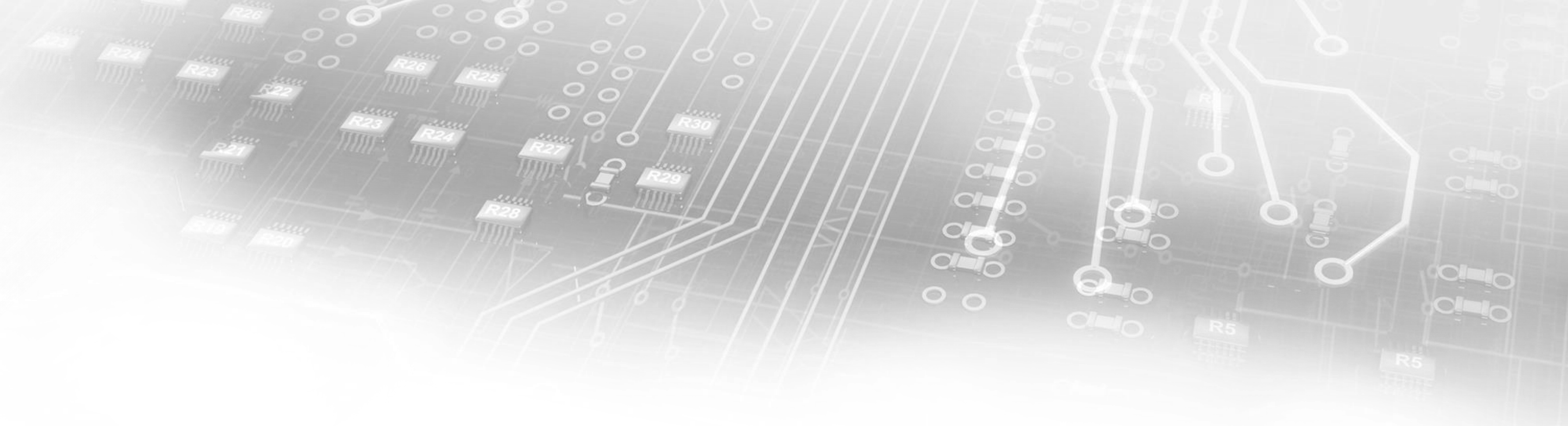
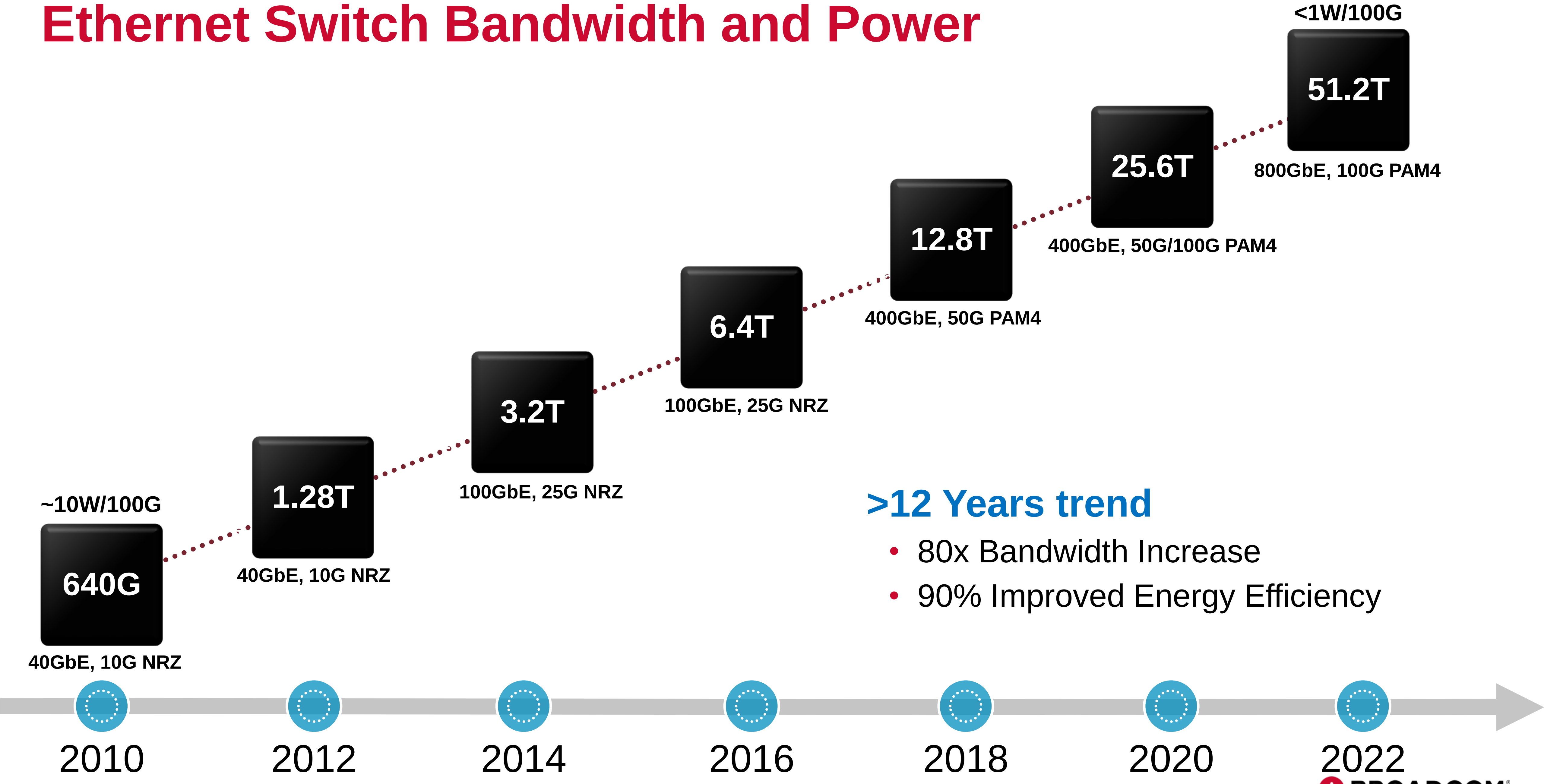


Evolution of switches power consumption

Itzik Kiselevsky

Carbon Aware Networks Workshop 2023

Ethernet Switch Bandwidth and Power



~10W/100G

640G

40GbE, 10G NRZ

1.28T

40GbE, 10G NRZ

3.2T

100GbE, 25G NRZ

6.4T

100GbE, 25G NRZ

12.8T

400GbE, 50G PAM4

25.6T

400GbE, 50G/100G PAM4

<1W/100G

51.2T

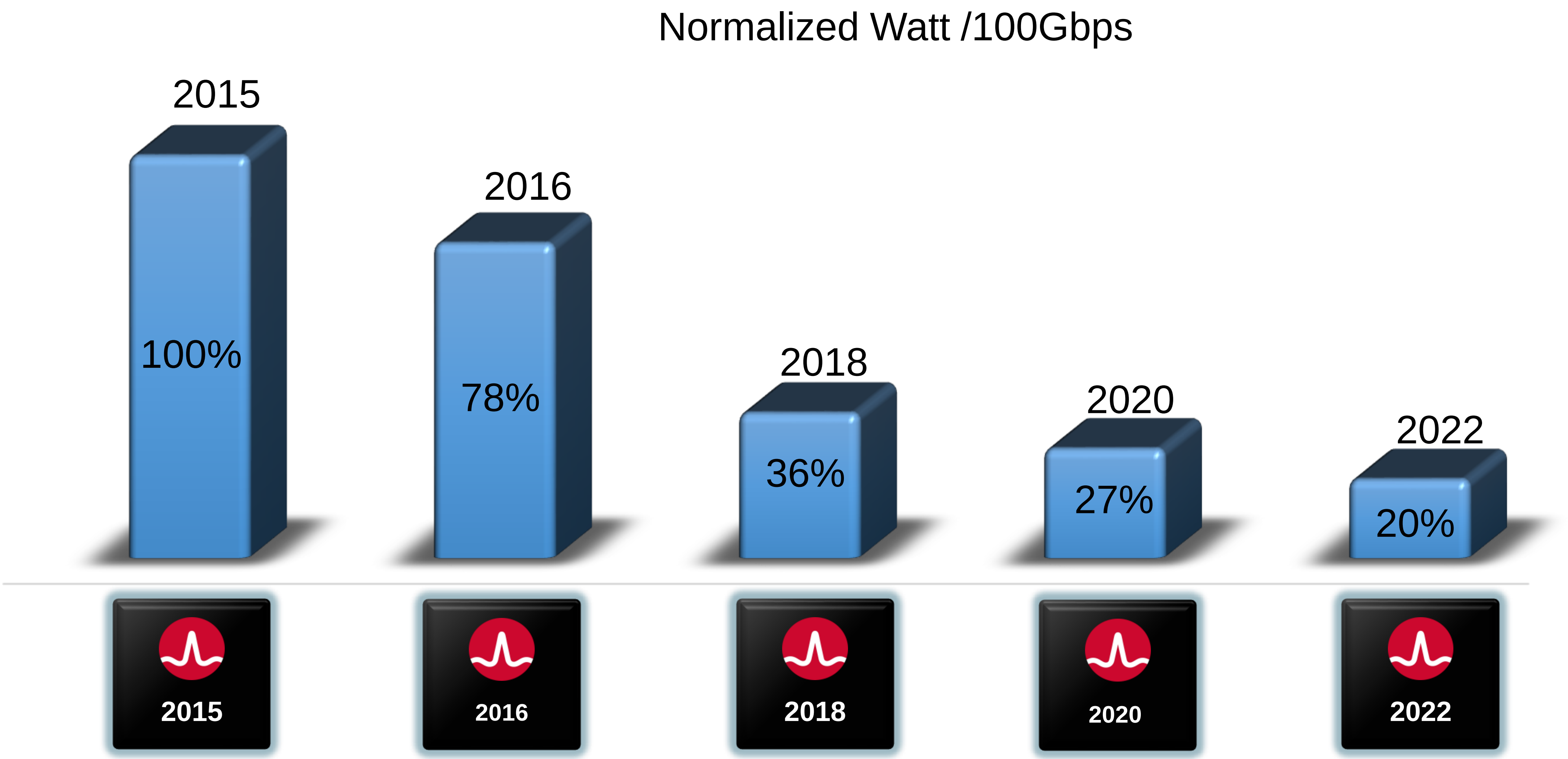
800GbE, 100G PAM4

2010 2012 2014 2016 2018 2020 2022

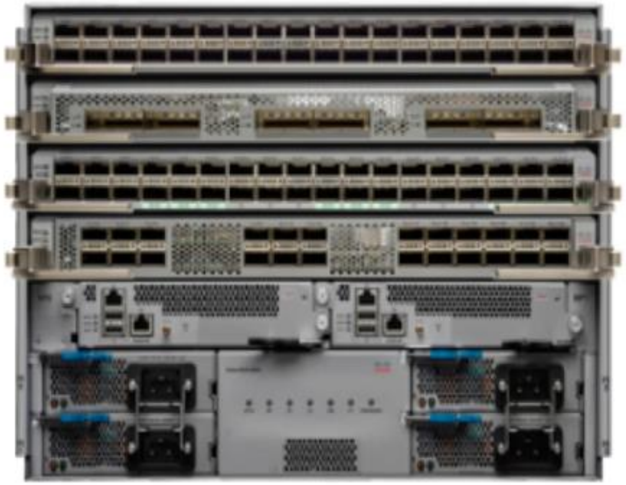
>12 Years trend

- 80x Bandwidth Increase
- 90% Improved Energy Efficiency

Reducing Power Consumption per 100Gbps



Changing The Network – Lower Footprint and Power

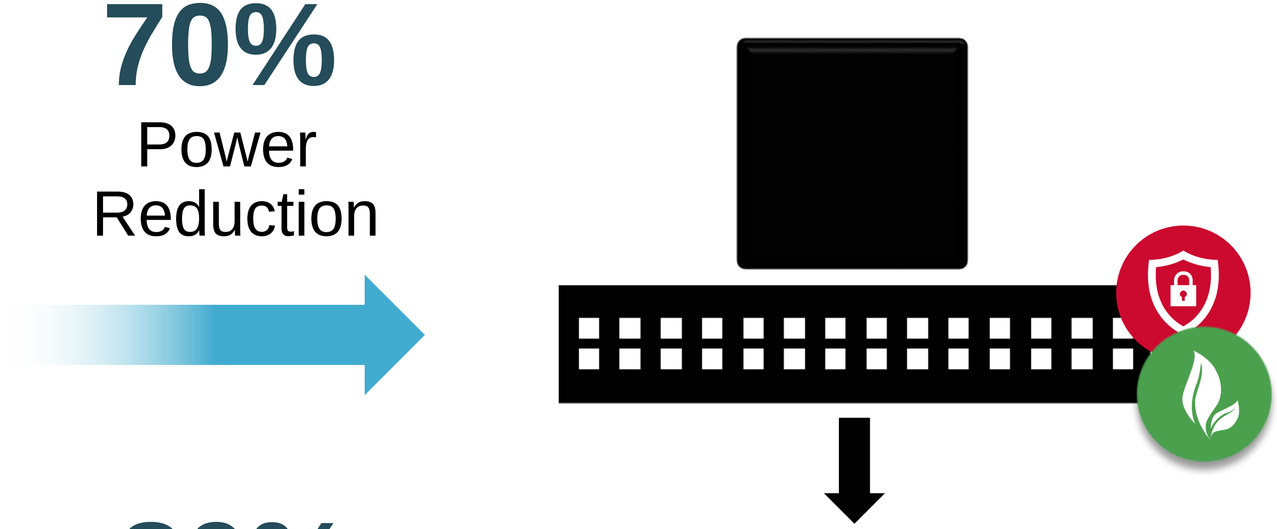


7 RU
3 of 4 Line Cards
6 Fabric Cards

64 x 400GE

70%
Power
Reduction

80%
Space
Savings
(modular to fixed)

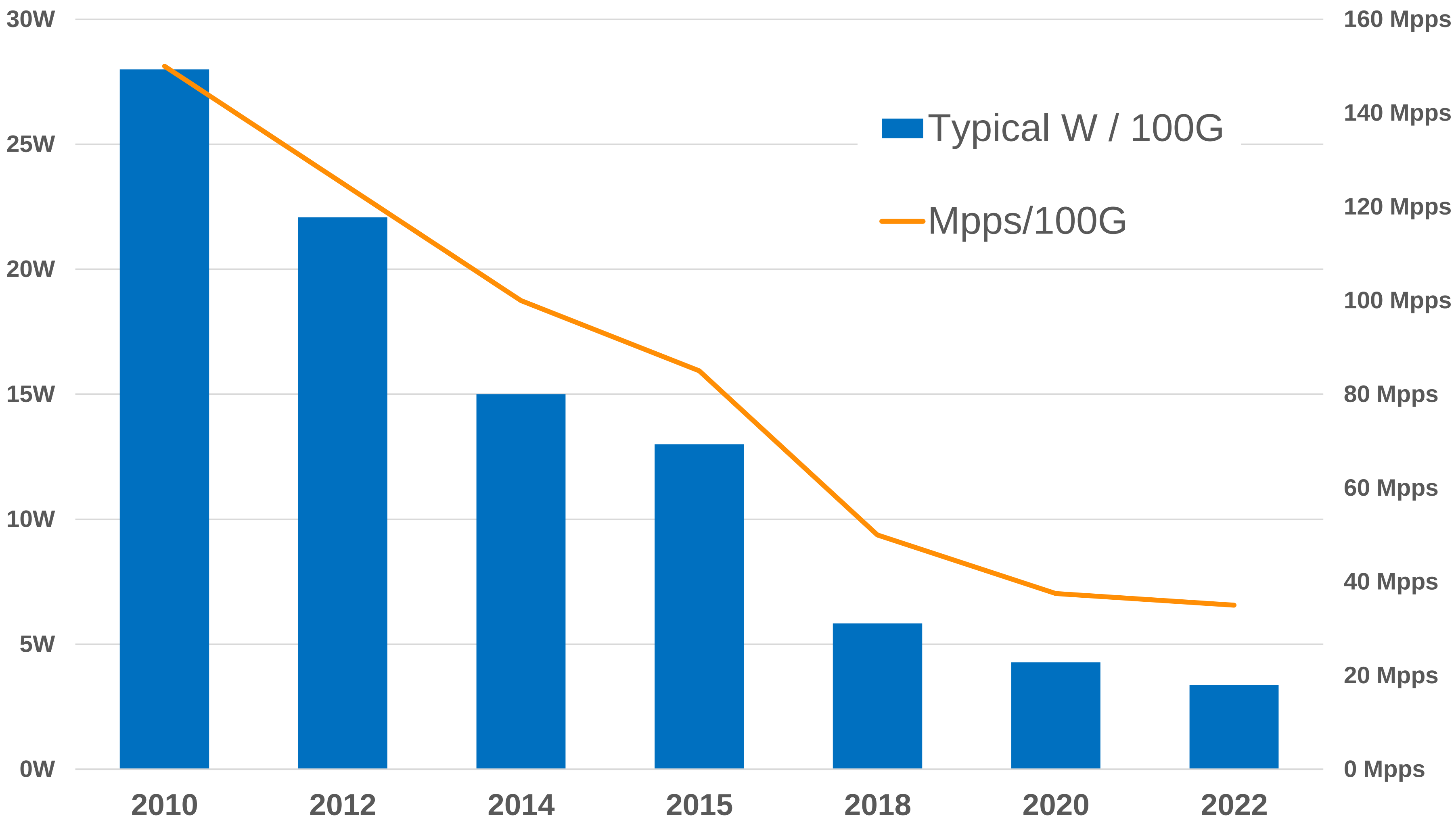


32 x 800GE /
64 x 400GE

Switch devices power optimization

- Focus on power reduction and optimizations
- Enable long term power consumption reduction
  - Measured in Watts per 100Gbps
  - As appears on previous slides
- Architecture optimized for power
  - Higher port radix, Enables reducing tiers
  - Optimization for higher average packet sizes
- Higher integration level
  - HBM instead of external DRAM devices
  - Internal MACsec cores instead of external MACsec PHYs
  - Co-Packaged Optics
- Control of clock trees
  - Clock shutdown of unused blocks
  - Save on dynamic idle power
- Clock gating
  - If not toggling no power is consumed
  - Reduced packet rate / data rate reduces dynamic power proportionally
- Power throttling
  - Static or dynamic option to control the switch power
  - Power control achieved by performance throttling, limiting packet rate and/or data rate

Power-Driven Roadmap / Spec



Power types

- Leakage – Power consumed by the device when held in HW Reset
  - Variability due to process and temperature
- Idle Power
  - Power consumption after SW initialization, with no traffic.
  - Include Leakage, Clock Trees, Interfaces power (including SerDes), control plane (e.g. PCIe)
  - Device configuration has significant effect
    - Mainly due to number of Interfaces (number and rate of ports, number of HBM)
    - Unused blocks – e.g. MACsec, Fabric

- Dynamic power – Power consumed by activity of flops
  - Affected by traffic packet rate and traffic BW
- MAX device power – maximum power consumed by device across all configurations, traffic patterns, process, voltage, temperature

Example for device power partition

	%
Max power	100%
Leakage	16%
Full configuration Idle power (excluding Leakage)	47%
Fabric (including all Fabric SerDes)	13%
NIF SesDes	13%
Dynamic activity power	36%

– Typical device, 105˚c die temp

Power Throttling

- cTDP - Configurable Thermal Design Power
- cTDP in Processors:
  - Modern processors come with a specific Base Frequency and a specific TDP
  - Configurable TDP options mean the computer manufacturer can modify the Base Frequency and TDP of the CPU within the specific values
  - Depending on the design of the chassis/cooler, the computer manufacturer may increase or decrease Base Frequency and TDP
- Power throttling – enbles similar capabilities as cTDP in switches
- Applications
  - Tune max power by packet rate or BW capping
  - Dynamically reduce performance to limit power consumption or temperature based on
    - Power supply current sensing
    - Device temperature
    - External to the device information, Fan fault as an example

Three-Pronged Approach to Reducing Interconnect Power

Linear Drive Optics

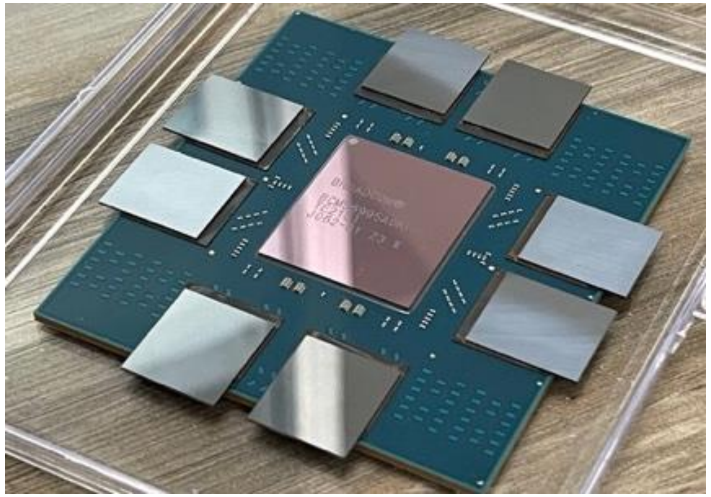
(in addition to retimed pluggables)

25% - 50% Reduction in Optics Power

Extended Reach for Copper Cables

Four Meter DAC (2x IEEE spec)

Co-Packaged Optics

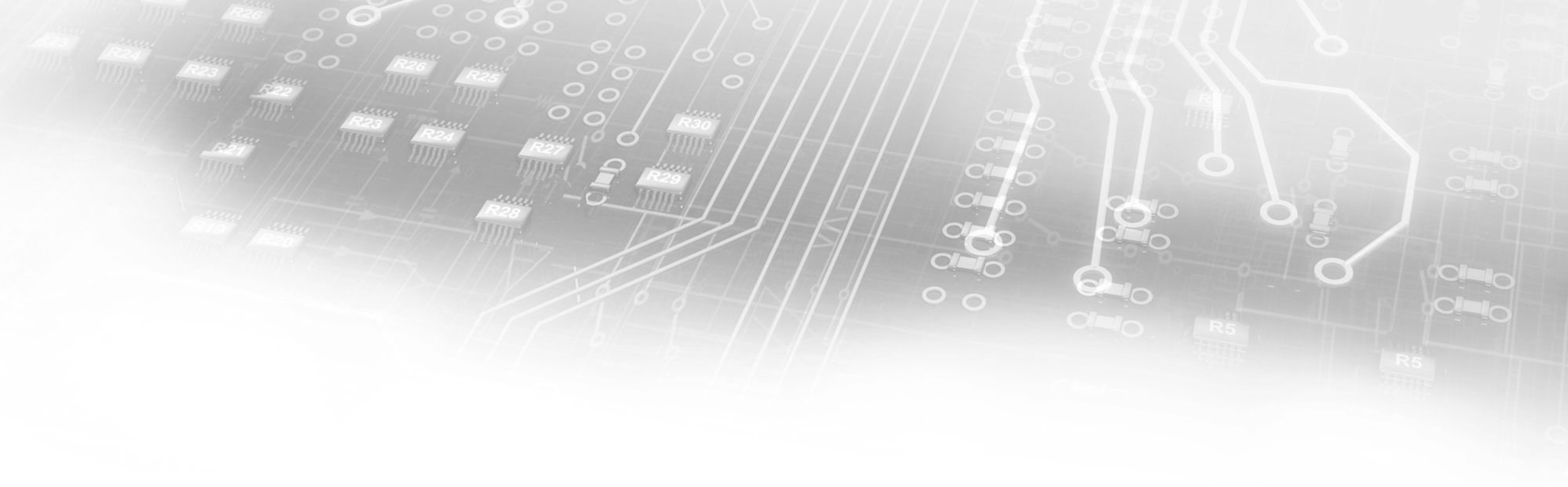


Lowest Power and Cost Optics

New Paradigm for Network Interconnect: Includes Features from Copper SerDes and Optical DSPs

Summary

- Switches silicon has a long trend of doubling BW every 18-24 Months
- Switches energy efficiency (W /100G) has decreased by 90% from 2010
- Switches Architecture optimized for power
  - Higher port radix, Enables reducing tiers
  - Optimization for higher average packet sizes
  - Higher level of integration
  - Higher BW per port
  - Power throttling
- Keeping Power consumption as low as possible is one of main BRCM goals
  - the others are increasing the BW and time to market



Thanks

BROADCOM®