# Multiprocessor Architectures Using Multi-Hop Multi-OPS Lightwave Networks and Distributed Control

David Coudert
LIP ENS-Lyon
46, allée d'Italie
69364 Lyon Cedex 07, France
David.Coudert@lip.ens-lyon.fr

Afonso Ferreira
CNRS-Projet SLOOP
Laboratoire I3S & INRIA
Sophia-Antipolis, France
Afonso.Ferreira@sophia.inria.fr

Xavier Muñoz
Matemàtica Aplicada i Telemàtica
UPC
08034 Barcelona, Spain
xml@mat.upc.es

## Abstract

*Advances in optical technology have increased the interest for multiprocessor architectures based on lightwave networks because of the vast bandwidth available. In this paper we propose a passive star multi-hop lightwave network called stack-Kautz, based on the Kautz graph. We show that this architecture is very cost-effective with respect to its resources requirements. We also propose control protocols for accessing the optical passive star couplers, which improve on the bit complexity of the control sequence proposed in the literature for the Partitioned Optical Passive Star network. Finally, we show through simulation that these control protocols efficiently implement shortest path routing on the stack-Kautz network.*

## 1 Introduction

The advances in optical technology, such as low energy loss optical passive star couplers (OPS) [8], as well as tunable optical transmitters and receivers [13], have increased the interest for optical interconnection networks for multiprocessor systems because of their large bandwidth [4, 5].

The topologies proposed for such networks can be divided in two classes, according to the number of intermediate processors a message has to visit before delivery [10, 11]. In a *single-hop* network, the nodes communicate with each other in only one step. Such topologies require either a large number of transmitters and receivers per node, or rapidly tunable transmitters and receivers. Unfortunately, the delay which is presently needed to tune the transmitters is quite large (roughly a micro-second) [13] compared to the transmission delay for a message, which causes important latencies in communication.

In a *multi-hop* architecture, there is no direct path between all pairs of nodes: a communication should use intermediate nodes to reach the destination. This allows the usage of statically tuned transmitters and receivers, but on the other hand the processing of the information by the intermediate nodes causes a loss of speed. Furthermore, it is clear that the smaller the number of intermediate nodes, the faster communications take place in the network. The choice of the topology for the network, at both physical and logical levels, is thus crucial. For a fixed number of transmitters and receivers per node, and a fixed number of nodes in the network, the number of intermediate nodes a message is required to hop through should be minimized. However, other parameters have to be taken into account, like the simplicity of control and routing protocols, as well as the easiness of the optical realization.

Networks based on OPS's can be further classified according to the number of optical couplers used, being single-OPS or multi-OPS [6]. Under current optical technology, however, the latter seem more viable and cost-effective. Therefore, in this paper we address design issues in multi-hop multi-OPS architectures and introduce a new optical interconnection network for multiprocessor systems, called *stack-Kautz*, allowing *one-to-many* communications at every communication step. It is based on the Kautz graphs (constant degree $d$ and diameter $\lfloor \log_d N \rfloor$, for $N$ nodes) [1, 7, 9, 12] and on the stack-graphs [4, 6]. We study design characteristics of the stack-Kautz network, its scalability, and give control protocols for accessing the shared OPS's. Finally, we show through simulation that these control protocols efficiently implement shortest path routing on this network.

## 2 Preliminaries

In this section we present the OPS coupler, the POPS network, stack-graphs and the Kautz-graph.

## 2.1 Optical passive star

An **optical passive star coupler** is a single-hop 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 Figure 1). When one of the input processors sends a message through an OPS coupler, the $s$ output processors have access to it. Throughout this paper, we will use **single-wavelength** OPS couplers of degree $s$. Consequently, only one processor can send an optical signal through it per time step. An OPS coupler is a **passive** optical system, i.e. it requires no power source. It is composed of an optical multiplexer followed by an optical fiber or a free optical space and an optical de-multiplexer that divides the incoming light signal into $s$ equal signals of a $s$-th of the incoming optical power. Note that only one optical beam has to be guided through the circuit [8]. A practical realization of an OPS coupler using a hologram at the outputs, is described in [3].

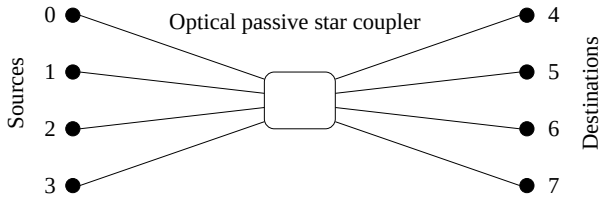


**Figure 1. An OPS coupler of degree 4.**

## 2.2 A single-hop multi-OPS network

The **Partitioned Optical Passive Star network** $POPS(t, g)$, introduced in [5], is composed of $N = tg$ processors and $g^2$ OPS couplers of degree $t$. The processors are divided into $g$ groups of size $t$ (see Figure 2). 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 $i$-th group of processors, and the output to the $j$-th group of processors.

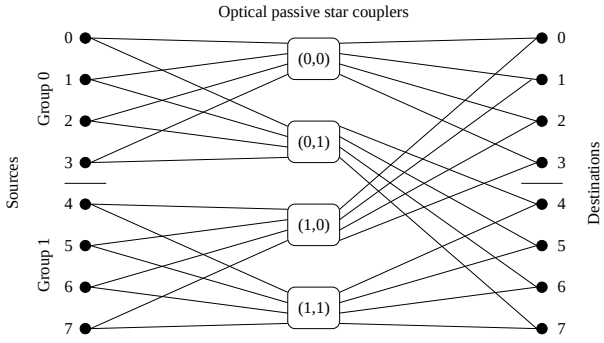


**Figure 2. POPS(4,2).**

## 2.3 A model for multi-OPS networks

In this section we recall a model which allows us to study more easily multi-OPS networks. The stack-graphs were first defined in an early version of [4] as hyper-graphs built from graphs. Briefly, stack-graphs are obtained by piling up copies of the original graph and subsequently replacing each stack of edges by one hyper-edge. A formal definition is as follows.

**Definition 1** *[4] Let $G(V, A)$ be a directed graph. The* **stack-graph** *$\varsigma(s, G) = (V_\varsigma, A_\varsigma)$ is as follows, $s$ being called* **stacking-factor** *of the stack-graph.*

1. *The set of nodes $V_\varsigma$ of $\varsigma(s, G)$ is $V_\varsigma = \{0, \cdots, s-1\} \times V$, $s \geq 1$.*
2. *Let $\pi$ be the projection function defined from $V_\varsigma$ onto $V$ such that $\pi((i, v)) = v$, for $0 \leq i < s$ and $v \in V$.*
3. *The set of hyper-arcs $A_\varsigma$ of $\varsigma(s, G)$ is then $A_\varsigma \stackrel{def}{=} \{a_\varsigma = (\pi^{-1}(u), \pi^{-1}(v)) | (u, v) \in A\}$.*

An OPS coupler can thus be modeled as a hyper-arc as shown below.

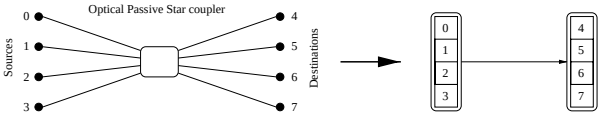


**Figure 3. Modeling an OPS by a hyper-arc.**

We refer the interested reader to [2], where modeling POPS by stack-graphs can be found, as well as further applications.

## 2.4 Kautz graphs

Stack-graphs represent a powerful tool for modeling multi-OPS networks and also for guiding their design. Indeed, we can use them to build one-to-many lightwave networks based on graphs having good properties, like the connection of a large number of nodes with respect to the constant degree and a small diameter. The Kautz graph has such good properties and is defined as follows.

**Definition 2** *[9] The* **directed Kautz graph** *$K(d, k)$ of degree $d$ and diameter $k$ is the digraph defined as follows (see Figure 4).*

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

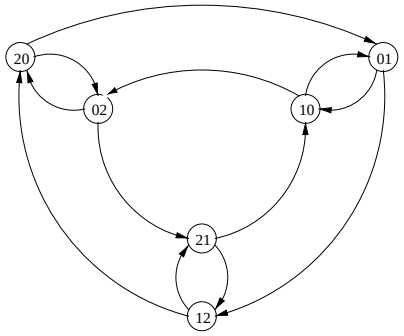


**Figure 4. The Kautz graph $K(2,2)$.**

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

The Kautz graph $K(d,k)$ has $N = d^{k-1}(d+1)$ nodes, constant degree $d$ and diameter $k = \lfloor \log_d N \rfloor$. It is both Eulerian and Hamiltonian and the best known with respect to the number of nodes if $d > 2$ [9]. As an example, $K(5,4)$ has $N = 3750$ nodes, degree $5$ and diameter $4$.

It is important to note that routing on the Kautz graph 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 [7, 12].

# 3 A multi-hop multi-OPS network

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

Let the **reflective Kautz graph** $K_r(d,k)$ be the Kautz graph $K(d,k)$ of degree $d$ and diameter $k$ in which we add an arc from every node to itself (loop). The number of nodes in $K_r(d,k)$ is the same as in $K(d,k)$, each node has degree $d+1$, and the node labels are the same. A reflective Kautz graph has the same properties as a Kautz graph $K(d,k)$.

Thus, we can define the stack-Kautz graph as follows.

**Definition 3** *The* **stack-Kautz graph** $SK(s,d,k)$ *is the stack-graph* $\varsigma(s, K_r(d,k))$ *of stacking-factor* $s$*, degree* $d+1$ *and diameter* $k$.

The *stack-Kautz* network has the topology of the stack-Kautz graph $SK(s,d,k)$ and $N = sd^{k-1}(d+1)$ nodes (see Figure 5). Each node is a processor labeled by a pair $(x,y)$ where $x$ is the label of the stack in $K(d,k)$ and $y$ is an integer $0 \leq y < s$, i.e., $x$ is the label of a processor group and $y$ is the label of a processor in this group. Since the stack-Kautz network inherits most of the properties of the Kautz graph, like shortest path routing, fault tolerance and others, we chose it as a good candidate for the topology of an OPS-based lightwave network. In the following we will investigate further these properties.

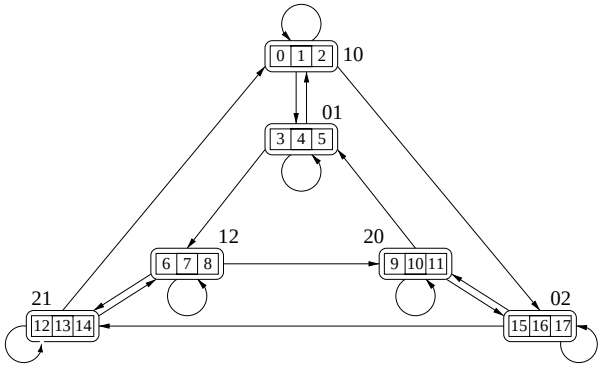


**Figure 5. The stack-Kautz network $SK(3,2,2)$.**

## 3.1 Characteristics

An OPS-based network with the topology of a stack-Kautz network $SK(s,d,k)$ has $N = sd^{k-1}(d+1)$ processors 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, $SK(12,5,5)$ has $N = 45000$ processors and diameter 5.

Each group of $s$ processors 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.

Each processor has one transceiver per link. So there are $d+1$ transmitters and receivers per processor and a total of $sd^{k-1}(d+1)^2$ transceivers in the network.

For the sake of illustration, the network $POPS(t,g)$ with $g$ groups of size $t$ has $g^2$ OPS couplers for $N = tg$ processors and each processor has $g$ transmitters and receivers. It is clear that, for a same number of nodes, the number of OPS couplers in $SK(s,d,k)$ is smaller than in $POPS(t,g)$ and analogously for the individual and total number of transceivers.

## 3.2 Power budget and scalability

The power budget corresponds to the cost of sending a message from one processor to another, in terms of energy. The goal is to minimize this value, that roughly equals the

number of OPS couplers crossed by the message times the degree of each coupler (i.e., $sk$ in the stack-Kautz).

In order to proportionally decrease the power budget in a stack-Kautz network $SK(s, d, k)$, 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 processor. Thus, by increasing the diameter of the network, the power budget and the resources are proportionally reduced with respect to the number of processors. However it is necessary to preserve $s \geq d$ to have more processors in the network than OPS couplers.

Table 1 gives numerical evidence of the resources required by the two multi-OPS networks (POPS and stack-Kautz) in order to interconnect 1800 processors with 900 OPS's.

| | $SK(12, 5, 3)$ | $POPS(60, 30)$ |
|---|---|---|
| Groups | 150 | 30 |
| Processors | 1800 | 1800 |
| Degree OPS's | 12 | 60 |
| OPS's | 900 | 900 |
| Tr.\|Rec. per proc. | 6 | 30 |
| Tr.\|Rec. total | 10800 | 54000 |
| Power budget | 48 | 60 |
| Broadcast (time steps) | 4 | 2 |
| Advanced control (# of bits) | 106 | 561 |

**Table 1. Examples of required resources.**

### 3.3 Control protocols

A single wavelength OPS coupler can transmit only one message per communication step. Since $s$ processors share $d + 1$ OPS couplers, $s \geq d$, an efficient control protocol is required. One such protocol was proposed in [5] for the POPS network $POPS(t, g)$ with $g$ groups of size $t$. It supposes that a processor can receive messages on all its receivers at the same time and that it can send a message on only one link per communication step. Each group of $t$ processors contains one node in charge of the control of the group. The total bit complexity of the control protocol is $t \log g + g \log t + t + g$ bits.

For the control protocols of our mono-wavelength stack-Kautz network, we suppose, as in [5], that a processor can receive messages on all its links at the same time. By reading the header of a message, a processor can decide whether it has to process it or not. The control protocol has thus just to avoid local conflicts, inside a group of processors.

A **simple control protocol** can be implemented in any multi-OPS network with a bit complexity of $s \log(d+1)+s$ bits: The processor which is in charge of the control of the group has $s$ counters, one for each processor. Each counter is increased of 1 when the corresponding processor receives a refusal. A counter is set to 0 when the corresponding processor receives an acknowledgment. The larger the counter value, the higher the corresponding processor priority. Its time complexity for a group of size $s$ and degree $d + 1$ is $O(s)$.

An **advanced control protocol** for multi-OPS networks can be considered under the following hypothesis.

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

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

Let $p_0$ be the processor in charge of the control of a group of $s$ processors. The protocol is as follows.

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

The bit complexity of this control protocol is $s(d+1) + s \log(d+2)$ bits. Since the time complexity of the maximum matching algorithm is $O(s(d + 1)^2)$, we have

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

## 4 Simulation

We built a simulator for the stack-Kautz network which implements a shortest path routing algorithm and guarantees that a path is shorter or equal to the diameter of the network. Our simulator can use either the simple or the advanced control protocol.

We kept the load $\gamma$ of the network $SK(12, 5, 3)$ at $\gamma = 0.5$ by "injecting" new messages, during 1000 communication steps. Figure 6 shows the accumulated percentage of

delivered messages out of the total number of messages (left curve), as a function of the number of steps needed, as well as the total number of the delivered messages (right curve). 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, the advanced control protocol is much better than the simple control protocol, with respect to the number of delivered messages.

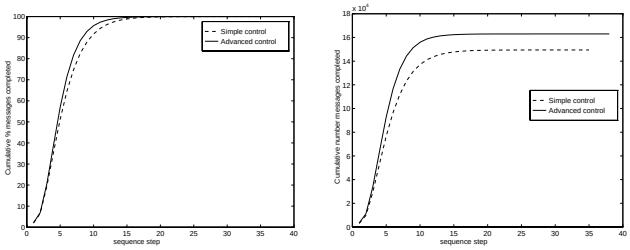


**Figure 6. Cumulative percentage and number of messages completed.**

Finally, it is also interesting to study the load of the networks $SK(12,5,3)$ and $SK(12,5,2)$, when the probability of having a new message is sharply increasing. This models, for instance, cases where a global exchange is performed in the middle of a normal state of the network.

In Figure 7, the load $\gamma$ of the network is induced by the probability $p$ that a processor creates a new message at every step. For the curve on the left, we set $p = 0.1$ 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.

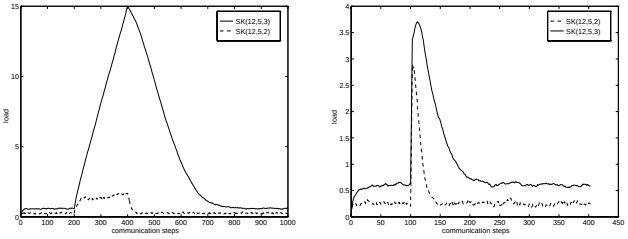


**Figure 7. Network load evolution for $SK(12,5,2)$ and $SK(12,5,3)$, when the probability of message creation changes.**

## References


[1] C. Berge. *Hypergraphs*. North Holland, 1989.

[2] P. Berthomé and A. Ferreira. Improved embeddings in POPS networks through stack-graph models. *Third Internationnal Workshop on Massively Parallel Processing using Optical Interconnections,pp. 130-135*, July, 1996. IEEE Press.

[3] M. Blume, F. McCormick, P. Marchand, and S. Esener. Array interconnect systems based on lenslets and CGH. *SPIE International Symposium on Optical Science, Engineering and Instrumentation, Paper 2537-22, San Diego (USA)*, 1995.

[4] H. Bourdin, A. Ferreira, and K. Marcus. A performance comparison between graps and hypergraph topologies for passive star WDM lightwave networks. *Computer Networks and ISDN Systems*, To appear. A preliminary version appeared in the *Proceedings of the 2nd IEEE MPPOI*, pp 257-264.

[5] D. Chiarulli, S. Levitan, R. Melhem, J. Teza, and G. Gravenstreter. Partitioned Optical Passive Star (POPS) Topologies Multiprocessor Interconnection Networks with Distributed Control. *IEEE Journal of Lightwave Technology*, 14(7):1601–1612, 1996.

[6] A. Ferreira. Towards effective models for optical passive star based lightwave networks. In P. Berthomé and A. Ferreira, editors, *Optical Interconnections and Parallel Processing: Trends at the Interface*. Kluwer Academic Publisher, Boston (USA), 1997.

[7] M.A. Fiol, J.L.A. Yebra, and I. Alegre. Line digraphs iterations and the (d,k) digraph problem. *IEEE Trans. on Computers C-33, 400-403*, 1984.

[8] S. Gardner, P. Harvey, L. Hendrick, P. Marchand, and S. Esener. Photorefractive Beamsplitter For Free Space Optical Interconnections. *SPIE Photonics West'97, Optoelectronics'97, Diffractive and Holographic Optics Technology IV*, San Jose (USA), February 1997.

[9] W.H. Kautz. Bounds on directed (d,k) graphs. Theory of cellular logic networks and machines. *AFCRL-68-0668 Final report, 20-28*, 1968.

[10] B. Mukherjee. WDM-Based Local Lightwave Networks Part I: Single-Hop Systems. *IEEE Networks*, 6(3):12–27, may 1992.

[11] B. Mukherjee. WDM-based local lightwave networks part II: Multihop systems. *IEEE Networks, pp. 20-32*, July 1992.

[12] G. Smit, P. Havinga, and P.Jansen. An Algorithm for Generating Node Disjoint Routes in Kautz Digraphs. *Proceeding Fifth International Parallel Processing Symphosium, pp. 102-107*, 1991.

[13] F. Sugihwo, M. Larson, and J. Harris. Low Threshold Continuously Tunable Vertical-Cavity-Surface-Emitting-Lasers with 19.1nm Wavelength Range. *Applied Physics Letters 70*, page pp. 547, 1997.