Connecting Switch to Fiber: The Energy Efficiency Challenge

Davide Tonietto
Huawei Technologies Canada Ltd., Ottawa, Canada
davide.tonietto@huawei.com

***Abstract*—** Pressure for energy efficiency in distributed computing systems has put in sharp focus ASIC to fiber efficiency as an area needing improvement. What are the origin of the problem and the possible solutions?

1. Introduction

SerDes are a foundation building block of the ICT industry. Several coinciding trends are placing energy efficiency, in particular for the front panel, or switch to fiber, at the center of attention.

1. Complexity and efficiency trends of data rate scale-up as usual are not sustainable
2. Computing-AI has replaced Ethernet as driving force & requirements
3. Disaggregation is replacing integration as system scaling strategy
4. Electrical interconnect capacity has reach its limits in BW, cost, physical density
5. Optical technologies are on the cusp of replacing copper at very short distances of few meters or below

2. The Canary in the Coalmine

In this paper we will focus on high performance fabric switch as an application example, since it is a large ASIC which typically integrates the highest number of SerDes and uses the highest ratio of SerDes vs. core logic resources among all ASICs and thus is the perfect “canary in the coalmine”. In fact, between 2013 and 2023 the share of resources used by SerDes skyrocketed to roughly 30-40% for a fabric switch.

3. Definition of Front Panel Efficiency

If we focus our attention on a relatively simple example like a switch and assume at least half of its I/O will connected to fiber via front panel, it is easy to understand why energy efficiency of the front panel is important. The fundamental pre-requisite of any energy efficiency definition is to make sure that all estimates and comparisons between different solutions include all the blocks and components from the core of the ASIC to the fiber. Also, energy spent for common resources should be accounted for, for example shared laser sources, laser cooling etc. when efficiency is calculated. Where the energy is used, i.e. module vs. ASIC, is very important because it will affect the energy overhead involved in cooling the components and providing regulated power to them at the required voltage and associated energy losses.

4. The Energy Efficiency Challenge: DSP vs. non-DSP SerDes Trends

Intuitively, efficiency gets progressively worse as the channel gets more challenging. It is interesting to observe the progression of this worsening and draw some empirical observations. For this purpose we plot, in Fig. 1, SerDes energy efficiency in pJ/bit as a function of equivalent channel complexity, for simplicity using electrical channel insertion loss (IL) at Nyquist frequency in dB for 112Gbps PAM4. As CMOS technology scales from 7nm to 3nm, the shape of the curve remains largely unchanged but efficiency improves, with larger gains at high IL where a larger DSP benefits more from CMOS scaling. Energy efficiency translates into ASIC resources as shown in Fig.2 example of a 5nm, 50Tbps, 112Gbps/lane Switch Fabric, with a total expected power in the range of 600 to 800W.

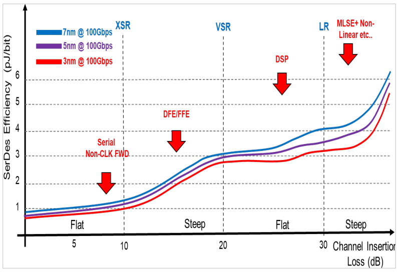

Fig1: Energy Efficiency vs. Channel IL

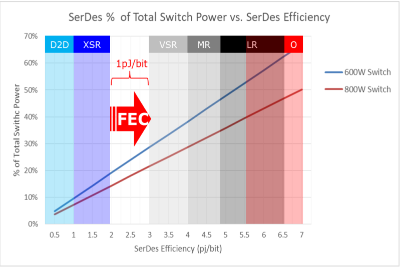

Fig. 2. Switch resources vs. SerDes Type

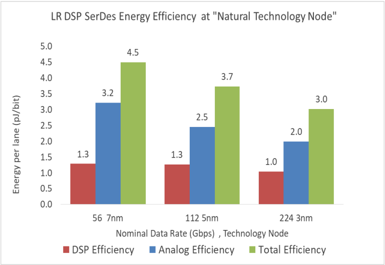

Fig 3. LR DSP SerDes Efficiency Trends

At 56Gbps and even more at 112Gbps most SerDes developed to be used above 20-25dB channels are based on DSP [1]][2][3][4] and we expect this trend to continue beyond 112Gbps. DSP intrinsically scales well beyond 7nm, but the AFE only receives marginal benefit from scaling. We estimate that energy efficiency will improve node over node by ~15-20% (Fig.3). Even with these improvements, power and area increase dramatically at 224Gbps. Based on the trends above and on our expectations for a 100T 3nm Switch a DSP LR SerDes would consume more than 40% of the total power. While DSP LR SerDes are becoming unaffordable, XSR [5][6] SerDes, which are not DSP based but use a more traditional analog architecture, could provide an effective alternative to LR, with ¼ of the power and less than ½ of the area, inclusive of all overheads [7][8]. Also, XSR could increase the available space for core logic which could in turn help reduce the number of dies required for a given ASIC core.

5. Multi Stage vs. Single Stage Interconnect

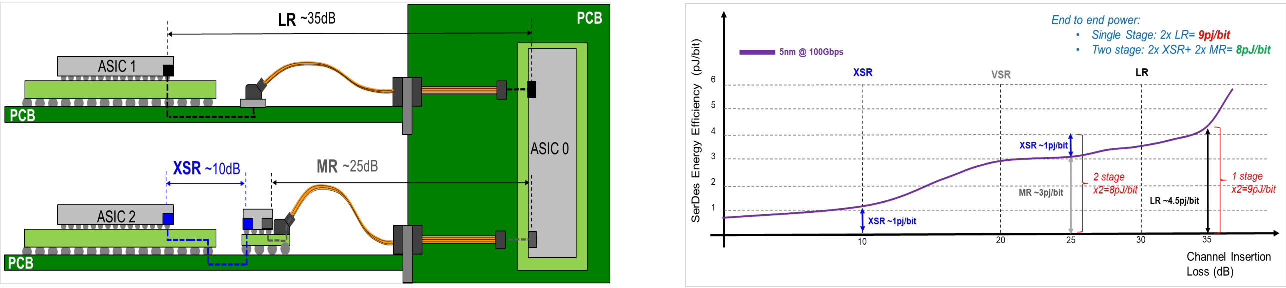

Fig.4. Single vs. 2 stage interconnect

XSR cannot cover the same channel an LR SerDes could, so a possible approach is to split it into 2 or more shorter links. While this may sound counterintuitive, when a channel becomes extremely challenging, this solution becomes more energy efficient as outlined in Fig. 4. This applies to the more complex and even less intuitive case of ASIC to fiber links in the front panel. In particular the recently introduced concept of "Linear" pluggable modules where the retimed IC inside the module is replaced by a non-retimed one, thus replacing two links with a more challenging and less efficient one. In an apple to apple comparison it has been shown [9] that a front panel ASIC to Fiber based on Linear approach is less efficient than an optimized retimed one. Moreover there are other very significant system level benefits in splitting very complex links [10].

6. Fast vs. Slow

A common misconception is that increasing data rates will intrinsically improve energy efficiency. This is due to an erroneous interpretation of the high speed interconnect trends in the last 25 years. While it is true efficiency improved as data rate increased, the reason is not intrinsic but due to other factors, such as CMOS scaling, vastly improved passive channels, better architectures and design methodologies and recently the shift from NRZ to PAM4, which improved SerDes efficiency. However it increased error floor and required FEC, which in turn decreased efficiency and increased latency by roughly an order of magnitude. In reality faster is not more efficient rather, for same node, passive channel technology and architecture, the opposite is true as empirical data shows for 200Gbps vs. 100Gbps in Fig. 5 below. At 200Gbps, even for equivalent electrical channel complexity (impossible to achieve at same cost!), DSP needs to work harder and use advanced equalization at lower insertion loss.

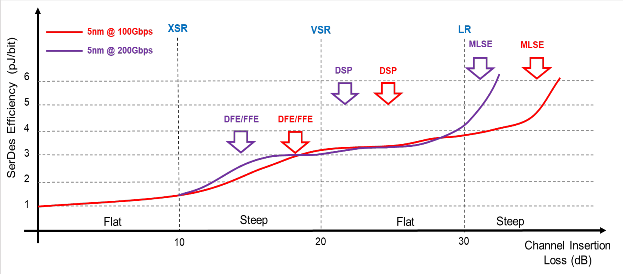

Fig. 5. 100Gbps vs. 200Gbps efficiency

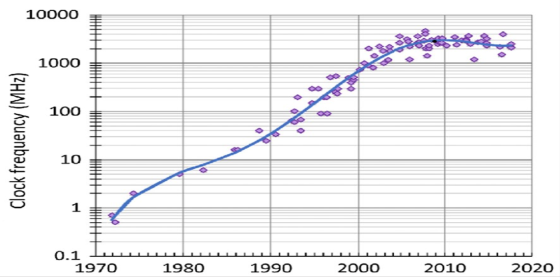

Fig 6. Microprocessor clock speed over the years

A useful analogy is microprocessor clock speed growth which flattened out around 3-4Ghz over 15 years ago and even decreased since to increase efficiency (Fig.5). Technologies and link architectures that likewise increase density and reduce speed of interconnect will provide significant energy efficiency advantages in the future [11][12][13] [14].

7. Optical Densification

Optical densification is the trend that leads to a progressive increase in energy efficiency of optical engines, enabled by and combined with, an increase in in physical density. In the table below we summarize the possible steps from state of the art pluggable modules to NPO (near Package Optics) to 2D socketed and 2.5D soldered CPO (Co-Packaged Optics) and finally 3D CPO, indicating possible energy efficiency targets (7nm node).

TABLE 1. Optical Engines A2F energy efficiency

Type	*A2F (pj/bit)*	*Location*	*Host Interface*	*OE*
Pluggable	17	Front Panel	VSR, PCB	EIC+ Discrete TOSA/ROSA
Pluggable	15	Front Panel	VSR, Flyover	EIC w/ Integrated TIA/DRV Discrete EML
NPO	12	PCB	XSR+, Flyover	EIC w Integrated TIA/DRV +PIC, ELS
2D CPO	11	ASIC PKG[a]	XSR, LD sub.	
2.5D CPO	10	ASIC PKG[b]	D2D, HD sub.	
3D CPO	<8	ASIC PKG[c]	D2D, 3D	

[a] Socketed assembly, [b] soldered assembly, [c] 3D stacking

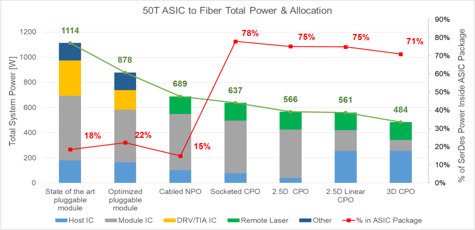

Fig. 7. 50T ASIC to Fiber Total Power and Allocation

Significant energy efficiency gains could be achieved by optical densification, however, while system power could be significantly reduced, ASIC power is not (Fig. 7) With CPO, the portion of the current and power allocated to components inside the ASIC package is more than 70% compared to less than 20% with modules and NPO. This can lead to significant hurdles in designing and cooling the ASIC and possibly increase overall costs.

8. Power scaling

The most important tool we have to improve overall efficiency of interconnects is power scaling. Most ASIC today interface the front panel with the same LR SerDes used for backplanes. The host interface on the module side is often a scaled down version of the same LR SerDes (to save IP development costs…) and the DSP driving the optics is a similar SerDes usually with special features geared toward optical impairments. All of the SerDes in this chain have little or none flexibility in adjusting performance and power to the real requirements of the channel and often these adjustments are manual and "open loop" hence require significant expertise to be used. The result is often overkill where a lot of energy is wasted by overpowered LR DSP SerDes operating on channels both electrical and optical that would require much less resources. While of course utilizing the "perfect" SerDes with design specification exactly tailored for the target channel would lead to the best efficiency, dynamic power scaling can come really close [1], [2].

9. Conclusions

Significant changes in front panel ASIC to fiber approach are required to meet energy efficiency challenges. Focus on density rather than speed, simplicity rather than complexity and finally power scaling will benefit efficiency and latency which are fundamental for dominant computing and AI applications.

10. References

[1] M. LaCroix et al., "A 60Gb/s PAM-4 ADC-DSP Transceiver in 7nm CMOS with SNR-Based Adaptive Power Scaling Achieving 6.9pJ/b at32dB Loss," *ISSCC*, pp. 114-115, Feb. 2019.

[2] M. LaCroix et al., "A 116Gb/s DSP-Based Wireline Transceiver in 7nm CMOS Achieving 6pJ/b at 45dB Loss in PAM-4/Duo-PAM-4 and 52dB in PAM-2," *ISSCC*, pp.132-133, Feb. 2021.

[3] T. Ali et al., "A 460mW 112Gb/s DSP-Based Transceiver with 38dB LossCompensation for Next-Generation Data Centers in 7nm FinFET Technology," *ISSCC*, pp.118-119, Feb. 2020.

[4] Z. Guo et al., "A 112.5Gb/s ADC-DSP-Based PAM-4 Long-Reach Transceiver with >50dB Channel Loss in 5nm FinFET," *ISSCC*, pp. 116-117, Feb. 2022.

[5] "IEEE Standard for Ethernet," IEEE Std 802.3-2018 (Revision of IEEE Std 802.3-2015), pp. 1-5600, 31 Aug. 2018.

[6] N. Tracy et al., "112 Gbps Electrical Interfaces – An OIF Update on CEI-112G," Panel session, *OFC*, San Jose, 2020.

[7] G. Gangasani et al., "A 1.6Tb/s Chiplet over XSR-MCM Channels using 113Gb/s PAM-4 Transceiver with Dynamic Receiver-Driven Adaptation of TX-FFE and Programmable Roaming Taps in 5nm CMOS," *ISSCC*, pp. 122-123, 2022.

[8] Ramy Yousry et al., " A 1.7-pJ/b 112Gbps XSR Transceiver for Intra-package Communication in 7nm FinFETtechnology," *ISSCC*, pp. 180-181, 2021.

[9] Davide Tonietto, "Framework for linear pluggable energy efficiency estimates", oif2023.241.02, June 21st 2023

[10] Davide Tonietto, " The The Future of Short Reach Interconnect", ESSCIRC, pp 1-8, Sept. 2022

[11] R. Farjadrad, M. Kuemerle and B. Vinnakota, "A Bunch-of-Wires (BoW) Interface for Interchiplet Communication," *IEEE Micro*, vol. 40, no. 1, pp. 15-24, 1 Jan.-Feb. 2020.

[12] A. Chandrasekhar et al., "Server CPU Package Design Using PoINT Architecture," *ECTC*, pp. 2180-2185, 2019.

[13] P. K. Huang et al., "Wafer Level System Integration of the Fifth Generation CoWoS®-S with High Performance Si Interposer at 2500 mm2," *ECTC*, pp. 101-104, 2021.

[14] https://www.samtec.com/solutions/flyover