Energy Efficiency and 40GBASE-T

Hugh Barrass

Cisco

IEEE P802.3bq

PHY ad-hoc

Supporters and Contributors

Your name

Could be here

Energy Efficiency & 40GBASE-T

- Typical power a critical parameter
  - Perception failure for 10GBT – learn lesson
  - Aim for 40GBT competitive with QSFP

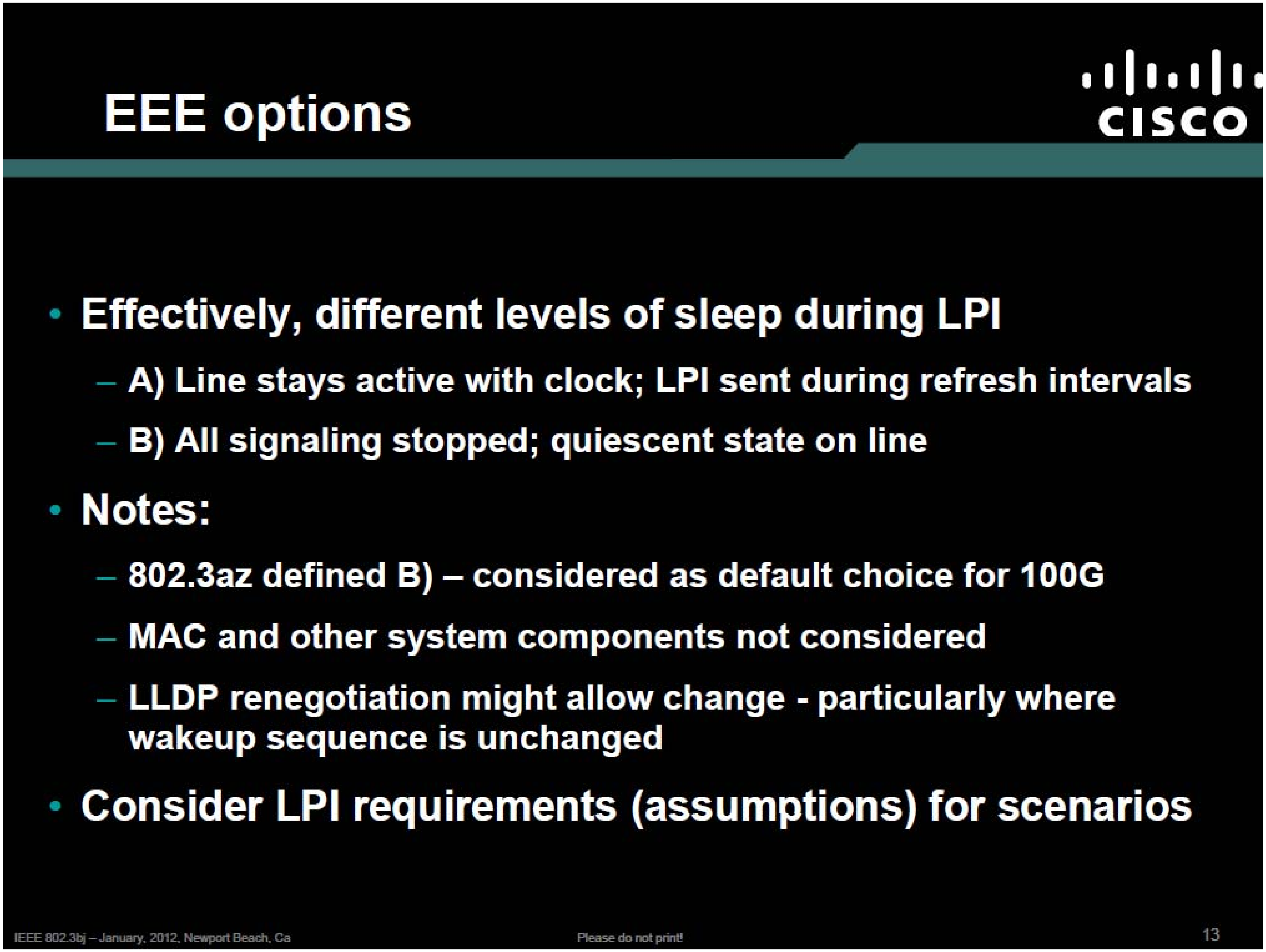
- Worst case PHY power still matters
  - Limiting factor for port density

- EEE – LPI power levels critical
  - Both deep sleep and fast wake
  - Trade energy savings vs. usage

Consider arguments for Fast Wake

- Fast Wake was introduced in 802.3bj
  - Justifications based on backplane copper
- Fundamental principles the same:
  - Fast Wake – more effective at high util
  - Also, less disruption to applications
- Review what was done in 100G backplane…

Extract from 100G twinax proposal



EEE options

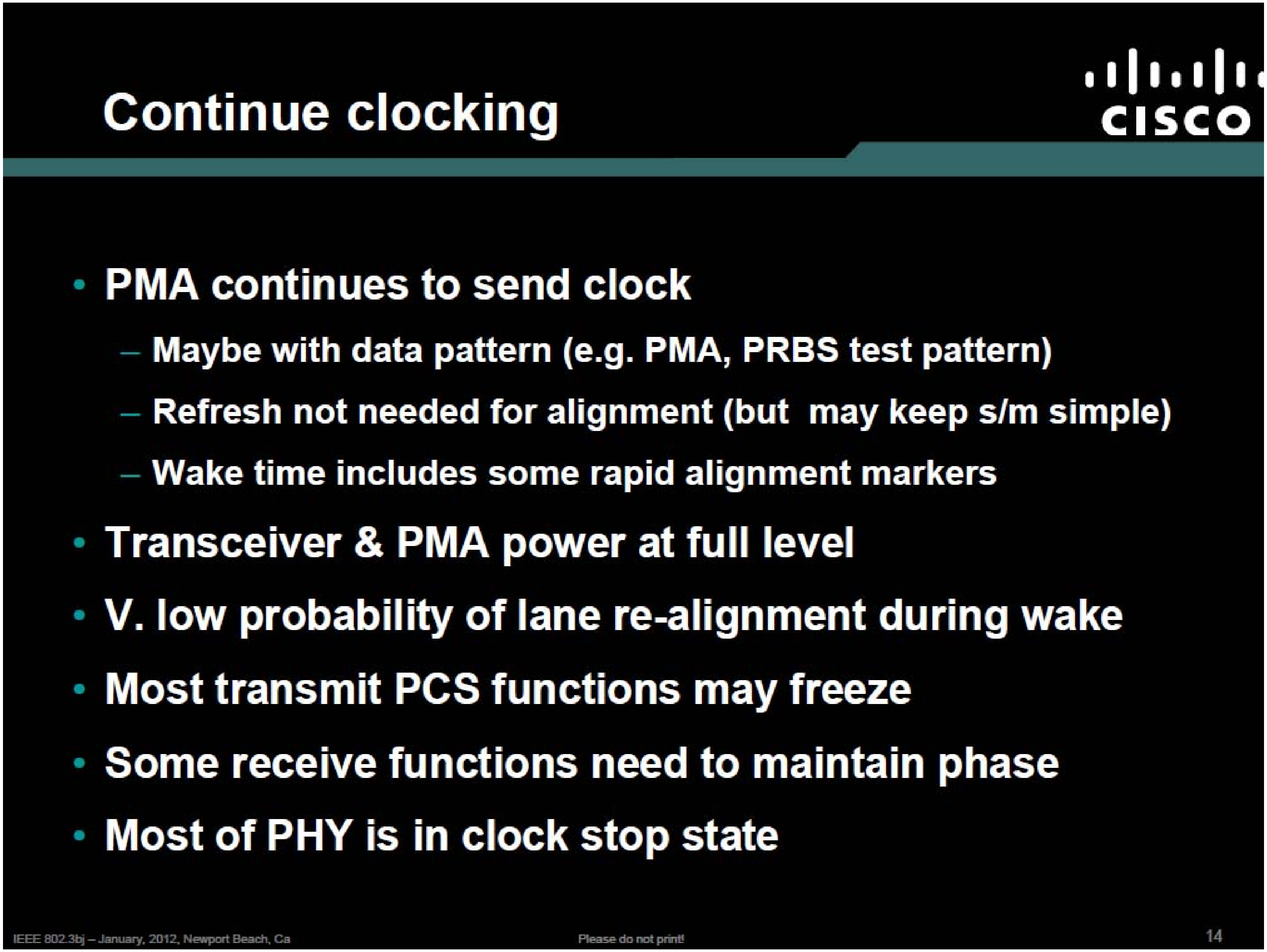
- Effectively, different levels of sleep during LPI
  - A) Line stays active with clock; LPI sent during refresh intervals
  - B) All signaling stopped; quiescent state on line
- Notes:
  - 802.3az defined B) – considered as default choice for 100G
  - MAC and other system components not considered
  - LLDP renegotiation might allow change - particularly where wakeup sequence is unchanged
- Consider LPI requirements (assumptions) for scenarios

IEEE 802.3bj – January, 2012, Newport Beach, Ca

Please do not print!

13

Extract from 100G twinax proposal



Continue clocking

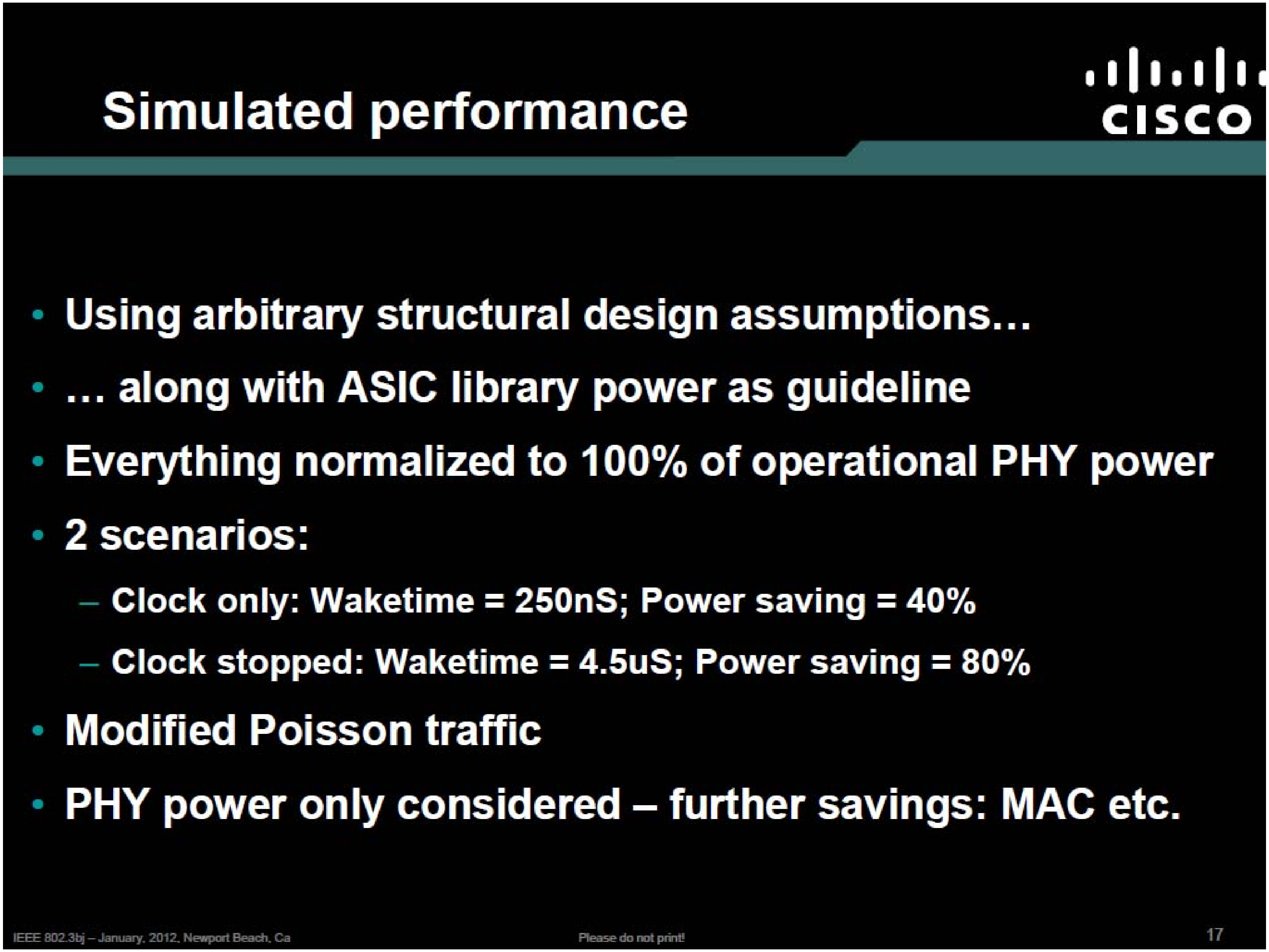
- PMA continues to send clock
  - Maybe with data pattern (e.g. PMA, PRBS test pattern)
  - Refresh not needed for alignment (but may keep s/m simple)
  - Wake time includes some rapid alignment markers
- Transceiver & PMA power at full level
- V. low probability of lane re-alignment during wake
- Most transmit PCS functions may freeze
- Some receive functions need to maintain phase
- Most of PHY is in clock stop state

IEEE 802.3bj – January, 2012, Newport Beach, Ca

Please do not print!

14

Extract from 100G twinax proposal



Simulated performance

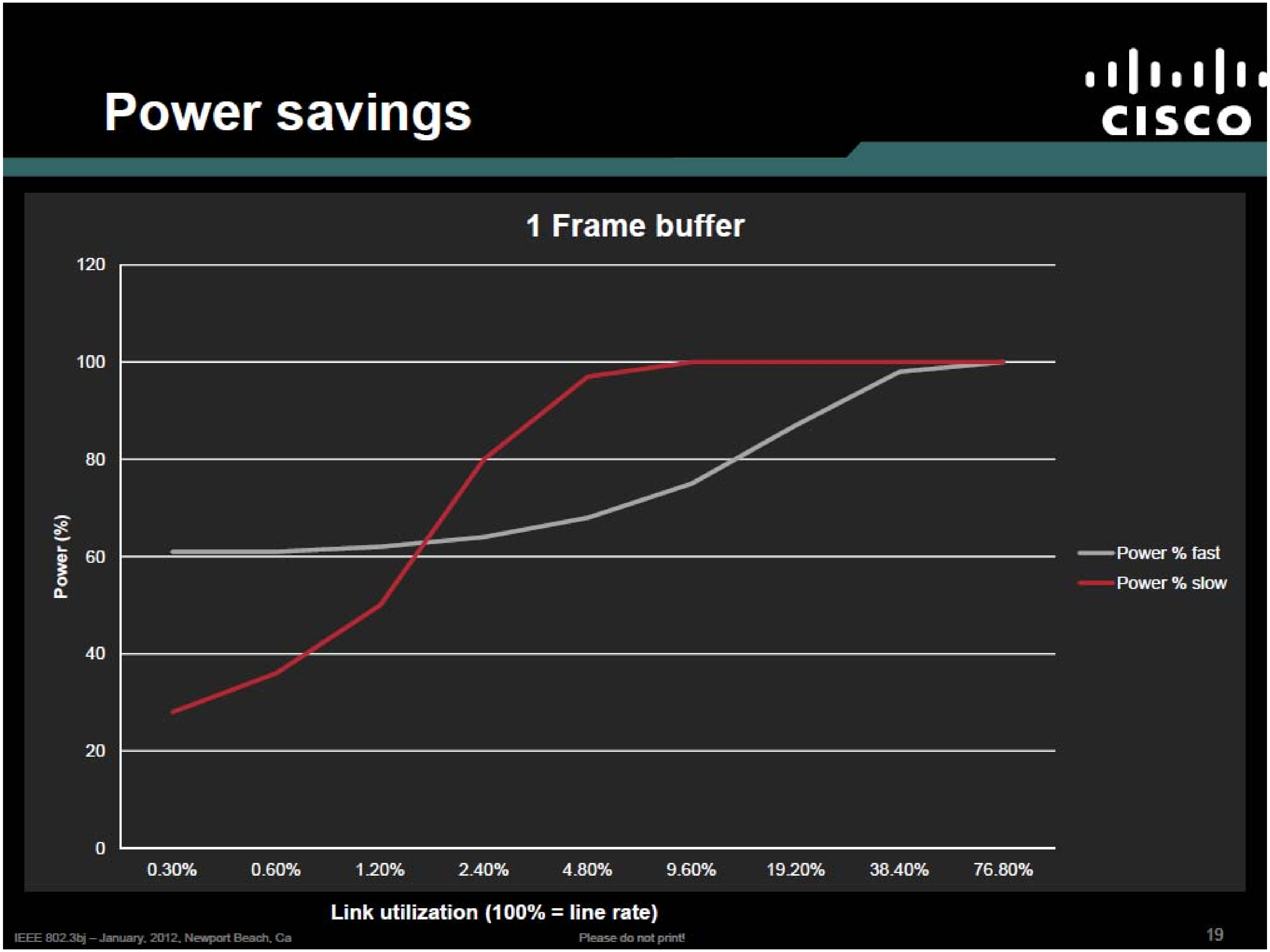
- Using arbitrary structural design assumptions…
- … along with ASIC library power as guideline
- Everything normalized to 100% of operational PHY power
- 2 scenarios:
  - Clock only: Waketime = 250nS; Power saving = 40%
  - Clock stopped: Waketime = 4.5uS; Power saving = 80%
- Modified Poisson traffic
- PHY power only considered – further savings: MAC etc.

IEEE 802.3bj – January, 2012, Newport Beach, Ca

Please do not print!

17

Extract from 100G twinax proposal



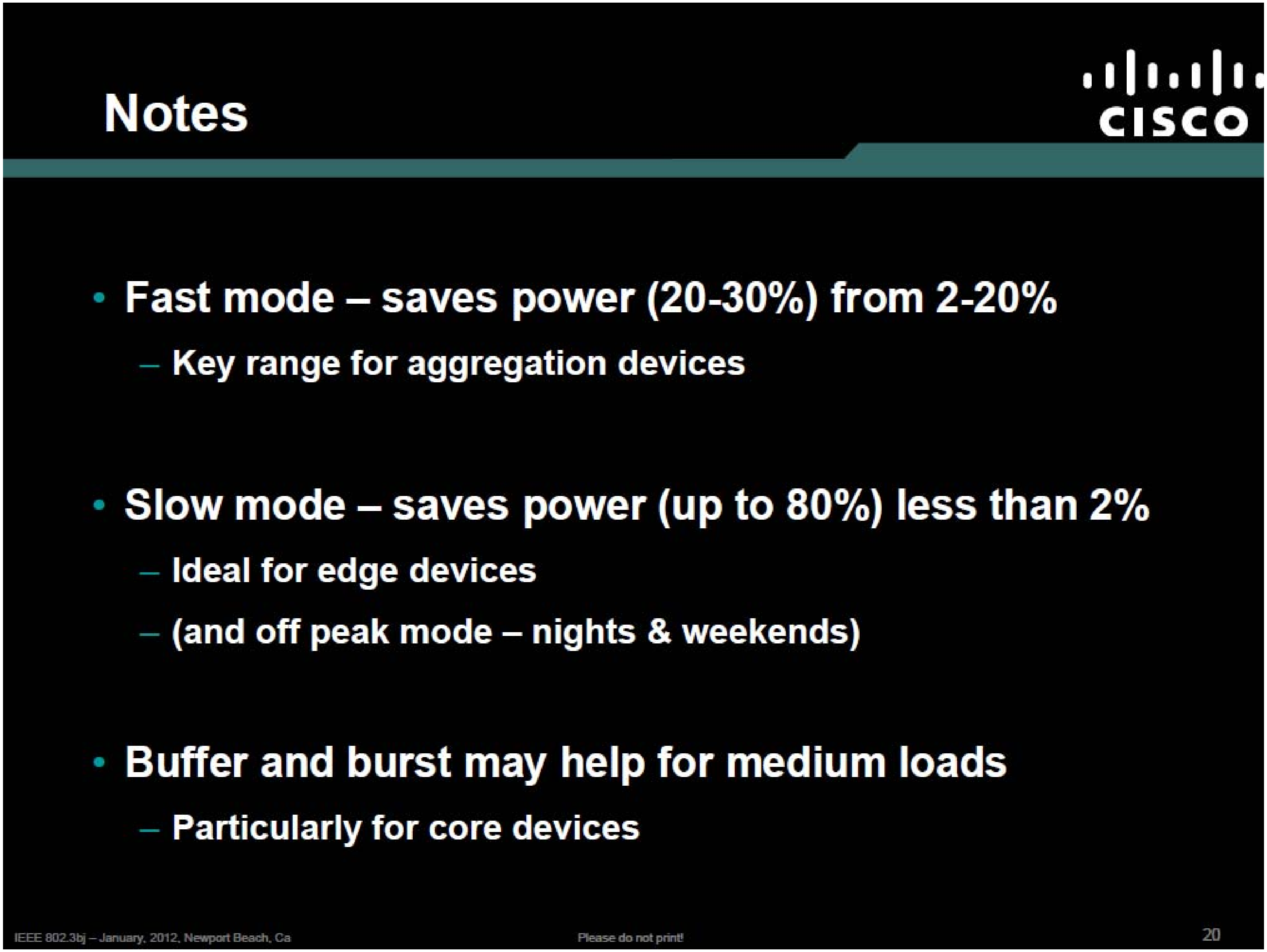
Power savings

IEEE 802.3bj – January, 2012, Newport Beach, Ca

Please do not print!

19

Extract from 100G twinax proposal



Notes

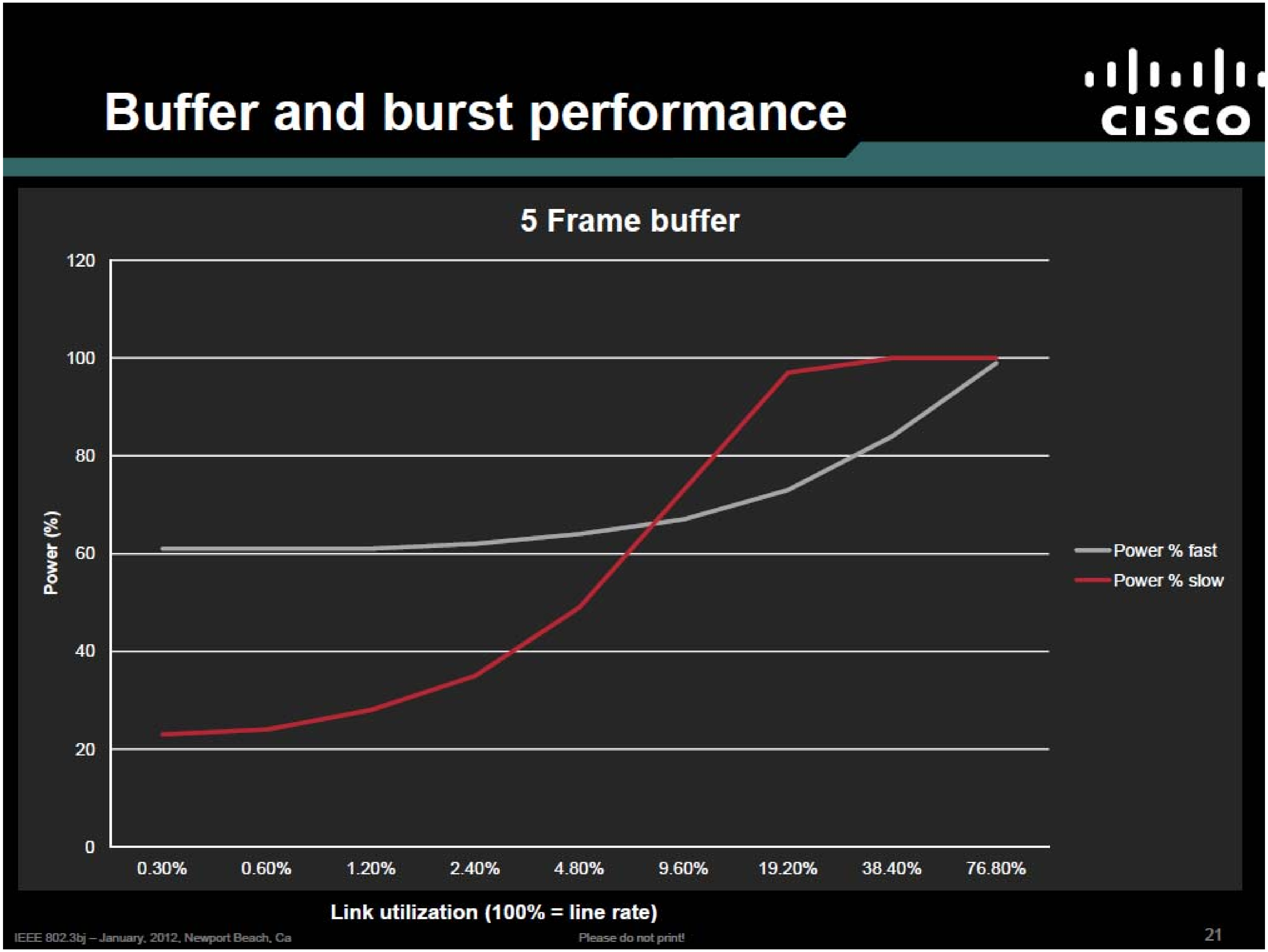
- Fast mode – saves power (20-30%) from 2-20%
  - Key range for aggregation devices
- Slow mode – saves power (up to 80%) less than 2%
  - Ideal for edge devices
  - (and off peak mode – nights & weekends)
- Buffer and burst may help for medium loads
  - Particularly for core devices

IEEE 802.3bj – January, 2012, Newport Beach, Ca

Please do not print!

20

Extract from 100G twinax proposal



Buffer and burst performance

5 Frame buffer

Link utilization (100% = line rate)

IEEE 802.3bj – January, 2012, Newport Beach, Ca

Please do not print!

21

EEE goals, with Fast Wake

- EEE Deep Sleep – similar to 10GBT
  - Transmission ceases, except refresh
  - Up to 80% PHY power reduction in LPI
  - Scaling BT: wake time ~ 1.9us
    - (may be too aggressive)

- EEE Fast Wake – continue sending signal
  - Aim for >40% PHY power reduction
  - Wake time ~ 250ns

40GBASE-T Fast Wake baseline choices

- **First consider analog front-end operation**
  - (~ 50% of power - zimmerman_3bqah_01_1213.pdf)

- Currently defined: DSQ128 (or similar) @ 3200 GBaud

- Options to reduce power:
  - Change to PAM4
  - Reduce Tx power
  - Reduce frequency (e.g. 1600 Gbaud)
- Will these interfere with ability to wake efficiently?

FW coding and information content

- **Fundamentally – only Idle & LPI needed**
  - Possibly: refresh, sleep & wake for transitions
- Needs to retain/re-establish cadence of FEC framing
- Very low information content (< 0.001 bits/baud)
  - FEC not required
  - Predictable data patterns allow robust operation
- Spectral content for filter/canceller maintenance

EEE (LPI) mini-frame

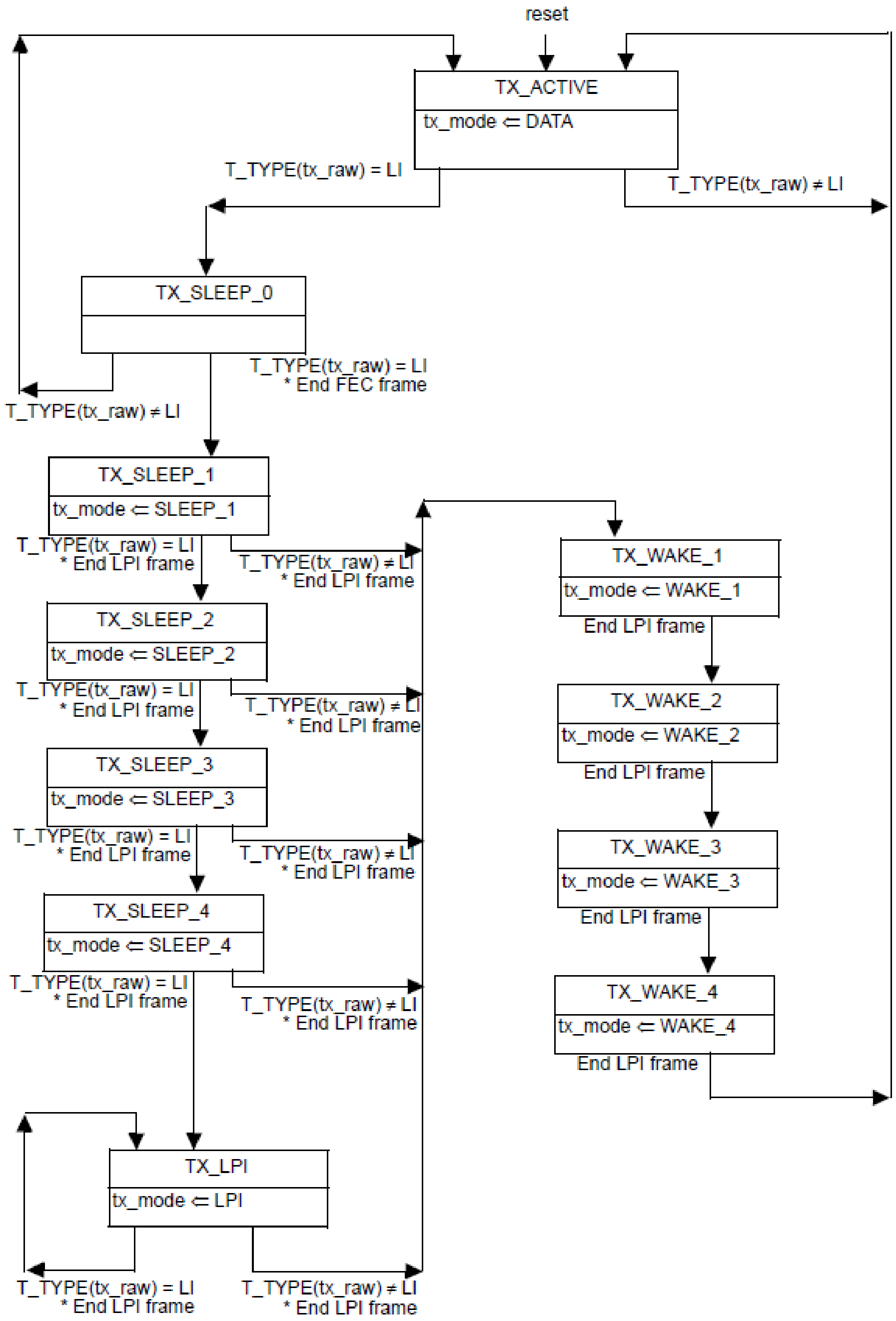
- **Define a frame which is ¼ of LDPC frame size**
  - Allows faster response
  - But has v. low information density
  - 128 x 7 bit symbols (or equivalent)
  - No FEC, just frame structure

- Specific frame types: Sleep-1, Sleep-2, Sleep-3, Sleep-4, LPI, Wake-1, Wake-2, Wake-3, Wake-4

- State machine defines when to send frames
  - Transitions aligned with FEC frame – stop & resume
  - Only ever 2 possible receive options (i.e. 1 bit/frame)

FW Tx state machine

- **Simplified – only FW shown**
  - **(sleep states useful for deep sleep)**
- **The 9 different LPI frames still need to be defined – each one is 128 symbols, easily distinguished.**
  - **Predefined data, but using a scrambled pattern**
  - **Symbols could be 7 bit (DSQ-128), or 4 bit (2D-PAM4) or other depending on analog choices**
- **Rx state machine TBD – should be straightforward**



TBD's

- **Choices for FW analog behavior**
  - How much power can be saved?
  - How much can be changed (keeping ~200ns wake)

- **Definitions for PCS/FEC data coding & framing**

- **Deep sleep behavior**
  - Current assumption – same as 10GBT ~4x faster
  - What can be improved (starting from blank sheet)?
  - Where will #BT have to increase?

Thanks!