EEE for 40G/100G NGOPTX

Open Issues and Objective Proposal

IEEE 40G and 100G Next Generation Optics Study Group

Michael J. Bennett
Lawrence Berkeley National Laboratory

Wael William Diab
Broadcom Corporation

Contributors and Supporters

Brad Booth, Dell

Oren Sela, Mellanox

Steve Carlson, HighSpeed Design

Ryan Latchman, Mindspeed

Alexander Umnov, Huawei

Hugh Barrass, Cisco

Topics

- Brief Recap of Motivation/Application of EEE in This Project
- Open Items
- Objective Proposal

Motivation/Application Recap

- Reduced power consumption adds to market potential
  - Users want energy-proportionality[1,2]
- System power-savings potential is much greater than PHY power-savings
  - See dove_02_05_08.pdf (slide 5)
- More efficient to consider EEE in the initial specification
- There is consensus
  - Straw Poll at May Interim:
    - Support the consideration of EEE in this project Y: 56 N: 3 A: 10

[1] L. Barroso and U. Hölzle, The Case for Energy-Proportional Computing. Computer, 40(12):33-37, December 2007

[2] http://www.ethernetalliance.org/wp-content/uploads/2012/02/EATEF_Panel-3_Power_12_0216.pdf (slides 51, 52, 56)

Open Items

- Several questions were raised at the May interim meeting. The following are addressed in diab_01_0712_optx.pdf:

- L2 education
- Capability exchange
- System savings
- Alternative to Auto Negotiation

Open Items

- The following were addressed by Hugh Barrass and are shown in the backup slides
- Utilization
- Traffic profiles

- Fast Wake and Sharing Capability Exchange with P802.3bj can be discussed and developed further if there is interest and consensus, but it should not stop the work from moving forward

Objective Proposal

Proposed text for objective:

Specify optional operation for exchange of EEE capability and state on Next Gen 40G and 100G optical interfaces

Thank You!

Backup

Motivation for EEE in this project

- EEE could help make the datacenter more energy proportional to load[1].
- End users are asking developers to "make better energy proportionality a primary design objective" for future systems[1].
- Savings for the IEEE 802.3az PHY alone should be around 90% and energy reduced by up to 70% for the NIC when in LPI mode[2].
  - much greater savings possible in systems using LLDP
    - See dove_02_05_08.pdf (slide 5)

1. L. Barroso and U. Hölzle, The Case for Energy-Proportional Computing. Computer, 40(12):33-37, December 2007
2. P. Reviriego, K. Christensen, J. Rabanillo, and J. A. Maestro, 'An Initial Evaluation of Energy Efficient Ethernet' in IEEE communications letters, VOL. 15, NO. 5, May 2011

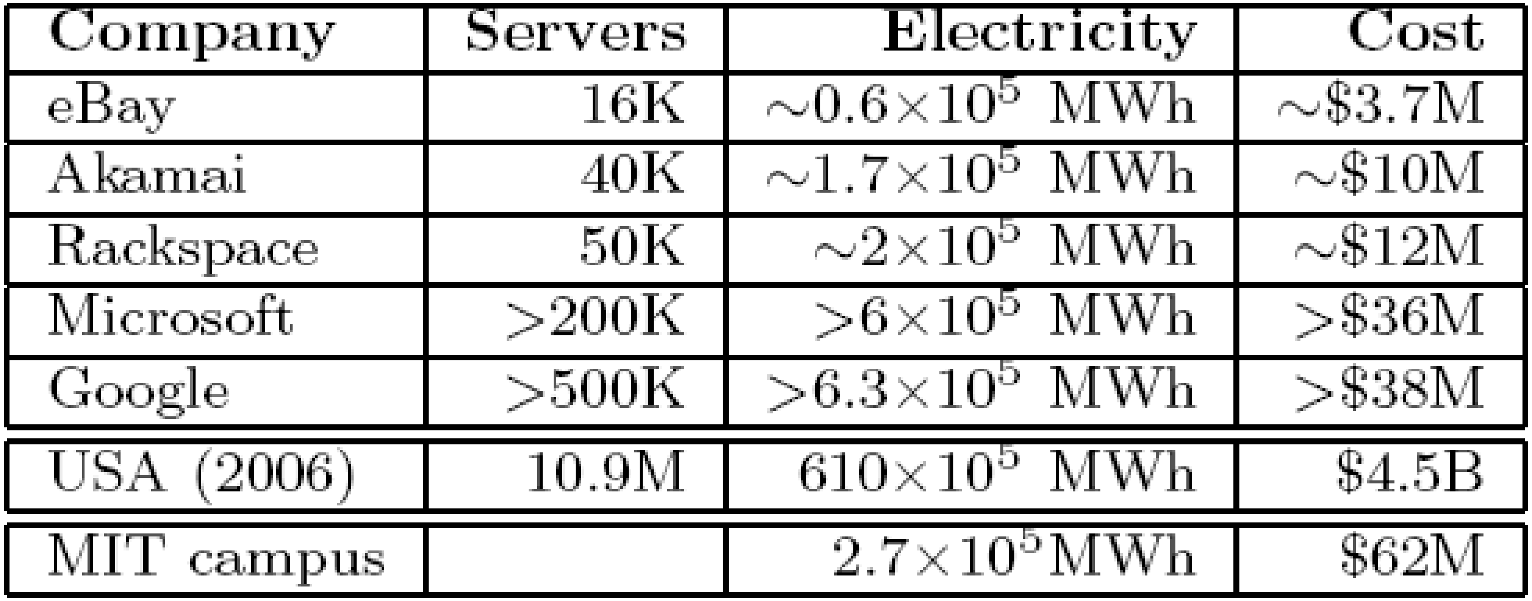
Motivation for EEE in this project

- Data center operators are very much interested in using power efficiently as energy-use impacts operational expense
  - E.g. Google spent ~$200M on energy in 2010
    - Note that Google's data centers are roughly 50% more efficient than others
- Data Center Operators want energy-proportional equipment
- Larger data centers use optical links
  - "Likely a lot of value in figuring out EEE for optical links"

Source: http://www.ethernetalliance.org/wp-content/uploads/2012/02/EATEF_Panel-3_Power_12_0216.pdf slides 51,52,56

Motivation for EEE in this project

- Energy cost is still a significant operational expense in data centers [1]



Company	Servers	Electricity	Cost
eBay	16K	$\sim 0.6 \times 10^5$ MWh	~\$3.7M
Akamai	40K	$\sim 1.7 \times 10^5$ MWh	~\$10M
Rackspace	50K	$\sim 2 \times 10^5$ MWh	~\$12M
Microsoft	>200K	$> 6 \times 10^5$ MWh	>\$36M
Google	>500K	$> 6.3 \times 10^5$ MWh	>\$38M
USA (2006)	10.9M	610×10^5 MWh	\$4.5B
MIT campus		2.7×10^5MWh	\$62M

[1] Cutting the Electric Bill for Internet-Scale Systems, Qureshi et. al, SIGCOMM '09 Proceedings of the ACM SIGCOMM 2009 conference on Data communication, ISBN: 978-1-60558-594-9. Estimated annual electricity costs for large companies (servers and infrastructure) @ \$60/MWh (6 cents / KWh)

Motivation for EEE in this project

- Even in high transaction-rate networks, utilization is not 100% 24 hours/day, 365 days/year = opportunity to save energy[1]

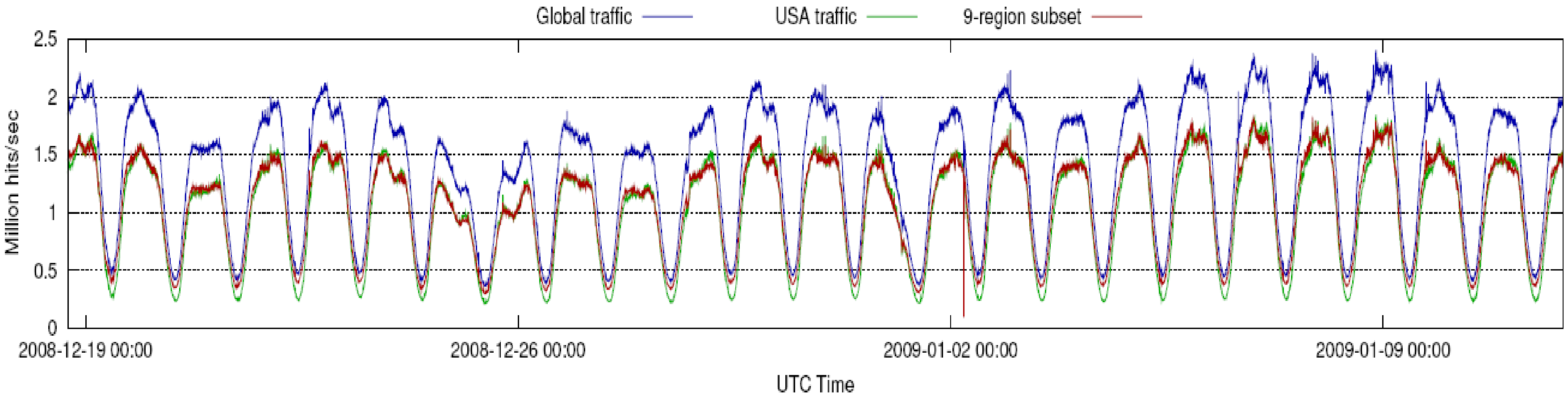


Figure 14: Traffic in the Akamai data set. We see a peak hit rate of over 2 million hits per second. Of this, about 1.25 million hits come from the US. The traffic in this data set comes from roughly half of the servers Akamai runs. In comparison, in total, Akamai sees around 275 billion hits/day.

1. Cutting the Electric Bill for Internet-Scale Systems, Qureshi et. al, SIGCOMM '09 Proceedings of the ACM SIGCOMM 2009 conference on Data communication, ISBN: 978-1-60558-594-9

Motivation for EEE in this project

- Another example of an energy saving opportunity

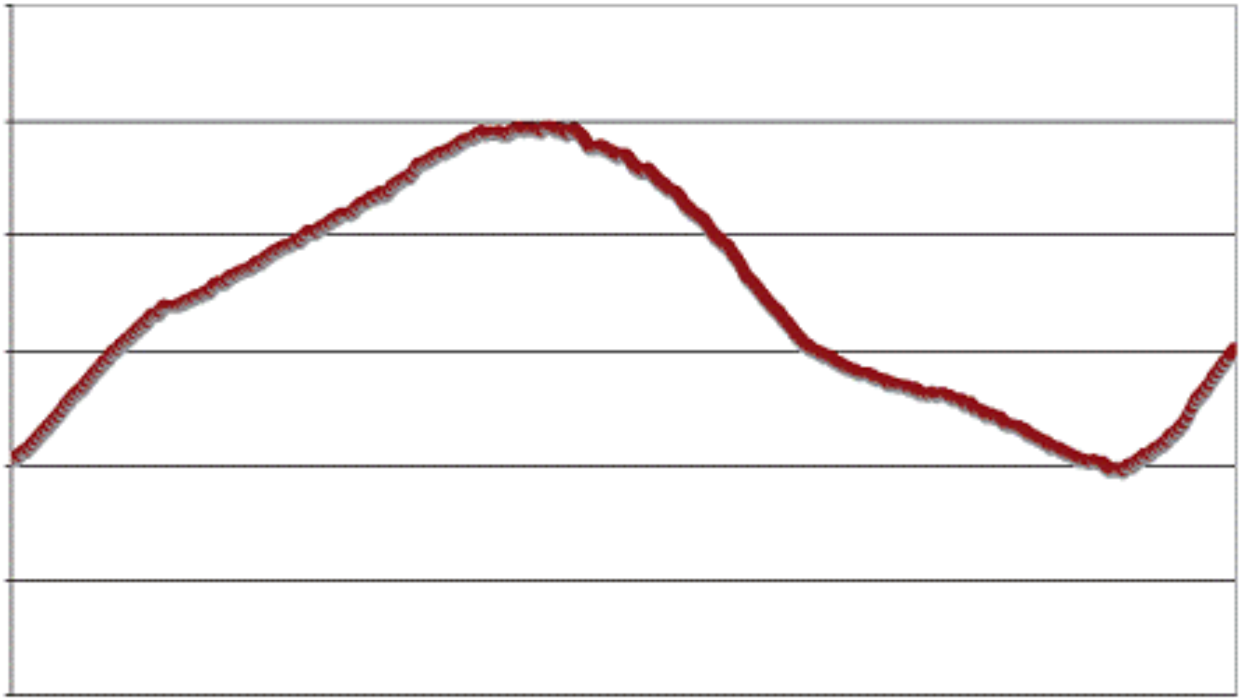


FIGURE 2.2: Example of daily traffic fluctuation for a search service in one datacenter; *x*-axis is a 24-h period and the *y*-axis is traffic measured in queries per second.

1. http://research.google.com/pubs/pub35290.html

Motivation for EEE in this project

- Energy Efficiency is a priority for regulators
  - EU CoC on Energy Consumption of Data Centers
  - Energy Star specs for Small Network Equipment
    - Large Network Equipment coming
  - Policy will encourage technologies like EEE
  - Can support that by including EEE in the specification
- Therefore EEE is a "must" for a new specification

Source: http://www.itu.int/dms_pub/itu-t/oth/09/05/T09050000010004PDFE.pdf
Source: http://www.energystar.gov/index.cfm?c=new_specs.small_network_equip

Motivation for EEE in this project

- EEE should be included *at the beginning of projects*
  - Very difficult and time consuming task to retrofit EEE into completed specifications
  - Run the risk of breaking things
  - Much more efficient to consider EEE in the initial specification

How could this apply to 40G/100G NG Optics?

- Lowest hanging fruit for 40G and 100G NG Optics
  - Use LPI codewords for signaling – no PMD power-down
  - Bulk of the work is being done in P802.3bj
- EEE is showing up in switches and will likely be a feature in most, if not all Ethernet switches by the time this project is finished
  - Including EEE in this project enhances market potential
- Is there interest in working on EEE?

Issues for Optical EEE

- EEE “classic” requires quiescent state on line
- May not be feasible for (all) optics:
  - V. long time to restart lasers after shutdown …
  - … holding at static levels problematic
  - Unknown issues with reliability with power/temperature cycles

- EEE “Fast Wake” as introduced in 802.3bj does not require quiescent state on line – normal Tx continues
  - No Changes to the PMD
  - Useful for saving energy on high utilization links
  - PCS, MAC & other systems savings still apply
  - Longer system-level wake negotiation still possible

Reduced power scenarios

- Data presented in Jan 2012 for 802.3bj still relevant
  - Based on PHY power savings in Copper, fast mode (& normal)
  - Fast mode power similar for fiber & copper (PCS, scrambler, lane alignment, etc.)
  - System level savings in addition – MAC, lookup tables, etc.
- Buffer & burst relevant for core optics
  - Large buffers and latency tolerant traffic common
  - Fiber latency alone often >> 10x max frame delay
  - Buffer & burst works very well for moderately high utilization

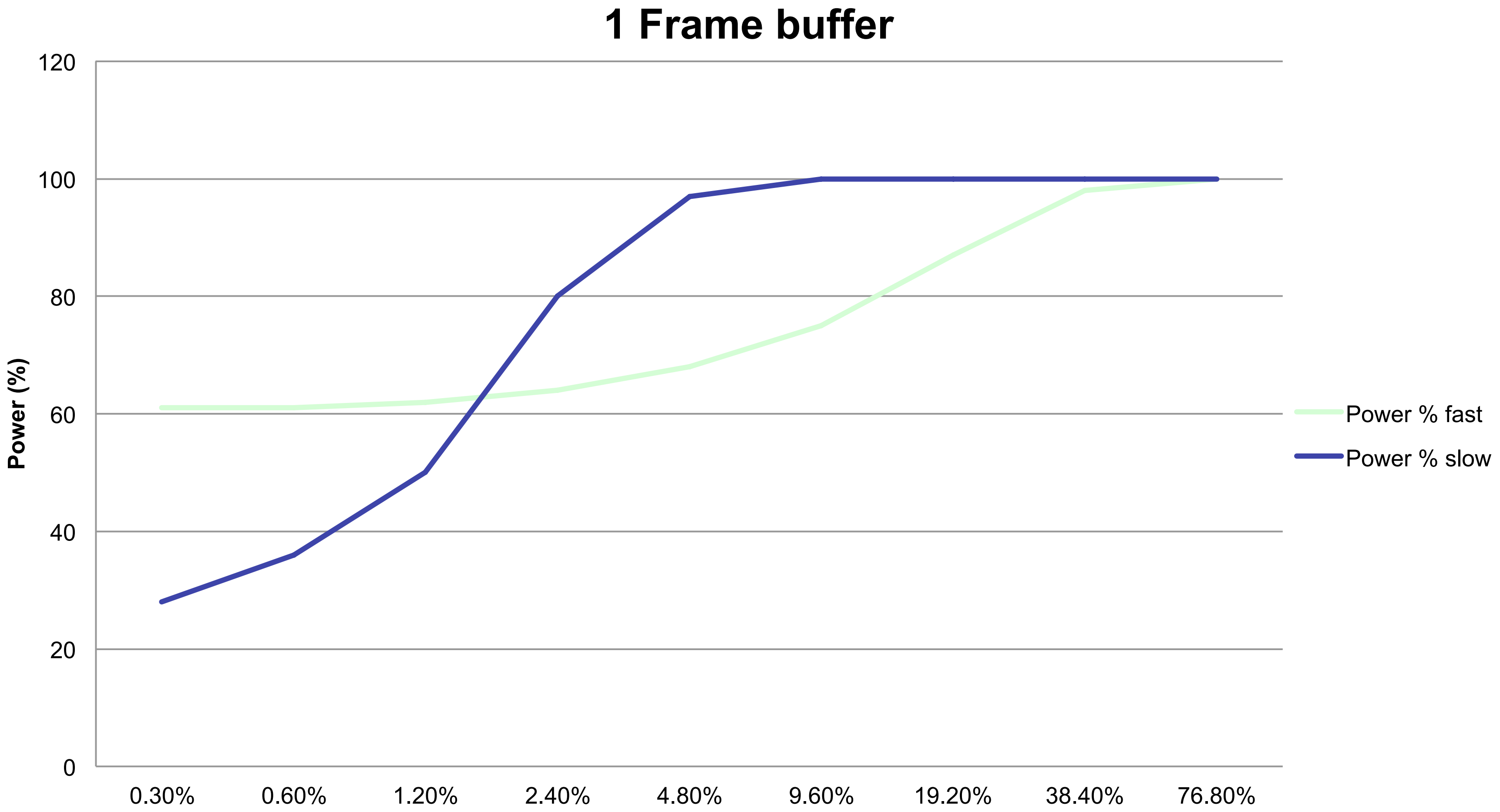
Simulated performance

- 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.

Simulation provisos

- Traffic model scaled up from much slower
  - Results in very pessimistic savings (no long IPGs)
- Heuristic simulation, v. simplistic behavior
- Actual power savings, v. design dependent
  - Leakage losses, fast/slow power switching, etc.
- Other assumptions can be explored
- Effect of buffer & burst
  - Modeled simply as longer packets
  - May be useful for core devices

Power savings

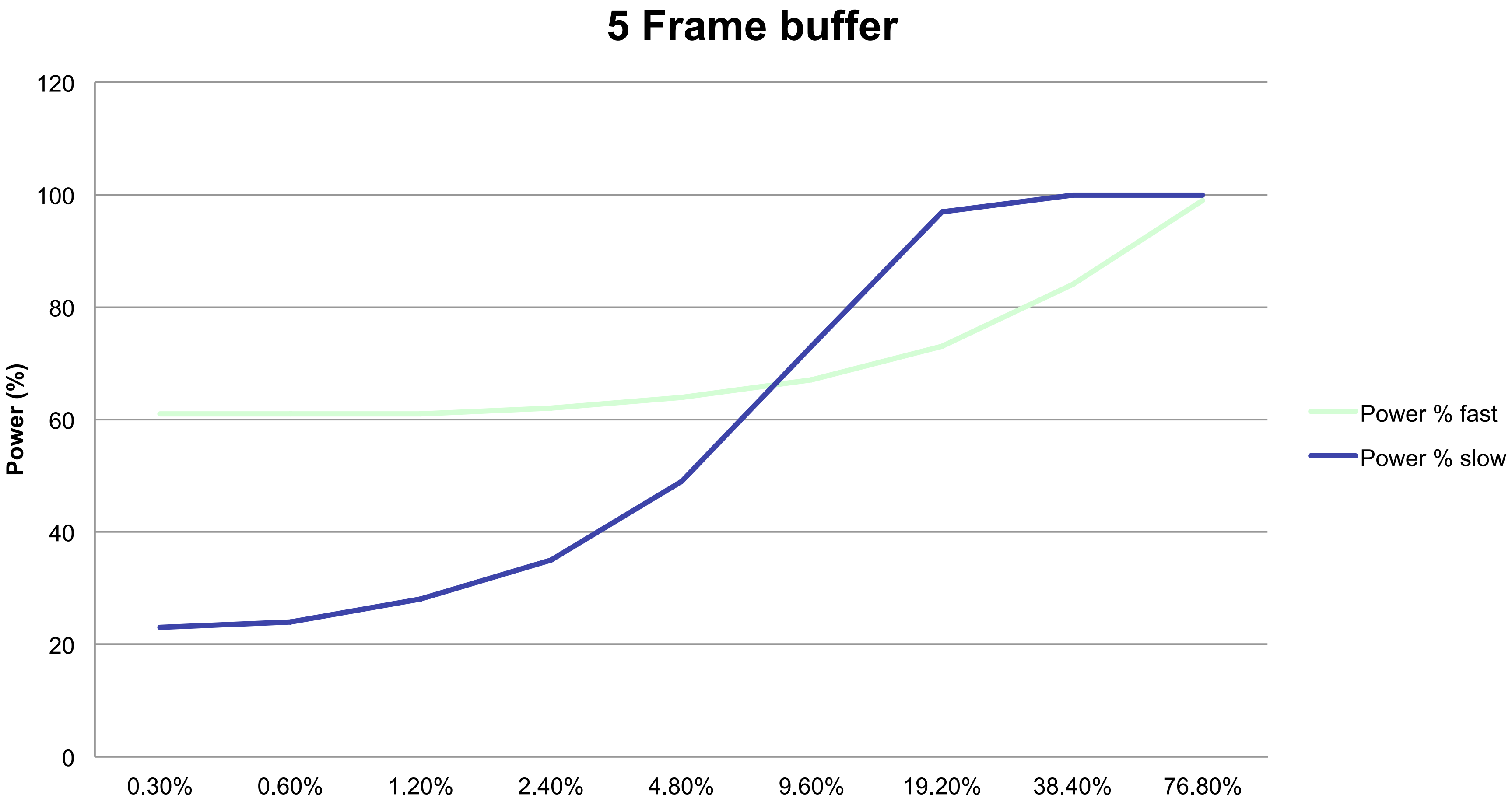


Link utilization (100% = line rate)

Notes

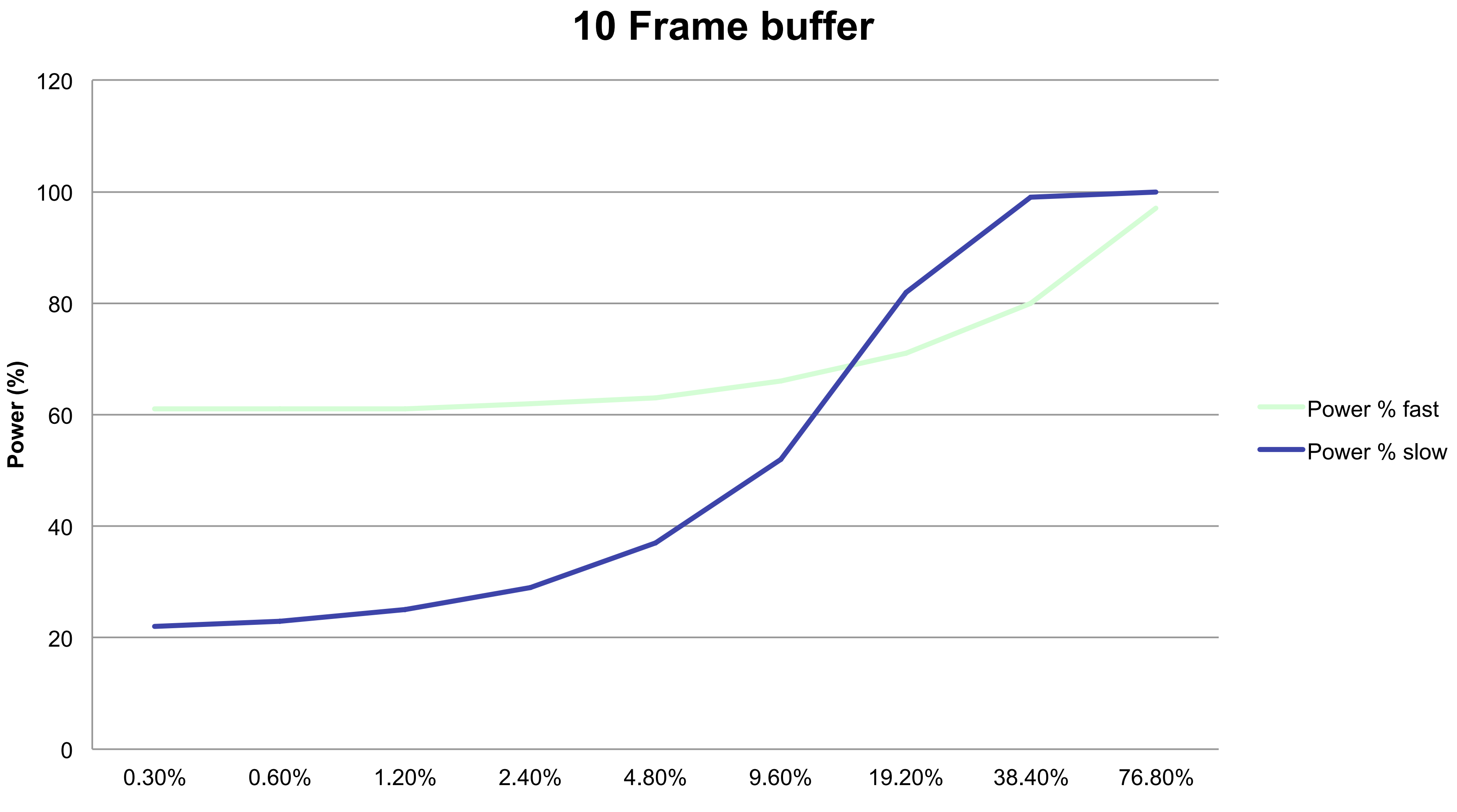
- Fast mode – saves power (20-30%) from 2-20%
  - Key range for aggregation devices
- Slow mode – saves power (up to 80%) less than 2%
  - Ideal for edge devices
  - (and off peak mode – nights & weekends)
- Buffer and burst may help for medium loads
  - Particularly for core devices

Buffer and burst performance



Link utilization (100% = line rate)

Buffer and burst performance



Link utilization (100% = line rate)

Conclusions…

- EEE Fast Wake provides useful solution for optical interface
- Very small additional work required for incorporation
  - Already defined for 100G (& maybe 40G) copper
- No significant impact to optical operation
  - Some study of effect of RAMs on clock quality