

Internal Backpressure for Terabit Switch Fabrics

Anna V. Manolova, *Member, IEEE*, Sarah Ruepp, *Member, IEEE* Andreas Rytlig,
Michael Berger, *Member, IEEE*, Henrik Wessing, *Member, IEEE*, and Lars Dittmann, *Member, IEEE*

Abstract—This paper proposes and analyzes the efficiency of novel backpressure schemes for Terabit switch fabrics. The proposed schemes aim at buffer optimization under uniform traffic distribution with Bernoulli packet arrival process. Results show that a reduction of the needed maximum buffer capacity with up to 47% can be achieved with switch-internal backpressure mechanisms at the expense of a small control overhead.

Index Terms—Terabit switches, 100 Gigabit Ethernet, backpressure, flow control, switch architecture.

I. INTRODUCTION

THE Internet is increasingly populated by bandwidth-demanding applications. Following the evolution of 10 Gigabit Ethernet (GE), 100 Gigabit Ethernet is currently emerging as a promising candidate to fulfill the request for increased line speed.

Switching 100 GE requires that the selected switch architecture is scalable enough to accommodate high capacity transmission [1], as the 100 GE line speed quickly aggregates to the Terabit scale within the switch fabric. Single-stage switch fabrics, e.g. crossbar switches, do not scale up to a Terabit system. Hence, a multi-stage switch fabric, where the traffic can be distributed on different chips, is a promising approach. In particular, the Clos' architecture relies on three stages of switching modules, where each module connects to all the modules in the adjacent stages via a unique path [2], as illustrated in Figure 1. Each of these modules may then be constructed as an individual crossbar switch. However, the sheer amount of data that must be switched in a Terabit system faces a number of obstacles. Since the different traffic flows arriving at the input of a switch are independent of each other, it may happen that two cells are destined for the same destination simultaneously. This causes contention and requires the use of queues and buffers and to avoid packet loss. However, the available memory in a Terabit switching system is scarce. In order to avoid buffer overflow and associated packet loss, backpressure (BP) mechanisms can be applied [3].

In general two types of flow control can be applied: internally (link-level) between the individual stages (modules) of a switch; and end-to-end between the input and output traffic managers [4]. Another classification can be done based on the operation. Backpressure (BP) is the simplest flow control, which is based on sending a queue status indicator from a downstream queue to an upstream queue: 0 (operational status) and 1 (overloaded condition). Under such a scheme,

the upstream queue stops sending traffic to the particular downstream queue [3]–[5]. Credit based flow control schemes are another alternative, where downstream queues signal back to upstream queues special "credits" which are used by the upstream queues in order to decide if and how much traffic can be forwarded downstream [6], [7]. This type of flow control is more fine-grained since it operates on per virtual connection (per flow) basis. Despite its flexibility and efficiency the approach requires higher complexity. Book-keeping and credit control for switches with many inputs/outputs become a hurdle especially at high speeds.

In this paper, we propose and evaluate the performance of a novel BP scheme for Terabit switches. The most significant contribution of our work is that unlike the standard queue-to-queue operation of the BP schemes, our scheme is queue-to-module, where the controlled module does not employ queueing. This gives us the ability to apply finer analysis and control. We consider a specific Clos network architecture, which influences the design of the mechanisms. Furthermore, we focus on link-level BP which facilitates independent implementation of the switch fabric and the traffic manages and makes interoperability easy. Last, we use discrete-event simulation for the evaluation of the BP schemes, not mathematical formulations as in the majority of the existing work.

The remainder of the paper is organized as follows: Section II explains the switch architecture. Section III details the proposed BP mechanisms. In Section IV the simulation study is presented and in Section V the results are discussed. Section VI concludes the paper.

II. SWITCH ARCHITECTURE

We focus on a Clos-based packet switch architecture due to its potential to provide adequate throughput at high bit rates [8]. Clos switches have three stages which can be either buffered or bufferless. Depending in which stage the memory is allocated, different architectures are possible, e.g., Space-Memory-Space, Memory-Space-Memory [9], Memory-Memory-Memory, Space-Memory-Memory (SMM) [8] (adopted in this work). For the first architecture to be practical advanced schedulers are required [10], whereas the second architecture requires complex dispatching schemes [9]. The Memory-Memory-Memory is considered the best performing Clos architecture, achieving 100% throughput, but it has been suggested recently that the memory in the first stage is not necessary [8]. This led to the consideration of the SMM architecture where the purpose of the buffers in the central modules is to resolve contention amongst cells from different input modules and the scheduling in the bufferless first stage is reduced to a simple load balancing scheme.

This work has been partially supported by the Danish Advanced Technology Foundation (Højteknologifonden) through the research project "The Road to 100 Gigabit Ethernet".

A. V. Manolova, S. Ruepp, A. Rytlig, M. Berger, H. Wessing and L. Dittmann are with DTU-Fotonik, Technical University of Denmark, Kgs. Lyngby, Denmark; e-mail: {anva,srru,s052800,msbe,hewe,ladit}@fotonik.dtu.dk

Several Input Module (IM) connection matrix configurations for the SMM architecture are possible [11]. The Desynchronized Static Round Robin (DSRR) [8] and the simple Random scheme are dynamic control schemes, where the connection matrix changes every time-slot. Such schemes though increase the complexity of the system and introduce an undesired out-of-sequence cell delivery, which leads to increased end-to-end delay [11]. In this work we adopt a static IM connection matrix configuration, which results in a cheaper switch with reduced requirements for re-sequencing buffers (since no out-of-sequence cell delivery is observed), which lowers the end-to-end delay [11]. Under this scheme, the inputs of an IM are statically (permanently) connected to specific outputs. We use Virtual Output Queueing (VOQ) in all Central Modules (CMs) and Output Modules (OMs) (Figure 1), where each module has one Input queue per incoming link, which is logically separated in virtual output queues (one queue per output link from the module). Furthermore, simple round-robin schedulers are used for traffic forwarding in CMs and OMs.

III. BACKPRESSURE SCHEMES

Our proposed BP signals are operated between the CMs and the IMs (see Figure 1). An advantage of keeping the BP operation within the switch is that it is independent of the traffic manager. Unlike standard BP schemes, which operate on a queue-to-queue association, the BP we propose operates on a queue-to-module association, where a downstream queue controls the connectivity matrix of an upstream IM. In this way we achieve buffer overflow control in the CMs via traffic redistribution (i.e. load balancing) amongst the other CMs and not through buffering as such. A BP signal is sent when the queue length in a CM exceeds a certain threshold value. The following BP schemes are proposed:

- Coarse control:
 - When a CM detects total buffer overflow (accumulated over all Input queues), it sends a 1-bit signal to all IMs which change their connection matrix, following a round-robin principle, i.e. a connection $i \rightarrow j$ is changed to connection $i \rightarrow j + 1 \bmod N$ (N - number of outputs from the IM) (**BP_coarse**);
- Fine control: When a CM detects a buffer overflow of an Input queue, it sends a 1-bit signal to the corresponding IM, which changes its connection matrix:
 - by following a round-robin principle (**BP_fine**);
 - by redirecting the traffic from the line card with the minimum number of cells, received within the last 100 time-slots, towards the CM sending the BP signal (**BP_fine_load_deflect**);
 - each time slot following a round-robin principle for a limited duration of time (**BP_fine_load_balance**).

The control method for the BP scheme is time-based. When a BP signal is sent, a back-off timer is activated during which no other CM can send a BP signal to a given IM. This is required in order for the system to stabilize and the effects of the change in the connection matrix to become effective. For the **BP_fine_load_balance**, the back-off timer also controls the duration for which the IM changes its connection matrix.

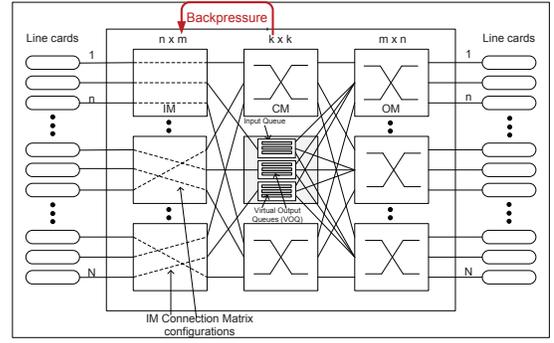


Fig. 1. Switch fabric based on 3-stage Clos network. The arrow denotes backpressure signals. Example VOQ configuration in one CM indicated.

IV. SIMULATION STUDY

We evaluate the efficiency of the proposed BP schemes using Opnet Modeler [12]. We utilize a 9x9 switch with three IMs, three CMs and three OMs. Links run at 100 Gbps speed. The generated packets have uniformly distributed lengths between 512 and 12176 bits and are segmented into 256 bit cells for forwarding through the switch. We employ Bernoulli traffic generators, i.e., the generated traffic is smoother than Poisson, which corresponds to the application scenario for a core 100G Ethernet switch. The traffic is uniformly distributed between all source/sink pairs. 100 ms of operation is simulated, resulting in amount of generated packets between $5.5 \cdot 10^6$ and $13 \cdot 10^6$ with a step of $1.5 \cdot 10^6$ for the indicated loads.

The efficiency of the schemes is evaluated in terms of the maximum needed buffer space (in number of cells) without cell loss. No upper bound for the Input queues is set, instead - the maximum required buffer space for any of the Input queues in all CMs is observed. BP is initiated when the buffer space in a CM exceeds a given threshold. The threshold value is set to be 60% lower than the maximum needed buffer space if no BP is applied, averaged over five independent simulation runs. The used back-off timer for all schemes is equal to 500 time-slots. Five independent simulation runs have been performed with different random seeds. The presented results are the average of all runs. 95% confidence intervals have been calculated. The obtained maximum needed buffer size for the different BP schemes is compared to a case where no BP is used and the IMs have static configuration. The needed control overhead (number of one-bit signals for BP control) is obtained as well.

V. RESULTS

Figure 2 illustrates the maximum amount of needed buffer space (in number of cells) for all investigated schemes. As can be expected, the Static case without BP has the worst performance especially at high loads. All BP schemes perform better than the Static case, which clearly illustrates their efficiency. In order to exactly evaluate the improvement each BP scheme can offer, Figure 3 presents the average improvement factor in % for each BP scheme compared to the Static case, based on the obtained average values from Figure 2. We can see that **BP_fine_load_balance** significantly outperforms the rest of the schemes. This is due to the specifics of the scheme's operation, where the load from all input line cards, connected

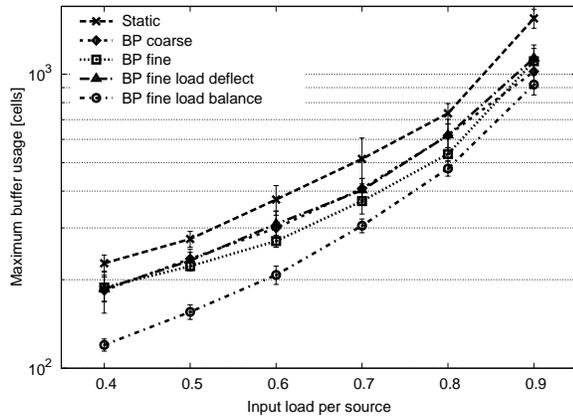


Fig. 2. Maximum buffer usage among all Input queues of all CMs.

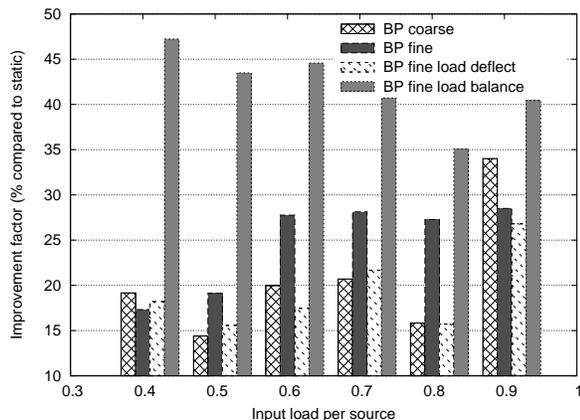


Fig. 3. Improvement of the CMs' buffer usage compared to Static scheme.

to a single IM is balanced among all CMs for a certain duration of time. The efficiency of the scheme will thus depend on the duration of the backoff period and the dynamics of the incoming traffic flow. It is interesting to observe that **BP_fine** outperforms **BP_fine_load_deflect** (though only marginally), which is due to the fact that modules operate and send BP signals independently. This lack of coordination may lead to cases where two of the IMs might target their most loaded interfaces to only one CM, effectively overloading it.

Table I shows the cost of using the BP schemes in terms of control overhead. As expected the coarse-grained BP scheme has the lowest overhead since it activates BP when the total buffer size of the CM has exceeded the threshold value (i.e. the reaction is per averaged buffer size for all input queues), whereas the fine-grained schemes activate the BP on a per input queue overflow basis. All fine-grained schemes require approximately the same amount of control overhead (100 – 150 kbps). The observed overhead does not follow a particular tendency because of the chosen threshold values. Since the threshold value is different for each tested load condition and is correlated to the performance of the static case (i.e. it is not an independent value), this leads to different conditions for activating the BP schemes for each load point. Nevertheless, considerably higher overhead for the fine-grained schemes is expected, which is supported by the results. It is important to notice that more control information

TABLE I
CONTROL OVERHEAD (IN KILO BITS) PER BP FOR ALL TESTED LOADS.

Input load per node	BP_coarse	BP_fine	BP_fine load_deflect	BP_fine load_balance
0.4	30.5 ± 0.2	121.6 ± 0.3	124.4 ± 0.3	110.2 ± 0.6
0.5	27.3 ± 0.3	122.2 ± 0.7	126.7 ± 0.5	108.3 ± 0.3
0.6	35.1 ± 0.4	140.1 ± 0.4	147.4 ± 0.8	121 ± 0.7
0.7	29.7 ± 0.6	121.7 ± 1.1	129.9 ± 0.6	107 ± 0.9
0.8	23.1 ± 0.5	97.8 ± 0.5	107.8 ± 1	90.3 ± 0.6
0.9	48 ± 0.3	134.1 ± 2.2	158.6 ± 1.4	134.6 ± 1.5

does not guarantee better performance. For example, at 0.9 input load the **BP_coarse** scheme has a higher improvement factor than the **BP_fine_load_deflect** (about 10%) and still generates about 3 times lower overhead. On the other hand, at the same load **BP_coarse** has lower improvement factor than **BP_fine_load_balance** (about 10%), but has 2.8 times lower overhead. Since the statistical aspects of the results are not taken into account (only the average values), none of the BP schemes can be conclusively declared to be the best, especially at high loads.

The standard trade-off between efficiency and cost is evident. The overall evaluation of the efficiency of the schemes includes not only the achieved performance but also their complexity and the application scenario. The BP schemes need to be properly tuned to the given traffic conditions for achieving the best possible performance.

VI. CONCLUSION

In this paper, we propose and evaluate the efficiency of four novel backpressure flow control schemes for SMM Clos-based Terabit switches. Unlike existing schemes our proposals operate on a queue-to-module principle. The main focus of the schemes is buffer-space optimization which is paramount for improved system integration, scalability and cost-efficiency. Simulation results show that at the cost of about 110kbps control overhead we can achieve between 35% to 47% reduction in the needed maximum buffer space.

REFERENCES

- [1] C. Hermsmeyer et al., "Towards 100G packet processing: Challenges and technologies," *Bell Labs Technical Journal*, vol. 14, no. 2, 2009.
- [2] C. Clos, "A study of non-blocking switching networks," *Bell Systems Technical Journal*, pp. 406–424, 1953.
- [3] T. Kanazawa et al., "Input and Output Queueing Packet Switch with Backpressure Mode for Loss Sensitive Packets in Threshold Scheme," in *IEEE PACRIM*, 1997, pp. 527–530.
- [4] H. J. Chao, "Flow Control in a Multi-Plane Multi-Stage Buffered Packet Switch," in *High Performance Switching and Routing, HPSR*, 2007.
- [5] F. Chiussi et al., "Backpressure in shared-memory-based ATM switches under multiplexed bursty sources," in *IEEE INFOCOM*, 1996.
- [6] H. T. Kung and R. Morris, "Credit-based flow control for atm networks," *IEEE Network*, vol. 9, pp. 40–48, March/April 1995.
- [7] R. Schoenen and A. Dahlhoff, "Closed Loop Credit-Based Flow control with Internal Backpressure in Input and Output Queued Switches," in *HPSR*, 2000, pp. 195–203.
- [8] X. Li, Z. Zhou, and M. Hamdi, "Space-Memory-Memory Architecture for CLOS-network Packet Switches," in *IEEE ICC*, 2005.
- [9] R. Rojas-Cessa and H. Chao, "Maximum Weight Matching Dispatching Scheme in Buffered Clos-Network Packet Switches," in *IEEE ICC*, 2004.
- [10] J. Kleban and S. Piotrowski, "Performance Evaluation of Selected Packet Dispatching Schemes for the CBC Switches," in *Communications Letters, IEEE*, 2009.
- [11] S. Ruepp et al., "Performance Evaluation of 100 Gigabit Ethernet Switches under Bursty Traffic," in *ONDM*, 2011.
- [12] OPNET Technologies, Inc., <http://www.opnet.com>.