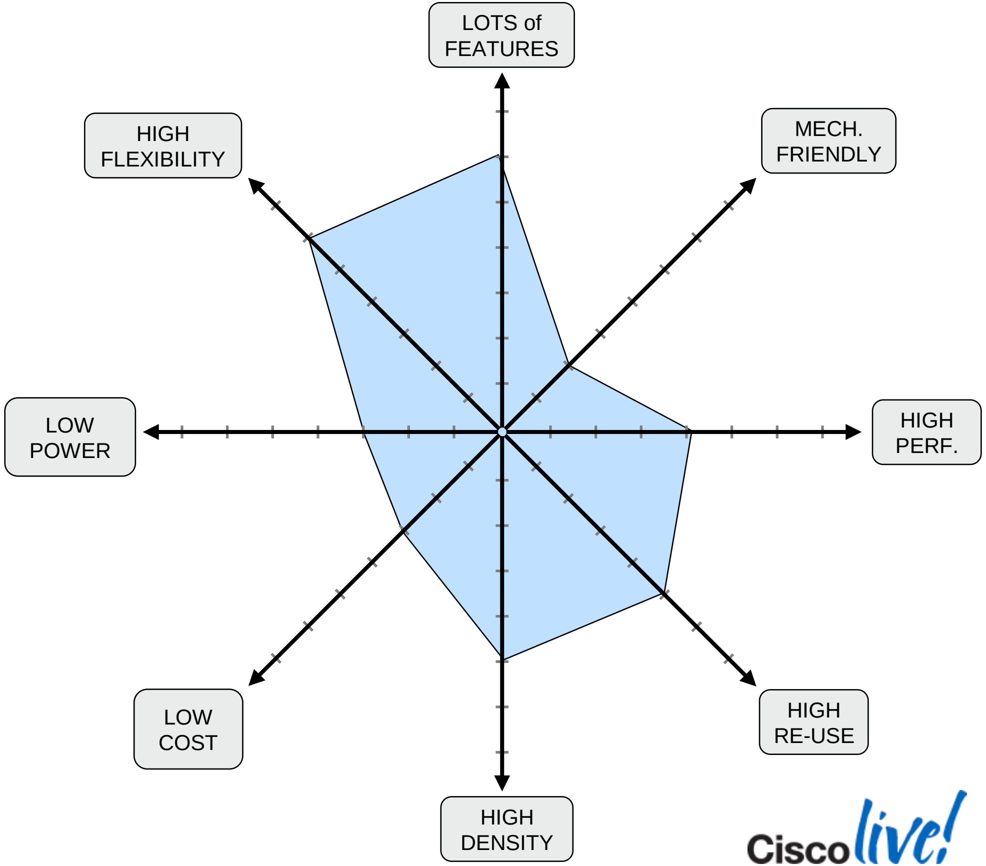
Router Engineering : Small Fixed L3 Switch

- Let's look at something like a Nexus 3k switch...
- Ripping fast for the features it has
- All ASIC: no adding features!
- One-time shot (no re-use!)



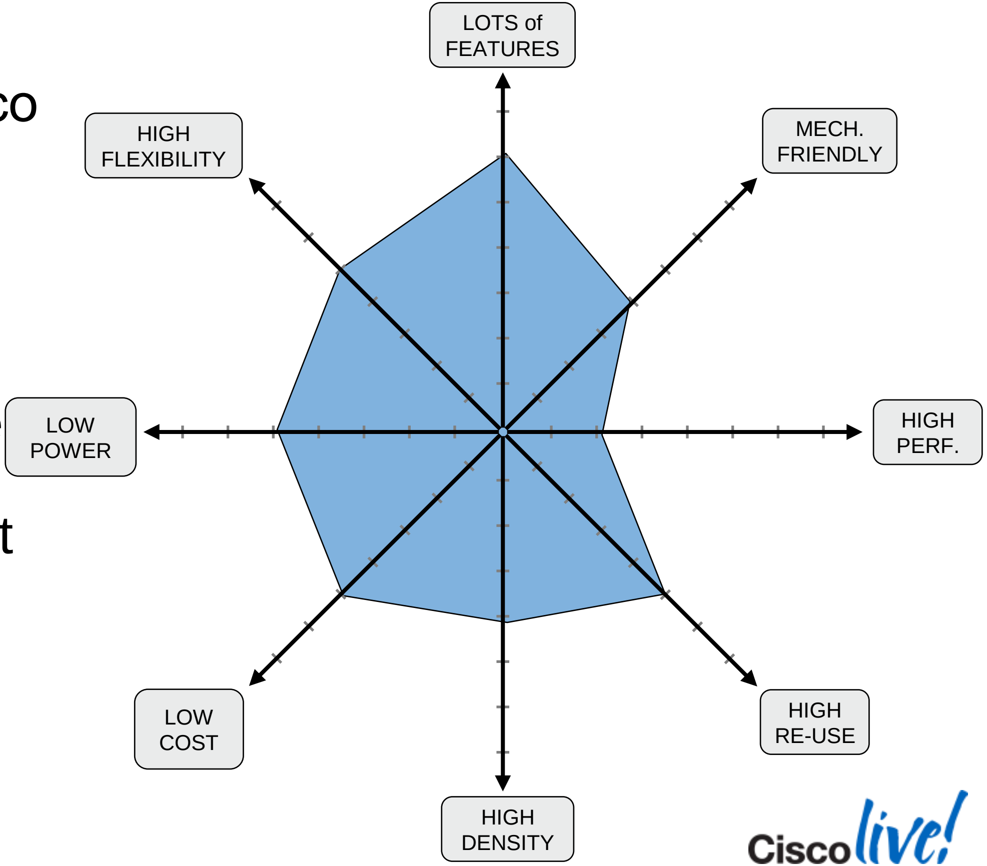
Router Engineering : Big Modular Routing System

- Let’s look at something like a NCS-6000 or ASR-9000...
- Dense, very flexible (expansion), lots of features
- Big, loud, heavy...



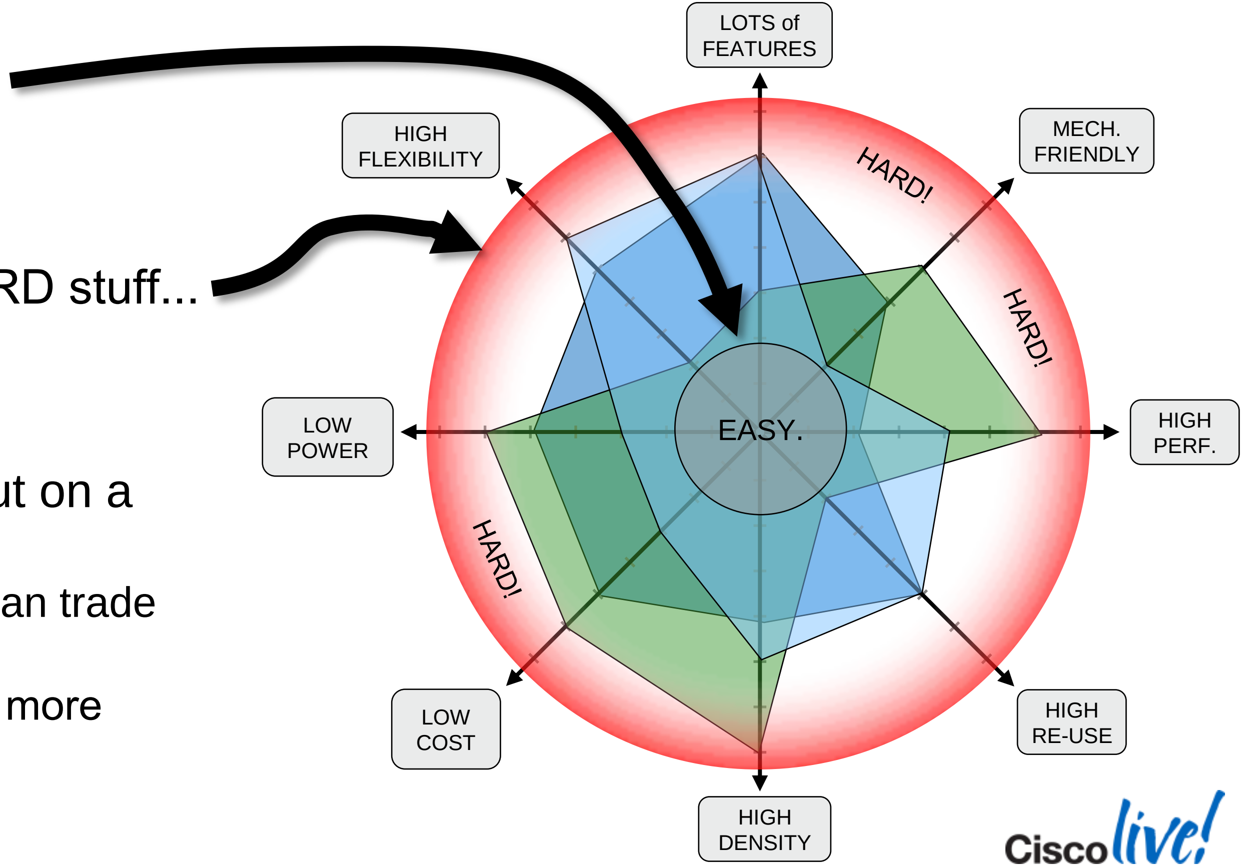
Router Engineering : Mid-Range/Access System

- Let's look at something like a cisco 3800...
- Flexibility & features
- Not particularly high performance
- Takes up a lot of space for what it does **for any one user**



Router Engineering : The Aggregate...

- There’s the easy stuff...
- And then there’s the HARD stuff...
- The cost to get further out on a line is NOT constant...
  - close to the middle, you can trade something almost 1:1
  - out at the ends, it’s much more expensive.



Software –Virtualised IOS XR

Software Infrastructure Goals

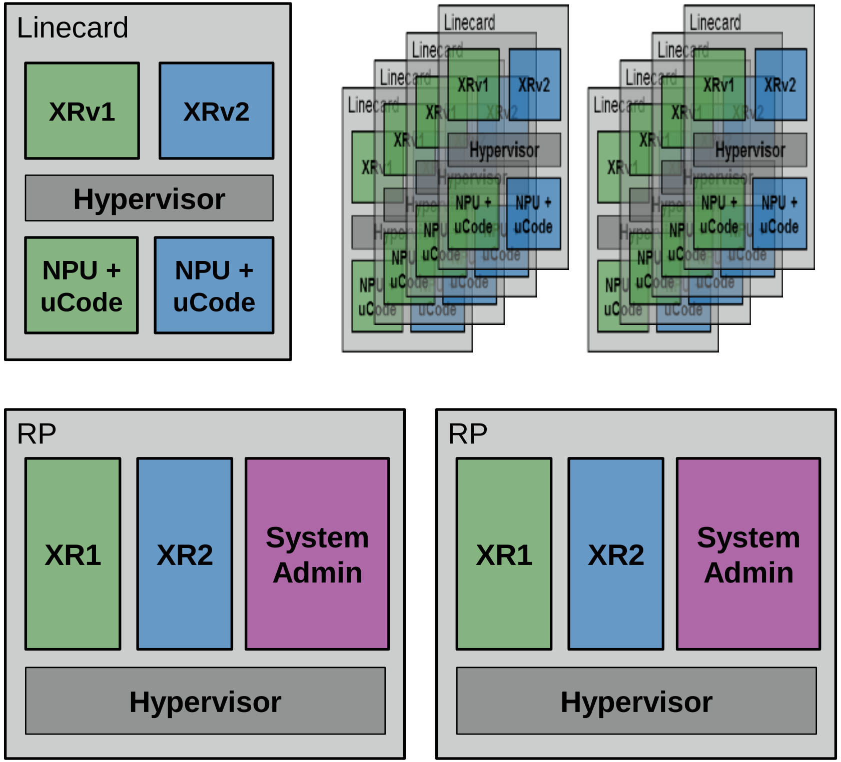
- Scale & Performance
- Virtualisation
- Availability
- Flexibility
- Security

Software Scale and Performance

- Scale can run into **tens of millions** of prefixes
- Millions of MAC addresses (which must be learned **much** faster)
- Distributed systems can be **hundreds of GB** of memory
- Must boot, converge, and absorb configurations **exceptionally** fast
- Solutions:
  - Good code, good hardware, and **lots** of architectural choices
    - Data structure choice, lots of CPU and memory, smart hardware interconnections

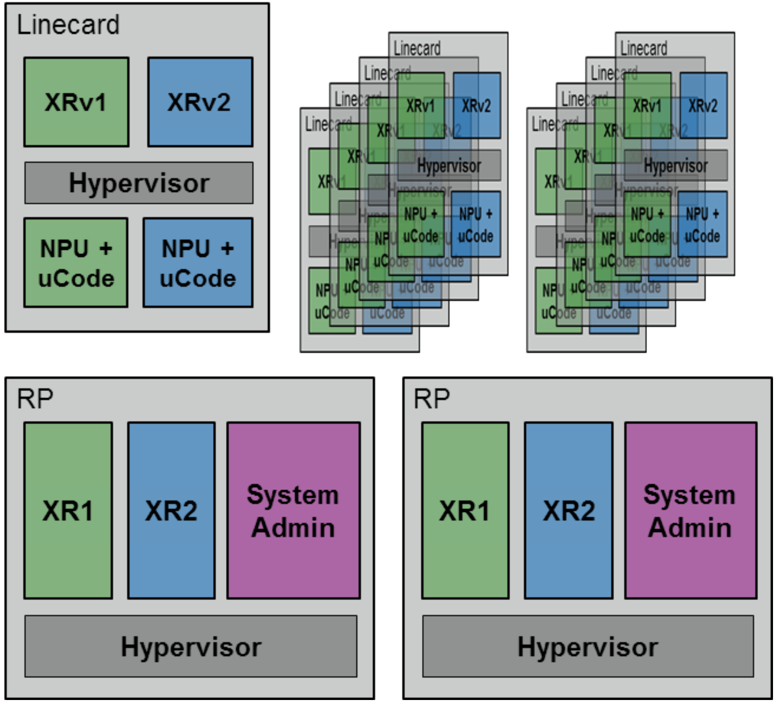
Software Virtualisation Model : Moving Forward

- Thin hypervisor layer
- High performance
- Direct hardware access
- Parallel VMs



Software Availability / Reliability

- Independent versioning
- Protocol and internal state synchronisation
- Parallel protection
- Software flexibility



Linecard

XRv1 XRv2

Hypervisor

NPU + uCode NPU + uCode

RP

XR1 XR2 System Admin

Hypervisor

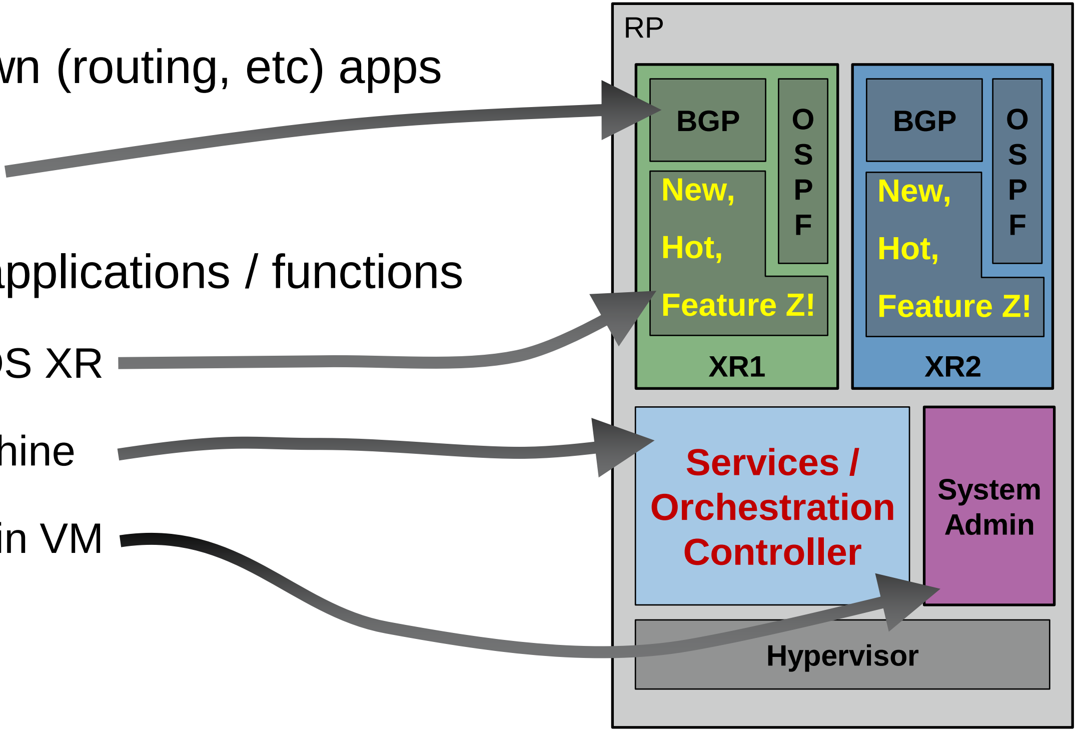
RP

XR1 XR2 System Admin

Hypervisor

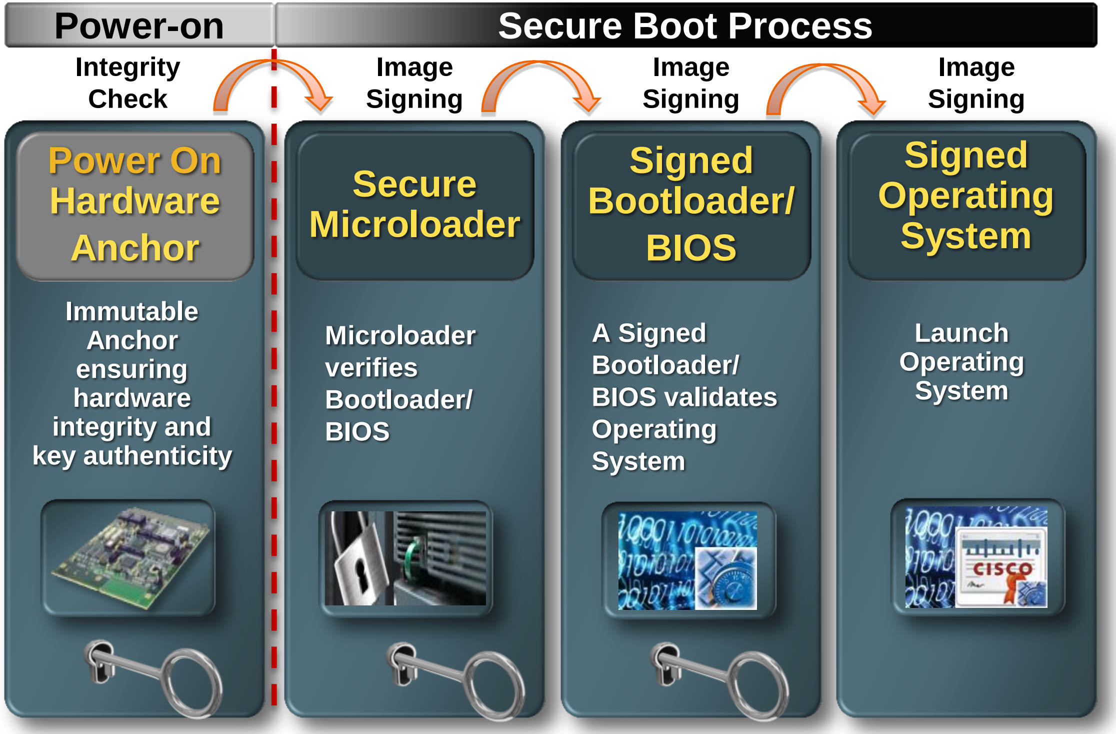
Software Flexibility and Feature Consistency

- Use existing XR for well known (routing, etc) apps
- BGP is "just an application"
- Choice of where to put new applications / functions
  - In a separate module within IOS XR
  - In a new, separate virtual machine
  - (less likely) in the system admin VM



Software Security : Secure Boot

- Authenticate and sign
- Independent security for each phase of the boot process
- Hardware and software
- Crypto-grade randomness



Power-on

Secure Boot Process

Integrity Check

Power On Hardware Anchor

Immutable Anchor ensuring hardware integrity and key authenticity

Image Signing

Secure Microloader

Microloader verifies Bootloader/ BIOS

Image Signing

Signed Bootloader/ BIOS

A Signed Bootloader/ BIOS validates Operating System

Image Signing

Signed Operating System

Launch Operating System

Packet Forwarding Silicon & Network Processors

The ASIC's Role in Networking

Moore's Law still holds ... mostly ☺

- Driving down power, increasing density

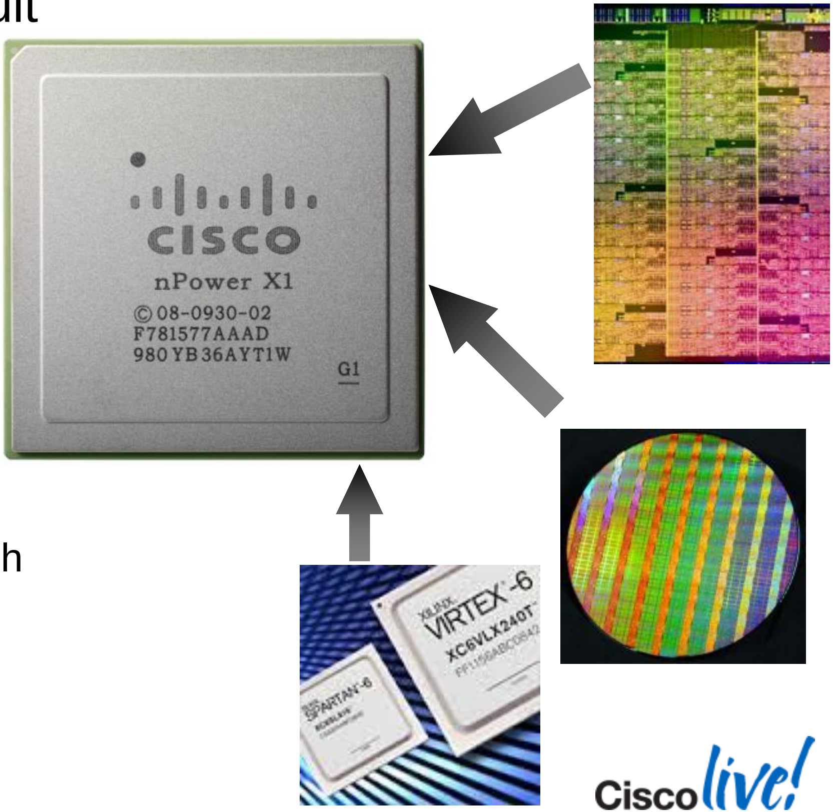
  "The # of transistors on an IC doubles every 2 years" - Stated by George Moore, Intel founder, in 1965

- Many-to-1 Integration – simplifies systems
  - saves power, space, cost, complexity
- ASICs used in many different networking applications
  - Digital Signal Processing for optics
  - NPU
  - Fabric
  - Phy & Framing for I/O

Moore's Law
In 1965, Intel co-founder Gordon Moore predicted
of transistors on a piece of silicon would double e
"Moore's Law."

The Building Blocks

- ASIC: Application Specific Integrated Circuit
  - Exactly what it sounds like...
  - Fast, cheap, low power, ultra high risk.
- FPGA: Field Programmable Gate Array
  - re-programmable, lower risk, lower performance
  - limited by size of logic structures
  - “libraries” from FPGA vendors
- CPU: ‘general purpose’ processor
  - programmable, very flexible, but large & slow & high power
- NPU : Network Processing Unit
  - combines (attempts?) a balance of all



Cisco nPower X1™ NPU - Detail

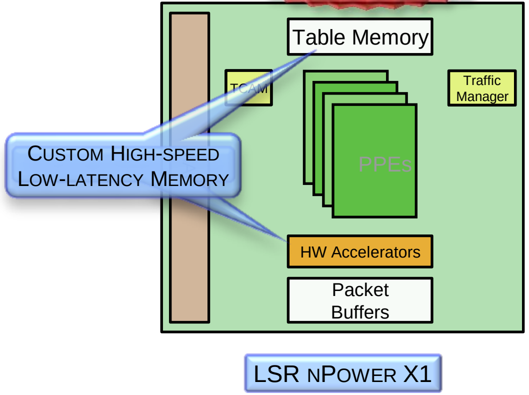
The industry's most advanced Network Processing Unit

Custom High-perf memory

3+ Billion Transistors

HW Accelerators

- **Hardware integration** drives density and power efficiency
- Hybrid NPU/ASIC design
  - 336x 800 MHz custom **programmable** packet processors
  - **Hardware accelerators** for core operations (CEF lookup, WRED, stats, ACL engine . . .)
- 140 Mpps, 200 Gbps full duplex
  - (140 Mpps ingress + 140 Mpps egress on same NPU)
- Traffic Manager for Hierarchical QoS
  - 4K queues, 3+ level QoS
- Integrated MAC silicon, custom lookup RAM
  - Higher bandwidth and lower latency than DDR3

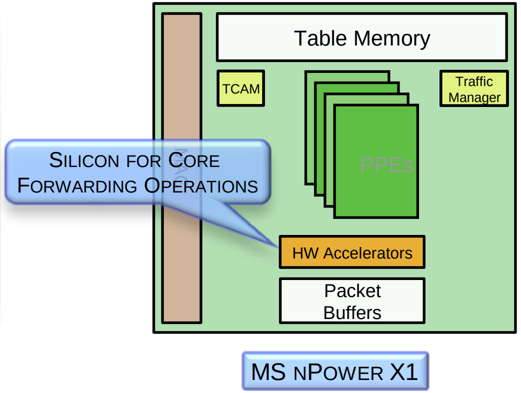


Custom High-speed Low-latency Memory

LSR nPower X1

10x 100Gbps CPAK LC w/ nPower X1

nPower X1 NPU



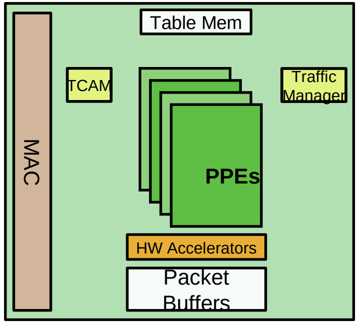
Silicon for Core Forwarding Operations

MS nPower X1

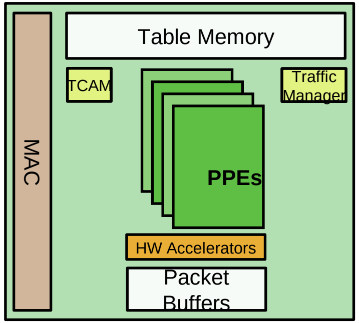
Network Processing Unit Comparison

re-use of silicon “blocks” is key to efficiency

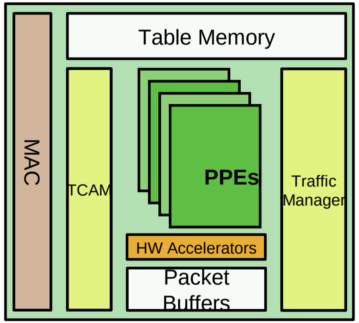
- X1 LSR cards are optimised for cost, performance & density
- X1 multi-service cards are optimised high performance & scaled core services
  - Increased Table Memory for forwarding databases
- X1e Line Cards are optimised for highest scale
  - 32x queue density, 8xTCAM density compared to X1 series linecards
  - 5+ Levels of Hierarchical QoS
  - Lower per-card density due to faceplate real estate (i.e., size of 60 SFP+ optics)



LSR NPOWER X1



MS NPOWER X1

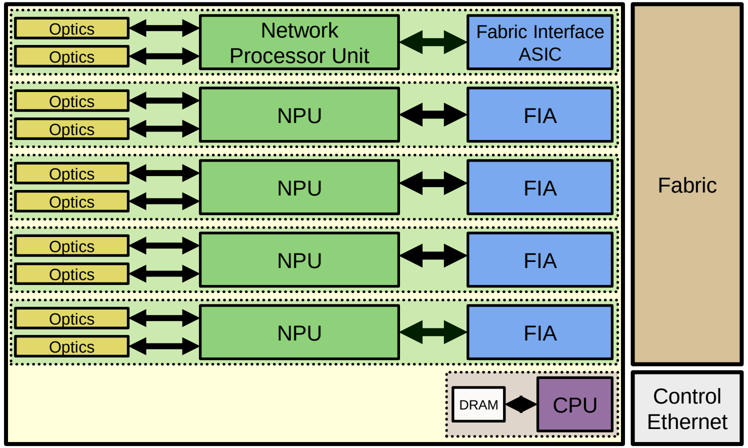


NPOWER X1E

NCS 6000 Line Cards

- Route Processor builds forwarding tables and distributes to Line Cards
- Line Cards contain a variable number of **slices**
- Each slice is an autonomous set of optics and forwarding ASICs
  - Packet forwarding, features, and QoS
  - Fabric access, including fabric queuing, segmentation/reassembly, and scheduling

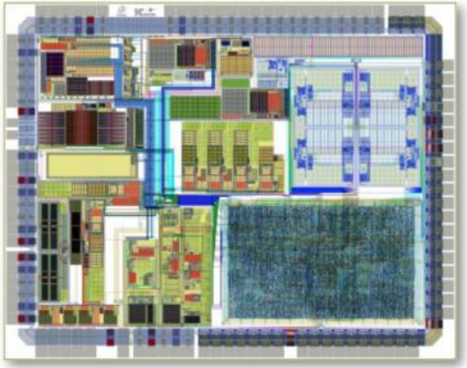
5-SLICE 100G LINE CARD



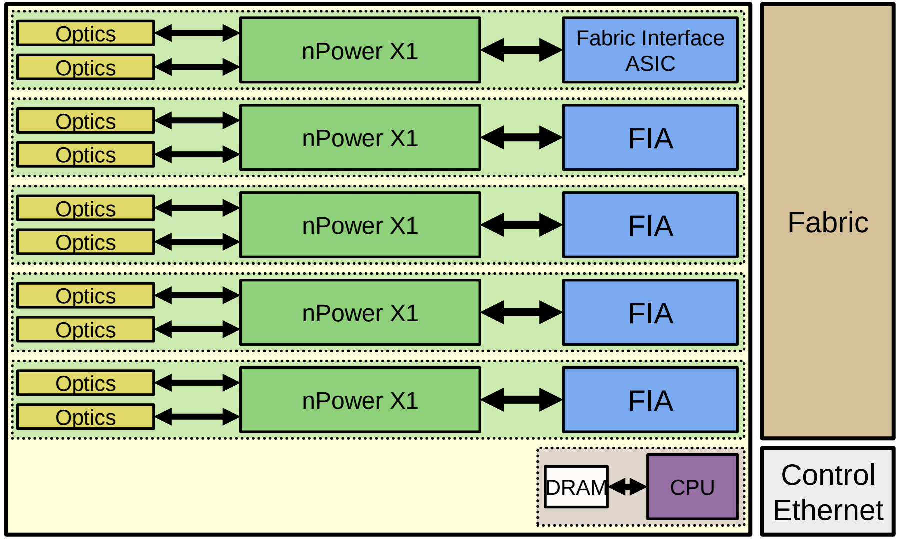
Line Cards with Cisco nPower X1™ NPUs

- 1 Tbps Line Cards have 5 data-plane slices with nPower X1 NPUs
- 700 Mpps ingress + 700 Mpps egress per LC
- 200Gbps bi-dir + tip + tax

NPOWER X1
FORWARDING ASIC



CORE LINE CARD
WITH SLICE ARCHITECTURE



Optics
Optics
nPower X1
Fabric Interface ASIC

Optics
Optics
nPower X1
FIA

Optics
Optics
nPower X1
FIA

Optics
Optics
nPower X1
FIA

Optics
Optics
nPower X1
FIA

DRAM
CPU

Fabric

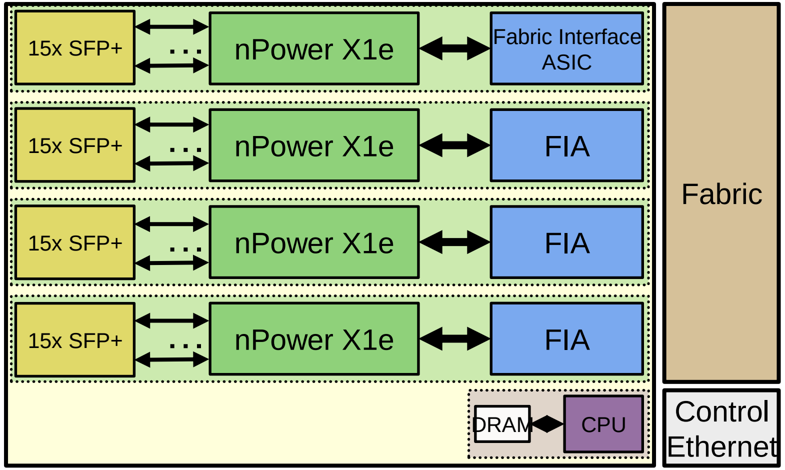
Control Ethernet

Cisco nPower X1e™ Line Card Architecture

- Line Cards with nPower X1e have 4 slices
- 480 Mpps ingress + 480 Mpps egress per LC
- Extended Traffic Manager & TCAM

  Highest QoS & Scale

NPOWER X1E FORWARDING ASIC



NPOWER X1E CARD ARCHITECTURE

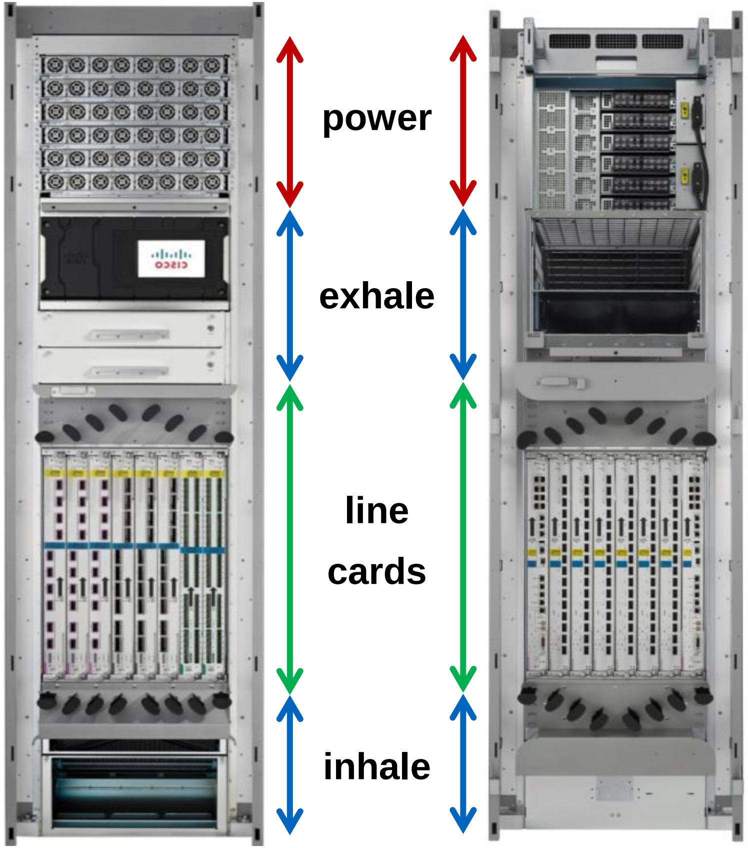
Mechanical, Thermal, and Power

Mechanical, Thermal, and Power Considerations

- My goodness, these are big boxes...
- How do we build them?
  - (never mind move them...)
- How do we cool them?
- How do we power them?

Thermal Design Considerations

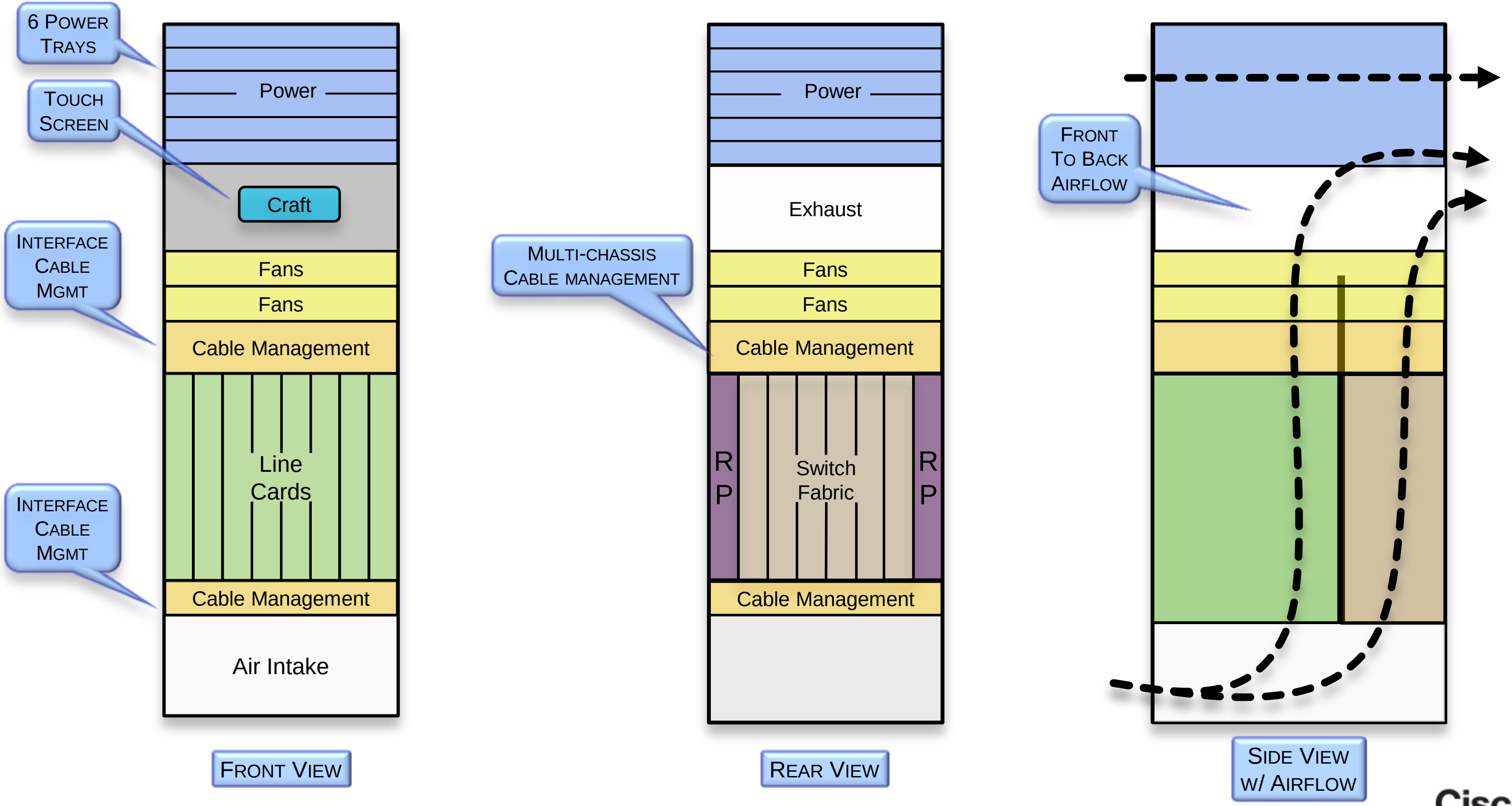
- Heatsinks == heat exchangers
  - delta(temp) * delta(time)
  - limits on max ASIC/component temps
- Fan laws (these are not your friends):
  - flow ~= rpm^2
  - power ~= rpm^3
- → Wide, tall slots and large input/exhaust plenums to reduce air velocity



Chassis Design and Fabrication

- Torsional strength / rigidity
- Seismic stability
- weight (total and per-component)
- Cable management
- Serviceability
- Power distribution
- it-is-a-rack vs rack-mountable

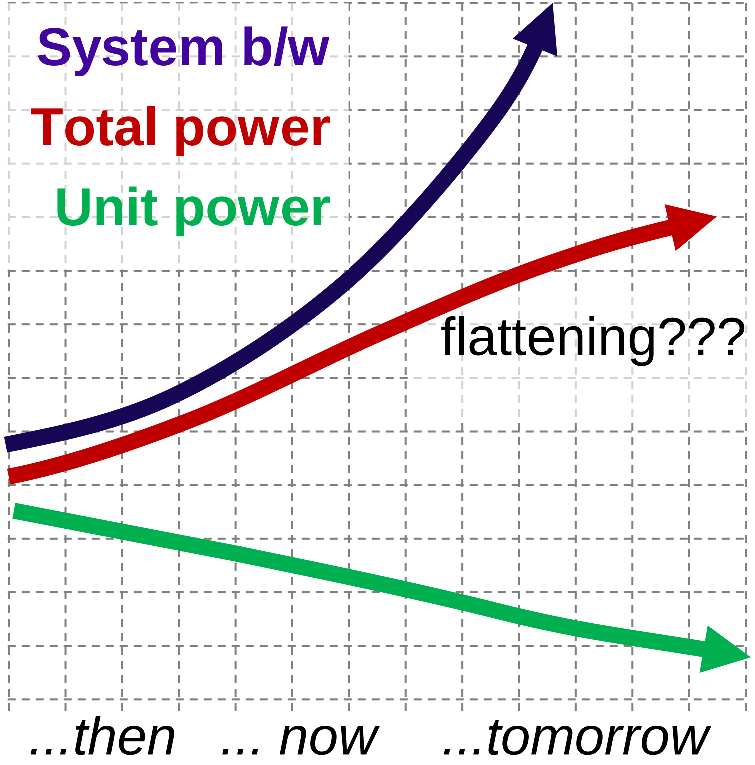
NCS 6008 Physical Layout



System Power Realities

- Business demands higher bw/rack
- System bandwidth increases FASTER

  than unit power decreases...
- Power / rack can continue to go up
  - or
- equivalent systems will get smaller



Optical Density & Flexibility

Delivering Tomorrow’s Technology Today
100G Transceiver CMOS Photonics

100G-SR10 CPAK

100G-LR4 CPAK

- **Standards Based**

  100GBASE-LR4 / SR10 compliant

  OTU4 compliant

- **High Density**

  Next-gen devices yield >70% size Reduction

  Drives density on core routers to 10 ports of 100GE

  Utilises 4x 25G electrical interface (common to other technologies)

- **Low Power Consumption**

  Next-gen devices can yield 70% decrease in power consumption

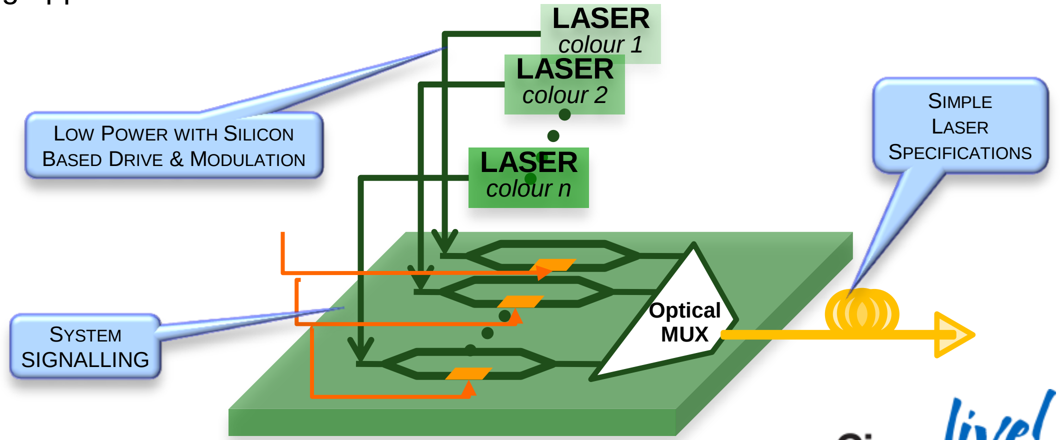
  Leveraging next-gen optics & next-gen IC technology for efficiency

- **Leveraging Silicon / Moore’s Law**
  - CMOS Photonics leverages massive industry investment in CMOS mfg

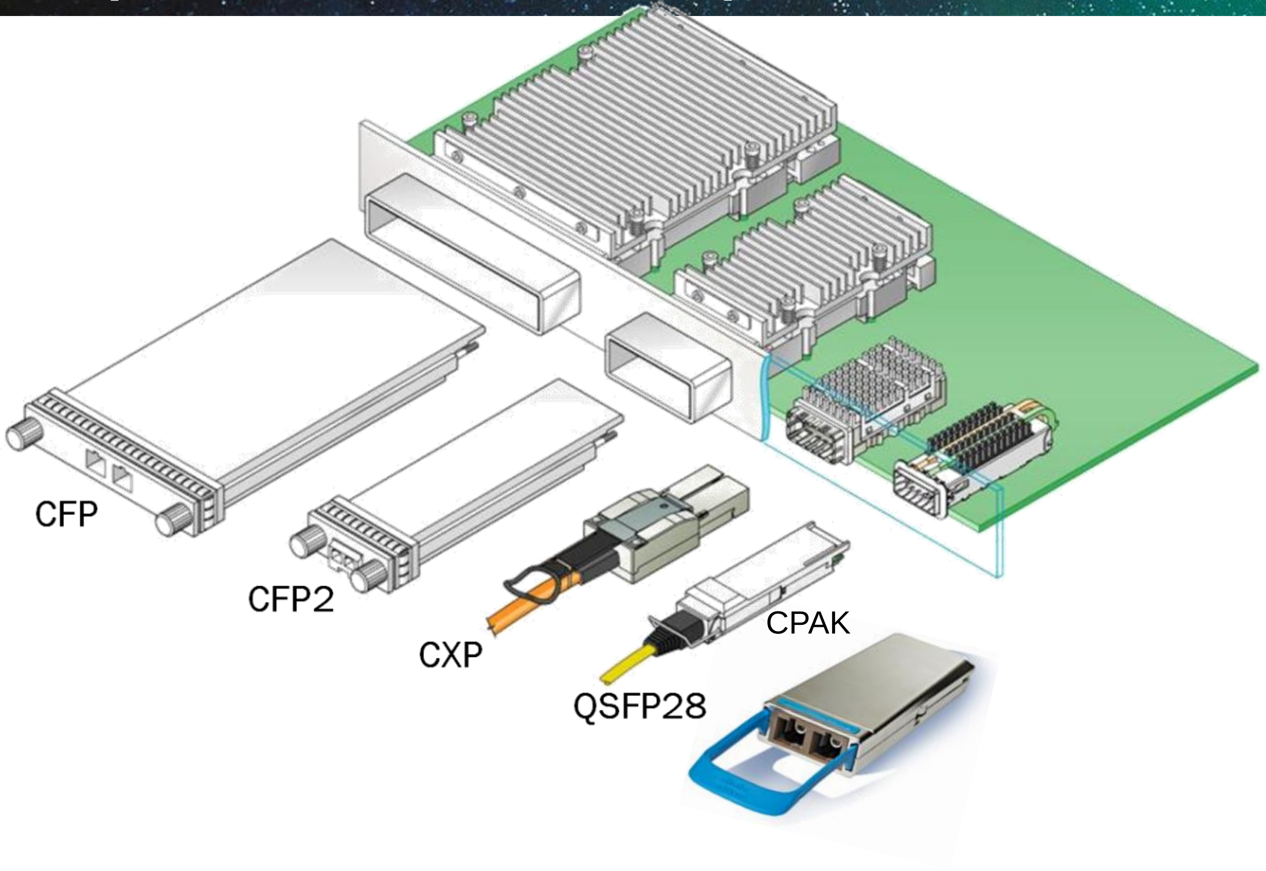
CMOS Photonics Technology

- CMOS Photonics grown in Silicon fabs like other ASICs & commercial ICs
- Light is Continuous Wave Power (DC in electricity) supplied by laser
- Light is modulated in CMOS photonics & coupled into fibre for transmission
- CMOS Photonics is designed using standard IC design tools

  Traditional photonics is designed with analog approximations & customised tools


LASER colour 1
LASER colour 2
LASER colour n
Low Power with Silicon Based Drive & Modulation
System SIGNALLING
Optical MUX
Simple Laser Specifications

Optics Form Factor Comparison

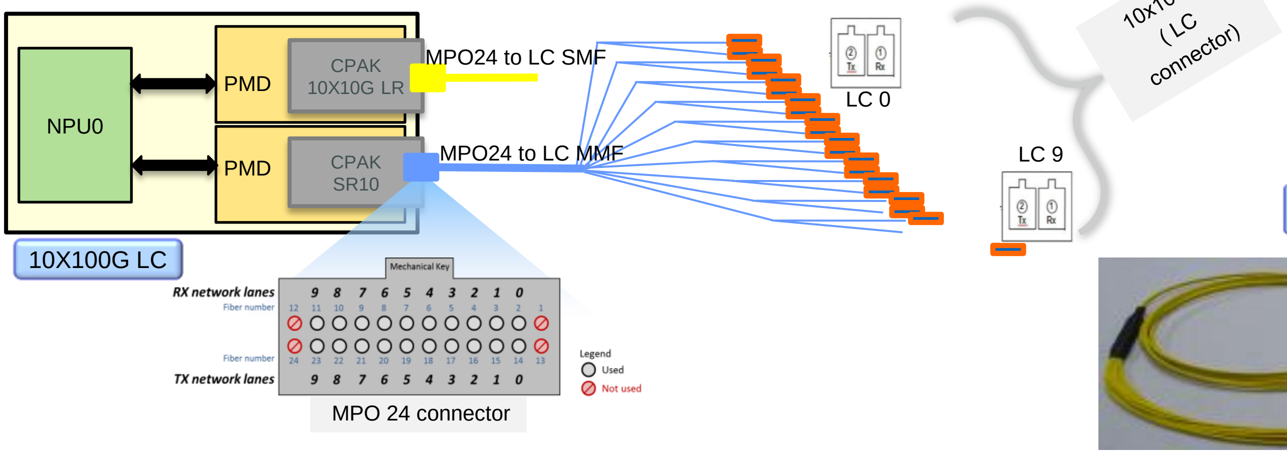


- **choices, choices**
- **power**
  - implies density
- **size**
  - **can’t have density and backwards compatibility...**

10G Breakout Cable Solution

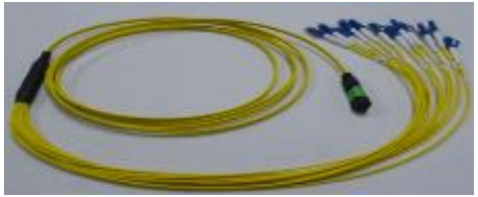
```
RP/0/RP0/CPU0:ios(config)# hw-module location 0/0/CPU0 slice 0 breakout 10G
RP/0/RP0/CPU0:ios#sh interfaces * br
               Intf          Intf        LineP              Encap  MTU        BW
               Name          State       State              Type  (byte)    (Kbps)
--------------------------------------------------------------------------------
          Te0/0/1/0            up          up               ARPA   1514    10000000
           ~ snip ~
          Te0/0/1/9            up          up               ARPA   1514    10000000
```

- 10 : 1 Breakout Cable with a MPO24 MMF or SMF
- Only 20 fibres are used from 24 fibres of one MPO24 cable



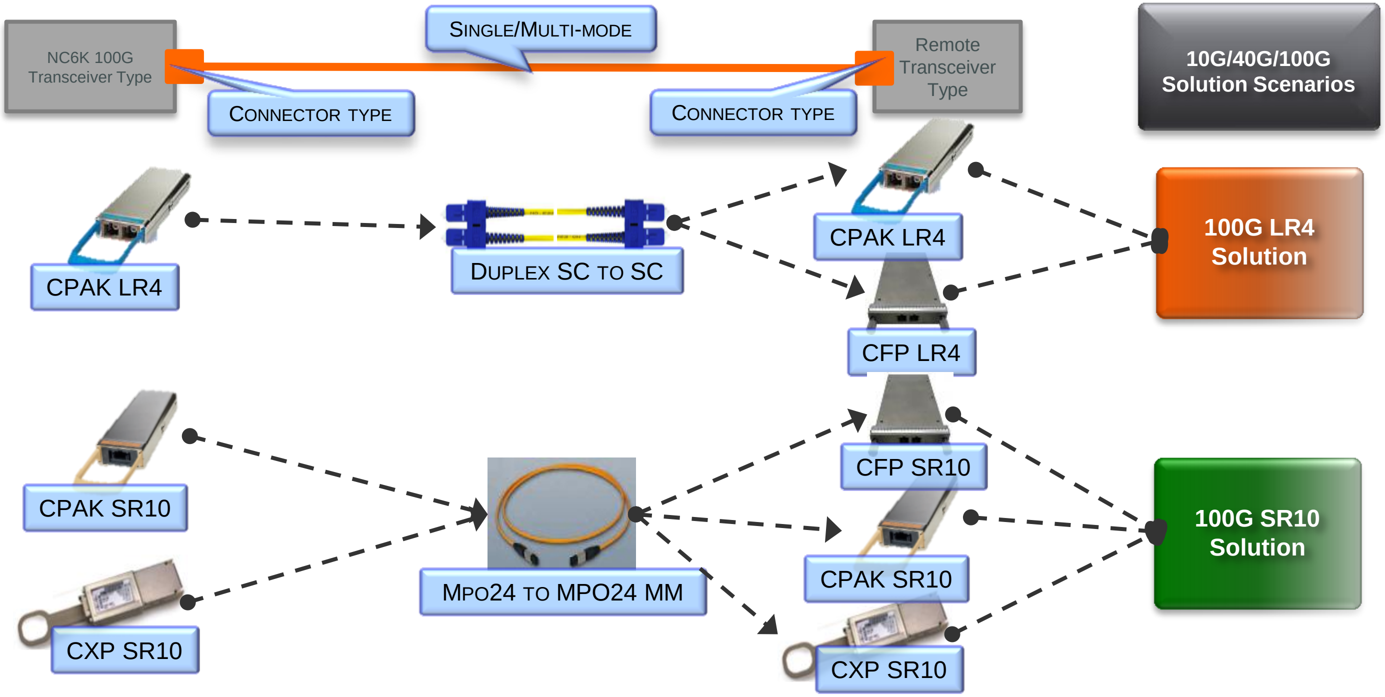
MPO 24 connector

MPO24 TO 10X DUPLEX LC MM



MPO24 TO 10X DUPLEX LC SM

100G Solution Scenarios

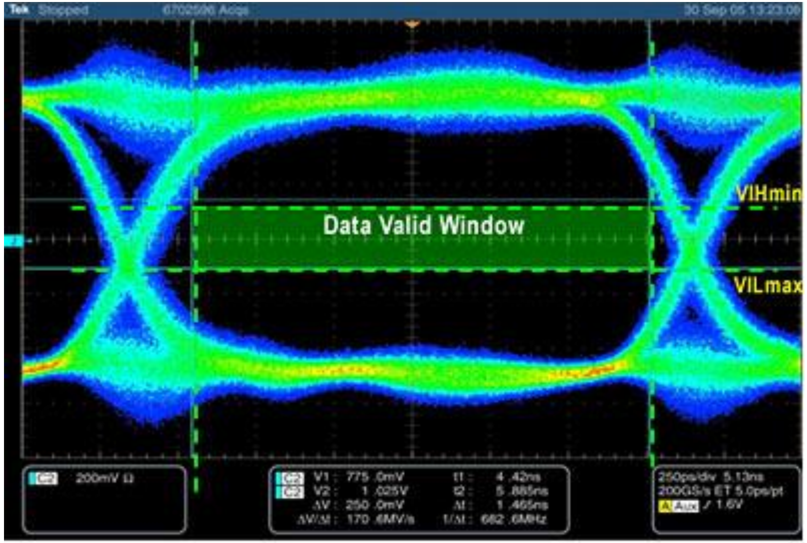


- **Part numbering:**
- **CPAK-100G-LR4**
  - **CPAK = Form Factor**
    - **local to system**
  - **100G = link speed**
  - **LR = range standard**
  - **4 = # of channels**
    - **standard**

Long Haul Optical Performance & Granularity

Long Haul Optical Challenges

- Non-linear effects
- Signal integrity over long distances
  - amplification and/or regeneration
- At very high bit rates
  - 40 -> 100 -> Nx100G
- Maintain reasonable system power



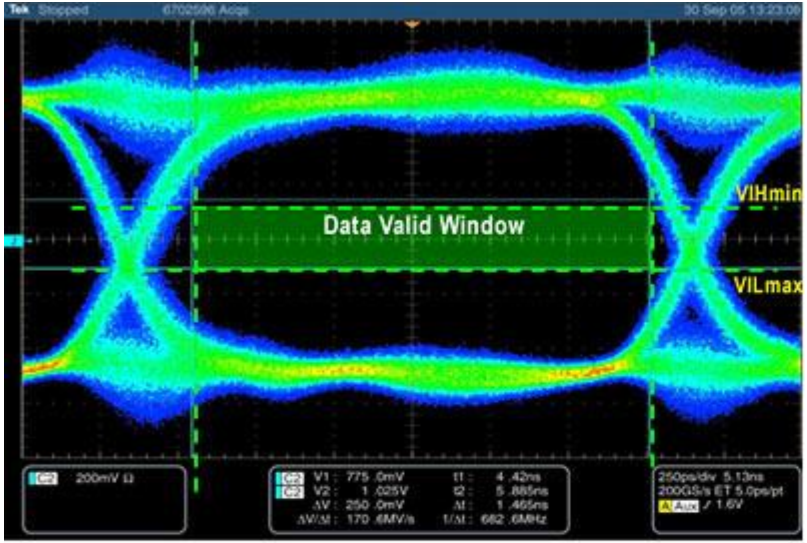
Long Haul Optical Challenges

- spectral efficiency

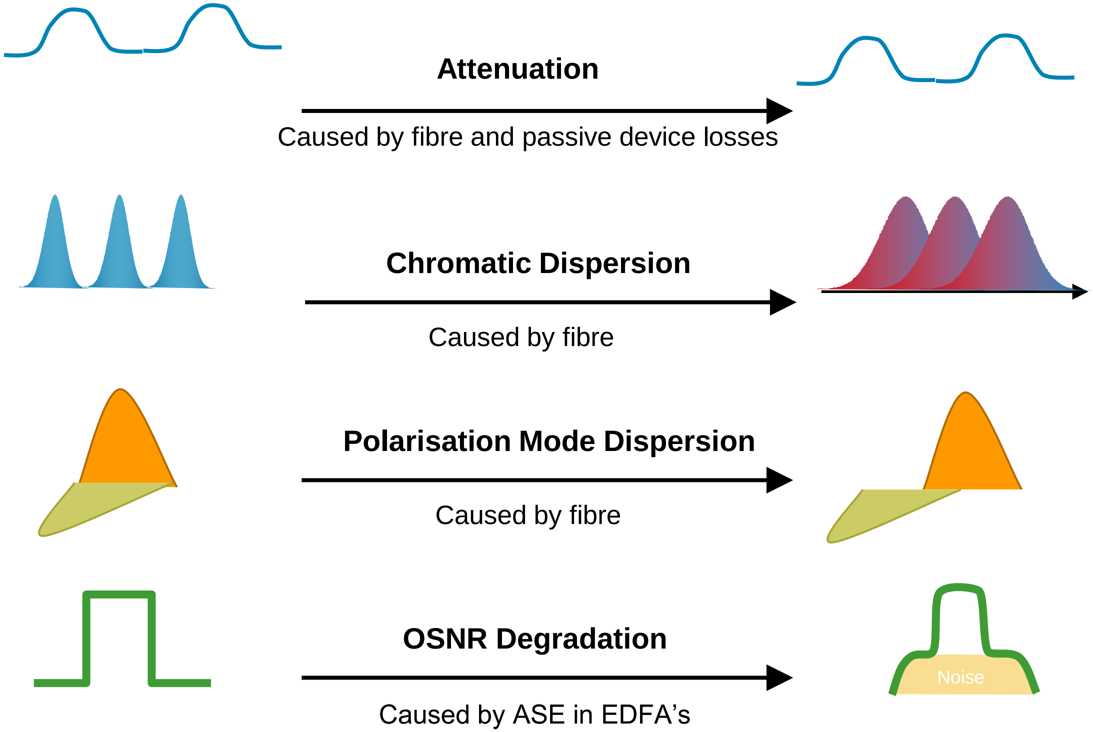
- modulation techniques
- demodulation techniques
  - amplification and/or regeneration

- sub- and super-channels
  - 40 -> 100 -> Nx100G

- reach for granularity tradeoff



Linear Channel Impairments



Attenuation

Caused by fibre and passive device losses

Chromatic Dispersion

Caused by fibre

Polarisation Mode Dispersion

Caused by fibre

OSNR Degradation

Caused by ASE in EDFA’s

Linear Optical Impairments

Solutions

Attenuation

EDFA's can help overcome attenuation, applied per span, but add noise

…Hybrid Raman/EDFA amplification can overcome attenuation with minimal noise

Chromatic Dispersion

DCU's can help mitigate dispersion problems, applied per span, but add cost, latency, and loss

…Now compensated for in Digital Signal Processing via Coherent Detection

Polarisation Mode Dispersion

Generally have to live with it. Regenerate signal when required.

…Now compensated for in Digital Signal Processing via Coherent Detection

OSNR

Nothing can overcome losses in OSNR! Must regenerate!

…But advanced Forward Error Correction can lower OSNR requirements

Why is 100G an Inflection Point?

- Dramatic improvement in performance over 10G
  - Higher CD tolerance – no DCU needed (70,000 ps/nm)
  - Higher PMD tolerance – tolerant of old / bad fibre (100 ps DGD)
  - Manageable OSNR requirement – similar to 10G, minimize regen
  - Compatible with existing 10G – maximize wavelength fill

- Ultimately – what do we mean by performance?

Performance = cost / bit / km

Capacity + Performance = Rapid Adoption

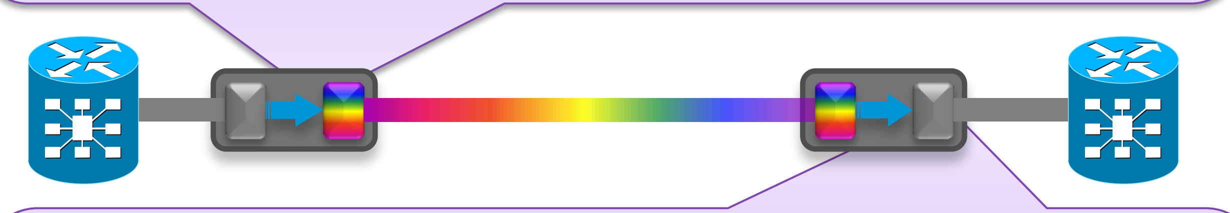
100 Gigabit DWDM Transmission

Overcoming the Challenges

Problem: DSP electronics are not yet capable of processing 100Gb/s serial data rates

Solution: Dual Polarisation **QPSK** Modulation, allows single wavelength 100G transmission with a baud rate of ~28-32 Gbaud/s

This is a **Modulation** (TX) function



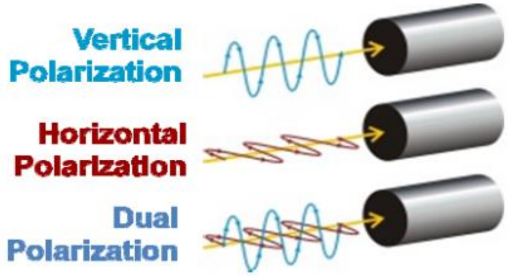
Problem: Transmission impairments increase significantly at high bit rates (CD, PMD, non-linear effects)

Solution: Compensate for these impairments with intelligent Digital Signal Processing, enabled by **Coherent Detection**

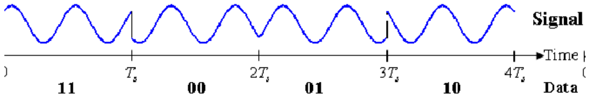
This is a **Demodulation** (RX) function

100G DWDM Transmission

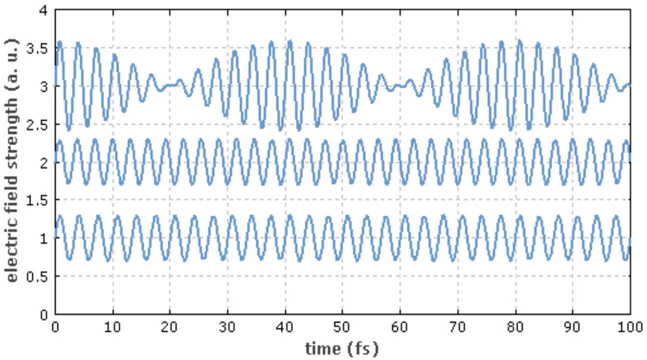
Dual Polarisation



Quadrature Phase Shift Keying



With Coherent Detection



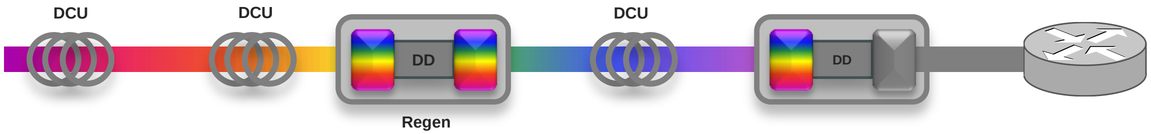
100G Technology – Coherent Detection

Direct Detection

Must correct for impairments in the physical domain (insert DCU’s)

Forced to live with non-correctable impairments via network design (limit distance, regenerate, adjust channel spacing)

Dumb detection (OOK), no Digital Signal Processing, only FEC

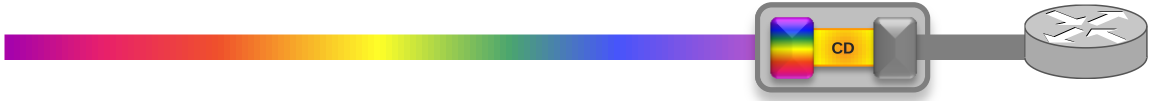


Coherent Detection

Moves impairment correction **from the optical** domain **into the digital** domain

Allows for digital correction of impairments (powerful DSP) vs. physical correction of impairments (DCU’s). Adds advanced FEC.

Massive performance improvements over Direct Detection.

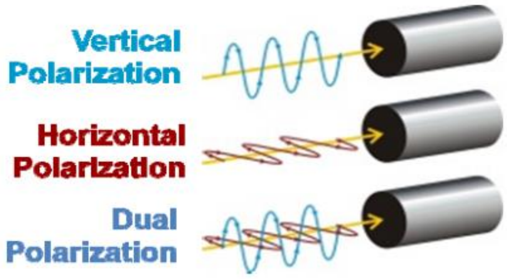


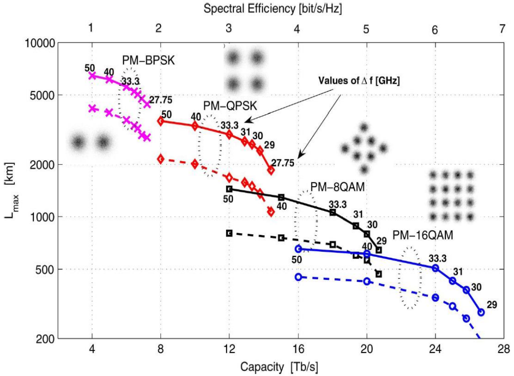
DSP / Silicon-driven data recovery

... In 2014 it’s now easier to use transistor/CMOS based chips to do better signal recovery than it is to improve the optics themselves...

100G Technology - Summary

- After modulation, resulting baud rate is ¼ of the bit rate
- Minimizes optical bandwidth requirements, maintain 50GHz spacing
- Receiver only requires DSP technology at 28-32 Gb/s, available today
- Coherent Detection results in awesome tolerance to optical impairments



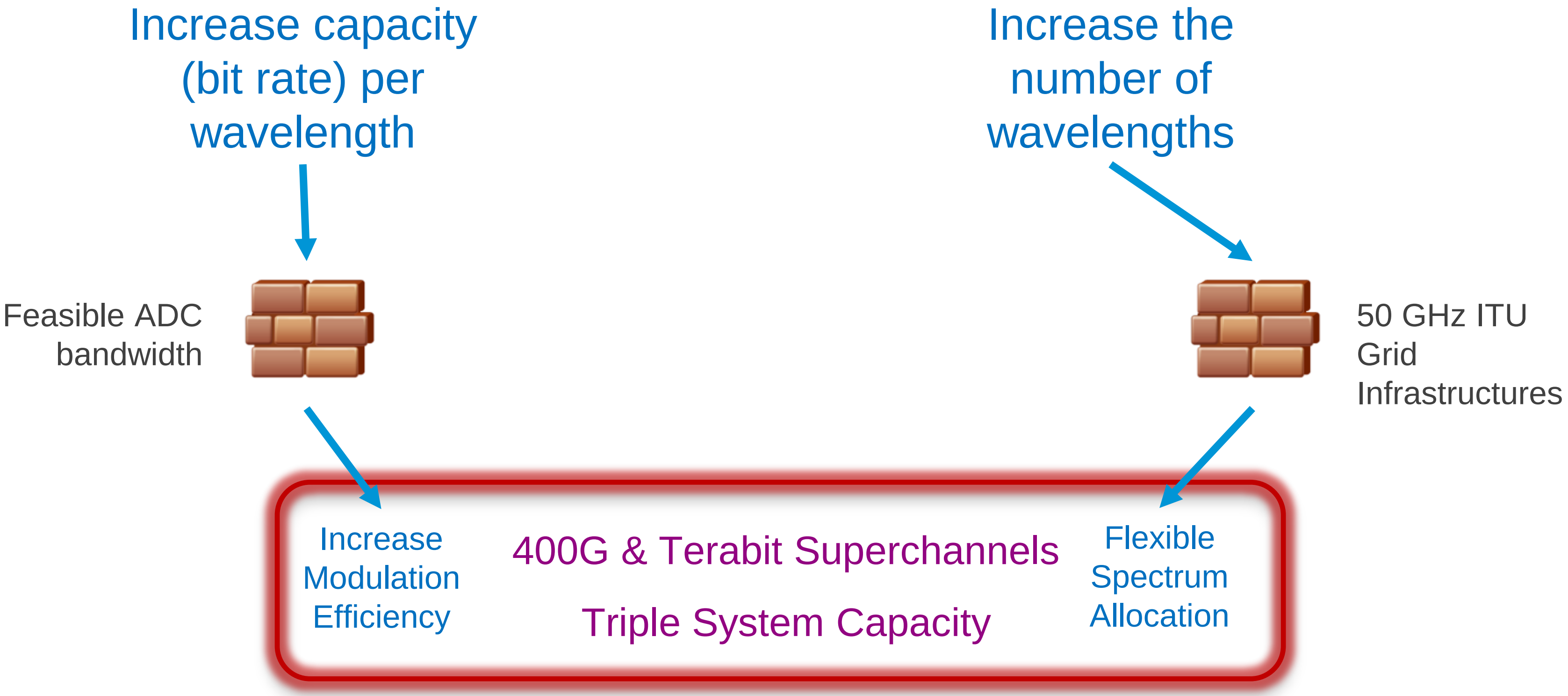


Drivers to go Beyond 100G

- Mobile, Video, Data Centre bandwidth growing exponentially
- Continual router switching / port capacity increases
  - Next generation NPU’s can drive 200G+ per interface
  - 1Tb/s Line Cards coming
  - IEEE working on 400G and/or 1Tb Ethernet
- DWDM System Capacity at 100G, 8-10 Tb/s, is not enough

We Need More Transport Capacity

How to Increase Transport Capacity?



Increase capacity (bit rate) per wavelength

Feasible ADC bandwidth

Increase the number of wavelengths

50 GHz ITU Grid Infrastructures

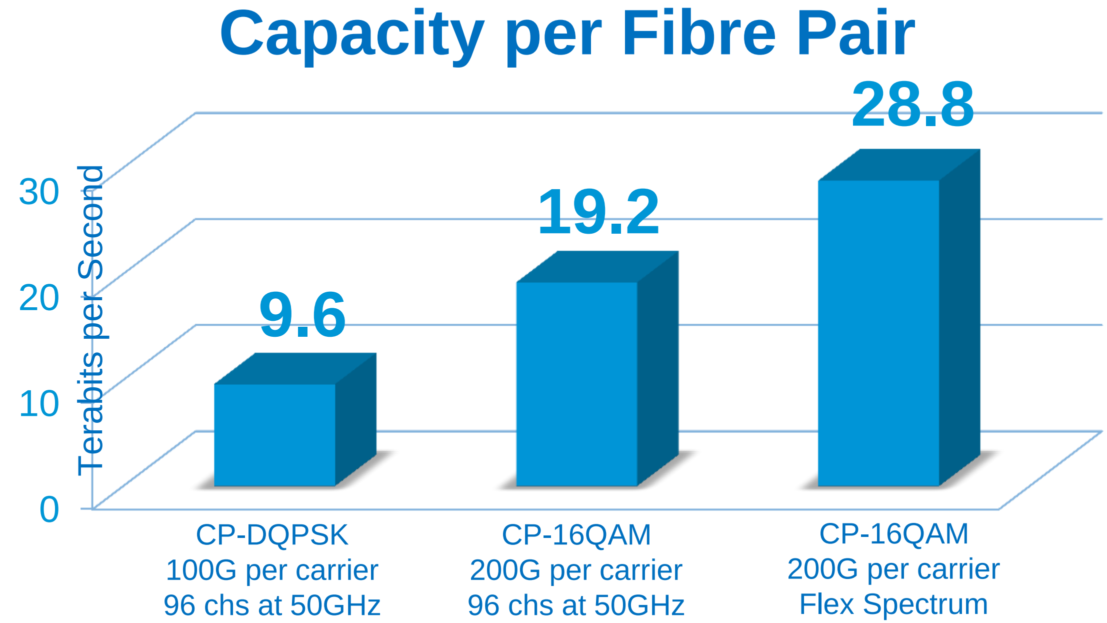
Increase Modulation Efficiency

400G & Terabit Superchannels

Triple System Capacity

Flexible Spectrum Allocation

DWDM System Capacity



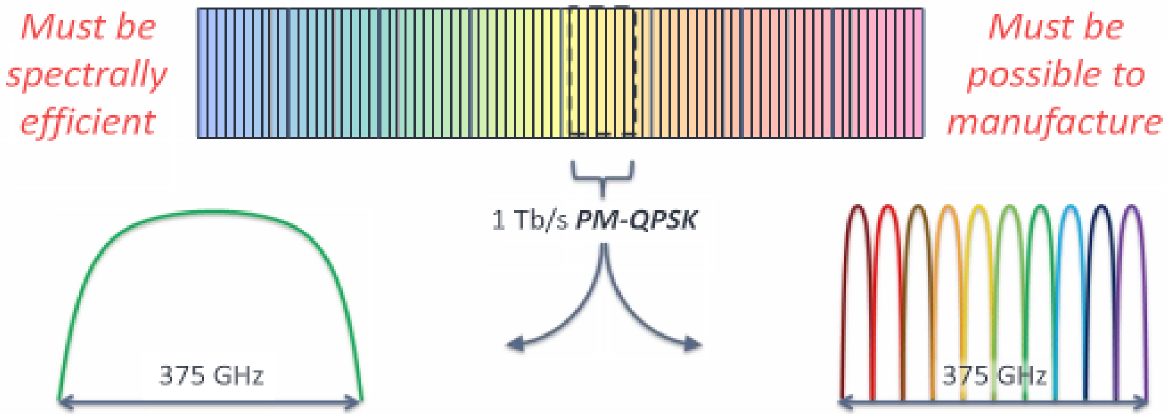
Today → **Next Gen Silicon** → **Next Gen Silicon & ROADM**

The Superchannel Concept

- Information distributed over a few subcarriers spaced as closely as possible forming a variable rate superchannel
- Each subcarrier working at a lower rate, compatible with current ADCs and DSPs

Must be spectrally efficient

Must be possible to manufacture

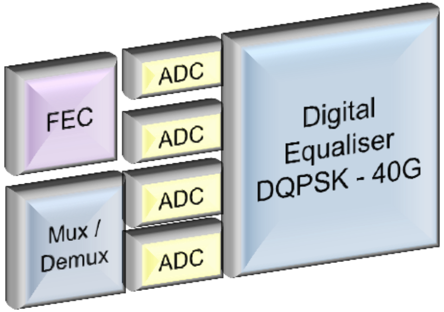


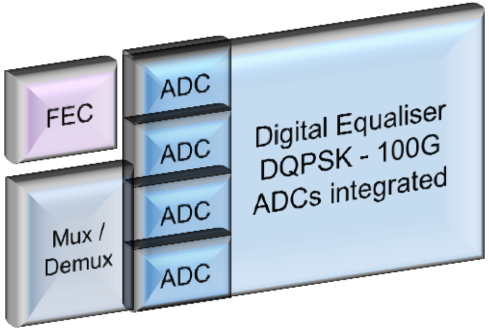
1 Laser
4 modulators
320 GBaud Electronics
~ 11 nm Si
Time to Market: ~10 years

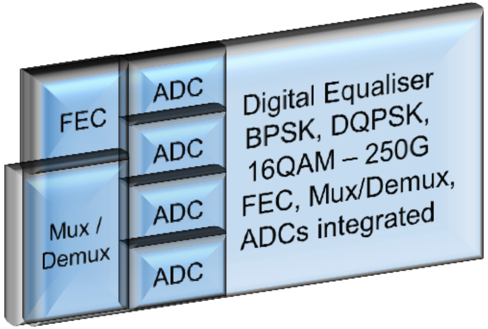
10 Lasers
40 Modulators
32 GBaud electronics
Photonic ICs
Time to Market: ~2 years

DSP Evolution – Many to 1 Integration

Enabling the Next Generation of Terabit Networks






FEC
Mux / Demux
ADC
ADC
ADC
ADC
Digital Equaliser BPSK, DQPSK, 16QAM – 250G FEC, Mux/Demux, ADCs integrated

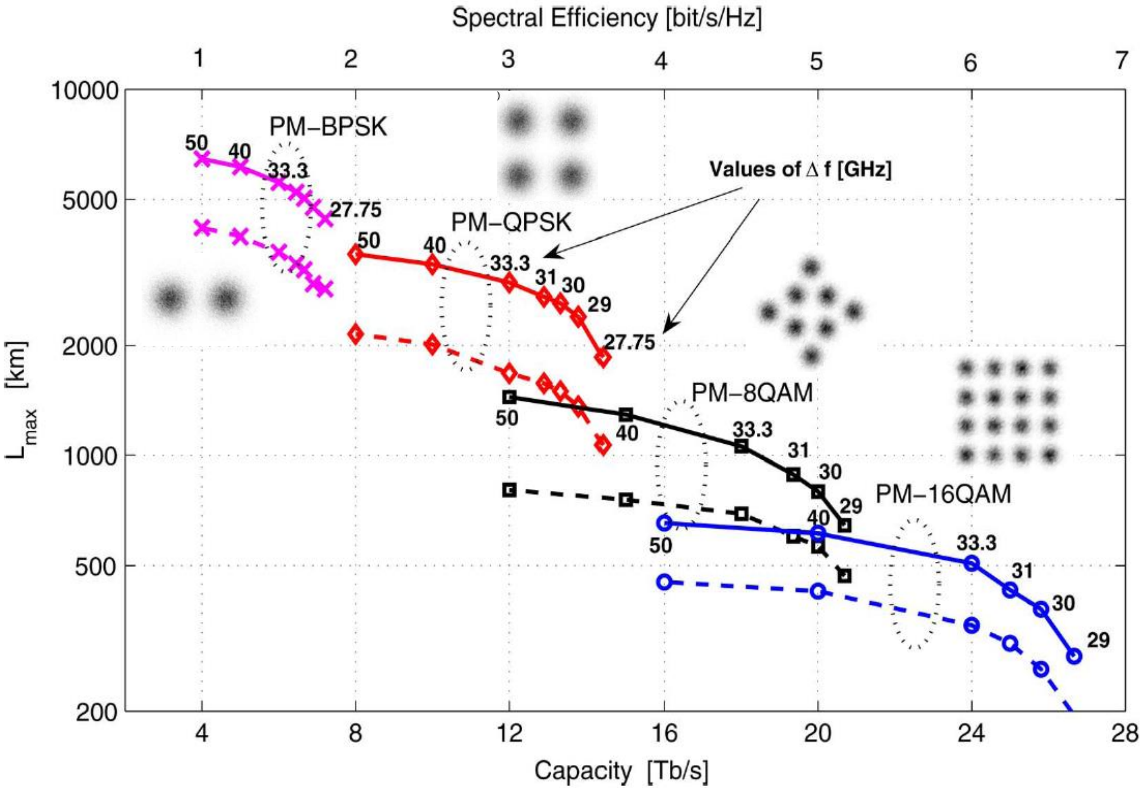
40G DSP

DQPSK – 40G

100G DSP
chipset
8-Bit ADC
DQPSK – 100G
Digital Equaliser
65 nm CMOS

250G DSP chipset
Digital Equaliser
Software Selectable Modulation
BPSK-50G, QPSK-100G, 16QAM-250G
TX / RX DSP, Soft Decision FEC
28 nm CMOS

Trade off: Reach vs. Capacity



- Tighter spacing:
  - more data but shorter reach
- Higher modulation:
  - more data, but shorter reach
- Better fibre:
  - longer reach, but $$$$$
- **More efficiency....**
  - **but lots more engineering!**

nLight Next Generation ROADM

Add/Drop without physical contact

Provide Programmability

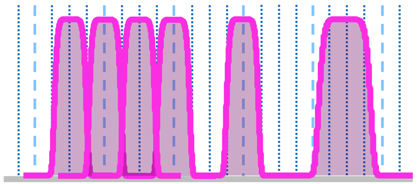
- Impairment awareness
- Automatic optical restoration

Meshing & Scaling challenges

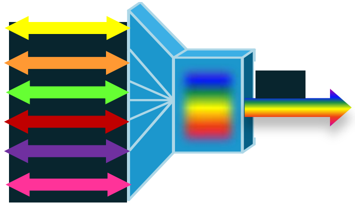
- Wavelength and degree contention

Amplification

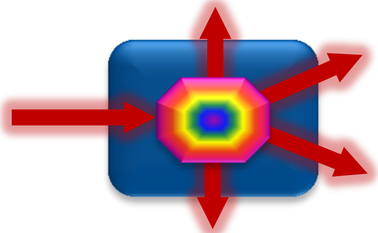
- Silicon integration (EDFA + Raman(
- reduce power and improve Rx



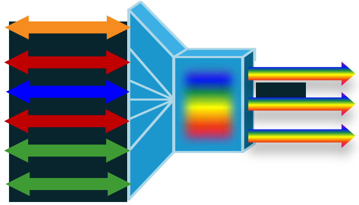
Flex Spectrum (96 chs @ 50GHz)



Colourless

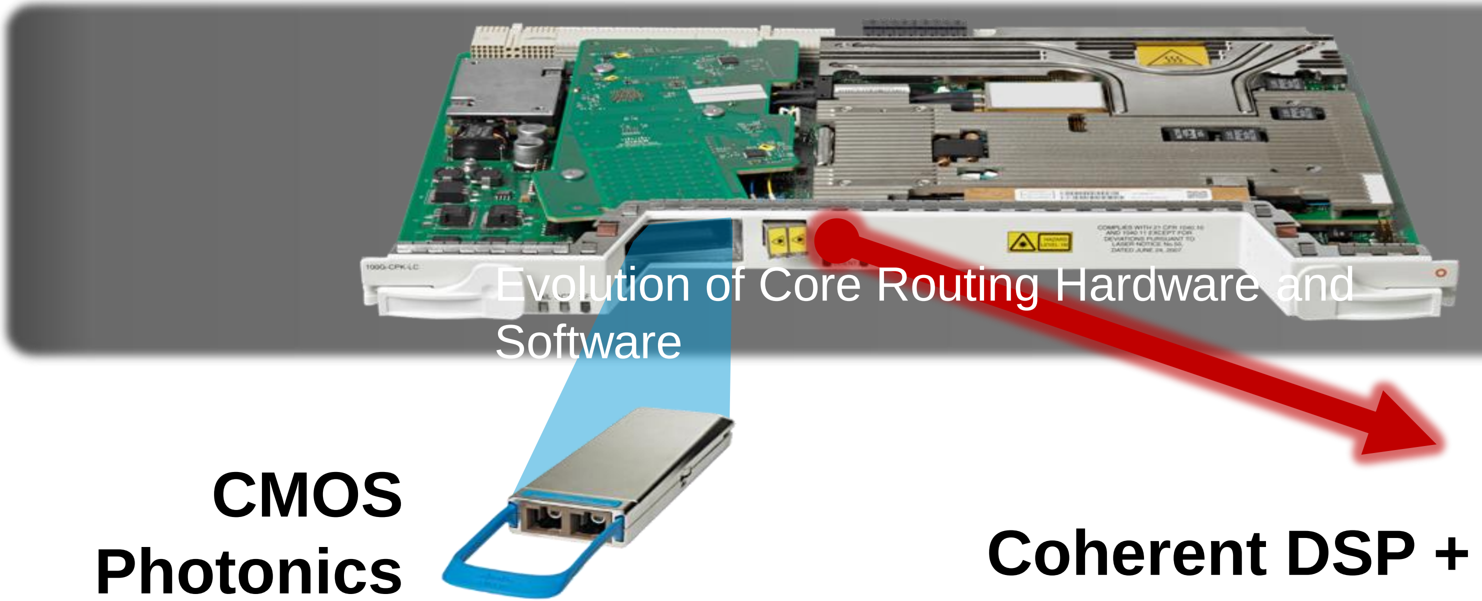


Omni-Directional 16 (32) Degrees



Contentionless

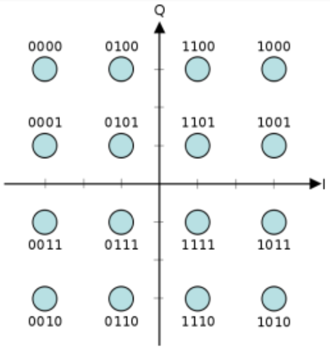
nLight 100G CPAK Coherent Transponder



Evolution of Core Routing Hardware and Software

CMOS Photonics

Coherent DSP + Advanced modulations



Single slot transponder with LR4 or SR10 client interface

Q & A

Complete Your Online Session Evaluation

Give us your feedback and receive a Cisco Live 2014 Polo Shirt!

Complete your Overall Event Survey and 5 Session Evaluations.

- Directly from your mobile device on the Cisco Live Mobile App
- By visiting the Cisco Live Mobile Site www.ciscoliveaustralia.com/mobile
- Visit any Cisco Live Internet Station located throughout the venue

Polo Shirts can be collected in the World of Solutions on Friday 21 March 12:00pm - 2:00pm

Learn online with Cisco Live!

Visit us online after the conference for full access to session videos and presentations. www.CiscoLiveAPAC.com cisco™