A multihop multi-OPS optical interconnection network

David Coudert, Afonso Ferreira, Xavier Munoz

A Multihop Multi-OPS Optical Interconnection Network

David Coudert, Afonso Ferreira, and Xavier Muñoz

***Abstract*—In this paper, we study the design of regular multicast networks implemented with Optical Passive Star (OPS) couplers. We focus on an architecture based on both Kautz graphs and stack-graphs, and show that it is very cost-effective with respect to its resources requirements, namely the number of OPS couplers, power budget, scalability and number of transceivers, and presents a large ratio number-of-nodes/diameter. The important issue of medium access control is also addressed and control protocols for accessing the optical couplers are given and analyzed. Finally, we show through simulation that these control protocols efficiently implement shortest path routing on these networks.**

***Index Terms*—Control protocols, hypergraphs, Kautz digraphs, lightwave networks, logical topologies, optical passive star (OPS), routing simulations, stack-graphs.**

I. Introduction

OPTICAL interconnection is establishing itself as the most efficient technique for implementing communications in high-performance computing systems, because the maturity of optical and optoelectronic device technologies supports the deployment of very large lightwave networks which can provide a huge bandwidth.

Such networks are usually divided into two classes, according to the number of intermediate nodes a message has to visit before delivery [19]. In a *singlehop* network, the nodes communicate with each other in only one step. Unfortunately, such topologies require either a large number of transceivers per node, or rapidly tunable transmitters and receivers. On the other hand, in a *multihop* network, there is no direct path between all pairs of nodes and a communication should use intermediate nodes to reach the destination. This allows to reduce the number of (statically tuned) transmitters and receivers per node, but the processing of the information by the intermediate nodes may cause a reduction on the transmission speed.

These networks can be implemented using existing optical technology, like the low-energy-loss Optical Passive Star (OPS) coupler [10], [17], which allows incoming optical signals to be broadcast to all output ports. OPS-based networks are further classified according to the number of optical couplers used [3], being *single-OPS* [23] or *multi-OPS* [5], [6]. Although a great deal of research effort have been concentrated on single-OPS topologies [12], [19], [23], they present a severe drawback: since an OPS coupler splits an incoming optical signal in several outgoing optical signals, without signal amplification, the technological constraints related to these splitting capabilities limit its size [11] and consequently, the size of the network.

Therefore, our work focus on multi-OPS networks, which seem more viable and cost-effective under current optical technology [6], [19]. In multi-OPS networks, given a fixed number of transmitters and receivers per node, and a fixed number of nodes in the network, one should try to minimize the number of intermediate nodes a message is required to hop through. Furthermore, other parameters have also to be taken into account, like the growth capability, the simplicity of control and routing protocols, as well as the easiness of the physical architecture.

One remarkable proposal for the Optical Interconnection of Multiprocessor Systems was the Partitioned Optical Passive Star (POPS) topology described in [6], and properly modeled in [3]. A POPS network is a singlehop multi-OPS network, in which the nodes are partitioned into groups, each group of nodes being connected to all other groups through OPS couplers. In the POPS network, each OPS coupler connects one group of nodes to one other group. This network benefits of the routing facilities of singlehop networks and of the broadcast capabilities of OPS-based networks. On the other hand, it requires elaborate control protocols [6]. The POPS network represents an advance compared to a single-OPS network with respect to the total number of nodes. However, like a single-OPS network, the POPS network also presents technological limitations: the size of the groups of nodes is limited by the size of the OPS coupler, as explained before, and the maximum number of groups in the network is equal to the maximum number of transceivers per node, as allowed under current technologies. Hence, the scaling capabilities of the POPS network are also limited, motivating us to study alternatives to singlehop multi-OPS networks.

In this paper, we study the logical topology of regular multihop multi-OPS optical interconnection networks. We concentrate on topologies based on the family of Kautz digraphs[1] [16]. The Kautz digraphs have a large number of nodes, $n = d^D + d^{D-1}$, for given constant degree d and diameter D. The eccentricity[2] of Kautz Networks, for various routing protocols, were compared to different network topologies, namely, Shufflenet,

Manuscript received June 15, 1999; revised July 25, 2000. This work was supported in part by a CTI CNET and the RNRT project PORTO *Planification et Optimisation des Réseaux de Transport Optiques* with Alcatel and France Telecom. Support from the Spanish Research Council (CICYT) under project TIC97-0963 is also acknowledged.

D. Coudert and A. Ferreira are with CNRS, INRIA-I3S, F-06902 Sophia-Antipolis, France.

X. Muñoz is with DMAT, UPC, 08034 Barcelona, Spain.

Publisher Item Identifier S 0733-8724(00)10539-0.

[1]A digraph is a directed graph.

[2]The average distance taken between each pair of nodes, defined as follows. Let $G = (V, E)$ be a digraph with $|V| = n$ nodes and $|E| = m$ arcs, the eccentricity of G is $e(G) = (1/(n(n-1))) \sum_{u \in V} \sum_{v \in V - \{u\}}$ shortest-path-length (u, v)

GEMNET, and de Bruijn, in [20]. This study showed that Kautz networks are very attractive and more efficient than the others, in case it would be chosen as the logical topology of a communication network. Moreover, Kautz digraphs were used as the logical topology of the Rattlesnake ATM switch presented in [13], which is a cost effective switching system designed to build a local area ATM network that supports multimedia applications. The Kautz digraph was chosen because of its simple routing mechanism, efficient broadcast and multicast protocols, and the possibility to generate node-disjoint routes, which makes it node (or link) fault tolerant [21].

Notwithstanding, although graphs and digraphs play an important role in the analysis and synthesis of *point-to-point* networks [19], [22], OPS networks feature *one-to-many* communications where messages sent by the nodes can be broadcast to all outputs of the OPS couplers. Therefore, they can be better modeled by hypergraphs, a generalization of graphs, where edges may connect more than two nodes [1]. Hence, hypergraphs can be used to design one-to-many communication networks, as the *HyperKautz* proposed in [2]. In a nutshell, a HyperKautz network, with n nodes of degree d, uses m OPS couplers of size s, such that $dn = sm$, and has the same routing mechanism as a Kautz network with n nodes of degree sd and diameter $\log_{sd} n$. Thus, a HyperKautz network can be viewed as a Kautz network with n nodes of degree sd, in which groups of s arcs, coming from s different nodes, are merged into OPS couplers of size s (see [2] for more details). On the other hand, since the nodes connected to the inputs of any OPS coupler are different of the nodes connected to its outputs, HyperKautz networks are difficult to control and to implement, because they lack regularity. Nevertheless, this topology was evaluated as an optical network in [15], and shown to be more efficient than the Shufflenet—a point-to-point network—, at least with respect to the number of nodes in the network for constant degree and diameter, and average network load.

In this work, we introduce regular multihop multi-OPS networks based on the Kautz topology using the concept of stack-graphs, presented in [5], which allow us to model and manipulate such networks. The *stack-Kautz* network presented here, is obtained from a Kautz graph by replacing each node by a group of s nodes and each arc by an OPS coupler connecting two groups of s nodes. Hence, this network is regular with the small constant degree and the low diameter of the underlying Kautz digraph and has a large number of nodes, equal to sn, where n is the number of nodes of the Kautz digraph and s is the size of the groups that replaced the nodes. The stack-Kautz network has also a simple routing protocol and an efficient and optimal broadcast algorithm.

In the remainder of this paper, we start by recalling, in Section II, the results from the literature upon which we constructed ours. In particular, we present the OPS, we recall the stack-graphs from [5], and we give an example of a POPS network. We also recall the definition of the Kautz graph from [15]. Then, we introduce the *stack-Kautz*, in Section III. We study its characteristics and scalability, and compare it to the POPS, by studying the resources required by each network. Then, in Section IV, we give control protocols for accessing the shared OPS's and show that the bit complexity of our protocols improve on the control sequence proposed in [6] for the POPS network. Finally, we show through simulation that these control protocols efficiently implement shortest path routing on the stack-Kautz network. We close the paper with some concluding remarks and directions for further research.

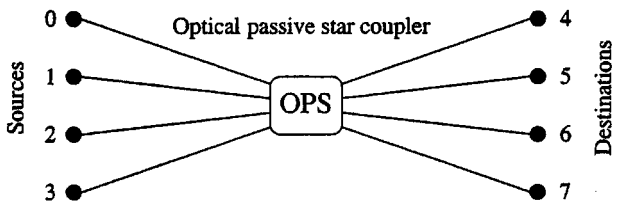


Fig. 1. Optical passive star coupler of degree 4.

II. Preliminaries

For the sake of completeness, in this section we recall the concepts of the OPS coupler, stack-graphs, the POPS network and the Kautz-graph.

A. Optical Passive Stars

An **optical passive star coupler** is a singlehop one-to-many optical transmission device. An OPS (s, z) has s inputs and z outputs. In the case where s equals z, the OPS is said to be of degree s (see Fig. 1). When one of the input nodes sends a message through an OPS coupler, the s output nodes have access to it. An OPS coupler is a **passive** optical system, i.e., it requires no external power source. It is composed of an optical multiplexer followed by an optical fiber or a free optical space and a beam-splitter that divides the incoming light signal into s equal signals of a sth of the incoming optical power. The interested reader can find in [4] a practical realization of an OPS coupler using a hologram [10] at the outputs. Another realization using optical fiber is described in [17]. Throughout this paper, we will deal with **single wavelength** OPS couplers of degree s, implying that only one optical beam can be guided through each device. Consequently, no two nodes can have concurrent access to any OPS.

B. Stack-Graphs

We saw in the introduction that one-to-many networks (e.g., OPS-based networks) are better modeled by hypergraphs. With respect to regular OPS-based networks, a particular class of directed hypergraphs, called stack-graphs, was defined in [5]. Informally, they can be obtained by means of piling up s copies of a digraph G and subsequently viewing each stack of s nodes as a hypergraph node and each stack of s arcs as a hyperarc. The value s is called the stacking factor of the corresponding stack-graph, and $\varsigma(s, G)$ denotes the stack-graph of stacking factor s, obtained from the digraph G. A formal definition can be found in [5].

Therefore, in stack-graphs an OPS coupler of degree s is modeled by a hyperarc linking two hypergraph nodes composed of s vertices each, meaning that the processors of one set (the OPS sources) can send messages only through the hyperarc,

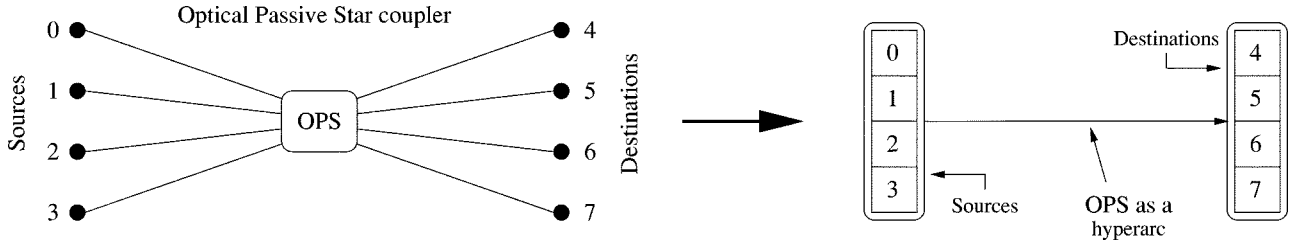


Fig. 2. Modeling an OPS by a hyperarc.

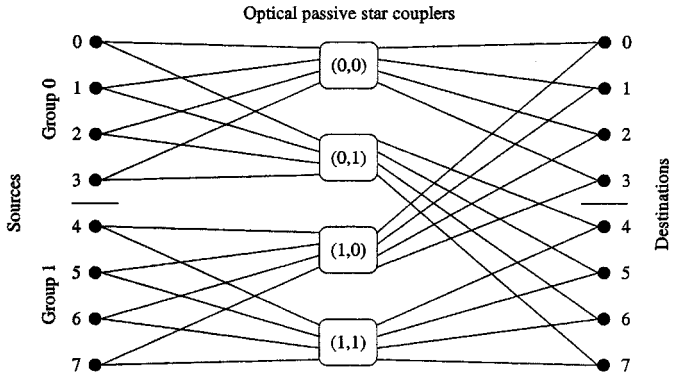


Fig. 3. POPS(4,2) with eight nodes.

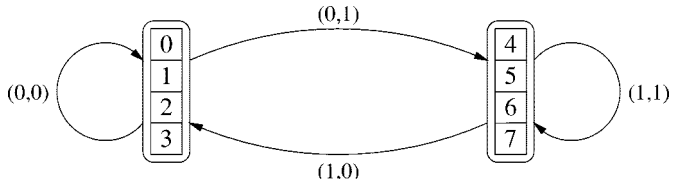


Fig. 4. POPS(4,2) modeled as $\varsigma(4,\ K_2^*)$.

whereas the other set (the OPS destinations) can receive messages only through the same hyperarc. Fig. 2 shows an OPS coupler modeled by a hyperarc.

C. Partitioned Optical Passive Star Network

The POPS network POPS$(t,\ g)$, introduced in [6], is composed of $N = tg$ nodes and g^2 OPS couplers of degree t. The nodes are divided into g groups of size t (see Fig. 3). Each OPS coupler is labeled by a pair of integers $(i,\ j)$, $0 \leq i,\ j < g$. The input of the OPS $(i,\ j)$ is connected to the ith group of nodes, and the output to the jth group of nodes. The POPS is a singlehop multi-OPS network.

Since an OPS coupler is modeled as a hyperarc, the POPS network POPS$(t,\ g)$ can be modeled as a stack-K_g^* (or $\varsigma(t,\ K_g^*)$, for short) of stacking-factor t, where K_g^* is the complete digraph with loops[3] having g nodes and g^2 arcs (see Fig. 4), as proposed in [3].

D. Kautz Digraphs

Stack-graphs represent a powerful tool for building multi-OPS networks based on graphs presenting good topological characteristics. The Kautz digraph, defined below, is one such graph.

Let us start with the definition of the Kautz digraph.

[3]A loop is an arc from a node to itself.

1) Definition: [16] The directed Kautz graph $KG(d,\ k)$ of degree d and diameter k is the digraph defined as follows (see Fig. 5).

a) A vertex is labeled with a word of length k, $(x_1,\ \ldots,\ x_k)$, on the alphabet $\Sigma = \{0,\ \ldots,\ d\}$, $|\Sigma| = d+1$, in which $x_i \neq x_{i+1}$, for $1 \leq i \leq k-1$.

b) There is an arc from a vertex $x = (x_1,\ \ldots,\ x_k)$ to all vertices y such that $y = (x_2,\ \ldots,\ x_k,\ z)$, $z \in \Sigma$, $z \neq x_k$.

Another definition of directed Kautz graph, in terms of *line digraph iteration*,[4] was proposed in [9]. It is shown that $KG(d,\ 1)$ is the complete digraph without loops K_{d+1}^+ and that $KG(d,\ k) = L^{k-1}(K_{d+1}^+)$. Fig. 5 shows $K_3^+ = KG(2,\ 1)$ and two iterations of the line digraph.

The Kautz digraph $KG(d,\ k)$ has $N = d^{k-1}(d+1)$ nodes, constant degree d and diameter k (*hence*, $k = \lfloor \log_d N \rfloor$). It is both Eulerian and Hamiltonian and optimal with respect to the number of nodes if $d > 2$ [16]. As an example, $KG(5,\ 3)$ has $N = 150$ nodes, degree 5, and diameter 3.

Notice that routing on the Kautz digraph is very simple, since a shortest path routing algorithm (every path is of length at most k) is induced by the label of the nodes.

III. A Multihop Multi-OPS Network Based on the Stack-Kautz

We now have a good model for multi-OPS networks (the stack-graphs) and also a digraph having good properties as a multihop network model (the Kautz digraph). In this section we introduce a multihop multi-OPS architecture based on the *stack-Kautz*.

A. Definition and Main Characteristics

In order to define the optical interconnection network called stack-Kautz, we use the Kautz digraph with loops $KG^*(d,\ k)$, where every node has a loop and hence degree $d+1$. Thus, we can define the stack-Kautz as follows.

1) Definition: The **stack-Kautz** $SK(s,\ d,\ k)$ is the stack-graph $\varsigma(s,\ KG^*(d,\ k))$ of stacking-factor s, degree $d+1$ and diameter k (see Fig. 6).

The stack-Kautz network has the topology of $SK(s,\ d,\ k)$ and $N = sd^{k-1}(d+1)$ nodes. Each node is labeled by a pair $(x,\ y)$ where x is the label of the stack in $KG(d,\ k)$ and y is an integer $0 \leq y < s$, i.e., x is the label of a node group and y is the label of a node in this group. Since the stack-Kautz network

[4]Given a digraph G, the *line digraph* operation allows to define a new digraph $L(G)$ whose vertex set is in one to one correspondence with the set of arcs of G. A vertex u of $L(G)$, representing the arc $u = (x,\ y)$ in G, is adjacent to a vertex v if and only if v represents the arc $v = (y,\ z)$ in G. We denote by $L^2(G)$ the digraph $L(L(G))$ and by $L^k(G)$ the digraph $L(L^{k-1}(G))$.

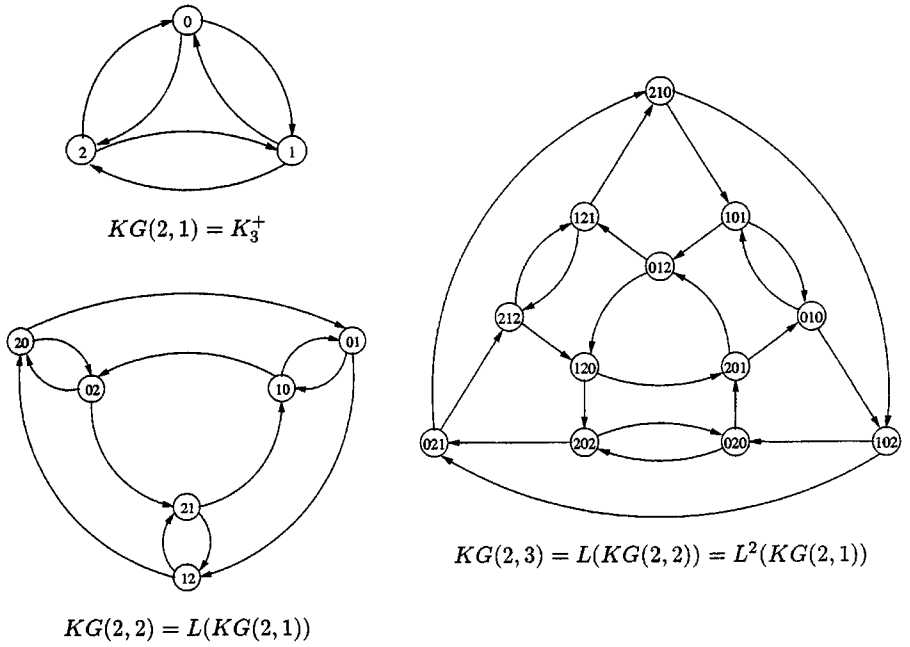


Fig. 5. Three Kautz digraphs: $KG(2,\ 1) = K_3^+$ and two iterations of the line digraph, $KG(2,\ 2) = L(KG(2,\ 1))$ and $KG(2,\ 3) = L(KG(2,\ 2)) = L^2(KG(2,\ 1))$.

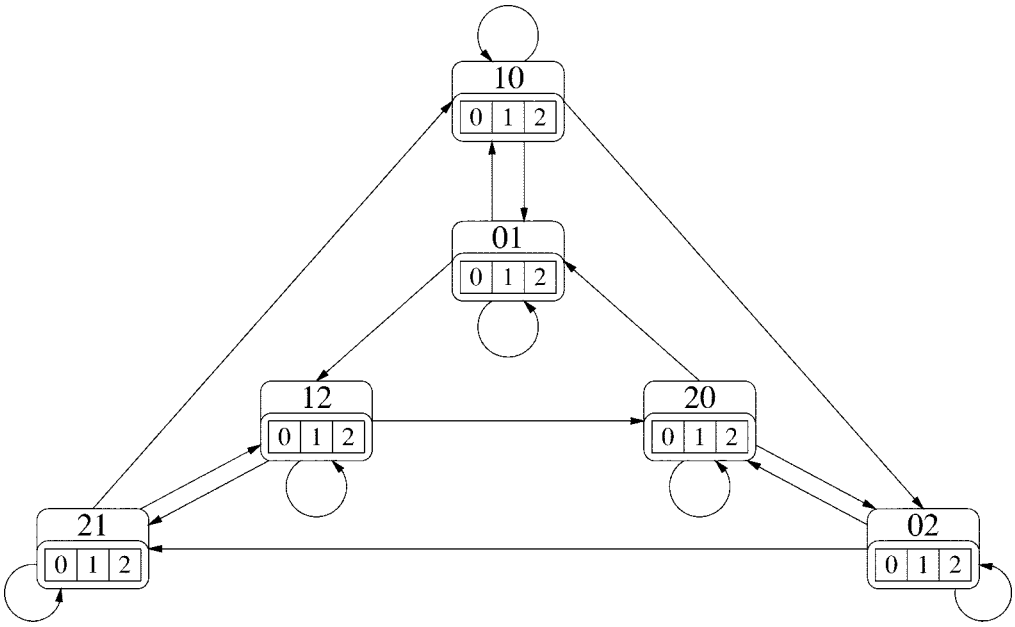


Fig. 6. Stack-Kautz network $SK(3,\ 2,\ 2)$.

inherits most of the properties of the Kautz digraph, like shortest path routing, fault tolerance and others, it is a good candidate for the topology of an OPS-based lightwave network.

2) Diameter: An OPS-based network with the topology of a stack-Kautz network $SK(s,\ d,\ k)$ has $N = sd^{k-1}(d+1)$ nodes divided in $g = d^{k-1}(d+1)$ groups of size s. It is possible to preserve a small diameter k and have a large number of nodes. For instance, $SK(12,\ 5,\ 3)$ has $N = 1800$ nodes and diameter 3.

3) Number of OPS Couplers: Each group of s nodes has an output degree $d+1$, hence it is connected to the input of $d+1$ OPS couplers of degree s. The stack-Kautz network $SK(s,\ d,\ k)$ requires $d^{k-1}(d+1)^2$ OPS couplers of degree s. Notice that the number of OPS's is independent of the stacking-factor.

4) Number of Transceivers: Each node has one transmitter and one receiver per link. Hence there are $d+1$ transmitters and $d+1$ receivers per node and a total of $2sd^{k-1}(d+1)^2$ transceivers in the network.

In the following we discuss ways to optimize the assignment of parameters s, d and k. Table I and Fig. 7 show examples of such assignments.

TABLE I
RESOURCES EXAMPLES FOR SK AND POPS

	$POPS(60,30)$	$SK(20,9,2)$	$SK(12,5,3)$	$SK(12,5,5)$
Groups	30	90	150	3750
Nodes	1800	1800	1800	45000
Diameter	1	2	3	5
Degree OPS's	60	20	12	12
OPS's	900	900	900	22500
Tr.|Rec. per proc.	30	10	6	6
Tr.|Rec. total	54000	18000	10800	270000
Power budget	60	20	12	12
Broadcast (time steps)	2	3	4	6
Advanced Control (# of bits)	570	280	108	108

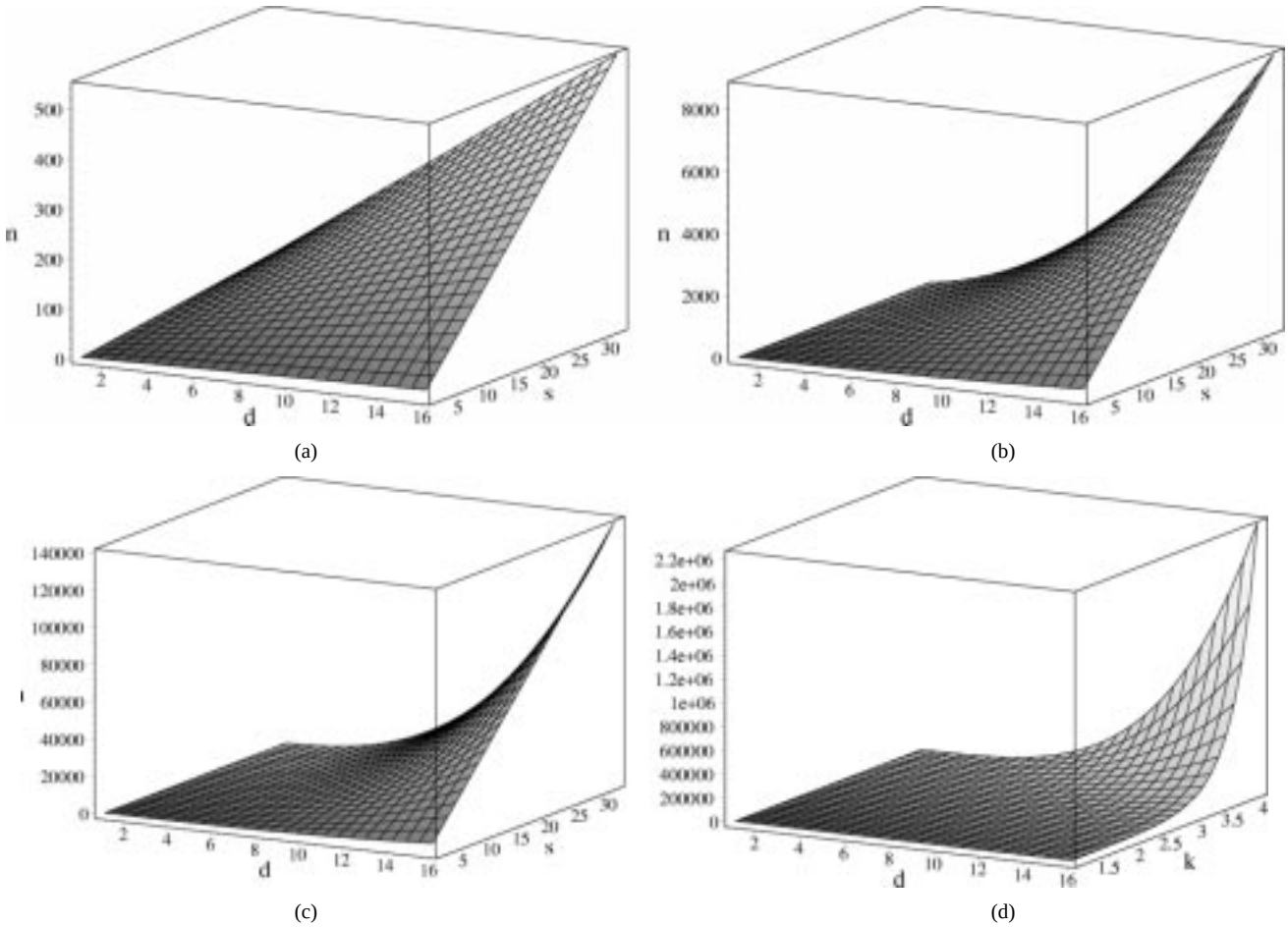


Fig. 7. Plots (a), (b), and (c) depict (a) n the number of nodes of the POPS, and of (b) the stack-Kautz of diameter 2 and of (c) diameter 3, when both the degree d and the stacking-factor s vary. Plot (d) shows the behavior of n the number of nodes of the stack-Kautz of stacking-factor $s = 32$, when both the degree d and the diameter k vary.

B. Power Budget and Scalability

The scalability analysis of a network allows us to find the best way to build it in terms of number of nodes/network diameter/stacking factor/degree/technologies/power budget/cost. In a multi-OPS network like the stack-Kautz $SK(s, d, k)$, it corresponds to the choice of the three parameters s, d and k.

The power budget corresponds to the cost of sending a message from one node to another through an OPS coupler in terms

of energy. When a node sends a message through an OPS coupler of degree s, s nodes receive it. Let the optical power required by a receiver to detect a message be normalized to 1 and for the sake of simplicity, make the assumption that no loss of light power occur during a transmission. Therefore the power budget of sending one message through an OPS coupler is s, the stacking-factor. Consequently, the maximum optical power which can be delivered by an optical transmitter divided by the minimum value of the optical power required by a receiver to detect a message gives the maximum value of the stacking-factor. As an example, an OPS coupler of degree 16 can be realized based on VCSEL's and diffractive optics [11].

The stack-Kautz network $SK(s, d, k)$ has $N = sd^{k-1}(d+1)$ nodes divided in $g = d^{k-1}(d+1)$ groups of size s, and $(d+1)g$ OPS couplers of degree s. Each node has degree $d+1$ (i.e., $2(d+1)$ transceivers) and the total number of transceivers in the network is $2(d+1)N$. The diameter, k, of the network corresponds to the maximum number of nodes a message has to jump through before delivery.

When the **stacking-factor,** s, increases from s_1 to s_2, then:

- the degree of the OPS couplers also increases from s_1 to s_2;
- the total number of transceivers increases from $2s_1d^{k-1}(d+1)^2$ to $2s_2d^{k-1}(d+1)^2$;
- the number of nodes in the network increases from $s_1d^{k-1}(d+1)$ to $s_2d^{k-1}(d+1)$ [see Figs. 7(a)–(c)].

Therefore, the above three resources are increased by a factor s_2/s_1. The other parameters remain unchanged.

When the **degree**, $d+1$, of the nodes increases from d_1+1 to d_2+1, then:

- the number of transceivers per node also increases from $2(d_1+1)$ to $2(d_2+1)$;
- the number of groups increases from $d_1^{k-1}(d_1+1)$ to $d_2^{k-1}(d_2+1)$;
- the number of OPS couplers increases from $d_1^{k-1}(d_1+1)^2$ to $d_2^{k-1}(d_2+1)^2$;
- the total number of transceivers increases from $2sd_1^{k-1}(d_1+1)^2$ to $2sd_2^{k-1}(d_2+1)^2$;
- the number of nodes in the network increases from $sd_1^{k-1}(d_1+1)$ to $sd_2^{k-1}(d_2+1)$ [see Figs. 7(a)–(d)].

This means that the above four resources are increased by a factor $(d_2/d_1)^{k-1}((d_2+1)/(d_1+1))$. The other parameters remain unchanged.

Finally, when the **diameter** k of the network increases from k_1 to k_2, then:

- the number of groups increases from $d^{k_1-1}(d+1)$ to $d^{k_2-1}(d+1)$;
- the number of OPS couplers increases from $d^{k_1-1}(d+1)^2$ to $d^{k_2-1}(d+1)^2$;
- the total number of transceivers increases from $2sd^{k_1-1}(d+1)^2$ to $2sd^{k_2-1}(d+1)^2$;
- the number of nodes in the network increases from $sd^{k_1-1}(d+1)$ to $sd^{k_2-1}(d+1)$ [see Fig. 7(d)].

Hence, the above four resources are increased by a factor $d^{k_2-k_1}$. The other parameters remain unchanged.

Therefore, in order to proportionally decrease the power budget, with a fixed number of nodes, the number of groups must be large with respect to the group size. Also, it is better to increase the diameter of the network in order to minimize the number of transmitters and receivers per node. Thus, by increasing the diameter of the network, the power budget and the resources are proportionally reduced with respect to the number of nodes. However it is necessary to preserve $s \geq d$ to have more nodes in the network than OPS couplers.

As an example, given in Table I below, $SK(20, 9, 2)$ has $N = 1800$ nodes divided in 90 groups of size 20 and 900 OPS couplers for a diameter 2. Each node has 10 transceivers and the power budget is equal to 20. Let the diameter increase of one unit and let the number of nodes and OPS couplers stay unchanged. Thus, we obtain $SK(12, 5, 3)$ which is composed of 150 groups of 12 nodes. The power budget is decreased to 12 and the number of transceivers per node is now equal to 6.

Table I also gives numerical evidence for both POPS and stack-Kautz networks. The three networks POPS$(60, 30)$, $SK(20, 9, 2)$ and $SK(12, 5, 3)$ have the same number of nodes for different diameters. $SK(12, 5, 3)$ appears to be better than the others in terms of number of transceivers per node and total number of transceivers, degree of its OPS couplers, power budget and bit complexity of its medium access control protocol (see Section IV). Furthermore, broadcasting arbitrary but fixed size messages takes only two times steps more than in POPS, as shown below.

C. Broadcasting Algorithm

An *all-port* broadcast algorithm for the Kautz digraph completes in optimal k steps, since in the all-port model, a node transmits the message through all its links simultaneously. Thus, the broadcast on any digraph G with diameter D, completes in optimal D steps.

Assuming that the size s of a group is greater than the degree, $s \geq d$, one can design a simple and optimal broadcast algorithm for the stack-Kautz. It suffices to make the broadcast initiator use in the first step the loop, in order to broadcast the information inside its group, then make each element of the group inform all the elements of a different contiguous group, and this until completion. A direct analysis shows that broadcast completes in optimal $k+1$ steps.

IV. Medium Access Control Protocols

A single wavelength OPS coupler can only be accessed by one node at a time. Since s nodes share $d+1$ OPS couplers, $s \geq d$, an efficient medium access control protocol is required. One such protocol was proposed in [6] for the POPS network $POPS(t, g)$ with g groups of size t. It supposes that a node can use all its receivers at the same time, but only one output port. Each group of t nodes contains one node in charge of the control of the group.

Their control protocol consists of two phases. During the first phase, each node sends to the node in charge of the control of the group a word of $\log g$ bits, encoding the index of the OPS coupler it wants to use. Then, the node in charge of the control decides which node is going to gain access to the OPS couplers and sends a word of t bits encoding an acknowledgment or a refusal to each node.

During the second phase, the control protocol performs a *hand-shake* between senders and receivers with acknowledgment or refusal. This second phase requires two global exchanges of information. The first one is to ensure that the receiving groups have the index of the node which is going to receive the message, encoded in $g \log t$ bits. The second global exchange is needed to route back acknowledgment or refusals, encoded in g bits. The bit complexity of the first phase is $t \log g + t$ bits and $g \log t + g$ bits for the second. The total bit complexity of the control protocol is $t \log g + g \log t + t + g$ bits. For instance, take POPS$(60, 30)$, which has 1800 nodes divided in 30 groups of size 60. The bit complexity of this control protocol is 570 bits, which is small compared to the size of the network.

For the control protocols of our stack-Kautz network, we suppose, as in [6], that a node can receive messages on all its links at the same time. By reading the header of a message, a node can decide whether it has to process it or not. We also suppose that a node can always receive a message, contrary as in [6] in which acknowledgment or refusals are transmitted during the second global exchange (i.e.: a node has the opportunity of refusing a new message when its buffers are full in order to avoid a loss of messages). Thus, in our case, the control protocol has just to avoid local conflicts, inside a group of nodes.

A. Simple Control

1) Hypothesis: Each node has a buffer of messages to be transmitted (FIFO). A node can request one OPS per communication step, in order to transmit the message which is at the top of the buffer of messages.

The node which is in charge of the control of the group has s counters, one for each node. Each counter is increased of 1 when the corresponding node receives a refusal. A counter is set to 0 when the corresponding node receives an acknowledgment.

A simple control protocol for a group of nodes, under hypothesis Section IV-A1, is as follows. Let p_0 be the node in charge of the control of its group of s nodes.

a) All nodes of the group send successively the index of the OPS coupler that each one wants to use to p_0. This index is a word of $\log(d+1)$ bits.
b) Node p_0 assigns an OPS coupler to a node if it is the only one which wants to use it or if its counter is the highest. In case of conflict (two or more equal counters) one of the nodes is randomly chosen.
c) Node p_0 sends a word of s bits (one bit per node), encoding acknowledgment and refusals, to all the nodes of the group. It adds 1 to each counter corresponding to a node receiving a refusal and sets to 0 each counter corresponding to an acknowledgment.

The bit complexity of this control protocol is $s \log(d+1) + s$ bits. The time complexity is computed in the following proposition.

2) Proposition: The time complexity of the simple control protocol for a group of size s and degree $d+1$ is $O(s)$.

Proof: Step 1 takes time $O(s)$ and Step 3 is composed of a single send. With respect to Step 2, once node p_0 has received the indexes of the OPS couplers from the s nodes of the group, p_0 puts the node indexes in a table with $d+1$ entries (one per OPS). Each entry of this table contains x_i indexes, such that $\sum_{i=0}^{d} x_i \leq s$. Node p_0 then chooses one of the x_i elements for each entry of the table. Since it takes $O(\log x_i)$ comparisons per entry, the time complexity of Step 2 is proportional to $\sum_{i=0}^{d} \log x_i \leq \sum_{i=0}^{d} x_i \leq s$.

B. Advanced Control

Another control protocol can be considered under the following hypothesis.

1) Hypothesis: Each node has $d+1$ buffers of messages to be transmitted, one per OPS coupler. At most $d+1$ messages can be proposed per node and per communication step.

The node in charge of the control of the group has $s(d+1)$ counters ($d+1$ per node), i.e., 1 per OPS coupler for each node. It adds 1 to each counter which corresponds to a node receiving a refusal and sets it to 0 when the node receives an acknowledgment.

An advanced control protocol of a group of nodes, under hypothesis Section IV-B1, is as follows. Let p_0 be the node in charge of the control of a group of s nodes.

a) All nodes of the group send successively a word of $d+1$ bits, encoding the presence or not of a message to be transmitted in each of its $d+1$ buffers of messages (one for each OPS), to p_0.
b) Node p_0 realizes a maximum weight matching between the nodes and the OPS couplers using the weight induced by the counters. This maximum weight matching is realized using a standard algorithm [7].
c) Node p_0 sends a word of $s \log(d+2)$ bits, encoding for each node an acknowledgment (index of an OPS coupler) or a refusal (special word), to all nodes in its group.

The bit complexity of this control protocol is $s(d+1) + s \log(d+2)$ bits.

2) Proposition: The **time complexity** of the advanced control protocol for a group of size s and degree $d+1$ is $O(s(d+1)^2)$.

Proof: Step 1 takes time $O(s)$ and Step 3 is composed of a single send. With respect to Step 2, the time complexity of the maximum weight matching algorithm is $O(s(d+1)^2)$. □

A short numerical comparison between this two control protocols, on $SK(12, 5, 3)$, shows that the bit complexity of the simple protocol is 48 bits, and it is equal to 108 bits with the advanced protocol. $SK(12, 5, 3)$ has 1800 nodes divided in 150 groups of size 12.

The advanced control protocol is more complex than the simple protocol but it allows us to maximize the utilization of the OPS couplers at each communication step. As it will be shown in Section IV-C, we obtain a good average delivery time for messages with both the simple and the advanced protocols.

C. Performance Issues

We built a simulator for the stack-Kautz network which implements a shortest path routing algorithm which guarantees that path-lengths are bounded by the diameter of the network.

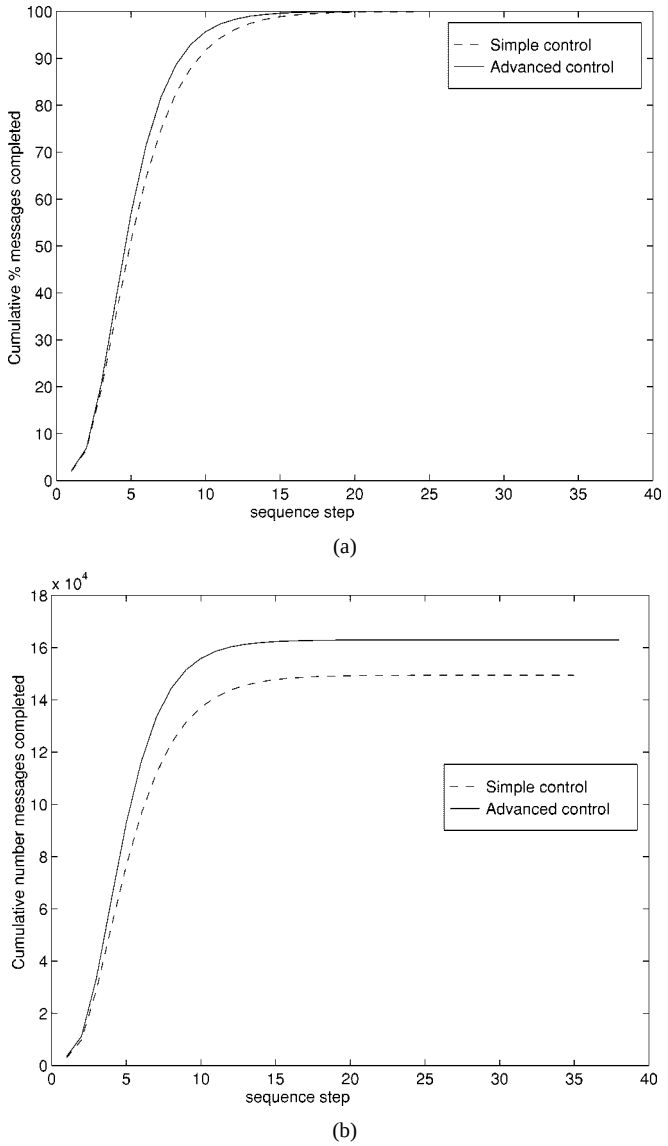


Fig. 8. Left curve: cumulative percentage of messages completed. Right curve: cumulative number of messages completed.

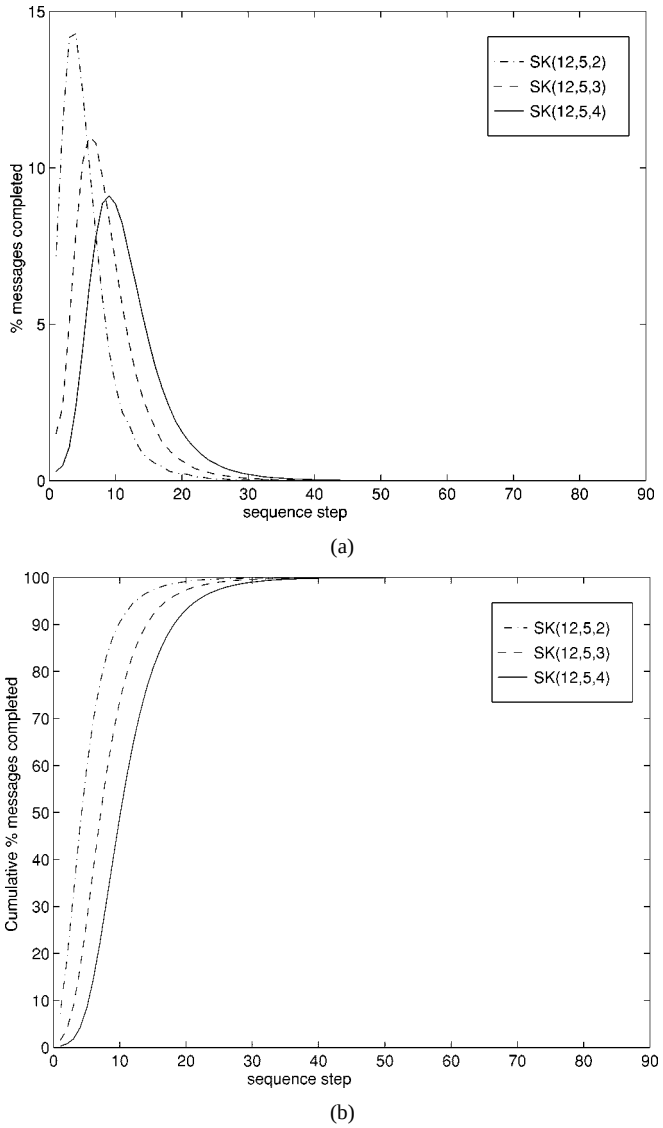


Fig. 9. Percentage of messages completed for different diameters.

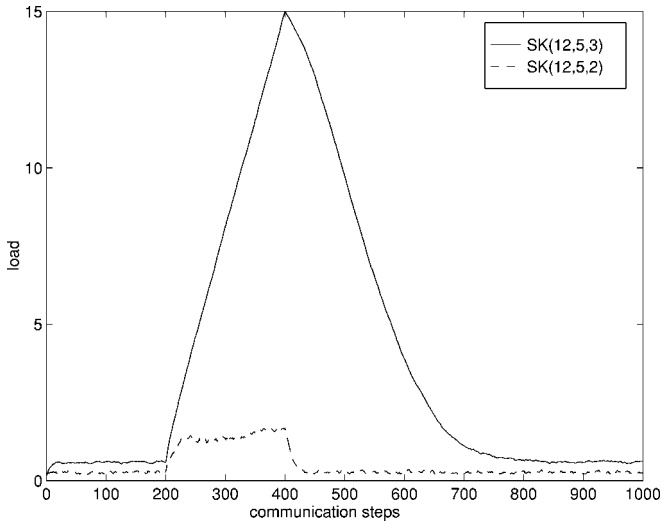
Our simulator can use either the simple or the advanced control protocol. Messages are considered to have a normalized unit size.

In order to compare the stack-Kautz control protocols, we kept the load γ of the network $SK(12, 5, 3)$, with 1800 nodes, at $\gamma = 0.5$ by "injecting" new messages, during 1000 communication steps. Fig. 8 shows, for our protocols, the accumulated percentage of delivered messages out of the total number of messages, as a function of the number of steps needed, as well as the total number of the delivered messages. We remark that the percentages are the same for the two protocols, but not the total number of delivered messages. The difference between the total number of delivered messages in the right curve is explained by the fact that even though the load is kept at the same value for the two protocols, the "speed" of the messages is not the same, and therefore, the total number of injected messages is not the same either. Thus, it becomes clear that the advanced control protocol is much better than the simple control protocol, with respect to the number of delivered messages.

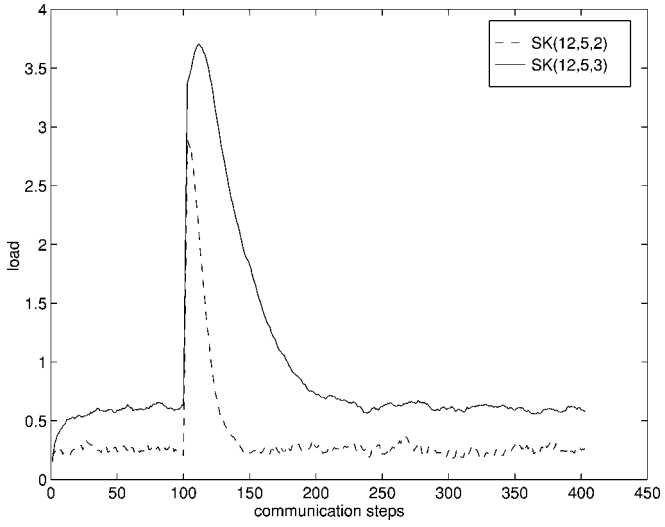
Fig. 9 gives the result of the simulation of the networks $SK(12, 5, 2)$, $SK(12, 5, 3)$ and $SK(12, 5, 4)$, with 360, 1800, and 9000 nodes respectively, for which the load γ has been kept at $\gamma = 1$ during 1000 communication steps, i.e., in average there is always one message per node in the network.

Note that the theoretical delivery delay is $\gamma s\epsilon/d+1$. Since the eccentricity of $SK(12, 5, 4)$ is $\epsilon = 3.74$, the average theoretical delivery delay is 7.5 steps, while Fig. 9 gives us an average delay of ten steps. This can be explained by the low load, which implies that all links are not necessarily used at each communication step.

Finally, it is also interesting to study traffic situations in which the network could fall in crisis. Hence we plotted the load of the networks $SK(12, 5, 3)$ having 1800 nodes and $SK(12, 5, 2)$ having 360 nodes, when the probability of arrival of a new message is sharply increased. This could model, for instance, cases where global message exchanges are performed in the middle of a normal state of the network. In Fig. 10, the load γ of the network is induced by the probability p that a node creates a new message at every step. For the curve on the left, we set $p = 0.1$



(a)



(b)

Fig. 10. Network load evolution for $SK(12,\ 5,\ 2)$ and $SK(12,\ 5,\ 3)$, when the messages arrival rate drastically changes.

during 200 steps, then $p = 0.2$ during another 200 steps, and finally back to $p = 0.1$. For the curve on the right, we set $p = 0.1$, then $p = 1$ for three steps and back to $p = 0.1$. The results show that the stack-kautz is not blocked by either slow or sharp rises of the network load, and that the load stabilizes again in time, as long as the messages arrival rate goes back to normal.

V. Conclusion

In this paper, we introduced a regular multihop multi-OPS optical network: the stack-Kautz. This network makes possible the interconnection of a large number of nodes because it is based on a low degree and low diameter graph. We have proposed control and routing protocols which guarantee an upper bound for the delivery delay of the messages through the network.

One conclusion of our work is that the stack-Kautz compares very well against other multi-OPS logical topologies. Moreover, the stack-Kautz may actually be built, since we showed in [8] its optical layout using existing optoelectronic technologies, namely the *Optical Transpose Interconnection System* from [18], [22].

Acknowledgment

The authors are grateful to the anonymous referees for their thorough reading of the manuscript and very helpful comments.

References

[1] C. Berge, *Graphs and Hypergraphs*. Amsterdam: North-Holland, 1973.
[2] J.-C. Bermond, R. Dawes, and F. Ergincan, "De Bruijn and Kautz bus networks," *Networks*, vol. 30, pp. 205–218, 1997.
[3] P. Berthomé and A. Ferreira, Eds., *Optical Interconnections and Parallel Processing: Trends at the Interface*. Norwood, MA: Kluwer, 1997.
[4] M. Blume, F. McCormick, P. Marchand, and S. Esener, "Array interconnect systems based on lenslets and CGH," in *SPIE Int. Symp. Optical Science, Engineering, and Instrumentation*, San Diego, CA, 1995, Technical Report 2537-22.
[5] H. Bourdin, A. Ferreira, and K. Marcus, "A performance comparison between graph and hypergraph topologies for passive star WDM lightwave networks," *Computer Networks and ISDN Systems*, vol. 30, pp. 805–819, 1998.
[6] D. Chiarulli, S. Levitan, R. Melhem, J. Teza, and G. Gravenstreter, "Partitioned Optical Passive Star (POPS) topologies for multiprocessor interconnection networks with distributed control," *J. Lightwave Technol.*, vol. 14, no. 7, pp. 1601–1612, 1996.
[7] W. Cook, W. Cunningham, W. Pulleyblank, and A. Schrijver, *Combinatorial Optimization*, ser. Discrete Mathematics and Optimization: Wiley Interscience, 1998.
[8] D. Coudert, A. Ferreira, and X. Muñoz, "Topologies for optical interconnection networks based on the Optical Transpose Interconnection System," *OSA Appl. Opt.*, vol. 39, no. 17, pp. 2965–2974, June 2000.
[9] M. A. Fiol, J. L. A. Yebra, and I. Alegre, "Line digraphs iterations and the (d, k) digraph problem," *IEEE Trans. Comput.*, vol. 33, pp. 400–403, 1984.
[10] D. Gardner, P. Marchand, P. Harvey, L. Hendrick, and S. Esener, "Photorefractive beamsplitter for free space optical interconnection systems," *OSA Appl. Opt.*, vol. 37, no. 26, pp. 6178–6181, September 1998.
[11] M. Ghisoni, H. Martinsson, N. Eriksson, M. Li, A. Larsson, J. Bengtsson, A. Khan, and G. Parry, "4 × 4 fan-out spot generator using GaAs based VCSEL's and diffractive optical element," *IEEE Photon. Technol. Lett.*, vol. 9, p. 508, 1997.
[12] E. Hall, J. Kravitz, R. Ramaswani, M. Halvorson, S. Tenbrink, and R. Thomsen, "The Rainbow-II gigabit optical network," *IEEE J. Select. Areas Commun.*, vol. 14, no. 5, pp. 814–823, June 1996.
[13] P. Havinga and G. Smit, "Rattlesnake—A single chip high-performance ATM switch," in *Proc. Int. Conf. Multimedia and Networking (MmNet)*, Japan: IEEE Press, 1995, pp. 208–217.
[14] M. Imase, T. Soneoka, and K. Okada, "A fault-tolerant processor interconnection network," *Syst. Comput. Jpn.*, vol. 17, no. 8, pp. 21–30, 1986.
[15] S. Jiang and T. Stern, "Regular multicast multihop lightwave networks," in *IEEE InfoCom'95*, 1995, pp. 692–700.
[16] W. H. Kautz, "Bounds on directed (d, k) graphs. Theory of cellular logic networks and machines," *AFCRL-68-0668, SRI Project 7258, Final Report*, pp. 20–28, 1968.
[17] Y. Li, T. Wang, and K. Fasanella, "Inexpensive local interconnect solutions using side-coupling polymer optical fibers," in *Massively Parallel Processing using Optical Interconnections*, Canada: IEEE Press, June 1997, pp. 45–51.
[18] G. Marsden, P. Marchand, P. Harvey, and S. Esener, "Optical Transpose Interconnection System architectures," *OSA Opt. Lett.*, vol. 18, no. 13, pp. 1083–1085, July 1993.
[19] B. Mukherjee, *Optical Communication Networks*, ser. Computer Communications. New York: McGraw-Hill, 1997.
[20] G. Panchapakesan and A. Sengupta, "On multihop optical network topology using Kautz digraphs," in *IEEE INFOCOM'95*, 1995, pp. 675–682.
[21] G. Smit and P. Havinga, "Multicast and broadcast in the rattlesnake ATM switch," in *Proc. Int. Conf. Multimedia and Networking (MmNet)*, Japan: IEEE Press, 1995, pp. 218–226.
[22] F. Zane, P. Marchand, R. Paturi, and S. Esener, "Scalable network architectures using the Optical Transpose Interconnection System (OTIS)," *J. Parallel Distrib. Comput.*, vol. 60, no. 5, pp. 521–538, 2000.

[23] Z. Zhang and A. S. Acampora, "Performance analysis of multihop lightwave networks with hot potato routing and distance-age-priorities," *IEEE Trans. Commun.*, vol. 42, pp. 2571–2581, Aug. 1994.

David Coudert received the M.Sc. degree in computer science from the Ecole Normale Superieure of Lyon, France, in 1997 and is currently working towards the Ph.D. degree candidate in the joint project MASCOTTE of the CNRS/INRIA/UNSA, located at the INRIA Sophia-Antipolis.

His research interests include graph theory, free space and WDM optical networks, network designs, and combinatorial optimizations.

Afonso Ferreira received his M.Sc. from the University of São Paulo, Brazil, in 1986 and the Ph.D. degree from the Institut National Polytechnique de Grenoble, France, in 1990, both in computer science.

In October 1990, he joined the French CNRS (Centre National de la Recherche Scientifique) as a Researcher. Since September 1997, has been with the INRIA Sophia Antipolis, where he works at the joint project MASCOTTE of the CNRS/INRIA/UNSA. His research focuses on the rich interdependence between algorithms design and analysis, modeling, and experimental analysis. Areas of interests include computational telecommunications, optimisation, discrete mathematics, and parallel processing. His current application areas concentrate on communication issues in optical and wireless networks.

Dr. Ferreira serves on the editorial boards of the *Journal of Parallel and Distributed Computing*, *Journal of Interconnection Networks*, *Parallel Processing Letters*, and *Parallel Algorithms and Applications*. He was Guest Editor for special issues of *Mobile Networks (MoNet)*, *Parallel Computing*, *Parallel Processing Letters*, *Theoretical Computer Science*, and *Journal of Parallel and Distributed Computing*, on discrete algorithms and methods for mobile communications, and for parallel systems. He has published over 100 papers in international journals and conferences, has coauthored one book (on satellite networks, in French), has edited 10 books, and has organized and served in Programme Committees for more than 50 international conferences and workshops. He is General co-Chair for the IEEE IPDPS (International Parallel and Distributed Processing Symposium) in 2001 and 2002, and Programme Vice-Chair for the 18th STACS (Symposium on Theoretical Aspects of Computer Science) in 2001 and Programme Chair for the 19th STACS in 2002. He served as Publicity Chair for IPDPS'98, and Workshops Chair for the ACM/IEEE Mobicom'98 (International Conference on Mobile Computing and Networking). He is a member of the Steering Committees of STACS, and of the annual Workshops: ACM Dial M for Mobility (Discrete Algorithms and Methods for Mobile Computing and Communications), IEEE Irregular (Solving Irregularly Structured Problems in Parallel), and IEEE WOCCS (Optical Communications for Computing Systems). Among other activities, he is currently member of the Advisory Boards of the IEEE Technical Committee for Parallel Processing, and of the conferences EuroPar and PDCS.

Xavier Muñoz was born in Barcelona, Spain, in 1967. He received the Telecommunications Engineering and Ph.D. degrees in 1992 and 1996, respectively, both from the Universitat Politecnica de Catalunya, Barcelona, Spain.

He is an Assistant Professor in the Departament de Matematica Aplicada i Telematica of the Universitat Politecnica de Catalunya. His research interests include the applications of graph theory to interconnection networks and comunications in networks.