Facebook repense l'architecture réseau de son dernier datacenter

Quand on s'appelle Facebook la question de la topologie réseau prend un sens particulier lors de la construction d'un datacenter. Les problématiques sont connues, un grand volume de données et de requêtes qui circulent tous les jours, un besoin de haute disponibilité en temps réel, la prise en charge des liens vers des clients et prestataires externes. Le réseau social a donc repensé son architecture traditionnelle de réseau avec l'extension de son datacenter à Altoona dans l'Iowa.

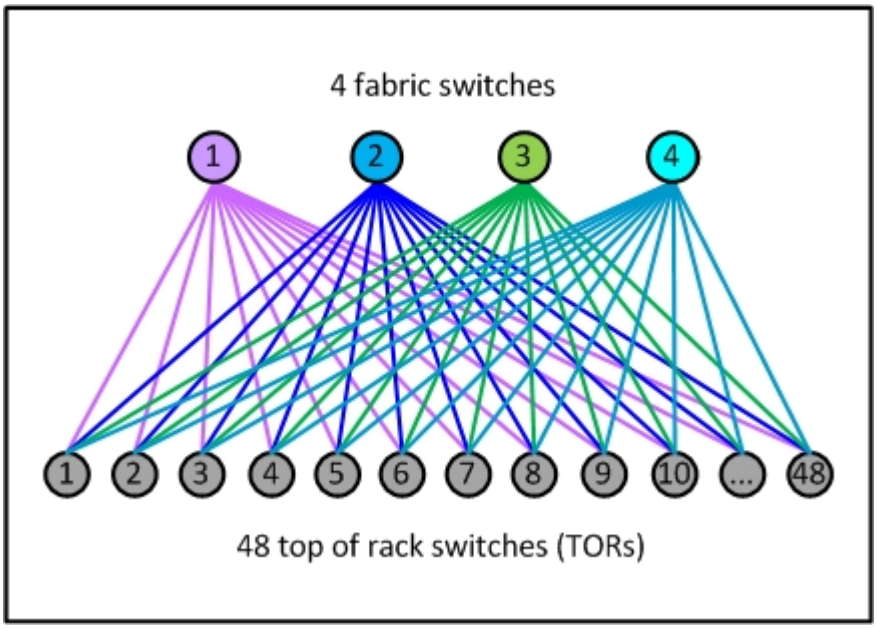
Ce dernier qui est alimenté par des énergies 100% renouvelables, via un parc éolien, a construit son réseau sur la technologie dite « data center fabric ». Alexey Andreyev, ingénieur réseau chez Facebook, explique dans un blog qu'il faut distinguer « *le trafic « machine to user » qui se définit comme les requêtes, la création de contenu ou d'applications est très important, mais il s'agit de la partie émergée de l'iceberg. Le trafic « machine to machine » au sein du datacenter est en croissance exponentielle et le volume double en moins d'une année* ». C'est sur ce dernier que datacenter fabric apporter une solution.

La limite des clusters

Pour faire face à ce niveau de trafic, Facebook avait habituellement une architecture sous forme de cluster de racks de serveurs avec un switch top of rack pour agréger différents commutateurs à forte densité de ports. Or cette topologie pose quelques problèmes en matière d'évolutivité, « *la taille du cluster est limitée par le nombre de ports des commutateurs* », précise l'ingénieur réseau. De même, il existe peu de produits fournis par les équipementiers capables de répondre aux exigences de la société de Menlo Park en termes de bande passante, mais aussi de maintenance opérationnelle du réseau.

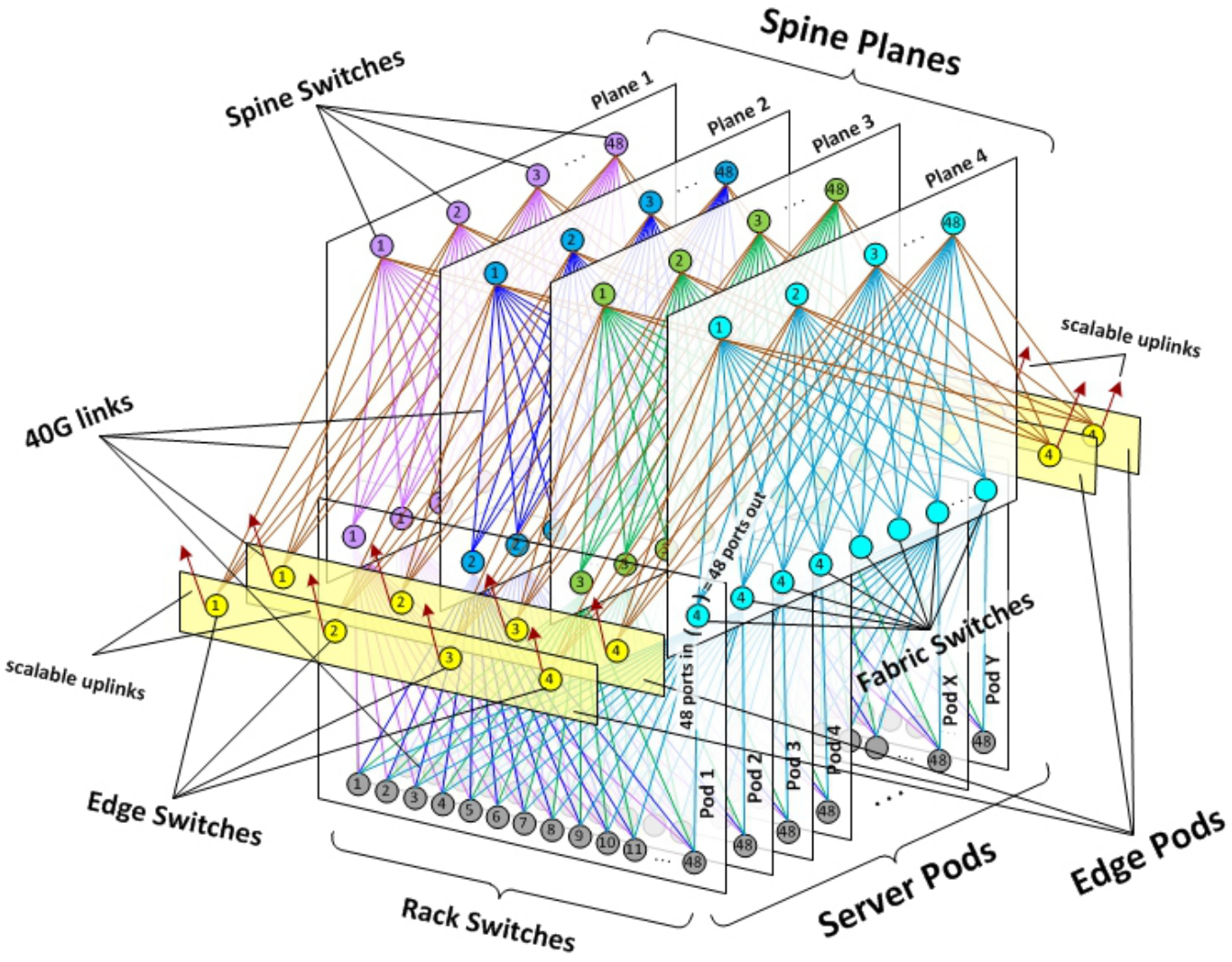
Un saucissonnage en pod de 48 racks serveurs

D'où l'idée d'avoir une approche désagrégée de l'architecture réseau, « *au lieu d'avoir des clusters avec beaucoup d'équipements réseaux, nous avons cassé le réseau en plusieurs petites unités identiques, des serveurs pods et créer une connectivité haute performance entre les pods au sein du datacenter* ». La taille de ces petites unités est de 48 racks de serveurs. Chaque Pod est relié à 4 commutateurs fabric (spécialement élaboré par Facebook) avec des liens 40 G permettant d'atteindre une capacité de bande passante de 160 G pour un rack de serveurs connectés en 10 G.



Une partie software adaptée

L'avantage de cette solution modulaire est de pouvoir la répliquer au sein du datacenter et de requérir que des commutateurs basiques pour gérer l'agrégation top of rack. Cette simplicité autorise plusieurs options de routage via le protocole BGP 4 (le seul retenu par l'entreprise). Dans le même temps, Facebook souligne avoir travaillé sur un contrôleur BGP centralisé qui permet de contourner les chemins de routage via un logiciel. Cette approche flexible est dénommée « DCCO (distributed control, centralized override) ». Cela signifie aussi que le réseau social a élaboré son propre logiciel de gestion et de configuration du datacenter fabric. Quand la firme veut intégrer un nouvel équipement, il est automatiquement reconnu et configurer. Idem en cas de problème, cela ressemble au même processus que le décommissionnement d'une machine virtuelle.



A lire aussi :

[Réseau : Facebook nargue Cisco et Juniper avec son switch Wedge]

[Facebook renforce la sécurité des datacenters avec PrivateCore]

Crédit Photo: Facebook