Power Management for 10GBASE-T (learning from 802.3af)

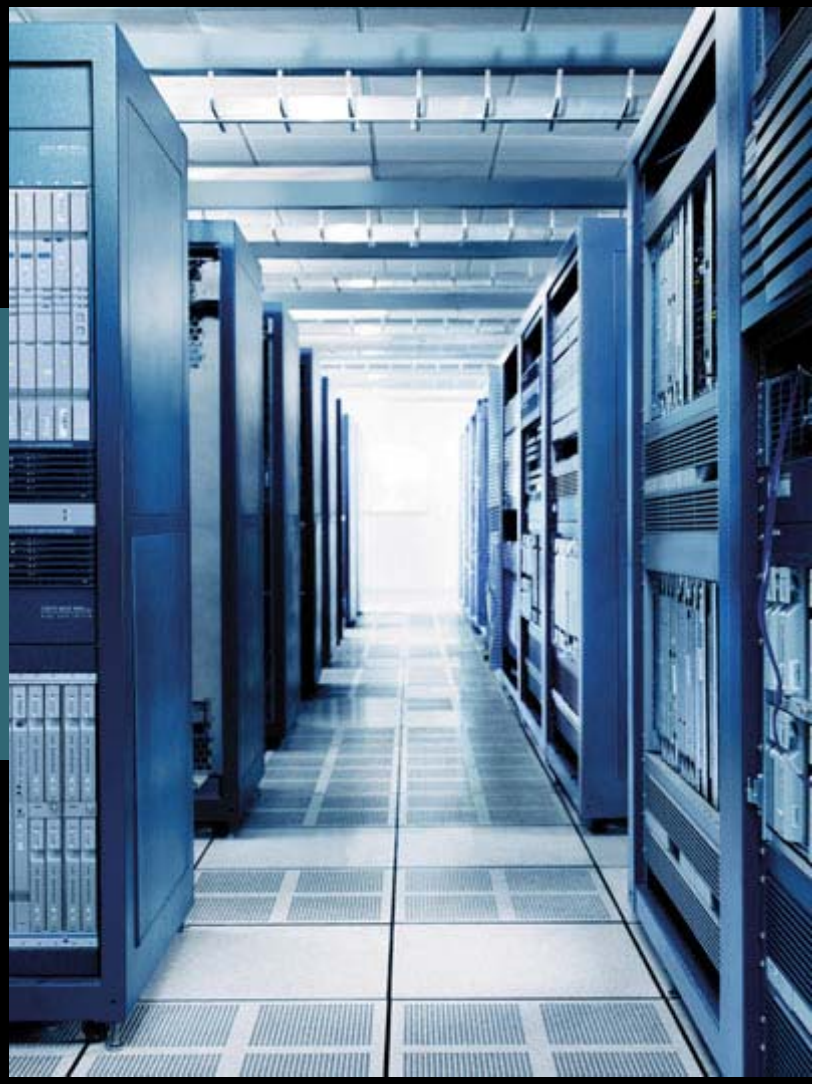
Hugh Barrass (Cisco Systems)

Power management of 10GBASE-T PHYs

10GBASE-T PHY power levels are high enough to warrant power management methods in system

- Early silicon may be as high as 15W (100m), 7W (55m) or 4W (30m)…
  - … same power levels as 802.3af
- Expected distribution of cable lengths centered around 30m
  - Most links don't need full length / full power
- Systems could use power management similar to .3af
  - e.g. oversubscribed design, allocate power / link
- Consider power management methods & control
  - Optimal (but practical) power configurations
  - Usage scenarios

Example

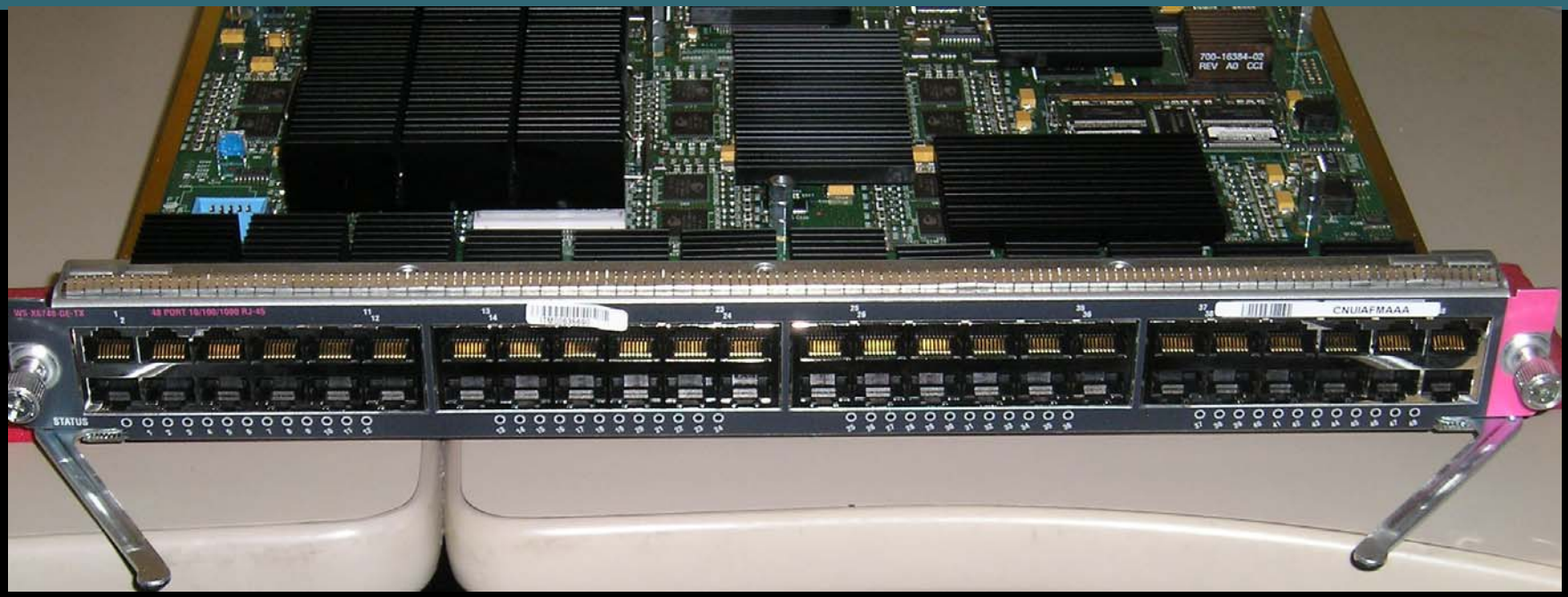


48 port line card (sometime in the future)

Limited power available = oversubscribed design

- Assume quad 10GBASE-T with power save features
  - Full power = 5W; 3.5W for 55m; 2.5W for 30m; 1.5W for 15m
- Power available = 200W / slot
- Package limitation = 15W / quad
  - Allows use of cheaper packaging
- Typical data center deployment
  - Cable lengths distributed with 30m median
- Examine mechanism and s/w behavior
  - (using similar strategy to 802.3af mechanisms)
- Note that pwr saving beneficial – even without power limit

Implementation



- 200W blade, 50W fabric, 25W sundries (BP i/f, conversion etc.)
  - 125W / 48 ports = ~2.5W per port average
- Thermal restriction for packages – 15W per quad
- Sophisticated management s/w allows port prioritization

In practice…

Line card is configured and enabled…

- Ports in a quad start to link, s/w tracks of pwr requirements
  - 1 x 100m link, 2 x 30m link, 1 x 15m link = 11.5W OK
  - 2 x 100m link, 1 x 30m link = limit 4th port to 30m or 15m
  - 2 x 100m link, 1 x 55m link = 4th link at 55m refused
- S/w also tracks total power
  - 3 x 100m link, 7 x 55m link, 26 x 30m link, 12 x 15m link = 122.5W
  - 3 x 100m link, 8 x 55m link, 26 x 30m link, 10 x 15m link = limit next port @ 15m
  - 3 x 100m link, 8 x 55m link, 26 x 30m link, 10 x 15m link = next link @ 30m refused
- S/W may use autoneg state machine to allow fallback operation at 1Gbps
- Port prioritization means that some ports may be “pre-approved”
  - Preferable to forced removal

Other considerations

Some assumptions made for this proposal…

- Power requirements of a 10GBASE-T PHY ~proportional to length
  - TD cancellers; FFT length; Tx power (maybe Rx ADC precision)
  - Gross simplification: Pwr = k x length + const
  - Heavily dependant on particular implementation
- PHY is able to determine length prior to link up
  - TDR during autoneg; link characterization by DSP during training, etc.
  - Could be built in as a separate step if needed – according to implementation
- Detailed control left to implementation – various possibilities
  - PHY automatically selects reduced power mode;
  - System s/w sets limit & PHY determines length before enabling link
  - Intelligence may be in PHY device or with more s/w intervention
- Far end devices do not need any intelligence, or power save features
  - No need to consider two-ended problems, states etc.

Proposal

Sponsor ballot comment already submitted…

- Add register to show length based power save modes
  - Register bits may be read-only, if PHY is fully automatic…
  - … or read/write if system s/w intervention required
  - Also allows force power save mode for power management
- Add test scenario to ensure that system works as advertized
  - Test reduced all length modes supported by system
- Lengths chosen : 55m; 30m; 15m
  - Approximately logarithmic power distribution…
  - …most efficient for power management
- All power save modes optional
  - PHY/system may implement some, none or all modes in any combination
  - No need for far end system to know or care

Q and A


CISCO SYSTEMS