2ème année 2023-2024

# Réseaux TCP/IP

9 Novembre 2023

▷ **Exercice 1 : DataCenter** TCP

Data Center TCP *est une proposition de modification de* TCP *visant à résoudre certains problèmes apparaissant en particulier dans les* datacenters *[1].*

*De façon simplifiée, la situation peut être décrite comme suit. Un grand nombre de serveurs sont interconnectés au travers d'un commutateur ethernet. Ils mettent en place la stratégie suivante : l'un d'entre eux (appelé ici* aggregator*) reçoit une requête d'un client et décompose le travail en un certain nombre de requêtes transmises à autant d'autres serveurs (appelés ici* workers*). Ces derniers réalisent le travail puis envoient le résultat en retour à l'*aggregator *qui peut alors répondre au client. Ces requêtes et réponses sont de petite taille (1 ou deux paquets) et nécessitent une latence aussi faible que possible (typiquement de l'ordre de 200 ms).*

*Parallèlement, du traffic est transmis aux serveurs afin d'alimenter en permanence leurs bases de données de sorte à maintenir la qualité des réponses. Ce traffic est très exigeant en débit.*

*Le commutateur ethernet dispose d'une mémoire limitée partagée par tous les liens. Pour des raisons de simplicité et d'équité, chaque lien est doté d'une file d'attente ne pouvant contenir qu'un nombre limité de trames.*

*Le protocole de transport utilisé est une variante de* TCP *mais pourrait également être fondé sur* UDP*. Le but de cet exercice est d'étudier* DCTCP*, une proposition de modification de* TCP *optimisée pour un tel contexte. Nous suivrons pour cela la logique et les notations du papier décrivant ce protocole [1].*

*Trois types de problèmes peuvent apparaître dans les files d'attente du* switch*.*

**1.1** ***Incast*** *— Le premier problème (appelé* Incast*) ne concerne que les communications entre* aggregator *et* workers *: malgré le faible volume des échanges, des paquets peuvent être perdus.*

*Comment cela peut-il s'expliquer ?*

*Chaque perte entraine une retransmission sur* time out *dont le délai est de 300 ms. Les échéances requises par l'application ne peuvent donc pas être respectées*

*Pourquoi* TCP *n'arrive pas à retransmettre plus rapidement ?*

*Quelles solutions pourrait-on imaginer pour résoudre ce problème d'*Incast *?*

*Réponses :*

*(suite)*

**1.2 Saturation des files** —

*Le deuxième problème survient lorque les deux types de traffic coexistent. Les réponses des workers subissent alors des retards relativement importants, sans nécessairement subir de pertes.*

*Expliquer ce phénomène. Quelles solutions sont envisageables ?*

*Réponses :*

*Le troisième problème est lié à la gestion globale de la mémoire du switch, nous n'en parlerons pas ici.*

*Les principes de* DCTCP *sont simples. Le commutateur est configuré pour positionner un bit*[1] *dans l'entête* IP *de chaque paquet lorsque le taux d'occupation instantané de la mémoire est supérieur à une valeur $K$ fixée.*

*Pour chaque segment reçu, le récepteur* DCTCP *transmet dans l'accusé de réception correspondant la valeur de ce bit*[2].

*L'émetteur* DCTCP *reçoit donc une séquence de bits correspondant à l'état de la mémoire du*

1. On utilise pour cela le bit CE d'ECN dont DCTCP peut être vu comme une variante.
2. Une fois encore, on utilise de façon un peu détournée l'intégration de ECN dans TCP.

*switch "observée" par la séquence des segments émis. Il calcule alors sur chaque fenêtre le taux $\alpha$ des paquets marqués (par un bit à 1) dans la dernière fenêtre.*

*$\alpha$ peut donc être vu comme un indicateur de la probabilité que l'occupation de la mémoire du switch soit supérieure à $K$.*

*L'émetteur* DCTCP *met alors à jour* CWND *de la façon suivante*[3]

$$cwnd \leftarrow cwnd.(1 - \alpha/2)$$

**1.3 Évolution de la taille de fenêtre** — *Supposons que $N$ connexions* DCTCP *de longue durée avec un même* RTT *soient en concurrence sur un lien de débit $C$.* **Imaginons qu'elles se soient synchronisées,** *c'est-à-dire que leurs tailles de fenêtre croissent ensemble additivement (en* congestion avoidance*) puis décroissent multiplicativement (avec le coefficient $1 - \alpha/2$) au même instant.*

*La taille de la fenêtre de chaque connexion suit alors une évolution telle que celle décrite par la figure 1 (gauche). Nous la supposerons périodique de période $T_c$ et d'amplitude $D$.*

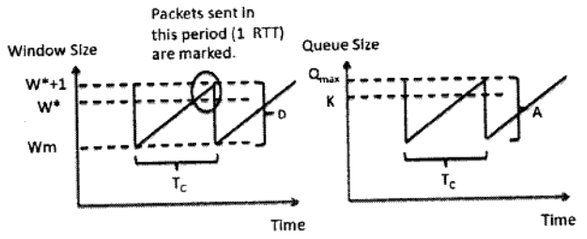


FIGURE 1 – Évolution de la taille de fenêtre d'une connexion (gauche) et de l'occupation de la file du *switch* (droite)

*On cherche une expression approchée de $S(W_1, W_2)$, le nombre de segments émis par une connexion lorsque la taille de fenêtre croît de $W_1$ à $W_2$ (deux valeurs quelconques). Pour cela :*

- *Combien de* RTT *sont nécessaires pour passer de $W_1$ à $W_2$ ?*
- *Quelle est la valeur moyenne de la taille de la fenêtre sur cette période ?*
- *En déduire une expression simple de $S(W_1, W_2)$.*

*Réponses :*

3. Tous les autres mécanismes de TCP sont préservés.

*(suite)*

*Appelons $W^* = (C.RTT + K)/N$ la taille de fenêtre à partir de laquelle l'occupation de la mémoire dépasse $K$. Le switch va donc marquer les paquets envoyés après le dépassement de ce seuil par l'émetteur.* DCTCP *pourra donc réagir au bout de 1* RTT *(donc lorsque la taille de fenêtre sera de $W^* + 1$).*

**1.4** **Calcul de $\alpha$** — *Lorsque* DCTCP *va réagir, quelle nouvelle valeur va-t-il alors donner à la taille de la fenêtre ? Appelons $W_m$ cette valeur.*

*Quel est le nombre (exprimé avec la fonction $S(.,.)$) de segments émis sur une période $T_c$ ?*

*Quel est le nombre de ces segments qui sont marqués par le switch (exprimé avec la fonction $S(.,.)$) ?*

*Utiliser ces deux nombres pour établir une équation dont $\alpha$ est solution.*

*En supposant que $\alpha$ soit très faible, en donner une approximation simple en fonction de $W^*$.*

*Réponses :*

*Du fait du comportement des diverses connexions, le taux d'occupation de la file d'attente du commutateur ethernet suit lui aussi une évolutoin en dent de scie. Ce comportement, illustré par la partie droite de la figure 1, est d'une amplitude désignée par $A$ que l'on cherche maintenant à calculer.*

**1.5 Conséquences sur la mémoire du commutateur ethernet** — *D'après la question précédente, quelle est la valeur de $D$, l'amplitude de la taille de fenêtre d'une connexion ? On en donnera une expression exacte en fonction de $\alpha$ et $W^*$ et une valeur approchée en fonction de $W^*$.*

*En déduire une valeur de $A$ que l'on exprimera en fonction des paramètres du système : $N, C, RTT, K$.*

*Réponses :*

**1.6 Comparaison avec TCP** — *Différents modèles de TCP montrent que les oscillations de taille de file induites par ce dernier sont en $O(C.RTT)$.*

*Quelle conclusion peut-on en tirer sur DCTCP ?*

*Réponses :*

**1.7 Pertinence du modèle** — *La figure 2, (issue de [1]) compare les résultats théoriques de DCTCP fournis par le modèle que nous venons de développer à ceux obtenus par simulation pour différentes valeurs de $N$. Commenter ces résultats en expliquant pourquoi les différences sont de*

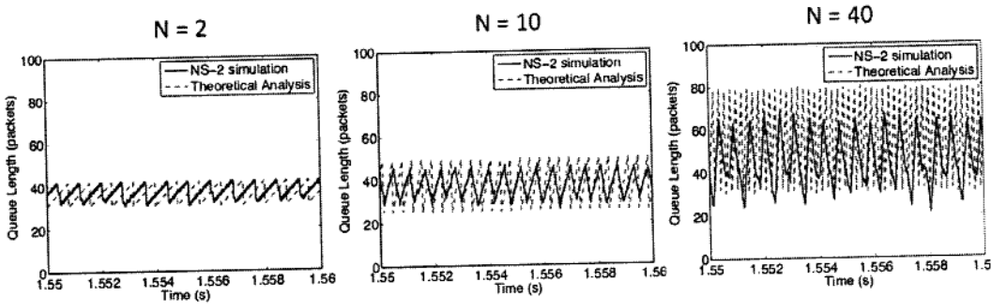


FIGURE 2 – Occupation de la file

*plus en plus grandes entre les deux modèles? Est-ce génant que le modèle théorique ne donne pas les bonnes valeurs?*

***1.8* Analyse des résultats** — *La figure 3 reprend les principaux résultats publiés dans [1]. Que peut on conclure sur* DCTCP *?*

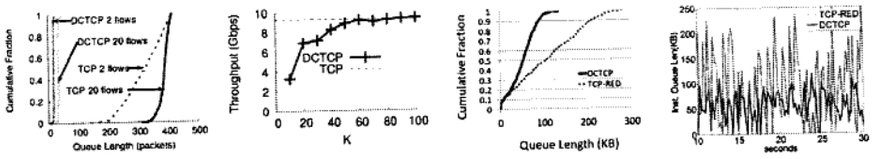


**Figure 13: Queue length CDF (1Gbps)** **Figure 14: Throughput (10Gbps)** (a) CDF of queue length (b) Time series of queue length **Figure 15: DCTCP versus RED at 10Gbps**

FIGURE 3 – Mesures par simulation

*Réponses :*

# Références

*[1] Mohammad Alizadeh, Albert Greenberg, David A. Maltz, Jitendra Padhye, Parveen Patel, Balaji Prabhakar, Sudipta Sengupta, and Murari Sridharan.* Data center tcp (dctcp). SIGCOMM Comput. Commun. Rev., *40(4) :63–74, August 2010.*

■

▷ **Exercice 2 : MultiPath** TCP

MP-TCP *est une extension de* TCP *visant à utiliser plusieurs chemins (et plusieurs adresses) pour une même connexion* TCP. *Pour des raisons de compatibilité, elle est implantée notemment par le biais d'une option et offre aux applications la même interface que* TCP.

**2.1** **Performances** — *Quels peuvent être les apports d'une telle extension en termes de performances et fonctionnalités offertes à l'application ?*

*Réponses :*

**2.2** **Difficultés d'implantation** — *Décrire comment les principaux mécanismes de* TCP *peuvent être impactés par une telle extension.*

*Quelles considérations doivent être prises en compte pour les implanter de façon compatible avec cette extension ?*

*Réponses :*

▷ **Exercice 3 : Synthèse**

*Sur la machine* `graham.enseeiht.fr`*, un utilisateur tape la commande* `ping www.ensica.fr`.

*En supposant un démarrage à froid, que tout fonctionne correctement et que tous les éléments impliqués soient correctement configurés pour permettre à cette commande d'aboutir, décrire l'enchainement des différents protocoles impliqués.* ■

Réponses :