NOKIA Bell Labs

Spécialité : Electronique et communications

Ecole doctorale : Informatique, Télécommunications et Electronique de Paris

Présentée par

Miquel Angel Mestre Adrover

Pour obtenir le grade de
DOCTEUR DE TELECOM SUDPARIS

Data center optical networks: short- and long-term solutions

Réseaux optiques pour les centres de données: solutions à court et long terme

Soutenue le 21/10/2016

Devant le jury composé de:

Directeur de thèse
Prof. Badr-Eddine Benkelfat

Encadrants de thèse
Dr. Yann Frignac
Dr. Yvan Pointurier

Rapporteurs
Prof. Lars Dittmann
Prof. Christophe Peucheret

Examinateurs
Prof. Delphine Marris-Morini
Dr. Cédric Ware

N° NNT : 2016TELE0022

Réseaux optiques pour les centres de données : Solutions à court et long terme

Résumé

Les centres de données forment aujourd'hui le cerveau qui rend possible tous les services Web et en nuage largement utilisés dans le monde entier. De telles installations comprennent un grand nombre de serveurs interconnectés qui stockent et traitent toutes les informations disponibles sur le Web et donnent lieu à des applications en nuage / Internet que nous utilisons au jour le jour (par exemple, le stockage en nuage, la lecture de flux vidéo, le partage d'images et de vidéos, les réseaux sociaux, etc.).

Le trafic dans les centres de données progresse à un rythme très rapide. La Fig. 1 montre l'évolution du trafic global des données et les prévisions à partir de 2010 jusqu'en 2019. Comme le montre la figure, le trafic généré dans un centre de données peut être divisé en trois groupes selon son origine/destination : 1) le trafic restant à l'intérieur des centres de données (« within DC » en bleu), 2) le trafic échangé

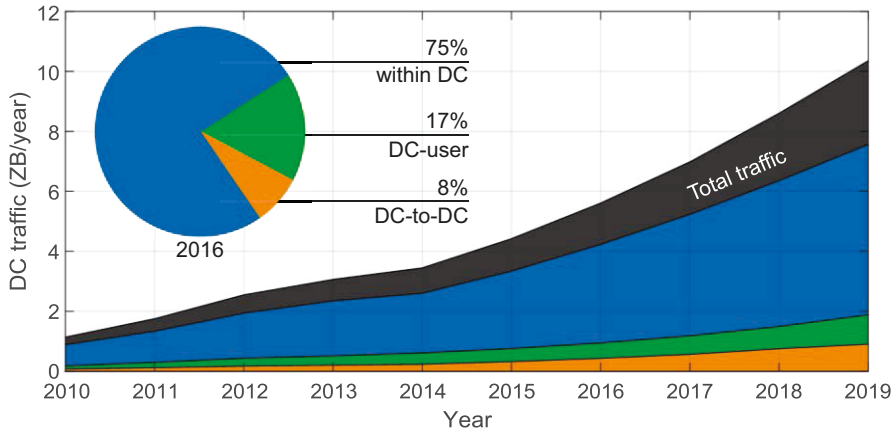


Fig. 1: Évolution du trafic des centres de données et prévisions pour 2010-2019. Données extraites de l'index global des nuages de Cisco.

entre les utilisateurs externes et les centres de données (« DC-user » en vert) et 3) le trafic échangé entre les centres de données (« DC-to-DC » en orange). Le graphique principal représente l'évolution du trafic pour chaque type (en couleur) et au total (en gris foncé). On peut clairement remarquer la prédominance du trafic interne du centre de données, qui représente 75% du trafic total (voir graphique sectoriel montrant le pourcentage pour la prévision du trafic en 2016). Ce pourcentage met en lumière la grande quantité de trafic échangée entre les serveurs et/ou les unités de stockage. Le trafic restant sort du centre de données pour établir la communication avec les utilisateurs (17%) et les autres centres de données (8%). L'analyse de l'indice global de cloud de Cisco montre que la quantité totale de trafic générée globalement dans les centres de données affiche une croissance d'un facteur 10 (2010-2019) atteignant plus de 10 ZB/an en 2019.

À la recherche d'une solution passant à l'échelle, les centres de données ont adopté une topologie « Folded Clos », incluant un grand nombre de commutateurs électroniques, organisés en une architecture multi-niveaux pour fournir une connectivité complète entre les serveurs, comme illustré dans la Fig. 2. Afin de suivre la grande demande de capacité, les centres de données augmentent rapidement la capacité de leur réseau de matrices de commutation: après 1 Gb/s,

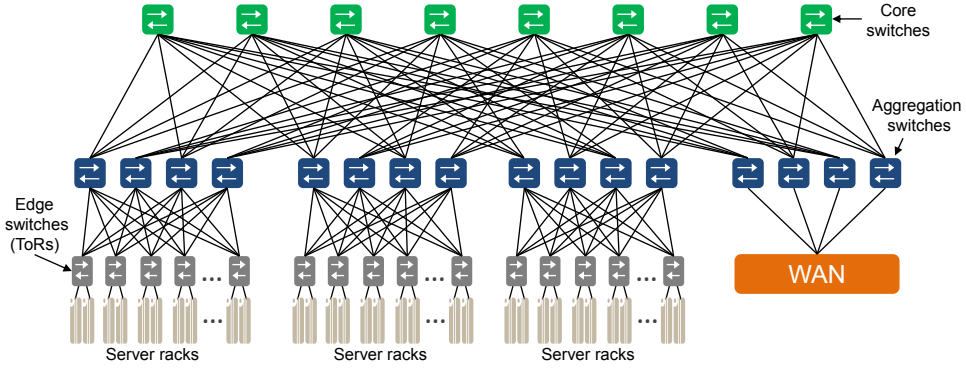


Fig. 2: Centre de données avec une topologie Folded Clos.

ils sont passés à des ports et des interfaces de commutation de 10 Gb/s et 40 Gb/s, ces vitesses étant les plus fréquemment utilisées dans les réseaux de centres de données. Néanmoins, depuis l'apparition de 100-GE (le plus souvent basé sur 4 voies à 25-Gb/s), le couple (25, 100) Gb/s devrait remplacer le couple (10, 40) Gb/s.

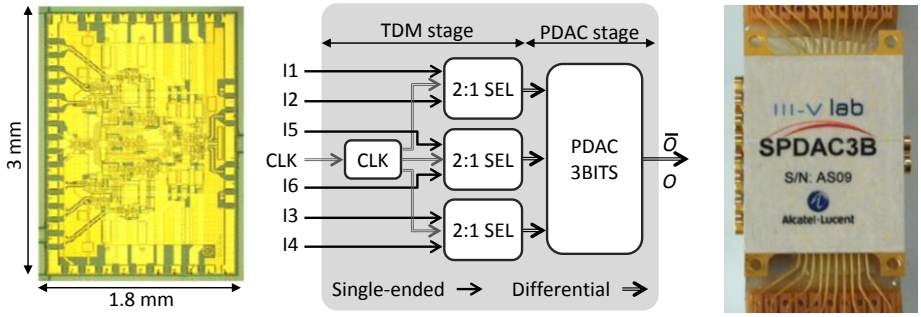
Les centres de données actuels sont confrontés à quatre défis qui doivent être résolus pour continuer à augmenter la capacité : 1) à court terme, afin de upgrader la capacité des centres de données sans augmenter considérablement l'infrastructure du centre de données, les principaux fournisseurs de services nécessiteront le développement de nouveaux commutateurs, supportant un plus grand nombre de ports et de plus en plus de capacité. En conséquence, des interfaces optiques avec des débits de données dépassent à chaque génération des barrières de 25, 50 et 100 Gb/s par voie. Néanmoins, la vitesse de chaque voie et le nombre de ports ne peuvent pas augmenter arbitrairement. Bien que la topologie de commutation électronique Folded Clos offre une évolutivité théoriquement arbitraire, lorsqu'elle passeront aux capacités sans précédent attendues à l'avenir, une telle architecture introduiront d'autres défis qui devront être résolus à plus long terme: 2) énorme complexité du réseau, incluant des centaines de milliers ou même des millions d'éléments; ce qui entraîne des coûts élevés de développement, de maintenance et d'exploitation; 3) une grande consommation d'énergie, causée par la grande quantité de commutateurs et d'interfaces requises pour assurer l'interconnexion des serveurs, qui coûte annuellement des millions de dollars aux grands fournisseurs de services; 4) une latence de bout en bout élevée, provoquée par les nombreux passages à travers des commutateurs électroniques placés dans le réseau à grande échelle; ce qui peut limiter les nouvelles applications en temps réel en nuage.

Dans cette thèse, nous proposons des solutions optiques, du point de vue de la couche physique, qui visent à relever les défis mentionnés ci-dessus à court et à long terme. La couche physique des communications optiques étant l'épine dorsale de ce travail, nous

fournissons dans le Chapitre 2 les connaissances basiques sur les communications optiques. Dans ce chapitre, nous présentons d'abord un large vue d'ensemble des réseaux optiques à l'échelle mondiale pour ensuite nous concentrer sur des aspects plus techniques, tels que les techniques existantes de communication/commutation et la mise en œuvre de différents types d'émetteurs-récepteurs qui seront utilisés dans les chapitres suivants.

Puis, dans le Chapitre 3, nous visons à augmenter la capacité des interfaces optiques. Nous démontrons expérimentalement des émetteurs-récepteurs à modulation d'intensité à haute vitesse et détection directe (IM-DD) capables d'atteindre une capacité de 100 Gb/s par voie et au-delà. Afin d'obtenir des débits de données aussi élevés, nous avons utilisé des formats de modulation avancés tels que la modulation d'amplitude à 4 niveaux (PAM-4) et la modulation d'amplitude à 8 niveaux (PAM-8) qui permettent de doubler et de tripler les débits de données par rapport à la modulation d'amplitude plus basique à 2 niveaux (PAM-2), utilisée aujourd'hui dans les centres de données. Des signaux électriques multi-niveaux à très grande vitesse (jusqu'à 100 GBd) sont fournis par un convertisseur numérique/analogique à haute puissance que l'on appelle *selector power DAC* (SP-DAC), voir la photographie de la puce, le schéma et le module en boîtier de la Fig. 3.

Comme représenté dans le schéma fonctionnel, ce circuit multiplexe



a) SP-DAC chip b) SP-DAC schematic c) SP-DAC module

Fig. 3: (a) photographie de la puce, (b) schéma fonctionnelle et (c) module emballé.

d'abord temporellement les signaux d'entrée (jusqu'à six) par paires (SEL), fournissant jusqu'à trois signaux PAM-2 électroniques au double du débit des données d'entrée. Par la suite, de tels signaux sont utilisés comme bits de codage dans l'étage DAC suivant, générant ainsi des signaux électriques jusqu'à 8 niveaux avec une grande puissance de sortie.

Nous rapportons d'abord un émetteur laser à modulation externe (EML), comprenant un laser à rétroaction distribuée (DFB) et un modulateur d'électro-absorption (EAM) avec une bande passante de 50 GHz, voir le module en boîtier sur la Fig. 4(a). Ce module peut transmettre avec succès 112 Gb/s sur une distance de 2 km avec des marges de fonctionnement très élevées en utilisant des signaux PAM-4 à 56 GBd comme illustré sur la Fig. 4(b-c). Les grandes marges atteintes par notre émetteur permettent de travailler dans une large gamme de configurations des récepteurs, supportant des récepteurs à faible bande passante (jusqu'à 18 GHz) ou une égalisation de faible

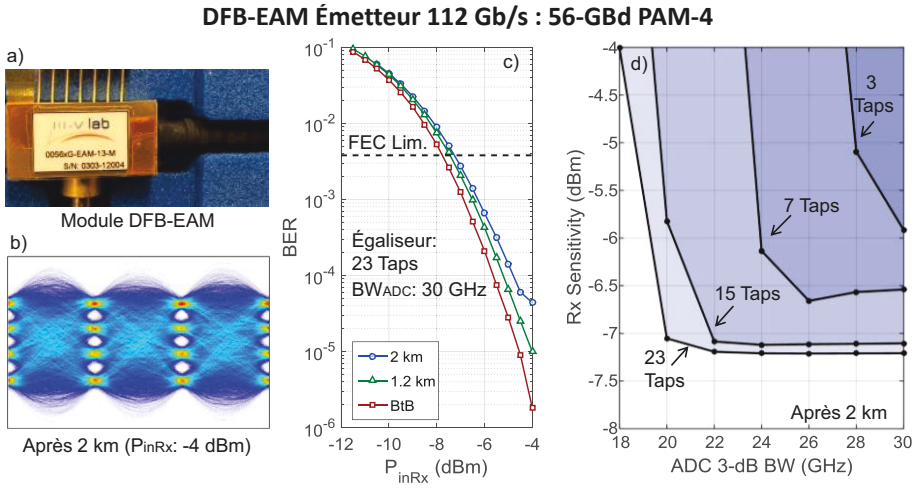


Fig. 4: (a) module en boîtier, (b) diagramme de l'œil après 2 km, (c) BER en fonction de la puissance d'entrée dans le récepteur, (d) sensibilité du récepteur en fonction de la bande passante de l'ADC tout en utilisant des égaliseurs de différentes longueurs.

complexité (seulement trois échantillons ou *taps*), voir Fig. 4(d).

Ensuite, nous rapportons plusieurs solutions capables de surpasser 100 Gb/s, toutes réalisées avec un émetteur IM-DD à base d'un modulateur Mach-Zehnder (MZM). D'abord, nous démontrons un débit de données de 168 Gb/s avec deux approches différentes: PAM-4 à 84-GBd et PAM-8 à 56-GBd. Comme le montre la Fig. 5(a-b), la première approche (vitesse de transmission plus élevée avec un ordre de modulation plus faible) offre les meilleures performances en terme de taux d'erreur binaire (BER) et de sensibilité, mais avec une portée limitée à 1 km en raison de l'impact critique de la dispersion chromatique. A l'inverse, la deuxième approche (ordre de modulation plus élevé avec une vitesse de transmission plus faible) offre une performance inférieure pour des distances inférieures à 1 km, mais permet une portée supérieure (jusqu'à 2 km de transmission).

Enfin, nous démontrons un émetteur-récepteur PAM-4 travaillant à 100 GBd qui émet un débit de données jusqu'à 200 Gb/ s sur une distance de 500 m, voir Fig. 5(c). Un détecteur de séquence de

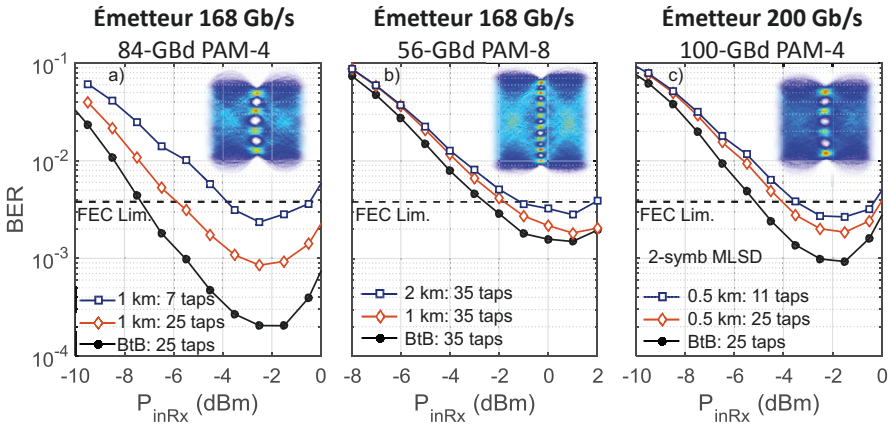


Fig. 5: BER en fonction de la puissance d'entrée dans le récepteur pour une émetteur à (a-b) 168 Gb/s tout en utilisant PAM-4 à 84 GBd (a) ou PAM-8 à 56 Gbd (b), et (c) à 200 Gb/s en utilisant PAM-4 à 100 GBd.

maximum de vraisemblance (MLSD) à 2 symboles permet de détecter un signal à bande passante extrêmement élevée incluant une marge de fonctionnement, tout en utilisant des composants optoélectroniques 40G disponibles dans le commerce et une égalisation de moins de 11 échantillons (taps).

Afin de surmonter les défis introduits antérieurement (2: complexité du réseau, 3: consommation d'énergie et 4: latence), notre équipe a proposé en 2014 un réseau pour l'intra-connexion des centre de données appelé *Burst Optical Slot Switching* (BOSS), qui est décrit de manière détaillée dans le Chapitre 4. Cette proposition remplace le réseau de commutation électronique, représenté sur la Fig. 6(a) par une nouvelle architecture qui relie des serveurs à des nœuds BOSS, qui sont interconnectés par des anneaux de fibre formant une topologie de tore, voir la Fig. 6(b). Le long de chaque anneau, les données se déplacent de manière transparente (c'est-à-dire sans conversion optoélectronique) à travers tous les nœuds intermédiaires (de la source à la destination), encapsulées dans des intervalles de multiplexage temporel de durée fixe de quelques µs et sur des longueurs d'onde différentes.

La combinaison d'émetteurs-récepteurs à grande vitesse (au-delà de 150 Gb/s), du multiplexage statistique et de la transparence optique, permet de réduire de 100 à 500 fois le nombre d'interfaces optiques et

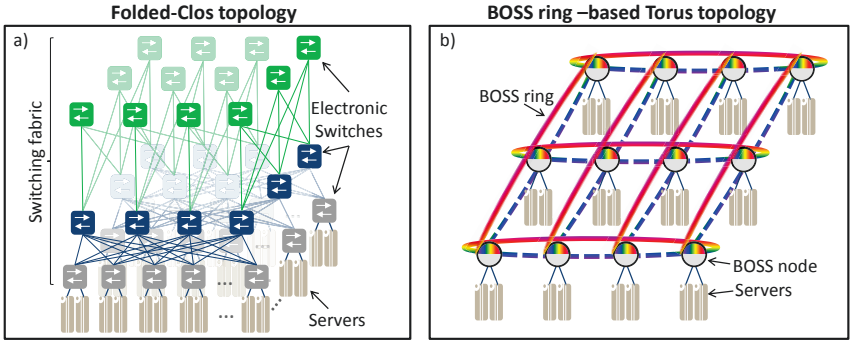


Fig. 6: (a) Topologie Folded Clos traditionnelle et (b) topologie BOSS.

de câbles, diminuant ainsi la complexité du réseau, et donc les coûts de développement et d'exploitation (réponse au défi 2); permet de réduire la consommation électrique par un facteur 2 à 3 par rapport aux réseaux tout électroniques actuels (défi 3) et a le potentiel de diminuer la latence (défi 4).

Comme le montre la Fig. 7, un nœud BOSS gère typiquement le trafic traversant plusieurs anneaux. À travers chaque anneau, des canaux de données à longueurs d'onde multiples (paquets colorés) se propagent avec un canal de contrôle (flux gris) qui transporte les en-têtes des paquets (y-compris, par exemple, les nœuds sources et de destination) et des informations de planification et de gestion de réseau. Les nœuds BOSS comprennent trois éléments matériels principaux: le récepteur en mode rafale (BM-RX), le bloqueur de slot et l'émetteur (BM-TX). Lors de l'arrivée à un nœud, tous les canaux de données traversent un répartiteur de puissance, qui répartira toutes les longueurs d'onde vers les récepteurs et le bloqueur de slots. Les BM-RX utilisent des lasers à accord rapide pour sélectionner, par détection cohérente, la longueur d'onde désirée. De plus, les BM-RX utilisent des algorithmes adaptatifs rapides capables de récupérer les paquets courts (quelques microsecondes).

Le bloqueur de slots est utilisé principalement pour effacer les slots

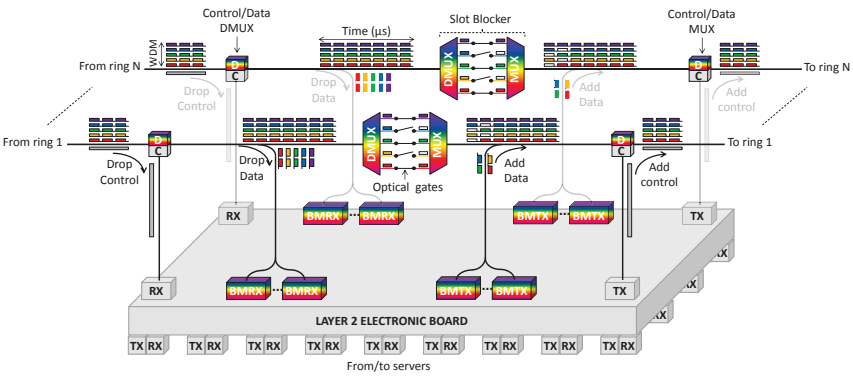


Fig. 7: Schema d'un nœud BOSS.

sur les longueurs d'onde qui ont déjà été reçues pour réutilisation ultérieure. Une manière possible de mettre en œuvre un tel dispositif est représenté sur la Fig. 7. Dans un tel schéma, un démultiplexeur de longueur d'onde (DMUX) est utilisé pour séparer toutes les longueurs d'onde. Ensuite, chaque longueur d'onde atteint une porte optique, ce qui permet de laisser passer ou d'effacer certains slots. Enfin, un multiplexeur de longueur d'onde (MUX) combine tous les canaux. Avant de quitter le nœud, les BM-TX insèrent de nouveaux paquets de données dans les slots disponibles.

Dans le Chapitre 4, nous explorons plusieurs approches pour l'implémentation physique des nœuds BOSS, tout en étudiant les dégradations dépendantes de la technologie qui se produisent lors de la traversée d'un grand nombre de nœuds. Tout d'abord, nous évaluons l'utilisation des amplificateurs optique à semi-conducteurs (SOA) comme portes optiques dans les nœuds BOSS, qui sont intéressants pour leur capacité d'amplification. Dans la Fig. 8, nous montrons le facteur Q^2 en fonction de la puissance d'entrée dans chaque SOA (P_{inSOA}) évalué le long d'une grande cascade de SOA en utilisant différents formats de modulation : QPSK, 8-QAM et 16-QAM. Nous pouvons observer que même pour le format de modulation le plus robuste (QPSK), les distorsions non linéaires induites par les SOA limitent la portée à environ 40 nœuds, ce qui

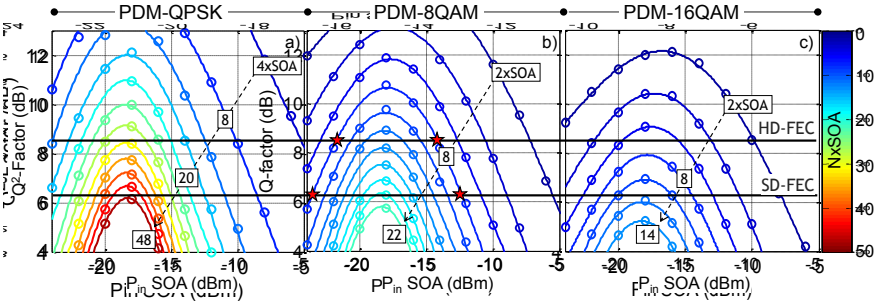


Fig. 8: Facteur Q^2 en fonction de la puissance d'entrée dans chaque SOA (P_{inSOA}) évalué le long d'une grande cascade de SOA en utilisant différents formats de modulation: (a) QPSK, (b) 8-QAM et (c) 16-QAM.

conduit à une mauvaise évolutivité du centre de données.

Par conséquent, nous nous dirigeons vers des portes optiques à base d'atténuateurs optique variables (VOA), qui n'introduisent pas de distorsions non linéaires et peuvent être intégrés monolithiquement avec d'autres dispositifs sans avoir besoin de plates-formes hybrides.

Ensuite, nous étudions l'impact de la cascade de différents types de de/multiplexeurs de longueur d'onde (D/MUX) pour plusieurs configurations de grille pour différents formats de modulation entre QPSK et 32-QAM, et des formes d'impulsions NRZ ou Nyquist (NPS). La Table 1 montre le débit de données brut et net réalisable

			Portée (nombre de nœuds)		
Format de modulation	Débit brut (Gb/s)	Débit net (Gb/s)	NRZ WSS à grille de 100 GHz	NRZ WSS à grille de 100 GHz	NPS WSS à grille de 50 GHz
32-QAM	325	250	20	14	15
16-QAM	260	200	40	27	40
8-QAM	195	150	60	40	53
QPSK	130	100	>100	79	>100

Table 1 : Portée des formats de modulations mesurées pour différentes configurations de D/MUX et formes d'impulsion.

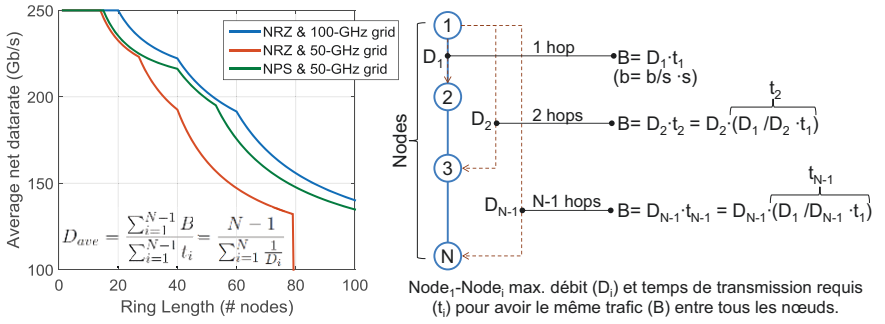


Fig. 9: Débit de données moyen en fonction de la longueur des anneaux en supposant une distribution de trafic uniforme.

pour chaque format de modulation et la portée pour différentes formes d'impulsions et configurations des D/MUX.

Par la suite, on montre dans la Fig. 9 le débit net moyen de données réalisable en fonction de la longueur de l'anneau (en nombre des nœuds) pour les différentes configurations. Nous démontrons que, lorsqu'on utilise des nœuds à commutation sélective de longueur d'onde (WSS) de 50 ou 100 GHz, les transpondeurs N-QAM flexibles peuvent atteindre un nombre de nœuds très élevé (plus de 100 nœuds) et des débits nets moyens supérieurs à 200, 150 et 100 Gb/s pour des anneaux de 40, 70 et 100 nœuds, respectivement.

Néanmoins, lors du passage à des D/MUX à faible coût, tels que *array waveguide gratings* (AWG) à grille de 100 GHz, le filtrage accumulé affecte sévèrement les performances des signaux NPS N-QAM, présentant des portées limitées à 70 nœuds, voir la zone grise et les points jaunes sur la Fig. 10(a) indiquant la portée de différents formats de modulation, et un débit moyen 30% inférieur par rapport aux nœuds à base de WSS, voir lignes bleues grises sur la Fig. 10(b).

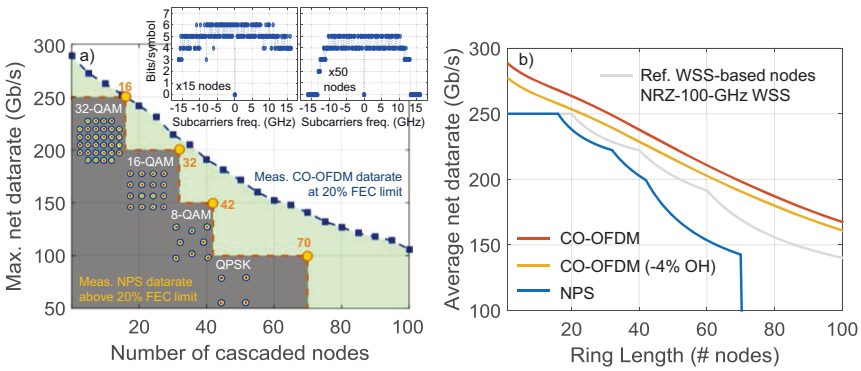


Fig. 10: (a) Comparaison du débit de données net mesuré expérimentalement au-dessus de la limite FEC par rapport au nombre de nœuds et (b) du débit net moyen de données par rapport à la longueur de l'anneau pour les signaux CO-OFDM et NPS. (a) montre des exemples de paramètres de chargement en bits.

Afin de surmonter de telles limitations, nous proposons l'utilisation d'un multiplexage orthogonal par répartition en fréquences cohérent (CO-OFDM), qui est capable de s'adapter de façon spectrale à la plupart des canaux de fréquence détériorés par l'optimisation du chargement en bits des sous-porteuses, conduisant à des capacités améliorées, voir la ligne pointillée bleue et la zone verte sur la Fig. 10(a), indiquant la capacité maximale par nœud et le gain de capacité, respectivement. Comme représenté sur la Fig. 10(b), nous démontrons que CO-OFDM peut parcourir des anneaux basés sur AWG avec un débit de données moyen supérieur à celui des signaux NPS ou NRZ lorsqu'ils passent par des nœuds haut de gamme basés sur WSS. Lorsqu'on compare NPS et CO-OFDM dans un environnement à faible coût, ce dernier fournit un taux de données moyen augmenté de 30% et une portée de 40% plus élevée, ce qui permet l'utilisation de composants à faible coût sans encourir des pénalités supplémentaires.

En résumé, nous proposons dans cette thèse un ensemble de solutions pour surmonter les limitations des centres de données actuels à court et à long terme. Nous répondons d'abord à la demande urgente d'interfaces optiques à haut débit et à faible coût ; puis nous fournissons une solution disruptive fondée sur des anneaux à commutation optique qui aborde les plus grands défis des centres de données tels que l'évolutivité, la complexité, la consommation d'énergie et la latence.