Green Networking: end to end design and operation

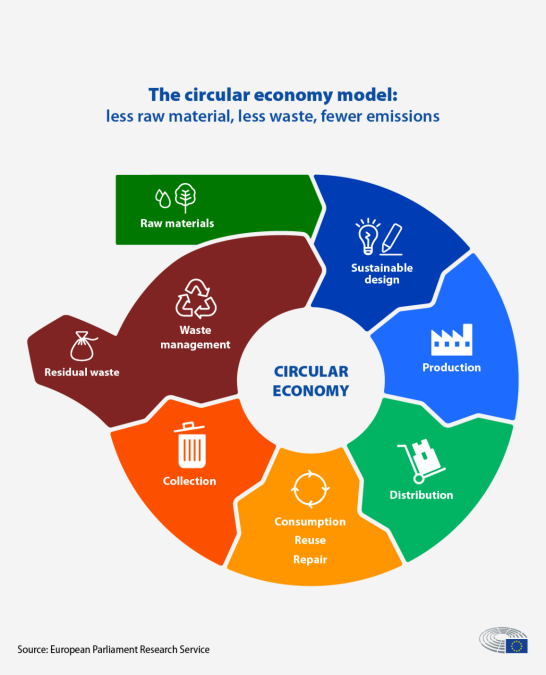
Chip, Device, Protocol and Management level optimizations

AusNOG 2024, Sydney, 5-6 September 2024

Green Networking ESG and Regulation

ESG/Regulation | Power Usage | OPEX cost

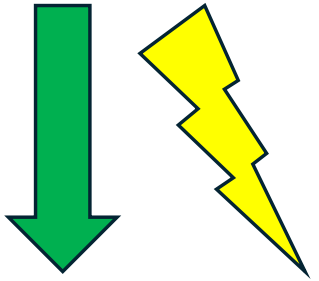
- Increasing pressure on companies and governments to meet green or carbon targets
  - Paris climate agreements -> Greenhouse Gas Reduction -> CO2 Reduce -> Cleaner power -> Less Power
  - Carbon footprint is greenhouse gases emitted to do an activity: From a communications equipment perspective how much power does the device use and what source does the power come from. Companies are increasing investment into renewable energy sources or buying carbon credits, farms, sinks, etc.
  - **Scope 1** emissions caused by own usage (direct power usage)
  - **Scope 2** emissions caused indirectly (where does your power come from)
  - **Scope 3** emissions from supply chain (production, delivery and disposal of products that the company uses).
    - Your customers are looking to you for this.
- Green – reduce unnecessary components and packaging, improve product and packaging materials green credentials, reduce weight and excess packaging, improve, enable circular economy



Green Operations in networking mostly = Less power

ESG/Regulation | Power Usage | OPEX cost

- Most focus is on high power utilization areas
  - Various studies show the biggest power savings are in the **access/mobile networks (70%)** vs **Transport (10%)**
  - Other areas of focus are the facilities themselves (efficiency of the data centers and exchanges). From a measurement perspective this is PUE (power usage effectiveness is a multiplier for every bit of power that equipment uses the facility requires PUE more to feed and cool that piece of equipment)
- Increasing discussions and pressure on OPEX cost of which power utilization cost is an element
  - Cost inflation for power and additional cost in investing into renewable energy or carbon offsets



From a power perspective: Priority of cares

Basically…

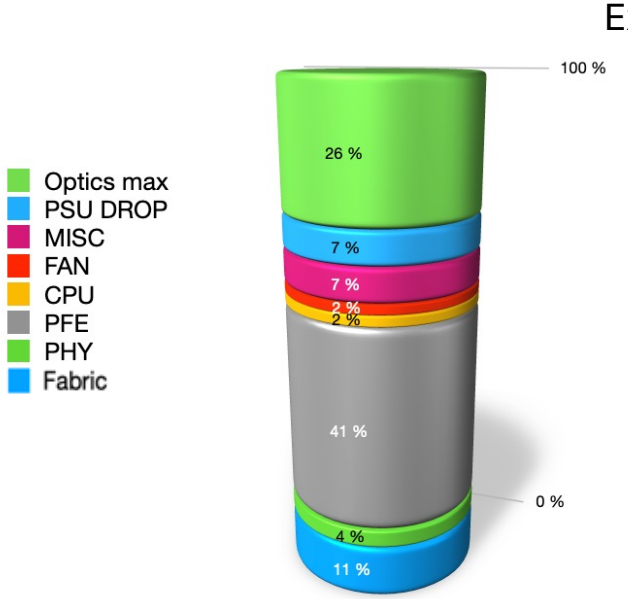
1. Uninterrupted traffic (Highest priority)
2. Design simplicity & SW quality
3. Resiliency & error handling
4. Redundancy
5. Ease of operation
6. Network topology
7. Power Saving (lowest priority)

OPEX cost important and green importance growing

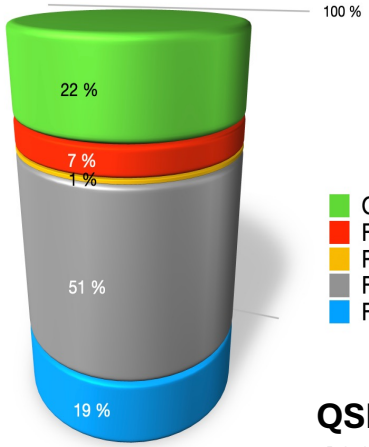
We can target static and dynamic optimization through planning and operation

Device level

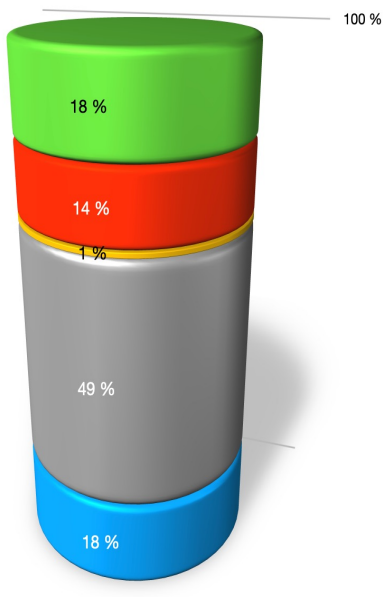
Examples power usage of components



Appliance style device at 20C and 400GE FR / 100G LR optics



25C / 30% IMIX



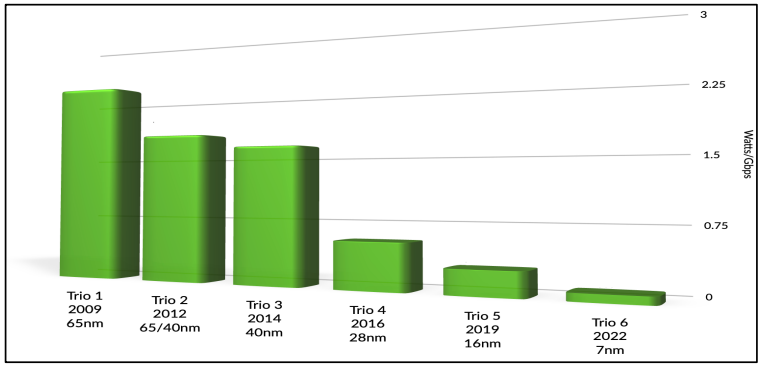
QSFP-DD 400G LR8

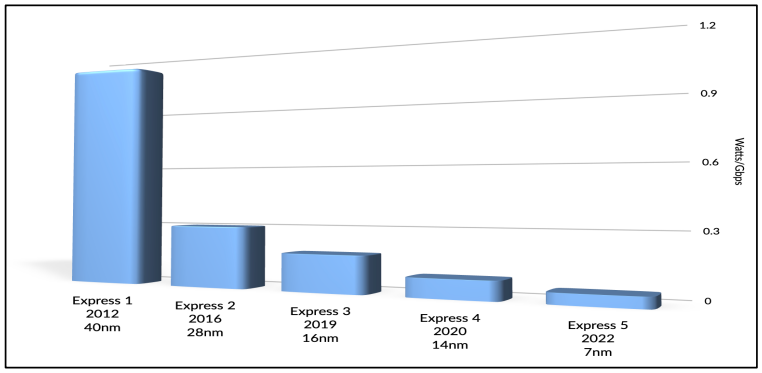
40C / Line Rate IMIX

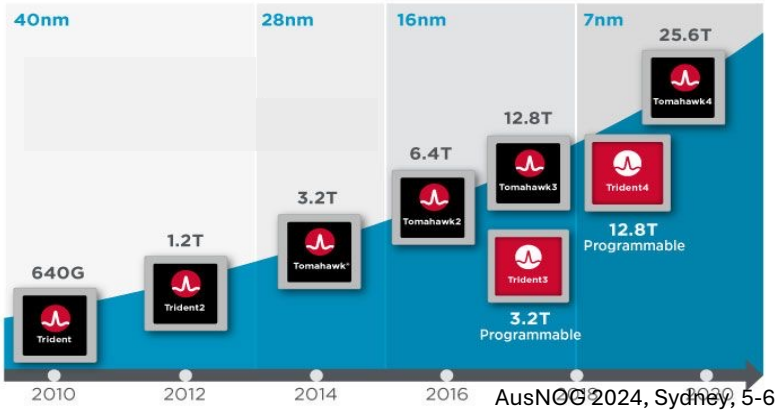
- ASICs and increasingly optics taking up most power usage
  - Most component power usage is non-linear (Idle is maybe 70-80% of maximum power)
  - Better to be off than in a low load/no load state
- Load/Features and temperature have an effect

Chip level

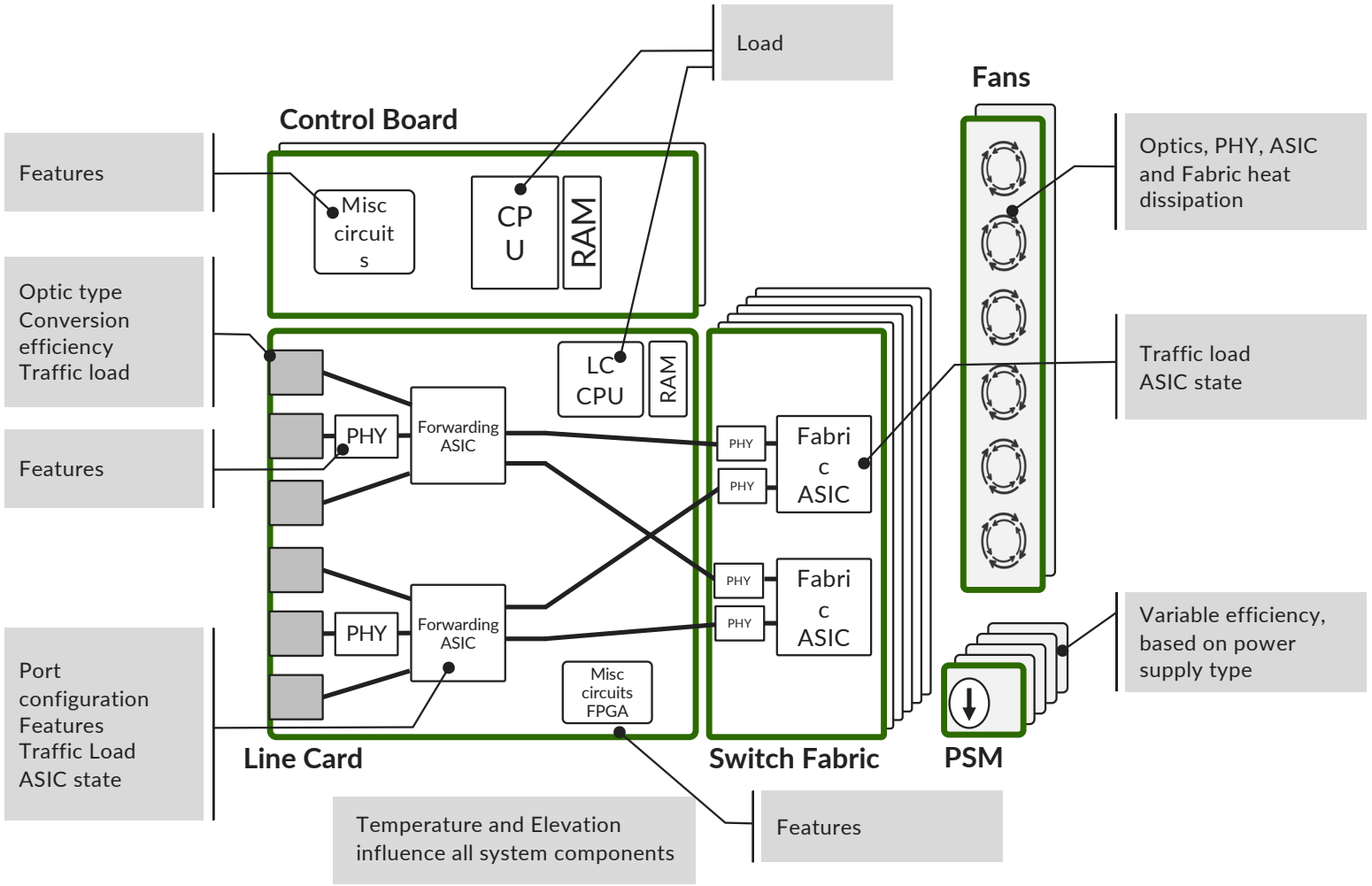
- **Smaller nanometer chips**
  - More efficient in power usage
  - Less charge required to change transistor state
  - Can fit more transistors in the same space
- **Consolidation of functions**
  - Previously: blocks of chips for a single function
  - Separate ASICs/FPGAs/CPUs
  - MACSEC/IPSEC, OAM, Sync etc.
- Clock gating
- Memory access







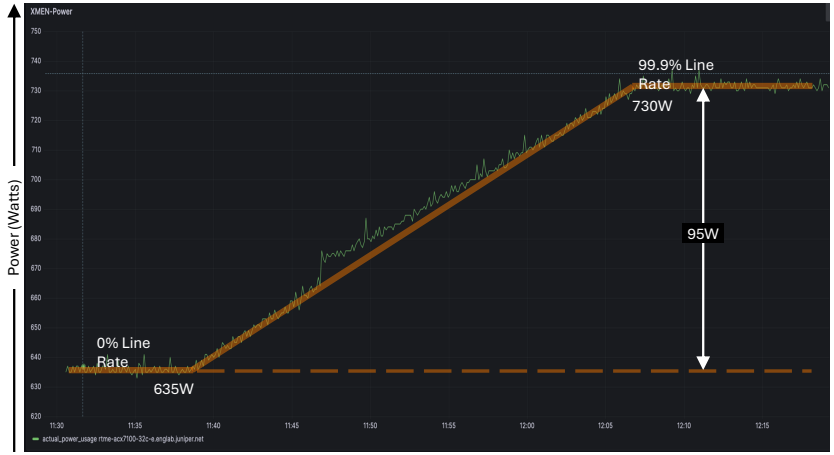
Device Level
Abstracted Modular Chassis Example



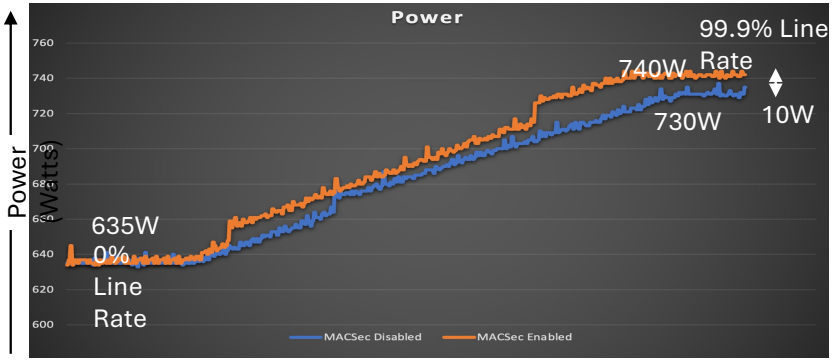
- Traffic load of all components
- Feature usage of all components
- Memory type usage
- Power supply efficiency and conversions
- Optics type and conversion efficiency
- Temperature and Elevation
- Fan load

Device level impact of features traffic Load

Traffic load impacts power utilization in this case single ASIC Broadcom Jericho 2 ~%14



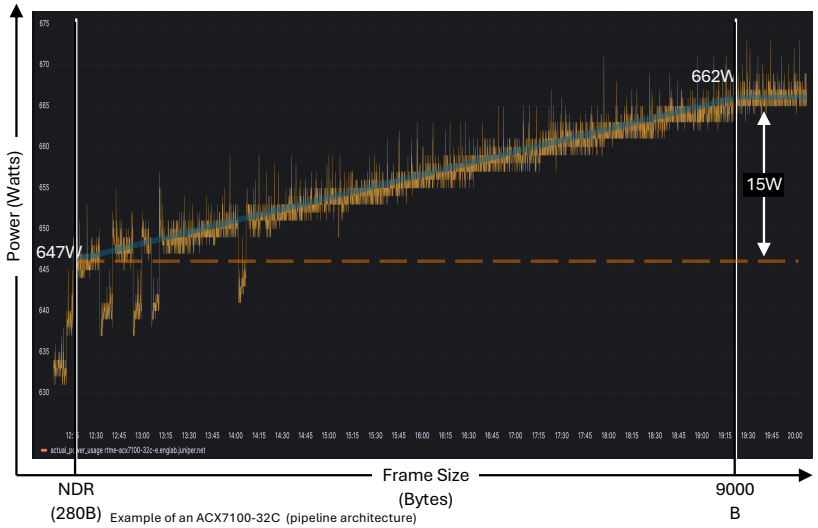
Example of an ACX7100-32C (pipeline architecture)



Example of an ACX7100-32C

MACSEC in Phy didn't result in major power drain on varying load

Varying frame size doesn't result in major power usage difference. fps/pps counts.



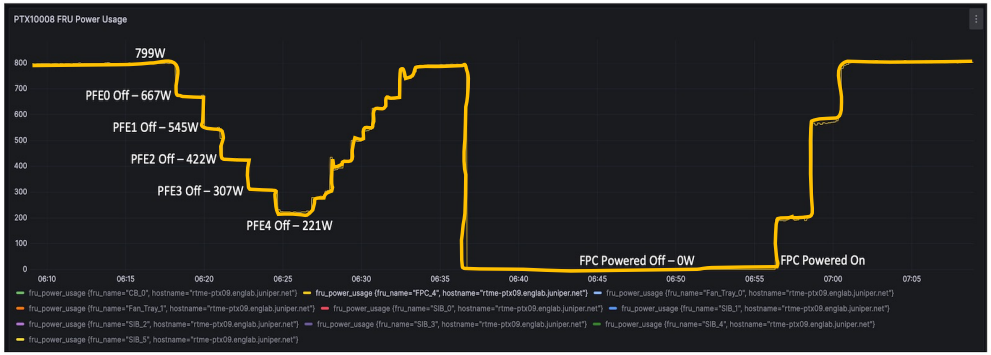
Example of an ACX7100-32C (pipeline architecture)

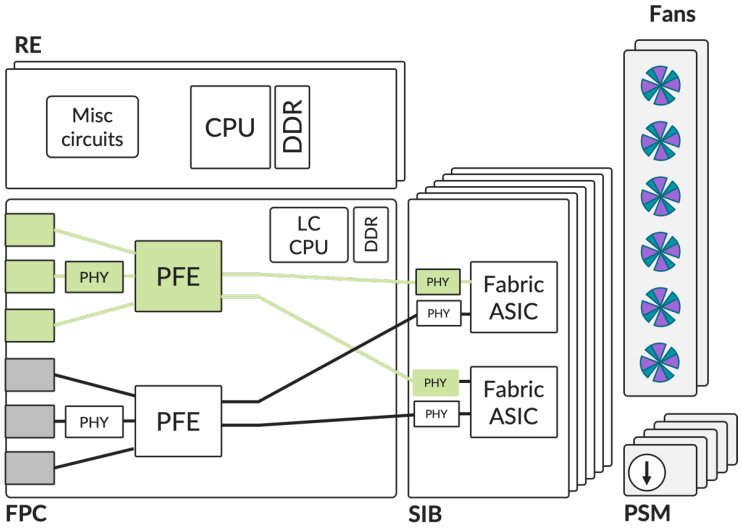
Power utilization fluctuates, changes per design, per chip.

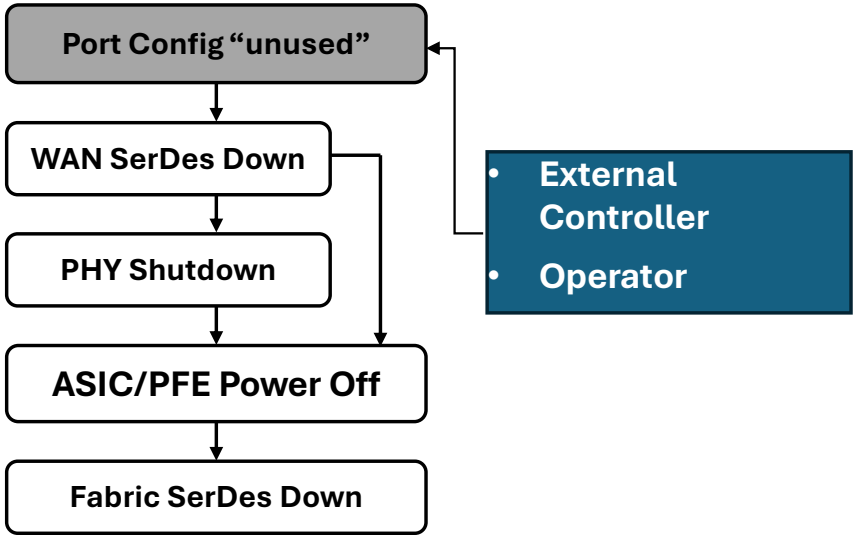
Turn on turn off components at device level

Optimizing power on a device

- Turn off chain, Interface -> PHY -> ASIC/PFE -> Line card -> Fabric





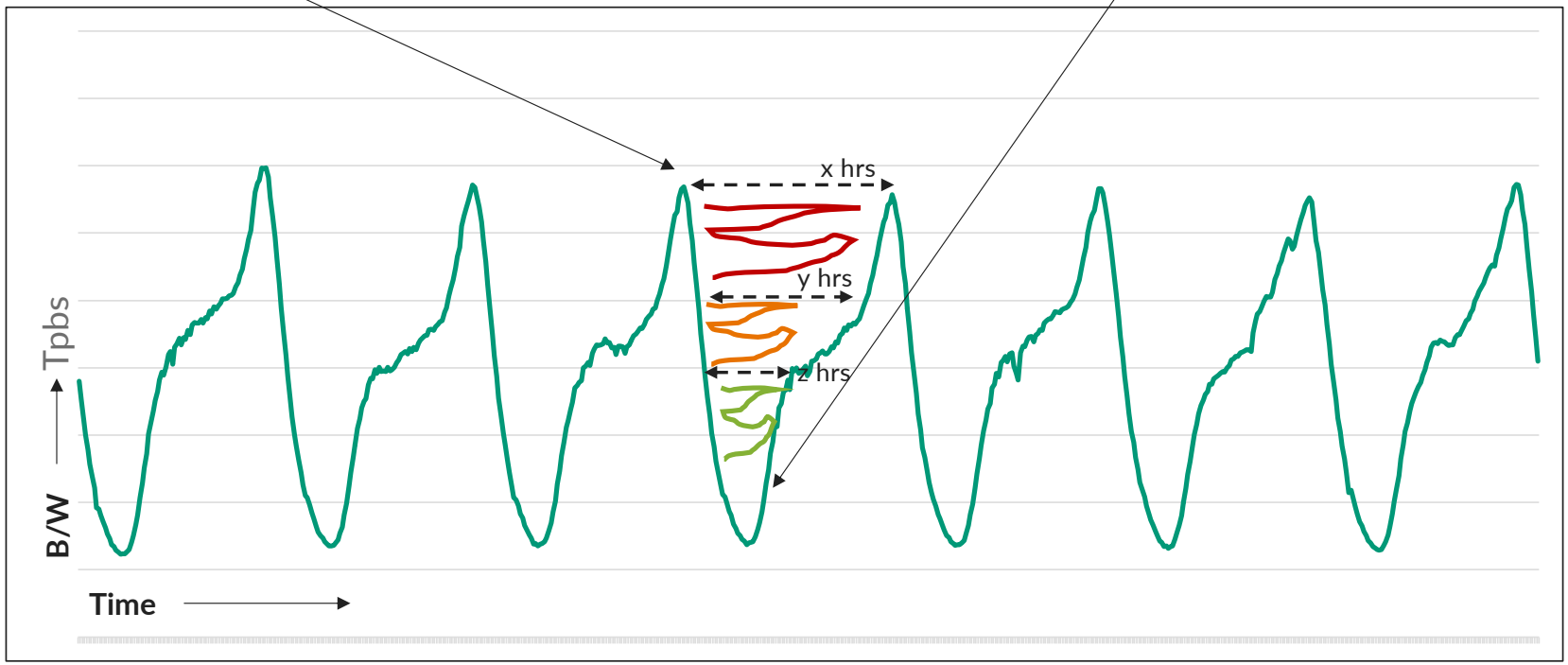


What about in the network?

Networks are designed to Peak load (while also accommodating various failures scenarios)

Large majority of time, the network is highly under utilized

Power (CO_2 & $$$) saving opportunities



Optimise standardization development

Network Level
Green Networking: Observability, Design & Intent

Observability

Insights into the current state of the network:

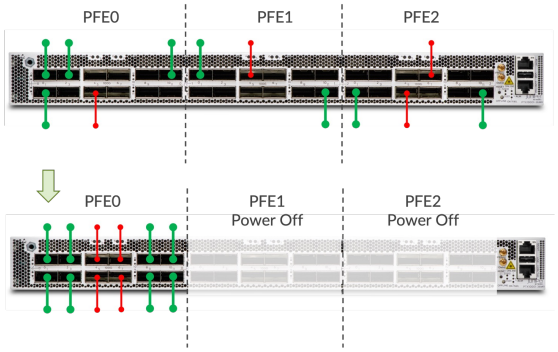
- Component & path level power utilization
- Component & path level CO_2 contribution
- Power / Energy costs ($$$)

Sustainable Operations Dashboard w/ Green TE Insights

Design

Turning off connectivity (IF<-> PHY/GB <-> ASIC) between routers can result in good – medium – bad energy savings
Every device is different and understanding all options and implications for scale, resiliency and power is difficult – Power save modes

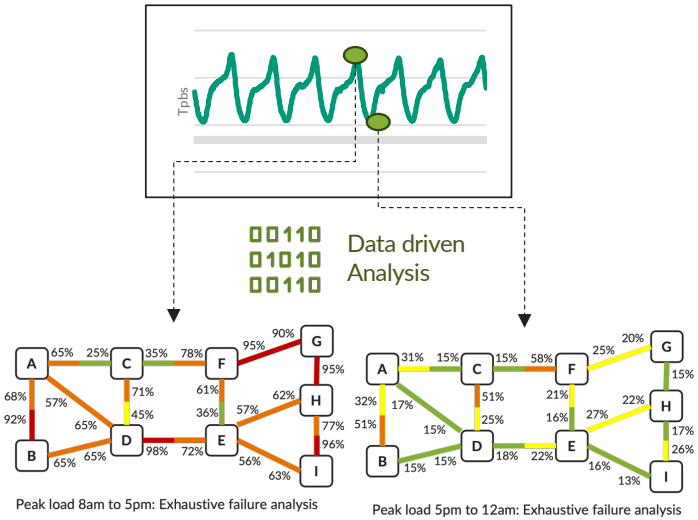
Sustainable systems with Intelligent Power Management



Policy Intent

Never isolate a node, Never compromise HA, …

Accommodate worst case and/or predicted load for a given ToD via Exhaustive failure analysis



Network Level
YANG for power modeling & control

Standardization under development

- Chassis (436W)
  - RE0 (66W)
  - RE1 (66W)
    - State: ON
  - FEB 0 (75W)
    - State: ON
    - PFE 0 (200W)
      - State: ON
    - PFE 1 (200W)
      - State: ON
      - FPC …
        - Interfaces…

- PCE/controller can use YANG to learn about node power architecture and state
- Use the same for control
  - More granularity of control if needed
- More granularity of sensors.
- Allow devices to describe power state and power saving modes
- Standardized for interoperability
- draft-li-ivy-power

Standardization under development

Network Level
Tactical green TE: Modelling power efficiency
Intent based Power efficiency policy

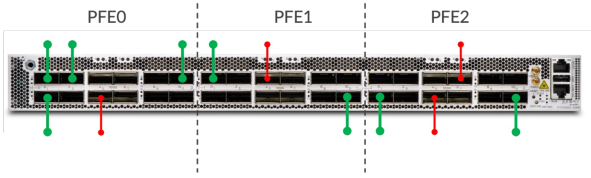
- Modelling H/W power-usage characteristics
  - Precise modelling of power is not recommended as there are factors beyond control of routing protocols
  - '`power-band`': derived minimum / average / maximum power usage of platform components
  - Allows for the definition of an Intent based policy
- Modelling power efficiency as a TE attribute
  - '`power-group:cost`': computed and made available by platform software to routing protocols
  - Represents the relative efficiency of a group within a power-band
  - Used as constraints within TE path profiles in a set of ordered constraints (more detail next slide)

Example power-bands

Band0: Un-rated	Power-group without power-save capability – Always ON
Band1: 5-star energy rating	<= 30W/Tbit
Band2: 4-star energy rating	[30W/Tbit – 75W/Tbit]
Band3: 3-star energy rating	[75W/Tbit – 200W/Tbit]
Band4: 2-star energy rating	[200W/Tbit – 300W/Tbit]
Band5: 1-star energy rating	>= 300W/Tbit

Example power-groups

Pwr-grp1 (PFE0)	EAG 100	Cost 23 (w/g)
Pwr-grp2 (PFE1)	EAG 101	Cost 33 (w/g)
Pwr-grp3 (PFE2)	EAG 102	Cost 30 (w/g)



Network Level
Tactical green TE: TE & state transitions
Supports distributed (on-box), centralized (PCE) & hybrid solutions

Standardization under development

- Power-group based TE path placement
  - Optimize for `power-group:cost` & `bandwidth` using an ordered set of constraints

Rule 0: Place path with only Grey links. If failure, go to next rule

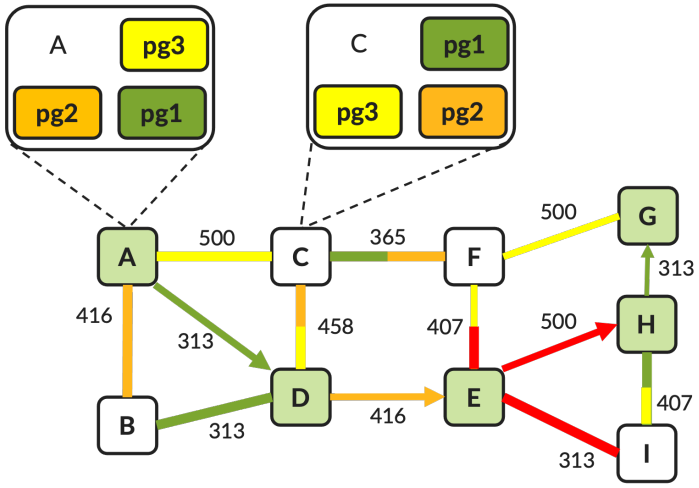
Rule 1: Include links of next band-level (grey + green). If failure, go to next rule

Rule 2: Include links of next band-level (grey + green + yellow). If failure, next rule

... repeat by including next band-level until path is placed.

  - Multiple paths within the same pwr-grp are differentiated by cost
  - '`most-fill`' is the tie-breaker for ECMP

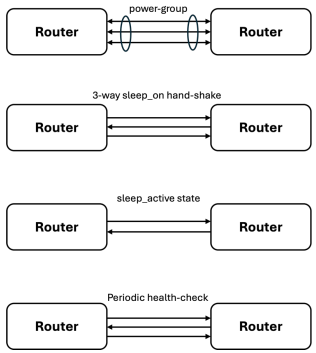
Power aware path placement



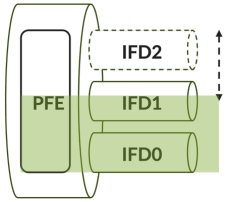
Network Level
Tactical green TE: TE & state transitions
Supports distributed (on-box), centralized (PCE) & hybrid solutions

Standardization under development

- Coordinated power state management
  - Power Management Protocols (LACP &/or LLDP extensions)
  - Discern between '`power sleep`' & DOWN
  - Graceful, reliable, load estimation based state transitions
    - Available capacity is monitored (e.g. max-flow alg) & sleeping links can be proactively awakened (e.g. RscrNotify) to bring back the overall capacity to within a required threshold



Coordinated IFD sleep_state



Pm - Maximum *Predicted* pwr-grp utilization for the binning interval

Pg - Guard Bandwidth. Margin of safety.

Pt = Pm + 2x Pg - high threshold for the decision algorithm reaction

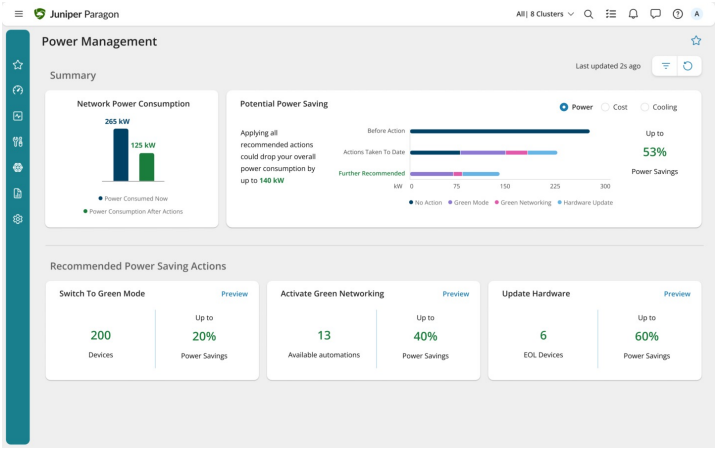
Pc - Instantaneous pwr-grp bandwidth utilization.

L_{bw} - Total pwr-grp bundle bandwidth.

Load estimation / prediction

How to be greener, what steps to take…

- Measure and track usage
- Improve facility efficiency
- Refresh equipment
  - turn off or retire legacy
  - replace with green/power efficient equipment
- Optimize Statically
  - understand & plan device design level optimizations to improve power utilization of devices
- Optimize dynamically
  - explore network level time of day optimization
- Design for low latency
  - at network node level (reduced hops) is also power efficient



An access network in a US Tier 1 ISP
85 nodes, 1500 links
Savings of 58KW out of 207KW
28.0% of PFE/ASIC & link power