A Simple Analytical Model for Energy Efficient Ethernet

Marco Ajmone Marsan, Antonio Fernández Anta, Vincenzo Mancuso, Balaji Rengarajan, Pedro Reviriego Vasallo, and Gianluca Rizzo

***Abstract*—The recently approved Energy Efficient Ethernet standard IEEE 802.3az achieves energy savings by using a low power mode when the link is idle. However, those savings heavily depend on the traffic patterns, due to the overhead inherent in transitions between active and low power modes. This makes it impractical to estimate energy savings through measurements or simulations in all relevant scenarios. In this letter we present an analytical model to estimate the energy consumption of an Energy Efficient Ethernet link, based on simple traffic parameters. The model is validated through simulation and experimental data.**

***Index Terms*—Energy management, modeling.**

I. Introduction

THE recently approved IEEE 802.3az Energy Efficient Ethernet (EEE) standard [1] is expected to provide significant energy savings in local area networks over the coming years [2]. EEE saves energy by operating the physical layer of a link in low power mode when it is not carrying traffic. The EEE standard specifies different low power modes for widely used physical layers of Ethernet over Unshielded Twisted Pairs (UTP), namely 100BASE-TX (100 $Mbps$), 1000BASE-T (1 $Gbps$) and 10GBASE-T (10 $Gbps$).

Transitions to and from the low power mode are not instantaneous. The minimum transition times specified in the standard are different for each speed, but always significantly larger than the transmission time of the typical maximum-size frame [1]. Therefore, the energy overhead due to transitions can be relevant even when the traffic load is low [3]. This means that, in addition to traffic load, the traffic pattern characteristics, e.g., frame lengths and interarrival times, are key to determine the energy savings that will be obtained with EEE. For example, in the case of a 1000BASE-T link, a load as low as 3 $Mbps$ can prevent the link from ever switching to low power mode if frames are small and evenly spaced, while for bursty arrivals of large frames the link would be in low power mode most of the time.

In this letter we propose the first analytical model to estimate the energy consumption of an EEE link, based on simple traffic parameters, namely the average frame size and the first two moments of frame interarrival times. Our model compares favorably with alternative methods, like actual energy measurements or packet level simulations [3], which are impractical or computationally demanding [4].

The rest of this letter is organized as follows. Section II describes the analytical model that we propose for the estimation of the behavior of EEE links. Section III shows that, notwithstanding the simplifications used in the model, it can be used to rather accurately predict the behavior of links with real traffic. Section IV concludes the letter.

Manuscript received May 10, 2011. The associate editor coordinating the review of this letter and approving it for publication was J. Murphy.

M. Ajmone Marsan, A. Fernández Anta, V. Mancuso, B. Rengarajan, and G. Rizzo are with the Institute IMDEA Networks, Madrid, Spain (e-mail: vincenzo.mancuso@imdea.org). M. Ajmone Marsan is also with Politecnico di Torino, Italy.

Pedro Reviriego Vasallo is with the University Antonio de Nebrija, Madrid, Spain.

Digital Object Identifier 10.1109/LCOMM.2011.060111.110973

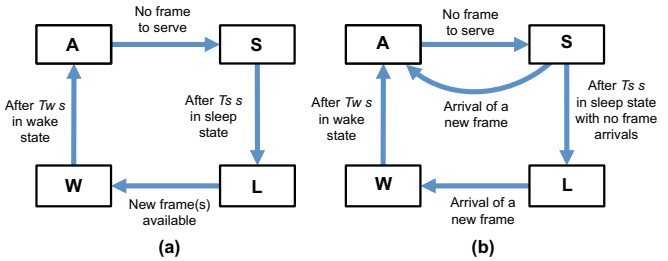

Fig. 1. State transition diagrams for (a) an EEE 10GBASE-T link, and (b) for an EEE 100BASE-TX or 1000BASE-T link.

II. Model

Behavior of EEE links: EEE links can be in one of the following four states: Active (A), Sleep (S), Wake (W) and Low Power (L). The state transition diagrams are illustrated in Fig. 1. Frame transmissions only occur in state A. When the link completes transmitting all the buffered frames, it enters state S as a transition to state L. If no frame arrives for a time period of T_s seconds while in state S, the link enters state L (during which power consumption is minimized). A frame arrival in state L results in the link transitioning to state W which lasts T_w seconds. After this wake interval, the link transitions to state A and begins transmitting.

The behavior of EEE 10GBASE-T links is different from EEE 100BASE-TX and 1000BASE-T links in the case of a frame arrival in state S. In EEE 100BASE-TX and 1000BASE-T links, a frame arrival in the sleep interval causes an immediate transition to state A (see Fig. 1b), without incurring in delays of up to approximately 200 μs due to large standard values for T_s. In contrast, immediate transition to state A is not supported in EEE 10GBASE-T links, since the standard value for T_s is less than 3 μs. Therefore, in 10 $Gbps$ links, arrivals have to be buffered while the sleep operation is completed in T_s seconds, and the link then goes through state W for T_w seconds before frames can be served (see Fig. 1a). Note that EEE mechanisms for 100 $Mbps$ and 10 $Gbps$ links are defined for unidirectional links, while 1 $Gbps$ EEE links can transition to low power mode only when there is no traffic in both link directions. Our analysis considers unidirectional links, and thus reflects the behavior of 100BASE-TX and 10GBASE-T EEE links. Moreover, as shown in Section III, it also provides a reasonable estimate of the energy savings in 1000BASE-T EEE links with strongly asymmetric traffic load, by just looking at the direction with the highest load.

Queue model: We model unidirectional links, and evaluate the average time spent in each of the four possible link states by means of an $M/G/1$ queue model with infinite waiting

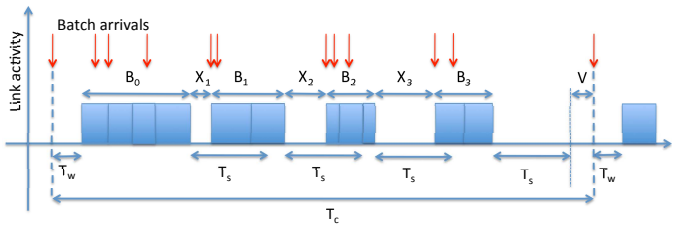

Fig. 2. System cycle for a 1000Base-T EEE link.

room and with server timeout and activation times. The packet service rate is non-zero only in the active state, where it equals a constant R, corresponding to the link speed.

We denote by S_p the size of a single frame and by $E[S_p]$ the average frame size. Frames arrive in batches of random size $N_b \geq 1$, according to a Poisson process with rate λ [5]. Each *batch* arrival, i.e., a burst of received frames, represents a queue customer in our model. The motivation to use batch arrivals lies in the bursty nature of packet arrivals in real networks: packets can arrive at a network interface so close in time that they are transmitted back-to-back. Batch arrivals allow this behavior to be captured in the model. However, by setting $N_b = 1$, Poisson *frame* arrivals can be modeled as well.

Cycle analysis: The sample path of the queue can be viewed as a sequence of cycles as illustrated in Fig. 2. A cycle starts with the batch arrival that induces the transition from state L to state W. This is followed by a busy period with the link in state A, whose duration is denoted by B_0. The initial busy period is followed by a random number N of sleep/active interval pairs $(X_i,\ B_i)$, $i \geq 1$, with each pair corresponding to a batch arrival in state S before a time T_s has elapsed. Note that the sleep time is $X_i = T_s + T_w$ in the case of 10GBASE-T links while it is the random time interval between the start of state S and the next frame arrival in 100BASE-TX or 1000BASE-T EEE links. Finally, a sleep interval of duration T_s precedes the idle period V, whose random duration corresponds to the time interval before the beginning of a new cycle. We denote the length of a cycle by T_c and its average by $E[T_c]$.

Using results from renewal theory [6], we can focus on the system cycle, and compute the fraction of time spent in each link state as the ratio between the average time in each state in a cycle and the average cycle duration. We denote the fraction of time in the cycle spent in state α as f_α, for $\alpha \in \{A, S, L, W\}$.

We derive system cycle parameters for two cases: (i) arrivals in S are served immediately (e.g., 100BASE-TX and 1000BASE-T EEE links), and (ii) arrivals in S are served with delay (e.g., 10GBASE-T EEE links). Results for the two cases are denoted as $()'$ and $()''$, respectively.

Theorem 1. *For unidirectional links in which arrivals in S are served immediately, the average cycle duration and the fraction of time spent in states A, L, S, and W are given by* $E[T_c'] = \frac{\lambda T_w + e^{\lambda T_s}}{\lambda(1-\rho)}$, $f_A' = \rho$, $f_L' = \frac{1}{\lambda E[T_c']}$, $f_S' = \frac{e^{\lambda T_s} - 1}{\lambda E[T_c']}$, *and* $f_W' = \frac{T_w}{E[T_c']}$, *with* $\rho = \frac{\lambda}{R} E[S_p] E[N_b]$.

Proof: Consider the different intervals included in T_c. B_0 is the busy period of an $M/G/1$ queue with batch arrivals, and its average depends on the batch arrival rate λ, the mean batch service time $E[\tau]$, and the queue size Z_0 at the beginning of the busy period [6]:

TABLE I
COMPARISON OF SIMULATION WITH REAL UNIDIRECTIONAL TRACES (FROM CAIDA), AND MODEL FOR 10GBASE-T EEE LINKS ($T_w = 4.48\mu s$, $T_s = 2.88\mu s$) WITH BATCH POISSON ARRIVALS

ρ $(= f_A'')$	$E[S_p]$ [Bytes]	m_Y $[\mu s]$ trace	σ_Y $[\mu s]$ trace	f_L'' [%] trace	f_L'' [%] model
0.032	563.4	14.13	16.13	60.99 $\pm$ 0.30	62.88
0.075	768.1	8.17	9.27	43.21 $\pm$ 0.36	44.63
0.147	423.2	2.30	2.62	9.84 $\pm$ 0.13	9.18
0.150	636.4	3.40	3.78	17.57 $\pm$ 0.78	16.58
0.191	844.6	3.54	3.95	17.68 $\pm$ 0.27	16.79
0.251	587.2	1.87	1.97	5.06 $\pm$ 0.09	4.39
0.469	735.4	1.26	1.38	2.20 $\pm$ 0.03	1.23

$$E[B_0] = \frac{E[Z_0]\, E[\tau]}{1-\rho}; \tag{1}$$

where $E[Z_0] = 1 + \lambda T_w$ (the batch arrival that triggers the wake-up, plus the average number of Poisson batch arrivals during the wake-up interval T_w). The average batch service time accounts for the batch size distribution: $E[\tau] = E[S_p]\, E[N_b]/R$.

Each busy interval B_i, $i \geq 1$, occurs if the residual interarrival time X_i, counted from the end of the previous busy period, does not exceed the sleep time T_s. Since arrivals are Poisson, the probability of having no arrivals in T_s is $P_0 = e^{-\lambda T_s}$. Thereby, the number $N \geq 0$ of busy periods in a cycle, not counting B_0, can be seen as the number of consecutive successes of a geometric $r.v.$ with success probability $1 - P_0$. Hence, its average value is: $E[N] = \frac{1-P_0}{P_0} = e^{\lambda T_s} - 1$.

Given that the sleep time of T_s seconds is not completed, the duration of X_i is that of a truncated exponential, and hence its average is $E[X_i | X_i \leq T_s] = \frac{1}{\lambda} - \frac{T_s}{e^{\lambda T_s} - 1}$, $i \geq 1$.

As for the busy periods B_i, $i \geq 1$, an expression like (1) holds, with $E[Z_i] = 1$, since the service starts immediately in state S. Hence, all B_i, as well as all X_i, $i \geq 1$, are $i.i.d.$ $r.v.$'s. The idle period V has mean $1/\lambda$ due to the lack of memory of the exponential distribution. Thereby the average cycle duration is as follows:

$$E[T_c'] = T_w + E[B_0] + E[N](E[X_1 | X_1 \leq T_s] + E[B_1]) + T_s + \frac{1}{\lambda}.$$

Considering that the link is in state A during busy periods B_i, in state L during V, in state S during X_i plus T_s seconds before V, and in state W during the initial T_w seconds of each cycle, the result follows. ■

Theorem 2. *For unidirectional links in which arrivals in S are served after the completion of the sleep time T_s, the average cycle duration and the fraction of time spent in states A, L, S, and W are given by* $E[T_c''] = \frac{1 + \lambda(T_s + T_w)\, e^{\lambda T_s}}{\lambda(1-\rho)}$, $f_A'' = \rho$, $f_L'' = \frac{1}{\lambda E[T_c'']}$, $f_S'' = \frac{e^{\lambda T_s}}{E[T_c'']} T_s$, *and* $f_W'' = \frac{e^{\lambda T_s}}{E[T_c'']} T_w$.

Proof: The proof resembles that of Theorem 1. The cycle structure is exactly as in Theorem 1. However, since now $X_i = T_s + T_w$ is a constant, the average number of batches queued at the beginning of B_i, $i \geq 1$, is $1 + \lambda(T_s + T_w - E[X | X \leq T_s])$, with X the exponential time of arrival of the first batch in S. Thus, the expression for $E[T_c'']$, which leads to the result, is:

$$E[T_c''] = T_w + E[B_0] + E[N](T_S + T_w + E[B_1]) + T_s + \frac{1}{\lambda}.$$

■

III. VALIDATION

We validated our model against synthetic traces, i.e., batch Poisson arrivals generated by means of a simulator, and real

TABLE II
COMPARISON OF SIMULATION WITH REAL BIDIRECTIONAL TRACES (FROM GOOGLE DATA CENTERS), AND MODEL FOR UNIDIRECTIONAL 1000BASE-T EEE LINKS ($T_w = 16\mu s$, $T_s = 182\mu s$) WITH BATCH POISSON ARRIVALS

ρ in/out	$E[S_p]$ [Bytes]	m_Y $[\mu s]$ trace with higher ρ	σ_Y $[\mu s]$ trace with higher ρ	f'_A [%] traces	f'_A [%] model	f'_L [%] traces	f'_L [%] model
0.015 / 0.528	1497.3 (out)	22.68 (out)	185.20 (out)	52.87 $\pm$ 5.55	52.81	34.80 $\pm$ 5.38	36.63
0.087 / 0.074	944.4 (in)	87.01 (in)	307.88 (in)	15.31 $\pm$ 1.77	8.68	52.19 $\pm$ 1.83	65.70
0.008 / 0.041	748.4 (out)	145.17 (out)	233.96 (out)	4.78 $\pm$ 0.50	4.12	37.68 $\pm$ 0.42	46.05

traffic traces of 10 $Gbps$ and 1 $Gbps$ links obtained from the CAIDA archive [7], [8], and from Google data centers, respectively.[1] Synthetic traces were generated with geometrically distributed burst sizes, with success probability p_b, so that $E[N_b] = 1/(1-p_b)$. CAIDA traces were collected on a 10 $Gbps$ backbone optical link, where the high aggregation level makes traffic characteristics not far from Poisson, with an arrival rate that changes over time but that can be considered constant over a time scale of tens of minutes [9]. Google data center traces were collected on 1 $Gbps$ server links, and they are highly bursty.

Results for synthetic traces, not shown here due to lack of space, match the model results with very high accuracy (as expected, since our model is exact under such assumptions).

A good model accuracy is obtained also when considering real traces. We use those real traces as input for an EEE link simulator that computes the transmission time of each packet in the input trace according to EEE specifications. Tables I and II illustrate the results obtained with traces at different rates and loads. We first use the real trace to compute the offered load ρ, the average packet size $E[S_p]$, and the first two moments of the packet interarrival time Y, namely the average m_Y and the standard deviation σ_Y, over the entire input trace. Then we simulate the performance of the EEE link, and compute the fraction of time spent in states A, L, S, and W. Results are shown in the tables with 99% confidence intervals. Later, we evaluate the model based on the values of $E[S_p]$, m_Y, and σ_Y. To this purpose, we consider that, under the batch Poisson arrival approximation, and with geometrically distributed burst sizes with parameter p_b, Y is 0 with probability p_b, and is exponentially distributed with rate λ with probability $1-p_b$. Thus, $m_Y = (1-p_b)/\lambda$, and $\sigma_Y = \sqrt{1-p_b^2}/\lambda$, which allows us to compute λ and p_b.

Table I shows that the model yields accurate estimates of the fractions of time spent in state L for 10 $Gbps$ links with a large spectrum of loads and average packet sizes. Noticeably, the time spent in low power mode is deeply affected by the average packet size, e.g., with a load equal to 0.15 in a 10 $Gbps$ link, f''_L can be as high as 17.68% when the packet size is 636.4 bytes, while a very similar load value $\rho = 0.147$ leads to $f''_L = 9.84\%$ when the packet size decreases to 423.3 bytes.

Table II displays the results obtained by simulating bidirectional 1000BASE-T links based on real traces. We compare the simulation against an approximation, where we only model the traffic in the direction with higher load. The table reports the load measured for each link direction (in and out), and the statistics that are used in the model. Note that the simulated link can switch to low power mode only if the link is idle in *both* directions, hence, f'_A, as estimated from the traces, does not correspond to the load. Model's results are quite accurate in case of highly asymmetric links, as shown in the first row of Table II; instead, when the loads in the two directions are comparable, the model overestimates the time spent in state L, nonetheless providing a reasonable estimate.

As expected, the amount of energy saving that can be achieved heavily depends on the traffic burstiness. In fact, considering as index of burstiness the ratio σ_Y/m_Y, we found that f'_L and f''_L can be very high either when the load is very low (see traces with $\rho < 0.1$ in the tables) or when the standard deviation of interarrivals is significantly higher than the average (see Table II). E.g., assume that, as suggested in [3], the energy consumption in state L is 10% with respect to state A, while in states S and W the consumption is practically the same as in state A; since non-EEE links are always in state A, the energy saving due to EEE for the cases of Table I, first row, and Table II, first row, is $\sim$55% and $\sim$31%, respectively.

These results, and others for a large set of CAIDA traces not shown here due to lack of space, show that the model can be conveniently adopted to predict the energy saving that can be achieved with EEE in a large range of scenarios.

IV. CONCLUSIONS AND FUTURE WORK

In this letter, we have presented and validated a model that accurately predicts the time that a unidirectional EEE link spends in each of its four possible states. The model can be used with 100BASE-TX, 1000BASE-T and 10GBASE-T EEE links, and its applications are twofold: (i) evaluate the energy saving achievable by replacing existing links with EEE links; (ii) drive the design of traffic shaping mechanisms that would increase the achievable energy saving. In fact the model can be used to analyze the consumption of links in which a timer is used to collect frames before starting a link activation.

[1]Thanks to K. Fu, G. Chesson, L.A. Barroso and U. Holzle from Google for providing the traces from their data centers, and to D Larrabeiti (Univ. Carlos III, Madrid) for pre-processing Google and (part of) CAIDA traces.

REFERENCES

[1] IEEE Std 802.3az: Energy Efficient Ethernet-2010.
[2] K. Christensen, P. Reviriego, B. Nordman, M. Bennett, M. Mostowfi, and J. A. Maestro, "IEEE 802.3az: the road to energy efficient Ethernet," *IEEE Commun. Mag.*, vol. 48, no. 11, pp. 50-56, Nov. 2010.
[3] P. Reviriego, J. A. Hernández, D. Larrabeiti, and J. A. Maestro, "Performance evaluation of energy efficient Ethernet," *IEEE Commun. Lett.*, vol. 13, no. 9, pp. 697-699, Sep. 2009.
[4] A. Nucci and K. Papagiannaki, *Design, Measurement and Management of Large-Scale IP Networks: Bridging the Gap Between Theory and Practice*. Cambridge University Press, 2008.
[5] R. Jain and S. Routhier, "Packet trains–measurements and a new model for computer network traffic," *IEEE J. Sel. Areas Commun.*, vol. 4, no. 6, pp. 986- 995, Sep. 1986.
[6] L. Kleinrock, *Queueing Systems: Theory*. J. Wiley and Sons, 1975.
[7] C. Walsworth, E. Aben, K. C. Claffy, and D. Andersen, "The CAIDA anonymized 2009 Internet traces, http://www.caida.org/data/passive/passive_2009_dataset.xml
[8] K. C. Claffy, D. Andersen, and P. Hick, "The CAIDA anonymized 2010 Internet traces, http://www.caida.org/data/passive/passive_2010_dataset.xml
[9] T. Karagiannis, M. Molle, M. Faloutsos, and A. Broido, "A nonstationary Poisson view of Internet traffic," in *Proc. INFOCOM*, pp. 1-12, Mar. 2004.