Front. Electr. Electron. Eng. China (2006) 3: 291–296
DOI 10.1007/s11460-006-0040-1

RESEARCH ARTICLE

SUN Chang-hua, LIU Bin, LI Wen-jie

Research on switching throughput of traffic manager in core routers

Abstract A general model is made to analyze switching throughput of traffic manager in core routers. By designing a real traffic manager that uses the OC-48c interface, the whole system is analyzed and it is pointed out that at least four HSSLs should be employed per CSIX interface when using Vitesse's GigaStream switch chipset. Meanwhile, at the CSIX interface, the CFrame should be constructed according to the actual size of the last cell of each IP packet. The above principles can guarantee forwarding of IP packets at line rate. A general relationship between throughput and buffering scheme of IP packets in the external memory is given, which is useful in the design of switch fabric in core routers.

Keywords throughput, GigaStream, CSIX interface, traffic manager

1 Introduction

Switch fabric is widely used in core routers to replace shared bus and memory. Between switch fabric and traffic manager, there is a standard interface: common switch interface (CSIX) [1]. The CSIX-L1 specification defines a physical interface, including data format and flow control. According to the specification, CFrame is the data format and fixed cell is required in the internal of the switch fabric. Maximum CFrame Payload size depends on the length of the internal cell in the fabric [1, 2]. Therefore, changing of data formats and possible padding will waste bandwidth, which affects the throughput of routers. References [3–5] discuss this problem, but only provide a qualitative description and analysis. A quantitative analysis is needed when designing a router. In addition, how to store IP packets are stored in the external memory affects the throughput of routers. Thus, we will analyze these two factors and give a general conclusion.

Figure 1 shows the structure of a traffic manger (THUNP) designed for OC-48c interface. GigaStream switch chipsets (VSC872/VSC882) are used and reduced latency DRAM RLDRAM II is used as external memory in the design. In the input direction, a variable length IP packet is first processed by the network processor (NP) and then transmitted to traffic manager (TM). TM stores the IP in RLDRAM II (position *C* in Fig. 1) and then schedules it to position *B*. Then, the packet is formed to CFrame and transmitted to the switch fabric. Finally, the switch fabric transmits it to the destination line card. So at position *A*, *B* and *C*, the IP packet is added some overhead, which would reduce the system's throughput. And the throughput performance in *A* and *B* is close related to the maximum CFrame Payload size. Then through mathematical modeling and analysis, we give the principles of how to choose the maximum CFrame Payload size in THUNP, which can be used in other general systems.

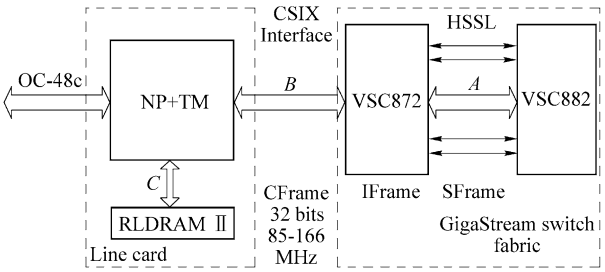

Fig. 1 The structure of THUNP line card system for OC-48c

When storing an IP packet in the external memory, it is necessary to segment the IP into a fixed cell in favor of easy memory management. At position *C*, the choice of maximum CFrame Payload size depends on different design approaches:

1) *BC* Separate. As shown in Fig. 2(a), variable length IP packets are first processed by NP, then segmented into fixed cells and stored in the external memory. At the CSIX interface, all the fixed cells belonging to the same IP are reassembled and then segmented to form a separate CFrame. In this approach, *B* has little relationship with *C*.

Translated from *Acta Electronica Sinica*, 2005, 33(7): 1243–1246 (in Chinese)

SUN Chang-hua (✉), LIU Bin, LI Wen-jie
Department of Computer Science and Technology,
Tsinghua University, Beijing 100084, China
E-mail: sch04@mails.tsinghua.edu.cn

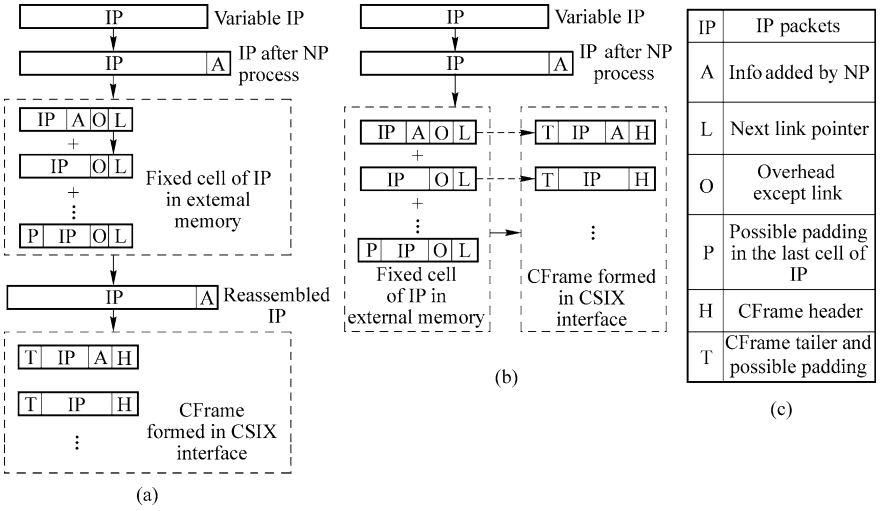

Fig. 2 Two design approaches. (a) *BC* separate; (b) *BC* together; (c) Acronym

2) *BC* Together. As shown in Fig. 2(b), variable length IP packets are first processed by NP, then segmented into fixed cells and stored in the external memory. At the CSIX interface, the cells are formed into CFrames separately. In this approach, *B* has close relationship with *C*. Maximum CFrame Payload size depends on the unit of memory management.

The choice of the two approaches depends on the following factors: system throughput, design difficulty, flexibility and scheduling algorithm. We mainly consider system throughput and design difficulty to analyze positions *ABC*.

Intel IXP 2400 is a widely used network processor, whose maximum CFrame Payload size is 64–120 bytes when connecting to GigaStream switch chip sets [5]. Users can learn to configure it with this paper. The remainder of the paper is organized as follows: Sect. 2 analyzes and gives the model of the system, and Sect. 3 analyzes the real system THUNP and discusses approaches to improve the throughput. Finally, the conclusion is presented in Sect. 4.

2 Modeling and analysis of traffic manager and switch fabric

A general system structure, as shown in Fig. 3, can be obtained by abstracting the line card system in Fig. 1. Figure 3 is the input direction, and the output direction is almost the same. In the figure, a general or special DRAM such as RLDRAM II, can be used as the external memory. As stated in Sect. 1, at position *A*, *B* and *C*, an IP packet may not be transmitted at line rate due to overhead. This means that possible variable length IP packets would not be forwarded at line rate in the system. These variable length IP packets are called DROP IP. This process is considered as little feedback and the final number of DROP IP is the sum of DROP IP at position *A*, *B* and *C*, separately.

At the same time, when computing the real bandwidth of IP packets at position *A*, *B* or *C*, we do not consider the IP packet that may not be forwarded at line rate at the other two positions. We simply compare the real bandwidth at that position and the input line rate. Obviously, this modeling ignores some unimportant factors, but can also give us interesting results of the real system.

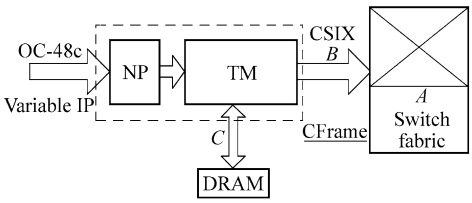

Fig. 3 The abstract structure of OC-48c line card system

The bit rate at OC-48c interface R(OC-48c) is:

$$R(\text{OC-48c}) = 155.52 \times 16 \times \left(\frac{260}{270}\right) = 2.396\,16\ \text{Gb/s} \tag{1}$$

According to PPP Protocol [6], in the HDLC Frame of OC-48c, there is a 9-byte overhead: 1 byte Flag, 1 byte Address, 1 byte Control, 2 bytes Protocol, and 4 bytes FCS. Therefore, for an IP packet with length y bytes, the real bandwidth at OC-48c R(IP@OC-48c) is:

$$R(\text{IP@OC-48c}) = R(\text{OC-48c})\frac{y}{(y+9)} \tag{2}$$

Due to overhead at *A*, *B* and *C*, the effective bandwidth obtained is related to the raw bandwidth, IP length, and the overhead length.

$$\text{Effective}_{\text{BW}} = \text{Raw}_{\text{BW}} \frac{\#\text{IP}}{\#\text{IP}+\#\text{overhead}} \tag{3}$$

IP packets can be forwarding at line rate, if satisfying:

$$\text{Effective}_{\text{BW}} \geq R(\text{IP@OC-48c}) \tag{4}$$

We classify the IP packets by their length. If one packet size cannot satisfy Eq. (4) at position *A*, *B* or *C*, we say that IP packet of this length (this IP class) could not be forwarded at line rate. This hypothesis has limitations, but we could consider extreme situations such as system test. The

goal of the system design is to make the number of IP class that could not be forwarded at line rate as few as possible, which would improve the system throughput.

3 Analysis of the throughput of THNPU

In this section, we analyze the real system THNPU and discuss the approaches of improving the system throughput.

Figure 1 shows the structure of THNPU. VSC872 connects with the CSIX interface of THNPU and receives the CFrame from Traffic Manager. Then, in VSC872, the header (CH) and tailer (CT) of CFrame are taken out. If the payload size is less than the maximum CFrame Payload size, padding is added to make it equal to maximum CFrame Payload size. Then, the message is converted into internal data segment format: IFrame. High-speed serial link (HSSL) is used between VSC872 and VSC882. And 2 or 4 HSSLs can be used per CSIX interface. An HSSL can run at 2.125 Gb/s or 2.643 84 Gb/s, but the efficient data bandwidth is 2.0 Gb/s or 2.5 Gb/s due to overhead. SFrame is the data segment format at HSSL and its length equals to maximum CFrame Payload size plus 12 bytes. Figure 4 shows the data format between CFrame, IFrame and SFrame. In the figure, CPayload represents CFrame's payload, including some padding. IH is the IFrame's header and IF is the tailer. CP is the padding in IFrame to make CFrame payload size equal to maximum CFrame Payload size. And IFP is the other padding. SH is the SFrame's header. The maximum CFrame Payload size is allowed in the discrete range from 40 to 120 bytes in increments of 4 bytes in GigaStream [7, 8].

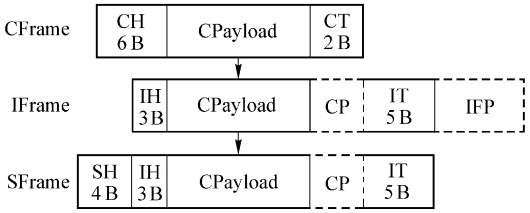

Fig. 4 Change of the data segment formats

As shown in Fig. 1, the maximum CFrame Payload size affects the real bandwidth that an IP packet can obtain at position *A*, *B* and *C*. Therefore, in this section, we first discuss how to choose the maximum CFrame Payload size at position *A* and *B*, then analyze the throughput at position *C* with different *BC* design approaches. Finally, a conclusion is made based on the analysis of the three positions.

3.1 Choice of maximum CFrame Payload size at *A*

Since the SFrame size is equal to maximum CFrame Payload size plus 12 bytes, it is easy to compute the real bandwidth of IP packet at *A*. In the best situation (ideal system), CFrame Payload has no overhead and we do not consider any information that may be added to the IP packet by NP. Then, if the maximum CFrame Payload size is x bytes, the real bandwidth of IP packet with y bytes obtained at *A* is:

$$R(\text{IP@HSSL}) = R(\text{HSSL})\frac{y}{(x+\text{SF})\left\lceil \frac{y}{x} \right\rceil} \tag{5}$$

$\lceil n \rceil$ represents minimum integer greater than n. R(HSSL) is the data rate at HSSL. (x+SF) is SFrame size and SF equals 12.

When per CSIX interface uses two HSSLs, R(HSSL) equals 2×(2.643 84 × 16 ÷ 17). Then R(HSSL) is almost two times as R(OC-48c) and it is considered that two HSSLs would forward all class of IP packets. But that is not the truth.

For any x, we can compute R(IP@HSSL) and R(IP@OC-48c) of all classes of IP packets(classified by length, 40–655 35 bytes). Then we obtain the number of classes that could not be forwarded at line rate according to Eq. (4) and show them in Fig. 5(a). The best choice of x is 44 to 88 bytes. However, in the real system, as shown in Fig. 2, CFrame Payload needs some overhead a bytes, including information for IP reassembly and for CFrame format. NP also adds some information to IP packets, such as next IP address (NIP). Supposing their length is b bytes, and then at HSSL, the real bandwidth of IP packet obtained is shown in Eq. (6). In THUNP, a equals 4 and b equals 8. Then the number of IP classes that could not be forwarded at line rate according to Eq. (4) is shown in Fig. 5(b). At this time, any maximum CFrame Payload size could not forward all length of IP packets. The best choice of x is 40. If the smallest IP packets could form a single CFrame, x should be greater than 52 and this will do favor to the other system module. Thus the good choice of x is 52 bytes to 88 bytes.

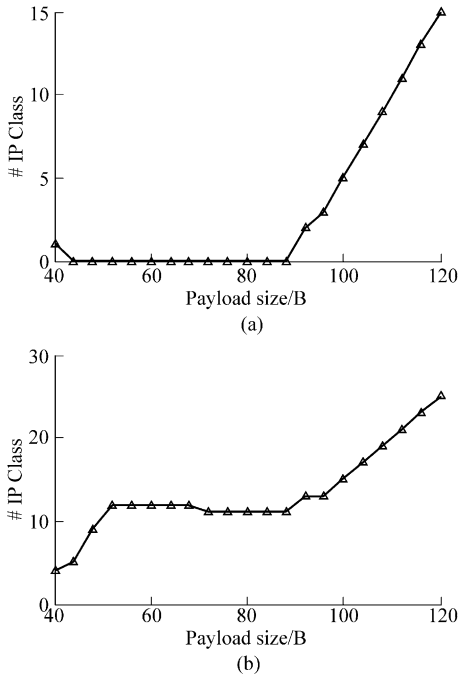

Fig. 5 No. of IP class without line rate at HSSL. (a) Ideal system; (b) THUNP system

$$R(\text{IP}@\text{HSSL}) = R(\text{HSSL})\frac{y}{(x+\text{SF})\left[\frac{y+b}{x-a}\right]} \tag{6}$$

If a equals 2 and b equals 0 (or a=0, b=4), any maximum CFrame Payload size could also not forward all length of IP packets. Therefore, in real systems, due to over head, it is impossible to use two HSSLs per CSIX to forward all length of IP packets. We need four HSSLs per CSIX. At this time, any maximum CFrame Payload size is met the requirement in spite of running at 2.125 Gbps or 2.643 84 Gbps. As shown in Fig. 6, if we consider as many classes as possible as of IP that could obtain speedup 2, and also the smallest IP packets that could form a single CFrame, the good choice of x is 52 bytes to 88 bytes.

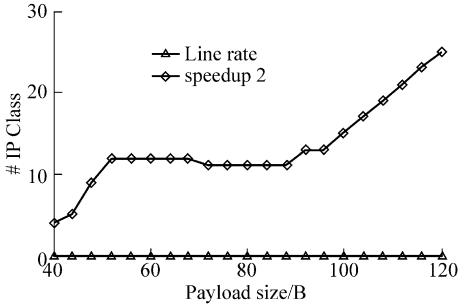

Fig. 6 No. of IP class without line rate or speedup two when using four HSSLs per CSIX

3.2 Choice of maximum CFrame Payload size at *B*

At CSIX interface, the clock frequency GigaStream supporting is between 85 MHz to 166 MHz. As shown in Fig. 2, there are two approaches to form CFrame. Fixed Formed CFrame: The IP packet is segmented into fixed cell. If needed, the last cell is padded to the length of fixed cell. And the fixed cell is formed CFrame separately. Variably Formed CFrame: The last cell of IP packet is formed CFrame according to its real length and possible padding is taken out. The former method is very simple but there is waste of bandwidth due to padding of the last cell, especially when the IP packet is short. The latter method is somewhat complicated, but the effective bandwidth is higher.

In Fixed Formed CFrame, the effective bandwidth of IP packet obtained at CSIX interface is computed like Eq. (6). Just SF is replaced with CF, and R(HSSL) with R(CSIX). R(CSIX) is the bandwidth of CSIX and (x+CF) is the lengthe of CFrame. CF equals 8. When CSIX runs at 166 MHz, for any maximum CFrame Payload size x, we obtain the number of classes that could not be forwarded at line rate and show them in Fig. 7. The best choice of x is 92, 96 and 100 (Assuming x is greater than 52). However, through deep research, this system is not stable. If CSIX runs at 165 MHz, any x cannot forward all length of IP packets. When CSIX runs at the recommended clock frequency 100 MHz by the CSIX-L1 Specification, the minimum class of IP that cannot be forwarded at line rate is 251. When running at 125 MHz, the minimum number is 191.

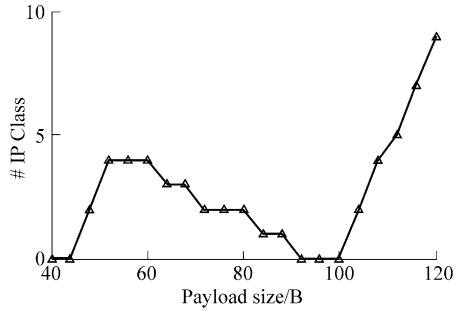

Fig. 7 No. of IP class without line rate at Fixed Formed CFrame

In Variably Formed CFrame, the effective bandwidth of IP packet with length y is:

$$R(\text{IP}@\text{CSIX}) = R(\text{CSIX})\frac{y}{(x+\text{CF})(n-1)+\text{last}}$$
$$\text{last} = \left(\left[\frac{y}{4}\right]4+b\right)-(x-a)(n-1)+\text{CF}+a \tag{7}$$

In Eq. (7), $n = \left[\frac{y+b}{x-a}\right]$ (n-1) cells are formed with the maximum CFrame Payload size and the last cell with its actual size. When CSIX runs at 166 MHz, any x could forward all length of IP packets. Figure. 8 shows the minimum effective bandwidth of all maximum CFrame Payload size when running at 166 MHz. Through computation, when CSIX runs at 125 MHz, any x can forward all length of IP packets. Therefore, it is necessary to use Variably Formed CFrame to achieve a good system throughput.

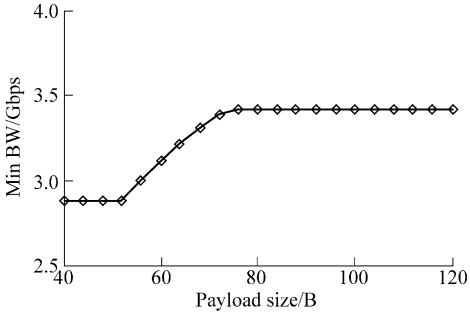

Fig. 8 Minimum effective bandwidth

Just considering position A and B, the following conclusion can be made. At least four HSSLs should be employed per CSIX interface. At the CSIX interface, the CFrame should be constructed according to the actual size of the last cell of each IP packet. These principles make the system design stable and flexible. And any maximum CFrame Payload size could forward all length of IP packets. If we consider as many classes as possible of IP packets that can

obtain speedup 2, and also the smallest IP packets that can form a single CFrame, the good choice of x is 52 bytes to 88 bytes.

After analysis of the real trace Auckland-II [9], there are many IP packets with length less than 64 bytes and their packet size is about 60 % [10]. Therefore, if the IP packet with length less than 64 bytes can form a single CFrame, it would do favor to the other modules like network processor and traffic manager. In this sense, the good choice of x is greater than 76 bytes. And we can choose maximum CFrame Payload size as 76 bytes to 88 bytes.

3.3 Throughput consideration at *C*

As stated in Sect. 1, there are two design approaches between *B* and *C*: *BC* Separate and *BC* Together. In *BC* Separate, a reassembly and segmentation operations are added, but the overhead in external memory is less. In *BC* Together, the design is simple but overhead is higher. In two approaches, the effective bandwidth of the external memory is computed as in Eq. (6), as follows:

$$R(\mathrm{IP}@\mathrm{Mem}) = R(\mathrm{Mem})\frac{y}{\mathrm{Mem}\left\lceil \frac{y+b}{\mathrm{Mem}-c} \right\rceil} \tag{8}$$

R(Mem) is the total bandwidth of the external memory when reading or writing data. Mem is the length of the fixed cell and also the unit of memory management. c is overhead in the cell and b is the information added by NP.

In THUNP, RLDRAM II [11, 12] is used as external memory. Its data bus of read and write is separated, but it is not a real dual port memory because read/write command can not be issued in the same clock. At least burst 4 is needed to simultaneously do read/write of data and data bus not empty. The configuration of RLDRAM II is 16 M with 18 bits data bus, and it has 8 different banks. In order to avoid bank conflicts, 8 banks are needed to read/write together. Therefore, Mem is 64 bytes when burst is 4 and Mem is 128 bytes when burst is 8. After computation, Mem with 64 bytes is better. One clock is needed to refresh memory in 64 clocks. R(Mem) is computed as follows when RLDRAM runs at 166 MHz.

$$R(\mathrm{Mem}) = \frac{1}{6}\times 16\times 2\times \frac{63}{64}\ \mathrm{Gb/s} \tag{9}$$

In THUNP, with *BC* Separate, b = 12, c = 4; with *BC* Together, b = 8, c = 8. Due to more overhead, the effective bandwidth of *BC* Together is less than *BC* Separate. Also, *BC* Together affects the choice of maximum CFrame Payload size because the separate fixed cells are formed to CFrame. Figure 9 shows the effective bandwidth in *BC* Separate.

In *BC* Separate, reassembly and segmentation operations are added as shown in Fig. 10. We implement them with Quartus 4.2 on Stratix EP1S80 [13]. The LE (logic element) of Reassembly is less than 1% and memory less than 1%. LE of Segmentation is less than 1% and memory less than 1%. So the added operations are not very complicated. But when using *BC* Separate, *C* would not affect the choice of maximum CFrame Payload size. We can choose it just according to *A* and *B*, and system design could be more flexible.

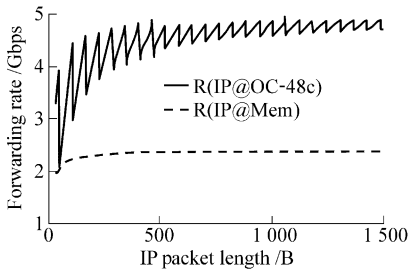

Fig. 9 Effective bandwidth using *BC* Separate

Fig. 10 Added operations in *BC* Separate

From Fig. 9, we can see the effective bandwidth is not very high with some lengths of IP packets. The method to improve the bandwidth is to reduce overhead especially the padding in the last cell. Bank reordering and conflicts solution in RLDRAM II should also be considered, but this is beyond the scope of this paper and left for further research.

3.4 Put *ABC* together

When using the *BC* Together design approach, the maximum CFrame Payload size depends on the external memory. One cell is formed to CFrame and is used as unit of memory management. When using *BC* Separate, the system is much flexible and choice of maximum CFrame Payload size is just according to *A* and *B*. 76 bytes to 88 bytes are good choices.

4 Conclusions

We analyze the factors affecting throughput of traffic manger and make a general model. By designing and analyzing a real OC-48c traffic manager, THUNP, we point out that at least four HSSLs should be employed per CSIX interface when using Vitesse's GigaStream switch chip sets. Meanwhile, at the CSIX interface, the CFrame should be constructed according to the actual size of the last cell of each IP packet. The above principles can guarantee forwarding of IP packets at line rate. If we consider as many classes as possible of IP packets that can obtain speedup 2, and also the smallest IP packets that can form a single CFrame, maximum CFrame Payload size with 76 bytes to 88 bytes is a good choice. Also, external memory does affect the system throughput. We recommend the separate design of storing IP packets in the external memory with CSIX inter-

face. This will make the system flexible.

The input direction is analyzed in this paper and the output direction is similar. If we use a switch system other than GigaStream VSC872/882, similar analysis can be made according to the model of this paper. There is much further work to do especially improving bandwidth of external memory.

Acknowledgements This work was supported by the National Natural Science Foundation of China (No. 60173009 and No. 60373007), the Hi-Tech Research and Development program of China (No. 2002AA103011-1 and No. 2003AA115110), and China/Ireland Science and Technology Collaboration Research Foundation (CI-2003-02). Authors would like to thank Mr. Hu Chengchen and Mr. Li Jing for their discussions.

References

1. CSIX-L1 Specification V1.0, http://www.npforum.org/techinfo/csixL1.pdf, Aug. 5, 2000
2. CSIX-L1 Frequently asked questions, http://www.npforum.org/techinfo/CSIX_FAQ_D1.0.pdf, Dec. 5, 1999
3. Why modern switch fabrics use a fixed-size frame format, Vitesse Semiconductor Corporation, white paper, V1.0, http://www.vitesse.com, Jan. 27, 2004
4. Collier S., Grudsky M., The switch fabric multiservice dilemma, Vitesse Semiconductor, CommsDesign, http://www.commsdesign. com/story/OEG20020702S0035, July 2, 2002
5. Intel IXP2400 network processor / Vitesse GigaStream Switch Fabric Solution White Paper, Intel Corporation, V 1.0, http://www.intel.com/design/network/papers/25214201.pdf, November, 2002
6. Perkins D, PPP: the point-to-point protocol for the transmission of multi-protocol datagrams over point-to-point links, IETF RFC 1171, 1990
7. Gigastream intelligent switch fabric VSC872/VSC882 Design Manual, http://www.vitesse.com, 2002
8. VSC872 and VSC882 data sheet, http://www.vitesse.com, 2003
9. National Laboratory for Applied Network Research (NLANR), Auckland-II, http://pma.nlanr.net/Special, 2004
10. Li wen-jie, Liu Bin, Preemptive short-packet-first scheduling in input queueing switches, Acta Electronica Sinica, 2005, 33(4): 576–583 (in Chinese)
11. RLDRAM II Data Sheets and Technical Notes, http://www.rldram.com/datasheets/index.html, 2003
12. 288 Mb SIO REDUCED LATENCY (RLDRAM II) Datasheet, http://www.micron.com, 2003
13. Stratix device handbook, http://www.altera.com, April, 2004