On contention resolution in the data vortex optical interconnection network

Assaf Shacham* **and Keren Bergman**

Department of Electrical Engineering, Columbia University, New York, New York 10027, USA
Corresponding author: assaf@ee.columbia.edu

Received February 17, 2007; revised April 24, 2007; accepted April 24, 2007; published May 29, 2007 (Doc. ID 80158)

Alternative contention resolution techniques are studied in the data vortex interconnection network, namely, the insertion of fiber-delay-line (FDL) buffers into the switching nodes. The performance of each technique is evaluated according to relevant performance metrics: acceptance rate, mean latency, and latency variance. A detailed discussion concludes that while traditional data vortex networks perform better in terms of throughput, FDL-based switching nodes have a favorable impact in reducing the system latency. © 2007 Optical Society of America

OCIS codes: 060.0060, 060.4250.

1. Introduction

Modern high-performance computing (HPC) systems are increasingly facing a communications bottleneck. Traditional electronic interconnection networks [1] reach fundamental physical limitations set by transmission-line losses, which lead to bandwidth limits and increased power consumption [2]. It has recently been suggested by several research efforts that optical interconnection networks (OIN) have the potential to mitigate or solve most of these problems by capitalizing on the inherent large bandwidth, low loss, and bit-rate transparency of optical transmission and switching systems [3–7].

While its advantages are well known, the employment of optical technology in the design of high-performance interconnection networks requires careful consideration of several factors that do not exist in the design of electronic systems. These factors include the lack of adequate buffering capability, comparable with registers and random access memory (RAM), and the limited amount of processing that can be done all optically. Additionally, optics presents challenges such as signal impairments, noise, and nonlinearities, which are typically negligible in complementary metal-oxide semiconductor (CMOS)-based systems [8].

Data vortex (DV) is an OIN architecture, designed to exploit the advantages of optical switching systems while properly addressing the limitations of optical technology [9–11]. It capitalizes on wavelength division multiplexing (WDM) to offer a large transmission bandwidth and to simplify switching-node design. Its unique, distributed topology, comprising a large number of 2×2 wideband switching nodes, facilitates contention resolution in the space domain and scalability to a large number of ports. The limits on optical processing are addressed in the switching nodes [12,13] by electronically decoding a part of the optically encoded address of the packets and routing the packets using electronically switched semiconductor optical amplifier (SOA) gates. A transparent end-to-end photonic path is thus formed, enabling exploitation of the immense transmission bandwidth offered by WDM and alleviating the need for costly O/E/O conversions. The distributed structure and the high-speed wideband switching of the SOAs suggest that a DV interconnection network can scale both in terms of bandwidth and port count while maintaining nearly time-of-flight latencies [9–11].

One of the main concerns in the design of OINs is the contention resolution mechanism. A state of contention occurs when a network resource (e.g., an output port or an internal path) is requested by more packets than it can simultaneously serve. A simple and typical example for contention is a case where two packets are addressed

1536-5379/07/060777-12/$15.00

to the same output port at the same timeslot in a slotted, packet-switched router. In electronic switching, fabrics contentions are trivially resolved by using RAM to buffer the packets and then to read them sequentially. In optical networks, since photonic memories are unavailable, the problem becomes more challenging, and other means of resolving contentions have to be considered.

Contention resolution methods in OIN can be categorized according to the following taxonomy, suggested in [14]:

1. *Wavelength domain*: Contending packets are transmitted on the same fiber at the same time encoded on different wavelengths.
2. *Time Domain*: Some of the contending packets are delayed and are then transmitted sequentially. This is typically implemented using a fiber delay line (FDL).
3. *Space Domain*: Some of the contending packets are deflected to an undesired output port, from which an alternative path to the final destination must exist.

The contention resolution approach used in a DV interconnection network is a combination of two of the above-mentioned methods. In the switching-node scale, contentions are resolved in the space domain, and packets are deflected to an undesired port of the switching node when the requested port is not available. The deflected packets then reach their destination, perhaps at a later time, by traveling on a different path, which always exists in the network. The space domain node-scale technique translates to a time-domain contention resolution in the system scale as contending packets reach their final destination at different times. This property is termed virtual buffering as packets are virtually buffered in internal paths until their destination becomes available. More on the DV architecture can be found in Section 2.

In this paper, we evaluate the effectiveness of the virtual buffering mechanism by comparing it with an alternative contention resolution method: time-domain contention resolution at the node scale. Two approaches for the insertion of FDL-based recirculating buffers into the switching nodes are investigated. The performance effects are studied and compared through simulations, and a concluding discussion is provided.

The rest of this paper is organized as follows: a brief overview of the DV architecture is given in Section 2 along with the performance metrics by which the comparison is made. In Section 3, we describe the two design approaches of the FDL-based nodes compared with the original DV node. The performance study and its results are detailed in Section 4. Finally, Section 5 provides a discussion and concluding remarks.

2. Architecture Overview

The DV topology is a multistage interconnection network (MIN), composed of optical bufferless 2×2 switching nodes. The network topology is organized such that contentions may occur only when packets attempt to move ahead between stages. Since the contentions cannot be resolved in the bufferless switching nodes, contending packets are delayed in the same stage until they can progress to the subsequent one. To facilitate the optical buffering of packets within stages, the traditional MIN structure is modified by the addition of switching nodes at every stage and optical fibers connecting the added switching nodes. Packets are deflected between switching nodes in the same stage when their progression path is occupied. The packets are therefore buffered in the fibers which connect nodes within stages (i.e., deflection fibers). This technique is termed virtual buffering [9].

The added switching nodes are organized in circles to guarantee the availability of deflection paths, so the stages become 3D cylinders (Fig. 1). A data vortex interconnection network can therefore be viewed as a set of concentric cylinders (or stages) defined by three structural parameters: H, the height of the network, C, the number of stages ($=\log_2 H+1$), and A (the angle parameter), the number of switching nodes along the circumference of each stage. Figure 1 illustrates an exemplary DV network. The switching nodes at the outer cylinder are used as network input ports, for packet injection, and the ones at the inner cylinder are used for packet ejection (network output ports).

A given DV implementation is structurally defined by the 3-tuple (A,H,C). Network configurations can also vary in the fraction of populated input–output (I/O) ports to the total number of possible of nodes that can serve as ports. In general, all nodes

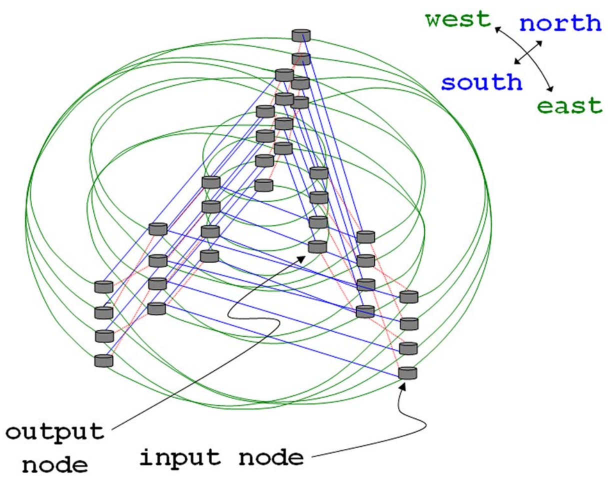

Fig. 1. Example DV topology with the dimensions $C=3$, $H=4$, $A=3$. This network has $N=36$ switching nodes and can have $N_T=4$ or 12 I/O ports, in asymmetric or symmetric injection, respectively.

on the input cylinder (i.e., $A \times H$ switching nodes) can be used for packet injection. It may be desired, in some cases, to use asymmetric injection where a fraction of these switching nodes are used as ports. A' is defined as the number of active angles through which injection–ejection is performed in the case of asymmetric injection [9,15]. The special case where $A'=A$ is actually symmetric injection, where all the input cylinders are used as injection ports. The number of switching nodes in a DV system is $N \equiv A \times H \times C$, and the number of I/O ports is therefore given as $N_T \equiv A' \times H$.

Packets are injected into the network in a time-slotted manner through the switching nodes in the input stage (outer cylinder) and travel in a circular motion within stages and inward, one node hop per timeslot. At every switching node, a single address bit is inspected and compared with a reference value configured in the node. The interconnection patterns within stages are designed such that packets travel through different height levels while traversing a stage and progress to the next stage when they are at an appropriate height (i.e., the appropriate address bit is equal to the reference value). When progression paths are blocked the packets remain in the same stage until another progression path is found.

The configured reference values and the inspected address bit are dependent on the switching-node cylinder spatial coordinates: the cylinder (c), height (h), and angle (a). By appropriately selecting the reference values, packets are routed in a binary-tree fashion to a height that corresponds to their optically encoded address. Destination tag routing, where a single height bit is used to encode all the information required for the routing in each stage [1], is employed to simplify the address decoding and the switching-node design [12]. For a more detailed presentation of the DV architecture the reader is referred to [11].

The switching nodes [12] are connected in a manner facilitating both progression of packets between stages and deflection within stages. One of the input ports is connected to a switching node in the same stage (West port) and the other input port is connected to a switching node in a previous stage (North port). Accordingly, the output ports are connected such that one leads to a switching node at the same stage and one leads to a switching node at the subsequent stage (East and South ports, respectively). So at every switching node the South port is the progression port and the East port is the deflection port to which packets are routed in case the South port is unavailable, or if the node's reference value does not match the packet's address.

The DV switching nodes [Fig. 2(a)] are 1×2 switches, implemented using two SOAs as switching elements. A passive network of couplers and filters, along with photodetectors and high-speed electronics are used to decode the necessary bits of the optical header, process them to generate a routing decision, and drive the SOAs with electrical current accordingly. The exact implementation of the switching node is reported in [12,13] and is beyond the scope of this paper.

To ensure that only a single packet is received in each switching node at a timeslot (i.e., avoid collisions) the nodes can exchange deflection signals. For example (Fig. 3),

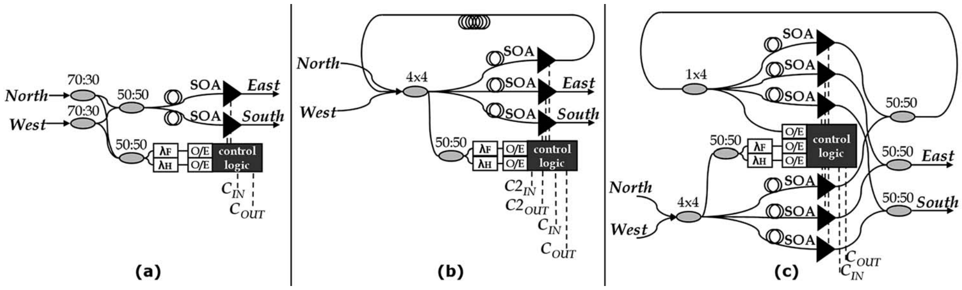

Fig. 2. Switching node designs. (a) Original DV switching nodes, (b) the blocking switch node, which can handle a single packet in a timeslot, using three SOAs, (c) the nonblocking switch node, which can handle one buffered packet and one incoming packet in each timeslot, and requires six SOAs.

to avoid collisions in node A, an electronic control cable connects node B and node C. Whenever node B, which is connected to node A's West port transmits a packet to node A (i.e., East) it also emits a signal to node C, to mark node A's North port as blocked. Upon receiving the signal, node C regards its South port as blocked, and any packet which is received in that timeslot is sent to the East port. Once a simple timing requirement is met, this structure can be used across the network to prevent packet collisions in all nodes [11,12].

The cylindrical structure and the deflection signaling mechanism in the DV architecture guarantee that packets are not lost inside the system, and congestion is manifested as blocked progression paths and as backpressure on the input ports. When the network becomes congested, packets may be delayed within stages. When the pressure propagates to the input stage (the outer cylinder) and it becomes congested the injection of new packets is blocked, and they are either dropped at the inputs or required to reattempt injection using injection control modules [16]. The acceptance rate is therefore defined as the ratio of the successfully injected packets (packets not blocked by internal traffic) to the number of total injection attempts.

The packet latency in DV interconnection networks, measured in node hops, is nondeterministic and relies to a large extent on the rate of contentions in the network and the congestion level. The more frequent the contentions, the longer it takes for packets to find available progression paths, so additional hops are taken in each stage. As the load increases, the mean packet latency grows larger, and the latency distribution becomes wider (i.e., the latency variance increases) [9,10].

The acceptance rate, the mean latency, and the latency variance are used in this study to evaluate the performance of DV interconnection networks. In an HPC interconnection network, it is desirable to keep the latency as low as possible, for a minimal memory access time. Limiting the latency variance is considered necessary to allow for efficient programming and predictable performance [17]. The latency and latency variance should therefore be minimized when designing a DV interconnection network. Since every denied packet has to be reconstructed and therefore consumes expensive transmitter time and power, a high acceptance rate is also required. In Section 3, we evaluate several methods to alter the DV architecture by the incorporation of alternative means of contention resolution at the node level.

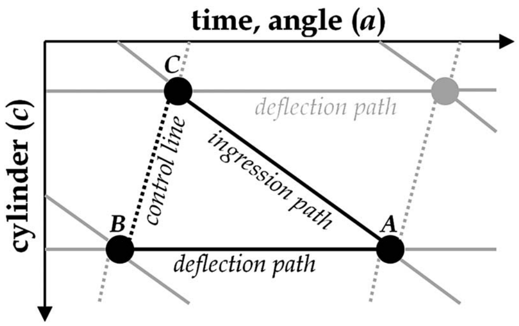

Fig. 3. To avoid packet collision at node A, an electronic signal is transmitted by node B on the control line, to node C [11].

3. Alternative Contention Resolution Mechanism

While they are simpler to implement optically, systems that employ deflection routing suffer from the disadvantage of a nondeterministic and fairly unpredictable behavior [1]. The positive feedback process where congestion is followed by packet deflections, which in turn cause increased congestion, may lead to a low saturation load in deflection-based networks. In the DV, this effect is manifested as increased backpressure generated when packets spend more hops in each stage, having difficulties finding available progression paths, and may lead to a lower acceptance rate and increased latency. To mitigate this effect, we investigate an alternative method of resolving contentions in the switching nodes and thus limit the use of deflection routing.

In the DV switching node [see Fig. 2(a)], contentions are resolved solely by deflection: When the South port is blocked packets are deflected to the East port. We suggest an alternative means of resolving the contention. A recirculating FDL buffer is added to each switching node. When the South port is blocked, as indicated by the control signal, a packet whose desired output port is South is sent to a FDL and is handled by the switching node later. The FDL is used to resolve contentions in the time domain.

Two approaches are considered: In the first approach [18] [Fig. 2(b)] a node may handle only a single packet per timeslot, similarly to the original DV. As seen in the block diagram in Fig. 2(b) this approach requires a 1 × 3 SOA-based switch to route the packet from any of the three inputs (West, North, or FDL) to one of three outputs (East, South, or FDL). To guarantee that only a single packet is received in each timeslot, a new set of control cables have to be added, to enable the transmission of intrastage deflection signals. The new signals are required because whereas in the original DV the West port of the packet is always available for packet reception, in the new setting a packet in the FDL may block the West port. The intrastage signal is sent to the switching node, which drives the West port, letting it know that its East output port is blocked. To ensure that the intrastage signal reaches its destination node in time to block a packet, the FDL must be made two slots long. This approach will be referred in this paper as the blocking switching node.

The drawbacks of the blocking switching node approach are evident: it requires additional hardware (the intrastage cabling), introduces additional blocking to the network, and each FDL traversal is two slots long. To overcome these shortcomings, we suggest a second approach [19], based on a 2 × 3 SOA switch [Fig. 2(c)]. This switching node can handle two packets at the same time: one packet, which is received from either the West or the North, input ports and one packet received from the FDL. Although this node still has internal blocking between the West and North ports (like the original DV node), the FDL does not block nor can it be it blocked by any on the inputs. The node is therefore termed the nonblocking switching node.

Since the nonblocking switching node does not introduce any blocking to the West input port, it does not require the introduction of intrastage signaling. The FDL length can therefore be set at one timeslot reducing the latency penalty incurred by output port blocking. On the other hand, the nonblocking switching node naturally requires more hardware to implement the node, e.g., six SOAs are required to implement the 2 × 3 switch. Table 1 summarizes the differences between the nodes.

In Table 1, we introduce a cost factor as a means of assessing the cost and complexity of each node. Several cost models have been suggested for optical interconnection networks (see, for example, [20]). We use a simple model where the cost and complex-

Table 1. Switching Nodes Comparison

	Original DV	FDL Blocking	FDL Nonblocking
Number of SOAs	2	3	6
Number of control I/O ports	1	2	1
FDL length [slots]	N/A	2	1
Number of photodetectors	2	2	3
Number of 2 × 2 couplers	4	5	11
Normalized cost	1	1.5	3

ity of an optical switching node scales linearly with the number of switching elements (i.e., SOAs). This model is based on the assumption that the optical switch is the main cost component of the switching node and that the rest of the hardware (passive optics, electro-optic components, and electronic circuits) also grows when a larger switch is used. The cost factors in Table 1 are calculated according to this model and are normalized such that the original DV nodes have a cost factor of 1.0.

4. Comparative Performance Study

To evaluate the performance of a DV with the suggested improvements, and compare them with the original architecture, a custom cycle-accurate simulation program was developed. The program, written in the C++ programming language, provides extensive flexibility to model structural modifications within the switching nodes as well as topological variations. The simulation program also enables the user to choose traffic patterns and to set the injection load for an individual simulation run or a set of runs.

The simulations conducted in this study use synthetic Bernoulli uniform random traffic, where each input port injects traffic at a constant rate of $p \in [0,1]$ packets per slot (i.e., the injection rate) and each packet's destination address is selected with equal probability among all possible output ports. While the actual traffic patterns in HPC systems are rarely uniform, uniform traffic serves well in revealing the raw capacity of a network and providing an upper bound on the network performance.

All the simulation runs start with a warm-up period of 400 slots where packets are injected but no measurement is taken. Following the warm-up period, packets injection and latency statistics are recorded during a 9200 slot measurement period. For each measurement point, ten such runs are conducted and ensemble averages are calculated. This measurement technique follows the procedure recommended in [1], has been thoroughly tested, and has constantly provided reliable results.

In the following simulations, we study the performance effect of each of the new node design approaches (the blocking node and the nonblocking node) and compare them with the DV virtual buffering mechanism. As a basis for comparison, we choose the following design point: The DV structural parameters are $H=16$, $A=5$, $C=5$, and both symmetric and asymmetric injection are studied. The systems therefore have $N_T=80$ input and 80 output ports (16×5). When simulating the FDL-based nodes we replace the switching nodes in the original topology with one of the proposed designs. In the case of the blocking nodes, the system is simulated with the necessary intrastage deflection signals.

As discussed above, the metrics according to which each configuration is evaluated are the following:

1. *Acceptance rate*: The ratio of successful injections to attempted injections. It is assumed that packets that fail injection are lost.
2. *Mean Latency*: In node hops from the input node to the output node, averaged across all injected packets.
3. *99.9 Percentile Latency*: As a measure of the latency variance.

4.A. Optimization of Buffer Hop Limits

When designing the FDL-based switching nodes, it is of interest to study the effects of limiting to the number of times a packet may traverse the loop in each node. When such a hop limit is set, in cases when it is exceeded and the South output port is still blocked, the packet is deflected to the East port. A switching node without a FDL traversal limit will buffer the packet until the South port blocking is released. A switching node whose FDL traversal limit is 0 is equivalent to the original DV switching node, which uses only deflection routing. Limiting the FDL traversals can therefore be interpreted as a moving scale balancing the newly introduced time-domain contention resolution and the original DV space-domain contention resolution. Both FDL-based nodes are simulated and compared to the original DV with symmetric injection at 0.3 and 0.7 injection rates with varying values of FDL hop limits. The results are given in Fig. 4.

An inspection of the results reveals that most of the performance gain is achieved even when a single buffer hop is allowed. For the network based on the nonblocking switching node the acceptance rate increases by 40% to 50% and then rises very

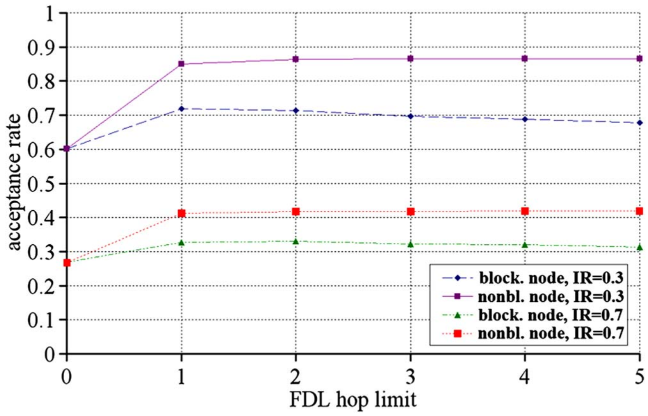

Fig. 4. Effect of FDL hop limit on acceptance rate at injection rates of 0.3 and 0.7.

slowly when the hop limit is increased. In the case of the blocking node, the maximum increase of 20% is reached with a one-hop limit, and the acceptance rate then decreases slowly.

The difference in behavior can be explained by the different designs of switching nodes. In the network based on the blocking node as packets are allowed to spend more time in the node, they block other packets from progression, and this effect quickly translates to packet dropping at the network inputs and to a decrease in the acceptance rate. The nonblocking switching node acts as an additional buffering space and does use deflection to resolve contention when the buffer is full. Blocked packets typically spend 1–2 timeslots in the buffer and then either progress or are pushed East by other packets if the South port remains blocked. Only a minuscule fraction of the packets actually use the buffer to the full hop limit, and their effect on the performance is negligible. The performance is therefore nearly constant with a hop limit of 2 or greater.

The conclusion drawn from this study is that in networks based on the nonblocking nodes the network dynamics effectively limit the FDL traversal to 1–2 timeslots, so there is no need to externally enforce the limit. For the blocking node, conversely, superfluous blocking is introduced when an FDL limit is not enforced, so the limit should be set at one traversal.

4.B. Acceptance Rate and Throughput

The improvement in the acceptance rate resulting from the intranode buffering capabilities of the new switching nodes is unmistakable. This is especially true when nonblocking nodes are used. However, this improvement comes at the cost of adding complexity and hardware to the switching nodes and, in the case of the blocking nodes, also to the network. It will therefore be of interest to compare the performance gain offered by the FDL-based switching node to an alternative increase in the buffering capacity of the network, increasing the virtual buffering capacity. This increase is achieved by the insertion of additional buffering nodes as additional angles in the DV cylinders and switching to asymmetric injection mode.

In asymmetric injection only a fraction of the outer-cylinder nodes serve as injection ports, and a fraction of the inner-cylinder nodes serve as output ports. In the simulation configurations, we keep the number of input ports fixed ($A'=5$, $N_T=80$) and control the network angle parameter ($A=10$ and $A=15$) to vary the asymmetric injection fraction. The additional nodes do not serve as I/O ports, but merely provide additional buffering space and progression paths for packets, which are unable to make progress to subsequent stages. An overhead view of the network with and without additional nodes can be seen in Fig. 5.

The simulated configurations are chosen such that a comparison can be done between configurations of equal costs. Table 2 is a summary of the simulated DV configurations. For each configuration the table shows the topological parameters (A, H, and C), the number of injection angles (A'), the total number of nodes (N), the total number of ports (N_T), and the calculated cost factor.

The five data vortex configurations detailed in Table 2 are simulated under the conditions described in previously in this section: Bernoulli uniform traffic under varying

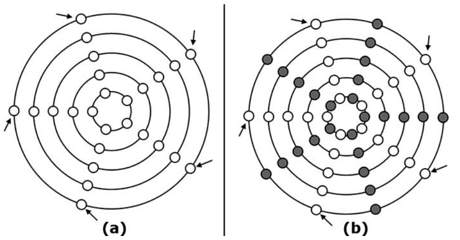

Fig. 5. Overhead view of a five-stage DV network (progression fibers are omitted for clarity). (a) Symmetric injection ($A'=5$, $A=5$), (b) asymmetric injection ($A'=5$, $A=10$).

Table 2. Simulated Configurations

Parameter	Original	Original × 2	Original × 3	Blocking Node	Nonblocking Node
Node Type	Original	Original	Original	Blocking	Nonblocking
A	5	10	15	5	5
C	5	5	5	5	5
H	16	16	16	16	16
A' 5	5	5	5	5	
N	400	800	1200	400	400
N_T	80	80	80	80	80
Normalized cost	400	800	1200	600	1200

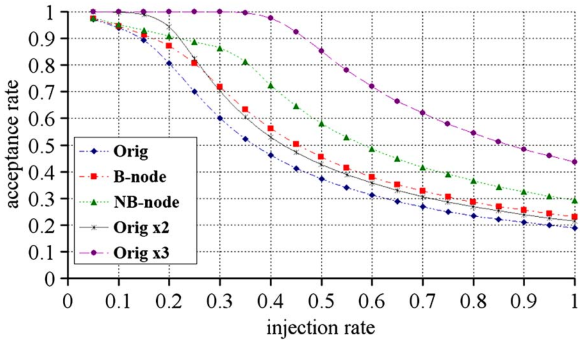

Fig. 6. Acceptance rate versus injection rate of the five considered DV configurations.

injection rates. Figure 6 shows the acceptance rate versus the injection rate curves for the five configurations. Because every failed injection has to be reattempted, and consumes additional I/O time, it is desired to achieve an acceptance rate as high as possible.

An inspection of Fig. 6 leads to several interesting conclusions: First, as observed in previous work [9,15], the acceptance rate function is monotonically decreasing with the injection rate. Second, the FDL-based configurations improve the acceptance rate only at medium to high injection rates while asymmetric injection improves the acceptance rate at low injection rates, raising it to nearly 1.0 when the injection rate is low (e.g., smaller than 0.3 in the Original ×3 configuration).

A different view, yielding more information, is provided by plotting the throughput versus the injection rate, as appears in Fig. 7. The throughput is the steady-state number of packets injected into the system per unit time, normalized such that a throughput of 1.0 means that every port successfully injects a packet at every slot. In Fig. 7, the saturation value of every configuration can easily be seen. The saturation value can be helpful in determining the injection rate that should serve as an upper boundary for the healthy operation of a given configuration. Operating above saturation will not yield additional throughput but can result in increased latency and latency variance as will be shown in Subsection 4.C.

4.C. Latency

The latency in the DV is measured in node hops and is considered an important performance metric. The transmission latency in HPC systems translates to memory

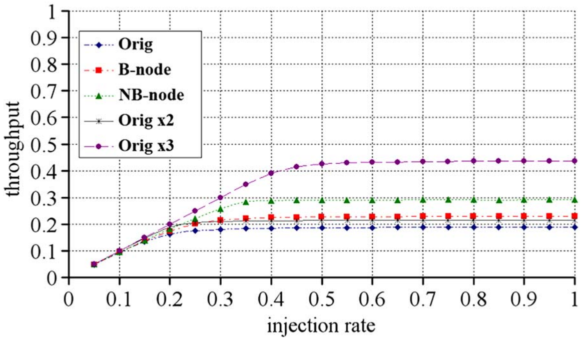

Fig. 7. Throughput versus injection rate of the five considered DV configurations.

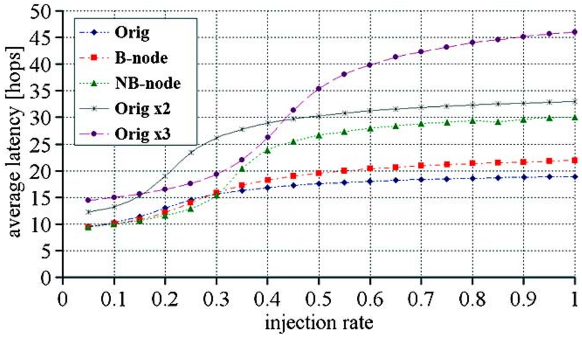

Fig. 8. Average latency versus injection rate of the five considered data vortex configurations.

access latency, which is crucial to application execution times [17]. In optical interconnection networks, specifically, the number of node hops also affects the signal integrity of the optical packet, which passes through several SOAs where amplified spontaneous emission (ASE) accumulates [21]. To minimize the noise buildup and other effects, minimal hop latency is also desired. To facilitate the design of network interfaces, it is desired that all packets undergo similar length paths and appear with similar levels of noise and distortion at the outputs, so a small latency variance is sought.

The average hop latency of the five considered configurations as a function of the injection rate is given in Fig. 8. When inspecting Fig. 8 in light of the saturation loads obtained from Fig. 7, it is evident that when the network is saturated, and increased injection does not yield additional throughput, the latency still rises. For example, a network based on the nonblocking node reaches saturation at an injection rate of 0.35 (see Fig. 7) when the average latency is 20 hops (Fig. 8). Increasing the injection rate does not contribute to throughput but increases the average latency as high as 30 hops. The conclusion drawn is that a network should be operated below the saturation load.

Figure 8 can be confusing because some of the configurations have higher average latencies at the same load, and this may be interpreted as performance degradation. This is not necessarily the case since some of the networks, which have higher acceptance rates and throughputs, actually route more packets. Figure 9 therefore presents a more useful view of the average latency versus throughout.

Generally, a latency versus throughout plot can be ambiguous in interconnection networks that have a concave throughput versus injection rate curve. In these networks, the same throughput may be attained in two different network states, producing different latencies and ambiguous results [1]. Figure 8 shows that this is not the case in any of the DV networks, and the throughput function is monotonically increasing with the injection rate. A latency versus throughput plot is therefore valid.

Figure 9 provides better insight on the latency dynamics of the considered configurations. It is clear that while the asymmetric injection configurations extended-angle configurations provide higher throughput, the price is an increased latency even when compared with the original DV. This can be explained by the fact that the extended-angle networks are substantially larger so a path that has to be taken to reach the

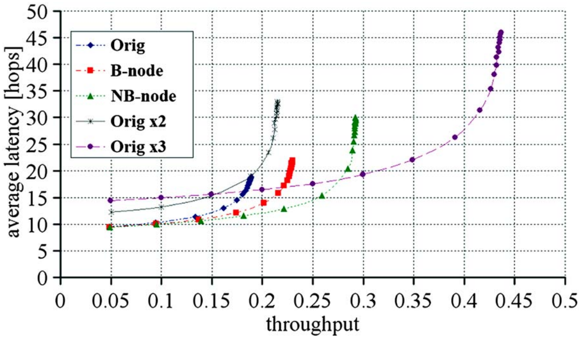

Fig. 9. Average latency versus throughput of the five considered DV configurations.

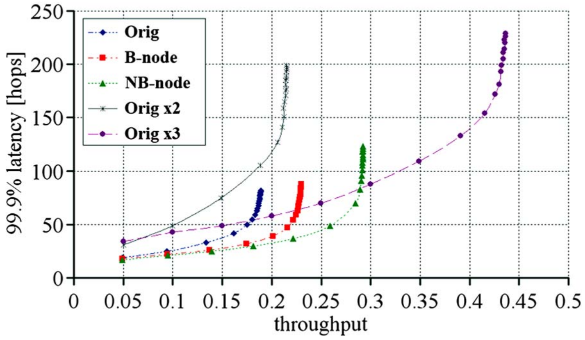

Fig. 10. Throughput versus 99.9% latency of the five considered DV configurations.

desired angle is longer. The networks composed of FDL-based nodes have lower subsaturation latency and perform better than the DV network with symmetric and asymmetric injections. As expected, the nonblocking node provides improved performance when compared with the blocking node.

Another interesting observation is the clear appearance of the saturation throughput of each configuration. It can be seen that the symmetric injection systems have lower latency and lower saturation, while the asymmetric injection systems do have a higher throughput at the expense of high subsaturation latency.

Finally, the latency variance is studied by plotting the 99.9% latency against throughput in Fig. 10. The resemblance between the average packet latency curve and that of the latency of the 99.9% packet is evident. The need to operate well below saturation is even more obvious as the latency may rise quickly to hundreds of hops when the network is improperly loaded. On the other hand, the latency for an 80-port network, with the nonblocking switching node, can be limited to less than 50 hops when the network operates below saturation (an injection rate of 0.3).

As described in the beginning of this section, the latency measurement is also important for physical-layer reasons. Some optical effects in SOAs, such as the accumulation of ASE noise and polarization dependent gain (PDG), can lead to signal degradation. Recent studies on the propagation of optical packets through SOA-based multistage interconnection networks show that while the hop count should certainly be minimized, there is still room to construct large-scale networks. For example, in [21] wavelength-striped packets with eight payload wavelengths have been shown to have a bit error rate (BER) of 10^{-9} or better after 58 SOA hops. The power penalty of an SOA-based switching node was measured in [22] to be lower than 0.8 dB. Finally, the use of polarization maintaining fiber is suggested in [23] as a means of mitigating the cumulative PDG effect.

5. Conclusion

Several methods of improving the original data vortex (DV) architecture are considered in this study. Whereas the original DV switching nodes offer only space-domain contention resolution, a new node design is suggested, where a combined time–space-domain contention resolution approach is used. The performances of several configurations are studied.

The original DV topology offers several methods to increase the network's throughput, mainly by increasing the network dimensions (i.e., the number of switching nodes). Two of these methods, namely, varying the asymmetric injection-ratio and injection-rate limiting of the input ports, are studied in [24]. The impact of asymmetric injection on performance is also studied in [9,15]. These methods, in general, are all based on choosing an operating point on the latency versus load curve, such that a lower load offers improved acceptance rate and latency. These approaches have two main drawbacks: First, as they simply move on a fixed curve, they cannot offer a breakthrough in terms of performance per unit cost; second, as networks grow larger, their zero-load latency is also increased. A latency reduction, desired both for application-driven reasons and physical-layer-based reasons, cannot be attained using these approaches.

Replacing the bufferless switching nodes with switching nodes that contain FDL buffers can break this equation. This is a method that can be used to increase the networks' buffering capacity without contributing to the zero-load latency. While FDL-based DV networks can be more expensive in terms of throughput per unit cost, they can offer a solution for latency-sensitive applications where cost may not necessarily be the most important concern.

Two main conclusions should be taken from this study. First, a well-known fact is strengthened and demonstrated: It is evident that any DV network should operate well below saturation, and such an injection policy should be strictly enforced. While technically the injection rate can be increased above the saturation load, doing so does not contribute to throughput and has a detrimental effect on latency.

Second, when performance improvement methods are considered, a network designer should define the main goal—throughput or latency. When the main target is throughput, traditional DV networks operated under asymmetric injection offer a higher throughput per unit cost. If, on the other hand, the main design goal is low latency, then a lightly loaded DV using the nonblocking FDL-based nodes, is the preferred solution. While this solution may be more expensive in terms of cost per unit throughput, it enables the construction of low latency interconnection networks, which may not be achievable using more traditional means.

The authors acknowledge the support of the U.S. Department of Defense under subcontract B-12-664.

References

1. W. J. Dally and B. Towles, *Principles and Practices of Interconnection Networks* (Morgan Kaufmann, 2004).
2. D. A. B. Miller, "Rationale and challenges for optical interconnects to electronic chips," Proc. IEEE **88**, 728–749 (2000).
3. R. Luijten, C. Minkenberg, B. R. Hemenway, M. Sauer, and R. Grzybowski, "Viable optoelectronic HPC interconnect fabrics," in *Proceedings of ACM/IEEE Conference on Supercomputing* (SC|05) (IEEE Computer Society, 2005), p. 18.
4. A. Shacham and K. Bergman, "Building ultra low latency interconnection networks using photonic integration," IEEE Micro (to be published).
5. T. Lin, K. A. Williams, R. V. Penty, I. H. White, M. Glick, and D. McAuley, "Performance and scalability of a single-stage SOA switch for 10×10 Gb/s wavelength striped packet routing," IEEE Photon. Technol. Lett. **18**, 691–693 (2006).
6. R. D. Chamberlain, M. A. Franklin, and C. S. Baw, "Gemini: an optical interconnection network for parallel processing," IEEE Trans. Parallel Distrib. Syst. **13**, 1038–1055 (2002).
7. A. K. Kodi and A. Louri, "Design of a high-speed optical interconnect for scalable shared-memory multiprocessors," IEEE Micro **25**, 41–49 (2005).
8. G. P. Agrawal, *Fiber-Optic Communication Systems* (Wiley, 2002).
9. Q. Yang and K. Bergman, "WDM packet routing for high-capacity data networks," J. Lightwave Technol. **19**, 1420–1426 (2001).
10. Q. Yang and K. Bergman, "Performances of the data vortex switch architecture under nonuniform and bursty traffic," J. Lightwave Technol. **20**, 1242–1247 (2001).
11. A. Shacham, B. A. Small, O. Liboiron-Ladouceur, and K. Bergman, "A fully implemented 12×12 data vortex optical packet switching interconnection network," J. Lightwave Technol. **23**, 3066–3075 (2005).
12. B. A. Small, A. Shacham, and K. Bergman, "Ultra-low latency optical packet switching node," IEEE Photon. Technol. Lett. **17**, 1564–1566 (2005).
13. A. Shacham, B. A. Small, O. Liboiron-Ladouceur, J. P. Mack, and K. Bergman, "An ultra-low latency routing node for optical packet interconnection networks," in *Proceedings of IEEE/LEOS 17th Annual Meeting* (IEEE, 2004), pp. 565–566, paper WM2.

14. S. Yao, B. Mukherjee, S. J. B. Yoo, and S. Dixit, “A unified study of contention-resolution schemes in optical packet-switched networks,” J. Lightwave Technol. **21**, 672–683 (2003).
15. C. Hawkins and D. S. Wills, “Impact of number of angles on the performance of the data vortex optical interconnection network,” J. Lightwave Technol. **24**, 3288–3294 (2006).
16. A. Shacham, B. A. Small, and K. Bergman, “A wideband photonic packet injection control module for optical packet switching routers,” IEEE Photon. Technol. Lett. **17**, 2778–2780 (2005).
17. D. Dai and D. K. Panda, “How much does network contention affect distributed shared memory performance?” in *Proceedings of International Conference on Parallel Processing*, Bloomington, Ill., Aug. 1997, pp. 454–461.
18. A. Shacham and K. Bergman, “An FDL-based photonic switching node for a data vortex optical packet switched interconnection network,” in *Proceedings of European Conference on Optical Communications (ECOC 2006)*, Sept. 2006, paper We3.P.138.
19. A. Shacham and K. Bergman, “An enhanced buffered switching node for a data vortex interconnection network,” in *Proceedings of IEEE/LEOS 19th Annual Meeting* (IEEE, 2006), paper WW2.
20. B. A. Small and K. Bergman, “Optimization of multiple-stage optical interconnection networks,” IEEE Photon. Technol. Lett. **18**, 238–240 (2006).
21. O. Liboiron-Ladouceur, B. A. Small, and K. Bergman, “Physical layer scalability of WDM optical packet interconnection networks,” J. Lightwave Technol. **24**, 262–270 (2006).
22. B. A. Small, T. Kato, and K. Bergman, “Dynamic power considerations in a complete 12 × 12 optical packet switching fabric,” IEEE Photon. Technol. Lett. **17**, 2472–2474 (2005).
23. O. Liboiron-Ladouceur, K. Bergman, M. Boroditsky, and M. Brodsky, “Polarization-dependent gain in SOA-based optical multistage interconnection networks,” J. Lightwave Technol. **24**, 3959–3967 (2006).
24. A. Shacham and K. Bergman, “Optimizing the performance of a data vortex interconnection network,” J. Opt. Netw. **6**, 369–374 (2007).