Structure and Latency Analyses for High Performance Computing System Based on Asynchronous Optical Packet Switching

Zhao Jun and Sun Xiaohan
Department of Electronic Engineering
Southeast University
Nanjing 210096, China
E-mail: zhaojun1115@seu.edu.cn
xhsun@seu.edu.cn

KEYWORDS
High performance computing system, Asynchronous optical packet switching, Distributed management structure, Stability, Latency

ABSTRACT

A novel high performance computing system based on optical packet switching and optical multicast technologies is presented. Distributed management architecture is used to alleviate the storing and computing pressures of every stage, which is easy to realize all-optical scalability. Asynchronous switching mode is accepted at every stage for high-speed and huge-capacity burst services transmission. The system scale and the stability features are analyzed, and a two stage system which interconnects 38,400 CPUs is adopted. Moreover, the average packets waiting latencies caused by the scheduling units and the recycling-fiber-delay-line based collision resolution units are simulated as 12.9ns and 0.63ns, respectively.

1. INTRODUCTION

High performance computing systems (HPCS) use high-bandwidth and low-latency links to interconnect huge amounts of distributed microprocessors for providing timely exchanging of high-speed and large-capacity services[1,2]. The BlueGene/L System, a joint development of IBM and the Department of Energy's (DOE) National Nuclear Security Administration (NNSA), has been significantly scaled up from 65,536 to 106,496 nodes and now has achieved a Linpack benchmark performance of 478.2 TFop/s. Accordingly, the development of HPCS is in the tendency of higher speed and more processors, so more pressures are placed on the performance of the interconnection network [3].

The traditional electrical link is becoming the bottleneck for high-speed and high-capability data transmission, on the other hand, the static optical interconnection technology, such as optical circuit switching, can not meet the need for burst services transmission [4]. Optical packet switching (OPS) technology which has the effective bandwidth utilization ability and fine exchanging granularity is becoming the most promising one in next generation optical network. HPCS based on OPS technology can improve the parallel exchanging ability for the system and is very attractively in HPCS design [5, 6]. However, the switching unit as well as the collision resolution module are still not mature, also, synchronous switching technology and centralized management structure are commonly used which are not easy to be constructed and not favorable for system scalability and will also induce more queue latencies [7].

This paper proposes a HPCS system based on asynchronous OPS and optical multicast (AOPS & M-HPCS) technologies. Distributed management structure is used to alleviate the storing and computing pressures of every stage and can easily to realize all-optical scalability. Multiple CPUs share one optical transceiver which can increase the system scale. The optical-switch (OS) based on semiconductor optical amplifiers (SOA) combining with optical splitter are used as switching units, meanwhile, the recycling fiber delay lines (Rec-FDL) are used as collision resolution units. The system scale, stability and delays induced by scheduling-unit and Rec-FDL are simulated.

2. ARCHITECTURE ANALYSES FOR AOPS & M-HPCS SYSTEM

Figure 1 shows two-stage AOPS & M-HPCS system structure where the master-node which has higher level controls n slave-nodes only, meanwhile, each slave-node manages n scheduling-units (SU) which controls m CPUs respectively and ordinal sends packets from CPUs to the optical transceiver (TX/RX). By using the 80 wavelength dense wavelength division multiplexing (DWDM) links with single channel capacity of 40 Gbit/s, the value of n can be confirmed as 80, and the two-stage system can interconnect $6,400\times m$ CPUs.

The Edge-note (EN) which constituted by electrical buffers and packets assembly units can assemble the data from CPUs into packet-payloads, which will be exchanged by the optical switching units (OSU) in optical domain. Information such as storage capacities and computing abilities needed by each service is carried in the packet-labels which will be extracted and processed electronically by the controlling-units (CU).Each packet is transmitted asynchronously for

Proceedings of the 2008 High Performance Computing & Simulation Conference ©ECMS
Waleed W. Smari (Ed.)
ISBN: 978-0-9553018-7-2 / ISBN: 978-0-9553018-6-5 (CD)

decreasing the queue delay, and is exchanged firstly among CPUs inside the slave node where it is generated in, which can reduce the transmission latencies and arbitration complexities of every stage. If no destinations can be found in this slave-node, the packet will be exchanged to the master-node, and then will be allocated to other slave-nodes.

The SOA has lots of advantages such as nanosecond high-speed switching ability, low controlling voltage, easily been integrated, and so on. Consequently, it can improve the switching speed by using the SOA based optical-switches and a 1×81 optical splitter to construct the OSU. 80 of the splitter output are used to interconnect the SUs in the slave-node, while the additional one is dedicated to the master-node connection. A packet can be switched to one or more destinations by controlling the on-off states of all the SOAs which can realize the optical multicast.

It may cause wrong receiving if two or more packets arrive at the SOA-switch in its once tuning-time. Here we use a 2×2 optical switch (OS) combine with a Rec-FDL as the collision resolution units, which will send the lower priority packets into the Rec-FDL for delaying. The tunable wavelength converters (TWC) can avoid wavelength conflicts between the downstream signals and the slave-node signals and can also avoid collisions among upstream signals in the master-node.

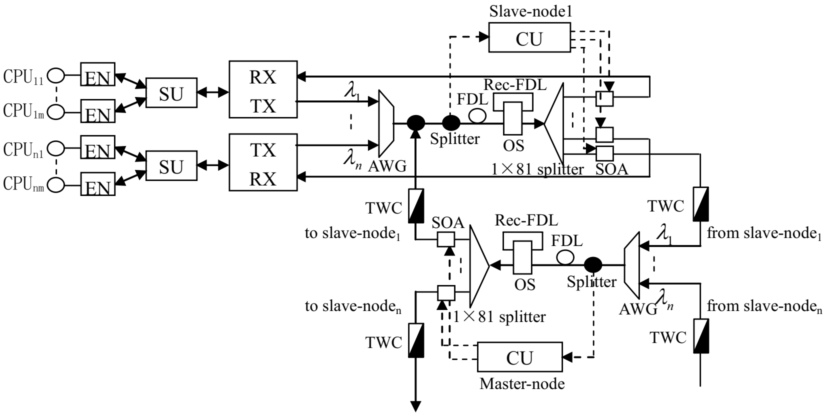

Figure 1: The Infrastructure for Two-stage AOPS＆M-HPCS (EN: Edge Node; SU: Scheduling Unit; RX/TX: Optical Transceiver; CU: Controlling Unit; Rec_FDL: Recycling Fiber Delay Line; OS: Optical Switch).

3. PERFORMANCE ANALYSES FOR AOPS & M-HPCS SYSTEM

3.1. Latencies Caused by the Scheduling Units

The SU transmit the packets from all the attached ENs in a polling mode. The inquiry time t' for one EN is L/V, where the packet length L is 256Byte in this paper, and the optical transmitter rate V is 40Gbit/s. If the SU finds no packets in the EN, then t' equals to 0, and the probability distribution for t' is approximated expressed as follows:

$$P=\begin{cases}1-e^{-\lambda mL/V}, t'=L/V \\ e^{-\lambda mL/V}, t'=0\end{cases} \tag{1}$$

Here, λ represents the new packets arriving rate. Therefore, the average value for t' can be obtained from (1), which is shown as $\bar{t}=(1-e^{-\lambda mL/V})(L/V)$, then the polling cycle for all the m ENs is $m\bar{t}$. If one EN generates j packets in this period, then the packets longest waiting latency is $T_1=jm\bar{t}$, and this value will change in the subsequent polling cycle, the fluctuation value Δt_1 has two possibilities, which are shown as follows: $\Delta t_1=-m\bar{t}$, there have no new packets arrive at this EN in time $-m\bar{t}$, while $\Delta t_1=im\bar{t}$ (i=1,2,…) represent that i new packets are generated. Therefore, the mean value for Δt_1 is:

$$E(\Delta t_1)=-m\bar{t}e^{-\lambda m\bar{t}}+\sum_{i=1}^{\infty}im\bar{t}\frac{(\lambda m\bar{t})^i}{i!}e^{-\lambda m\bar{t}} \tag{2}$$

The system is stable only if the longest waiting latency is gradually decreasing. Accordingly, the expression of $E(\Delta t_1)<0$ must be contented for ensuring the system to be stable.

Assuming that the time required for T_1 to decrease to zero is represented as $T_1(j)$, which is schematically depicted by the calculation flow-chart as shown in figure 2, so the average latencies for the packets waiting in the SU is given by:

$$T_{SU}=\sum_{j=1}^{\infty}\frac{(\lambda m\bar{t})^j}{j!}e^{-\lambda m\bar{t}}T_1(j) \tag{3}$$

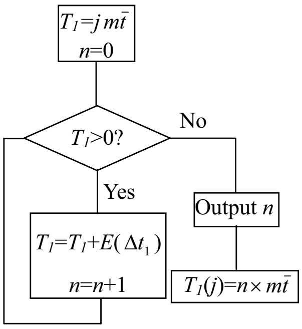

Figure 2: Calculation Flow-chart for $T_1(j)$

3.2. Latencies Caused by the Rec-FDL

A wrong receiving occurs whenever j+1 packets arrive simultaneously at a SOA-switch inside its once tuning-time t, then j lower priority packets will be switched into the Rec-FDL, with the arriving as well as departing time shown in figure 3. Here, $\varDelta T$ represents the interval of the packets arriving time, T' represents the time-slot that may generate packets in the period of t since the packet p_j enter the Rec-FDL.

The longest waiting latency for the j packets is T_2=jt, which will be changed if other packets arrive at the Rec-FDL inside the next period of T'. There also have two values for the fluctuation-value ($\varDelta t_2$), one is $\varDelta t_2$=-t, represents that no packets arrive at the Rec-FDL, and the other is $\varDelta t_2$=it (i=1,2,…), which shows that i packets are sent into the Rec-FDL. Accordingly, the mean value for $\varDelta t_2$ can be shown as follows:

$$E(\Delta t_2) = -te^{-80\lambda' T'} + \sum_{i=1}^{\infty} it \frac{(80\lambda' T')^i}{i!} e^{-\lambda' T'} \qquad (4)$$

Where, λ' represents the departing rate for the packets from the SU, and the value of 80 is the number of the SUs controlled by one slave-node.

If there have i CPUs generate packets in once polling cycle $m\bar{t}$, then $\lambda' = i/(m\bar{t})$, and the mean value can be shown in (5).

$$\lambda' = 80\sum_{i=0}^{m} \binom{m}{i} (1-e^{-\lambda m\bar{t}})^i (e^{-\lambda m\bar{t}})^{m-i} i/(m\bar{t}) \qquad (5)$$

The system is stable only if the T_2 can decrease to zero. Accordingly, the other stability condition for the system is $E(\varDelta t_2)$<0.

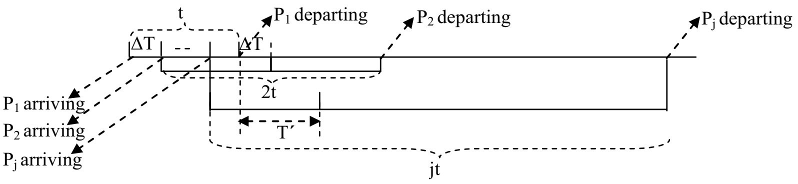

Figure 3: The Arriving and Departing Time for the j Packets inside the Rec-FDL.

The average waiting latencies ($T_{Rec\text{-}FDL}$) induced by the Rec-FDL can be expressed in equation (6), where $T_2(j)$ represents the time required for T_2 to decrease to zero, with the calculation flow-chart shown in figure 4.

$$T_{\mathrm{Rec}-FDL} = \sum_{j=1}^{\infty} \frac{(\lambda' t)^j}{j!} e^{-\lambda' t} T_2(j) \qquad (6)$$

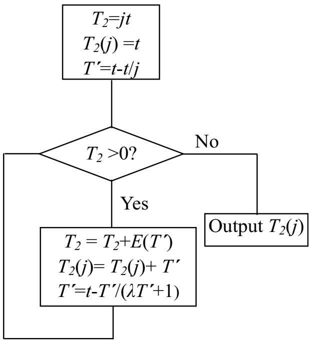

Figure 4: Calculation Flow-chart for $T_2(j)$

4. SIMULATION ANALYSES

The simulation parameters are shown as follows: t=2ns, T'=1ns. According to above analyses, $E(\varDelta t_1)$<0 as well as $E(\varDelta t_2)$<0 must be satisfied for ensuring the stabilities of the system. It can be seen from figure 5 that the number of the CPUs (m) controled by the SU varies inversely with λ, the maximum value of m is 19 as λ=1×10^{6}packets/s, and is 29 when λ decreases to 0.6×10^{6}packets/s. Accordingly, both the system scale and the packets arriving rate must be considered in this AOPS & M-HPCS design. Moreover, these values can always ensure $E(\varDelta t_2)$<0 according to figure 6. In this paper, the value of λ=1×10^{6}packets/s and m=6 are adopted, therefore, the two-stage system can interconnect 38,400 CPUs.

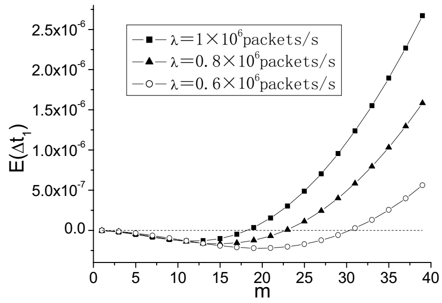

Figure 5: Relations of $E(\varDelta t_1)$ versus m and λ

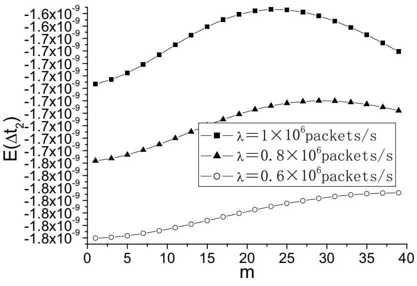

Figure 6: Relations of $E(\varDelta t_2)$ versus m and λ

As the system scale increases with m, and the number of the packets, which enter the system simultaneously, are also increased with λ, therefore, the larger value of the $m\lambda$ will cause more collisions, which will induce higher blocking rate and more waiting latencies. From figure 7 and figure 8 we can see that when $m\lambda$=6×10^{6}packets/s, the latencies caused by the SU collisions (T_{SU}) and the Rec-FDL collisions ($T_{Rec\text{-}FDL}$) equals to 12.9ns and 0.63ns, respectively.

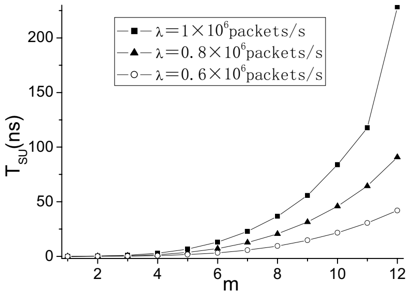

Figure 7: Relations of T_{SU} versus m and λ

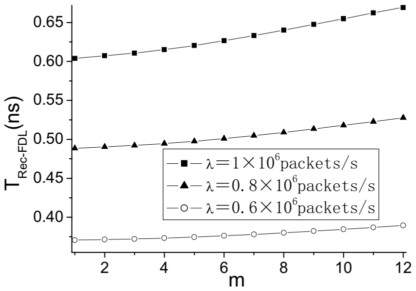

Figure 8: Relations of $T_{Rec\text{-}FDL}$ versus m and λ

Furthermore, the latencies caused by the Rec-FDL in the master-node can be analyzed with the same methods as described above, and the simulation results are influenced by the system parameters, such as the packet length, optical transceiver rate, as well as the tuning-time for the SOA-switch, and so on.

5. CONCLUSIONS

A novel HPCS based on asynchronous OPS and optical multicast technologies is presented. The dense wavelength division multiplexing transmission technologies together with the multistage distributed management topologies are used to construct a scalable interconnection network, which is suitable for timely and stochastic accesses for high-speed and massive burst services. The collision resolution unit based on the Rec-FDL are described in detail. The stabilities and the scale of the system are analyzed, and the latencies caused by the the Rec-FDL and the collisions in the scheduling units for a 38,400 CPUs interconnection system are simulated, which is 0.63ns and 12.9ns, respectively. Moreover, the experimental researches will be done later.

REFERENCES

[1] Hawkins C, Small B A, Wills D S, et al, "The data vortex, an all optical path multicomputer interconnection network [J]", *IEEE Transactions on Parallel and Distributed Systems*, 2007, 18(3), pp. 409-420.

[2] Hawkins C, Wills D S, "Impact of number of angles on the performance of the data vortex optical interconnection network [J]", *Journal of Lightwave technology*, 2006, 24(9), pp. 3288-3294.

[3] Drost R, Forrest C, Guenin B, et al, "Challenges in building a flat-bandwidth memory hierarchy for a large-scale computer with proximity communication [C]", *Proceedings of the 13th Symposium on High Performance Interconnects*, Palo Alto, CA, August 2005, Page(s):13-22.

[4] Barker K J, Benner A, Hoare R, "On the feasibility of optical circuit switching for high performance computing systems [C]", *IEEE, Conference on Supercomputing*. Seattle, USA, 2005, pp. 1-22.

[5] Masetti F, Chiaroni D, Dragnea R, et al, "High-speed high-capacity packet-switching fabric: a key system for required flexibility and capacity [J]", *Journal of Optical Networking*, 2003, 2(7), pp. 255-265.

[6] Hemenway R, Richard R. Grzybowski, "Optical-packet-switched interconnect for supercomputer applications [J]", *Journal of Optical Networking*, 2004, 3(12), pp. 900-913.

[7] Minkenberg C, Abel F, Muller P, et al, "Designing a Crossbar Scheduler for HPC Applications [J]", *IEEE Micro* , 2006, 26(3), pp. 58-71.

AUTHOR BIOGRAPHIES

JUN ZHAO, he is currently pursuing his Ph.D degree in the lab of optical communications, department of electronics engineering, Southeast University, Nanjing, China. His research interests include the high performance computing system design, optical packet switching, high speed optical signal processing, and transparent optical networks.

XIAOHAN SUN, a professor of electronics engineering, Southeast University, Nanjing, China. She was the visiting professor at Research Lab of Electronics, MIT, USA from 2002 to 2004. Her current research interests are photonic devices, high-speed photonics systems and next generation optical networks (NGON), including (a) optical pulse propagation in WDM systems influenced by nonlinear effects, PMD, crosstalk and so on, (b) management and control for NGON, (c) reliability and survivability for NGON, (d) Optical fiber sensor technology and optical imaging, and (e) design and measurement for semiconductor materials based PLCs.