Προγραμματισμός και **Συστήματα στον Παγκόσμιο Ιστό**

Κεφάλαιο 4:
Κρυφές Μνήμες Αντιπροσώπων:
Προσδοκίες, Μυστικά και Παγίδες

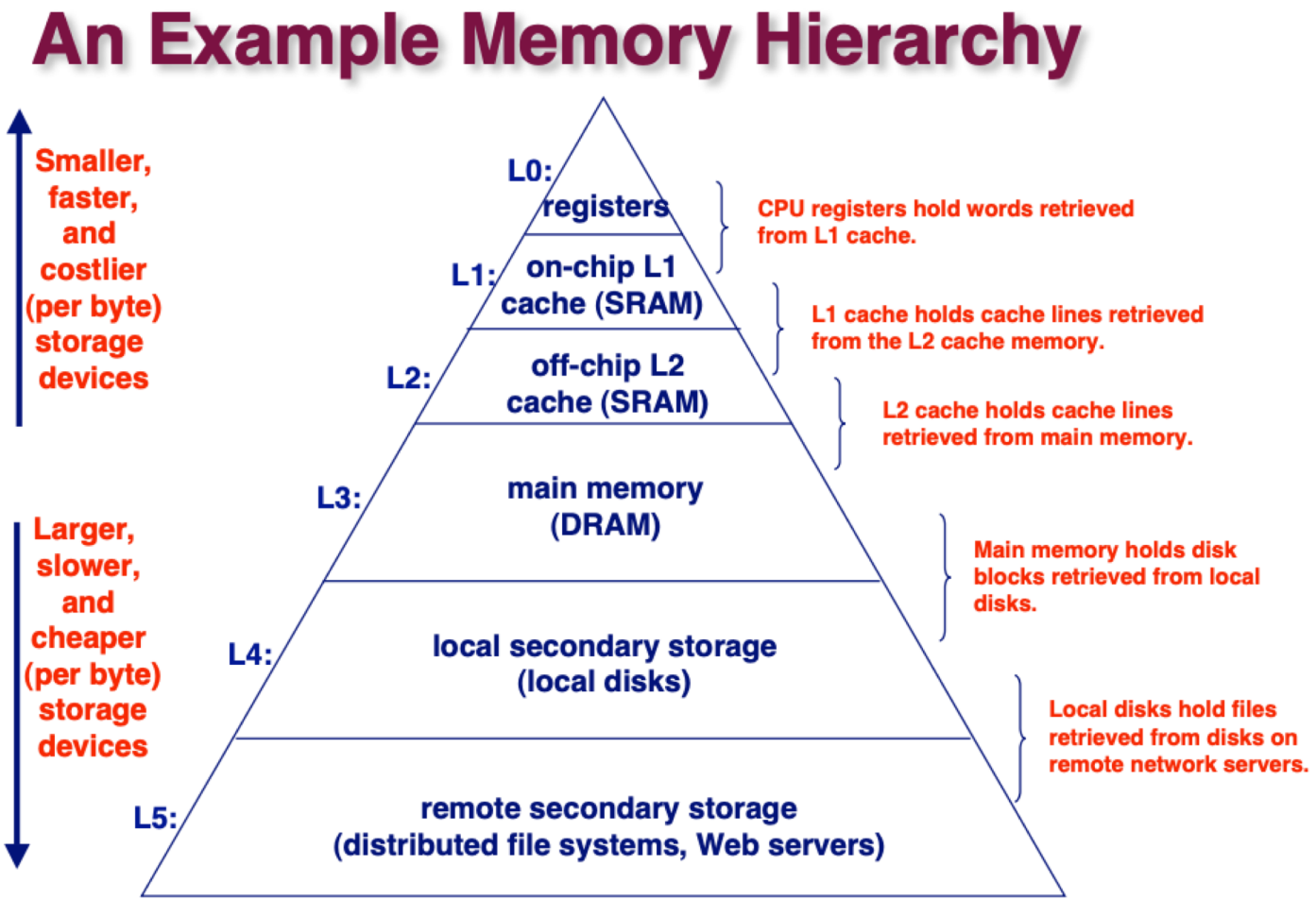
Ωφέλη

- Χρήστες:
  - Άμεσα: μειωμένος χρόνος απόκρισης (καθυστέρηση - latency)
  - Έμμεσα: μειωμένη κατανάλωση εύρους ζώνης (bandwidth) ➔ λιγότερα σημεία συμφόρησης, ...
- Εταιρείες:
  - μειωμένη κατανάλωση εύρους ζώνης ➔ μικρότερο κόστος αγοράς εύρους ζώνης από ΕΠΥΔ
  - Ύπαρξη αντιπροσώπων στις εταιρείες ➔ δυνατότητα να ελέγχουν προσπελάσεις εργαζομένων στον Ιστό ... ➔ αύξηση παραγωγικότητας ...

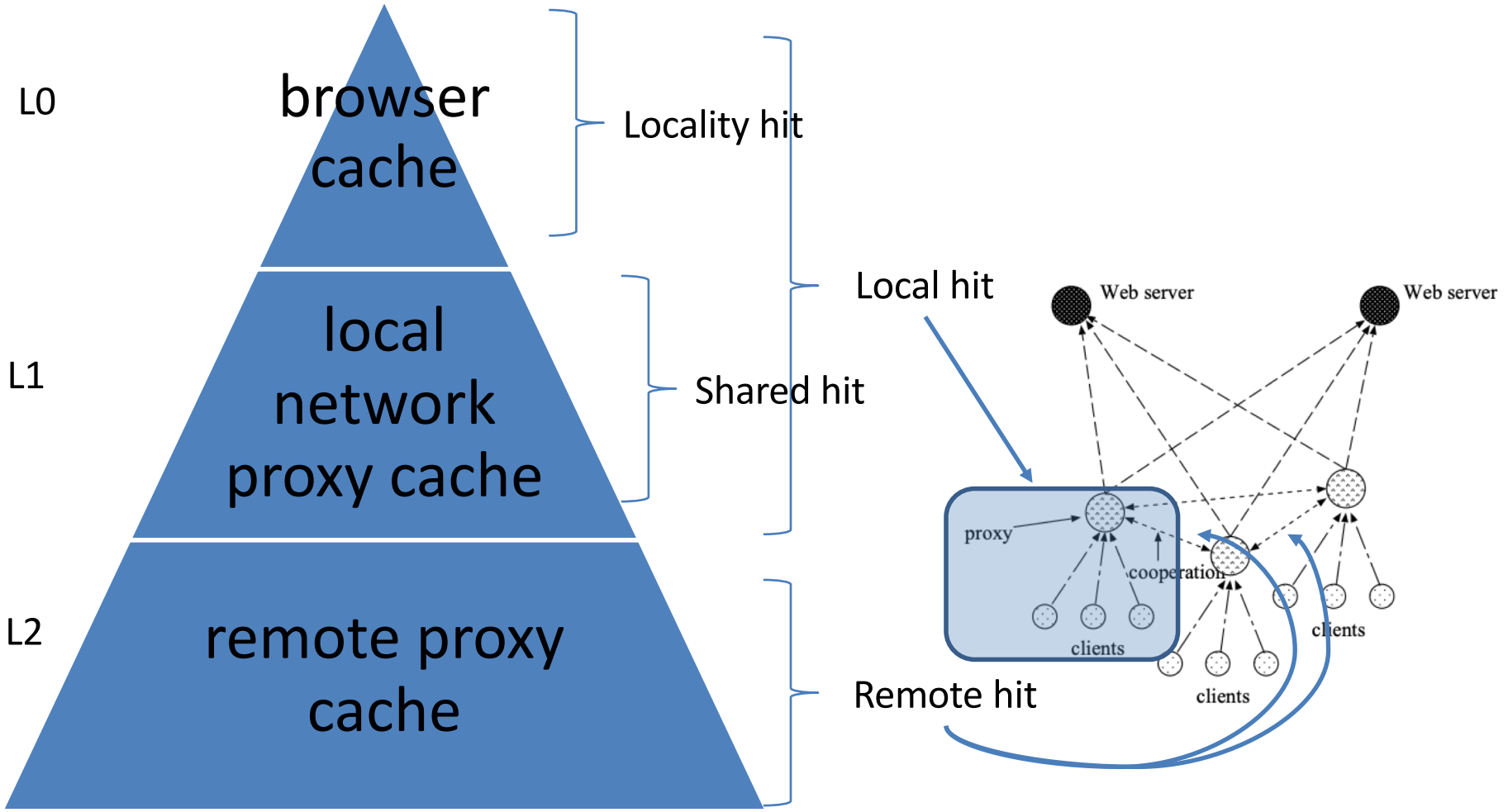
Ωφέλη

- Εταιρείες Παροχής Υπηρεσιών Διαδικτύου (ISPs):
  - αν αιτήσεις χρηστών εξυπηρετούνται νωρίς στη ραχοκοκαλιά του δικτύου της ΕΠΥΔ
    - ➔ μειωμένη κατανάλωση εύρους ζώνης
    - ➔ δυνατότητα υποστήριξης περισσότερων πελατών
    - ➔ περισσότερα έσοδα
  - Αν ερωτήσεις πελατών εξυπηρετούνται από αντιπροσώπους στο δίκτυο της ΕΠΥΔ
    - ➔ μικρότερη κίνηση εισρέει από άλλες ΕΠΥΔ
    - ➔ μικρότερο κόστος λειτουργίας για την ΕΠΥΔ!
  - Επίσης, με αντιπροσώπους στο δίκτυο της ΕΠΥΔ, η ΕΠΥΔ προσφέρει καλύτερης ποιότητας υπηρεσίες
    - Εξαρτάται λιγότερο από την ποιότητα υποδομών άλλων ΕΠΥΔ και εξυπηρετητών-πηγών

Κρυφές μνήμες



Κρυφές μνήμες



Τοπολογίες

- Ιεραρχική
  - Τοπική -> οργανισμός -> περιοχή -> χώρα
  - Τοποθέτηση σε σημεία κλειδιά του δικτύου (προσοχή στο σχεδιασμό)
  - Κάθε επίπεδο συνεισφέρει πρόσθετη καθυστέρηση
  - Τα υψηλά επίπεδα προκαλούν «μποτιλιάρισμα»
  - Τα ίδια αντικείμενα σε πολλαπλούς κόμβους

Τοπολογίες

- Κατανεμημένη
  - Μόνο σε επίπεδο οργανισμού
  - Κάθε ΚΜ συνεργάζεται με τις άλλες
  - Πρέπει να γνωρίζει το περιεχόμενό τους
  - Καλύτερη εξισορρόπηση φόρτου εργασίας
  - Αλλά περισσότερη καθυστέρηση (συνδέσεις), υψηλή χρήση εύρους ζώνης, διαχείριση κλπ.
- Υβριδική
  - Συνδυασμός ιεραρχικής και κατανεμημένης

Μετρικές

- Τα ωφέλη κυρίως προκύπτουν από τη μείωση στην καθυστέρηση (χρόνο απόκρισης) και στο εύρος ζώνης.
- Ποιές είναι οι μετρικές που κυρίως χρησιμοποιούνται για να μετρήσουν αυτά τα ωφέλη κρυφών μνημών;
  - Hit Rate: Ρυθμός χτυπημάτων (ΡΧ) της κρυφής μνήμης
    - Ποσοστό αιτήσεων που εξυπηρετούνται από την ΚΜ
  - Byte Hit Rate: Ρυθμός χτυπημάτων ψηφίων (ΡΧΨ) της κρυφής μνήμης
    - Ποσοστό των ψηφίων που ανακτήθηκαν από την ΚΜ
- Προσοχή:
  - Αυτό που ενδιαφέρει είναι όμως η **πραγματική** μείωση στην καθυστέρηση και στο εύρος ζώνης και όχι η **εικαζόμενη** μέσω ΡΧ, ΡΧΨ.

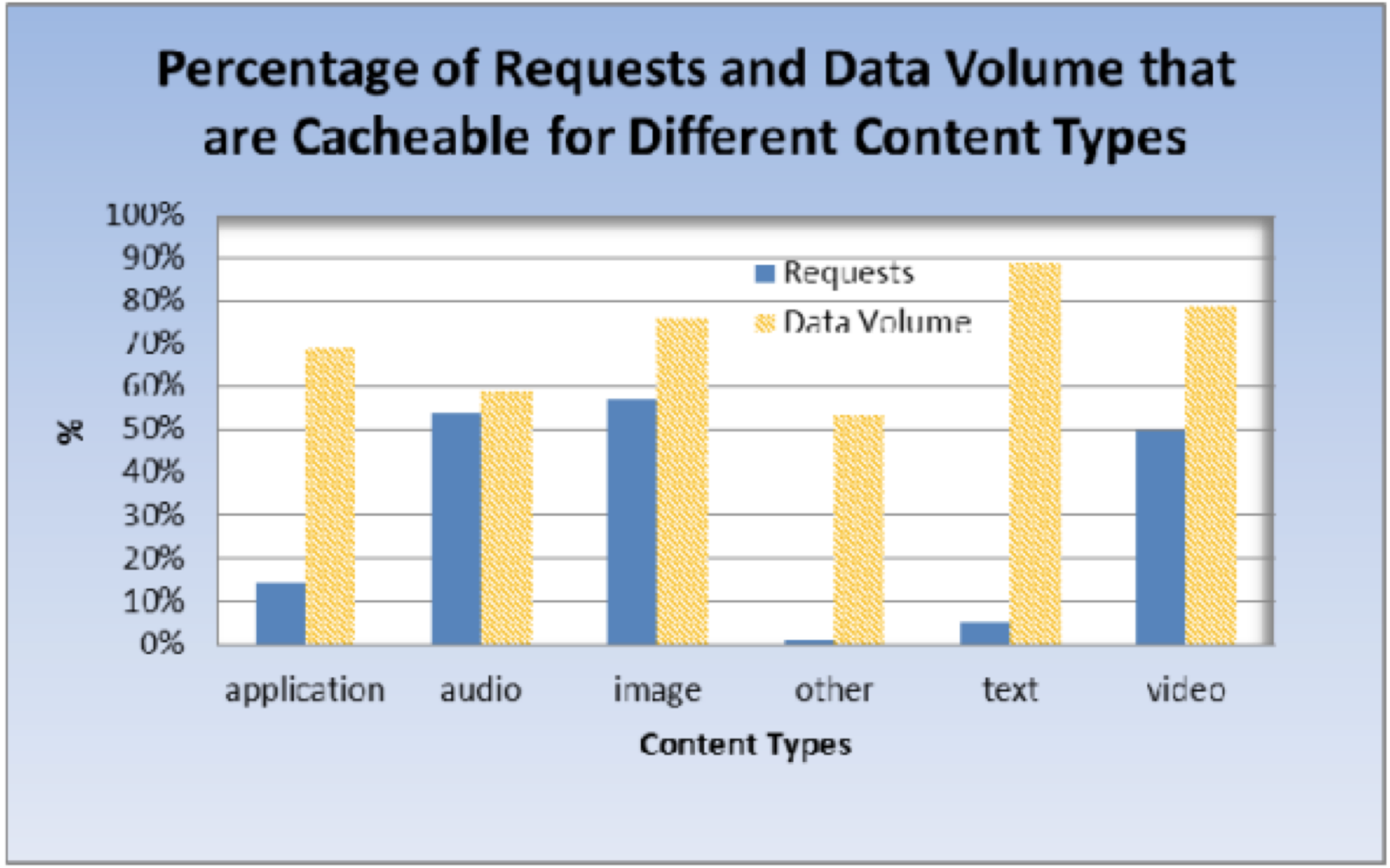
Πότε βοηθούν οι ΚΜ στους Αντιπροσώπους

- Γιατί μια μεγάλη ΚΜ στους Περιηγητές δεν αρκεί;
  - Τοπικά χτυπήματα (locality hits): χτυπήματα για ίδιες σελίδες από τον ίδιο χρήστη
  - Διαμοιραζόμενα χτυπήματα (shared hits): χτυπήματα για ίδιες σελίδες από διαφορετικούς χρήστες
- Διαμοιραζόμενα χτυπήματα μπορούν μόνο να εξυπηρετηθούν από ΚΜ αντιπροσώπων
  - οι ΚΜ στους περιηγητές δεν αρκούν
  - Δεν είναι προσπελάσιμοι από άλλους χρήστες!
- Κομβική ερώτηση: είναι τα διαμοιραζόμενα χτυπήματα αρκετά, έτσι ώστε να δικαιολογούνται τα κόστη αντοπροσώπων;
  - Μελέτες από πραγματικές λίστες ερωτημάτων (traces) έδειξαν
    - τα διαμοιραζόμενα χτυπήματα είναι >80% όλων των χτυπημάτων
    - Ο ΡΧ και ΡΧΨ ήταν > 50%
  - Εμπειρικός κανόνας είναι ότι για να έχει νόημα μια διαμοιραζόμενη ΚΜ, απαιτείται ένας ΡΧ > 30-40 %.

Παράδειγμα

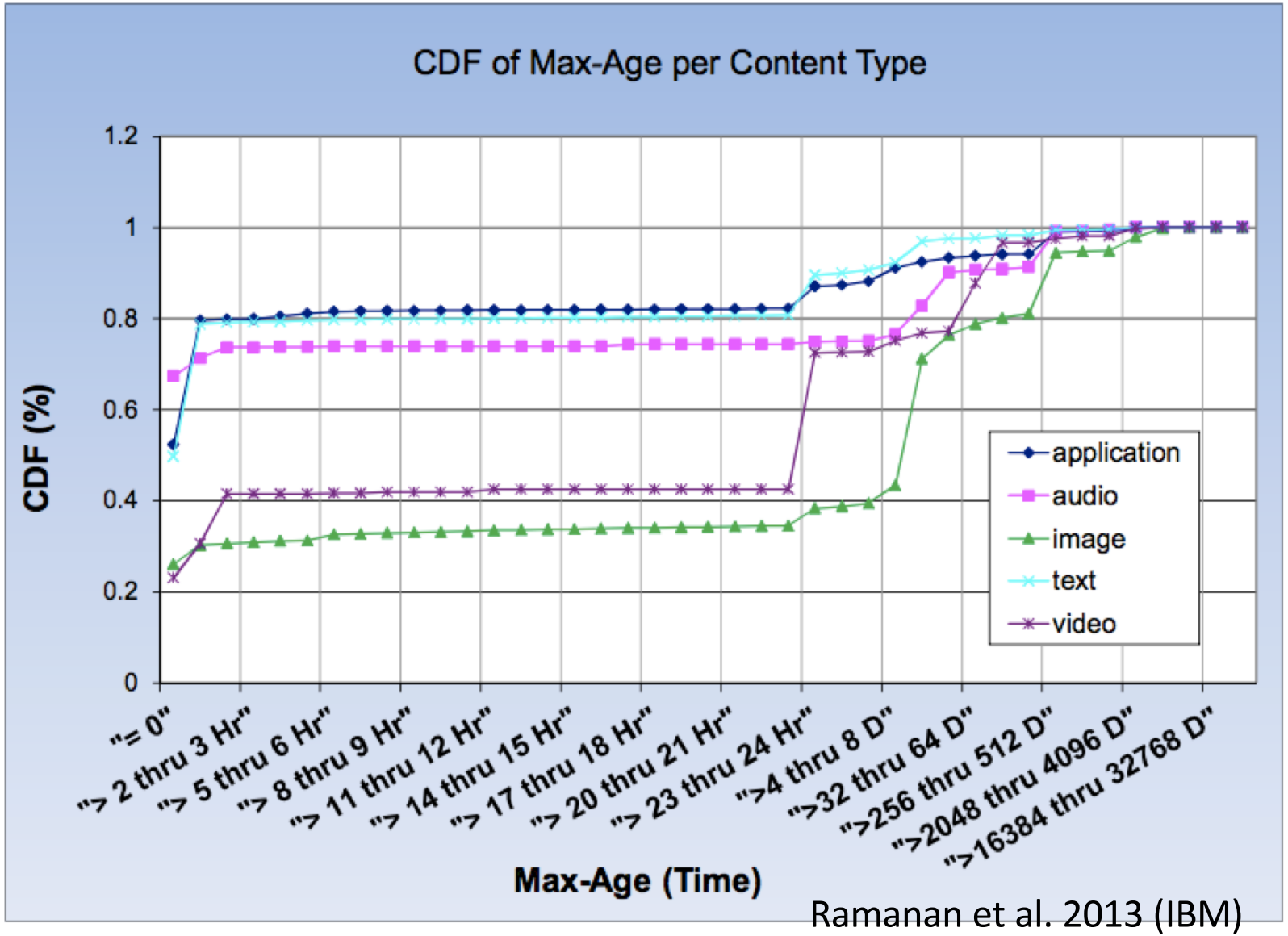
- Παραδοχές:
  - μέσο κόστος μετάδοσης είναι $150 per gigabyte,
  - Ο ISP «αγοράζει» 1000 gigabytes το μήνα από κάποιο πάροχο (upstream ISP)
  - 70% κίνηση διαδικτύου (ιστοαντικείμενα)
- Τότε
  - Μια ΚΜ που έχει 25% byte hit rate εξοικονομεί για τον ISP ένα επαναλαμβανόμενο κόστος $26,250 το μήνα.
- Αν
  - η ΚΜ κοστίζει $100,000 σαν κεφάλαιο κτήσης και $2000 το μήνα λειτουργικά κόστη,
- Τότε
  - η λειτουργία της ΚΜ έχει καθαρό όφελος περίπου $18,000 το μήνα για την επιχείρηση.

Trace ενός κινητού δικτύου (LTE)



Ramanan et al. 2013 (IBM)

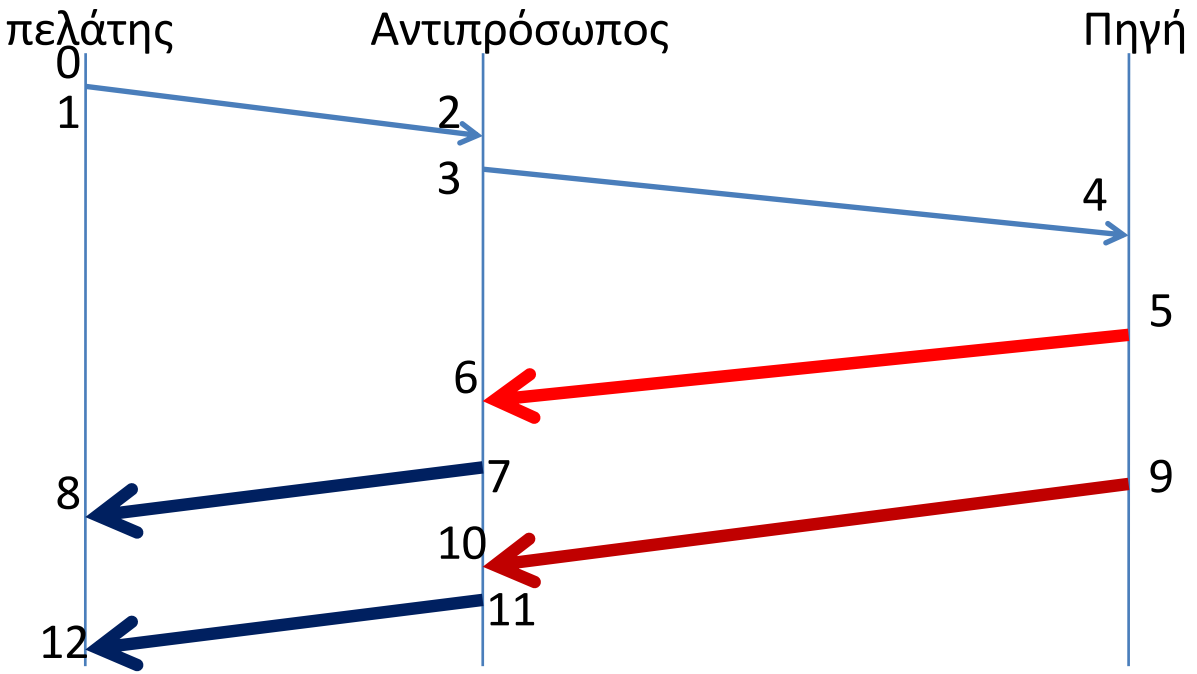
TTLs στο δίκτυο LTE



Ramanan et al. 2013 (IBM)

Καθυστέρηση (Latency)

- Εξωτερική καθυστέρηση: Μεταξύ Αντιπροσώπου-Πηγής:
  - Χρονικά σημεία: 3 - 10
- Εσωτερική καθυστέρηση: Χρονικά σημεία: 0-1 και 10-12



Καθυστέρηση – Αισιόδοξο Σενάριο

- Η εξωτερική καθυστέρηση συνεισφέρει >80% της συνολικής καθυστέρησης
- Με (εύστοχο) χτύπημα (shared hit) στην ΚΜ, γλυτώνουμε την εξωτερική καθυστέρηση
  - το μέγιστο όφελος προκύπτει όταν δεν υπάρχει επικάλυψη μεταξύ εξωτερικής και εσωτερικής καθυστέρησης (δηλ. τη γλυτώνουμε ολόκληρη)
- Αν υποθέσουμε ότι ένα άστοχο χτύπημα της ΚΜ (miss) , δεν κοστίζει παραπάνω ➔
  - Με ΡΧ ~50% θα εισπράταμε μείωση καθυστέρησης ~ 80% × 50% = ~40%.
- Μελέτες, όμως, έχουν δείξει ότι ακόμα και με αυτές τις παραδοχές, η βελτίωση στην καθυστέρηση ήταν 20-25%.
  - Που οφείλεται στο ότι τα πιο συχνά ζητούμενα ιστο-αντικείμενα είναι μικρού μεγέθους, που έχουν μικρότερη εξωτερική καθυστέρηση.
- ➔ όχι και τόσο ... Αισιόδοξο σενάριο...

Καθυστέρηση – Απαισιόδοξο Σενάριο

- Οι προηγούμενες παραδοχές ήταν υπεραισιόδοξες
  - Η TCP σύνδεση απαιτείται να γίνει
    - Σε χτύπημα, 1 σύνδεση
    - Σε αστοχία, 2 συνδέσεις
    - ➔ μικρότερα ωφέλη
  - Εσωτερική και εξωτερική καθυστέρηση επικαλύπτονται χρονικά
  - Όταν η σύνδεση πελάτη-αντιπροσώπου είναι αργή, τότε αυτή είναι το πρόβλημα ➔ το να γλυτώσουμε την εξωτερική καθυστέρηση είναι επουσιώδες
  - Μερικά αντικείμενα δεν μπορούν να αποθηκευτούν σε ΚΜ (un-cacheable).

Καθυστέρηση – Απαισιόδοξο Σενάριο

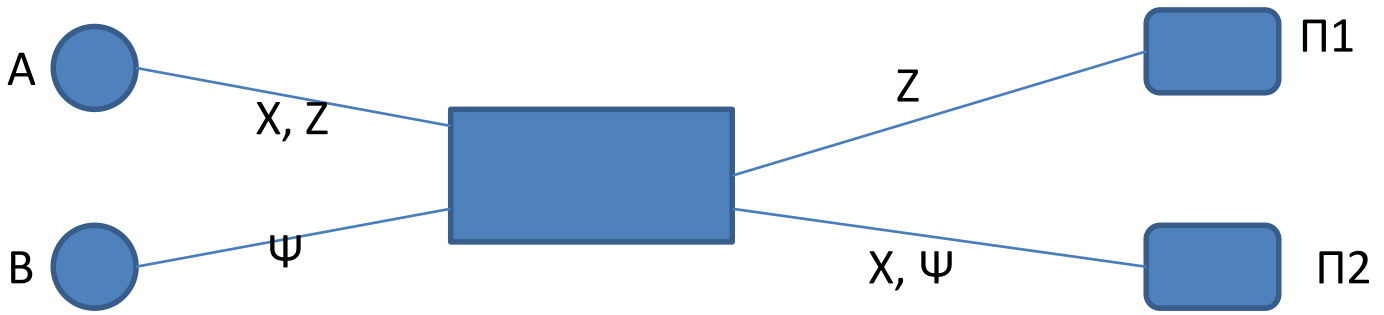
- Μελέτες που έλαβαν υπόψη τα παραπάνω
  - Έδειξαν ότι **ΚΜ δεδομένων (data caching)** συνεισφέρει αμελητέα ωφέλη (< 10%) κυρίως
    - Λόγω αργής σύνδεσης πελάτη-αντιπροσώπου
    - Λόγω έξτρα καθυστέρησης στον αντιπρόσωπο
- Και αυτές οι μελέτες όμως ήταν υπερ-Απαισιόδοξες: Δεν έλαβαν υπόψη
  - προβλήματα καθυστέρησης στο Διαδίκτυο (άρα κάποια ωφέλη ΚΜ δε φάνηκαν)
  - Κάποιες από τις συνδέσεις πελατών-πηγών επαναχρησιμοποιούσαν TCP συνδέσεις, ενώ στην περίπτωση Πελάτη-Αντιπρ.-Πηγής αυτό δεν ίσχυε.

Καθυστέρηση και Προσδοκίες: Συμπέρασμα

- Το αναμενόμενο είναι τα ωφέλη να κυμαίνονται ανάμεσα στην αισιόδοξη και στην απαισιόδοξη πρόβλεψη.
- Πάντως, οι προσδοκίες είναι χαμηλές!
- Ευτυχώς, η κατάσταση είναι πιο ... Ρόδινη.
- Προσέξτε: Αυτά αφορούν σε ΚΜ δεδομένων (κρύψιμο δεδομένων).
  - Τί άλλη ΚΜ θα μπορούσε να υπάρξει;

Καθυστέρηση: Κρύβοντας τις TCP Συνδέσεις

- TCP connection caching
- Σημαντικό κομμάτι της συνολικής καθυστέρησης
- Από το HTTP 1.1 επιτρέπεται η χρήση της ίδιας σύνδεσης TCP για μεταφορές πολλαπλών αντικειμένων
- Παράδειγμα:



Καθυστέρηση: Κρύβοντας τις TCP Συνδέσεις

Σενάριο 1:

1. Ο Α ζητά το αντικείμενο Χ από τον Π2 (μέσω του αντιπροσώπου).
2. Ο Β ζητά το αντικείμενο Ψ από τον Π2.
3. Χωρίς αντιπρόσωπο, ο Β πρέπει να χτίσει νέα σύνδεση TCP με τον Π2.
4. Με αντιπρόσωπο, όμως, αν έχει κρατήσει ανοικτή τη σύνδεση TCP,

   ➔ αποφεύγεται το σχετικό κόστος

Θυμηθείτε: επίμονες συνδέσεις

Καθυστέρηση: Κρύβοντας τις TCP Συνδέσεις

Σενάριο 2:

1. Ο Α ζητά το αντικείμενο Χ από τον Π2 (μέσω του αντιπροσώπου).
2. Ο Α ζητά το αντικείμενο Ζ από τον Π1.
3. ➔ ο Α μπορεί να επαναχρησιμοποιήσει τη σύνδεση TCP με τον αντιπρόσωπο
4. Επίσης, αν ο Β είχε ήδη ζητήσει το αντικείμενο Ψ από τον Π1, ο αντιπρόσωπος δε χρειάζεται να χτίσει νέα σύνδεση TCP με τον Π1

   ➔ αποφεύγεται το σχετικό κόστος

Καθυστέρηση: Κρύβοντας τις TCP Συνδέσεις

- Συναθροίζοντας αιτήσεις πελατών προς την ίδια πηγή, ένας αντιπρόσωπος ελαχιστοποιεί τον αριθμό των ανοικτών συνδέσεων TCP που πρέπει να διαχειρίζεται μια πηγή !
- Δηλαδή, αντί να κρατούμε και να χρησιμοποιούμε τα ίδια αντικείμενα (data) για μελλοντικές αιτήσεις, χρησιμοποιούμε την ίδια σύνδεση για μελλοντικές αιτήσεις →κρύψιμο σύνδεσης (connection caching).
- Επιπλέον, συνήθως, οι εξυπηρέτες-πηγές δεν δέχονται αιτήσεις για επίμονες συνδέσεις από πελάτες γιατί το διαχειριστικό κόστος γίνεται μεγάλο...
- Φυσικά, οι αντιπρόσωποι πρέπει να επαναχρησιμοποιούν μόνο ανενεργές συνδέσεις

➔ Οι Αντιπρόσωποι μπορούν να κρύβουν

- ➔ Δεδομένα ΚΑΙ συνδέσεις
- ➔ Όταν συναντάται ο μέγιστος αριθμός ανοικτών συνδέσεων σε έναν αντιπρόσωπο, τότε χρειάζεται πολιτική για να αποφασιθεί ποια σύνδεση θα κλείσει – αντίστοιχα με την πολιτική ανταλλαγής δεδομένων στην ΚΜ.

Καθυστέρηση: Κρύβοντας τις TCP Συνδέσεις και Δεδομένα

- Μελέτες έχουν δείξει ότι το κρύψιμο συνδέσεων αποφέρει μεγαλύτερα ωφέλη από το κρύψιμο δεδομένων.
- Συνολικά επιτεύχθηκε ~25% βελτίωση της καθυστέρησης με
  - 20% να προέρχεται από κρύψιμο συνδέσεων
  - 5% από κρύψιμο δεδομένων
- Επίσης σημαντικά ωφέλη προέκυψαν ακόμα και όταν **μόνο** η σύνδεση πελάτη-αντιπροσώπου ήταν επίμονη/κρυμμένη).

Καθυστέρηση: Διαιρώντας τις TCP Συνδέσεις

- Γεγονός: η ρυθμο-απόδοση (throughput) κατά την ανάκτηση ενός ιστοαντικειμένου, καθορίζει σε μεγάλο βαθμό την καθυστέρηση που βλέπει ο πελάτης (ειδικά για μεγάλα ιστοαντικείμενα).
- TCP congestion control (ελέγχος συμφόρησης): Βασικά:
  - SLOW START
    - Αποστολέας πρώτα στέλνει 2 πακέτα και περιμένει 2 ACKs (ρυθμοαπόδοση 2 πακέτα / RTT - παράθυρο συμφόρησης CWND = 2)
    - Όταν λάβει τα 2 ACKs, στέλνει 4 πακέτα και περιμένει 4 ACKs,... (ρυθμοαπόδοση 4 πακέτα / RTT, CWND = 4)
    - Μετά 8 πακέτα και περιμένει 8 ACKs... (ρυθμοαπόδοση 8 πακέτα / RTT , CWND = 8)...
  - Συνεχής διπλασιασμός του CWND μέχρι κάποιο όριο (sshthresh)
    - Κατόπιν, γραμμική αύξηση (+1): **congestion avoidance**
  - Όταν αρχίσει να βλέπει αύξηση αριθμού χαμένων πακέτων, αρχίζει να μειώνει το ρυθμό μετάδοσης και μετά τον αυξάνει γραμμικά
- ➔ : εξαιτίας του TCP congestion control (ελέγχου συμφόρησης) η ρυθμοαπόδοση μιας σύνδεσης εξαρτάται από το μέγεθος του χρόνου RTT (round-trip time)
- ➔ μεγαλύτερα αντικείμενα απολαμβάνουν μεγαλύτερη ρυθμοαπόδοση.

Καθυστέρηση: Διαιρώντας τις TCP Συνδέσεις

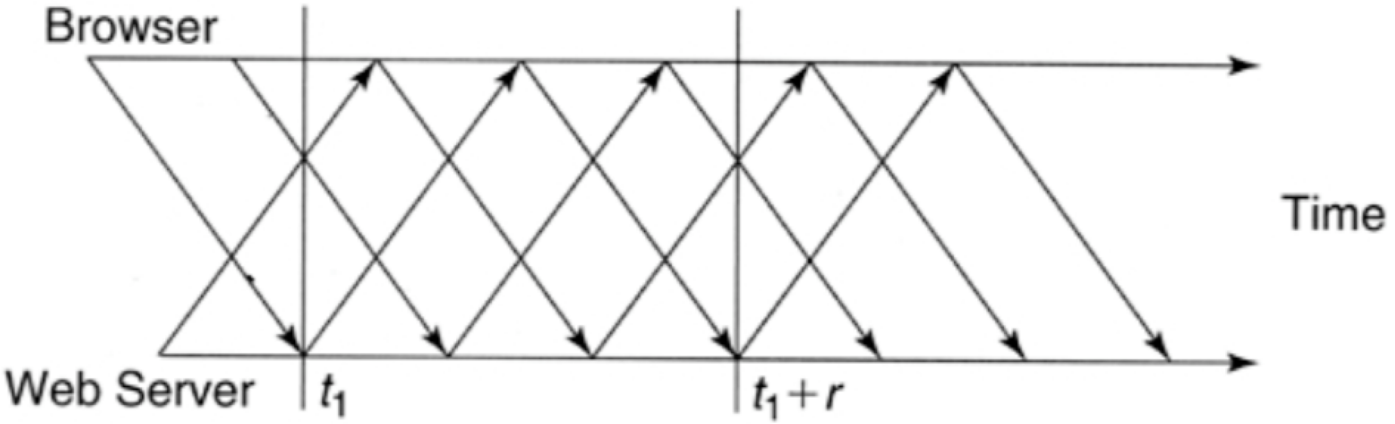
- ➔ αν μικρύνουμε το RTT, θα αυξήσουμε τη ρυθμοαπόδοση!
- Με τη χρήση αντιπροσώπων ανάμεσα σε πελάτες / εξυπηρέτες-πηγές, αυτό ακριβώς επιτυγχάνεται
  - Οι αποστάσεις πελάτη – αντιπρόσωπου και αντιπρόσωπου-πηγής είναι μικρότερες από αυτήν πελάτη – πηγής

  Παράδειγμα: αν ο αντιπρόσωπος είναι στη μέση της απόστασης πελάτη – πηγής, το RTT θα πέσει περίπου στο μισό
  - ➔ η ρυθμοαπόδοση θα διπλασιαστεί!

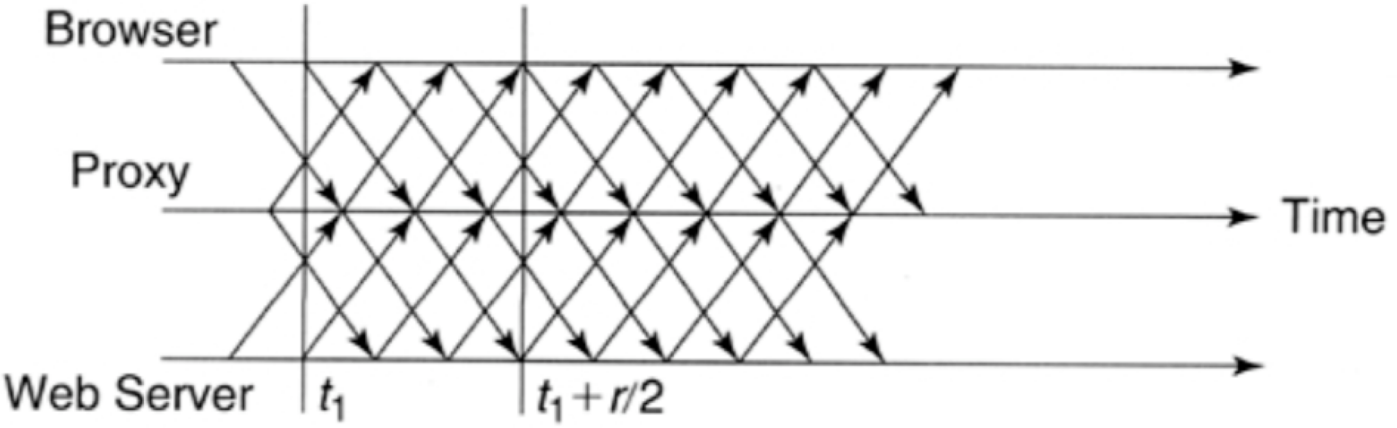
  (ο αποστολέας πάντα στέλνει π πακέτα ανά RTT , αλλά τώρα το RTT είναι το μισό αυτού που θα ήταν χωρίς αντιπρόσωπο.)

➔ Με διαιρεμένες TCP Συνδέσεις, (TCP Splitting) μέσω ΚΜ και αντιπροσώπων, η καθυστέρηση μειώνεται.

Καθυστέρηση: Διαιρώντας τις TCP Συνδέσεις



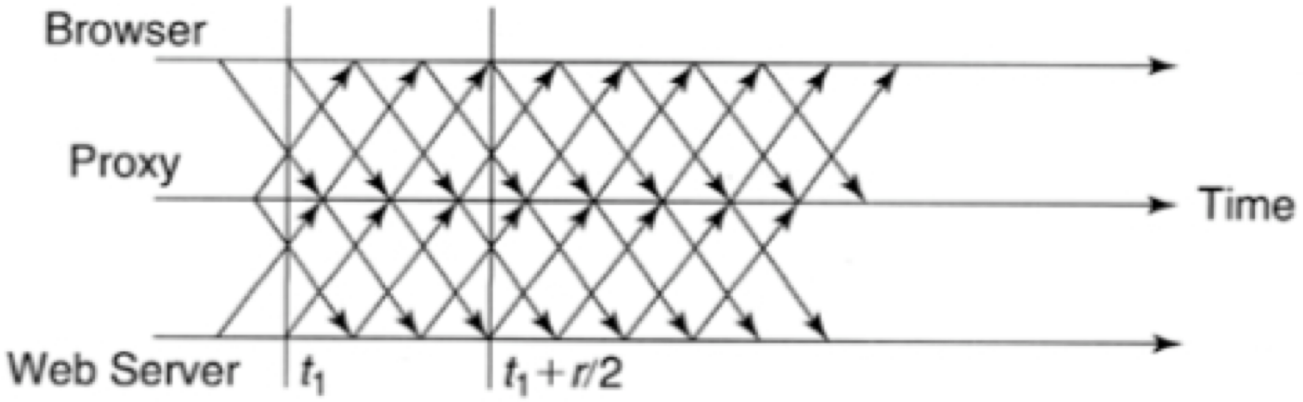
(*a*) No proxy; link RTT is *r*



(*b*) Proxy present; link RTT is *r*/2

Διαίρεση TCP συνδέσεων

- Αφού το RTT μεταξύ server & proxy είναι r/2, ο proxy θα λάβει το πακέτο p_1 στο χρόνο t_1+r/4.
- Την ίδια στιγμή, ο proxy στέλνει το ack πίσω στο server, και το πακέτο στον πελάτη.
- Καθώς το πακέτο ταξιδεύει στον πελάτη, το ack ταξιδεύει προς τον server.
- Αν δεν είχαμε τον ενδιάμεσο, το ack στον server θα στελνόταν μόνο όταν το πακέτο έφτανε στον πελάτη.
- Ο παραλληλισμός αυξάνει τη ρυθμοαπόδοση.



(*b*) Proxy present; link RTT is *r*/2

Εύρος Ζώνης

- Η κίνηση HTTP είναι ένα μεγάλο μέρος της κίνησης στο διαδίκτυο (το άλλο είναι P2P)
- Μειώνοντας απαιτήσεις στο εύρος ζώνης, μειώνονται κόστη (Ευρώ) και έμμεσα και η καθυστέρηση!
- Μελέτες έχουν βρει 30-40% ΡΧΨ που βρέθηκε να μειώνει τις απαιτήσεις στο εύρος ζώνης κατά ~25%.
- Αλλά, όπως και με τα της καθυστέρησεις, υπάρχουν ... Μυστικά...και περιπλοκές.

Εύρος Ζώνης

- Ο βασικός παράγοντας που περιπλέκει οφείλεται στις διακοπτόμενες αιτήσεις (aborted object transfers).
  - Κλικ STOP στον περιηγητή
  - Κλικ άλλο URL πριν τελειώσει η προηγούμενη μεταφορά...
- Όταν ο αντιπρόσωπος μάθει για τη διακοπή, μπορεί είτε
  - να την παραβλέψει και να συνεχίσει,
  - Είτε να τη μεταφέρει στην πηγή (abort forward).
- Η πρώτη επιλογή έχει αρνητικές επιπτώσεις στο εύρος ζώνης
  - Αφού ο ΡΧΨ είναι < 50%, τα ψηφία που μεταφέρονται μάλλον δε θα επαναχρησιμοποιηθούν...
  - Άρα, προώθηση της διακοπής στην πηγή;

Εύρος Ζώνης

- Ακόμα όμως και με προωθούμενες διακοπές, υπάρχει πρόβλημα
  - Εξαρτάται από το πόσο μεγάλο τμήμα του αντικειμένου είχε ήδη μεταφερθεί από την πηγή στον αντιπρόσωπο.
  - Αυτό εξαρτάται από τη σχέση ανάμεσα
    - στο εύρος ζώνης ανάμεσα στον πελάτη και στον αντιπρόσωπο (EZ1) και στο
    - στο εύρος ζώνης ανάμεσα στον αντιπρόσωπο και στην πηγή (EZ2).
    - Αν EZ1 << EZ2, τότε μέχρι να λάβει τη διακοπή ο αντιπρόσωπος, ένα μεγάλο μερός του αντικειμένου (που θέλει, εν τέλει, ο πελάτης) έχει ήδη κατέβει στον αντιπρόσωπο, χαραμίζοντας το αντίστοιχο εύρος ζώνης.
- Σε αυτές τις περιπτώσεις, συνήθως ο αντιπρόσωπος μειώνει το ρυθμό ζήτησης πακέτων από την πηγή για να καλυφθεί η μεγάλη διαφορά ανάμεσα στα EZ1 και EZ2.

Αντιπρόσωποι και Δεδομένα Συνεχούς Ροής

- Βλέπε βίντεο, audio κλπ.
- 2 κατηγορίες:
  - On-demand (κατ' απαίτηση)
  - Live streaming (ζωντανές ροές)
- Για «κατ' απαίτηση» εφαρμογές, η ΚΜ μπορεί κάλλιστα να χρησιμοποιοηθεί, όπως πριν
  - Το πρόβλημα όμως είναι το πολύ μεγάλο μέγεθος αυτών των αρχείων που θα γεμίσουν τις ΚΜ.
  - Αλλά, αν δεν αποθηκευτούν στην ΚΜ, τότε η απόδοση θα είναι άσχημη...
  - Η λύση είναι στο να αποθηκεύονται στην ΚΜ ένα αρκετά μεγάλο τέτοιο τμήμα του αρχείου, έτσι να αρχίζει άμεσα η αποστολή του στον πελάτη από την ΚΜ.
  - Παράλληλα, ο αντιπρόσωπος εδίδει αιτήσεις για το υπόλοιπα τμήματα να έρχονται άπό την πηγή.

Αντιπρόσωποι και Δεδομένα Συνεχούς Ροής

- Για Live streaming εφαρμογές:
  - Θα μπορούσαν να χρησιμοποιηθεί IP multicast, αλλά για πολλούς λόγους δεν αρκεί
    - Δεν υποστηρίζεται παντού
    - Ομάδες πελατών μεταβάλλονται δυναμικά και συχνά → routers έχουν μεγάλο κόστος διαχείρισης...
    - Επίσης, πολλοί routers θεωρούν multicast πακέτα ως χαμηλότερης προτεραιότητας → μεγάλο ποσοστό χαμένων πακέτων...
- Έτσι, η πηγή ανοίγει διαφορετικές συνδέσεις για διαφορετικούς πελάτες → το διαδίκτυο μεταφέρει στην ουσία την ίδια πληροφορία πάνω από πολλές διαφορετικές συνδέσεις.
- Με τη χρήση αντιπροσώπων σε ISP, συναθροίζονται όλες οι συνδέσεις των πελατών του ISP σε μια σύνδεση με την πηγή... Ελαχιστοποιώντας έτσι τις απαιτήσεις σε εύρος ζώνης!

Συνεργατικές ΚΜ

- Κάποιος ISP λειτουργεί αρκετές τοπικές ΚΜ
- Θα μπορούσε σε περίπτωση miss η αίτηση να πάει σε κάποια άλλη ΚΜ αντί για την πηγή;
  - Hits
    - Local hit (L0 + L1)
    - Remote hit (L2)
  - Misses
    - Global miss
    - Freshness miss (επικύρωση από άλλη ΚΜ ή πηγή)

Ζητήματα

- Διαχείριση τοποθεσίας
  - Πως μπορεί μια ΚΜ να γνωρίζει αν έχει το αντικείμενο κάποια άλλη;
- Αποφυγή ΚΜ
  - Μήπως είναι προτιμότερη η απευθείας πρόσβαση στην πηγή;
- Κατανάλωση ΕΖ
  - Πως μειώνουμε την επικοινωνία μεταξύ ΚΜ;

Διαχείριση τοποθεσίας

- Ιεραρχικές ΚΜ
  - Οργάνωση σε «δένδρο» και αποστολή αιτήματος υψηλότερα στην ιεραρχία
- Κατανεμημένες ΚΜ
  - Broadcast query
    - Αποστολή αιτήματος προς όλες τις ΚΜ
  - URL hashing
    - Διαμοιρασμός των URL στις ΚΜ και ανακατεύθυνση στη σχετική ΚΜ
  - Directory-based
    - Διατήρηση ευρετηρίου ιστοαντικειμένων (κεντρικά, ή κατανεμημένα με summaries)

Αποφυγή ΚΜ

- Η ανάκτηση από την πηγή μπορεί να είναι πιο γρήγορη
- Η συνδεσιμότητα μπορεί να ποικίλλει
- Πρέπει να λαμβάνεται υπόψιν ο φόρτος στους proxies
- Λύση: συλλογή στατιστικών στοιχείων

Pre-caching / pre-fetching

- Η μελέτη των logs μπορεί να οδηγήσει και σε πληροφορίες για ιστοαντικείμενα που μπορεί να χρειαστούν στο μέλλον
  - Ανάκτηση χωρίς αίτηση από κάποιον browser
  - Το ίδιο αναλύοντας και το περιεχόμενο ενός ιστοαντικειμένου που μόλις ανακτήθηκε
- Μπορεί να δουλέψει στο επίπεδο του
  - Browser - server (εξατομικευμένες προβλέψεις)
  - Browser – proxy (ο proxy χρησιμοποιεί το νεκρό χρόνο για να φέρει πιθανώς χρήσιμα επόμενα αντικείμενα).
  - Proxy – server (με ή χωρίς προώθηση του περιεχόμενου στον browser)

Μερικές ενδιαφέρουσες πηγές

- Fan, L., Cao, P., Almeida, J., & Broder, A. Z. (2000). Summary cache: a scalable wide-area web cache sharing protocol. *IEEE/ACM Transactions on Networking (TON)*, *8*(3), 281-293.
- Wang, J. (1999). A survey of web caching schemes for the internet. *ACM SIGCOMM Computer Communication Review*, *29*(5), 36-46.
- Vesuna, J., Scott, C., Buettner, M., Piatek, M., Krishnamurthy, A., & Shenker, S. (2016). Caching doesn't improve mobile web performance (much). In *2016 {USENIX} Annual Technical Conference ({USENIX}{ATC} 16)* (pp. 159-165).
- Zakhary, V., Agrawal, D., & Abbadi, A. E. (2017). Caching at the web scale. *Proceedings of the VLDB Endowment*, *10*(12), 2002-2005.
- Sulaiman, S., Shamsuddin, S. M., & Abraham, A. (2013). A Survey of Web Caching Architectures or Deployment Schemes. *International Journal of Innovative Computing*, *3*(1).
- Tatar, A., De Amorim, M. D., Fdida, S., & Antoniadis, P. (2014). A survey on predicting the popularity of web content. *Journal of Internet Services and Applications*, *5*(1), 8.