CENTRES DE DONNEES A BASE DE PAQUETS OPTIQUES ET DE TRANSPONDEURS ELASTIQUES

Miquel A. Mestre, Guilhem de Valicourt, Philippe Jennevé, Haik Mardoyan, Sébastien Bigo et Yvan Pointurier

Alcatel-Lucent Bell Labs France, Route de Villejust, 91620 Nozay

Miquel_Angel.Mestre_Adrover@alcatel-lucent.com

RÉSUMÉ

Nous proposons un nouveau concept de réseau pour les centres de données combinant la transmission de paquets optiques et l'utilisation de transpondeurs élastiques capables de varier leur débit de 100 à 250 Gb/s en fonction du nombre de nœuds à traverser. Notre scenario requiert 340 fois moins de transpondeurs qu'un réseau actuel de centre de données à base de commutateurs électroniques et de transpondeurs à 10 Gb/s.

MOTS-CLEFS : *Transpondeur élastique ; paquets optiques ; centre de donnés.*

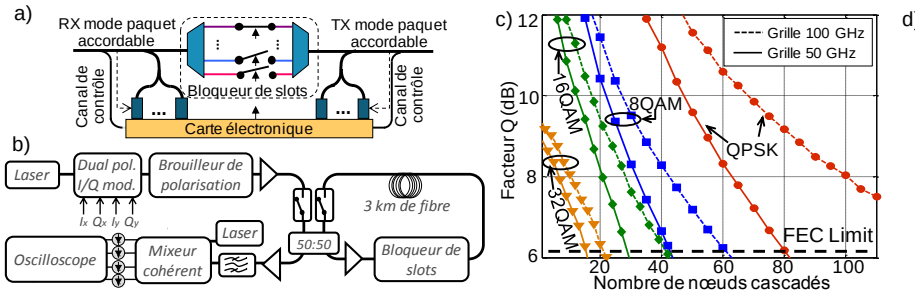
1. INTRODUCTION

Aujourd'hui, les centres de données (DCs) deviennent de plus en plus importants, allant de petites fermes de serveurs distribuées à des grandes fermes dédiées à des tâches spécifiques comme le traitement de données, le calcul, le stockage de données, etc. Le besoin de capacité à l'intérieur du DC augmente très rapidement (p. ex. un factor 20 tous les 4 ans pour les super-calculateurs) et conduit à une forte augmentation du coût et de la consumation d'énergie des équipements [1]. Un moyen de réduire la consommation d'énergie est l'introduction de la transparence, c'est-à-dire, la suppression des conversions optoélectroniques énergivores. Cependant, les DCs peuvent devenir si grands que leur interconnexion transparente avec des circuits optiques est impossible à cause du nombre limité de longueurs d'onde disponible dans la bande spectrale optique (bande C). La commutation de paquets optiques (OSS, Optical Slot Switching) dans des anneaux a été développée pour résoudre ce problème en introduisant la granularité de commutation sous-longueur d'onde [2].

Nous proposons un réseau de commutation à base d'OSS pour DCs de grande envergure. Néanmoins, dans les grands anneaux d'OSS, les signaux modulés à très haut débit ont le risque d'être trop dégradés quand ils sont reçus par les nœuds les plus éloignés. Pour la première fois, nous caractérisons expérimentalement la dégradation des signaux à très haut débit (32 Gbaud) et au format de modulation avancé (jusqu'à PDM-32-QAM) traversant une cascade de nœuds OSS. Nous montrons aussi une meilleure mise à l'échelle des DCs en utilisant l'anneau OSS avec des transpondeurs élastiques, par rapport à un anneau avec des transpondeurs à débit fixe, ou à l'architecture actuelle des réseaux DC à base de commutation électronique.

2. CASCADE DE NŒUDS DE COMMUTATION DE PAQUETS OPTIQUES

Nous représentons un nœud OSS dans la Fig. 1(a), où les données sont encapsulées dans des paquets de durée fixe (« slots ») multiplexés dans le temps et en fréquence (longueur d'onde). L'OSS permet une communication transparente entre tous les nœuds avec une granularité d'un slot (quelques µs). Ce nœud est composé par un bloqueur de slot, un récepteur (RX) et un transmetteur (TX) chacun capable d'accorder sa longueur d'onde rapidement [3]. Le bloqueur peut supprimer chaque slot sur chaque longueur d'onde, et peut être implémenté avec un démultiplexeur de longueur d'onde (DEMUX), une série de portes optiques telles que des atténuateurs optiques variables (VOA) rapides, et un multiplexeur de longueur d'onde (MUX). Le filtrage produit dans le MUX/DEMUX peut fortement dégrader la qualité du signal. Une étude expérimentale de la robustesse des signaux qui traversent une longue cascade de nœuds est donc nécessaire pour évaluer



Nœuds cascadés		
Mod. (Gb/s)	Nœuds 100 GHz	Nœuds 50 GHz
QPSK (100)	120	79
8QAM (150)	60	40
16QAM (200)	40	27
32QAM (250)	20	14

Figure. 1: a) Nœud OSS, b) Montage expérimental, c) Factor Q en fonction de nombre de nœuds cascadé et d) Nombre de nœuds cascadé pour cheque format de modulation avec grilles de 50 et 100 GHz.

les avantages des OSS dans des grands DCs. Dans cette étude, nous considérons divers formats de modulation (QPSK, 8-QAM, 16-QAM et 32-QAM) et un débit-symbole de 32 GBaud.

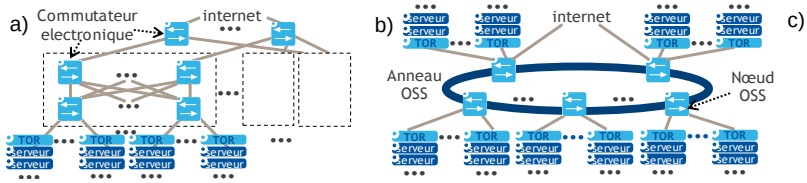
Nous représentons notre montage expérimental dans la Fig. 1(b). On commence par passer la lumière d'un laser à cavité externe (ECL) dans un modulateur I/Q à diversité de polarisation, afin de produire des paquets de durée 2 µs. Ceux-ci sont formés d'une séquence d'apprentissage de 256 symboles QPSK, suivie de 62080 symboles porteurs d'information utile. Ils traversent un brouilleur de polarisation puis un amplificateur, avant de pénétrer dans une boucle à recirculation. Celle-ci comprend deux commutateurs sélectifs flexibles en longueur d'onde, qui émulent les DEMUX/MUX d'un nœud OSS en utilisant une grille à 50 ou 100 GHz. Elle comprend aussi une bobine de fibre standard de 3 km en guise de ligne à retard optique pour faciliter le déclenchement de la boucle, et un amplificateur optique qui compense les pertes de 20 dB équivalente à celle d'un nœud OSS typique. La lumière est extraite et envoyée à un récepteur cohérent où elle est échantillonnée à 80 GS/s par un oscilloscope Lecroy Teledyne. Les paquets reçus sont traités hors-ligne avec des algorithmes de démultiplexage de polarisation rapides [4]. Dans cette expérience nous nous concentrons sur les effets de cascade de filtres et la porte n'est pas commutée.

La Fig. 1(c) montre la performance (facteur Q) de chaque format de modulation en fonction du nombre de nœuds OSS traversés, tandis que la Fig. 1(d) résume le nombre maximum de nœuds atteint pour un facteur Q de 6.25 dB, correspondant à la limite d'un code de correction d'erreurs (FEC) incluant 20% de sur-débit. Les paquets QPSK sont les plus robustes aux distorsions induites par les nœuds, et peuvent traverser jusqu'à 120 nœuds avec une grille de 100 GHz (79 nœuds avec une grille de 50 GHz). Comme prévu, le nombre maximum de nœuds atteint est réduit lorsque l'on augmente l'ordre de modulation. Au plus 60, 40 et 20 nœuds peuvent être cascadés pour des signaux PDM 8-,16- et 32-QAM, respectivement, sur la grille de 100 GHz. En passant à la grille de 50 GHz on observe une nouvelle réduction du nombre maximal de nœuds traversés à cause du rétrécissement spectral produit par la concaténation des filtres. Ces résultats suggèrent l'utilisation de transpondeurs élastiques capables d'adapter leur modulation (et donc leur débit de données) en fonction du nombre de nœuds à traverser dans des grands réseaux de DC.

3. APPLICATION AUX CENTRES DE DONNEES

Un DC se compose généralement de serveurs fixés dans des racks, chacun équipé d'un commutateur « Top of Rack » (ToR). Des commutateurs supplémentaires assurent la connectivité entre les ToRs, comme le montre la Fig. 2(a) pour un réseau de commutation typique (Folded Clos) [5]. En outre, les DCs sont généralement sous-dimensionnés, en particulier aux niveaux supérieurs de la hiérarchie des commutateurs, afin de réduire leur coût. Un sous-dimensionnement de 1 signifie que le réseau peut supporter la demande de tous les périphériques. Dans cette étude, nous considérons des racks de 20 serveurs (chacun avec une interface Ethernet de 1 Gb/s), chaque rack est équipé d'un ToR avec une capacité de commutation de 20 Gb/s et de 2 cartes de ligne à 10 Gb/s.

Comme le montre la Fig. 2(b), nous proposons une architecture où les ToRs sont connectés à un anneau de nœuds OSS. Comme un slot se déplace en moyenne sur la moitié de l'anneau, chaque longueur d'onde est partagée (en moyenne) par 2 nœuds. D'après la Fig. 1(d), il n'est pas possible de cascader plus que 79 nœuds avec une grille de 50 GHz (soit une utilisation de $\lceil 79/2 \rceil = 40$



c)

	OSS (élastique)	OSS (fixe)	Commutation électronique
Cartes	88 x 250 Gb/s	88 x 100 Gb/s	30180 x 10 Gb/s
Câbles	88	88	30180
Serveurs	137280	88000	137180

Figure 2: Centre de données avec architecture Folded Clos (a) et à base d'anneau OSS (b). c) Comparaison des architectures pour un grand centre de donnés.

longueurs d'onde sur les 88 disponibles) alors qu'avec un grille de 100 GHz, 44 longueurs d'onde sont disponibles correspondant à un anneau de 2x44=88 nœuds, lesquels peuvent être cascadés d'après la Fig. 1(d). Ainsi, nous limitons la conception du réseau à la grille de 100 GHz et à un maximum de 88 nœuds OSS. On avait déjà montré dans [4] que, pour assurer une utilisation équitable du canal par tous les nœuds, chaque nœud doit allouer une capacité moyenne de $1/\Sigma_r(N_r/B_r)$ (Gb/s) par longueur d'onde, où N_r est le nombre de nœuds (normalisé par le nombre de nœuds dans l'anneau) pour lequel le débit atteignable maximal est B_r. Par conséquent, en utilisant la Fig. 1(d), le débit de données moyen d'un transpondeur élastique est de 156 Gb/s (incluant un sur-débit de 20% pour le FEC et 8% pour l'encapsulation). En supposant un sous-dimensionnement de 10, chaque nœud OSS peut supporter une demande de 1,56 Tb/s et donc interconnecter 78 racks par nœud (=1,56 Tb/s/20 Gb/s) et un total de 137280 serveurs (88 nœuds x 78 racks x 20 serveurs/rack). Par comparaison, si des transpondeurs non élastiques sont déployés, le format PDM-QPSK à 100 Gb/s est requis pour que les signaux puissent traverser 87 nœuds. Ainsi, avec le même sous-dimensionnement de 10, chaque nœud OSS peut soutenir une de demande de 1 Tb/s, ce qui limite le nombre de racks par nœud à 50, conduisant à un total de 88000 serveurs seulement.

Pour interconnecter (presque) la même quantité de serveurs, l'approche typique d'aujourd'hui repose sur des commutateurs électroniques de degré k dans une topologie « Folded Clos », de sorte que $k^3/4$ interfaces peuvent être supportées. Dans ce cas, des commutateurs électroniques avec k=38 ports sont nécessaires pour relier $38^3/4=13718$ interfaces de ToR, soit 137180 serveurs. En supposant un sous-dimensionnement de 10, on peut montrer qu'un réseau DC entièrement électronique nécessite $1.1k^3/2=30180$ interfaces à 10 Gb/s. On note que les interfaces des ToRs et ToRs-commutateurs, également à 10 Gb/s, ne sont pas comptabilisées dans ce total car elles sont nécessaires dans tous les types de réseau considérés. La Fig. 2(c) compare les trois topologies. Une topologie en anneau OSS simplifie énormément l'architecture, en diminuant le nombre de transpondeurs de ~30000 (interfaces à 10 Gb/s) à 88 (élastiques ; jusqu'à 250 Gb/s) pour supporter un nombre similaire de serveurs. Cette réduction massive se traduit par une réduction du même ordre du nombre de câbles d'interconnexion, ce qui facilite considérablement l'installation du réseau. D'autre part, les transpondeurs élastiques permettent une augmentation de 49280 serveurs supportés par rapport au cas où les transpondeurs ne sont pas élastiques.

CONCLUSION

Nous avons proposé l'utilisation de la commutation de paquets optiques dans un anneau avec des transpondeurs cohérents élastiques capables d'adapter leur débit de 100 à 250 Gb/s en fonction du nombre de nœuds à cascader, avec une granularité de l'ordre de la microseconde. Le nombre de cartes de ligne et de câbles nécessaires pour équiper un centre de données avec ~140000 serveurs est divisé par 340, rendant possible l'interconnexion à l'intérieur de très grands centres de données.

Ce travail a été partiellement financé par le projet CELTIC+ SASER-SAVENET.

REFERENCES

[1] P. Pepeljugoski et al., Proc. OFC, OThX2 (2010).
[2] D. Chiaroni et al., Bell Labs Technical Journal, vol. 14, no 4, pp. 263-285, Winter 2010.
[3] J. Simsarian et al., Proc OFC, PDP.B.5 (2010).
[4] F. Vacondio et al., Proc. ECOC, We.1.F.2 (2013).
[5] A. Vahdat et al., in Proc. SIGCOMM, 2008