Communications and Network, 2014, 6, 29-42
Published Online February 2014 (http://www.scirp.org/journal/cn)
http://dx.doi.org/10.4236/cn.2014.61005

Scientific Research

Deployment of the Content-Based Switching Network

Christophe Michard, Yosuke Tanigawa, Hideki Tode
Department of Computer Science and Intelligent Systems, Osaka Prefecture University, Sakai, Japan
Email: michard@com.cs.osakafu-u.ac.jp, tanigawa@cs.osakafu-u.ac.jp, tode@cs.osakafu-u.ac.jp

Received October 20, 2013; revised November 20, 2013; accepted November 27, 2013

ABSTRACT

In this paper, we propose a hybrid network architecture, called Content-based Switching Network (CSN), and its signaling scheme, which addresses the issues inherent to conventional hybrid networks which implement a horizontal separation over the entire network (from edge to edge). We will show how CSN nodes can flexibly choose their switching paradigm (store-and-forward, optical bypass, electrical bypass) during a path establishment. Contents being transferred in one piece from end-to-end, the concept of packet can be eluded in our network, and, in particular, the user is able to avoid complicated transport layer functions, like TCP, if they are not essential. In CSN, very large contents have a special status, since they cannot be store-and-forwarded. We will show how the resource management has been designed in order to deal with such contents. A section is dedicated to deployment and feasibility issues. Simulation results will show that CSN can successfully transfer contents at 1 Gbps and 10 Gbps, the maximum speed being limited by the state-of-the-art device technologies when buffering is required (memory speed), while no major limit is observed in the case of all-optical transfers other than the optical fiber speed. Other results concern the deployment of CSN from an unclean slate approach. They will show how beneficial can be the deployment of CSN from an Optical Circuit Switching network.

KEYWORDS

Content-Based Switching Network; Optical Network; Wavelength Division Multiplexing; Hybrid Network

1. Introduction

Future networks will have to face a traffic implosion due to, for instance, the transmission of ultra high definition videos (4K, 8K) or the next-generation 3D and holographic imaging [1]. All-optical solutions have been proposed to provide very high capacity networks able to deal with such a demand. However, the actual state of the art is not mature enough to provide these attractive solutions. Translucent networks, allowing optoelectronic conversions at some points of the network, are now considered as the most viable solutions. In these networks, it is essential to minimize the amount of optoelectronic conversions. This problem is usually related to the knowledge of *where* conversions should be achieved in the network. As for the transmission of huge contents, we can distinguish two main applications, each one using a different kind of network. The first one is the streaming. Since it is a real-time application, Optical Circuit Switching (OCS) [2] is classically preferred for keeping a high Quality of Service (QoS). The second one concerns data exchange (between data centers for instance). In such a case, Optical Packet Switching (OPS) [3] is desirable. Existing hybrid network solutions [4,5] consist of a circuit-switching plane and a packet-switching plane in order to deal with both applications. Nevertheless, the network should decide through which one of them a content will be transmitted before the transmission. This rigid separation is not effective enough to deal with the increasing demand and the massive fluctuations of the forthcoming traffic.

We propose in this paper a hybrid network, called Content-based Switching Network (CSN), which can be considered as a fusion of OCS and store-and-forward based switching networks. Unlike the aforementioned clear plane separation solutions, each one of the CSN nodes has the capability of flexibly choosing its switching paradigm for a specific transfer, depending on its status, and reserves the resources accordingly. Switching paradigms are optical bypass, electrical bypass, and store-and-forward (description are given in Section 3.1). This leads to a network which is able to dynamically adapt itself to the demand.

The main benefits of our proposal are as follows. Under low traffic load condition, CSN can use a circuit-switching approach for all contents, which is the best if resources are available, leading to a better use of the resources of the network. Under higher traffic condition, each node can perform an OEO conversion on the fly for a specific content if needed (resource shortage, physical impairments), solving the questions of *when* and *where* a translucent bridge should be made. This virtually maximizes the merit of the optical transmission, in contrast to usual packet-switching networks which forcibly buffer the content at each node while the header is checked, introducing undesirable latency. From the viewpoint of service provisioning, real-time capabilities are fundamental, for instance for online games or video on demand services. If a specific content does not require these capabilities, CSN can provide delay tolerant services if needed. Finally, since in CSN a content is always transferred in one piece, the user does not need to worry about cumbersome packet interleaving.

The rest of this paper is organized as follows. In Section 2, we compare CSN to related works. In Section 3.1, we make a global description of the network. Section 3.2 introduces its signaling. In Sections 3.3 and 3.4, we explain how nodes are working from their internal point of view. In Section 4, deployment issues of the Content-based Switching Network are highlighted. Section 5 presents simulation results which concern the deployment of CSN from an OCS network (unclean slate approach), and show the superiority of our proposal compared to OCS in both 1 Gbps and 10 Gbps network speed scenarios.

2. Distinction with Other Works

2.1. Common Technologies

OCS is probably the best network when resources are available. When an end-to-end path is reserved, best throughput and QoS can be obtained. Nevertheless, its major weakness actually lies in the fact that end-to-end transfers are the only ones available. This lack of flexibility makes this network weak regarding the granularity. Additionally, long routes are penalized, since it is more difficult to reserve resources for them when the traffic increases, than for short ones: OCS lacks scalability. CSN tries to address this issue by offering intermediate buffering capabilities when resources are not sufficient. Basically, CSN tries to match OCS as much as possible when the traffic is low, making their throughputs very similar. The more the traffic is increased, the more buffering capabilities of CSN are used. Therefore latency tends to make the throughput of longer routes worse than in the case of OCS. Nonetheless, even if this is true for an individual successful transfer, it is not the case when we take path establishment rejections into account. For a certain amount of time, CSN is able to transfer more contents than OCS, even if the transfer speed of a particular content transferred through CSN is slower than the one of a content transferred through OCS. OCS does not transfer a content if the throughput cannot be maximum, while CSN accepts it if possible at a slower transfer speed, making a better use of the network resources for higher traffic.

Usual packet-switching networks forcibly perform header check and buffering at each node. On the contrary, CSN is flexible and selects the appropriate switching paradigm dynamically, maximizing the optical transmission capability if possible. In case of resource shortage, CSN can still choose between electrical bypassing and store-and-forwarding, while packet-switching networks always uses store-and-forwarding. CSN also allows partial buffering (a content is forwarded as soon as the subsequent path is achieved), reducing the effect of this particular latency. Finally, CSN allows the user not to divide a content in many packets as in packet-switching if its functions are not needed, avoiding its common issues: overhead and interleaving. We are not against the fact of switching packets, and even if we would not recommend it in our network, CSN can transfer packets. Nevertheless, we think that the aforementioned issues inherent to common packet-switched networks could be avoided. The Content-based Switching Network embeds capabilities to dynamically adapt its behavior to the demand, acting as OCS when resources are available, and more like packet-switching networks when the traffic is too high. Additionally, it offers a specific way to transfer very large contents, as described in Section 3.4 of the paper.

More precisely, as for the size of the contents, we think that small packet transfers as the ones performed in actual IP networks should be avoided in CSN, because of their granularity and the waste of bandwidth they implies. On the contrary, due to its content abstraction, the Content-based Switching Network provides a way to avoid complicated transport layer functions, like TCP [6,7] if they are not essential. However, CSN could be used to transfer packet bursts [8] when small packets cannot be avoided. CSN does not encapsulate contents, letting the user the choice of the complete nature of the content. If target users are end-users, details should be delegated to the application layer, as in P2P applications [9] for instance. If users are data centers, contents could be a part of a file, a file or several files, avoiding the need of creating a new path for each file in the latter case.

Now, we would like to discuss OPS. As written above, OCS and CSN are closely related. CSN tries to act as OCS as much as possible. If all-optical transfers cannot be achieved, either intermediate buffering or electrical bypassing is *adaptively* performed in as least nodes as possible. Consequently, we could see CSN as a network between OCS and store-and-forward based switching networks, with the condition of matching OCS behavior when possible, *i.e*. maximizing the merit of optical transmission. In order to do that, the CSN signaling reserves a path before transferring data (two-way signaling). Due to their different signaling approaches, the direct comparison between CSN and OPS is difficult. On the contrary, it is easier for us to compare CSN to its target technology, OCS, and we focused this paper on their comparison under similar signaling conditions.

2.2. Hybrid Technologies

CSN is more likely to be compared to other hybrid approaches. For instance, in [4,5] such a system was developed. In this Miyazawa and Furukawa's work, according to the network state, a content is transferred using packet- or circuit-switching from end to end. Resources are split, basically dedicated for one plane or the other, with additional ones shared by both planes in order to adapt the efficiency of a particular node to the traffic load. The idea is attractive, but is an all-optical approach, which leads to issues related to the actual efficiency of optical buffering [10]. The physical (therefore static) separation of the resources does not allow the network to get rid of the disadvantages of each plane: no intermediate buffering in circuit-switching, overhead and interleaving in packet-switching. CSN allows a smoother approach. In CSN, hardware resources are not split, but their attribution can be freely and dynamically assigned to any kind of application during the network use, depending on its policy (e.g. the shared/large wavelengths attribution mechanism presented in Section 3.4). Hardware resources are not reconfigured at all. This behavior allows each node, individually, to choose the switching paradigm adapted to its resources.

We only found a unique approach similar to ours. That emphasizes the novelty of CSN. In [11-13], the authors use OpenFlow [14] in order to unify the packet- and circuit-switching planes. They effectively blur the distinction between packets and circuits, viewing them as flows of different granularity in a flow-switched network. Information needed for switching are gathered into a centralized point, called unified-control-plane (UCP), which allows them to easily reserve paths according to users requirements (bandwidth, QoS, etc.) and the status of the entire network. Additionally, they can dynamically reassign an optical path to a better one during the transfer if a circuit is transferring a packet-flow. Their work is mainly focused on the development of IP networks, and they remarkably simplified the path and service managements. Their design methodology is detailed and comprehensive in its steps, but their extensions to the protocol for circuit-switching are experimental and not comprehensive, since they are considering the transfer of IP packets more than any other kind of content. For example, Optical Transport Network (OTN) switching is not supported and wavelength switching has only rudimentary support. They are actually expecting that with the formation of the Open Networking Foundation, the ability to control all kinds of circuit switches will eventually be included in the OpenFlow specifications as well. Contrary to OpenFlow, our control plane is not centralized, and simulation results given in Section 5 will demonstrate the fairness of CSN regarding the route length, making our proposal suitable for large-scale networks. Also, we are focusing on circuit-switching and the transfer of entire contents, discouraging the use of packets. That is why wavelength switching is fully supported in CSN, and why we also allow electrical bypassing in order to deal with physical impairments for long-reach transmissions.

3. Description of the Network

3.1. Overview

The Content-based Switching Network differentiates three different entities, which are shown in Figure 1.

- **Users** exchange contents thanks to CSN. The National Institute of Information and Communications Technology (NICT), Japan, assumes that all the end-users of the future Internet should be equipped with all-optical interfaces. We will make the same assumption in this paper. Target users in this paper are not limited to end-users, but could also be nodes from another network, or data centers. It depends on the use of the network;
- **Access nodes** are the interfaces between the Users and the Core nodes. They are in charge of the flow control of the incoming transfer requests from the Users. Contrary to TCP/IP networks which accept all incoming packets, CSN performs a content admission control;

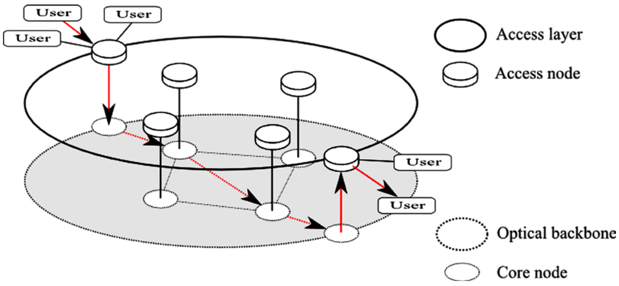

Figure 1. **Overview of the CSN network.**

- **Core nodes** are in charge of the resources reservations and the content switching. Considered resources are wavelengths, memory space, memory input speed and memory output speed (RAM devices are used for store-and-forwarding). A control plane is dedicated to the reservations, while contents are transparently (*i.e.* as they are) transferred through a content plane.

As for the switching paradigms, three different ones are available in each Core node. Figure 2 summarizes all of them ("S" stands for *Source*, while "D" stands for *Destination*). Figure 3 shows their implementation in a Core node.

1) **Optical bypassing (OB)** is achieved using an optical switch. The content is transferred through a node using the same wavelength;

2) **Electrical bypassing (EB)** is achieved using an electrical switch. The content is OEO converted in order to change the wavelength on the fly, avoiding buffering. It can be done to extend a path or in order to deal with physical impairments (signal regeneration);

3) **Store-and-forwarding (S&F)** is achieved using RAM devices in the nodes in order to buffer the incoming content until the path is extended. An electrical switch is needed to send the content to the output.

3.2. Signaling

3.2.1. Four Phases

The CSN signaling is based on the two-way signaling. A control message is sent from the source in order to reserve the lightpath and set the switches before the content transmission. Control and content planes are separated. Once a path is created, the content can be transferred in one piece, without any inspection. Security details are delegated to the application layer if the user fears any data interception. Control messages carry different kind of information, such as a transfer id, the content size, the source and destination addresses as well as the followed route, the gathered wavelengths availability, the chosen wavelengths and fibers for the reservations. Additional information can be transferred in order to achieve various services such as enhanced QoS for instance. The sender can specify timeouts in order to allow CSN to safely drop a content if the complete path reservation is too long to be performed. We distinguish four different phases in the signaling. Figure 4 illustrates them in a simple way.

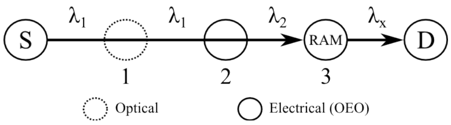

Figure 2. Available switching paradigms in each core node.

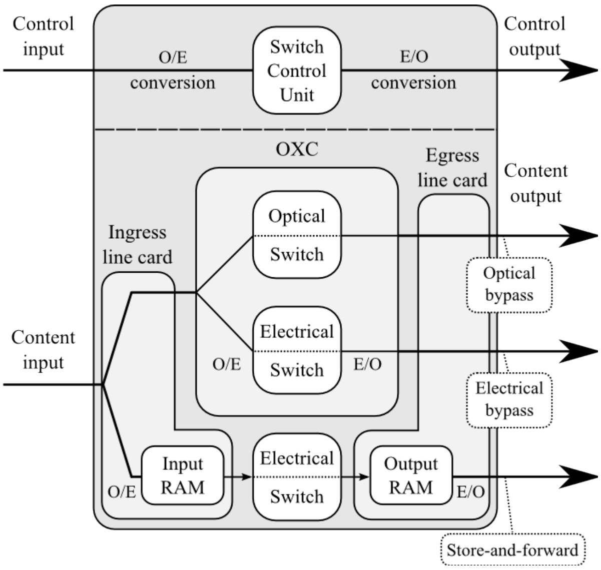

Figure 3. Switching architecture of a core node.

1) **Data collection**—During this phase a control message is sent in direction of the destination. Resources availability is checked, and data (resource availability, content status) are merged into the message;

2) **Resource reservation**—When the previous phase finishes, a control message is sent back in direction of the source. Each time it reaches a node, resource reservation and switch setup are performed depending on the previously collected data. Each intermediate node decides which switching paradigm will be used independently of the others;

3) **Content transfer**—A content transfer is initiated as soon as a path is created;

4) **Resource release**—Once the content transmission is over, a control message is sent from the sender in order to release the path.

The choice of the switching paradigm is not made before, but *after* the data collection phase. That is why it is easy to choose the right paradigm, and makes CSN different from usual hybrid networks.

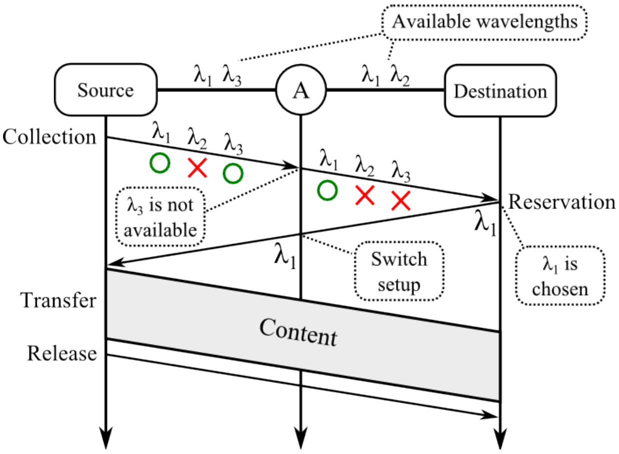

Figure 4. Optical bypassing is set up in node A.

3.2.2. Content Status

In order to avoid content dropping in S&F, memory resources are reserved during a path establishment. For instance, if the content size is 1 MiB, 1 MiB of memory will be reserved beforehand. In OCS, where S&F is not considered, the longer the route is, the less the probability of establishing an end-to-end path is, by lack of wavelengths. Since they cannot be store-and-forwarded, huge contents (e.g. 1 TiB) require a special attention in CSN. We address this issue by distinguishing two kinds of contents: the *small* ones, and the *large* ones. The content status is dynamically determined during the data collection phase of the CSN signaling. Basically, some wavelength resources are dedicated to large contents, and we try to establish an end-to-end path at all costs. Details on wavelength resources attribution are given in Section 3.4.

The content status is set to *unknown* when the data collection phase starts. Each time the control message reaches a node, memory resources are checked in order to know if the content could be buffered or not. If it can be buffered, the content status is updated to *small*. If a wavelength discontinuity is detected, the status is updated to *large*. Otherwise, the status remains *unknown*. At the end of the data collection phase, if the status is still *unknown*, it will be automatically set to *large*, as S&F cannot be achieved in any node of the path. The signaling tries to achieve OB at each node if possible. When resource shortage (no output wavelengths or wavelength discontinuity) is detected, the signaling has a different behavior depending on the content status.

- **Large**—We try to achieve an end-to-end path by any means. During the data collection phase, when a wavelength discontinuity is detected, we decide which wavelength will be used for the path until this node (first-fit), and the data collection continues. During the resource reservation phase, EB will be chosen at the node where the discontinuity was detected. If at a point no output wavelengths are available at all, a rejection will occur: a message will be sent to notify the source;
- **Small**—EB is not allowed for small contents (reasons are given in Section 3.2.4). Consequently, wavelength discontinuity is not allowed during the data collection phase. When either wavelength discontinuity is detected or no outputs wavelengths are available, the data collection phase ends, and the resource reservation phase is initiated. S&F is set up if possible, and the path reservation can continue. If not, a control message (*alternative S&F reservation*) is sent to the previous node where S&F was possible (we only remember the last available location in the control message, not all of them). If S&F cannot either be done in that alternative node, rejection occurs.

The alternative S&F reservation mechanism was designed for *medium*-sized contents which are detected *small* in a non congested part of the network, but whose reservation always fails because of a congested part of the network. Doing this way, they could enter the network via S&F in an uncongested part of it, then be detected *large* later by a congested part of the network and use the dedicated wavelengths for large contents in order to achieve their transfer.

3.2.3. Paradigm Selection

During the reservation phase, the control message carries which wavelength(s) should be used along the path. In practice, we only remember where (which node) a discontinuity occurs, and the associated wavelengths. Each time the message reaches a node, this particular node decides which paradigm will be used. If input and output wavelengths are the same, OB is chosen. If input and output wavelengths are different, EB is chosen. If the message is an *alternative S&F reservation* one, S&F is chosen.

In practice, the choice is very simple and made as late as possible. If a particular reservation cannot be achieved because a particular resource is not available anymore when the control message reaches the node, a rejection occurs. In the case of S&F, the intermediate node acts like a source access node: path establishment request management. Depending on the quickness of the subsequent path reservation, this intermediate source could release memory resources and becomes a simple OB or EB node. **Figures 4-6** show different cases of the signaling. Only one intermediate node is shown for simplicity. Each node between the source and node A, as well as between node A and the destination could have chosen any of the three proposed switching paradigms.

Figure 4 shows the case matching OCS, where OB can be set up. During the collection phase, λ_1 is available at the input and the output of node A. At the beginning of the reservation phase, λ_1 is chosen (first-fit) to establish the lightpath. When the control message reaches node A, since λ_1 is still available, OB is set up and the message is forwarded to the source. Then, the content transfer can be successfully initiated. Resources are released when the transfer ends.

Figure 5 shows a common S&F case. A wavelength discontinuity is detected at node A. If the content status had been *large*, EB would have been allowed in node A. In such a case, λ_1 is chosen for the previous path, and the collection can resume similarly to **Figure 4**. In **Figure 5**, node A has enough resources to buffer the content, and consequently the content status is *small*. Memory resources are reserved, and a control message is sent in direction of the source to reserve the path from the source to node A (subpath I). At this point, node A has become an intermediate source from the point of view of

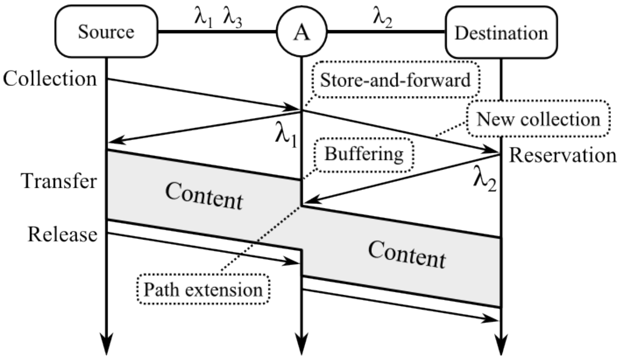

Figure 5. Store-and-forwarding is set up in node A.

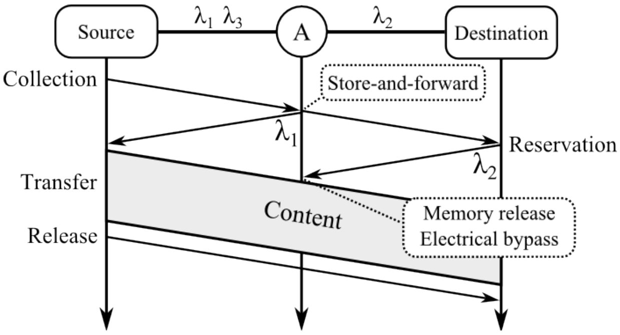

Figure 6. Shift from store-and-forwarding to electrical bypassing in node A.

the signaling. Node A initiates a new collection phase in order to establish a path until the destination (subpath II). The content status is set to *unknown*, and a control message is sent to the destination. When subpath II is achieved, node A has already received some bits of the contents, for instance 0.2 MiB out of 1 MiB. Node A releases the unused 0.8 MiB of reservation, and starts to forward the content before its complete reception. Once the source finished the content transmission, a resource release message is transmitted to node A in order to release subpath I. The same behavior occurs for node A and subpath II when the content forwarding is achieved.

In Figure 6, subpath II is achieved before any bits of the content have been received. In this case, memory resources can be fully released, and both subpath I and II can be glued together using EB.

3.2.4. Small Contents and Electrical Bypassing

In this section, we will give the reasons why, in the small contents case, S&F is chosen instead of EB during the data collection phase when a wavelength discontinuity is detected.

The main reason why S&F is favored is related to resource management. Let us suppose that in Figure 4 electrical bypass is chosen at node A, and let us compare that figure to Figure 6. In the latter, EB is eventually used. The difference lies in how long resource reservation for the path from node A to the destination lasts. When S&F is selected, content transmission can be initiated sooner, leading in a better utilization of the resources. Consequently, when the traffic increases, the network has potentially more wavelengths available: OB is more likely to occur for other requests. In contrast, if we had preferred EB, it would have happened at each hop, leading to resource wasting (reservation duration) and increased energy consumption. Finally, the longer the route is, the less the probability of successfully reserving a path is. For higher traffic, even if we can select EB at each node during the data collection phase, we are not guaranteed to be able to reserve the selected wavelengths during the resource reservation phase if the route is long. Favoring EB practically decreases the scalability of our network, which is unwanted.

3.3. Nodes Internal

Access nodes are in charge of the flow control of the incoming transfer requests as shown in Figure 7. A transfer request from a User is added in a backlogged queue, waiting to be handled. The backlogged queue is critical in order to avoid Denial of Service when too many requests are made in a short period of time. Additionally, the queue itself can be managed as a priority queue based on QoS or service attributes if needed. Once a path has been established by the Core nodes, the corresponding content can be transferred. At the end of the transfer, the *active* request is deleted and a new request waiting in the queue can be handled. The maximum amount of active requests is limited to the total number of available wavelengths (we cannot transfer simultaneously more contents than the total number of wavelengths). This is the criterion chosen for the simulations presented in this paper.

If a path reservation fails, the request will be added at the end of the queue after a certain delay, calculated as follows:

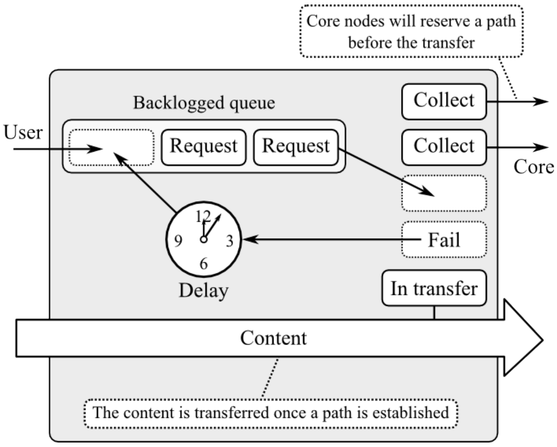
Core nodes will reserve a path before the transfer
Backlogged queue
Collect
Collect
User
Request
Request
Core
12
9
3
6
Fail
Delay
In transfer
Content
The content is transferred once a path is established

Figure 7. Transfer request management in an access node.

$$\Delta t_{\text{add}} = \frac{T_{\text{sender time out}}}{N_{\text{attempts}}} - EWMA_{\alpha}\left(\Delta t_{\text{queueing}}\right) \qquad (1)$$

Δt_{add} is the calculated delay. $T_{\text{sender time out}}$ is a time out delay fixed by the sender of the content. It allows the access node to drop the request if the path creation requires too much time. N_{attempts} is the maximum number of data collection attempts the access node can initiate. $\Delta t_{\text{queueing}}$ is the queuing delay of a request. Here, we compute the average of this delay using the EWMA (Exponentially-Weighted Moving Average), defined by

$$\begin{cases} EWMA \leftarrow \alpha \cdot \Delta t_{\text{queueing}} + (1-\alpha) EWMA \\ 0 < \alpha \le 1 \end{cases} \qquad (2)$$

The new average is equal to the weighted new queuing delay plus the weighted previous average. The smaller α is, the less the average takes into account the new delay. On the contrary, the larger α is, the quicker old values are forgotten.

Core nodes consist of a control plane and a content plane. The control plane is in charge of the path establishment. Some wavelengths are dedicated to the communication between nodes. The content plane is used for transferring the contents as they are without inspection. This plane consists of ingress/egress line cards, an OXC for optical and electrical bypassing, a separated electrical switch which interconnects the RAM devices of the line cards in the case of store-and-forwarding, add/drop multiplexers and several additional devices which are usually needed to achieve WDM. **Figure 3** summarizes the path taken by a content in a Core node depending on the chosen switching paradigm.

Similarly to the Access nodes, Core nodes implement a specific request management in the case of store-and-forwarding. Since an intermediate store-and-forwarding node acts as a new source from the signaling point of view, the mechanism is similar to the backlogged queue presented above, but renamed *forwarding queue*. There is one forwarding queue by output. The adding delay is calculated similarly to the one presented previously, except that we use the receiver time out instead of the sender one. The receiver time out is the delay the network has to be able to send the first bit of the content to the destination. If the time out is reached before the first bit has been received, the content can be safely dropped from the network. The receiver time out should also be transmitted by the source. If not, the owner of the access node can arbitrarily establish a policy in order to determine it.

3.4. Dealing with Large Contents

In Section 3.2.2, we distinguished two kinds of contents, the *small* ones and the *large* ones. We also mentioned that some wavelengths are dedicated to the transfer of *large* contents, in order to increase their probability of path reservation, since S&F is not allowed for them. In this section, we will explain how resources are shared out.

Wavelengths are basically divided into two groups. The ones dedicated to *large* content transmissions, and the shared ones which can be used for any transmission (*large* ones included). The objective is to guarantee a minimum amount of wavelengths dedicated for *large* contents, and adjust this value over the time, depending of the nature of the traffic. In parallel, we also guarantee a minimum amount of wavelength which can be used for any transmission. Each time a reservation is attempted in a node, the thresholds of guaranteed wavelengths are adjusted according to these formulas:

$$R_{\text{large}} \leftarrow \frac{R_{\text{large}} \cdot T_{\text{window}} + L}{T_{\text{window}} + \delta t_{\text{update}}} \qquad (3)$$

$$R_{\text{shared}} \leftarrow \frac{R_{\text{shared}} \cdot T_{\text{window}} + S}{T_{\text{window}} + \delta t_{\text{update}}} \qquad (4)$$

R_{large} is the ratio of wavelengths dedicated for large transmissions. R_{shared} is the ratio of wavelengths which can be used for any transmission. They are determined using a moving average. T_{window} is the size of the window, in unit of time. L is equal to 1 (max. ratio) when the reservation attempt concerned a *large* content, 0 (min. ratio) otherwise. S is the contrary. Finally, δt_{update} is the difference between the time of the reservation and the last update time. It represents the window shift. The more path reservations for large contents occur, the more R_{large} increases. These ratios need to be normalized:

$$R_{\text{large}} \leftarrow \frac{R_{\text{large}}}{R_{\text{shared}} + R_{\text{large}}} \qquad (5)$$

$$R_{\text{shared}} \leftarrow \frac{R_{\text{shared}}}{R_{\text{shared}} + R_{\text{large}}} \qquad (6)$$

Finally, we can update the thresholds:

$$N_{\text{large}} = R_{\text{large}} \cdot N_{tot} \qquad (7)$$

$$N_{\text{shared}} = R_{\text{shared}} \cdot N_{\text{tot}} \qquad (8)$$

Naturally, the results are averaged to integers. N_{tot} is the total number of available wavelengths between the current node and the subsequent one.

Until now, the ratios could diminish to 0. We still need to set a minimum value for both ratios. Let us assume that the threshold is 5%. In such a case, if R_{large} is less than 5%, we would update the previous thresholds using these formulas:

$$N_{\text{large}} = 0.05 \cdot N_{\text{tot}} \qquad (9)$$

$$N_{\text{shared}} = 0.95 \cdot N_{\text{tot}} \tag{10}$$

Depending on the application of the network, transmitted contents could be mostly *small* or *large*. The wavelength management presented in this section essentially concerned the nodes where the traffic is high and which are closed to be congested. In other nodes, since wavelengths are not congested at all, we would always have enough wavelengths to deal with the traffic. It would also be great to decrease R_{large} through the time when no reservation has been made for a while (reset to a default value).

4. Applications and Deployment

4.1. Applications and Services

The Content-based Switching Network has been designed to be versatile, therefore, we will only give here three different examples of use we could made to show how the network implementation can be extended in order to implement additional services. The first example concerns content caching, while the second one deals with virtual networks and topology modification. The last one discuss a method to accelerate the transfer of large contents.

Content-centric networks (CCN) [15] are specific networks where the users do not know where the requested files are. A request is sent to the access node, which will ask where the file is stored: the closer, the better. Being able to consider files in their entirety, the proposal, Content-based Switching Network, could be adapted for this kind of networking. Connecting access nodes to shared memories would allow CSN to create cache nodes for popular contents (overlay using line cards). Figure 8 shows how it could be used in CCN. A user sends a request for a specific file. Usually, this file is in a server at Location B. Transferring it from there to Location A would require core resources. Hopefully, the requested content is popular and was already cached in the shared memory of Location A: the content is directly transmitted to the user, saving precious core resources. Naturally, apart from content-centric networking, cache nodes are popular for a lot of applications, like Video on Demand for instance. In this example, we can see that services could be implemented at the access node level. The access nodes would then keep their role of traffic regulator.

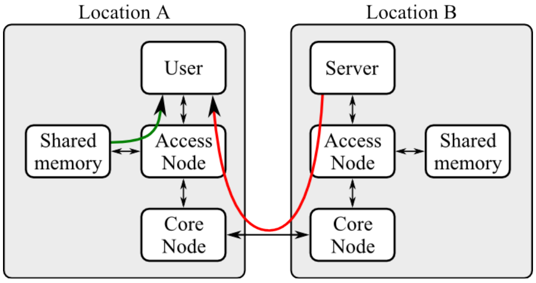

Figure 8. Content caching.

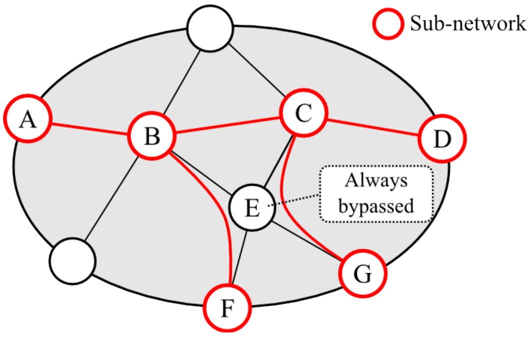

Figure 9. Topology modification.

Figure 9 shows how a network topology can be easily modified thanks to CSN. Let us assume that line cards linking nodes A, B, C, D, E, F and G have been rented out. We could set up node E in order to let all the traffic be bypassed between nodes B and F, as well as C and G. Node E being always optically bypassed, routes like A to F or D to G are one hop shorter. Moreover, the global topology is modified, since, for instance, the route from A to G is not A-B-E-G anymore, but A-B-C-G. A single core node could contain more than a thousand of line cards, in particular 1152 in the case of Cisco CRS-3 Multishelf System [16]. Since line cards and wavelengths can be independently managed in CSN (software configuration, not hardware), it would be easy for a company to rent them out for specific purposes. Research laboratories could also test specific networks over CSN. In this example, we emphasize the fact that physical parts of the network can be easily reconfigured for a specific purpose, thanks to their independency. In particular, we just illustrated a case where we need to adapt the mechanism presented in Section 3.4. Such rented out wavelengths could be exclusively reserved for the described service, and therefore should not be mixed with those which are managed by the global policy presented before.

Further research concerns different methods to use CSN in an efficient way. For instance, in order to reduce the wavelength contention costs and increase the transfer speeds of large contents, we could imagine a scenario where users split them in several parts and could send them simultaneously using different wavelengths. On contrary to usual packet-switching, the goal would be to enable parallel transfer over the network, where network resources would be reserved in this purpose.

4.2. Deployment Benefits and Issues

As written in the introduction, the Content-based Switching Network addresses the switching plane separation usually proposed by existing hybrid networks. Additionally, usual networks also distinguish three different layers, especially when it concerns large contents: the physical layer, the service network layer and the access network layer. In CSN, only one layer is used. We exemplified this fact in Section 4.1, showing how services, access nodes and physical resource management of the core nodes are tight together. Also, electrical and optical resource controls are commonly separated. Thanks to its signaling scheme and core node architecture, all these controls are unified for more simplicity.

As for the size of the contents, we think that small packet transfers as the ones performed in actual IP networks should be avoided in CSN, because of their granularity and the waste of bandwidth they implies. On contrary, due to its content abstraction, the Content-based Switching Network provides a way to avoid complicated transport layer functions, like TCP, if they are not essential. CSN does not encapsulate contents, leaving the user to choose the complete nature of the content. If target users are end users, details should be delegated to the application layer, as in P2P applications for instance. If users are data centers, contents could be a part of a file, a file or several files, avoiding the need of creating a new path for each file in the latter case.

We could think that the separate control plane of CSN may require a recovery procedure in case of congestion. Hopefully, considering that 1) we are avoiding dealing with small packets and their associated traffic and that 2) CSN control messages are small, we think that congestions should not usually occur; therefore a recovery procedure is not needed. Also, the physical layer should be robust enough to deal with bit losses, with functions to avoid message corruptions.

CSN performance is tightly related to the state-of-the-art devices. Since store-and-forwarding relies on RAM memory specifications, the network maximum reachable speed is limited by them. For instance, DDR3 devices can operate at 100 Gbps today. Since in CSN we reserve memory speed in order to avoid bit losses, and consequently content drops, it is trivial to figure why reaching 100 Gbps of network speed is compromised nowadays. If 100 Gbps are reserved to receive one content, remaining resources will not be sufficient to either forward the content before complete reception or deal with other contents. Simulation results presented in Section 5.2 show that CSN can already successfully handle 1 Gbps or 10 Gbps traffic though. All-optical transmissions do not have this issue, thus contents which can be optically bypassed from end-to-end are able to be transferred at very high speed. Additionally, we assumed during the simulations that DDR modules are used. This kind of RAM memory, having only one port to handle both read and write operations, is not suitable for networking applications [17]. Instead, QDR (Quad Data Rate) modules would be more adapted: the fact is that, due to the separation of input and output buffers, they are efficient when read and write operations are interleaved, exactly what is needed in our case. Unfortunately, QDR memory is still expensive and not developed enough to be used in actual networks.

Finally, the deployment of the CSN network from an unclean slate approach should be discussed. Due to the use of a new protocol, edges and the core parts of the network should be aware of it. In particular, simulation results of Section 5.3 deal with the deployment of CSN from an OCS network. In such a case, all access nodes are already aware of the CSN protocol in order to manage CSN control messages. OCS core nodes, which can only use optical bypassing, are progressively replaced by CSN ones. We will show that the progressive deployment can be beneficial for the whole network. Also, the use of the backlogged queue feature in access nodes can improve the OCS network performance, even if not even one CSN core node has been deployed yet.

5. Performance Evaluation

5.1. Simulation Model

In order to evaluate the proposed network, we designed a simulator using OMNeT++ 4.3 [18,19]. OCS network has also been implemented for the comparison with CSN. The design of CSN allows the use or not of the backlogged queue in the access nodes, since it seems to be unfair to compare directly OCS to CSN. From now, we will call *CSN With* the CSN implementation where the access backlogged feature is activated, while we will refer to *CSN Without* as the implementation where the feature is not used. We will consider two different scenarios. In both cases, the rejection ratio of the networks will be used for the comparison. The rejection ratio $R_{\text{rejection}}$ is defined as

$$R_{\text{rejection}} = \frac{N_{\text{rejected contents}}}{N_{\text{received contents}} + N_{\text{rejected contents}}} \qquad (11)$$

where $N_{\text{received contents}}$ and $N_{\text{rejected contents}}$ are respectively the number of successfully received and rejected contents.

As for the simulation settings, the COST 266 Core topology [20], which is a large optical backbone interconnecting 16 European cities, has been chosen to analyze their performances depending on the route length (from 1 to 6 hops) of a transfer. Four input and four output fibers, each of which carrying five wavelengths, interconnected two adjacent nodes. Each ingress line card contained a 120 Gbps, 1 GiB RAM memory, which corresponds to a single DDR3 SDRAM PC3-15000 DDR3-1866 device. The size of the control messages was arbitrarily fixed to 1 KiB. Traffic was generated at every node following an exponential distribution (classical one), the destination was uniformly chosen. End-to-end routes were computed so as to minimize the number of hops. Each simulation was performed several times using different seeds. Five seeds for the first scenario, twenty seeds in the second one. The reason of the numerous seeds in the second scenario will be given in Section 5.3. Results are the average of all the runs. Content sizes were generated using a log-normal distribution (mean = 1, variance = 0.7). The x-axis values were multiplied by 50 MiB, which led to a probability density as shown in **Figure 10**. Simulation duration was set to 3600 seconds, while a warming-up period of 30 seconds was observed at the beginning in order to avoid the transient period.

As for the backlogged and forwarding queues, the EWMA parameters were all set to 0.1. It is greater than the ones usually used in packet networks, since contents are less numerous than packets (*i.e*. the ratios are updated less frequently). The sender time out was fixed to 5 seconds and the maximum number of data collection attempts to 5, which leads to about one attempt per second during 5 seconds for a specific content in the access nodes. The receiver time out was set to 30 seconds, and the number of attempts to 30, which leads also to a frequency of one attempt per second if the queueing delays of the forwarding queues are not too long. Networks which do not make use of access backlogged queues can also retry a path establishment 5 seconds (=sender time out) after a fail. In such a case, it is assumed that the new attempt comes from the initiative of the user, and not from the network since the user request was previously rejected. It could be compared to page refreshing in web browsers for instance. Concerning the wavelength thresholds, the minimum ratio was set to 10%. Then the minimum amount of shared or large wavelengths would be 2 (= 0.1 × 4 line cards × 5 wavelengths) in the worst case.

In all graphs showing the simulation results, x-axis represents the traffic generated by each node, and not the global traffic of the network, which is the result of the multiplication of the value indicated by the x-axis multiplied by the number of nodes of the network. There are sixteen different nodes in the case of the COST 266 Core topology. The y-axis represents the aforementioned rejection ratio per route length. A logarithm scale is used.

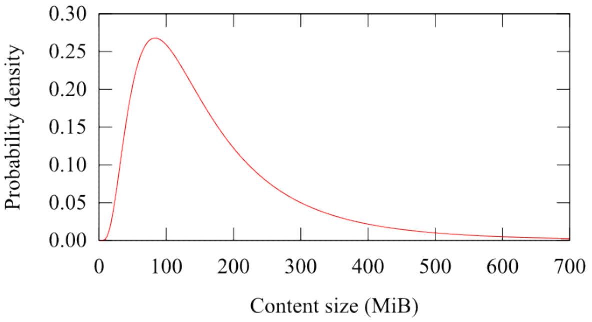

Figure 10. **Content size repartition (log-normal).**

5.2. First Scenario: Network Speed

In this first scenario, we modify the base speed of the network and compare the rejection ratios of OCS, CSN Without and CSN With. Network speed was set to 1 Gbps and 10 Gbps. Since results were similar from one seed to another, five different seeds (small number) were used for each point. **Figures 11-13** show the results for the 1 Gbps case. **Figures 14-16** show the ones for the 10 Gbps case. All transfers are limited to this speed, all-optical ones included.

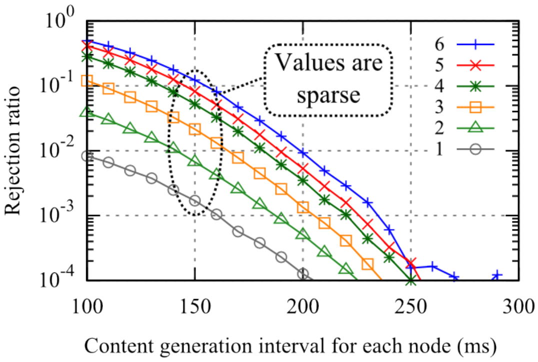

Figure 11. **OCS—1 Gbps.**

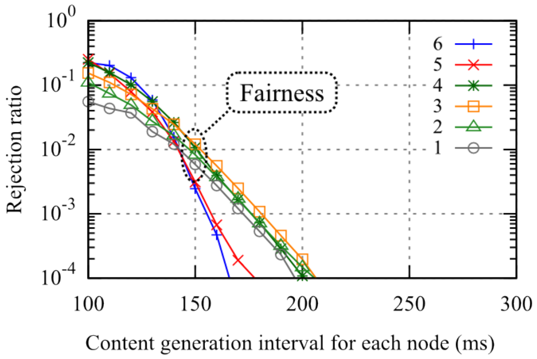

Figure 12. **CSN Without—1 Gbps.**

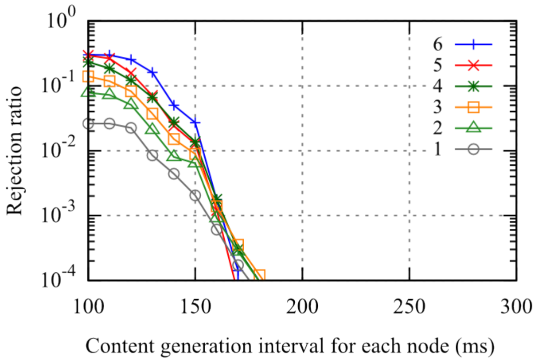

Figure 13. **CSN With—1 Gbps.**

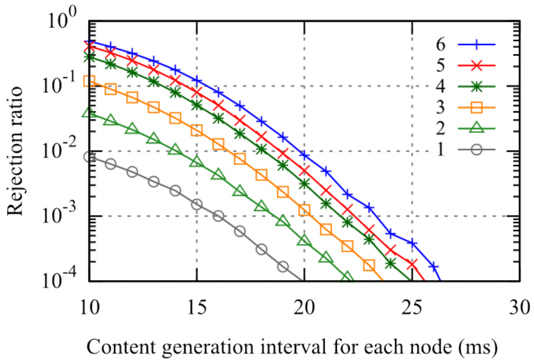

Figure 14. **OCS—10 Gbps.**

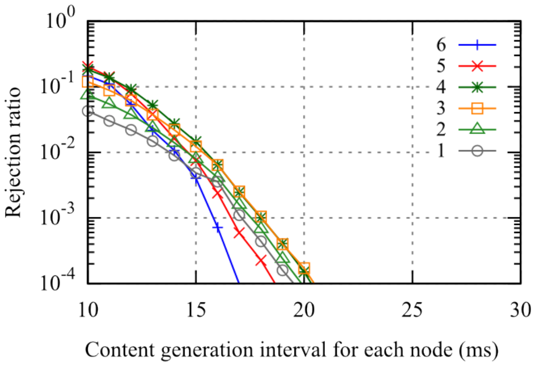

Figure 15. **CSN Without—10 Gbps.**

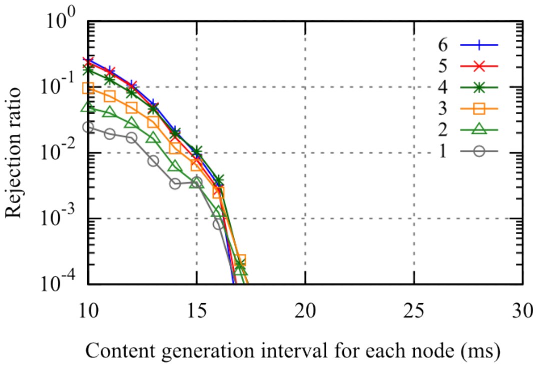

Figure 16. **CSN With—10 Gbps.**

The first thing we can notice is the fact that increasing the speed by a factor 10 allowed to increase the traffic by the same factor. Moreover, graph shapes are similar. That denotes that results are predictable, which is always a good thing. By comparing OCS to CSN, we can state that OCS results for each hop are sparse. The fact that lines are close one another in the CSN case shows that it has a better fairness regarding the route length. This behavior is greatly appreciated for large scale networks. Additionally, one and two hops routes excepted, OCS rejection ratios are greater than CSN ones. In particular, CSN was able to greatly decrease the rejection ratio of long routes. From 200 ms of traffic in the 1 Gbps case, we can see that CSN has no rejections anymore, contrary to OCS. The main differences between CSN With and CSN Without are 1) CSN With can stand a higher traffic before first rejections and 2) CSN With seems to lose fairness when the traffic is too high. Nevertheless, the second point is not an issue in the 10 Gbps case. If we compare Figures 15 and 16, CSN With results are all better than CSN Without ones at traffic of 10 ms. Shorter routes have simply fewer rejections. This can be explained by the fact that CSN With uses backlogged queues, which have a great influence on the probability of establishing a path.

5.3. Second Scenario: CSN Deployment

The second scenario concerns the deployment of CSN from an OCS network. All access nodes are assumed to know the CSN protocol; hence we will distinguish both the CSN With and Without cases. Only core nodes are full OCS (optical bypass is the unique available switching paradigm), and will be randomly replaced by CSN core nodes. In order to have reliable results, we chose to use 20 different seeds in order to cover several different cases. The speed of the network is set to 1 Gbps. Figure 11 shows the results for a raw OCS network. Figures 17 to 22 show the deployment results for the CSN Without network, from 0% to 100% of deployment. Figures 23 to 28 show the ones for CSN With.

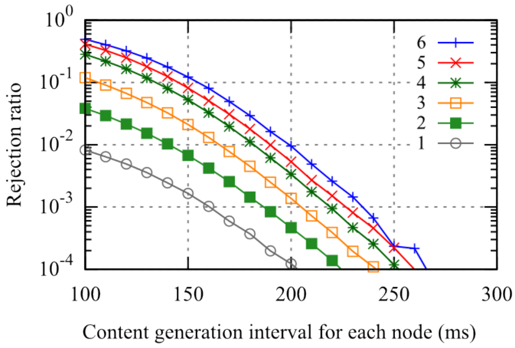

Figure 17. **CSN Without—0%.**

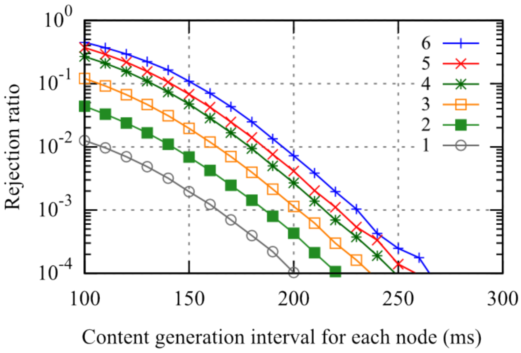

Figure 18. **CSN Without—20%.**

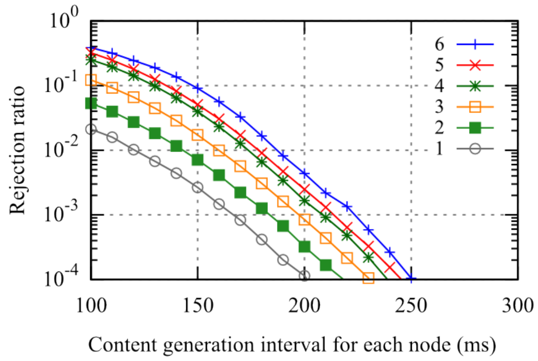

Figure 19. CSN Without—40%.

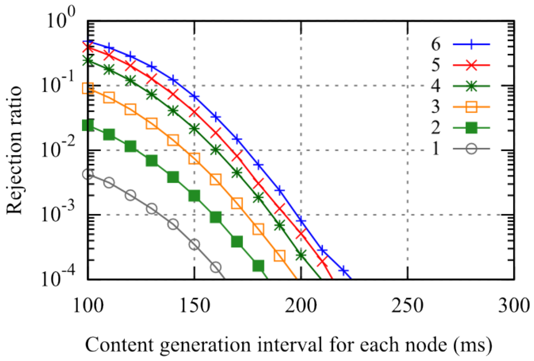

Figure 23. CSN With—0%.

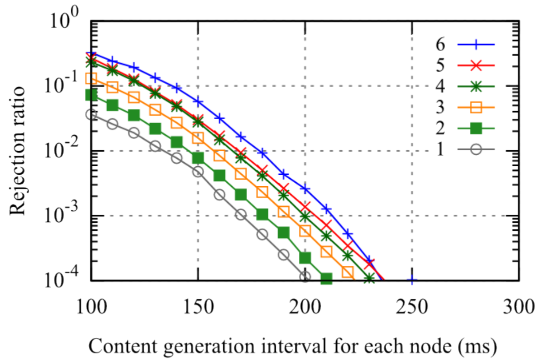

Figure 20. CSN Without—60%.

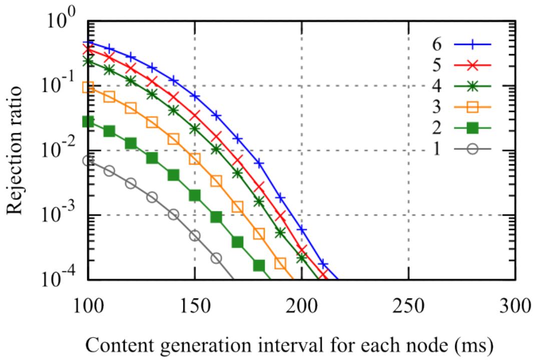

Figure 24. CSN With—20%.

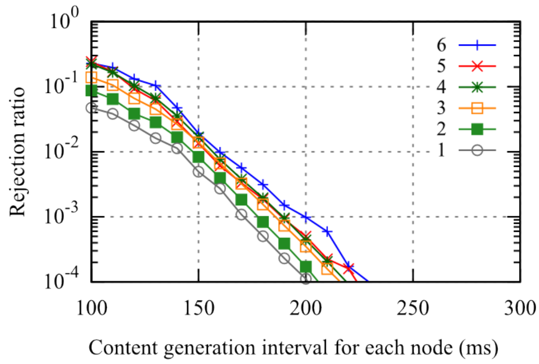

Figure 21. CSN Without—80%.

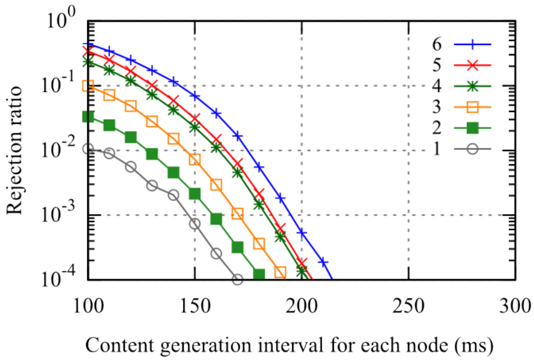

Figure 25. CSN With—40%.

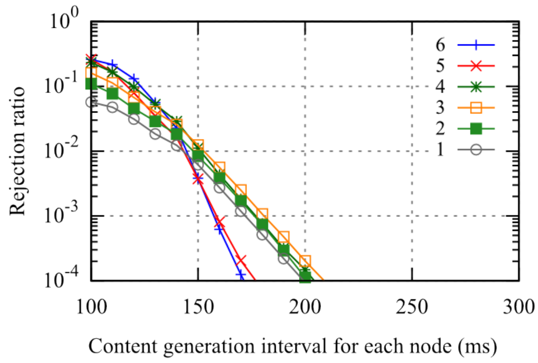

Figure 22. CSN Without—100%.

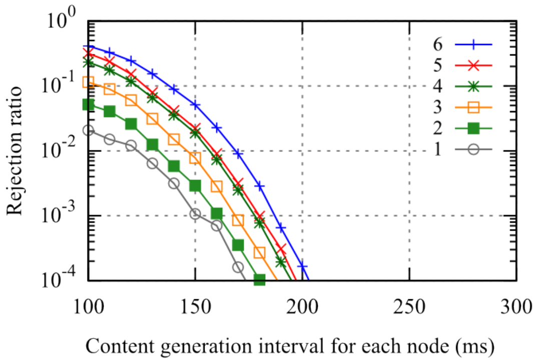

Figure 26. CSN With—60%.

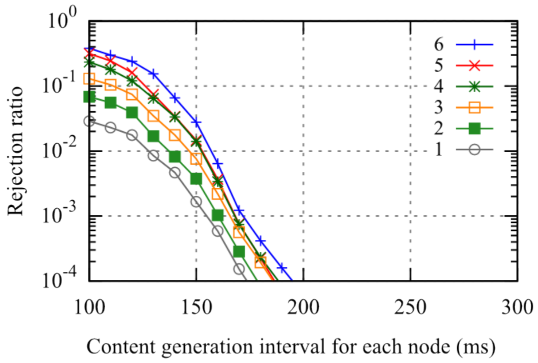

Figure 27. CSN With—80%.

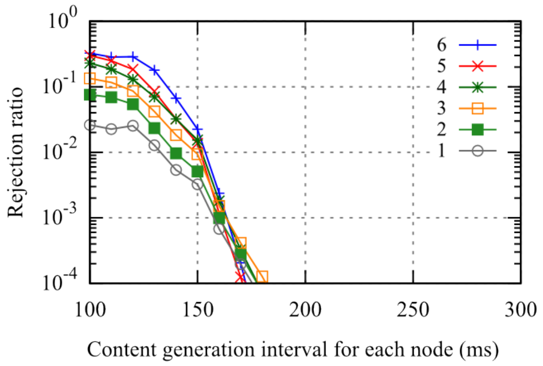

Figure 28. CSN With—100%.

As expected, OCS and CSN Without results at 0% of deployment are similar. Little differences are due to the number of seeds used. In both CSN With and CSN Without cases, we can see that the more the deployment progresses, the more fairness is achieved: lines are more and more closer one another. By comparing **Figures 17** and 23, we can notice that using a backlogged queue is beneficial for the network, even before core deployment: CSN With results globally prove fewer rejections at this point. Even if a clean slate approach is better, we can see that a progressive deployment of CSN core nodes from an OCS network is possible and simultaneously improves the performances of the global network.

6. Conclusion

We presented in this paper the Content-based Switching Network. Unlike conventional hybrid networks which implement a horizontal separation over the entire network (from edge to edge), CSN smoothly adapts the switching paradigm at the core node level in order to achieve as few separations as possible. The global concepts, architecture and signaling schemes of our proposal were given. Examples of services that can be implemented in CSN were shown. A discussion about the deployment of CSN, its benefits and some feasibility issues was done. We presented simulation results which showed to what extent the Content-based Switching Network can increase the fairness of the network regarding the route length and that it makes a better use of the optical resources of the network. These behaviors make CSN suitable for large scale applications in particular. Results also showed that CSN can successfully transfer contents at 1 Gbps and 10 Gbps. Progressive deployment of the Content-based Switching Network from an OCS network proved to be beneficial.

Acknowledgements

This research is partly supported by the National Institute of Information and Communications Technology, Japan.

REFERENCES

[1] Alcatel-Lucent, "Video Shakes up the IP Edge—A Bell Labs Study on Rising Video Demand and its Impact on Broadband IP Networks," Strategic White Paper, 2012. http://www3.alcatel-lucent.com/wps/DocumentStreamerServlet?LMSG_CABINET=Docs_and_Resource_Ctr&LMSG_CONTENT_FILE=White_Papers/Video_Shakes_Up_IP_Edge_EN_Whitepaper.pdf

[2] K. J. Barker, A. Benner, R. Hoare, A. Hoisie, A. K. Jones, D. K. Kerbyson, *et al.*, "On the Feasibility of Optical Circuit Switching for High Performance Computing Systems," *Proceedings of the* 2005 *ACM/IEEE Conference on Supercomputing*, 2005, pp. 16-38. http://dx.doi.org/10.1109/35.894388

[3] L. Xu, H. G. Perros and G. Rouskas, "Techniques for Optical Packet Switching and Optical Burst Switching," *IEEE Communications Magazine*, Vol. 39, No. 1, 2001, pp. 136-142. http://dx.doi.org/10.1109/35.894388

[4] T. Miyazawa, H. Furukawa, K. Fujikawa, N. Wada and H. Harai, "Development of an Autonomous Distributed Control System for Optical Packet and Circuit Integrated Networks," *Journal of Optical Communications and Networking*, Vol. 4, No. 1, 2012, pp. 25-37. http://dx.doi.org/10.1364/JOCN.4.000025

[5] T. Miyazawa, H. Furukawa, H. Harai and N. Wada, "Proposal and Implementation of an Autonomous Distributed Control for Elimination of Incomplete Lightpaths in Optical Packet and Circuit Integrated Networks," 16*th International Conference on Optical Network Design and Modeling*, Colchester, 17-20 April 2012, pp. 1-6. http://dx.doi.org/10.1109/ONDM.2012.6210263

[6] X. Wu, M. C. Chan, A. L. Ananda and C. Ganjihal, "Sync-TCP: A New Approach to High Speed Congestion Control," 17*th IEEE International Conference on Network Protocols*, Princeton, 13-16 October 2009, pp. 181-192. http://dx.doi.org/10.1109/ICNP.2009.5339684

[7] C. Ganjihal, "Experimental Evaluation of Sync-TCP and Other Highspeed Congestion Control Algorithms," Master's Thesis, National University of Singapore, 2009.

[8] Y. Chen, C. Qiao and X. Yu, "Optical Burst Switching (OBS): A New Area in Optical Networking Research," *IEEE Network Magazine*, Vol. 18, No. 3, 2004, pp. 16-23.

http://dx.doi.org/10.1109/MNET.2004.1301018

[9] A. Passarella, "Review: A Survey on Content-Centric Technologies for the Current Internet: CDN and P2P Solutions," *Computer Communications*, Vol. 35, No. 1, 2012, pp. 1-32. http://dx.doi.org/10.1016/j.comcom.2011.10.005

[10] K. Kitayama, A. Shinya, S. Matsuo, R. Takahashi, M. Murata and S. Arakawa, "Optical RAM Buffer for All-Optical Packet Switches," *Asia Communications and Photonics Conference and Exhibition*, Shanghai, 2-6 November 2009, pp. 5-6. http://dx.doi.org/10.1364/ACP.2009.FT1

[11] S. Das, G. Parulkar and N. McKeown, "Unifying Packet and Circuit Switched Networks," *IEEE GLOBECOM Workshops*, Honolulu, 30 November-4 December 2009, pp. 1-6. http://dx.doi.org/10.1109/GLOCOMW.2009.5360777

[12] V. R. Gudla, S. Das, A. Shastri, G. Parulkar, N. Mckeown, L. Kazovsky and S. Yamashita, "Experimental Demonstration of Open Flow Control of Packet and Circuit Switches," *Optical Fiber Communication Conference*, San Diego, 21-25 March 2010, pp. 1-3. http://dx.doi.org/10.1364/OFC.2010.OTuG2

[13] S. Das, "pac.c: A Unified Control Architecture for Packet and Circuit Network Convergence," Ph.D. Thesis, Stanford University, Stanford, California, 2012. http://archive.openflow.org/wk/index.php/PACC_Thesis

[14] N. McKeown, T. Anderson, H. Balakrishnan, G. Parulkar, L. Peterson, J. Rexford, S. Shenker and J. Turner, "Open Flow: Enabling Innovation in Campus Networks," *ACM SIGCOMM Computer Communication Review*, Vol. 38, No. 2, 2008, pp. 69-74. http://dx.doi.org/10.1145/1355734.1355746

[15] V. Jacobson, D. K. Smetters, J. D. Thornton, M. F. Plass, N. H. Briggs and R. L. Braynard, "Networking Named Content," *Proceedings of the* 5*th International Conference on Emerging Networking Experiments and Technologies*, 1 December 2009, pp. 1-12. http://dx.doi.org/10.1145/1658939.1658941

[16] Cisco Carrier Routing System, "Compare Models," 2006. http://www.cisco.com/en/US/products/ps5763/prod_models_comparison.html

[17] A. Chakrapani, "QDR SRAM and RLDRAM: A Comparative Analysis," White Paper, Cypress Semiconductor Corp., 2010. http://www.cypress.com/?docID=24581

[18] OMNeT++ Network Simulation Framework. http://www.omnetpp.org

[19] E. Weingärtner, H. vom Lehn and K. Wehrle, "A Performance Comparison of Recent Network Simulators," *Proceedings of the* 2009 *IEEE International Conference on Communications*, Dresden, 14-18 June 2009, pp. 1-5. http://dx.doi.org/10.1109/ICC.2009.5198657

[20] S. D. Maesschalck, D. Colle, I. Lievens, M. Pickavet, P. Demeester, C. Mauz, *et al*., "Pan-European Optical Transport Networks: An Availability-Based Comparison," *Photonic Network Communications*, Vol. 5, No. 3, 2003, pp. 203-225. http://dx.doi.org/10.1023/A:1023088418684