THE TIMELINESS OF ASYNCHRONOUS PACKET MULTIPLEXING IN SWITCHED ETHERNET

LI Qiao, ZHANG Xiaolin, XIONG Huagang, FEI Yuxia

Electrical and Information School, Beijing University of Aeronautics & Astronautics, Beijing, 100083, China

ABSTRACT

Powered by single-segment switched interconnection, Ethernet can be used in time-critical data acquisition applications. Unlike synchronous time division multiple access, asynchronous packet streams result in congestions and uncertain multiplexing delays. With the delay analysis in the *worst case* and *probabilistic guaranteeing* conditions, we restrict the packet-sizes, intervals or traffic burstiness *a priori* to regulate delay deviations within acceptable scales. Some methods of combinatorics and stochastic theory, e.g. *Cumulant Generating Function* and the *Large Deviation Principle*, are used and verified by some simulation-based computations. The influence of time varying delay for telemetry applications is also discussed in some sense.

KEYWORDS

Switched Ethernet, Packet Multiplexing, Queuing Delay, Large Deviation Principle, Stochastic Ordering

INTRODUCTION

Enhanced by several standardized mechanisms such as micro-segmentation, full-duplex operation and prioritization (specified by IEEE 802.1p, 802.3u *etc.*) [1], switched Ethernet can be used in some time-critical applications. As a mutual and one of the most widely supported COTS technologies, it had been adopted as industrial field bus throughout device, system domain and enterprise levels and also had been applied in platform electronics integrated systems, *e.g.* AFDX (Avionics Full Duplex) bus in civil aircrafts and dual fault-tolerance Ethernet within marine vessels *etc.*. Switched Ethernet

can be used as data acquisition network in industrial plants or platforms with acceptable features such as throughout, re-configurability, fault tolerance *etc*..

Due to price per port of Ethernet switched hubs is low, there is an increasing trend towards building single-segment networks. Accordingly, each host links in full duplex with one switched port exclusively. In data acquisition applications, data acquisition cards plugged in the host connect smart sensors, A/D converters or programmable controllers by stub cables.

Since the single-segment connection entirely avoids the non-determinant latency caused by CSMA/CD in media sharing Ethernet, CTOS Ethernets had been used in some real-time data acquisition applications such as a wind tunnel project [2]. However, unlike those specified systems, in integrated systems, different types of communicating tasks including command, supervisory control, monitoring and data accessing have various performance requirements, especially in real-time issues. It's necessary to propose an engineering solution strategy and some application-concerned tactics to overcome complexity and finally accomplish design objectives.

Synchronized by distributed soft-timers or global real timers, packets from distributed hosts can be multiplexed in time division without collisions. Supported by certain middleware, distributed tasks take dependable real-time communications in event-based synchronous manner. Well-designed synchronous accessing methods could tolerance temporal uncertainties, but incur overhead and complexity. In the other hand, asynchronous packet passing is in lightweight but implies traffic congestions that induce variable multiplexing queuing delays. To regulate delay deviations within acceptable scales, we restrict the packet-sizes, intervals or traffic burstiness by pre-planning.

Deterministic queuing theory developed primarily by [3] is referred to calculus or verify the delay bounds. As the most common implement of Ethernet switches, we consider the output queuing within an output port of the switch in FCFS (first come first served) and *work conserving* disciplines. Many input streams feed into the multiplexer (MUX) and are aggregated to the output port. Traffic burstiness causes buffering and a current input packet is delayed by the buffered data in $q(t)$ length at this instant t. Note that the interior caching inside switches is very fast and input and output rates are limited by the constant line rate c (c=100Mbps in fast Ethernet) , so that in a rough meaning, the queuing delay is deduced to $q(t)/c$. Considering the event that the $q(t)$ overrides a given threshold x, which corresponds with the queuing delay exceeding a certain deadline, we call the event "*hazard*" as a concise nomenclature with occurrence probability Q, and then term its counterpart as the "*guaranteeing probability*" (*GP*), certainly GP=1-Q.

For high-performance applications, the hazard is rare. Therefore instead of traditional queuing theory, we employ the *Large Deviation Principle* (*LDP*) to approximate the residual probability in the tail of the random variable's (*r.v.*) distribution far from its mean. *LDP* theory was founded in the mid 1960s[4] while its earlier results had been well known as the Chenoff Bound and the Cramer Theory. In around 1990, JosephY Hui, John T Lewis, Chang C-S and other researchers introduced it into

delay analysis in communication networks [5][6].

Computing the *Cumulant Generating Function*s (*CGF*) for the sums-in-byte of multiplexing arrivals (as *r.v.*s) in multiple time scales is the key step to approximate the hazard probability in term of *LDP*. We can directly deduce the *CGF* by enumerating all the feasible conditional combinations. To reduce the computation complexity, we construct two stochastic models by *stochastic ordering* theory as the bounds of the *r.v.*s and get the conservative but enough precise bounds of the hazard probability. In this paper, a case study is present to demo the stochastic bounds, and some results by simulated-based computations with the *Importance Sampling* (*IS*) technique are also illustrated.

The *Maximum Rate Function* [7] is also introduced to delay analysis in *worst case* condition and is derived through the limit of *CGF*s.

Additional, we also discuss the influence of time varying delay for telemetry applications in some sense and draw some delay tolerance solutions.

INTEGRATED REAL-TIME SWITCHED ETHERNET

Under integrated real-time LAN environments, different communication tasks have their own different quality requirements in multiple dimensions *i.e.* throughout, importance and timeliness *etc.*. The Ethernet protocols don't define sophisticated mechanisms for traffic isolation, service discrimination and flow control, and only support very coarse priority ranks (There are 8 priorities identified by a 3-bits field in IEEE 802.3 frame, and most of COTS industrial Ethernet switches usually have only 2~4 so called QoS ranks.). A practical solution is to afford differentiated service manners for different types of coarse classified tasks isolated by priority layers, in the meantime to take integrated resource allocations among the tasks of the same type. Therefore the schedule schemes for each layer can be designed nearly separately but the *worst case* temporal constraints of the higher layer should be reckoned in the lower as "backgrounds". Generally, the system assigns synchronous tasks with higher priority, while restricts the asynchronous packet passing delays in *probabilistic guaranteeing*.

1) The Synchronous Accessing Strategies

Supported by global real timers or distributed clock synchronization algorithms, packets from some hosts can be multiplexed collision-free in time division. TDMA is well used in telemetry domain but its off-line scheduling strategy is non-flexible.

Command/Response strategy seems to be synchronized by a centralized accessing controller. It is collision-free and also has inherent disadvantage of non-flexibility. To take the advantages of adaptation, flexibility and on-line re-configurability, event-based real-time

communication paradigms [8], which seem to be synchronized by distributed procedures or transactions, are developed above middleware with client/server (c/s), publisher/subscriber (p/s) or peer-to-peer (p2p) architectures.

In the p/s scenario the communicating job is invoked by publisher's pushing or by subscriber's pulling. Co-enlightened by the Just-in-Time and pull production system in manufacture domain, we are developing a kanban (token) based schedule scheme by a lightweight and real-time enhanced middleware in switched Ethernet.

2) The Asynchronous Packet Passing and the Credit Bucket Traffic Shaper

Asynchronous packet passing eliminates the overhead of clock synchronization or event acknowledgement. State messages are usually collected in data acquisitions in this mode. Periodic packet streams are the typical real-time traffic, while for the aperiodic packet streams we restrict their burstiness by the traffic shapers.

The literature [9] reports a Credit Bucket (*CB*) traffic shaper in sharing media Ethernet. In fact, the shaper is also useful for switched Ethernet. We re-implement it and insert it into the NIC (network interface card) driver routine at the host.

The *CB*, which is analogue with the well-known leaky bucket but is suitable for variable length packet streams, has two parameters: Credit Bucket Depth (CBD) and Refresh Period (RP). CBD limits the maximum number of credits that can be stored in the credit bucket. Up to CBD credits are added to the bucket every RP. If the number of credits exceeds CBD, overflow credits are discarded. When a packet arrives from the communicating task, if there is at least one credit in the bucket, the shaper forwards to the NIC and removes as many credits as the size of the packet (in bytes). By changing RP and CBD, one can control the burstiness of a packet stream while keeping the same average throughput. [9]

PROBABILISTIC GUARANTEEING FOR ASYNCHRONOUS PACKET MULTIPLEXING

According to the FCFS multiplexing model, make time discretization for slots, noted as t=0,1,2,… ; the ith arrival process is denoted as $\{a_i(t)\}$, i=1,2,…,n ; the capacity of MUX output link is nc, c is a constant. The length of buffering content is $q(t)$, $q(t)$ complies with the Lindlay recursion with initial condition $q(0)$=0:

$$q(t) \quad \max_{0 \ s \ t}\left[\sum_{i \ 1}^{n} (A_i(t) \quad A_i(s)) \quad nc(t \quad s)\right] \ , \quad q(0)=0 \qquad (1)$$

where $A_i(t) \quad \int_{s\ 0}^{t} a_i(s)$ is the cumulative arrival of the *i*th input. And the arrival processes satisfy *Independence*, *Stationarity*, *Traffic Characterization* assumptions [6] and as well as the *average rate condition* that $\sum_i r_i \quad n \ c$. Then for t> t , (t , refer to [6]), the distribution of $q(t)$ is same as q(t^{+}). When the hazard probability is rare, it can be approximated by (2) in term of *LDP*:

$$P\{q(t) \quad nx\} \quad \exp(\quad \int_{s\ 1}^{t} {}^{*}_{s}(x)) \qquad (2)$$

where ${}^{*}_{s}(x)$ is named the *rate function*, s is the *time scale*. ${}^{*}_{s}(x)$ is derived by the Legendre Transformation of the *CGF*. That is:

$${}^{*}_{s}(x) \quad \sup_{q\ 0}[n(cs \quad x)q \quad {}_{A,s}(q)] \qquad (3)$$

where the *CGF* of service process is $ncsq$; and the *CGF* of arrival process is ${}_{A,s}(q)$, which is denoted as ${}_{s}(q)$ or (s,q) hereafter as long as there is no confusion.

According to the *Contraction Principl*e, the rare probability is mainly decided by the minimum ${}^{*}_{s}$ in the exponential index (see (2)). s^{*}=arg$\max_{s}$(${}^{*}_{s}$) is called the *critical time scale*; and in practice:

$$P\{q(t) \quad nx\} \quad \exp(\quad {}^{*}(s^{*})) \qquad (4)$$

Remark: Denoting the output link as nc and the threshold as nx is only to coincide with the formats in our referred literatures; actually MUX's output aggregates all the inputs without division of n pieces.

For multiple class inputs, denote ${}_{i}(q)$ and n_i as the *CGF* and the amount of streams respectively of *i*th class. If the streams are independent, the normalized *CGF* is ${}_{x}(q) \quad \frac{n_1}{n} \ {}_{1}(q) \quad \ldots \quad \frac{n_m}{n} \ {}_{m}(q)$. Consider (3) and the *Contraction Principle*, rewrite (4) as:

$$1\text{-}GP = P(q(t) \quad nx) \quad \exp(\quad n \min_{1\ s\ t} {}^{*}(cs \quad x)) \qquad (5)$$

in which:

$${}^{*}(cs \quad x) \quad \sup_{q\ 0}((cs \quad x)q \quad (q(\mathrm{s})) \qquad (6).$$

The *CGF* of periodic inputs is showed in [6]. Herein we concern about the *CGF* computation of the *CB* shaped arrival.

Ethernet frame length is variable within maximum 1518 bytes and minimum 64 bytes. There are preamble signal and intervals outside a frame. To be simplified, our model defines that time slot is 40 m s for 100Mbps Ethernet and full load in one slot is 512 bytes. Regard 64byte/slot as a unit rate.

Ignoring all overhead signals, we told Ethernet frame "packet" in general terms. We also defined CBD, LC (link capacity), RP, and MF (max length of packet; note that the MF is restricted by given applications and is mostly less than 1518 bytes.).

The cumulative arrivals during various time scales are *r.v.s* : $A(s)$, s=1,2,…, t . *CGF* is logarithm of *moment generating function*. If $A(s)$'s sample space can be partitioned into some individual sets $\{A^{(k)}(s)\}$without intersection, and each set corresponds to a value $x^{(k)}$ and the probability $P\{A \quad A^{(k)}\}=p_k$, then the *CGF* is:

$$(q) \quad \log E(e^{qA}) \quad \log(\ \ p_k e^{qx_k}) \qquad (7)$$
$$k$$

For a *CB* shaped stream, there are two independent stochastic factors: (i) the random phase between the time-window (a certain time scale) and the RP (called "*outer phase*"), and (ii) the random relative position of the burstiness inside a RP (called "*inner phase*"). We get the joint probabilities by the *product formula* and compute the *CGF* directly by enumerating all the individual conditions.

To avoid the time-consuming computations, we present "3-s-5-p" and "4-s-7-p"models to provide estimators $\hat{A}$ for $A(s)$, (with fixed s). In the meaning of *convex ordering*, $\hat{A}$ is " $_{cx}$ " than $A(s)$, therefore, the *CGF* of $\hat{A}$ works as the conservative assessments for $A(s)$. We have:

Definition 1. 3-state-5-parameter model, ("3-s-5-p")

$$\hat{A} \quad \begin{array}{ll} 0, & (1 \quad m_H/\tilde{r}) \ (1 \quad p_T) \\ \tilde{r}, & (m_H/\tilde{r}) \ (1 \quad p_T) \\ \tilde{r}, & (1 \quad (m_T \quad \tilde{r})/(\hat{r} \quad \tilde{r})) \ p_T \\ \hat{r}, & ((m_T \quad \tilde{r})/(\hat{r} \quad \tilde{r})) \ p_T \end{array} \qquad \begin{array}{ll} 0 & (1 \quad \frac{m_H}{\tilde{r}}) \ (1 \quad p_T) \\ \tilde{r} & \frac{m_H}{\tilde{r}} \ (1 \quad p_T) \ (1 \quad \frac{m_T \ \tilde{r}}{\hat{r} \ \tilde{r}}) \ p_T \\ \hat{r} & \frac{m_T \ \tilde{r}}{\hat{r} \ \tilde{r}} \ p_T \end{array} \qquad (8)$$

where m is the mean, $\hat{r}$ is the peak value, $\tilde{r}$ is a selected sample value near the tail of distribution; and $P\{\hat{A} \quad \tilde{r}\}=p_T$, $\mathrm{E}\{\hat{A} \mid \hat{A} \quad \tilde{r}\}=m_T$, $m_H(1\text{-}p_T)+m_T p_T=m$.

Definition 2. 4-state-7-parameter model, ("4-s-7-p")

$$\hat{A}: \quad \begin{array}{ll} 0 & (1 \quad m_{H1}/\tilde{r}\,)\ (1 \quad p_T \quad p_{\tilde{r}}) \\ \tilde{r} & (m_{H1}/\tilde{r}\,)\ (1 \quad p_T \quad p_{\tilde{r}}) \\ \tilde{r} & p_{\tilde{r}} + (1 \quad (m_T \quad \tilde{r})/(\hat{r} \quad \tilde{r}))\ p_T \\ \hat{r} & ((m_T \quad \tilde{r})/(\hat{r} \quad \tilde{r}))\ p_T \end{array} \qquad (9)$$

where (i) $\tilde{r} = \tilde{r} - a_{min}$; $a_{min}=\{\min(a_i) \mid a_i \quad 0,\ i=0,1\ldots,RP-1\}$ is the minimum length of non-zero arrival in one slot during RP; (ii) $p_{\tilde{r}}=P\{\hat{A}=\tilde{r}\}$; (iii) $(1-p_T - p_{\tilde{r}})m_{H1}+ p_{\tilde{r}}\,\tilde{r} + p_T m_T = m$; and the other parameters defined as the same as in **Def.1**.

Remark: The literature [6] present a "two-state" Bernoulli model which is too pessimistic for *CB* shaped inputs. The models of **Def.1** and **Def.2** refer to [10] with recasting for variable packet streams. In our implements, the parameters m_T, p_T, and $p_{\tilde{r}}$ are computed by Monte Carlo method.

The hazard probability also can be computed by Monte Carlo simulation. In high performance systems, hazard event is very rare so that the *variant reduction* techniques must be employed to speedup the simulations. In fact, even in the beginning of discrete event simulation, in 1948 von Neumann and Ulam had applied them in neutron diffusion research. *Importance Sampling* (*IS*) we employed is the popular one of them.

In crude Monte Carlo method, the statistic estimator is: $$\frac{1}{N}\sum_{i\ 1}^{N} I\{X_i \quad A\} \qquad (10)$$

where X_i comes from a ensemble with original distribution $\underline{r}$ and sample size is N ; $I(.)$ denotes the *indicator function*, which is equal to 1 when the expression inside brackets is true and to 0 otherwise. If the random event A is very rare, the num of samples must be large enough to reach the acceptable precision. On the contrary, *IS* approach take samples from the twisted distributions $p(x)$ and revises each trial's result by multiplying the *likelihood ratio* $L(x)=p(x)/r(x)$. For *CB* Shaped packet multiplexing, the *IS* estimator is:

$$\frac{1}{N}\sum_{j\ 1}^{N}\sum_{l\ 1}^{m}\sum_{i\ 1}^{n_l} L^{(l)}(\hat{U}_{i,j}^{(l)})\ I(q_j(t) \quad nx) \qquad (11)$$

where (i) $\hat{U}_{i,j}^{(l)}$ is the "*inner phase*" of *i*th stream of *l*th class in the *j*th sampling; (ii) if the given time scale $[1, t \quad 1]$ contains multiple RPs, $\hat{\mathbf{U}}_{i,j}^{(l)}$ is a vector, and $L^{(l)}(\hat{\mathbf{U}}_{i,j}^{(l)})$ is the product of every

likelihood ratio corresponding to each "*inner phase*".

The key and the most difficult step in *IS* procedures is to find the proper twisted distributions termed as the "*change of measure*". Because the *measure* can't be expressed by analytic formulae for *CB* shaped traffic, suggested by [11], we develop an iterated adaptive learning approach. The idea is to feed simulation outputs into a goal-seeker that takes charge of the model parameter adjustments.

Additionally we employ the quasi-Monte Carlo method for the "*outer phase*" sampling.

A CASE FOR GUARANTEEING PROBABILITY COMPUTATION

An example for *CB* shaped packet stream multiplexing is demonstrated as **Table 1**, and results of it in the hazard probabilities for various length thresholds nx=64,72,80,88,96 (every 8 unit corresponding to time delay 1ms in 100Mbps Ethernet) are illustrated in **Figure 1**. The hazard probabilities (definitely, in logarithm form of $1/n \log(P\{q(t)>nx\})$) are computed by enumeration ("direct"), "3-s-5-p", "4-s-7-p" and *IS* simulation approaches respectively. Regarding results from "direct" computation as the truly values, it is shown that the stochastic bounds are conservative but have deviation less than one magnitude, which can be acceptable in **rare** hazard estimating context. With a little increased overhead, "4-s-7-p" achieves tighter bound than "3-s-7-p" and is preferred to the latter. *IS* simulation results are also within acceptable deviation but with inherent underestimation. However, simulation may be the terminal approach to evaluate large and complicated systems.

Table 1. An Example for CB Shaped Packet Stream Multiplexing

Class	MF	CBD	RF	Num.
1	2	4	10	6
2	8	8	15	2
3	16	24	25	2
4	8	16	50	2

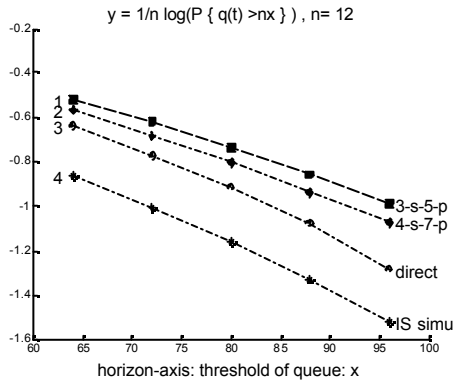

Figure 1. The Logarithm of Hazard Probabilities

MAXIMUM RATE FUNCTION FOR WORST CASE ANALYSIS

The *Maximum Rate Function* (s) defined as the maximum aggregated rate with respect to the time scale s in *worst case* condition provides a concise and unified *traffic characteristic description* framework for *service curve* based *admission control* and delay analysis in packet switched networks

[7]. The function can be deduced by the limits in positive infinity of partial derivatives of the arrival process *CGF*s with respect to q :

$$(s) \quad [\quad (q,s)/\ q\ ]_q \qquad [\quad (q,s)/q\ ]_q \qquad (12)$$

And in practice, we can get it by substituting q in the right side of (12) with 10~20 times value of the *critical time scale*. *CGF*s can also be got by empirical or measure-based methods.

SOME DISCUSSIONS ON TELEMETERING DELAY TOLERANCE

In integrated network-based data acquisition, drawback caused by fluctuations of asynchronous packet passing delay could be partly relieved by some engineering and theoretic approaches.

1) If packets are sent with fixed interval by source, for different usages, e.g.: (i) if the data are only used for off-line playback, variable communication delays don't affect negatively because playback clock is independent of the real-time clock; and (ii) for on-line processing, destination time axis is calibrated contrasting to the source by a fixed phase. Predictors and extrapolators are implemented through adaptive filtering algorithms.

2) If packet passing is driven by asynchronous events, on-line delay estimators such as the Smith predictor have to be used for time-critical applications. Another methods are highly related to the physical purposes, that is to say one should recognize the relationship between the real-time and the real world. For an instance, in a robostic tele-operation [12], the *torque* and the *acceleration* digital signals, the product of which is *power*, are transformed into *wave* variables before transmission to forbid the multiplicative errors.

CONCLUSION AND FUTURE RESEARCH

On the basis of switched Ethernet with some enhanced features in IEEE 802.1p and 802.3u, we describe a hybrid structure mixing either synchronous or asynchronous real-time packet streams. The timeliness of asynchronous multiplexing is investigated by analytic and simulation-based approaches. Enabled by scheduling strategies and quantitative performance evaluations, switched Ethernet based data acquisition systems could be used in certain real-time applications, in which real-time asynchronous transmissions are guaranteed in statistical or the *worst case* conditions.

Our future research on asynchronous multiplexing involves three topics: (i) extending *guaranteeing probability* computation to hierarchical switched environments, (ii) simulation-based performance optimization for complex configurations and (iii) delay tolerance solutions for telemetry systems.

REFERENCES

[1] Alves,M, Tover,E, Fohler,G, Buttazzo,G, "CIDER-Envisaging a COTS Communication Infrastructure for Evolutionary Dependable Real-Time Systems", The 12th Euromicro Conference on Real-Time Systems, Stockholm, Sweden, June 2000.

[2] Spataro,M, Kent,S, "Serial Communication Interface for Data Acquisition Instrumentation in a Wind Tunnel", DSTO-TR-0740, Defense Sci. and Tech. Org., Canberra, Australia, Nov., 1998.

[3] Cruz, R L. "A Calculus for Network Delay, Part I: Network Elements in Isolation", IEEE Trans. on Info. Theor., 37(1), Jan.,1991.

[4] Lewis,J.T, Russell,R, "An Introduction to Large Deviations for Teletraffic Engineers," ACM Performance'96, Lausanne, Switzerland, 1996.

[5] Chang,C-S, Thomas,J.A, "Effective Bandwidth in High-Speed Digital Networks", IEEE JSAC, 13(6), 1995.

[6] Chang,C-S, Chiu,Y-M, Song,T.W, "On the Performance of Multiplexing Independent Regulated Inputs," ACM SIGMETRICS Performance Evaluation Review, 29(1), 2001.

[7] Raha,A, Kamat,S, Zhao,W, "Guaranteeing End -to-End Deadlines in ATM Network", IEEE International Conference on Distributed Computing Systems, May., 1995.

[8] Meier,R, Cahill,V, "Taxonomy of Distributed Event-Based Programming Systems", ICDCSW'02, 2002.

[9] Kweon,S-K, Shin,K.G, Workman,G, "Archieving real-time communication over Ethernet with Adaptive Traffic Smoothing", RTAS'2000, 2000.

[10] Zhang,Z-L, Kurose,J, *et al.*, "Smoothing Statistical Multiplexing and Call Admission Control for Stored Video", IEEE JSAC, 15(6), 1997.

[11] de Boer,P.T, Kroese,D.P, Mannor,S, Rubinstein,R.Y, "A Tutorial on the Cross-Entropy Method," www.cs.utwente.nl/~ptdeboer/ce/, 2003.

[12] Munir,S, Book,W.J, "Internet-Based Teleoperation Using Wave Variables With Prediction," IEEE Trans. on Mechatronics, 7(2):124-133, Jun. 2002.