ΕΛΛΗΝΙΚΟ ΑΝΟΙΚΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ

ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ

ΜΕΤΑΠΤΥΧΙΑΚΗ ΕΞΕΙΔΙΚΕΥΣΗ
ΣΤΑ ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΠΑΙΓΝΙΟΘΕΩΡΗΤΙΚΗ ΜΕΛΕΤΗ ΠΟΛΥΕΠΙΠΕΔΩΝ ΑΡΧΙΤΕΚΤΟΝΙΚΩΝ WEB CACHING

ΦΩΚΑΪΔΗΣ ΠΑΝΑΓΙΩΤΗΣ

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ:
ΧΑΤΖΗΕΥΘΥΜΙΑΔΗΣ ΕΥΣΤΑΘΙΟΣ

ΠΑΤΡΑ
ΜΑΪΟΣ, 2009

Διπλωματική Εργασία

Παιγνιοθεωρητική Μελέτη Πολυεπίπεδων Αρχιτεκτονικών Web Caching

Φωκαΐδης Παναγιώτης

17-05-2009

Παιγνιοθεωρητική Μελέτη Πολυεπίπεδων Αρχιτεκτονικών Web Caching

Φωκαΐδης Παναγιώτης

Επιβλέπων	Μέλος	Μέλος
Χατζηευθυμιάδης Ευστάθιος	**Σκόδρας Αθανάσιος**	**Βασσάλος Βασίλειος**

ΠΕΡΙΛΗΨΗ

Ο Παγκόσμιος Ιστός είναι σήμερα, μια από τις πιο σημαντικές υπηρεσίες του Διαδικτύου και το Web Caching η πιο μελετημένη εφαρμογή του, η οποία προσεγγίζεται από διάφορα και ποικίλα επιστημονικά πεδία, ανάλογα με τα προβλήματα που κάθε φορά ανακύπτουν και την οπτική γωνία από την οποία αυτά εξετάζονται. Στην παρούσα Διπλωματική εργασία, ορίστηκε μια Ολιγοπωλιακή αγορά με χρέωση χωρίς διάκριση, για τη διαχείριση της υπηρεσίας caching σε μια ιεραρχία. Οι πωλητές είναι οι cache επιπέδου-2 (L2 cache), οι αγοραστές είναι οι cache επιπέδου-1 (L1 cache), το οικονομικό αγαθό είναι ο αποθηκευτικός χώρος και ο ανταγωνισμός γίνεται στις τιμές. Ως στόχος της εργασίας τέθηκε η διερεύνηση ύπαρξης και στη συνέχεια εύρεσης, πρώτον του βέλτιστου διανύσματος τιμών που θέτουν οι πωλητές, και δεύτερον των διανυσμάτων εκχώρησης χωρητικότητας στους αγοραστές. Με τη χρήση της Θεωρίας Παιγνίων, το ανταγωνιστικό ως προς τις τιμές Ολιγοπώλιο, μοντελοποιήθηκε ως μη συνεργατικό παίγνιο, και στη συνέχεια αποδείχθηκε η ύπαρξη μοναδικής Ισορροπίας Nash στο παίγνιο. Τα διανύσματα εκχώρησης χωρητικότητας έγινε δυνατόν να υπολογιστούν με αναλυτική μέθοδο, ενώ το βέλτιστο διάνυσμα τιμών υπολογίστηκε με τη βοήθεια διχοτομικού αλγόριθμου. Από την μελέτη των αποτελεσμάτων μέσω αριθμητικών παραδειγμάτων κατέστη δυνατόν να εξαχθούν χρήσιμα συμπεράσματα για την μελετούμενη δομή, αλλά και να επιβεβαιωθούν συμπεράσματα που διατυπώθηκαν κατά τη πορεία της επίλυσης των βασικών προβλημάτων.

Λέξεις-κλειδιά: Web Caching, Θεωρία Παιγνίων, Ολιγοπώλιο, Διαδικτυακές Εφαρμογές, Κατανεμημένα Συστήματα, Διαχείριση πόρων.

Περιεχόμενο: Κείμενο, Σχήματα.

Game Theoretic Analysis of Multi-Layer Web Caching Architectures

Fokaidis Panagiotis

Supervisor	Member	Member
Hadjiefthymiades Stathes	**Skodras Athanassios**	**Vassalos Vasilis**

ABSTRACT

The Web is nowadays, one of the most important services of the Internet, and Web Caching consists the most studied application of it, examined by various scientific approaches, depending on the problems arise at each particular case, as well as from the point of view examined. In this thesis, an Oligopolistic market with no price discrimination is set, to manage a Web Caching hierarchy. The sellers are the L2 caches, the buyers are the L1 caches, the economic good is the capacity and the competition is in prices. The aim of the work was to investigate the existence and then the determination of, firstly the optimal prices vector set by the sellers, and secondly the capacity vectors concerning the buyers. By means of Game Theory, the competitive in price Oligopoly, was modelled as non-cooperative game. Then, it was proved that the game has a unique Nash Equilibrium. The capacity vectors were estimated by means of an analytical method, while the optimal prices vector with a bisection algorithm. Examining the study results derived by several numerical examples, it was possible to draw useful conclusions for the examined structure. In addition a number of conclusions outlined during the solving procedure of the basic problems were, in this way, also confirmed.

Keywords: Web Caching, Game Theory, Oligopoly, Web Applications, Distributed Systems, Resource Management.

Content: Text, Figures.

Περιεχόμενα

1 Εισαγωγή

Ο Παγκόσμιος Ιστός (World Wide Web - WWW) είναι σήμερα, μια από τις πιο σημαντικές υπηρεσίες του Διαδικτύου (Internet). Η εντυπωσιακή ανάπτυξη και η διαρκώς αυξανόμενη χρήση της υπηρεσίας WWW κατέστησαν επιβεβλημένη τη βελτιστοποίηση στην πρόσβαση του πληροφοριακού περιεχομένου. Η χρήση των WWW cache συνεισφέρει σημαντικά προς αυτήν την κατεύθυνση, επιτυγχάνοντας τη μείωση, τόσο του χρησιμοποιούμενου εύρους ζώνης (bandwidth), όσο και του φόρτου στον πηγαίο διακομιστή (origin server), αλλά και του χρόνου εξυπηρέτησης του τελικού χρήστη, καθώς ένα μέρος του αιτούμενου περιεχομένου βρίσκεται ήδη πολύ κοντά σε αυτόν.

Η τεχνική του WWW caching, αφορά την αποθήκευση (cache) αντιγράφων των αντικειμένων που έχουν ζητηθεί από τους χρήστες του τοπικού δικτύου, έτσι ώστε, όταν κάποιο αντικείμενο ζητηθεί ξανά, αυτό να παραδίδεται κατευθείαν από τον caching proxy, χωρίς να γίνεται επικοινωνία με τον πηγαίο διακομιστή WWW περιεχομένου.

Τα ζητήματα που παρουσιάζονται στη διαχείριση των WWW cache είναι πολλά. Έχουν σχέση με την επιλογή διαφόρων παραμέτρων, αλλά και γενικότερων στρατηγικών, όπως σε ποιους διακομιστές να τοποθετηθούν κάποια αντίγραφα, ή από ποιον διακομιστή να ζητηθεί ένα αντίγραφο. Αυτά τα προβλήματα έχουν αποτελέσει αντικείμενο μελέτης από μεγάλο πλήθος ερευνητών καθώς κι αντικείμενο έρευνας κι ανάπτυξης προϊόντων από εμπορικούς οργανισμούς παγκοσμίως. Ανοικτά ζητήματα όμως παραμένουν αρκετά, καθώς υπάρχει σημαντικό περιθώριο για πιθανές βελτιώσεις υπαρχόντων αρχιτεκτονικών και στρατηγικών, αλλά και για πρόταση νέων.

Το WWW caching μπορεί να εφαρμοστεί και σε διάφορα επίπεδα σε ιεραρχία π.χ. $1^{ου}$ επιπέδου (Level 1 – L1), $2^{ου}$ επιπέδου (Level 2 – L2), κ.α. Όταν η χωρητικότητες των L2 cache διαμοιράζονται μεταξύ κάποιων L1 cache, είναι δυνατόν, αν μια L1 cache είναι ιδιαίτερα "επιθετική", σε σχέση με τις άλλες, να κυριαρχήσει επί της

διαμοιραζόμενης χωρητικότητας, με αποτέλεσμα, να επωφεληθεί εις βάρος των άλλων L1 cache.

Σε αυτήν την εργασία, για την αποφυγή "εγωιστικών" συμπεριφορών από τις L1 cache, προτείνεται ένα πλαίσιο, στο οποίο οι L2 cache παρέχουν τον αποθηκευτικό τους χώρο, με χρέωση. Οι L2 cache χωρίζουν το σύνολο του αποθηκευτικού τους χώρου σε διακριτά τμήματα, καθένα από τα οποία "νοικιάζουν" στις ενδιαφερόμενες L1 cache. Το περιγραφόμενο πλαίσιο μοντελοποιείται ως μια Οικονομία (Market), με σαφείς ρόλους για την κάθε οντότητα (οι L2 cache είναι οι πωλητές, ενώ οι L1 cache οι αγοραστές). Το κίνητρο για την εργασία αυτή, προέρχεται από την εργασία [1], σε κεφάλαιο της οποίας μελετάται το αναφερθέν πρόβλημα, χρησιμοποιώντας ως Οικονομικό μοντέλο το Μονοπώλιο με χρέωση με ή χωρίς διάκριση και με ή χωρίς ρυθμιστή.

Στην παρούσα εργασία, ως Οικονομικό μοντέλο εξετάζεται το Ολιγοπώλιο με χρέωση χωρίς διάκριση. Με τη χρήση της Θεωρίας Παιγνίων ως εργαλείο, ο ανταγωνισμός των πωλητών (L2 cache) για την προσέλκυση αγοραστών (L1 cache), μοντελοποιείται ως ένα μη συνεργατικό παίγνιο. Η επίλυση των προβλημάτων στα οποία οδηγεί το Παιγνιοθεωρητικό μοντέλο που προτείνεται, οδηγεί στον υπολογισμό των βέλτιστων τιμών ποσοτικών χαρακτηριστικών της εξεταζόμενης δομής, όπως το διάνυσμα χρέωσης και η κατανομή της χωρητικότητας.

Σχετικά με τη δομή της εργασίας, στο Κεφ. 2 μελετάται η γενικότερη κατάσταση σε θέματα τεχνολογιών WWW caching, όπως επίσης και κάποιες προηγούμενες προσπάθειες εφαρμογής Παιγνιοθεωρητικών προσεγγίσεων σε δικτυακά προβλήματα. Στο Κεφ. 3 γίνεται μια εισαγωγή σε έννοιες και συμβολισμούς σχετικά με τη Θεωρία Παιγνίων και στο Κεφ. 4 παρουσιάζονται κάποιες έννοιες της Οικονομικής Θεωρίας, όπως η αγορά, το προϊόν και τα μοντέλα αυτής ανάλογα με τον αριθμό των πωλητών. Ακολουθεί στο Κεφ. 5 η παρουσίαση, η ανάπτυξη και η προτεινόμενη επίλυση που δόθηκε στο υπό εξέταση πρόβλημα. Στο Κεφ. 6 γίνεται χρήση αριθμητικών παραδειγμάτων για την εξαγωγή και επιβεβαίωση συμπερασμάτων. Στο Κεφ. 7 παραθέτουμε συνολικά τα συμπεράσματα μας, στο Κεφ. 8 υπάρχουν οι αναφορές της εργασίας και ολοκληρώνοντας στο Κεφ. 9 (Παράρτημα) παρατίθενται οι αποδείξεις των Προτάσεων και Θεωρημάτων που χρησιμοποιήθηκαν στο Κεφ.5.

2 Web Caching

2.1 Γενικά

Η εντυπωσιακή αύξηση της ζήτησης Διαδικτυακών προϊόντων, είχε σαν αποτέλεσμα την μεγάλη αύξηση της δικτυακής κυκλοφορίας. Καθώς το διακινούμενο φορτίο πλησιάζει την χωρητικότητα του δικτύου, η δικτυακή κυκλοφοριακή συμφόρηση και η υπερφόρτωση των διακομιστών (servers) γίνεται συνηθισμένο φαινόμενο με αποτέλεσμα την αυξημένη καθυστέρηση της κυκλοφορίας. Καθυστέρηση στην κυκλοφορία μπορεί να προκληθεί και από μια ζεύξη χαμηλής ταχύτητας πάνω στην διαδρομή στην οποία κινείται το αιτούμενο αντικείμενο. Για την μείωση της εμφάνισης αυτών των μεγάλων καθυστερήσεων, μια στρατηγική είναι η δημιουργία αντιγράφων του αντικειμένου και η αποθήκευσή τους σε κατάλληλα σημεία μέσα στο δίκτυο. Κατόπιν η αίτηση του χρήστη μπορεί να κατευθύνεται εκεί, έχοντας με αυτόν τον τρόπο μείωση της δικτυακής κυκλοφοριακής συμφόρησης και ελαχιστοποίηση της καθυστέρησης στην εξυπηρέτηση του χρήστη.

Η παραπάνω διαδικασία υλοποιείται με την χρήση των caches. Το RFC 2616 καθορίζει ότι μια cache είναι ένας τοπικός χώρος αποθήκευσης μηνυμάτων απάντησης. Ένας λιγότερο αυστηρός ορισμός του caching είναι η μετακίνηση web περιεχομένου πλησιέστερα στους τελικούς χρήστες. Ως πλεονεκτήματα του caching μπορούν να θεωρηθούν τα παρακάτω:

- *Μείωση του χρησιμοποιούμενου εύρους ζώνης.* Αυτό έχει ως αποτέλεσμα την μείωση της κίνησης και της συμφόρησης στο δίκτυο όπως επίσης και της εξοικονόμησης χρημάτων από την πλευρά των εταιριών WWW hosting, καθώς αυτές πληρώνουν για το εύρος ζώνης που χρησιμοποιούν.
- *Μείωση του φόρτου στον πηγαίο διακομιστή.* Αυτό συμβαίνει καθώς η cache παρεμβάλλεται μεταξύ του χρήστη και του πηγαίου διακομιστή και χειρίζεται τις αιτήσεις.
- *Μείωση του χρόνου εξυπηρέτησης δηλαδή του χρόνου που αντιλαμβάνεται ο τελικός χρήστης από την υποβολή μιας αίτησης μέχρι την παρουσίαση σε αυτόν του*

ζητούμενου αντικειμένου. Η μείωση αυτή δημιουργεί ικανοποίηση στον χρήστη και συμβαίνει για δύο λόγους:

- τα συχνά ζητούμενα αντικείμενα, καθώς είναι αποθηκευμένα στην cache, προωθούνται στον χρήστη από αυτήν και όχι από τον πηγαίο διακομιστή, με αποτέλεσμα την ελαχιστοποίηση της καθυστέρησης παράδοσης.
- τα αντικείμενα που δεν είναι αποθηκευμένα στην cache, και προσκομίζονται από τον πηγαίο διακομιστή, μπορούν να ανακτηθούν σχετικά ταχύτερα από ότι χωρίς caching, καθώς υπάρχει μειωμένη κίνηση και συμφόρηση στο δίκτυο όπως και μειωμένος φόρτος στον πηγαίο διακομιστή.

2.2 Ποια αντικείμενα επιδέχονται caching

Ο πρωταρχικός στόχος μιας cache είναι να αποθηκεύει ορισμένες από τις απαντήσεις που λαμβάνει από τους πηγαίους διακομιστές. Μια απάντηση θεωρείται ότι επιδέχεται caching αν μπορεί να χρησιμοποιηθεί ως απάντηση σε μια μελλοντική αίτηση. Μια cache αποφασίζει εάν μπορεί να αποθηκεύσει μια συγκεκριμένη απάντηση, εξετάζοντας απαιτήσεις που σχετίζονται με το πρωτόκολλο HTTP καθώς και το υπό εξέταση περιεχόμενο. Σχετικά με το HTTP πρωτόκολλο απαιτείται οι caches να λαμβάνουν υπόψη κάποιες ντιρεκτίβες σχετικά με την δυνατότητα αποθήκευσης κάποιου συγκεκριμένου μηνύματος. Οι απαιτήσεις οι οποίες εξαρτώνται από το περιεχόμενο επηρεάζονται από τις επιχειρηματικές απαιτήσεις μιας cache καθώς και τις πολιτικές που επηρεάζουν την συχνότητα ελέγχου της ορθότητας των περιεχομένων της cache (revalidation). Οι πολιτικές αυτές, με την σειρά τους, επηρεάζονται από τα χαρακτηριστικά όπως π.χ., το μέγεθος και τον τύπο.

Το HTTP/1.1 καθορίζει απλούς κανόνες για το ποια αντικείμενα επιδέχονται αποθήκευση σε cache. Η μέθοδος αίτησης, τα πεδία επικεφαλίδας της αίτησης, ο κωδικός κατάστασης της απάντησης (response status) καθώς και τα πεδία επικεφαλίδας της απάντησης πρέπει να υποδεικνύουν εάν το αντικείμενο επιδέχεται αποθήκευση σε cache. Οι απαντήσεις στα αιτήματα OPTIONS, PUT και DELETE δεν επιδέχονται αποθήκευση σε cache. Οι απαντήσεις στα αιτήματα POST δεν επιδέχονται αποθήκευση σε cache εκτός εάν η απάντηση φέρει τις κατάλληλες Cache-Control και Expires επικεφαλίδες. Εάν η cache δεν υποστηρίζει το πεδίο επικεφαλίδα Range, όλες οι

απαντήσεις που φέρουν κωδικό κατάστασης 206 (Partial Content) δεν μπορούν να αποθηκευτούν στη cache.

Ορισμένες απαντήσεις περιέχουν πληροφορία από τον πηγαίο διακομιστή που αποκλείει την αποθήκευση του μηνύματος σε cache. Υπάρχουν 2 είδη τέτοιας πληροφορίας. Πληροφορίες σχετικά με την δυνατότητα αποθήκευσης και ντιρεκτίβες προς τις caches. Εάν η απάντηση περιέχει το πρώτο είδος πληροφορίας, η απόφαση για αποθήκευση σε cache θα πρέπει να βασίζεται σε αυτήν. Για παράδειγμα, ο πηγαίος διακομιστής μπορεί να καθορίσει ακριβώς το διάστημα για το οποίο το αντικείμενο θα πρέπει να θεωρείται έγκυρο μέσω του πεδίου Expires. Η ντιρεκτίβα Cache-Control μπορεί να αποκλείσει την αποθήκευση σε cache ορισμένων αντικειμένων. Για παράδειγμα η Cache-Control: private καθορίζει ότι μια διαμοιραζόμενη cache δεν μπορεί να αποθηκεύσει το αντικείμενο. Ένα μήνυμα απάντησης που περιέχει την ντιρεκτίβα Cache-Control: no-store δεν θα πρέπει να αποθηκευτεί καθόλου. Η ντιρεκτίβα Cache-Control: no-cache περιορίζει την πιθανότητα αποθήκευσης του αντικειμένου στη cache γιατί θα πρέπει να ελέγχεται η ορθότητα του αντικειμένου πριν από κάθε φορά που επιστρέφεται ως επιτυχία cache (cache hit). Οι ντιρεκτίβες δεν περιορίζονται μόνο στις απαντήσεις από τον πηγαίο διακομιστή. Μπορεί να ενσωματώνονται και στις ερωτήσεις από τον αντιπρόσωπο του χρήστη (user agent). Για παράδειγμα η Cache-Control: no-store μπορεί να εμφανιστεί τόσο σε απάντηση όσο και σε ερώτηση. Το πρωτόκολλο διαθέτει αυτές τις ντιρεκτίβες για να προστατεύει τον προσωπικό χαρακτήρα μιας απάντησης (privacy) και για να υποδείξει την πτητικότητα του αντικειμένου, δηλαδή ότι μπορεί να αλλάξει αμέσως μετά την αποστολή του.

Η παρουσία πεδίων επικεφαλίδας όπως τα Authorization και Vary ελαχιστοποιούν τις πιθανότητες αποθήκευσης του αντικειμένου σε cache. Η επικεφαλίδα αίτησης Authorization υποδεικνύει ότι το ζητούμενο αντικείμενο δεν είναι διαθέσιμο για όλους. Επίσης, η παρουσία του Vary δεικνύει ότι μια αποδεκτή απάντηση για cache θα πρέπει να περιορίζεται από τις τιμές που ορίζονται στο Vary πεδίο.

Μια cache μπορεί να έχει και τους δικούς της κανόνες για τον έλεγχο της δυνατότητας αποθήκευσης σε αυτήν κάποιου συγκεκριμένου αντικειμένου, άσχετα από τους περιορισμούς που επιβάλλονται μέσω του πρωτοκόλλου. Δηλαδή, αν και το αντικείμενο μπορεί να αποθηκευτεί στη cache, δεν σημαίνει κατ' ανάγκη, ότι τελικά θα αποθηκευτεί σε αυτήν. Τα μηνύματα μπορεί να είναι μεγάλα σε όγκο, να παράγονται

δυναμικά, να περιέχουν cookies, παράμετροι που επηρεάζουν σημαντικά την δυνατότητα αποθήκευσης του αντικειμένου σε cache. Οι πολιτικές αποθήκευσης σε cache επηρεάζονται από χαρακτηριστικές του μηνύματος και όχι από τους περιορισμούς του πρωτοκόλλου.

Οι επιχειρήσεις μπορεί να θέλουν να περιορίσουν το κόστος μεταφοράς αγνοώντας ορισμένους περιορισμούς που σχετίζονται με την cache. Για παράδειγμα μπορεί να αποθηκεύουν αντικείμενα που δεν θα πρέπει να αποθηκεύουν (π.χ., Cache-Control: private). Επίσης, οι caches μπορεί να λαμβάνουν υπόψη την επιβάρυνση αποθηκευτικού χώρου (storage overhead) και να μην αποθηκεύουν ορισμένα μεγάλα αρχεία παρά το γεγονός ότι επιδέχονται αποθήκευση σε cache. Εάν το αντικείμενο είναι μεγάλο, πολλά αντικείμενα θα πρέπει να εκτοπιστούν από την cache. Αναφορικά με την χρονική υστέρηση, το κόστος ανάκτησης από τους πηγαίους διακομιστές, μικρών σε όγκο αντικειμένων είναι μεγαλύτερο από το κόστος ανάκτησης τους από cache. Έτσι, μια cache μπορεί να αποφύγει την αποθήκευση μεγάλων αντικειμένων. Από την άλλη πλευρά, μια μεγάλη απάντηση αποθηκευμένη σε cache, εάν ζητηθεί ορισμένες φορές από τους χρήστες επιφέρει σημαντικά οφέλη εύρους ζώνης.

Πολλές caches δεν αποθηκεύουν τις απαντήσεις από scripts με το σκεπτικό ότι οι παράμετροι των σχετικών ερωτήσεων δεν πρόκειται να ξαναχρησιμοποιηθούν. Η παρουσία πρόσθετης πληροφορίας σχετικά με τη δυνατότητα αποθήκευσης σε caches στις επικεφαλίδες μιας δυναμικά διαμορφούμενης απάντησης (π.χ. Expires ή ETag) μπορεί να σημαίνει ότι αντικείμενο μπορεί να αποθηκευτεί σε cache (π.χ. ένα CGI script το οποίο επιστρέφει το n-οστό ψηφίο του αριθμού π). Πολλά www ερωτήματα συχνά καταλήγουν στις ίδιες απαντήσεις και υπάρχουν σήμερα caches που το λαμβάνουν αυτό υπόψη. Μια άλλη κατηγορία απαντήσεων που θεωρούνται ως μη επιδεχόμενες caching είναι οι απαντήσεις που εξατομικεύονται (tailored). Για παράδειγμα, οι απαντήσεις που συνοδεύονται από cookies θεωρούνται μη επιδεχόμενες caching, γιατί εκτιμάται ότι διαφέρουν από χρήστη σε χρήστη.

Η απόφαση για την αποθήκευση κάποιου αντικειμένου σε cache εξαρτάται από τον ρυθμό αλλαγών στο συγκεκριμένο αντικείμενο. Ορισμένα αντικείμενα μεταβάλλονται σπάνια, ίσως και καθόλου (π.χ., ηλεκτρονικές εκδόσεις βιβλίων). Μια παλιά ευρεστική τεχνική για τον προσδιορισμό της δυνατότητας αποθήκευσης σε cache ενός αντικειμένου ήταν η ημερομηνία τελευταίας αλλαγής. Υιοθετούνταν η υπόθεση

ότι εάν το αντικείμενο δεν έχει μεταβληθεί για μεγάλο χρονικό διάστημα είναι χαμηλή η πιθανότητα αλλαγής του στο άμεσο μέλλον. Αυτό το αντικείμενο θεωρούνταν υποψήφιο για αποθήκευση στην cache. Σε περίπτωση που το αντικείμενο αποθηκεύονταν στην cache, ο χρόνος τελευταίας αλλαγής υποδεικνύει το διάστημα στο οποίο θα έπρεπε να επανελεγχθεί η ορθότητα (revalidate) του αντικειμένου. Αντίστροφα, η λογική αυτή υποθέτει ότι αν το αντικείμενο έχει αλλαχθεί πρόσφατα είναι υψηλή η πιθανότητα αλλαγής του σε σύντομο χρονικό διάστημα. Το αντικείμενο αυτό δεν θα μείνει επίκαιρο (fresh) στην cache για μεγάλο διάστημα. Ένας άλλος προβληματισμός που λαμβάνεται επίσης υπόψη για τη δυνατότητα αποθήκευσης σε cache ενός αντικειμένου είναι ότι τα αντικείμενα που αλλάζουν συχνά είναι δημοφιλή και, κατ' επέκταση, χαρακτηρίζονται από υψηλό ρυθμό προσπέλασης (rate of access). Αυτά τα αντικείμενα θα πρέπει να αποθηκευτούν στην cache.

2.3 Που γίνεται το caching

Τα web αντικείμενα μπορούν να αποθηκεύονται στον υπολογιστή του χρήστη ή σε έναν διακομιστή στον Ιστό. Υπάρχουν διάφοροι τρόποι για το caching των web αντικειμένων:

- *Web browser cache*: Ο web browser αποθηκεύει τα αντικείμενα στον υπολογιστή του χρήστη. Ο web browser, πριν ζητήσει τα αντικείμενα από το κατάλληλο πηγαίο διακομιστή, ψάχνει για αυτά στην δική του cache. Η αποθήκευση συχνά ζητούμενων αντικειμένων στο σκληρό δίσκο του χρήστη αυξάνει την ταχύτητα πλοήγησης στο Διαδίκτυο, όμως δεν γίνεται εκμετάλλευση των αντικειμένων που ζητούνται ιδιαίτερα συχνά από άλλους χρήστες στο ίδιο περιβάλλον.
- *Proxy cache*: Ο proxy cache διακομιστής παρεμβάλλεται ανάμεσα στις HTTP αιτήσεις των χρηστών, και αν βρει το αντικείμενο που ζητείται στη cache του τότε το επιστρέφει στον χρήστη. Αν το αντικείμενο δεν βρεθεί, η cache απευθύνεται στον πηγαίο διακομιστή του αντικειμένου και για λογαριασμό του χρήστη, παίρνει το αντικείμενο, πιθανόν το αποθηκεύει, και τελικώς επιστρέφει το αντικείμενο στον χρήστη. Οι proxy caches συνήθως αναπτύσσονται στις άκρες του δικτύου (π.χ. στις εταιρικές πύλες, ή στα firewall των hosts) έτσι ώστε να μπορούν να εξυπηρετούν ένα μεγάλο αριθμό από εσωτερικούς χρήστες. Η χρήση proxy caches τυπικά επιδρά θετικά στην μείωση του απαιτούμενου εύρους ζώνης, στην βελτίωση του χρόνου

απόκρισης και στην αύξηση της διαθεσιμότητας στατικών αντικειμένων του Παγκόσμιου Ιστού. Η σχεδίαση με ένα αυτόνομο (standalone) proxy cache έχει το μειονέκτημα ότι η cache αποτελεί ένα μοναδικό σημείο αστοχίας στο δίκτυο. Συνεπώς, όταν η cache δεν είναι διαθέσιμη, το δίκτυο επίσης εμφανίζεται μη διαθέσιμο στους χρήστες. Επίσης, αυτή η προσέγγιση απαιτεί όλοι οι web browsers των χρηστών να διαμορφωθούν κατάλληλα για την χρήση της κατάλληλης proxy cache. Στη συνέχεια, αν ο proxy cache διακομιστής δεν είναι διαθέσιμος, όλοι οι χρήστες πρέπει να αναμορφώσουν τους web browsers τους με στόχο να χρησιμοποιήσουν μια διαφορετική cache.

- *Reverse (inverse) proxy cache*: Μια ενδιαφέρουσα προσέγγιση της αρχιτεκτονικής μιας proxy cache είναι η έννοια του ανάστροφου (reverse) proxy caching, στην οποία οι caches αναπτύσσονται κοντά στον τόπο προορισμού του περιεχομένου σε αντίθεση με την πλευρά του client. Αυτή είναι μια ελκυστική λύση για διακομιστές που περιμένουν έναν μεγάλο αριθμό αιτήσεων από το Διαδίκτυο ενώ ταυτόχρονα θέλουν να εξασφαλίσουν υψηλό επίπεδο ποιότητας υπηρεσιών. Το Reverse proxy caching είναι επίσης ένας χρήσιμος μηχανισμός για την υποστήριξη των web hosting farms (εικονικά domains που αντιστοιχούν σε ένα και μόνο φυσικό site) που είναι μια συνεχώς αυξανόμενη υπηρεσία για πολλούς παροχείς υπηρεσιών του Διαδικτύου (ISP). Θα πρέπει να σημειωθεί ότι η ανάπτυξη του reverse proxy caching είναι εντελώς ανεξάρτητη από το proxy caching στην πλευρά του χρήστη. Στην πραγματικότητα μπορούν να συνυπάρχουν και συλλογικά να βελτιώνουν την απόδοση του Ιστού τόσο από την προοπτική του χρήστη, όσο και από αυτήν του δικτύου και του διακομιστή.
- *Transparent proxy cache*: Το διαφανές (transparent) proxy caching ελαχιστοποιεί ένα από τα μεγαλύτερα μειονεκτήματα στην προσέγγιση του proxy server: Την απαίτηση για διαμόρφωση των web browsers. Οι transparent caches ανακόπτουν τις HTTP αιτήσεις και τις ανακατευθύνουν σε web cache servers ή σε ένα σύμπλεγμα από caches. Αυτό το είδος του caching δημιουργεί ένα σημείο στο οποίο είναι πιθανά διαφορετικά είδη διαχειριστικού ελέγχου, για παράδειγμα η απόφαση πώς να κατανεμηθεί το ισοζύγιο φόρτου των αιτήσεων κατά μήκος πολλαπλών caches. Η δύναμη του transparent cache είναι επίσης και η αδυναμία του: Παραβιάζει την end-to-end συμφωνία, μη διατηρώντας σταθερά τα τελικά σημεία της σύνδεσης. Αυτό

είναι ένα πρόβλημα όταν μια εφαρμογή απαιτεί να διατηρείται το καθεστώς μέσω συνεχόμενων αιτήσεων ή κατά την διάρκεια που μια λογική αίτηση εμπλέκει πολλαπλά αντικείμενα. Υπάρχουν δύο τρόποι για να αναπτυχθεί το transparent proxy caching: Στο επίπεδο διακόπτη (switch level) και στο επίπεδο δρομολογητή (router level). Στην δεύτερη περίπτωση χρησιμοποιείται μια πολιτική βασισμένη στην δρομολόγηση για να κατευθύνει τις αιτήσεις στην κατάλληλη cache. Για παράδειγμα, αιτήσεις από συγκεκριμένους χρήστες μπορούν να συσχετιστούν με κάποια συγκεκριμένη cache. Στην πρώτη περίπτωση ο διακόπτης ενεργεί ως ένας αποκλειστικός εξισορροπητής φόρτου. Οι διακόπτες γενικώς, είναι λιγότερο ακριβοί από τους δρομολογητές. Επίσης η λύση είναι περισσότερη ελκυστική διότι δεν υπάρχει επιπλέον φόρτος όπως αυτός που εισάγεται στην πολιτική που είναι βασισμένη στην δρομολόγηση.

2.4 Υλοποίηση του caching

Η υλοποίηση του μηχανισμού του caching βασίζεται σε συγκεκριμένα βήματα:

- *Έλεγχος αν το μήνυμα επιδέχεται caching*. Διαφορετικές caches υλοποιούν διαφορετικές προσεγγίσεις για την απόφαση caching του αντικειμένου. Όπως έχει ήδη αναφερθεί, τα διαφορετικά κριτήρια για την απόφαση αυτή είναι:
  - Υπάρχουν απαιτήσεις του πρωτοκόλλου που καθορίζουν ότι το συγκεκριμένο αντικείμενο δεν μπορεί να αποθηκευτεί στην cache;
  - Επιδέχεται συνήθως το περιεχόμενο caching;
  - Είναι πιθανή η επαναχρησιμοποίηση του αντικειμένου;
  - Μπορεί η απόφαση για αποθήκευση της συγκεκριμένης απάντησης να οδηγήσει σε αντικατάσταση ενός ή περισσοτέρων αντικειμένων;

  Μια cache χρησιμοποιεί κάποια ή όλα τα παραπάνω κριτήρια για να αποφασίσει αν πρέπει να εφαρμόσει caching σε κάποιο αντικείμενο.
- *Έλεγχος για την ύπαρξη διαθέσιμου χώρου*. Η cache πρέπει να διαγνώσει αν υπάρχει διαθέσιμος χώρος για την αποθήκευση του αντικειμένου, και στην περίπτωση που δεν υπάρχει χώρος, να αποφασίσει ποια ήδη αποθηκευμένα αντικείμενα θα αντικατασταθούν. Στην τελευταία περίπτωση, ενεργοποιείται ο μηχανισμός αντικατάστασης cache (cache replacement). Ο μηχανισμός αυτός εισάγει κάποια επιβάρυνση, ειδικά αν μικρότερα αντικείμενα που είναι ήδη αποθηκευμένα πρέπει

να διαγραφούν. Πρόσθετη επιβάρυνση προκαλείται όταν λαμβάνονται, μελλοντικά, αιτήσεις για αντικείμενα τα οποία έχουν διαγραφεί. Σε αυτήν την περίπτωση πρέπει να εγκατασταθούν νέες συνδέσεις για την άντληση τους από τους πηγαίους διακομιστές. Συχνά, αντικείμενα τα οποία θεωρούνται μη-επίκαιρα (stale) διαγράφονται από την cache ακόμη και στην περίπτωση που αυτή δεν είναι πλήρης. Έτσι περιορίζεται η ανάγκη για ενεργοποίηση του μηχανισμού αντικατάστασης cache την στιγμή διεκπεραίωσης μιας αίτησης (μείωση της υστέρησης που εκλαμβάνεται από το χρήστη). Μόλις προκύπτει ελεύθερος χώρος, η cache εξάγει πληροφορία για το μήνυμα, όπως το χρόνο τελευταίας μεταβολής καθώς και πληροφορία για την λήξη του αντικειμένου. Επικεφαλίδες όπως οι Expire και Cache-Control: max-stale μεταφέρουν πληροφορία σχετική με την λήξη (expiration) του αντικειμένου. Αυτά τα πεδία συντελούν στο να είναι η cache συμβατή με τους περιορισμούς του πρωτοκόλλου HTTP για το βάθος χρόνου στο οποίο η απάντηση μπορεί να επιστραφεί ως σημασιολογικά έγκυρη. Μια cache η οποία είναι συμβατή με το πρωτόκολλο είναι υποχρεωμένη να εξασφαλίζει ότι οι απαντήσεις που επιστρέφει θεωρούνται από τον πηγαίο διακομιστή ως επίκαιρες. Αν απουσιάζουν πληροφορίες λήξης του αντικειμένου, η cache προσδιορίζει ευρεστικά ένα χρόνο λήξης για να αποφασίσει πότε γίνεται μη-επίκαιρο (stale). Ο αλγόριθμος μπορεί να βασίζεται στη τιμή του πεδίου Last-Modified που είναι συνυφασμένη με το αντικείμενο. Επίσης, παράγεται ένα κλειδί για χρήση σε μελλοντικές αναζητήσεις (lookups). Το κλειδί αυτό είναι μια τιμή κατακερματισμού (hash value) η οποία βασίζεται στο URL του αντικειμένου.

- *Επιστροφή του αιτούμενου αντικειμένου στον χρήστη*. Εάν ένα αντικείμενο που αντιπροσωπεύει κάποιο κλειδί που αναζητείται στην cache εντοπιστεί, θεωρείται ότι συνέβη ένα cache hit. Τότε, ανάλογα με την πολιτική της cache και ενδεχόμενους περιορισμούς που επιβάλλονται από πεδία επικεφαλίδας, ένας επανέλεγχος ορθότητας μπορεί να εκτελεστεί για να διαπιστωθεί εάν το αντικείμενο είναι επίκαιρο. Εάν ο έλεγχος αυτός αποβεί θετικός, η αίτηση ικανοποιείται από την cache. Διαφορετικά, θεωρείται ότι συνέβη ένα cache miss, η cache ανακτά ένα νέο αντίγραφο του αντικειμένου από τον πηγαίο διακομιστή, και εφαρμόζει την πολιτική της για να αποφανθεί αν το αντικείμενο θα πρέπει να αποθηκευτεί παράλληλα με την προώθηση του στον χρήστη που το αιτήθηκε.

2.5 Αντικατάσταση περιεχομένων cache

Αντικατάσταση αντικειμένων σε μια cache, μπορεί να γίνει σε περίπτωση:

- *Συντήρησης της cache*. Μια cache μπορεί να ελέγχει εάν τα αντικείμενα που είναι αποθηκευμένα σε αυτή είναι επίκαιρα και να διαγράφει τα "παλιά" αντικείμενα. Μια cache μπορεί επίσης να ελέγχει τον ρυθμό των αιτήσεων για αποθηκευμένα αντικείμενα ώστε να αποφασίζει ποια από αυτά είναι δημοφιλή και να προβαίνει σε ειδικές ενέργειες για λογαριασμό τους. Για παράδειγμα, μια cache μπορεί να προελέγχει την ορθότητα – εγκυρότητα (pre-validation) των αποθηκευμένων σε αυτήν αντικειμένων για να διαπιστώνει αν αυτά που ζητούνται περισσότερο, είναι επίκαιρα. Αυτός ο προέλεγχος μπορεί να υλοποιείται με την HTTP Head μέθοδο που αντλεί μόνο τα μετα-δεδομένα για τα υπό συζήτηση αντικείμενα. Μια cache μπορεί και προ-δραστικά, να επικοινωνεί με τον πηγαίο διακομιστή και να ελέγχει εάν το αντικείμενο έχει μεταβληθεί. Εάν όντως έχει μεταβληθεί μπορεί να ξεκινά τη διαδικασία της προανάκτησης (prefetching) αντικειμένων για την ενημέρωση της cache.
- *Η cache είναι πλήρης*. Σε αυτή τη περίπτωση, αντικείμενα πρέπει να διαγραφούν για να δημιουργηθεί χώρος για την αποθήκευση νέων απαντήσεων. Πολλές στρατηγικές για την αντικατάσταση αντικειμένων έχουν προταθεί. Ορισμένες προέρχονται από το παραδοσιακό χώρο της διαχείρισης cache σε συστήματα αρχείων, ενώ άλλες είναι εξειδικευμένες στο web περιβάλλον. Μια ιδιαίτερα γνωστή προσέγγιση είναι το LRU (Least Recently Used) (αντικατάσταση του αντικειμένου που χρησιμοποιήθηκε λιγότερο). Οι στόχοι του caching, δηλαδή η μείωση του όγκου της πληροφορίας που ανταλλάσσεται στο δίκτυο καθώς και της υστέρησης που αντιλαμβάνεται ο χρήστης, οδηγούν σε σύνθετες αποφάσεις για την αντικατάσταση περιεχομένου cache. Οι σύνθετες αποφάσεις αποτελούν ένα συνδυασμό μετρικών που περιλαμβάνουν το μέγεθος των απαντήσεων που αποθηκεύονται, τον τύπο αντικειμένου ακόμη και την έννοια της απόστασης προς τον πηγαίο διακομιστή. Η χρησιμότητα διατήρησης ενός αντικειμένου στην cache μπορεί να υπολογιστεί από πολλούς παράγοντες όπως:
  - *Το κόστος ανάκτησης του αντικειμένου*: το κόστος ανάκτησης ενός αντικειμένου από ένα πηγαίο διακομιστή προσδιορίζεται από την διασυνδεσιμότητα της cache

και την απόσταση που πρέπει να διανύσει το αντικείμενο μέχρι να καταχωρηθεί στην cache. Αντικαθιστώντας ένα αντικειμένου του οποίου η ανάκτηση ήταν "ακριβή", το ίδιο κόστος θα πρέπει να αντιμετωπιστεί στην περίπτωση που το αντικείμενο ζητηθεί πάλι στο μέλλον.

- *Το κόστος αποθήκευσης του αντικειμένου*: Μια cache έχει σταθερό μέγεθος και η αποθήκευση ενός αντικειμένου σημαίνει λιγότερο χώρο για άλλα αντικείμενα. Ένα μεγάλο σε όγκο αντικείμενο καταλαμβάνει σημαντικό χώρο αλλά ενδεχόμενη αντικατάσταση του σημαίνει ότι η ανάκτηση του πάλι θα κοστίσει σημαντικά.
- *Ο αριθμός των προσβάσεων στο αντικείμενο κατά το παρελθόν*: ένα αντικείμενο που έχει προσπελαστεί πολλές φορές στο παρελθόν είναι πολύ πιθανόν να προσπελαστεί και στο μέλλον και, κατά συνέπεια, είναι επωφελές να παραμείνει στην cache για μεγαλύτερο διάστημα.
- *Η πιθανότητα προσπέλασης του αντικειμένου στο άμεσο μέλλον*: Εάν το αντικείμενο είναι πιθανόν να ανακτηθεί στο άμεσο μέλλον δεν ενδείκνυται η απόρριψη του από την cache. Η πιθανότητα πρόσβασης σε ένα αντικείμενο θα μπορεί να είναι γνωστή a priori ή να προσδιορίζεται βάσει της ιστορικότητας προσπέλασης (access patterns).
- *Ο χρόνος από την τελευταία μεταβολή του αντικειμένου*: Ένα αντικείμενο που δεν έχει μεταβληθεί για μεγάλο διάστημα είναι λιγότερο πιθανό να αλλάξει στο κοντινό μέλλον. Ένα αντικείμενο που παρήχθη πρόσφατα μπορεί να είναι δυναμικό ή να υπάρχει μεγάλη πιθανότητα να αλλάξει πάλι στο μέλλον. Τα αντικείμενα που υπάρχει μεγάλη πιθανότητα να αλλάξουν είναι συνήθως δημοφιλή. Αυτά τα αντικείμενα μπορούν να μεταβληθούν σαν αποτέλεσμα του δημοφιλούς τους χαρακτήρα και έτσι είναι καλοί υποψήφιοι για caching. Η αποθηκευμένη απάντηση μπορεί όμως να πρέπει να αντικατασταθεί συχνά με το μεταβαλλόμενο αντικείμενο. Ο χρόνος τελευταίας μεταβολής ενός αντικείμενου μπορεί έτσι να χρησιμοποιηθεί για να προσδιοριστούν οι πιθανοί υποψήφιοι για αντικατάσταση.
- *Ο χρόνος λήξης που προσδιορίζεται ευρεστικά*: Εάν δεν υπάρχει χρόνος λήξης προσδιορισμένος από τον διακομιστή, η cache προσδιορίζει ευρεστικά ένα χρόνο λήξης. Εάν δεν υπάρχουν αντικείμενα για τα οποία έχει παρέλθει ο

χρόνος λήξης, τότε αυτά που βρίσκονται κοντινότερα στην λήξη τους αποτελούν υποψήφια για αντικατάσταση.

Οι αλγόριθμοι που χρησιμοποιούνται ως επί το πλείστον για αντικατάσταση (replacement) αντικειμένων στη cache είναι οι ακόλουθοι:

- ✓ Least Recently Used (LRU)
- ✓ Least Frequently Used (LFU)
- ✓ Size of object (SIZE)
- ✓ Hyper-G (LFU/LRU/SIZE): Το σύστημα Hyper-G συνδυάζει τις πολιτικές LRU/LFU και Size. Η πρώτη απόφαση για αντικατάσταση (replacement) βασίζεται στο LFU. Εάν υπάρχουν περισσότερα από ένα αντικείμενα που πληρούν το παραπάνω κριτήριο, εφαρμόζεται η πολιτική LRU. Εάν πάλι, δεν προσδιορίζεται ένα αντικείμενο προς αντικατάσταση, επιλέγεται το μεγαλύτερο σε όγκο αντικείμενο.
- ✓ GreedyDual-Size: Ο αλγόριθμος αυτός είχε προταθεί για αντικατάσταση σελίδων στην μνήμη υπολογιστικών συστημάτων. Ο αρχικός προσανατολισμός του δηλαδή αφορούσε ένα σύστημα με σταθερό μέγεθος αντικειμένων. Ο αλγόριθμος επεκτάθηκε για να καλύψει την ποικιλότητα των μεγεθών των www αντικειμένων. Ο μετασχηματισμένος αλγόριθμος συσχετίζει μια τιμή χρησιμότητας (utility value) και αντικαθιστά το αντικείμενο με την χαμηλότερη τιμή χρησιμότητας. Εκτός από το κόστος μεταφοράς του αντικειμένου στην cache και το μέγεθος του, η τιμή χρησιμότητας επηρεάζεται από τον παράγοντα παλαίωσης (age factor) που ενημερώνονται καθώς αντικείμενα απομακρύνονται από την cache.

2.6 Συνέπεια cache

Ο πηγαίος διακομιστής αποφασίζει την χρονική διάρκεια για την οποία το αντικείμενο θα πρέπει να θεωρείται έγκυρο (freshness duration). Μια cache θα πρέπει να εξασφαλίσει ότι μια αποθηκευμένη απάντηση είναι ακόμη έγκυρη πριν να απαντήσει σε κάποιον χρήστη που αιτείται το αντικείμενο. Η συνέπεια της cache είναι ένα σημαντικό πρόβλημα. Πολλοί σχετικοί αλγόριθμοι έχουν προταθεί κατά τα τελευταία χρόνια για το πρόβλημα της συνέπειας των web caches. Η ανάγκη για cache συνέπεια εξαρτάται από τα αντικείμενα και τις πολιτικές οι οποίες έχουν επιβληθεί στην cache.

Οι caches μπορεί απλά να επιστρέφουν μια παλιά αποθηκευμένη τιμή μαζί με μια αιτία για την απαξίωση του αντικειμένου. Μεταξύ των αιτίων είναι η αδυναμία εγκατάστασης σύνδεσης προς τον πηγαίο διακομιστή ή ο υψηλός φόρτος της cache. Η επικεφαλίδα Warning του HTTP/1.1 μπορεί να χρησιμοποιηθεί για να υποδηλώσει ότι επιστρέφεται μια μη-επίκαιρη απάντηση.

Το πρωτόκολλο HTTP/1.1 παρέχει πολλούς τρόπους για την διατήρηση της συνέπειας των caches. Εάν ο πηγαίος διακομιστής θέσει ένα συγκεκριμένο χρόνο λήξης για ένα αντικείμενο, ο proxy που παρέχει caching ανεξάρτητα από την σημασιολογία του αντικειμένου, είναι υποχρεωμένος να υιοθετήσει τον ίδιο χρόνο λήξης. Η μόνη διαφοροποίηση σε αυτό είναι ο περιορισμός που μπορεί να τεθεί στο αίτημα του χρήστη μέσω επικεφαλίδας Cache-Control: only-if-cached που αναγκάζει τον proxy να επιστρέψει μια ήδη αποθηκευμένη απάντηση χωρίς να ελέγξει την ορθότητα της στον πηγαίο διακομιστή. Εάν ο πηγαίος διακομιστής δεν θέσει ένα χρόνο λήξης, ο proxy μπορεί να προσδιορίσει ευρεστικά ένα χρόνο λήξης. Ο πλέον συνήθης τρόπος ελέγχου της συνέπειας στο Παγκόσμιο Ιστό είναι η αποστολή ενός GET ή HEAD αιτήματος με μια επικεφαλίδα if-modified-since. Η επικεφαλίδα μεταφέρει μια χρονοσφραγίδα (timestamp) που δεικνύει το χρόνο τελευταίας μεταβολής του αντικειμένου όπως αυτός υποδεικνύεται από τον πηγαίο διακομιστή. Σε ορισμένες περιπτώσεις, ο χρόνος παραγωγής της απάντησης μπορεί να είναι ο χρόνος τελευταίας μεταβολής του αντικειμένου. Οι ετικέτες οντότητας (entity tags) του HTTP/1.1 σε συνδυασμό με την επικεφαλίδα if-modified-since μπορούν να χρησιμοποιηθούν για την πραγματοποίηση ελέγχων συνέπειας. Ο πηγαίος διακομιστής μπορεί να απαντήσει με ένα πλήρες αντίγραφο του αντικειμένου ή με την απάντηση 304 Not Modified (και χωρίς σώμα στην απάντηση). Παρόλα αυτά ένας έλεγχος συνέπειας προϋποθέτει πλήρη διαδοχή μηνυμάτων HTTP request/response.

Εάν το caching proxy στέλνει ένα αίτημα για έλεγχο της ορθότητας του αντικειμένου κάθε φορά που συμβαίνει ένα cache hit, η πολιτική καλείται ισχυρή συνέπεια (strong concistency). Εάν το proxy χρησιμοποιεί ένα ευρεστικό αλγόριθμο για να αποφανθεί εάν το αντικείμενο είναι επίκαιρο, χωρίς να συμβουλευτεί τον πηγαίο διακομιστή, η πολιτική καλείται ασθενής συνέπεια (weak consistency). Οι δύο ευρεστικοί αλγόριθμοι για έλεγχο συνέπειας είναι οι: leased-based και time-to-live.

- *Leased-based προσέγγιση*: Η cache συμφωνεί να αποθηκεύσει ένα αντικείμενο για συγκεκριμένο χρονικό διάστημα (περίοδος χρονομίσθωσης – lease) χωρίς να ελέγχει την ορθότητα του. Ο διακομιστής "υπόσχεται" να ειδοποιήσει την cache για ενδεχόμενες αλλαγές στο αποθηκευμένο αντικείμενο κατά την διάρκεια της περιόδου χρονομίσθωσης. Εάν η περίοδος παρέλθει η cache μπορεί να ελέγξει την ορθότητα του αντικειμένου ή να ανανεώσει την χρονομίσθωση. Αυτή η προσέγγιση μεταφέρει το κόστος του revalidation στον πηγαίο διακομιστή ο οποίος θα πρέπει να γνωρίζει και παρακολουθεί όλους τους proxies στους οποίους έχει υποσχεθεί ενημερώσεις. Η προσέγγιση δεν μπορεί να κλιμακωθεί αν ο πηγαίος διακομιστής είναι υποχρεωμένος να ειδοποιήσει εκατοντάδες – χιλιάδες proxies.
- *Time-To-Live (TTL) προσέγγιση*: Τα αντικείμενα έχουν συσχετιστεί με ένα χρόνο λήξης αποθήκευσης. Όταν παρέλθει αυτό το χρονικό διάστημα, τα αντικείμενα παύουν να θεωρούνται επίκαιρα. Κατά την διάρκεια της περιόδου TTL, η cache δεν επικυρώνει τις απαντήσεις διασώζοντας εύρος ζώνης. Η απόδοση τιμής στον χρόνο TTL μπορεί να επηρεαστεί ένα πλήθος παραμέτρων όπως ο χρόνος λήξης που αναφέρεται στην επικεφαλίδα της απάντησης, η συχνότητα αναφοράς στην απάντηση, ο χρόνος τελευταίας αλλαγής του αντικειμένου.

2.7 Επικοινωνία μεταξύ cache

Ανάλογα με το πως οργανώνονται οι caches, μπορούν να δέχονται και να λαμβάνουν πληροφορίες για τα αντικείμενα για τα οποία ενδιαφέρονται. Αυτή η επικοινωνία είναι εξωτερική, ανεξάρτητη από τα request/response μηνύματα που ρέουν μεταξύ χρηστών και πηγαίων διακομιστών. Η επικοινωνία μεταξύ των caches μπορεί να βασίζεται στο HTTP αλλά συνήθως χρησιμοποιεί εξειδικευμένα, light-weight πρωτόκολλα. Εάν ένα σύνολο από caches οργανώνεται σε ιεραρχία, μια cache μπορεί να επικοινωνήσει με τις υπόλοιπες caches στο ίδιο επίπεδο να διαπιστώσει εάν το ζητούμενο αντικείμενο είναι διαθέσιμο σε αυτές. Ένα ερώτημα για κάποιο αντικείμενο μπορεί να απαντηθεί από μια ή περισσότερες caches που τυχαίνει να έχουν το αντικείμενο. Συχνά, η ανάκτηση ενός αντικειμένου από μια τοπική cache είναι προτιμότερη από την ανάκτηση από τον πηγαίο διακομιστή. Η αναμονή για την απάντηση από όλες τις caches στην ιεραρχία μπορεί να αυξήσει σημαντικά την

υστέρηση στον χρήστη. Για την υποβοήθηση της επικοινωνίας μεταξύ των caches εξετάζονται τα παρακάτω πρωτόκολλα:

- *Internet Cache Protocol (ICP)*. Μια cache που δεν διαθέτει το ζητούμενο αντικείμενο μπορεί να θέλει να ελέγξει την διαθεσιμότητα του σε μια άλλη γειτονική cache. Αυτή η επικοινωνία είναι διαφορετική από την παραδοσιακή αίτηση για ένα αντικείμενο από τον πηγαίο διακομιστή. Σε αυτή την περίπτωση οι caches αποτελούν την πηγή καθώς και τον προορισμό των ανταλλασόμενων μηνυμάτων. Ένα διαφορετικό πρωτόκολλο απαιτείται για την επικοινωνία μεταξύ των caches. Ένα από τα πρώτα πρωτόκολλα που καθιερώθηκαν σε αυτήν την επικοινωνία είναι το Internet Cache Protocol (ICP). Το ICP είναι ένα πρωτόκολλο ερωταποκρίσεων. Το μήνυμα που στέλνεται από μια cache-πελάτη είναι μια ερώτηση για το αν ο ομότιμος κόμβος έχει ένα αντίγραφο από το αντικείμενο που χρειάζεται η συγκεκριμένη cache. Ο δημοφιλής χαρακτήρας του ICP οφείλεται στο γεγονός ότι χρησιμοποιείται από το Squid. Το ICP χρησιμοποιείται σε ιεραρχίες από caches, σύνολα caches που συνδέονται μεταξύ τους και κάτω από ένα κοινό γονέα. Η διαδικασία αυτή επαναλαμβάνεται, και η μετακίνηση προς τα ανώτερα επίπεδα της ιεραρχίας σημαίνει μετακίνηση προς μια περισσότερο κεντρική cache. Οι κεντρικές caches μπορεί να έχουν μια περιφερειακή (regional) cache ως άμεσο πρόγονο ενώ οι περιφερειακές caches μπορεί να έχουν μια εθνική cache ως πρόγονο. Αν υποτεθεί ότι η cache OriginalCache δεν έχει κάποιο ζητούμενο πόρο, θα σταλούν ICP αιτήσεις (πάνω από UDP) σε όλους τους ομότιμους κόμβους ταυτόχρονα. Εάν κάποιος από τους ομότιμους κόμβους διαθέτει τον αιτούμενο αντικείμενο, η OriginalCache θα ενημερωθεί σχετικά και θα ζητήσει την ανάκτηση του χρησιμοποιώντας HTTP. Εάν κανείς από τους ομότιμους κόμβους δεν διαθέτει το αντικείμενο, η OriginalCache θα προωθήσει την αίτηση στον γονέα της. Ο γονέας της OriginalCache επαναλαμβάνει την διαδικασία. Εάν καμιά από τις caches δεν διαθέτει τον πόρο, η OriginalCache θα πρέπει να προωθήσει την αίτηση στον πηγαίο διακομιστή. Η φιλοσοφία βάσει της οποίας λειτουργεί το ICP είναι ότι η αποστολή των ICP queries, ακόμη και αν αυτή επαναληφθεί πολλές φορές σε διάφορα επίπεδα της ιεραρχίας, είναι σημαντικά γρηγορότερη την επικοινωνία με τον πηγαίο διακομιστή.

- *Cache Array Resolution Protocol (CARP)*. Το CARP καθορίζει ένα μηχανισμό μέσω του οποίου ένα σύνολο από caching proxies μπορούν να λειτουργήσουν ως μια λογικά ενιαία cache. Ο μηχανισμός χειρίζεται το σύνολο των απαντήσεων που αποθηκεύονται στη cache συλλογικά μεταξύ της ομάδα (array) των proxies ως μια μεγάλη cache. Μια συνάρτηση κατακερματισμού του κλειδιού (hash function) χρησιμοποιείται για να διαιρεθεί το σύνολο των URL μεταξύ των caches. Ένας χρήστης που προσπαθεί να εντοπίσει ένα αποθηκευμένο στη cache αντικείμενο μπορεί να κατευθύνει την αίτηση στην κατάλληλη cache εφαρμόζοντας την συνάρτηση κατακερματισμού του κλειδιού. Η συνάρτηση αυτή χρησιμοποιεί το αιτούμενο URL καθώς και την ταυτότητα των μελών του proxy για να διαμορφώσει ένα μονοπάτι επίλυσης (resolution path). Εάν συγκριθεί με το ICP, το CARP έχει ντετερμινιστικό μονοπάτι επίλυσης της αίτησης, εξαλείφοντας έτσι την ανάγκη για μηνύματα ερωταποκρίσεων (queries). Επίσης, υπάρχουν λιγότερα διπλότυπα αποθηκευμένων αντικειμένων στο CARP από το ICP. Το CARP χρησιμοποιεί το HTTP καθώς και απομακρυσμένες κλήσεις διεργασιών (Remote Procedure Calls) για την επικοινωνία μεταξύ των proxies. Ο κάθε proxy συσχετίζεται με ένα παράγοντα φορτίου (load factor) που λαμβάνεται υπόψη πριν μια αίτηση οδηγηθεί σε ένα συγκεκριμένο proxy. Το CARP διατίθεται ως προϊόν από την Microsoft.
- *Cache Digest Protocol (CDP)*. Το CDP αποτελεί μια επέκταση του ICP. Η βασική ιδέα στο CDP είναι η δυνατότητα ανταλλαγής μιας περίληψης (digest) των περιεχομένων της cache. Η περίληψη αποτελεί μια ένδειξη της συλλογής των αντικειμένων σε μια cache. Όταν μια cache έχει στην διάθεση της μια περίληψη από όλους τους ομότιμους κόμβους μπορεί, πολύ εύκολα να ανατρέξει στην περίληψη και να εξετάσει εάν το αιτούμενο αντικείμενο είναι διαθέσιμο σε μια από τις caches. Εάν η διερεύνηση αυτή επιτύχει, ο συγκεκριμένος ομότιμος κόμβος είναι υποψήφιος να δεχθεί μια αίτηση ανάκτησης του αντικειμένου. Εάν ο έλεγχος στις περιλήψεις αποτύχει, οι αντίστοιχες caches δεν ερωτούνται με προφανές αποτέλεσμα την δραστική μείωση των ερωτημάτων που απευθύνονται στο σύνολο των ομότιμων κόμβων. Ένα πρόβλημα του μηχανισμού CDP είναι η εγκυρότητα των περιλήψεων και τα λανθασμένα ερωτήματα τα οποία οφείλονται σε αυτήν. Ένα αντικείμενο μπορεί να αφαιρεθεί από μια cache μετά την διαμόρφωση της σχετικής περίληψης. Ένα ακόμη πρόβλημα είναι το μέγεθος των περιλήψεων και η

ανταλλαγή τους μεταξύ των ομοτίμων κόμβων. Οι περιλήψεις ανταλλάσσονται μέσα σε HTTP μηνύματα, πάνω από το TCP, για λόγους αξιοπιστίας. Μια περίληψη μπορεί να θεωρηθεί σαν ένας κανονικός αντικείμενο και οι τεχνικές ελέγχου της ορθότητας του HTTP (resource revalidation) μπορούν να χρησιμοποιηθούν για την διερεύνηση του επίκαιρου της περίληψης.

- *Web Cache Coordination Protocol (WCCP)*. Το WCCP είναι ένας μηχανισμός συντονισμού, στενά δεμένος με το επίπεδο δικτύου. Ο σκοπός του WCCP είναι η παρεμπόδιση (intercept) της HTTP αίτησης και η ανακατεύθυνση της στην μηχανή cache. Επειδή η αίτηση θα αποτύχει εάν η cache δεν είναι, για κάποιο λόγο, διαθέσιμη, ένας μηχανισμός συντονισμού απαιτείται. Το αντικείμενο του μηχανισμού συντονισμού είναι να εξισορροπεί τον φόρτο μεταξύ διαφορετικών caches, έχοντας πλήρη γνώση της διαθεσιμότητας τους. Ελέγχοντας περιοδικά την διαθεσιμότητα μιας cache, ο μηχανισμός εξασφαλίζει ότι δεν πρόκειται να προωθηθούν πακέτα σε μια cache που δεν ανταποκρίνεται σε έλεγχο διαθεσιμότητας (hearbeat check). Ένας τέτοιος μηχανισμός αποτελεί την βάση του πρωτοκόλλου WCCP που υλοποιείται σαν τμήμα της Cisco Cache Engine. Η μηχανή cache ρυθμίζεται ώστε να δέχεται WWW αιτήσεις που ανακατευθύνονται σε αυτήν από ένα δρομολογητή. Ο δρομολογητής, που έχει ενεργοποιημένο το WCCP, μπορεί να επεξεργάζεται όλες τις IP επικεφαλίδες. Ένα TCP πακέτο που στοχεύει στην θύρα 80 ανακατευθύνεται στη cache. Επιπλέον, οι δρομολογητές που διαθέτουν WCCP επικοινωνούν περιοδικά με τις μηχανές caching για να εξασφαλίσουν την διαθεσιμότητα τους.

2.8 Διανομή Περιεχομένου (Content Distribution)

Η τεχνική της διανομής περιεχομένου αναφέρεται στην εφαρμογή επιλεκτικού κατοπτρισμού (selective mirroring). Η βασική ιδέα στην διανομή περιεχομένου είναι να μειωθεί ο φόρτος στον πηγαίο διακομιστή. Αυτό μπορεί να επιτευχθεί παρέχοντας τμήμα ή όλο το περιεχόμενο από ένα σύνολο αντιγράφων (replicas). Διάφορες τεχνικές χρησιμοποιούνται για την ανακατεύθυνση των αιτήσεων στα αντίγραφα (π.χ., τεχνικές που βασίζονται στο DNS). Ένας τρόπος για την διαίρεση (partitioning) ενός αντικειμένου είναι στα συστατικά βάσης (base) και εμφωλευμένων στοιχείων (embedded). Ένα έγγραφο βάσης (base document) αποτελεί το container έγγραφο και

τα εμφωλευμένα συστατικά είναι οι εικόνες ή τα scripts τα οποία αποτελούν τμήμα της WWW σελίδας. Οι διακομιστές που χρησιμοποιούνται για να εξυπηρετήσουν το non-container τμήμα του αντικειμένου καλούνται διακομιστές διανομής περιεχομένου. Μπορεί να εντοπίζονται κοντά στον πηγαίο διακομιστή ή οπουδήποτε στο διαδίκτυο. Τα εμφωλευμένα στοιχεία υπάρχουν ως αντίγραφα (replicated) σε αυτούς τους διακομιστές. Κατά την αίτηση, η υπηρεσία διανομής περιεχομένου (content distribution service) προσπαθεί να εντοπίσει τον διακομιστή διανομής περιεχομένου που βρίσκεται "πλησιέστερα" στον χρήστη για να του επιστρέψει τα εμφωλευμένα στοιχεία. Η εγγύτητα ενός διακομιστή διανομής περιεχομένου μπορεί να αναφέρεται σε γεωγραφική απόσταση, σε δικτυακή απόσταση και μετρικές υστέρησης (latency metrics). Αυτή η προσέγγιση ελαττώνει τον φόρτο στον διακομιστή βάσης (base server) και βελτιώνει τον χρόνο απόκρισης για τον τελικό χρήστη.

Ο στόχος της διανομής περιεχομένου δεν είναι διαφορετικός από αυτόν του caching. Και οι δύο προσεγγίσεις μετακινούν το περιεχόμενο κοντά στον τελικό χρήστη φιλοδοξώντας να μειώσουν την υστέρηση που αντιλαμβάνεται ο χρήστης καθώς και τον φόρτο στον πηγαίο διακομιστή. Με το caching, τα proxies πρέπει να διατηρούν την συνέπεια (consistency) και να επαληθεύουν την ορθότητα των αποθηκευμένων αντικειμένων. Με το μηχανισμό διανομής περιεχομένου, οι διακομιστές διανομής περιεχομένου έχουν πλήρη έλεγχο επί του περιεχομένου και μπορούν να προβούν σε ρυθμίσεις με τους διακομιστές που διαθέτουν περιεχόμενο για λογαριασμό τους.

Στο μηχανισμό κατοπτρισμού cache (cache mirroring), μεγάλα τμήματα ενός δικτυακού τόπου (site) κατοπτρίζονται (mirrored) σε διάφορους κόμβους του διαδικτύου. Στην προσέγγιση διανομής περιεχομένου, οι πηγαίοι διακομιστές αποφασίζουν ποιοι πόροι μπορούν να αντιγραφούν, και μεταφέρουν το έργο του κατοπτρισμού σε άλλο οργανισμό. Οι πηγαίοι διακομιστές είναι υποχρεωμένοι να ειδοποιούν την εταιρίας που αναλαμβάνει τη διανομή περιεχομένου όταν οι πόροι τους υφίστανται μεταβολές.

Ένα παράδειγμα εταιρίας που αναλαμβάνει διανομή περιεχομένου είναι η Akamai. Ένας δικτυακός τόπος που θέλει να διανεμηθούν τα περιεχόμενα του μέσω του Akamai, μετονομάζει αυτά τα URLs με συγκεκριμένο πρόθεμα. Το πρόθεμα περιέχει το όνομα ενός κόμβου (hostname string). Η DNS επίλυση του ονόματος κόμβου επιστρέφει την IP διεύθυνση ενός Akamai διακομιστή αντικατοπτρισμού (mirror

server) που είναι πολύ πιθανό να περιέχει αντίγραφο του αντικειμένου. Η απόφαση για την επιστροφή μιας IP διευθύνσεως λαμβάνεται από τον DNS server του Akamai δικτύου. Σχεδιαστικά, ο προσδιοριζόμενος διακομιστής Akamai είναι πλησιέστερα στον τοπικό DNS server του χρήστη που αιτήθηκε το αντικείμενο. Η προσδοκία είναι ότι ο χρήστης είναι αρκετά κοντά στον διακομιστή DNS (δικτυακή απόφαση) και το αντικείμενο θα πρέπει να μεταφερθεί για μια μικρή σχετικά απόσταση. Επειδή πρέπει να απεικονιστεί το όνομα κόμβου που περιέχεται στο URL, είναι δυνατό για το Akamai να χρησιμοποιήσει τον διακομιστή DNS για να προσδιορίσει τον κατάλληλο διακομιστή Akamai που διαθέτει τον ζητούμενο πόρο.

Η τεχνική του Akamai πρέπει να εξασφαλίσει ότι η DNS αναζήτηση επιστρέφει το πλησιέστερο mirror site. Οι DNS τιμές TTL πρέπει να ρυθμιστούν κατάλληλα ώστε να αποφευχθεί να παραμένουν αποθηκευμένες στη cache οι DNS απαντήσεις για μεγάλο διάστημα. Διαφορετικά, ένας χρήστης που αναζητάει κάποιον κόμβο μπορεί να χρησιμοποιήσει μια παλιά IP διεύθυνση ενός διακομιστή Akamai που δεν αποτελεί, πλέον, την καταλληλότερη επιλογή για το ζητούμενο αντικείμενο. Υπάρχει ένα trade-off μεταξύ του εντοπισμού της καλύτερης επιλογής και του κόστους των DNS αναζητήσεων. Υπάρχουν όμως ορισμένα προβλήματα με την προσέγγιση διανομής περιεχομένου. Ο πηγαίος διακομιστής ωφελείται από τον περιορισμένο φόρτο ενώ οι τελικοί χρήστες ωφελούνται από την άντληση αντικειμένων από "κοντινούς" χρήστες. Η θέση των διακομιστών διανομής περιεχομένου μπορεί να είναι ένα πρόβλημα για τους χρήστες. Είναι δυνατόν ορισμένοι χρήστες να έχουν καλύτερη δικτυακή συνδεσιμότητα (χαμηλότερα RTT) προς τον πηγαίο διακομιστή από τους διακομιστές διανομής περιεχομένου. Τεχνικά, οι διακομιστές διανομής περιεχομένου εργάζονται για λογαριασμό των διακομιστών βάσης. Οι χρήστες εγκαθιστούν απευθείας συνδέσεις με τους διακομιστές διανομής περιεχομένου και αναμένουν να είναι αυτοί συμβατοί με το HTTP πρωτόκολλο. Οι δικτυακοί τόποι διανομής περιεχομένου χρησιμοποιούν διαφορετικά πρωτόκολλα στην επικοινωνία τους με τους διακομιστές βάσης και μπορούν να χρησιμοποιούν και άλλους μηχανισμούς για να διασφαλίσουν ότι διαθέτουν επικαιροποιημένο περιεχόμενο.

3 Θεωρία Παιγνίων

3.1 Γενικά

Η Θεωρία Παιγνίων μπορεί να οριστεί ως η μελέτη της αντιπαράθεσης (conflict) και της συνεργασίας. Οι Παιγνιοθεωρητικές έννοιες εφαρμόζονται κάθε φορά που υπάρχει αλληλεπίδραση στις ενέργειες οντοτήτων που ονομάζονται παίκτες (players). Οι παίκτες μπορεί να είναι άτομα, ομάδες, επιχειρήσεις ή και συνδυασμός αυτών.

Για τους παίκτες γίνεται, κατά κανόνα, η υπόθεση ότι είναι λογικοί (rational). Ένας λήπτης αποφάσεων χαρακτηρίζεται λογικός αν λαμβάνει αποφάσεις που βρίσκονται σε συνέπεια με τις επιδιώξεις του. Στη Θεωρία Παιγνίων, κάθε παίκτης έχει ως στόχο του την μεγιστοποίηση της ωφέλειας (payoff) του. Για κάθε λογικό λήπτη αποφάσεων υπάρχει ένας τρόπος ανάθεσης αριθμητικών τιμών ωφέλειας στα διάφορα πιθανά ενδεχόμενα που μπορεί να προκύψουν, έτσι ώστε να διαλέγει πάντοτε τη λύση που μεγιστοποιεί την αναμενόμενη ωφέλεια του. Αυτό αποτελεί το θεώρημα μεγιστοποίησης της αναμενόμενης ωφέλειας (expected utility maximization theorem).

Τα λογικά αξιώματα τα οποία υποστηρίζουν το θεώρημα μεγιστοποίησης της αναμενόμενης ωφέλειας είναι υποθέσεις ασθενούς συνέπειας (weak consistency assumptions). Η βασική υπόθεση είναι το αξίωμα αντικατάστασης (substitution axiom) το οποίο μπορεί να αποδοθεί ως εξής:

> *"εάν ένας λήπτης αποφάσεων προτιμάει την εναλλακτική λύση 1 από την 2, όταν συμβεί το συμβάν Α και προτιμάει την λύση 1 ακόμη και αν δεν συμβεί η Α, τότε θα προτιμήσει το ενδεχόμενο 1 ακόμη και αν δεν γνωρίσει εάν στο συμβάν Α συμβεί ή όχι".*

Αυτή η υπόθεση είναι σε θέση να εξασφαλίσει ότι υπάρχει κάποια κλίμακα ωφέλειας στην οποία ο λήπτης απόφασης πάντα προτιμά τις εναλλακτικές επιλογές που μεγιστοποιούν την αναμενόμενη συνάρτηση ωφέλειας.

Γενικά, η μεγιστοποίηση της αναμενόμενης ωφέλειας δεν συμβαδίζει με την μεγιστοποίηση της χρηματικής απολαβής (monetary payoff). Η ωφέλεια (utility payoff) ενός ατόμου μπορεί να εξαρτάται από πολλαπλές μεταβλητές και όχι μόνο την

χρηματική αξία. Όταν υπάρχει αβεβαιότητα, η αναμενόμενη ωφέλεια μπορεί να καθοριστεί και να υπολογιστεί μόνο αν έχουν ανατεθεί πιθανότητες στα αβέβαια ενδεχόμενα συμβάντα. Οι πιθανότητες αυτές προσδιορίζουν ποσοτικά την πιθανότητα ενός συμβάντος.

3.2 Ιστορικά Στοιχεία

Κάποιες πρώιμες Παιγνιοθεωρητικές ιδέες εμφανίστηκαν από τους James Waldegrave το 1713 και τον Antoine Cournot το 1838 (στη μελέτη ενός δυοπωλίου). Ωστόσο η ανάπτυξη της Θεωρίας ξεκίνησε τις πρώτες δεκαετίες του 20ου αιώνα από τους μαθηματικούς Ernst Zermelo, Emile Borel και John von Neumann. Η Θεωρία Παιγνίων καθιερώθηκε ως επιστημονικό πεδίο το 1944 μετά τη δημοσίευση του βιβλίου "Theory of Games and Economic Behavior" των John von Neumann και Oskar Morgenstern. Το βιβλίο αυτό καθιέρωσε ένα μεγάλο μέρος της βασικής ορολογίας της θεωρίας, η οποία χρησιμοποιείται έως σήμερα.

Το 1950, ο John Nash έδειξε ότι τα πεπερασμένα παιχνίδια έχουν πάντα ένα σημείο ισορροπίας (σημείο ισορροπίας Nash), στο οποίο όλοι οι παίκτες επιλέγουν ενέργειες που είναι βέλτιστες, δοθέντων των επιλογών όλων των άλλων παικτών. Αυτή η κεντρική έννοια της μη συνεργατικής Θεωρίας Παιγνίων αποτελεί από τότε βασικό σημείο ανάλυσης και μελέτης. Περιέγραψε επίσης μια ευρεία κλάση παιγνίων για τα οποία υπάρχει πάντα μια τέτοια ισορροπία. Η έννοια της ισορροπίας Nash διεύρυνε εντυπωσιακά τα όρια αντίληψης της Θεωρίας Παιγνίων. Αμέσως μετά τη δημοσίευση της μελέτης του Nash, διάφορα Παιγνιοθεωρητικά μοντέλα άρχισαν να χρησιμοποιούνται στην Οικονομική Θεωρία και την πολιτική επιστήμη, ενώ οι ψυχολόγοι ξεκίνησαν να μελετούν πώς συμπεριφέρονται οι άνθρωποι σε πειραματικά παίγνια.

Στις δεκαετίες του 1950 και 1960, η Θεωρία Παιγνίων διευρύνθηκε θεωρητικά και εφαρμόστηκε σε προβλήματα στρατηγικής πολέμου και πολιτικής. Το 1965 ο Reinhard Selten διατύπωσε μια νέα έννοια ισορροπίας, εκλεπτύνοντας την έννοια της ισορροπίας Nash. Το 1967 ο John Harsanyi ανάπτυξε τις αρχές που διέπουν τα παίγνια στα οποία οι παίκτες έχουν ατελή πληροφόρηση.

Από τη δεκαετία του 1970, έχει δημιουργήσει επανάσταση στην Οικονομική Θεωρία. Επιπλέον, έχει βρει εφαρμογές στην Ψυχολογία, την Κοινωνιολογία και την Βιολογία (κυρίως ως αποτέλεσμα της εργασίας του βιολόγου John Maynard Smith).

Η Θεωρία Παιγνίων κέρδισε το παγκόσμιο ενδιαφέρον το 1994 με την απονομή του βραβείου Νόμπελ στα Οικονομικά στους John Nash, John Harsanyi και Reinhard Selten. Στα τέλη της δεκαετίας του 1990 η Θεωρία Παιγνίων βρήκε εφαρμογή στο σχεδιασμό δημοπρασιών. Διακεκριμένοι μελετητές της Θεωρίας Παιγνίων ασχολήθηκαν με το σχεδιασμό δημοπρασιών για την κατανομή δικαιωμάτων χρήσης του ηλεκτρομαγνητικού φάσματος στη βιομηχανία των κινητών τηλεπικοινωνιών. Οι περισσότερες από αυτές τις δημοπρασίες έχουν σχεδιαστεί με σκοπό την κατανομή των πόρων αυτών με περισσότερο αποδοτικό τρόπο σε σχέση με τις παραδοσιακές κυβερνητικές πολιτικές.

Το 2005 τιμήθηκαν με το βραβείου Νόμπελ στα Οικονομικά οι Thomas Schelling και Robert Aumann με το σκεπτικό ότι εμπλούτισαν την αντίληψη μας σχετικά με τις έννοιες του ανταγωνισμού και της συνεργασίας μέσω της Παιγνιοθεωρητικής ανάλυσης. Το 2007 τιμήθηκαν με το ίδιο βραβείο οι Roger Myerson, Leonid Hurwicz και Eric Maskin για τη θεμελίωση της θεωρίας σχεδιασμού μηχανισμών.

3.3 Κατηγορίες Παιγνίων

Όπως έχει αναφερθεί, το αντικείμενο μελέτης της Θεωρίας Παιγνίων είναι το παίγνιο, στο οποίο εμπλέκονται δύο ή περισσότεροι παίκτες. Ένα παίγνιο με έναν μόνο παίκτη αντιμετωπίζεται από την Θεωρία Αποφάσεων. Ένα παίγνιο ορίζει το σύνολο των παικτών, το σύνολο των ενεργειών (δηλαδή των αποφάσεων) που μπορεί να πάρει ο κάθε παίκτης, καθώς και τους προσωπικούς στόχους που έχει ο κάθε παίκτης. Ωστόσο, ένα παίγνιο δεν καθορίζει τις ενέργειες που τελικά ακολουθούν οι παίκτες. Μια λύση στη Θεωρία Παιγνίων είναι η συστηματική περιγραφή των εκβάσεων που μπορεί να έχει ένα παίγνιο. Η Θεωρία Παιγνίων ορίζει εύλογες έννοιες λύσης για διάφορες οικογένειες παιγνίων και μελετά τις ιδιότητες των λύσεων αυτών. Τα παίγνια ταξινομούνται σε διάφορα είδη μέσω ποικίλων κριτηρίων.

3.3.1 Μη Συνεργατικά Παίγνια (Noncooperative Games) - Συνεργατικά Παίγνια (Cooperative Games)

Η βασική υπόθεση, στα μη συνεργατικά παίγνια, είναι ότι κάθε παίκτης δρα μόνος του, προσπαθώντας να μεγιστοποιήσει κάποια ατομική συνάρτηση οφέλους, δεδομένης της πρόβλεψης του για το πως θα δράσουν οι υπόλοιποι παίκτες. Ενώ στη περίπτωση των συνεργατικών παιγνίων αναπτύσσονται αξιώματα τα οποία προσεγγίζουν την έννοια της δίκαιης κατανομής των ωφελειών που προκύπτουν από τη συλλογική δράση ενός συνόλου παικτών.

3.3.2 Παίγνια σε Στρατηγική Μορφή (Strategic Games) - Παίγνια σε Εκτεταμένη Μορφή (Extensive Games)

Ένα στρατηγικό παίγνιο ή παίγνιο σε στρατηγική μορφή ή παίγνιο σε κανονική μορφή είναι ένα μοντέλο παιγνίου στο οποίο κάθε παίκτης επιλέγει το πλάνο δράσης του μια μόνο φορά, και όλοι οι παίκτες κάνουν τις επιλογές τους ταυτόχρονα. Αυτό σημαίνει ότι κάθε παίκτης λαμβάνει τις αποφάσεις του χωρίς να γνωρίζει τι έχουν αποφασίσει οι υπόλοιποι παίκτες. Αντίθετα, ένα παίγνιο σε εκτεταμένη μορφή καθορίζει τις δυνατές αλληλουχίες γεγονότων: κάθε παίκτης επιλέγει το πλάνο δράσης του όχι μόνο κατά την έναρξη του παιγνίου, αλλά και όποτε χρειαστεί να λάβει κάποια απόφαση.

3.3.3 Παίγνια Τέλειας Πληροφόρησης (Games with Perfect Information) - Ατελούς Πληροφόρησης (Games with Imperfect Information)

Σε ένα παίγνιο τέλειας πληροφόρησης οι παίκτες είναι πλήρως ενημερωμένοι σχετικά με τις κινήσεις των υπολοίπων παικτών, ενώ σε ένα παίγνιο ατελούς πληροφόρησης η ενημέρωση τους δεν είναι απαραίτητα πλήρης.

Εκτός από τις παραπάνω βασικές κατηγορίες, έχουμε και διάφορες άλλες υποκατηγορίες παιγνίων, όπως τα συμμετρικά παίγνια (Symmetric Games), στα οποία όλοι οι παίκτες έχουν τις ίδιες στρατηγικές και τα ίδια κέρδη ανά στρατηγική, τα παίγνια μηδενικού αθροίσματος (Zero-Sum Games), όπου κάθε παίκτης κερδίζει ότι χάνουν οι υπόλοιποι, έτσι ώστε οι συνολικές απολαβές να είναι σταθερές, κ.α. Επίσης

ανάλογα με τις καταστάσεις που μοντελοποιεί ένα παίγνιο, μιλάμε για εξελικτικά παίγνια (Evolutionary Games), παίγνια σε δίκτυα (Network Games), παίγνια συμφόρησης (Congestion Games), κ.α.

3.4 Συνάρτηση Ωφέλειας

Η συνάρτηση ωφέλειας του παίκτη αντιπροσωπεύει τις προτιμήσεις του, πραγματοποιώντας μια αντιστοίχηση από μια κατάσταση του κόσμου (ή έκβαση του παιγνίου) σε έναν πραγματικό αριθμό. Όσο μεγαλύτερος είναι ο αριθμός αυτός, τόσο περισσότερο ικανοποιημένος είναι ο παίκτης από τη συγκεκριμένη κατάσταση. Στη Θεωρία Παιγνίων η συνάρτηση αυτή είναι γνωστή ως η συνάρτηση ωφέλειας Neumann-Morgenstern. Συγκεκριμένα, δεδομένου ότι το S αντιπροσωπεύει το σύνολο των καταστάσεων που μπορεί να αντιληφθεί ο παίκτης, η συνάρτηση ωφέλειας του παίκτη i έχει την παρακάτω μορφή:

$$u_i : S \to \boldsymbol{R}$$

Πρέπει να σημειωθεί ότι οι καταστάσεις ορίζονται ως αυτές οι καταστάσεις που μπορεί να αντιληφθεί ο παίκτης. Η δημιουργία μιας συνάρτησης ωφέλειας μπορεί να είναι μια δύσκολη διαδικασία, καθώς απαιτεί εις βάθος κατανόηση του μελετούμενου προβλήματος.

Δεδομένης μιας συνάρτησης ωφελείας, είναι δυνατόν να οριστεί μια κατάταξη των πιθανών καταστάσεων, σε σχέση με το πόσο επιθυμητές είναι. Συγκρίνοντας τις τιμές ωφέλειας δύο καταστάσεων, μπορεί να καθοριστεί ποια προτιμάει ο χρήστης. Η εν λόγω κατάταξη έχει τις παρακάτω ιδιότητες:

- Αυτοπαθής : $u_i(s) \geq u_i(s)$
- Μεταβατική : Αν $u_i(a) \geq u_i(b)$ και $u_i(b) \geq u_i(c)$, τότε $u_i(a) \geq u_i(c)$
- Συγκρίσιμη : Για κάθε $a,b \in S$, είτε $u_i(a) \geq u_i(b)$, είτε $u_i(a) \leq u_i(b)$

Οι συναρτήσεις ωφέλειας μπορούν να χρησιμοποιηθούν για την περιγραφή της συμπεριφοράς κάθε παίκτη, ποσοτικοποιώντας τις διάφορες ανταλλαγές (tradeoff) που αντιμετωπίζει, μαζί με την τιμή (ή αναμενόμενη τιμή) των ενεργειών του. Εφόσον έχει οριστεί μια συνάρτηση ωφέλειας για όλους τους παίκτες, αυτό που έχουν να κάνουν είναι να επιλέξουν τις ενέργειες που μεγιστοποιούν την ωφέλεια τους. Όπως και στα Οικονομικά, η λέξη "εγωιστής" χρησιμοποιείται για να χαρακτηρίσει ένα λογικό παίκτη

που στοχεύει στη μεγιστοποίηση της ωφέλειας του. Πρέπει να σημειωθεί ότι αυτή η χρήση της λέξης διαφέρει ελαφρώς σε σχέση με την καθημερινή χρήση της που συχνά υπονοεί μια διάθεση πρόκλησης βλάβης στους άλλους. Ένας πραγματικά εγωιστής παίκτης ενδιαφέρεται αποκλειστικά για τη δική του ωφέλεια.

Πρέπει να σημειωθεί ότι η υπόθεση εγωιστών παικτών δεν αποκλείει τη μελέτη συνεργατικών δομών. Μια τέτοια δομή μπορεί να εξεταστεί ως μια κατάσταση όπου οι συναρτήσεις ωφελείας των παικτών έχουν οριστεί με τρόπο τέτοιο που οι παίκτες φαίνονται σαν να συνεργάζονται. Για παράδειγμα, αν ένας παίκτης λαμβάνει υψηλότερη ωφέλεια όταν βοηθάει τους άλλους παίκτες, η συμπεριφορά που θα προκύψει θα μπορεί να χαρακτηριστεί ως συνεργατική, από κάποιον εξωτερικό παρατηρητή, παρόλο που ο παίκτης λειτουργεί εγωιστικά.

3.5 Μη Συνεργατικά Παίγνια

3.5.1 Παίγνια σε Στρατηγική Μορφή

Η στρατηγική μορφή αποτελεί τον πιο απλό τρόπο αναπαράστασης ενός παιγνίου. Για να ορίσουμε ένα παίγνιο σε στρατηγική μορφή αρκεί να ορίσουμε το σύνολο των παικτών, το σύνολο των διαθέσιμων επιλογών για κάθε παίκτη, και τον τρόπο με τον οποίο εξαρτώνται οι ωφέλειες των παικτών από τις επιλογές τους.

Τυπικά, ένα παίγνιο σε στρατηγική μορφή ή στρατηγικό παίγνιο Γ είναι της μορφής

$$\Gamma = \left\{ N, \left(S_i, u_i\right)_{i \in N} \right\}$$

όπου N είναι ένα μη κενό σύνολο και, για κάθε $i \in N$, το S_i είναι ένα μη κενό σύνολο και $u_i : S \to \boldsymbol{R}$ είναι μια συνάρτηση.

Το $N = \{1, 2, \ldots, n\}$ είναι το σύνολο των παικτών στο παίγνιο Γ. Για κάθε παίκτη i, το σύνολο S_i είναι το σύνολο των καθαρών στρατηγικών που είναι διαθέσιμες στον i. Όταν παίζεται το στρατηγικό παίγνιο Γ, κάθε παίκτης i πρέπει να επιλέξει μια από τις στρατηγικές στο σύνολο S_i. Ένα προφίλ στρατηγικών είναι ένας συνδυασμός στρατηγικών (μια για κάθε παίκτη) που θα μπορούσαν να επιλέξουν οι παίκτες στο N. Συμβολίζουμε με S το σύνολο όλων των δυνατών προφίλ στρατηγικών, έτσι ώστε

$$S = \times_{i \in N} S_i = S_1 \times S_2 \times \ldots \times S_n$$

Επίσης συμβολίζουμε με $S_{-i} = \times_{j \neq i} S_j = S_1 \times \ldots \times S_{i-1} \times S_{i+1} \times \ldots \times S_n$, το καρτεσιανό γινόμενο των συνόλων των καθαρών στρατηγικών όλων των παικτών, εκτός του i.

Για κάθε προφίλ στρατηγικών $s = (s_1, s_2, \ldots, s_i, \ldots, s_n) \in S$, ο αριθμός $u_i(s)$ εκφράζει την αναμενόμενη ωφέλεια που θα έχει ο παίκτης i (δηλαδή το κέρδος του) αν το s είναι ο συνδυασμός στρατηγικών που επιλέγουν οι παίκτες. Όταν μελετάμε ένα στρατηγικό παίγνιο, υποθέτουμε ότι όλοι οι παίκτες επιλέγουν ταυτόχρονα τις στρατηγικές τους, και επομένως δεν υπάρχει η παράμετρος του χρόνου στην ανάλυση ενός στρατηγικού παιγνίου.

Ένα στρατηγικό παίγνιο είναι πεπερασμένο αν το σύνολο παικτών N και όλα τα σύνολα των στρατηγικών S_i, $i \in N$, είναι πεπερασμένα.

Επομένως:

Ορισμός 1. *Ένα παίγνιο σε στρατηγική μορφή* $\Gamma = \{N, (S_i, u_i)_{i \in N}\}$ *αποτελείται από*

- *ένα μη κενό σύνολο* N *των παικτών*
- *για κάθε παίκτη* $i \in N$ *ένα μη κενό σύνολο* S_i *των καθαρών στρατηγικών που έχει στη διάθεση του ο παίκτης* i
- *για κάθε παίκτη* $i \in N$ *μια συνάρτηση ωφέλειας* $u_i : S \to \boldsymbol{R}$.

Αν τα σύνολα N *και* S_i, $i \in N$, *είναι πεπερασμένα τότε το* Γ *είναι πεπερασμένο παίγνιο.*

Μικτή στρατηγική σ_i, $i \in N$, ενός στρατηγικού παιγνίου $\Gamma = \{N, (S_i, u_i)_{i \in N}\}$, είναι μια κατανομή πιθανότητας στο σύνολο των αγνών στρατηγικών του S_i. Στα επόμενα, για ένα πεπερασμένο σύνολο X, συμβολίζουμε με $\Delta(X)$ το σύνολο όλων των κατανομών πιθανότητας πάνω στο X. Επομένως

$$\sigma_i \in \Delta(S_i)$$

Ένα προφίλ μικτών στρατηγικών $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_n)$ είναι ένα διάνυσμα που καθορίζει μια μικτή στρατηγική για κάθε παίκτη. Γράφουμε $\sigma \in \Delta(S)$, όπου $\Delta(S)$ είναι το σύνολο όλων των δυνατών προφίλ μικτών στρατηγικών όλων των παικτών και είναι ίσο με $\Delta(S) = \times_{i \in N} \Delta(S_i) = \Delta(S_1) \times \Delta(S_2) \times \ldots \times \Delta(S_n)$. Δηλαδή το $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_n)$ είναι ένα προφίλ μικτών στρατηγικών στο $\Delta(S)$ αν και μόνο αν, για κάθε παίκτη i και για κάθε καθαρή στρατηγική $s_i \in S_i$, το σ καθορίζει έναν πραγματικό αριθμό $\sigma_i(s_i) \geq 0$ που εκφράζει την πιθανότητα ο παίκτης i να επιλέξει την στρατηγική s_i, έτσι ώστε

$$\sum_{s_i \in S_i} \sigma_i(s_i) = 1, \quad \text{για κάθε } i \in N$$

Επίσης συμβολίζουμε με $\Delta(S_{-i}) = \times_{j \neq i} \Delta(S_j) = \Delta(S_1) \times \ldots \times \Delta(S_{i-1}) \times \Delta(S_{i+1}) \times \ldots \times \Delta(S_n)$, το καρτεσιανό γινόμενο των συνόλων των μικτών στρατηγικών όλων των παικτών, εκτός του i.

Αν και δεν μπορούμε να είμαστε βέβαιοι σχετικά με το ποιο προφίλ στρατηγικών θα επιλεγεί από τους παίκτες όταν παίζεται το παίγνιο Γ, η Θεωρία αποφάσεων κατά Bayes εγγυάται ότι υπάρχει κάποια κατανομή πιθανότητας στο σύνολο των προφίλ στρατηγικών $S = S_1 \times S_2 \times \ldots \times S_n$ το οποίο εκφράζει ποσοτικά τις πεποιθήσεις μας σχετικά με τις στρατηγικές που θα επιλέξουν οι παίκτες. Επιπλέον, επειδή υποθέτουμε ότι όλοι οι παίκτες επιλέγουν ταυτόχρονα και ανεξάρτητα τις στρατηγικές τους, οι πεποιθήσεις μας σχετικά με το παίγνιο πρέπει να αντιστοιχούν σε κάποιο προφίλ μικτών στρατηγικών $\sigma \in \Delta(S)$.

Για κάθε προφίλ μικτών στρατηγικών σ, συμβολίζουμε με $u_i(\sigma)$ την αναμενόμενη ωφέλεια για τον παίκτη i όταν οι παίκτες επιλέγουν ανεξάρτητα τις καθαρές στρατηγικές τους σύμφωνα με το σ, δηλαδή

$$u_i(\sigma) = \sum_{s \in S} \left(\prod_{j=1}^{n} \sigma_j(s_j) \right) u_i(s), \quad \text{για κάθε } i \in N$$

Το διάνυσμα $\sigma_{-i} = (\sigma_1, \ldots, \sigma_{i-1}, \sigma_{i+1}, \ldots, \sigma_n)$ θα υποδηλώνει ένα προφίλ μικτών στρατηγικών όλων των παικτών πλην του i. Οπότε το προφίλ μικτών στρατηγικών $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_n)$, μπορεί να γραφεί

$$\sigma = (\sigma_i, \sigma_{-i}).$$

Ανάλογα, το διάνυσμα $s_{-i} = (s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_n)$ θα υποδηλώνει ένα προφίλ στρατηγικών όλων των παικτών πλην του i. Οπότε το προφίλ στρατηγικών $s = (s_1, s_2, \ldots, s_n)$, μπορεί να γραφεί

$$s = (s_i, s_{-i})$$

3.5.2 Ισορροπίες Nash

Η πιο διαδεδομένη αρχή λύσης στη Θεωρία Παιγνίων είναι αυτή της ισορροπίας Nash. Η έννοια αυτή εκφράζει μια σταθερή κατάσταση στο στρατηγικό παίγνιο, στην

οποία κάθε παίκτης κάνει τη σωστή πρόβλεψη για τη συμπεριφορά των άλλων παικτών και ενεργεί λογικά. Η ισορροπία Nash προτείνει μια στρατηγική σε κάθε παίκτη, έτσι ώστε να μην υπάρχει παίκτης που να μπορεί να βελτιώσει την ωφέλεια του αν μετακινηθεί μονομερώς σε κάποια άλλη στρατηγική. Με άλλα λόγια, η ισορροπία Nash είναι ένα προφίλ (καθαρών ή μικτών) στρατηγικών τέτοιο ώστε να μην υπάρχει κάποιος παίκτης που να έχει συμφέρον να αλλάξει τη στρατηγική του, δεδομένου ότι οι υπόλοιποι παίκτες παραμείνουν στις στρατηγικές τους. Επειδή οι υπόλοιποι παίκτες είναι επίσης λογικοί, είναι εύλογο ο κάθε παίκτης να περιμένει ότι οι αντίπαλοι του θα ακολουθήσουν όπως και ο ίδιος τις στρατηγικές που τους υποδεικνύει η ισορροπία.

Ορισμός 2 (Καθαρή Ισορροπία Nash). *Ένα προφίλ στρατηγικών* $s^* \in S$ *είναι καθαρή ισορροπία Nash για το στρατηγικό παίγνιο* $\Gamma = \{N, (S_i, u_i)_{i \in N}\}$ *αν, για κάθε παίκτη* $i \in N$*, ισχύει*

$$u_i(s_i^*, s_{-i}^*) \geq u_i(s_i, s_{-i}^*)$$

Ορισμός 3 (Μικτή Ισορροπία Nash). *Ένα προφίλ μικτών στρατηγικών* $\sigma^* \in \Delta(S)$ *είναι μικτή ισορροπία Nash για το στρατηγικό παίγνιο* $\Gamma = \{N, (S_i, u_i)_{i \in N}\}$ *αν, για κάθε παίκτη* $i \in N$*, ισχύει*

$$u_i(\sigma_i^*, \sigma_{-i}^*) \geq u_i(\sigma_i, \sigma_{-i}^*)$$

3.5.3 Ύπαρξη Ισορροπιών Nash

Από τους Ορισμούς 2 και 3 είναι προφανές ότι ο σύνολο των καθαρών ισορροπιών Nash είναι υποσύνολο του συνόλου των (μικτών) ισορροπιών Nash σε ένα στρατηγικό παίγνιο. Υπάρχουν παίγνια για τα οποία το σύνολο των καθαρών ισορροπιών Nash είναι κενό. Ωστόσο, κάθε παίγνιο με πεπερασμένο σύνολο παικτών και στο οποίο κάθε παίκτης έχει πεπερασμένο πλήθος στρατηγικών έχει τουλάχιστον μια ισορροπία Nash. Το ερώτημα που τίθεται είναι ποιες υποθέσεις εξασφαλίζουν την ύπαρξη ισορροπιών Nash. Σε αυτό το ερώτημα θα απαντήσουμε κάνοντας χρήση του Θεωρήματος Σταθερού Σημείου του Kakutani.

Θεώρημα 1 (Σταθερού Σημείου, Kakutani). *Έστω* $S \subset R^n$ *μη κενό, συμπαγές και κυρτό σύνολο. Έστω άνω ημι-συνεχής συνολο-συνάρτηση* $\gamma : S \to 2^S$*, όπου* $\gamma(s)$ *μη κενό, κυρτό και συμπαγές σύνολο. Τότε η* $\gamma(s)$ *έχει ένα τουλάχιστον σταθερό σημείο.*

Με βάση το παραπάνω Θεωρήματος Σταθερού Σημείου του Kakutani έχουμε το εξής:

Θεώρημα 2. *Έστω παίγνιο* $\Gamma = \{N, (S_i, u_i)_{i \in N}\}$, *όπου για κάθε* $i \in N$

(i) *το* S_i *είναι μη κενό, κυρτό και συμπαγές*

(ii) *η* u_i *είναι συνεχής ως προς* s *και ημι-κοίλη ως προς* s_i.

Τότε το Γ *έχει ένα τουλάχιστον σημείο ισορροπίας Nash.*

Απόδειξη: Θεωρούμε τη συνολο-συνάρτηση

$$b_i(s_{-i}) = \{ s_i \in S_i : u_i(s_i, s_{-i}) \geq u_i(s_i', s_{-i}), \forall s_i' \in S_i \}$$

την οποία ονομάζουμε συνολο-συνάρτηση βέλτιστης αντίδρασης του παίκτη i. Θεωρούμε επίσης τη συνολο-συνάρτηση

$$b(s) = b_1(s_{-1}) \times b_2(s_{-2}) \times \ldots \times b_n(s_{-n})$$

την οποία ονομάζουμε συνολο-συνάρτηση βέλτιστης αντίδρασης σε όλο το Γ. Παρατηρούμε ότι το διάνυσμα s^* αποτελεί σημείο ισορροπίας Nash αν και μόνο αν είναι σταθερό σημείο της $b(\cdot)$, δηλαδή, αν και μονό αν $s^* \in b(s^*)$. Πρέπει λοιπόν να εξακριβώσουμε αν η $b(\cdot)$ έχει σταθερό σημείο.

Παρατήρηση 1. $b_i(s_{-i}) \neq \emptyset$.

(Προκύπτει από την συνεχεία της $u_i(s_i, s_{-i})$ ως προς s_i και από το ότι το S_i είναι συμπαγές σύνολο).

Παρατήρηση 2. *Το* $b_i(s_{-i})$ *είναι κυρτό και κλειστό σύνολο.*

(Προκύπτει από την ημι-κοιλότητα και τη συνεχεία της u_i).

Παρατήρηση 3. *Το* $b_i(s_{-i})$ *είναι άνω ημισυνεχές σύνολο.*

(Προκύπτει από το ότι το σύνολο $b_i(s_{-i})$ είναι κλειστό και το S_{-i} συμπαγές).

Ως συνεπεία των Παρατηρήσεων 1-3, το σύνολο $b(s)$ είναι επίσης άνω ημισυνεχές, κυρτό και συμπαγές. Άρα έχει ένα τουλάχιστον σταθερό σημείο (από το Θεωρήματος Σταθερού Σημείου του Kakutani).

Μια ειδική περίπτωση ισορροπιών σχετίζεται με τις κυρίαρχες στρατηγικές.

Ορισμός 2. *Η καθαρή στρατηγική* s_i' *του παίκτη* i *κυριαρχεί επί της στρατηγικής* s_i'' *αν για κάθε* s_{-i}, *είναι* $u_i(s_i', s_{-i}) \geq u_i(s_i'', s_{-i})$ *και υπάρχει* $\hat{s}_{-i}$ *για το οποίο η ανισότητα ισχύει αυστηρά.*

3.6 Εφαρμογές της Θεωρίας Παιγνίων σε Δικτυακά Προβλήματα

Μια σημαντική εργασία στον χώρο της μελέτης δικτυακών προβλημάτων με εφαρμογή παιγνιοθεωρητικών αρχών είναι η [16]. Η εργασία αυτή ασχολείται με το πρόβλημα του διαμοιρασμού του διαθέσιμου εύρους ζώνης ενός εικονικού μονοπατιού σε δίκτυο ATM (Asynchronous Transfer Mode) μεταξύ διαφόρων χρηστών. Στην εν λόγω εργασία, οι χρήστες δεν είναι παθητικοί, αλλά, κατανεμημένα, είναι σε θέση να λαμβάνουν αποφάσεις σχετικές με τον διαμοιρασμό των πόρων. Ο κάθε χρήστης δεσμεύει τμήμα του εύρους ζώνης για να εγκαταστήσει ένα εικονικό μονοπάτι για τις εισερχόμενες κλήσεις του, με σκοπό την ελαχιστοποίηση της πιθανότητας μπλοκαρίσματος (blocking probability). Η αλληλεπίδραση μεταξύ των στρατηγικών των χρηστών έχει αντιμετωπιστεί ως παίγνιο. Αποδεικνύεται ότι αυτό το παίγνιο έχει μοναδικού σημείου ισορροπίας Nash. Επιπλέον, αποδεικνύεται ότι η κατάσταση ισορροπίας είναι "δίκαια" για τους εμπλεκόμενους παίκτες/χρήστες, με την έννοια ότι όποιος χρήστης έχει περισσότερη ανάγκη από τον διαμοιραζόμενο πόρο, θα λάβει περισσότερο εύρος ζώνης στην κατάσταση ισορροπίας. Στην ίδια εργασία μελετάται κι η δυναμική συμπεριφορά του παραπάνω μηχανισμού. Αποδεικνύεται ότι μπορεί να επιτευχθεί σύγκλιση στην κατάσταση ισορροπίας Nash με τη χρήση των επαναληπτικών σχημάτων Gauss-Seidel καιJacobi.

Μια επίσης ιδιαίτερα σημαντική εργασία στο χώρο της Παιγνιοθεωρητικής μελέτης δικτυακών προβλημάτων είναι η [23]. Σύμφωνα με την εργασία αυτή, οι χρήστες του δικτύου είναι "εγωιστικές" οντότητες, οι οποίες προωθούν τα προσωπικά τους συμφέροντα, ενώ ο ρόλος του σχεδιαστή του δικτύου περιορίζεται στον καθορισμό της συμπεριφοράς των στοιχείων μεταγωγής του δικτύου (network switches). Ο στόχος της συγκεκριμένης μελέτης είναι ο αποτελεσματικός σχεδιασμός της λογικής λειτουργίας, η οποία θα πρέπει να επιβληθεί στα στοιχεία του δικτύου, ώστε να επιτευχθούν αποδεκτές επιδόσεις, παρά την εγωιστική συμπεριφορά των τελικών χρηστών. Η Παιγνιοθεωρητική προσέγγιση εφαρμόζεται σε ένα απλό σύστημα ενός μεταγωγέα οι πόροι του οποίου μοιράζονται στους χρήστες. Ο κάθε χρήστης στέλνει κίνηση Poisson προς τον μεταγωγέα. Η σχεδιαστική προσπάθεια εστιάζεται στον καθορισμό πολιτικών λειτουργίας του μεταγωγέα, ώστε το σύστημα να εμφανίζει καλές επιδόσεις, παρά την εγωιστική συμπεριφορά των χρηστών. Η "καλή συμπεριφορά" συνίσταται στο να είναι η Nash ισορροπία δίκαια κι αποτελεσματική.

Επίσης, η Nash ισορροπία θα πρέπει να είναι γρήγορα κι εύκολα προσβάσιμη, μέσω απλών τεχνικών κατανεμημένης βελτιστοποίησης. Τέλος, το σύστημα θα πρέπει να προσφέρει ορισμένες εγγυήσεις επιδόσεων ακόμη κι αν λειτουργεί εκτός ισορροπίας. Πέραν από την αναφορά στις βασικές αρχές που διέπουν τις Nash ισορροπίες, στην [23] μελετάται ο μηχανισμός λειτουργίας των μεταγωγέων που είναι γνωστός ως proportional allocation (αναλογική εκχώρηση) καθώς κι ο fair share (δίκαιου μεριδίου). Στον πρώτο μηχανισμό, το μέσο μήκος της ουράς ενός χρήστη είναι ανάλογο του ρυθμού αυτού του χρήστη. Η λογική αυτή απεικονίζεται στην πολιτική First-In-First-Out (FIFO) των μεταγωγέων. Στον fair share μηχανισμό η λογική είναι αυτή ενός συστήματος χρονοδρομολόγησης με προτεραιότητες (preemptive priority scheduling). Με το μηχανισμό fair share υπάρχει πάντα μια Nash ισορροπία, η οποία είναι και κατά Pareto αποτελεσματική. Επίσης, η εν λόγω ισορροπία είναι "envy-free", καθώς κανένας παίκτης δε "ζηλεύει" την ανάθεση ισορροπίας ενός άλλου, επομένως, και δίκαια. Σε αντιδιαστολή, η λογική proportional allocation οδηγεί σε ισορροπίες Nash που δεν είναι κατά Pareto αποτελεσματικές και μπορεί να μην είναι δίκαιες. Για την ίδια πολιτική, η σύγκλιση στην κατάσταση ισορροπίας δεν είναι εγγυημένη. Όλοι οι συνήθεις αριθμητικοί αλγόριθμοι εντοπισμού μεγίστου, αν εφαρμοστούν στην περίπτωση fair share οδηγούν σε Nash ισορροπία. Επίσης, μελετάται μια παραλλαγή της κατάστασης ισορροπίας η οποία είναι γνωστή ως ισορροπία Stackelberg. Στην κατάσταση ισορροπίας Stackelberg, η ωφέλεια του ηγέτη δεν είναι χαμηλότερη από αυτή στην απλή Nash ισορροπία. Η Nash ισορροπία, η οποία επιτυγχάνεται από τον fair share αλγόριθμο είναι, επίσης, ισορροπία Stackelberg.

Μια άλλη σημαντική εργασία στον χώρο των παιγνιοθεωρητικών προσεγγίσεων δικτυακών προβλημάτων είναι η [15]. Η εργασία αυτή εστιάζει στο πρόβλημα της δρομολόγησης κι εισάγει δύο μεθοδολογίες για την συγκρότηση μη συνεργατικών δικτύων όπου οι χρήστες λαμβάνουν αποφάσεις με σκοπό τη μεγιστοποίηση της ωφέλειας τους. Οι μεθοδολογίες αναφέρονται στη φάση της διαστασιοποίησης (provisioning) του δικτύου, καθώς και στη φάση της λειτουργίας του. Κατά τη φάση διαστασιοποίησης, υπολογίζονται οι βασικές λειτουργικές παράμετροι του δικτύου, με σκοπό τη διάθεση της χωρητικότητας των συνδέσεων, ώστε το σύστημα να οδηγηθεί σε μια αποτελεσματική Nash ισορροπία. Τα κριτήρια για την αποτελεσματικότητα του συστήματος είναι η τιμή (οριακό κόστος – marginal cost), την οποία αντιλαμβάνεται ο

χρήστης, το συνολικό κόστος για τον κάθε χρήστη ή συνδυασμός των ανωτέρω. Η λύση του προβλήματος εκχώρησης χωρητικότητας δεν συμβαδίζει με τα αναμενόμενα, αφού η επαύξηση της χωρητικότητας των συνδέσεων μπορεί να οδηγήσει σε υποβάθμιση των επιδόσεων των χρηστών. Η επίδειξη αυτής της μη-αναμενόμενης συμπεριφοράς βασίζεται στο παράδοξο του Braess. Κατά την διάρκεια της φάσης λειτουργίας ένας συντονιστής ελέγχει τη δρομολόγηση τμήματος της δικτυακής ροής. Ο συντονιστής γνωρίζει την μη-συνεργατική συμπεριφορά των χρηστών του δικτύου και λαμβάνει αποφάσεις δρομολόγησης, βάσει αυτής της συμπεριφοράς, προσπαθώντας πάντα να βελτιώσει την επίδοση του συστήματος. Το σενάριο αυτό είναι ακριβώς το σενάριο ενός παιγνίου Stackelberg. Ο συντονιστής (ηγέτης) μπορεί να επιβάλλει μια κατάσταση ισορροπίας, η οποία ταυτίζεται με την βέλτιστη κατάσταση του δικτύου.

Στην εργασία [19] μελετάται η εφαρμογή των αρχών της Θεωρίας Παιγνίων σε θέματα τυχαίας πρόσβασης (random access) κι ελέγχου ισχύος. Μελετάται η συμπεριφορά "εγωιστικά" συμπεριφερόμενων χρηστών σε ένα απλοποιημένο σύστημα Aloha. Στην περίπτωση του ελέγχου ισχύος, επιτυγχάνεται, μέσω παιγνιοθεωρητικών τεχνικών, ένα βέλτιστο σημείο λειτουργίας του συστήματος, χωρίς να απαιτείται η παρέμβαση ενός εξωτερικού ελεγκτή (external controller).

Στην εργασία [9] μελετούνται αλγόριθμοι για τον επιμερισμό του κόστους που προκύπτει από μεταδόσεις πολυεκπομπής (multicast). Εξετάζονται δύο βασικοί μηχανισμοί, το οριακό κόστος κι η λύση Shapley.

4 Στοιχεία Οικονομικής Θεωρίας

4.1 Στοιχεία αγοράς, πωλητών και αγοραστών

Η λειτουργία των αγορών (markets) προσδιορίζεται από δύο βασικές δυνάμεις, τη ζήτηση (demand) και την προσφορά (supply). Η αλληλεπίδραση, μεταξύ της προσφοράς και της ζήτησης στις αγορές, για τον καθορισμό της τιμής ισορροπίας, εξετάζεται από τη Μικροοικονομική Θεωρία. Οι όροι της ζήτησης και της προσφοράς, αναφέρονται στη συμπεριφορά των ατόμων, καθώς αλληλεπιδρούν μεταξύ τους στις αγορές. Η αγορά αποτελείται από μια ομάδα αγοραστών και πωλητών ενός συγκεκριμένου αγαθού ή υπηρεσίας, και πιθανώς ενός ρυθμιστή (regulator). Οι αγοραστές προσδιορίζουν τη ζήτηση και οι πωλητές προσδιορίζουν την προσφορά. Η ανταγωνιστική αγορά είναι μια αγορά με πολλούς αγοραστές και πωλητές, την οποία δεν ελέγχει κανείς τους και στην οποία οι αγοραστές και πωλητές δρουν επιλέγοντας στα πλαίσια ενός περιορισμένου εύρους τιμών. Εναλλακτικά, ένας ρυθμιστής (δηλαδή ένας τρίτος, όπως το κράτος σε μια πραγματική Οικονομία) μπορεί να επηρεάσει τις αποφάσεις των πωλητών και αγοραστών. Συγκεκριμένα, ένας ρυθμιστής μπορεί να επιβάλλει διάφορους περιορισμούς, συνήθως στους πωλητές, έτσι ώστε να αυξήσει την κοινωνική ευημερία ή να επιβάλλει κοινωνική δικαιοσύνη.

Οι αποφάσεις των πωλητών κι των αγοραστών καθοδηγούνται, συνήθως, από διάφορα κίνητρα. Για τον αγοραστή, το κίνητρο είναι η συνάρτηση ωφέλειας (utility function), $u(y)$, που αντιπροσωπεύει το όφελος για τη χρήση μιας ποσότητας y ενός αγαθού ή μιας υπηρεσίας. Για τον πωλητή, το κίνητρο είναι η συνάρτηση κόστους (cost function), $c(y)$, που αντιπροσωπεύει το κόστος για τη παροχή μιας ποσότητας y ενός αγαθού ή μιας υπηρεσίας. Η χρέωση των αγαθών εισάγεται για να επιτευχθεί η αποτελεσματική χρήση τους. Το όφελος του αγοραστή μειώνεται λόγω της χρέωσης που υφίσταται, ενώ ο πωλητής συλλέγοντας τις πληρωμές μειώνει το κόστος παροχής του αγαθού.

Σύμφωνα με την Οικονομική ορολογία, η ικανοποίηση ενός αγοραστή, ο οποίος αγοράζει μια ποσότητα y ενός προϊόντος με τιμή p ανά μονάδα προϊόντος, ποσοτικοποιείται από το καθαρό όφελος (net benefit), δηλαδή την ωφέλεια που αποκομίζει έχοντας το αγαθό μείον τη χρέωση για να το αποκτήσει:

$$u(y) - p \cdot y$$

Επομένως, η χρέωση είναι ένας μηχανισμός που έχει τη δυνατότητα να επηρεάζει τη συμπεριφορά του αγοραστή και την κατάσταση όλης της αγοράς. Όταν ο πωλητής θέτει τιμή p για κάθε μονάδα του προϊόντος που πουλά, ο αγοραστής απαντά με ζήτηση y. Χαμηλές τιμές αυξάνουν τη ζήτηση του αγοραστή, ενώ υψηλές την περιορίζουν. Τα έσοδα του πωλητή επηρεάζονται κι από τη ζήτηση που επιδεικνύουν οι αγοραστές, αλλά κι από την τιμή του αγαθού, παράγοντες που είναι αμφίδρομα συσχετισμένοι. Η συνάρτηση κέρδους (profit function) του πωλητή, π, είναι η διαφορά μεταξύ των εσόδων που προέρχονται από την πώληση ποσότητας y του προϊόντος και του κόστους για την παροχή αυτή της ποσότητας:

$$\pi = p \cdot y - c(y)$$

Η ποσότητα του αγαθού που θα αγοράσει ο αγοραστής εάν η τιμή είναι p, συμβολίζεται με $y(p)$, και ονομάζεται συνάρτηση ζήτησης (demand function). Ενώ, η τιμή που οι αγοραστές είναι διατεθειμένοι να πληρώσουν για y ποσότητα προϊόντος, συμβολίζεται με $p(y)$, και ονομάζεται αντίστροφη συνάρτηση ζήτησης (inverse demand function).

Με δεδομένη την τιμή p, που θέτει ο πωλητής, ο αγοραστής προσπαθεί να λύσει το πρόβλημα βελτιστοποίησης:

$$\max_{y} [u(y) - p \cdot y]$$

Η εφαρμογή των συνθηκών πρώτης τάξης, δίνει την εξίσωση:

$$u'(y) = p$$

που η λύση της θα είναι η ποσότητα y του προϊόντος που μεγιστοποιεί το καθαρό όφελος του αγοραστή.

4.2 Το είδος του προϊόντος

Το προϊόν ενός κλάδου είναι ομοιογενές (homogeneous) όταν το προϊόν του κάθε πωλητή δεν διαφέρει από εκείνο των άλλων ούτε αντικειμενικά (από άποψη τεχνικών χαρακτηριστικών) ούτε υποκειμενικά (από άποψη καταναλωτικών χαρακτηριστικών). Εναλλακτικά, μπορούμε να πούμε ότι το προϊόν ενός κλάδου είναι ομοιογενές, όταν τα προϊόντα των ανταγωνιστών είναι τέλεια υποκατάστατα του προϊόντος κάθε πωλητή.

Το προϊόν ενός κλάδου είναι διαφοροποιημένο (differentiated) όταν το προϊόν του κάθε πωλητή διαφέρει από εκείνο των άλλων είτε αντικειμενικά είτε υποκειμενικά. Εναλλακτικά, μπορούμε να πούμε ότι το προϊόν ενός κλάδου είναι διαφοροποιημένο, όταν τα προϊόντα των ανταγωνιστών είναι ατελή υποκατάστατα του προϊόντος κάθε πωλητή.

Δεδομένου ότι στα αντικειμενικά χαρακτηριστικά του προϊόντος περιλαμβάνονται στοιχεία όπως η εξυπηρέτηση του αγοραστή, ο τόπος και ο χρόνος στον οποίο διατίθεται το προϊόν, είναι σπάνιο να θεωρηθεί το προϊόν ενός κλάδου ομοιογενές. Όταν το προϊόν είναι ομοιογενές, περιμένουμε την επικράτηση μιας τιμής και όταν το προϊόν είναι διαφοροποιημένο, περισσότερων τιμών.

4.3 Ο αριθμός των πωλητών και η στρατηγική αλληλεπίδραση

Τα έσοδα και το κέρδος του κάθε πωλητή, εξαρτώνται γενικά, τόσο από τις αποφάσεις του ίδιου του πωλητή όσο και από τις αποφάσεις των ανταγωνιστών του. Για παράδειγμα, τα έσοδα ενός πωλητή μπορεί να επηρεασθούν είτε γιατί αυτός μεταβάλλει την τιμή του είτε γιατί οι ανταγωνιστές του μεταβάλλουν, εκείνοι, την τιμή τους.

Αν ο αριθμός των πωλητών είναι αρκετά μεγάλος ώστε κάθε πωλητής να μην επηρεάζεται αισθητά από τις αποφάσεις των ανταγωνιστών του, τότε βρισκόμαστε σε καθεστώς τέλειου ανταγωνισμού (perfect competition). Αν ο αριθμός των πωλητών είναι αρκετά μικρός ώστε κάθε πωλητής να επηρεάζεται αισθητά από τις αποφάσεις των ανταγωνιστών του, τότε βρισκόμαστε σε καθεστώς Ολιγοπωλίου (Oligopoly). Αν ένας πωλητής διαθέτει ένα προϊόν του οποίου δεν υπάρχει κοντινό υποκατάστατο, τότε βρισκόμαστε σε καθεστώς μονοπωλίου (monopoly). Πιο συγκεκριμένα:

4.3.1 Τέλειος ανταγωνισμός

Ο τέλειος ανταγωνισμός δεν παρατηρείται συχνά, αλλά αποτελεί σημείο αφετηρίας για τη μικροοικονομική ανάλυση και σύγκριση με άλλες μορφές αγοράς. Χαρακτηριστικό του τέλειου ανταγωνισμού είναι ο μεγάλος αριθμός πωλητών και αγοραστών. Οι πωλητές έχουν τη δυνατότητα εισόδου και εξόδου από την αγορά, οποιαδήποτε στιγμή και είναι αποδέκτες τιμών. Το προϊόν που διαθέτουν είναι ομοιογενές. Επίσης υπάρχει πλήρης πληροφόρηση σχετικά με το τι συμβαίνει ανά πάσα στιγμή.

4.3.2 Μονοπώλιο

Μονοπώλιο είναι η κατάσταση κατά την οποία υπάρχει μοναδικός πωλητής ενός προϊόντος για το οποίο δεν υπάρχουν κοντινά υποκατάστατα. Η τιμή ορίζεται από τον πωλητή και δεν υπάρχει δυνατότητα η τιμή αυτή να μεταβληθεί από ανταγωνιστές γιατί απλά αυτοί δεν υπάρχουν και δεν μπορούν να εισέλθουν στην αγορά. Η κατοχή από ένα πωλητή του αποκλειστικού δικαιώματος εκμετάλλευσης ενός προϊόντος ή μιας μεθόδου παραγωγής, η κατοχή της γνώσης και της τεχνολογίας που απαιτείται για την παραγωγή ενός προϊόντος ή ο αθέμιτος ανταγωνισμός, είναι κάποιες από τις συνθήκες που οδηγούν στη δημιουργία μονοπωλίων.

Το βασικό πρόβλημα του μονοπωλητή είναι η μεγιστοποίηση του κέρδους του και μπορεί να διατυπωθεί ως:

$$\max_y \pi = \max_y [r(y) - c(y)] = \max_y [p(y)y - c(y)]$$

όπου $r(y) = p(y)y$ να είναι η συνάρτηση εσόδων του.

Με βάση τις συνθήκες πρώτης τάξης στη βέλτιστη ποσότητα, το οριακό έσοδο εξισώνεται με το οριακό κόστος:

$$r'(y) = c'(y)$$

ή

$$p(y) + p'(y)y = c'(y)$$

και χρησιμοποιώντας την ελαστικότητα της ζήτησης ως προς την τιμή, $\varepsilon = \frac{p}{y} \cdot \frac{dy}{dp}$, προκύπτει:

$$p\left(1 + \frac{1}{\varepsilon}\right) = c'(y)$$

Το πρόβλημα του μονοπωλητή μπορεί να διατυπωθεί και ως:

$$\max_{p} \pi = \max_{p} [y(p)p - c(y(p))]$$

Από τις συνθήκες πρώτης τάξης στη βέλτιστη τιμή, και την ελαστικότητα της ζήτησης ως προς την τιμή, είναι και πάλι:

$$p\left(1+\frac{1}{\varepsilon}\right) = c'(y)$$

Μια ενδιαφέρουσα περίπτωση μονοπωλίου, είναι το φυσικό μονοπώλιο (natural monopoly), και είναι η κατάσταση της αγοράς στην οποία ένας πωλητής μπορεί να εξυπηρετεί την αγορά με καλύτερο αποτέλεσμα (κέρδος ή ζημιά) απ' ότι περισσότεροι πωλητές. Φυσικό μονοπώλιο εμφανίζεται συνήθως στη παραγωγή και διανομή ενέργειας, στις τηλεπικοινωνίες, στις συγκοινωνίες κ.λ.π.

Στην χρέωση με διάκριση (price discrimination) ο μονοπωλητής διαθέτει το προϊόν του, έστω ποσότητες y_1 και y_2, σε δύο αγορές με αντίστροφες συναρτήσεις ζήτησης αντίστοιχα $p_1(y_1)$ και $p_2(y_2)$. Η κατανομή των πωλήσεων που μεγιστοποιεί το κέρδος του δίνεται από τη λύση του προβλήματος:

$$\max_{y_1, y_2} \pi = \max_{y_1, y_2} [p_1(y_1)y_1 + p_2(y_2)y_2 - c(y_1 + y_2)]$$

Από τις συνθήκες πρώτης τάξης, και τις ελαστικότητες της ζήτησης, είναι:

$$p_1\left(1-\frac{1}{\varepsilon_1}\right) = c'(y) \quad \text{και} \quad p_2\left(1-\frac{1}{\varepsilon_2}\right) = c'(y)$$

Όπως είναι φανερό από τις δύο σχέσεις, ο μονοπωλητής θα πωλεί ακριβότερα (φθηνότερα) το προϊόν του στην αγορά με την ανελαστικότερη (ελαστικότερη) ζήτηση.

Μια ειδική περίπτωση χρέωσης με διάκριση είναι όταν ο μονοπωλητής εξυπηρετεί, π.χ. δύο, γεωγραφικά διαφορετικές αγορές, τη μακρινή αγορά H και την κοντινή αγορά L, με διαφορετικό κόστος μεταφοράς ανά μονάδα προϊόντος, αντίστοιχα h και l και διαφορετικές αντίστροφες συναρτήσεις ζήτησης $p_1(y_1)$και $p_2(y_2)$.

Σε αυτή τη περίπτωση η κατανομή των πωλήσεων που μεγιστοποιεί το κέρδος του δίνεται από τη λύση του προβλήματος:

$$\max_{y_1, y_2} \pi = \max_{y_1, y_2} [p_1(y_1)y_1 + p_2(y_2)y_2 - hy_1 - ly_2 - c(y_1 + y_2)]$$

Από τις συνθήκες πρώτης τάξης, και τις ελαστικότητες της ζήτησης, είναι:

$$p_1\left(1-\frac{1}{\varepsilon_1}\right)=h+c'(y) \quad \text{και} \quad p_2\left(1-\frac{1}{\varepsilon_2}\right)=l+c'(y)$$

Όπως είναι φανερό από τις δύο σχέσεις, το συνολικό μεταφορικό κόστος δεν κατανέμεται ανάλογα με την απόσταση της αγοράς, αλλά ανάλογα με την ελαστικότητα ζήτησης της αγοράς. Μόνον αν οι καταναλωτές στις δύο γεωγραφικές αγορές έχουν ίδια ελαστικότητα ζήτησης, θα πληρώνουν το κόστος μεταφοράς που τους αναλογεί.

4.3.3 Ολιγοπώλιο

Μεταξύ του τέλειου ανταγωνισμού και του μονοπωλίου υπάρχει μια δομή της αγοράς που ονομάζεται Ολιγοπώλιο. Μια αγορά ονομάζεται Ολιγοπωλιακή όταν ένας μικρός αριθμός πωλητών ελέγχει μεγάλο μέρος της αγοράς και εμποδίζει νέους πωλητές να εισέλθουν στον κλάδο. Σε κάποιες Ολιγοπωλιακές αγορές το προϊόν είναι ομοιογενές ενώ σε κάποιες άλλες (που είναι η πλειοψηφία) διαφοροποιημένο.

Η αλληλεξάρτηση μεταξύ των πωλητών μπορεί να δημιουργήσει μια τάση για σύμπραξη μεταξύ τους. Αν καταφέρουν να συνεργαστούν και να ενεργήσουν σαν μονοπώλιο (δηλ. όταν παράγουν μικρότερη ποσότητα και την πωλούν σε τιμή πάνω από το οριακό κόστος) τότε μεγιστοποιούν τα κοινά κέρδη τους. Με τους συνασπισμούς πωλητών επιτυγχάνεται η εξαφάνιση του μεταξύ τους ανταγωνισμού και ακόμη επιτυγχάνεται η περισσότερο ορθολογική οργάνωση της παραγωγής και αποτελεσματική διοίκηση. Οι κυριότερες μορφές συνασπισμών είναι οι κοινοπραξίες, οι κερδοσκοπικές συμπράξεις, τα καρτέλ, τα τραστ κ.λ.π.

Από την άλλη πλευρά, κάθε Ολιγοπωλητής ενδιαφέρεται μόνο για το δικό του κέρδος, και αυτό τον ωθεί στο να ανταγωνίζεται με τους άλλους στην προσπάθεια να ιδιοποιηθεί ένα μεγαλύτερο μέρος από τα κέρδη του κλάδου. Μπορεί να υπάρχει ανταγωνισμός στις ποσότητες ή στις τιμές. Όσο πιο σκληρός είναι ο ανταγωνισμός, τόσο χαμηλότερα είναι τα κέρδη του κλάδου. Το κέρδος ενός πωλητή εξαρτάται από τις αποφάσεις του ίδιου αλλά και τις αποφάσεις των ανταγωνιστών του, είναι δηλαδή ένα παίγνιο.

Βασική έννοια της Οικονομικής ανάλυσης της Ολιγοπωλιακής συμπεριφοράς είναι η ισορροπία Nash. Η πωλητής στο Ολιγοπώλιο δεν θέλει να αλλάξει τη στρατηγική του μέσω της οποίας επιτυγχάνει τη μεγιστοποίηση των κερδών του. Θα αλλάξει τη στρατηγική του εάν και μόνον εάν μια διαφορετική στρατηγική αποφέρει

υψηλότερα κέρδη. Το πόσα κέρδη απολαμβάνει κάθε πωλητής εξαρτάται από την επιτυχή πρόβλεψη της συμπεριφοράς των άλλων πωλητών. Αυτή η επιτυχής πρόβλεψη της συμπεριφοράς των άλλων πωλητών γίνεται μέσω στατικών ή δυναμικών υποδειγμάτων.

Στην περίπτωση που δεν υπάρχει συνεργασία μεταξύ των πωλητών, στην αγορά επικρατούν, ανάλογα με τις υποθέσεις, διαφορετικά υποδείγματα που οδηγούν σε μια σειρά από συμπεράσματα. Η συμπαιγνία μεταξύ πωλητών είναι πιθανή σε αγορές που είτε διαρκούν για μεγάλο χρονικό διάστημα είτε για αβέβαιο χρόνο. Όσο αυξάνεται ο αριθμός των πωλητών τόσο αυξάνεται ο ανταγωνισμός.

Ένα ταυτόχρονο παίγνιο (simultaneous game) είναι μια κατάσταση στην οποία οι αποδόσεις π.χ. δύο παικτών, π_1 και π_2, είναι συνάρτηση των κινήσεων (π.χ. τιμολογήσεων) και του ενός, a_1, και το άλλου, a_2, δηλαδή $\pi_1(a_1, a_2)$ και $\pi_2(a_1, a_2)$.

Ορίζεται η ισορροπία Nash ενός ταυτόχρονου παιγνίου ως οι τιμές, a_1^*, a_2^*, των αποφάσεων που ικανοποιούν τις συνθήκες:

$$\pi_1(a_1^*, a_2^*) \geq \pi_1(a_1, a_2^*) \quad \text{και} \quad \pi_2(a_1^*, a_2^*) \geq \pi_2(a_1^*, a_2)$$

Αν οι συναρτήσεις απόδοσης είναι συνεχείς και παραγωγίσιμες ως προς τις μεταβλητές των αποφάσεων, τότε η ισορροπία Nash, a_1^*, a_2^*, ικανοποιεί τις συνθήκες:

$$\frac{d\pi_1(a_1, a_2)}{da_1} = 0 \quad \text{και} \quad \frac{d\pi_2(a_1, a_2)}{da_2} = 0$$

Με άλλα λόγια, μια ισορροπία Nash είναι ένα ζεύγος αποφάσεων από τις οποίες κανένας παίκτης δεν έχει συμφέρον να αποκλίνει μονομερώς.

Στο υπόδειγμα ανταγωνισμού Cournot, οι πωλητές μεταβάλλουν την ποσότητα πώλησης του προϊόντος. Έστω η απλή περίπτωση όπου δύο πωλητές, 1 και 2, επιλέγουν την ποσότητα ενός ομοιογενούς προϊόντος, y_1 και y_2, την οποία θα διαθέσουν στην αγορά. Η τιμή του προϊόντος προσδιορίζεται από την αντίστροφη συνάρτηση ζήτησης με βάση τη συνολική ποσότητα που διατίθεται στην αγορά, $p(y_1 + y_2)$.

Τα προβλήματα που καλούνται να λύσουν οι δύο ανταγωνιστές είναι:

$$\max_{y_1} \pi_1 = \max_{y_1} [p(y_1 + y_2)y_1 - c_1(y_1)] \quad \text{και} \quad \max_{y_2} \pi_2 = \max_{y_2} [p(y_1 + y_2)y_2 - c_2(y_2)]$$

και από τις συνθήκες πρώτης τάξης:

$$p(Y) + p'(Y)y_1 = c_1'(y_1) \quad \text{και} \quad p(Y) + p'(Y)y_2 = c_2'(y_2)$$

όπου $Y = y_1 + y_2$. Αποδεικνύεται ότι η ποσότητα Y την οποία θα διαθέσουν στην αγορά οι δύο ανταγωνιστές, είναι μικρότερη εκείνης του τέλειου ανταγωνισμού και μεγαλύτερη εκείνης του μονοπωλίου. Επίσης αποδεικνύεται ότι η τιμή του προϊόντος θα είναι μεγαλύτερη εκείνης του τέλειου ανταγωνισμού και μικρότερη εκείνης του μονοπωλίου.

Στο υπόδειγμα ανταγωνισμού Bertrand, οι δύο πωλητές, 1 και 2, επιλέγουν τις τιμές, p_1 και p_2, αντίστοιχα που θα θέσουν στην αγορά για ένα ομοιογενές προϊόν. Η συνολική ποσότητα που θα απορροφηθεί από την αγορά με συνάρτηση ζήτησης $D(p)$ είναι $D(\min[p_1, p_2])$, γιατί όλοι οι αγοραστές θα αγοράσουν από το φθηνότερο πωλητή, εφ' όσον η ποσότητα των πόρων του πωλητή αυτού μπορεί να ικανοποιήσει την αγορά. Αποδεικνύεται ότι η ισορροπία Nash στον ανταγωνισμό Bertrand είναι η τιμή και η ποσότητα που θα ίσχυαν σε καθεστώς τέλειου ανταγωνισμού.

Στην περίπτωση του υποδείγματος Cournot, και οι δύο πωλητές παίρνουν αποφάσεις ταυτόχρονα. Εάν υποθέσουμε ότι ένας από τους δύο πωλητές έχει κάποιο πλεονέκτημα (μεγέθους ή οποιοδήποτε άλλο) και λαμβάνει τις αποφάσεις του πρώτος, τότε έχουμε έναν ηγέτη (leader). Οι αποφάσεις μπορούν να αφορούν τις ποσότητες ή τις τιμές των προϊόντων. Το υπόδειγμα το οποίο περιγράφει την περίπτωση αυτή ονομάζεται υπόδειγμα Stackelberg. Ο ηγέτης παίρνει σαν δεδομένη την συνάρτηση αντίδρασης του άλλου πωλητή, που ονομάζεται (follower), όταν λύνει το πρόβλημα μεγιστοποίησης των κερδών του. Ας υποθέσουμε ότι ο πωλητής 1 είναι ο ηγέτης και ο πωλητής 2, ο ακόλουθος. Οι πωλητές αποφασίζουν για τις ποσότητες των προϊόντων που θα διαθέσουν. Το πρόβλημα που πρέπει να λύσει ο πωλητής 2, καθώς γνωρίζει την ποσότητα y_1 του προϊόντος που διαθέτει ο πωλητής 1, είναι:

$$\max_{y_2} \pi_2 = \max_{y_2} [p(y_1 + y_2)y_2 - c_2(y_2)]$$

Η συνθήκη πρώτης τάξης, για το πρόβλημα αυτό, είναι όπως ακριβώς στο υπόδειγμα Cournot, που περιγράφηκε παραπάνω :

$$p(Y) + p'(Y)y_2 = c_2'(y_2)$$

όπου $Y = y_1 + y_2$. Μπορούμε να χρησιμοποιήσουμε αυτή την εξίσωση για να έχουμε την συνάρτηση αντίδρασης, $f_2(y_1)$, του πωλητή 2. Πηγαίνοντας πίσω στην αρχή της διαδικασίας, ο πωλητής 1 θέλει να επιλέξει την ποσότητα y_1 που θα διαθέσει,

προβλέποντας την αντίδραση του πωλητή 2. Άρα το πρόβλημα που αντιμετωπίζει ο πωλητής 1, είναι:

$$\max_{y_1} \pi_1 = \max_{y_1} \left[p\left(y_1 + f_2(y_1) \right) y_1 - c_1(y_1) \right]$$

Αυτό οδηγεί σε συνθήκη πρώτης τάξης, της μορφής:

$$p(Y) + p'(Y)(1 + f_2'(y_1))y_1 = c_1'(y_1)$$

Οι παραπάνω εξισώσεις, που προκύπτουν από τις συνθήκες πρώτης τάξης για τα προβλήματα των δύο πωλητών, είναι αρκετές για τον καθορισμό των βέλτιστων ποσοτήτων y_1 και y_2. Ανάλογη ανάλυση γίνεται όταν οι πωλητές αποφασίζουν για τιμές που θα θέσουν για το προϊόν.

Η εκτενής εφαρμογή της Θεωρίας Παιγνίων με τη βοήθεια των παραπάνω Ολιγοπωλιακών υποδειγμάτων έχει συμβάλει σημαντικά στην καλύτερη κατανόηση και εκτίμηση της αξιοπιστίας της εφαρμογής των στρατηγικών καθώς και στον προσδιορισμό των αποτελεσμάτων διαφόρων στρατηγικών που επιλέγονται ταυτόχρονα από τους Ολιγοπωλητές.

5 Προτεινόμενη Λύση

5.1 Γενικά

Σε αυτήν την εργασία, προτείνεται ένα Οικονομικό σχήμα για τη διαχείριση της χωρητικότητας, για εφαρμογή σε ιεραρχίες caching. Στο περιβάλλον που εξετάζεται, το οικονομικό αγαθό είναι ο αποθηκευτικός χώρος. Οι L2 cache (πωλητές) παρέχουν χώρο από το δίσκο τους, ενώ οι L1 cache (αγοραστές) πληρώνουν για να αποκτήσουν μέρος του εν λόγω χώρου. Η ανάγκη για την εισαγωγή ενός Οικονομικού σχήματος, προκύπτει καθώς, όταν η χωρητικότητα μιας L2 cache διαμοιράζεται μεταξύ κάποιων L1 cache, είναι δυνατόν, αν μια L1 cache είναι ιδιαίτερα «επιθετική», σε σχέση με τις άλλες, να κυριαρχήσει επί της διαμοιραζόμενης χωρητικότητας, με αποτέλεσμα, να επωφεληθεί εις βάρος των άλλων L1 cache. Προβληματικός διαμοιρασμός ενός κοινού πόρου μπορεί να παρατηρηθεί και σε άλλα περιβάλλοντα, όπως δίκτυα P2P, και CDN.

Η εισαγωγή ενός Οικονομικού σχήματος εξασφαλίζει την εφαρμογή σαφών κανόνων που θα διέπουν τις αλληλεπιδράσεις των οντοτήτων, καθώς και τιμές που θα οδηγούν σε διαφορετικά επίπεδα ζήτησης χωρητικότητας. Ορίζεται μια Ολιγοπωλιακή αγορά και η εκχώρηση της χωρητικότητας γίνεται βάσει χρέωσης χωρίς διάκριση. Οι L2 cache ανταγωνίζονται μεταξύ τους, με σκοπό να προσελκύσουν περισσότερους αγοραστές και να μεγιστοποιήσουν τα έσοδα τους.

Έστω $I=\{1,\ldots,n\}$ το σύνολο των L1 cache και $J=\{1,\ldots,k\}$ το σύνολο των L2 cache. Η L1 cache i μπορεί να αγοράσει από τις L2 cache το διάνυσμα χώρου

$$\boldsymbol{x}^{(i)}=\left(x_1^{(i)},\ldots,x_k^{(i)}\right) \tag{5.1}$$

όπου $x_j^{(i)}\geq 0$, $j=1,\ldots,k$ είναι η χωρητικότητα που θα έχει η L1 cache i στο δίσκο της L2 cache j. Αν p_j, $j=1,\ldots,k$ είναι η τιμή χρέωσης που θέτει η L2 cache j για κάθε μονάδα αποθηκευτικού χώρου που διαθέτει στις L1 cache, τότε ορίζουμε

$$p\left(\boldsymbol{x}^{(i)}\right)=\boldsymbol{p}^{\mathrm{T}}\boldsymbol{x}^{(i)}=\sum_{j=1}^{k}p_j x_j^{(i)} \tag{5.2}$$

να είναι το χρηματικό ποσό που πρέπει να πληρώσει η L1 cache i για την αγορά του διανύσματος (5.1), με δεδομένο το διάνυσμα τιμών $\boldsymbol{p} = (p_1, \ldots, p_k)$ που θέτουν οι L2 cache. Εδώ έγινε η υπόθεση της γραμμικής χρέωσης, δηλαδή ο αγοραστής χρεώνεται αναλογικά με την ποσότητα του χώρου που αποκτά και δεν εφαρμόζονται πάγιες χρεώσεις.

Έστω $u_i\left(\boldsymbol{x}^{(i)}\right)$, η συνάρτηση ωφέλειας (utility function) της L1 cache i. Γενικά, η συνάρτηση $u_i(\cdot)$ ποσοτικοποιεί την ικανοποίηση του αγοραστή i λόγω της κατοχής ενός συγκεκριμένου ποσού του αγαθού. Ορίζουμε ως καθαρό όφελος (net benefit) του αγοραστή i, την συνάρτηση

$$b_i\left(\boldsymbol{x}^{(i)}\right) = u_i\left(\boldsymbol{x}^{(i)}\right) - \boldsymbol{p}^T \boldsymbol{x}^{(i)} \tag{5.3}$$

Ο αγοραστής (L1 cache) i, προσπαθεί να λύσει το παρακάτω πρόβλημα βελτιστοποίησης:

$$\text{ΑΓΟΡΑΣΤΗΣ}_i: \begin{cases} \underset{\boldsymbol{x}^{(i)}}{\text{maximize}} \left[u_i\left(\boldsymbol{x}^{(i)}\right) - \boldsymbol{p}^{\mathrm{T}} \boldsymbol{x}^{(i)}\right] \\ \text{μ.τ.π.} \\ x_j^{(i)} \geq 0 \end{cases} \tag{5.4}$$

Η λύση αυτού του προβλήματος, έστω $\boldsymbol{x}^{(i)}(\boldsymbol{p})$, ονομάζεται ζήτηση (demand function) του αγοραστή i με δεδομένο το διάνυσμα τιμών $\boldsymbol{p}$, και είναι το διάνυσμα χώρου, το οποίο μεγιστοποιεί το καθαρό όφελος του με τον περιορισμό ότι κάθε μια από τις συνιστώσες του διανύσματος αυτού πρέπει να είναι ποσότητες θετικές ή μηδέν.

Η L2 cache j επιλέγει μια τιμή χρέωσης, έστω p_j, για κάθε μονάδα αποθηκευτικού χώρου που διαθέτει στις L1 cache. Η χωρητικότητα της L2 cache j είναι $C_{2,j}$ και υποθέτουμε ότι διαχωρίζεται σε διαφορετικά (απομονωμένα) διαμερίσματα που εκχωρούνται αποκλειστικά σε ξεχωριστές L1 cache, δηλαδή ο χώρος που παίρνει μια L1 cache δε μπορεί να χρησιμοποιηθεί από κάποια άλλη L1 cache.

Έστω $y_j(\boldsymbol{p}) = \sum_{i=1}^{n} x_j^{(i)}$ η συνολική ζήτηση χώρου από τον δίσκο της L2 cache j. Η συνάρτηση κέρδους της L2 cache j είναι:

$$\pi_j(\boldsymbol{p}) = p_j \cdot y_j(\boldsymbol{p}) = p_j \cdot \sum_{i=1}^{n} x_j^{(i)} \tag{5.5}$$

Γίνεται η υπόθεση ότι οι L2 cache δεν έχουν λειτουργικά έξοδα σε σχέση με το χώρο που εκχωρούν. Τα κόστη που έχουν να αντιμετωπίσουν είναι εφάπαξ κόστη (sunk costs) και αφορούν στην εγκατάσταση της υποδομής τους.

Το ανταγωνιστικό, ως προς τις τιμές, Οικονομικό μοντέλο που χρησιμοποιείται (και περιγράφηκε παραπάνω), μπορεί να γραφεί ως παίγνιο στρατηγικής μορφής στο οποίο:

- Υπάρχουν k παίκτες, οι L2 cache.
- Το σύνολο στρατηγικών S_j του παίκτη j, είναι το σύνολο των διαφορετικών τιμών χρέωσης p_j, που η L2 cache j μπορεί να επιλέξει, για κάθε μονάδα αποθηκευτικού χώρου που διαθέτει στις L1 cache. Δηλαδή $p_j \in S_j$ για κάθε $j = 1,\ldots,k$.
- Η συνάρτηση απόδοσης του παίκτη j, είναι η συνάρτηση κέρδους $\pi_j(\boldsymbol{p})$ της L2 cache j και δίνει το κέρδος της για κάθε δυνατό συνδυασμό τιμών (προφίλ στρατηγικών) $\boldsymbol{p} = (p_1,\ldots,p_k) \in S = S_1 \times \ldots \times S_k$.

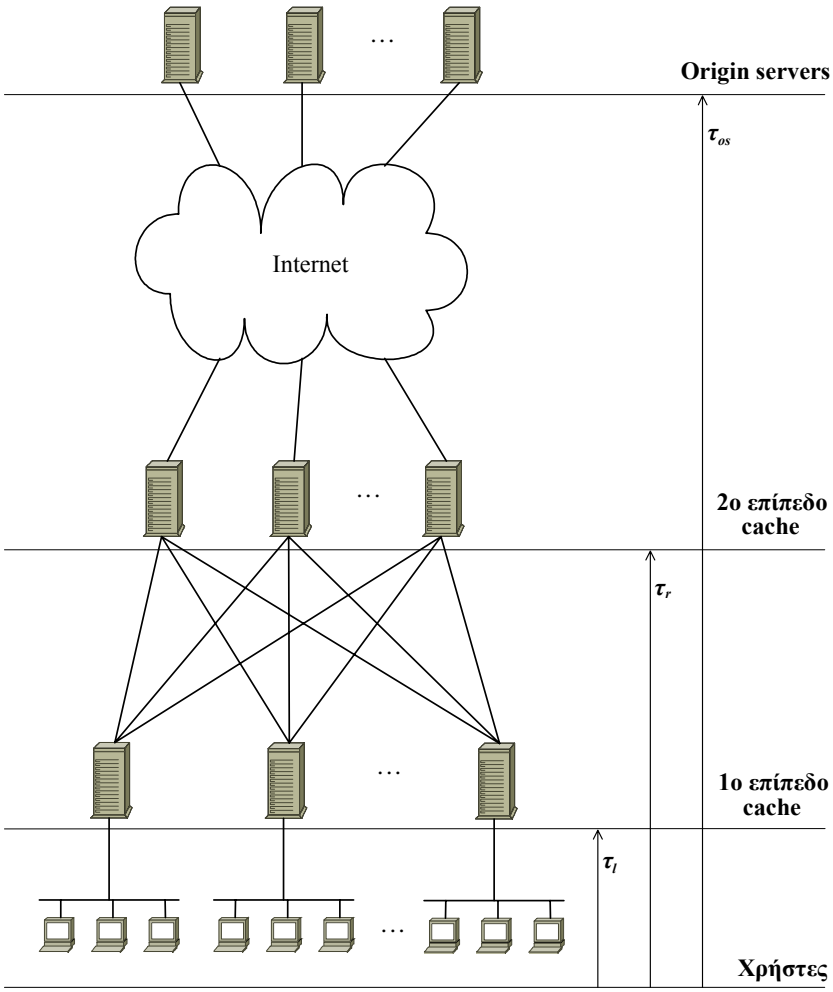
Ο πωλητής (L2 cache) j προσπαθεί να λύσει το παρακάτω πρόβλημα βελτιστοποίησης:

$$\text{ΠΩΛΗΤΗΣ}_j : \begin{cases} \underset{p_j}{\text{maximize}} \left[p_j \cdot \sum_{i=1}^{n} x_j^{(i)} \right] \\ \text{μ.τ.π.} \\ \sum_{i=1}^{n} x_j^{(i)} \le C_{2,j} \end{cases} \tag{5.6}$$

Δηλαδή ο πωλητής j προσπαθεί να βρει την τιμή χρέωσης p_j, η οποία προκαλεί συνολική ζήτηση y_j, και μεγιστοποιεί το κέρδος του με τον περιορισμό ότι η συνολική ζήτηση δεν πρέπει να ξεπερνά την χωρητικότητά του.

5.2 Σύστημα

Έστω O το σύνολο των αντικειμένων που υπάρχουν στο web και O_i το σύνολο των αντικειμένων στα οποία αναφέρονται οι χρήστες που συνδέονται με την L1 cache i. Εδώ υποθέτουμε ότι μια L1 cache εξυπηρετεί χρήστες που ενδιαφέρονται για ένα υποσύνολο του web, δηλαδή εξυπηρετεί μια ομάδα χρηστών που έχουν κάποιες ομοιότητες στα ενδιαφέροντα τους. Είναι $O_i \subseteq O$ για κάθε $i = 1,...,n$.



Σχήμα 5.1 Η αρχιτεκτονική Web caching του συστήματος

Η αίτηση του χρήστη για κάποιο αντικείμενο κατευθύνεται στην L1 cache με την οποία αυτός συνδέεται. Αν το αντικείμενο υπάρχει αποθηκευμένο εκεί, θεωρείται ότι επιτυγχάνεται μια τοπική ευστοχία (local hit) και ο χρήστης εξυπηρετείται σε χρόνο $\tau_\ell^{(\iota)}$. Αν το αντικείμενο δεν υπάρχει στην L1 cache, και υπάρχει στην L2 cache m, η

αίτηση προωθείται σε αυτήν, και τότε θεωρείται ότι επιτυγχάνεται μια απομακρυσμένη ευστοχία (remote hit), με αντίστοιχο χρόνο εξυπηρέτησης $\tau_{r_m}^{(i)}$. Αν το αντικείμενο δε υπάρχει ούτε σε κάποια από τις L2 cache, τότε ζητείται από τον πηγαίο διακομιστή, και ο χρόνος εξυπηρέτησης του χρήστη είναι $\tau_{os}^{(i)}$. Υποθέτουμε ότι οι χρόνοι $\tau_{\ell}^{(i)}$, $\tau_{r_m}^{(i)}$ και $\tau_{os}^{(i)}$ είναι σταθερές για κάθε $i=1,\ldots,n$, $m=1,\ldots,k$, και ισχύει:

$$\tau_{\ell}^{(i)} < \tau_{r_1}^{(i)} < \ldots < \tau_{r_k}^{(i)} < \tau_{os}^{(i)}$$

Στην εργασία αυτή, γίνεται η υπόθεση ότι η L1 cache i γνωρίζει ακριβώς τα αντικείμενα (URL) που είναι αποθηκευμένα στις k L2 cache. Τέτοια πληροφορία μπορεί να γίνει εύκολα διαθέσιμη, αν η L1 cache διατηρεί μια κατάλληλη δομή δεδομένων. Με τη δομή αυτή, η L1 cache i, όταν θα λαμβάνει μια HTTP αίτηση από ένα χρήστη, θα γνωρίζει αν το ζητούμενο αντικείμενο βρίσκεται αποθηκευμένο τοπικά, απομακρυσμένα (σε κάποια από τις L2 cache), ή θα πρέπει να μεταφερθεί από τον πηγαίο διακομιστή.

Η L1 cache i έχει χωρητικότητα τέτοια ώστε να έχει αποθηκευμένα σε αυτήν τα C_i πιο δημοφιλή αντικείμενα του συνόλου O_i. Τα υπόλοιπα αντικείμενα του συνόλου, αν θεωρηθούν διατεταγμένα κατά φθίνουσα σειρά δημοτικότητας, είναι αποθηκευμένα στις k L2 cache, ξεκινώντας από την πρώτη να περιέχει τα πιο δημοφιλή.

Έχει αποδειχθεί ότι η δημοτικότητα των web αντικειμένων ακολουθεί την κατανομή Zipf. Σύμφωνα με αυτήν, αν βάλουμε σε μια σειρά ένα σύνολο αντικειμένων (π.χ. τα αντικείμενα ενός web site) με βάση την δημοτικότητα τους (το αντικείμενο 1 είναι το πιο δημοφιλές), η πιθανότητα το h-ιοστό αντικείμενο να ζητηθεί είναι:

$$\text{prob}\{\text{να ζητηθεί το } h-\text{ιοστό αντικείμενο}\} = \frac{K}{h^a}$$

όπου $a \in (0,1)$ είναι ο παράγοντας της κατανομής Zipf, η τιμή του οποίου εξαρτάται από την τρέχουσα εφαρμογή και K είναι μια σταθερά κανονικοποίησης τέτοια ώστε:

$$\sum_{h=1}^{|U|} \frac{K}{h^a} = 1$$

όπου U είναι το σύνολο που περιέχει τα εξεταζόμενα αντικείμενα. Για την L1 cache i, στο σύστημα που έχουμε, το σύνολο U ταυτίζεται με το σύνολο των αντικειμένων στα οποία αναφέρονται οι χρήστες που συνδέονται με αυτήν. Δηλαδή $U \equiv O_i$.

Στις περισσότερες περιπτώσεις, το μέγεθος του αιτούμενου αντικειμένου είναι τάξεις μεγέθους μικρότερο από το μέγεθος της L1 cache. Οπότε μπορούμε να κάνουμε μια προσέγγιση, θεωρώντας τα περιεχόμενα της cache, όχι διακριτά αντικείμενα αλλά ένα συνεχές (continuum) δεδομένων. Επεκτείνουμε, λοιπόν, την $\frac{K}{h^a}$ και τα ακέραια σημεία h, στην συνεχή συνάρτηση $\frac{K}{x^a}$ με $x \in [0, +\infty)$.

Όταν ένας χρήστης της L1 cache i αιτηθεί ένα αντικείμενο, τότε αυτό βρίσκεται στην L1 cache με πιθανότητα ευστοχίας από την L1 cache i (local hit probability):

$$p_{h,\ell}^{(i)} = \sum_{h=1}^{C_i} \frac{K_i}{h^{a_i}}$$

Εφαρμόζοντας την παραπάνω προσέγγιση, η πιθανότητα παίρνει την μορφή:

$$p_{h,\ell}^{(i)} = \int_0^{C_i} \frac{K_i}{x^{a_i}} dx$$

και επειδή $\int_0^{O_i} \frac{K_i}{x^{a_i}} dx = 1$, θα είναι:

$$p_{h,\ell}^{(i)} = \left(\frac{C_i}{O_i} \right)^{1-a_i} \tag{5.7}$$

Θεωρούμε ότι η L1 cache i έχει αποθηκευμένο στις L2 cache το διάνυσμα χώρου $\boldsymbol{x}^{(i)} = \left(x_1^{(i)}, \ldots, x_k^{(i)} \right)$, όπου $x_j^{(i)}$ είναι το ποσό του χώρου που έχει η L1 cache i στο δίσκο της L2 cache j, για κάθε $i = 1, \ldots, n$ και $j = 1, \ldots, k$.

Η πιθανότητα, το αντικείμενο που αιτείται ο χρήστης, να βρίσκεται στην L2 cache m, δηλαδή η πιθανότητα ευστοχίας από την L2 cache m (m-remote hit probability) είναι:

$$p_{h,r_m}^{(i)} = \int_0^{C_i + \sum_{j=1}^{m} x_j^{(i)}} \frac{K_i}{x^{a_i}} dx - \int_0^{C_i + \sum_{j=1}^{m-1} x_j^{(i)}} \frac{K_i}{x^{a_i}} dx$$

η οποία μετά από πράξεις, παίρνει την μορφή:

$$p_{h,r_m}^{(i)} = \frac{\left(C_i + \sum_{j=1}^{m} x_j^{(i)}\right)^{1-a_i} - \left(C_i + \sum_{j=1}^{m-1} x_j^{(i)}\right)^{1-a_i}}{O_i^{1-a_i}} \tag{5.8}$$

Γενικά η πιθανότητα, το αντικείμενο που αιτείται ο χρήστης, να βρίσκεται στο δεύτερο επίπεδο των cache, είναι:

$$p_{h,r}^{(i)} = \int_0^{C_i + \sum_{j=1}^{k} x_j^{(i)}} \frac{K_i}{x^{a_i}} dx - \int_0^{C_i} \frac{K_i}{x^{a_i}} dx$$

ή ισοδύναμα:

$$p_{h,r}^{(i)} = \frac{\left(C_i + \sum_{j=1}^{k} x_j^{(i)}\right)^{1-a_i} - \left(C_i\right)^{1-a_i}}{O_i^{1-a_i}} \tag{5.9}$$

και ονομάζεται πιθανότητα ευστοχίας από το δεύτερο επίπεδο των cache (remote hit probability).

Ο αναμενόμενος χρόνος εξυπηρέτησης, δηλαδή ο χρόνος που αντιλαμβάνεται ο χρήστης από την υποβολή μιας αίτησης μέχρι την παρουσίαση σε αυτόν του αιτούμενου αντικειμένου, είναι:

$$E\{\tau_i\} = p_{h,\ell}^{(i)} \cdot \tau_\ell^{(i)} + \left(1 - p_{h,\ell}^{(i)}\right) \sum_{m=1}^{k} \left(p_{h,r_m}^{(i)} \cdot \tau_{r_m}^{(i)}\right) + \left(1 - p_{h,\ell}^{(i)}\right)\left(1 - p_{h,r}^{(i)}\right)\tau_{os}^{(i)}$$

και από τις (5.7), (5.8) και (5.9) προκύπτει:

$$\begin{aligned} E\{\tau_i\} = & \left(\frac{C_i}{O_i}\right)^{1-a_i} \tau_\ell^{(i)} \\ & + \left(1 - \left(\frac{C_i}{O_i}\right)^{1-a_i}\right) \sum_{m=1}^{k} \left\{ \frac{\left(C_i + \sum_{j=1}^{m} x_j^{(i)}\right)^{1-a_i} - \left(C_i + \sum_{j=1}^{m-1} x_j^{(i)}\right)^{1-a_i}}{O_i^{1-a_i}} \tau_{r_m}^{(i)} \right\} \\ & + \left(1 - \left(\frac{C_i}{O_i}\right)^{1-a_i}\right) \left(1 - \frac{\left(C_i + \sum_{j=1}^{k} x_j^{(i)}\right)^{1-a_i} - C_i^{1-a_i}}{O_i^{1-a_i}} \right) \tau_{os}^{(i)} \end{aligned} \tag{5.10}$$

5.3 Αγοραστής (L1 cache)

Καθώς ο αντικειμενικός σκοπός της L1 cache i, είναι να ελαχιστοποιήσει το χρόνο εξυπηρέτησης των χρηστών που συνδέονται με αυτήν, μια "λογική" επιλογή ως συνάρτηση ωφέλειάς της, είναι η

$$u_i\left(\boldsymbol{x}^{(i)}\right)=-E\{\tau_i\}$$

Ορίζοντας τις ποσότητες

$$\hat{\tau}_j^{(i)}=\begin{cases}\tau_\ell^{(i)} & , j=0\\ \tau_{r_j}^{(i)} & ,1\le j\le k\\ \tau_{os}^{(i)} & , j=k+1\end{cases}$$

και

$$v_j^{(i)}=\left(\frac{C_i+\sum_{m=1}^{j}x_m^{(i)}}{O_i}\right)^{1-a_i}\quad,\quad j=0,1,\ldots,k$$

έχουμε τη παρακάτω Πρόταση.

Πρόταση 5.1. *Ο αναμενόμενος χρόνος εξυπηρέτησης* $E\{\tau_i\}$ *που δίνεται από την (5.10), παίρνει τη μορφή:*

$$E\{\tau_i\}=v_0^{(i)}\hat{\tau}_0^{(i)}+\left(1-v_0^{(i)}\right)\hat{\tau}_{k+1}^{(i)}+v_0^{(i)}\left(1-v_0^{(i)}\right)\left(\hat{\tau}_{k+1}^{(i)}-\hat{\tau}_1^{(i)}\right)-\left(1-v_0^{(i)}\right)\sum_{m=1}^{k}\left(v_m^{(i)}\left(\hat{\tau}_{m+1}^{(i)}-\hat{\tau}_m^{(i)}\right)\right)$$

Απόδειξη: Η απόδειξη βρίσκεται στο Παράρτημα. ■

Με βάση τη Πρόταση 5.1, η συνάρτηση ωφέλειας $u_i\left(\boldsymbol{x}^{(i)}\right)=-E\{\tau_i\}$, γράφεται

$$u_i\left(\boldsymbol{x}^{(i)}\right)=\left(1-v_0^{(i)}\right)\sum_{m=1}^{k}\left(v_m^{(i)}\left(\hat{\tau}_{m+1}^{(i)}-\hat{\tau}_m^{(i)}\right)\right)-v_0^{(i)}\hat{\tau}_0^{(i)}-\left(1-v_0^{(i)}\right)\hat{\tau}_{k+1}^{(i)}-v_0^{(i)}\left(1-v_0^{(i)}\right)\left(\hat{\tau}_{k+1}^{(i)}-\hat{\tau}_1^{(i)}\right)$$

Θέτοντας

$$\hat{u}_i\left(\boldsymbol{x}^{(i)}\right)=\left(1-v_0^{(i)}\right)\sum_{m=1}^{k}\left(v_m^{(i)}\,\Delta\hat{\tau}_m^{(i)}\right)=\left(1-v_0^{(i)}\right)\sum_{m=1}^{k}\left(\Delta\hat{\tau}_m^{(i)}\left(\frac{C_i+\sum_{j=1}^{m}x_j^{(i)}}{O_i}\right)^{1-a_i}\right)\qquad(5.11)$$

με

$$\Delta\hat{\tau}_m^{(i)} = \hat{\tau}_{m+1}^{(i)} - \hat{\tau}_m^{(i)} > 0, \quad m = 1,\ldots,k$$

και ορίζοντας μια σταθερά B_i, $i = 1,...,n$ να δίνεται από τη σχέση

$$B_i = v_0^{(i)}\hat{\tau}_0^{(i)} + \left(1 - v_0^{(i)}\right)\hat{\tau}_{k+1}^{(i)} + v_0^{(i)}\left(1 - v_0^{(i)}\right)\left(\hat{\tau}_{k+1}^{(i)} - \hat{\tau}_1^{(i)}\right) > 0$$

η συνάρτηση ωφέλειας $u_i\left(\boldsymbol{x}^{(i)}\right)$ με $i = 1,...,n$, παίρνει τη μορφή

$$u_i\left(\boldsymbol{x}^{(i)}\right) = \hat{u}_i\left(\boldsymbol{x}^{(i)}\right) - B_i \tag{5.12}$$

Πρόταση 5.2. *Η πραγματική συνάρτηση* $g(x) = x^k$ *με* $x > 0$ *και* $0 < k < 1$, *είναι γνησίως κοίλη.*

Απόδειξη: Η απόδειξη βρίσκεται στο Παράρτημα. ■

Ορισμός 5.1 (Ανισότητα Jensen). *Η συνάρτηση* $f : \boldsymbol{R}^n \to \boldsymbol{R}$ *είναι γνησίως κοίλη εάν για κάθε* $\boldsymbol{x}, \boldsymbol{y}$ *του πεδίου ορισμού της, με* $\boldsymbol{x} \neq \boldsymbol{y}$, *και για κάθε* $0 < h < 1$, *ισχύει*

$$f\left((1-h)\boldsymbol{x} + h\boldsymbol{y}\right) = (1-h)f(\boldsymbol{x}) + hf(\boldsymbol{y})$$

Ορίζοντας τη συνάρτηση

$$g_i(y) = y^{1-a_i}, \; i = 1,...,n$$

η (5.11) γράφεται

$$\hat{u}_i(\boldsymbol{x}) = \left(1 - v_0^{(i)}\right)\sum_{m=1}^{k}\left(\Delta\hat{\tau}_m^{(i)} g_i\left(\frac{C_i + \sum_{j=1}^{m} x_j}{O_i}\right)\right)$$

Πρόταση 5.3. *Η συνάρτηση* $\hat{u}_i(\boldsymbol{x}) = \left(1 - v_0^{(i)}\right)\sum_{m=1}^{k}\left(\Delta\hat{\tau}_m^{(i)} g_i\left(\frac{C_i + \sum_{j=1}^{m} x_j}{O_i}\right)\right)$, $i = 1,...,n$, *είναι γνησίως κοίλη.*

Απόδειξη: Η απόδειξη προκύπτει εύκολα από τον Ορισμό 5.1 και την Πρόταση 5.2, και βρίσκεται στο Παράρτημα. ■

Θεώρημα 5.1. *Η συνάρτηση καθαρού οφέλους της L1 cache i (σχέση 5.3)*

$$b_i\left(\boldsymbol{x}^{(i)}\right)=u_i\left(\boldsymbol{x}^{(i)}\right)-\boldsymbol{p}^T\boldsymbol{x}^{(i)}=\hat{u}_i\left(\boldsymbol{x}^{(i)}\right)-B_i-\boldsymbol{p}^T\boldsymbol{x}^{(i)},\quad i=1,\ldots,n$$

είναι γνησίως κοίλη.

Απόδειξη: Η απόδειξη προκύπτει εύκολα, καθώς η συνάρτηση καθαρού οφέλους $b_i\left(\boldsymbol{x}^{(i)}\right)$ είναι το άθροισμα μιας γνησίως κοίλης, μιας σταθερής και μιας γραμμικής συνάρτησης, και βρίσκεται στο Παράρτημα. ■

Ας επανέλθουμε τώρα, στο πρόβλημα βελτιστοποίησης που αντιμετωπίζει η L1 cache i, και περιγράφει η σχέση (5.4). Όπως έχει αναφερθεί παραπάνω, ο αγοραστής i, ψάχνει να βρει το διάνυσμα χώρου

$$\boldsymbol{x}^{(i)}(\boldsymbol{p})=\arg\max_{\boldsymbol{x}^{(i)}} b_i\left(\boldsymbol{x}^{(i)}\right)=\arg\max_{\boldsymbol{x}^{(i)}}\left[u_i\left(\boldsymbol{x}^{(i)}\right)-\boldsymbol{p}^{\mathrm{T}}\boldsymbol{x}^{(i)}\right]$$

Από τις συνθήκες πρώτης τάξης $\nabla b_i\left(\boldsymbol{x}^{(i)}\right)=\mathbf{0}^{\mathrm{T}}$, είναι

$$\left(p_1,\,p_2,\ldots,\,p_k\right)=\left(\frac{\partial u_i\left(\boldsymbol{x}^{(i)}\right)}{\partial x_1^{(i)}},\frac{\partial u_i\left(\boldsymbol{x}^{(i)}\right)}{\partial x_2^{(i)}},\ldots,\frac{\partial u_i\left(\boldsymbol{x}^{(i)}\right)}{\partial x_k^{(i)}}\right)$$

και καθώς $\dfrac{\partial u_i\left(\boldsymbol{x}^{(i)}\right)}{\partial x_j^{(i)}}=\dfrac{\partial \hat{u}_i\left(\boldsymbol{x}^{(i)}\right)}{\partial x_j^{(i)}}$ για κάθε $j=1,\ldots,k$, από την (5.11) προκύπτει ότι

$$p_s=\left(1-v_0^{(i)}\right)\left(\frac{1}{O_i}\right)^{1-a_i}\left(1-a_i\right)\sum_{m=s}^{k}\left(\Delta\hat{\tau}_m^{(i)}\left(C_i+\sum_{j=1}^{m}x_j^{(i)}\right)^{-a_i}\right),\quad s=1,\ldots,k$$

και στη συνέχεια, αποσπώντας, από το εξωτερικό άθροισμα, τον όρο με $m=s$, είναι

$$p_s=\left(1-v_0^{(i)}\right)\left(\frac{1}{O_i}\right)^{1-a_i}\left(1-a_i\right)\sum_{m=s}^{k}\left(\Delta\hat{\tau}_m^{(i)}\left(C_i+\sum_{j=1}^{m}x_j^{(i)}\right)^{-a_i}\right)$$

$$=\left(1-v_0^{(i)}\right)\left(\frac{1}{O_i}\right)^{1-a_i}\left(1-a_i\right)\left[\sum_{m=s+1}^{k}\left(\Delta\hat{\tau}_m^{(i)}\left(C_i+\sum_{j=1}^{m}x_j^{(i)}\right)^{-a_i}\right)+\Delta\hat{\tau}_s^{(i)}\left(C_i+\sum_{j=1}^{s}x_j^{(i)}\right)^{-a_i}\right]$$

$$=p_{s+1}+\left(1-v_0^{(i)}\right)\left(\frac{1}{O_i}\right)^{1-a_i}\left(1-a_i\right)\Delta\hat{\tau}_s^{(i)}\left(C_i+\sum_{j=1}^{s}x_j^{(i)}\right)^{-a_i}$$

όπου $p_{k+1}=0$.

Από την θετική διαφορά $p_s - p_{s+1} = \left(1-v_0^{(i)}\right)\left(\frac{1}{O_i}\right)^{1-a_i}(1-a_i)\Delta\hat{\tau}_s^{(i)}\left(C_i+\sum_{j=1}^{s}x_j^{(i)}\right)^{-a_i}$,

με απλές πράξεις, παίρνουμε

$$\sum_{j=1}^{s}x_j^{(i)} = A_s^{(i)}\left(p_s - p_{s+1}\right)^{-\frac{1}{a_i}} - C_i \qquad (5.13)$$

όπου

$$A_s^{(i)} = \left(\left(1-v_0^{(i)}\right)\left(\frac{1}{O_i}\right)^{1-a_i}(1-a_i)\Delta\hat{\tau}_s^{(i)}\right)^{\frac{1}{a_i}}, \quad s=1,\ldots,k$$

και

$$A_0^{(i)}\left(p_0 - p_1\right)^{-\frac{1}{a_i}} = C_i$$

Είναι

$$x_s^{(i)} = \sum_{j=1}^{s}x_j^{(i)} - \sum_{j=1}^{s-1}x_j^{(i)}$$

και από την (5.13)

$$x_s^{(i)} = A_s^{(i)}\left(p_s - p_{s+1}\right)^{-\frac{1}{a_i}} - A_{s-1}^{(i)}\left(p_{s-1} - p_s\right)^{-\frac{1}{a_i}}, \quad s=1,\ldots,k \qquad (5.14)$$

Αν η διαφορά στο δεξί μέλος της παραπάνω εξίσωσης είναι μη αρνητική για όλες τις τιμές του s, το διάνυσμα ζήτησης χώρου $\boldsymbol{x}^{(i)}$, που έχει συνιστώσες $x_s^{(i)}$, $s=1,\ldots,k$ όπως βρέθηκαν στην (5.14), είναι ένα τοπικό μέγιστο της συνάρτησης καθαρού κέρδους $b_i\left(\boldsymbol{x}^{(i)}\right)$ της L1 cache i, και επειδή αυτή, όπως δείχθηκε, είναι μια κοίλη συνάρτηση, αυτό θα είναι και ολικό μέγιστο. Επίσης, πάλι λόγω της γνήσιας κοιλότητας της $b_i\left(\boldsymbol{x}^{(i)}\right)$, αυτό θα είναι και μοναδικό.

Αν όμως κάποιες από τις συνιστώσες $x_s^{(i)}$ του διανύσματος ζήτησης χώρου $\boldsymbol{x}^{(i)}$, είναι αρνητικές, η λύση που βρέθηκε είναι μη εφικτή και το μέγιστο πρέπει να αναζητηθεί σε ένα διάνυσμα που περιέχει μηδενικά στοιχεία.

Οπότε για την εύρεση του διανύσματος που μεγιστοποιεί τη συνάρτηση καθαρού κέρδους $b_i\left(\boldsymbol{x}^{(i)}\right)$ της L1 cache i, ακολουθούμε την παρακάτω διαδικασία:

1. Ελέγχουμε αν οι συνιστώσες του διανύσματος που βρέθηκε είναι μη αρνητικές, και αν ναι, τότε το διάνυσμα αυτό είναι το ζητούμενο.

2. Αν όχι, τότε βρίσκουμε την μικρότερη από τις αρνητικές συνιστώσες και την αντικαθιστούμε με μηδέν.
3. Λύνουμε από την αρχή τις εξισώσεις $\frac{\partial\left(u_i\left(\boldsymbol{x}^{(i)}\right)-p^{\mathrm{T}}\boldsymbol{x}^{(i)}\right)}{\partial x_s^{(i)}}=0$, που αντιστοιχούν στις συνιστώσες $x_s^{(i)}$ που απέμειναν.
4. Πηγαίνουμε στο βήμα 1.

Με δεδομένο, λοιπόν, το διάνυσμα τιμών $\boldsymbol{p}=(p_1,\cdots,p_k)$ που θέτουν οι k L2 cache, θα υπάρχει μοναδικό διάνυσμα ζήτησης χώρου $\boldsymbol{x}^{(i)}$, $i=1,...,n$ για κάθε μια από τις n L1 cache, το οποίο θα μεγιστοποιεί αντίστοιχα το καθαρό κέρδος τους.

5.4 Πωλητής (L2 cache)

Έστω $\boldsymbol{p}_{-j}=\left(p_1,p_2,\ldots,p_{j-1},p_{j+1},\ldots,p_k\right)\in S_{-j}$ το διάνυσμα τιμών που θέτουν οι L2 cache εκτός της j-ιοστής. Η συνάρτηση συνολικής ζήτησης χώρου $y_j(\boldsymbol{p})=\sum_{i=1}^{n}x_j^{(i)}$ από τον δίσκο της L2 cache j, με δεδομένο και σταθερό το διάνυσμα τιμών $\boldsymbol{p}_{-j}$, μπορεί να γραφεί

$$y_j\left(p_j,\boldsymbol{p}_{-j}\right)=\sum_{i=1}^{n}x_j^{(i)}$$

και από την σχέση (5.14), να πάρει την έκφραση

$$y_j\left(p_j,\boldsymbol{p}_{-j}\right)=\sum_{i=1}^{n}A_j^{(i)}\left(p_j-p_{j+1}\right)^{-\frac{1}{a_i}}-\sum_{i=1}^{n}A_{j-1}^{(i)}\left(p_{j-1}-p_j\right)^{-\frac{1}{a_i}} \qquad (5.15)$$

Πρόταση 5.4. *Υπάρχει μια μοναδική τιμή χρέωσης, έστω* p_j^*, *ώστε* $y_j\left(p_j^*,\boldsymbol{p}_{-j}\right)=C_{2,j}$, *για κάθε* $j=1,\ldots,k$ *και για κάθε* $\boldsymbol{p}_{-j}\in S_{-j}$.

Απόδειξη: Η απόδειξη βρίσκεται στο Παράρτημα. ■

Η Πρόταση 5.4 δείχνει ότι υπάρχει μια μοναδική τιμή, p_j^*, την οποία αν θέσει η L2 cache j, θα προκαλέσει ζήτηση χώρου ίση με την χωρητικότητα της, $C_{2,j}$.

Το κέρδος της L2 cache j, με δεδομένο και σταθερό διάνυσμα τιμών $\boldsymbol{p}_{-j}$ είναι

$$\pi_j\left(p_j,\boldsymbol{p}_{-j}\right)=\begin{cases} p_j\cdot y_j\left(p_j,\boldsymbol{p}_{-j}\right) & ,y_j\left(p_j,\boldsymbol{p}_{-j}\right)\le C_{2,j} \\ p_j\cdot C_{2,j} & ,y_j\left(p_j,\boldsymbol{p}_{-j}\right)> C_{2,j} \end{cases} \tag{5.16}$$

καθώς, όταν η συνολική ζήτηση χώρου είναι μικρότερη ή ίση από την χωρητικότητα της, τότε παρέχει το χώρο που της ζητήθηκε, και έχει κέρδος $\pi_j\left(p_j,\boldsymbol{p}_{-j}\right)=p_j\cdot y_j\left(p_j,\boldsymbol{p}_{-j}\right)$, ενώ όταν της ζητούν χώρο περισσότερο από όσο έχει, τότε παρέχει όλο το χώρο που κατέχει, και το κέρδος της είναι $\pi_j\left(p_j,\boldsymbol{p}_{-j}\right)=p_j\cdot C_{2,j}$.

Πρόταση 5.5. *Η συνάρτηση κέρδους $\pi_j\left(p_j,\boldsymbol{p}_{-j}\right)$ της L2 cache j, παίρνει την μέγιστη τιμή της στο p_j^*, για κάθε $j=1,\ldots,k$ και για κάθε $\boldsymbol{p}_{-j}\in S_{-j}$.*

Απόδειξη: Η απόδειξη βρίσκεται στο Παράρτημα. ■

Δηλαδή, με άλλα λόγια, η L2 cache j, επιτυγχάνει το μέγιστο κέρδος της, όταν θέσει τιμή p_j^*, τιμή από την οποία θα προκύψει ζήτηση χώρου που θα εξαντλήσει το διαθέσιμο αποθηκευτικό χώρο της $C_{2,j}$.

Θεώρημα 5.2. *Το παίγνιο των L2 cache έχει μοναδική ισορροπία Nash το διάνυσμα $\boldsymbol{p}^*=\left(p_j^*,\boldsymbol{p}_{-j}^*\right)=\left(p_1^*,\ldots,p_k^*\right)\in S$, το οποίο είναι εφικτό.*

Απόδειξη: Από την Πρόταση 5.5, για την L2 cache j, ισχύει $\pi_j\left(p_j^*,\boldsymbol{p}_{-j}\right)\ge\pi_j\left(p_j,\boldsymbol{p}_{-j}\right)$ για κάθε $p_j\in S_j$ και για κάθε $\boldsymbol{p}_{-j}\in S_{-j}$. Όμως το $\boldsymbol{p}_{-j}^*\in S_{-j}$, οπότε $\pi_j\left(p_j^*,\boldsymbol{p}_{-j}^*\right)\ge\pi_j\left(p_j,\boldsymbol{p}_{-j}^*\right)$. Άρα το διάνυσμα τιμών $\boldsymbol{p}^*=\left(p_1^*,\ldots,p_k^*\right)$ αποτελεί μια ισορροπία Nash για το παίγνιο. Για τις συνιστώσες p_j^* του διανύσματος $\boldsymbol{p}^*$, ισχύουν επίσης:

i) Κάθε μια αποτελεί μοναδική τιμή, μέσω της οποίας μεγιστοποιείται η συνάρτηση κέρδους της αντίστοιχης L2 cache, το οποίο σημαίνει ότι η παραπάνω αποδειχθείσα ισορροπία Nash είναι μοναδική.
ii) Κάθε μια αποτελεί τιμή, η οποία αν τεθεί από την αντίστοιχη L2 cache θα προκαλέσει συνολική ζήτηση χώρου ίση με τη χωρητικότητα της και όχι πάνω από

αυτή, οπότε ο περιορισμός της (5.6), $\sum_{i=1}^{n} x_j^{(i)} \le C_{2,j}$, ικανοποιείται και το διάνυσμα $\boldsymbol{p}^*$ είναι εφικτό. ■

Αυτό που μένει τώρα, είναι να βρεθεί το διάνυσμα τιμών $\boldsymbol{p}^* = \left(p_1^*,\ldots,p_k^*\right)$. Από την Πρόταση 5.4 και τη σχέση (5.15), τα p_j^* πρέπει να ικανοποιούν τις

$$\sum_{i=1}^{n} A_j^{(i)}\left(p_j^* - p_{j+1}^*\right)^{-\frac{1}{a_i}} - \sum_{i=1}^{n} A_{j-1}^{(i)}\left(p_{j-1}^* - p_j^*\right)^{-\frac{1}{a_i}} = C_{2,j}, \quad j=1,\ldots,k \tag{5.17}$$

όπου. $p_{k+1}^* = 0$ και $\sum_{i=1}^{n} A_0^{(i)}\left(p_0^* - p_1^*\right)^{-\frac{1}{a_i}} = \sum_{i=1}^{n} C_i$.

Θέτοντας για κάθε $j=1,\ldots,k$

$$K_j = \sum_{i=1}^{n} A_j^{(i)}\left(p_j^* - p_{j+1}^*\right)^{-\frac{1}{a_i}}$$

οι παραπάνω εξισώσεις (5.17) παίρνουν τη μορφή $K_j - K_{j-1} = C_{2,j}$. Δημιουργώντας τα αθροίσματα $\sum_{j=1}^{\ell}\left(K_j - K_{j-1}\right) = \sum_{j=1}^{\ell} C_{2,j}$, $\ell=1,\ldots,k$, προκύπτει η σχέση $K_\ell - K_0 = \sum_{j=1}^{\ell} C_{2,j}$. Με τον αρχικό συμβολισμό είναι $\sum_{i=1}^{n} A_\ell^{(i)}\left(p_\ell^* - p_{\ell+1}^*\right)^{-\frac{1}{a_i}} - \sum_{i=1}^{n} A_0^{(i)}\left(p_0^* - p_1^*\right)^{-\frac{1}{a_i}} = \sum_{j=1}^{\ell} C_{2,j}$, και τελικά

$$\sum_{i=1}^{n} A_\ell^{(i)}\left(p_\ell^* - p_{\ell+1}^*\right)^{-\frac{1}{a_i}} = \sum_{i=1}^{n} C_i + \sum_{j=1}^{\ell} C_{2,j} \tag{5.18}$$

Ορίζοντας τη συνάρτηση

$$f\left(\Delta p_\ell^*\right) = \sum_{i=1}^{n} A_\ell^{(i)}\left(\Delta p_\ell^*\right)^{-\frac{1}{a_i}} - \sum_{i=1}^{n} C_i - \sum_{j=1}^{\ell} C_{2,j}, \quad \ell=1,\ldots,k$$

με

$$\Delta p_j^* = p_j^* - p_{j+1}^* > 0, \quad j=1,\ldots,k \quad \text{με } p_{k+1} = 0$$

διαπιστώνουμε ότι αυτή είναι γνησίως φθίνουσα, καθώς $\frac{\partial f\left(\Delta p_\ell^*\right)}{\partial \Delta p_\ell^*} = -\sum_{i=1}^{n} \frac{1}{a_i} A_\ell^{(i)}\left(\Delta p_\ell^*\right)^{-\frac{1}{a_i}-1} < 0$. Οπότε, για κάθε $\ell=1,\ldots,k$, υπάρχει μοναδικό Δp_ℓ^*, έτσι ώστε $f\left(\Delta p_\ell^*\right) = 0$.

Η εύρεση του, θα γίνει με την εφαρμογή ενός διχοτομικού αλγόριθμου. Συγκεκριμένα, θεωρείται μια ελάχιστη διαφορά τιμών $\Delta p^*_{\min}$ και μια μέγιστη $\Delta p^*_{\max}$, με την αναζητούμενη διαφορά τιμών να βρίσκεται στο διάστημα $\left(\Delta p^*_{\min}, \Delta p^*_{\max}\right)$.

Ως $\Delta p^*_{\min}$ μπορεί να επιλεγεί το μηδέν, παρατηρώντας ότι τότε η συνάρτηση f γίνεται συν άπειρο.

Ως $\Delta p^*_{\max}$ θα πρέπει να επιλεγεί μια τιμή, η οποία θα κάνει την συνάρτηση f αρνητική. Μια τέτοια τιμή μπορεί να βρεθεί πειραματικά (αυξάνοντας σταδιακά την τιμή μέχρι να γίνει μικρότερη από το άθροισμα των χωρητικοτήτων που βρίσκονται στο δεξί μέλος της 5.18).

Έχοντας υπολογίσει, λοιπόν, και το $\Delta p^*_{\max}$, η εφαρμογή της διχοτομικής μεθόδου για τον υπολογισμό της τιμής, με την οποία η συνάρτηση f θα μηδενιστεί, μπορεί να γίνει άμεσα.

Ακολουθεί ο διχοτομικός αλγόριθμος:

$low = \Delta p^*_{\min}$; /* $f\left(\Delta p^*_{\min}\right) > 0$ */

$hi = \Delta p^*_{\max}$; /* $f\left(\Delta p^*_{\max}\right) < 0$ */

while *(1) {*

$mid = \dfrac{low + hi}{2}$;

if *(* $hi - low < \varepsilon$ || $f\left(mid\right) = 0$ *) {*

result = mid;

exit;

}

if $f\left(mid\right) > 0$

low = mid;

else

hi = mid;

}

Η σταθερά ε είναι ένας αρκετά μικρός, θετικός, πραγματικός αριθμός, ο οποίος χρησιμοποιείται ως κριτήριο τερματισμού του επαναληπτικού αλγόριθμου.

Η εφαρμογή του πιο πάνω αλγόριθμου για κάθε $\ell = 1,\ldots,k$, θα δώσει το διάνυσμα $\boldsymbol{\Delta p}^* = \left(\Delta p_1^*,\ldots,\Delta p_k^*\right)$. Και καθώς, για κάθε $\ell = 1,...,k$, είναι $p_\ell^* = \sum_{j=\ell}^{k} \Delta p_j^*$ και $p_{k+1} = 0$, υπολογίζεται το μοναδικό διάνυσμα τιμών $\boldsymbol{p}^* = \left(p_1^*,\ldots,p_k^*\right)$ που θα μεγιστοποιεί ταυτόχρονα τις συναρτήσεις ωφέλειας των L2 cache και είναι (όπως αποδείχθηκε παραπάνω) η ισορροπία Nash του παιγνίου των L2 cache.

6 Αριθμητική Επίλυση

Σε αυτό το κεφάλαιο, γίνεται η μελέτη ορισμένων επιλεγμένων χαρακτηριστικών της δομής των L1 cache – L2 cache, που εξετάστηκε σε αυτή την εργασία, και ο ρόλος που έχουν στη διαμόρφωση άλλων χαρακτηριστικών της δομής. Δόθηκε έμφαση σε χαρακτηριστικά, όπως τα έσοδα των L2 cache ή το όφελος που απολαμβάνουν οι L1 cache (δηλαδή το καθαρό όφελος).

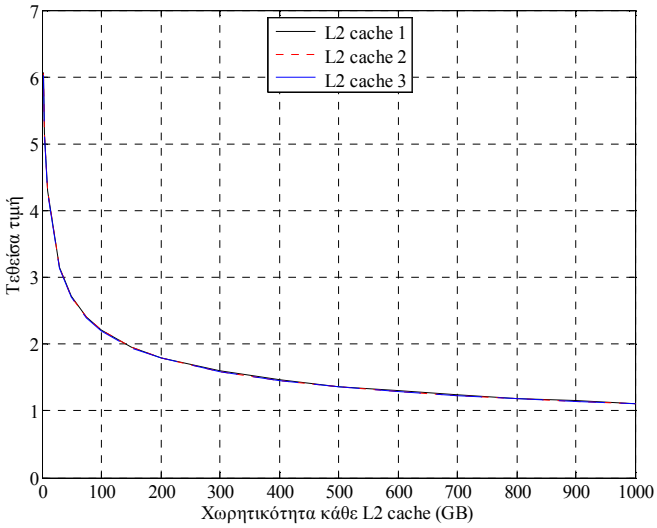
	L1 Cache 1	**L1 Cache 2**	**L1 Cache 3**
Χωρητικότητα L1 cache (C_i)	1GB	1GB	1GB
Συνολική χωρητικότητα αντικειμένων αναφερόμενα από την L1 cache (O_i)	500GB	500GB	500GB
Παράγοντας δημοτικότητας Zipf (a_i)	0.3	0.3	0.3
Χρόνος εξυπηρέτησης από L1 cache $(\tau_\ell^{(i)})$	5ms	5ms	5ms
Χρόνος εξυπηρέτησης από L2 cache 1 $(\tau_{r_1}^{(i)})$	10ms	10ms	10ms
Χρόνος εξυπηρέτησης από L2 cache 2 $(\tau_{r_2}^{(i)})$	13ms	13ms	13ms
Χρόνος εξυπηρέτησης από L2 cache 3 $(\tau_{r_3}^{(i)})$	16ms	16ms	16ms
Χρόνος εξυπηρέτησης από πηγαίο διακομιστή $(\tau_{os}^{(i)})$	1,000ms	1,000ms	1,000ms
	L2 Cache 1	**L2 Cache 2**	**L2 Cache 3**
Χωρητικότητα L2 cache $(C_{2,j})$	10GB	10GB	10GB

Πίνακας 6.1. Χαρακτηριστικά του μοντέλου του παραδείγματος

Η μελέτη γίνεται με τη βοήθεια ενός μοντέλου, με τρεις L1 cache και τρεις L2 cache, τα χαρακτηριστικά των οποίων παρουσιάζονται στον Πίνακα 6.1. Οι παράμετροι

που παρουσιάζει ο Πίνακας 6.1 περιγράφουν μια βασική ρύθμιση, η οποία είναι συμμετρική. Δηλαδή, τα χαρακτηριστικά των τριών L1 cache είναι ίδια, όπως επίσης και αυτά των τριών L2 cache. Η εισαγωγή κάποιας διαφορετικότητας μεταξύ των L1 cache ή (και) των L2 cache, θα δώσει τη δυνατότητα σύγκρισης του βαθμού της επίδρασης κάποιων χαρακτηριστικών του εξεταζόμενου μοντέλου, σε άλλα χαρακτηριστικά του μοντέλου.

Πριν ακόμη εισαχθεί η παραπάνω αναφερόμενη διαφορετικότητα που θα βοηθήσει στη σύγκριση, μπορεί κάποιος να δει το Σχήμα 6.1, στο οποίο φαίνεται η σχέση μεταξύ της χωρητικότητας των L2 cache και των τιμών που αυτές θέτουν. Όλες οι παράμετροι έχουν τις τιμές του Πίνακα 1, και η αύξηση των χωρητικοτήτων και των τριών L2 cache, γίνεται ταυτόχρονα. Η βασική παρατήρηση είναι ότι οι τιμές που θέτουν μειώνονται, όσο η χωρητικότητα τους αυξάνεται. Αυτή η μείωση των τιμών δεν προκαλεί μείωση των εσόδων τους, όπως θα φανεί παρακάτω.



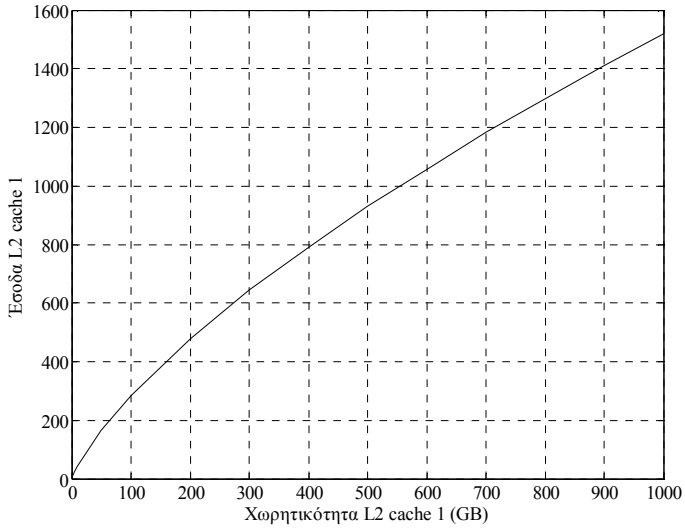
Σχήμα 6.1. Τεθείσα τιμή από τις L2 cache

Με βάση τις παραμέτρους του Πίνακα 6.1, ορίζεται ένα σύνολο σεναρίων με μεταβολή κάποιων από τις παραμέτρους. Στα παρακάτω, όπου υπάρχει μεταβολή της χωρητικότητας μιας L2 cache, θεωρούμε ότι οι χωρητικότητες των άλλων δύο L2 cache, παραμένουν σταθερές στην αρχική τιμή τους, εκτός εάν διαφορετικά αναφέρεται.

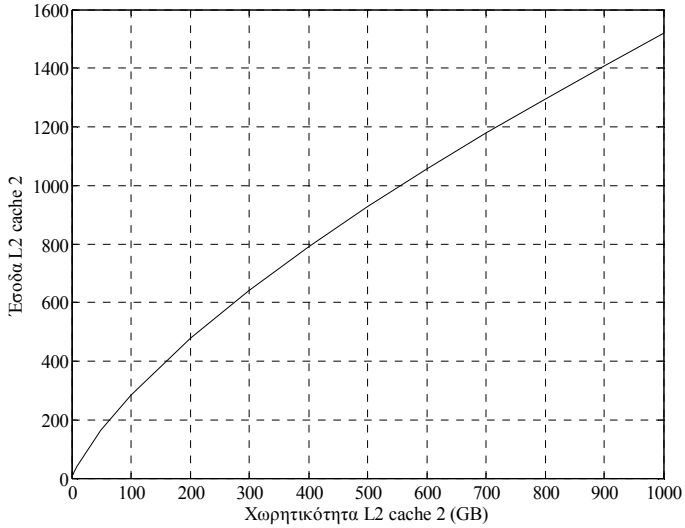
6.1 Σενάριο 1 (Μεταβολή της χωρητικότητας των L1 cache)

Για να διαφανεί η επίδραση της χωρητικότητας που διαθέτει κάθε L1 cache (C_i), στο χώρο που θα αποκτήσει από κάθε μια L2 cache $\left(x_j^{(i)}\right)$, καθώς και στην προκύπτουσα ωφέλεια, αντί του διανύσματος $(C_1, C_2, C_3) = (1\text{GB}, 1\text{GB}, 1\text{GB})$, χρησιμοποιείται το διάνυσμα $(C_1, C_2, C_3) = (1\text{GB}, 2\text{GB}, 3\text{GB})$.

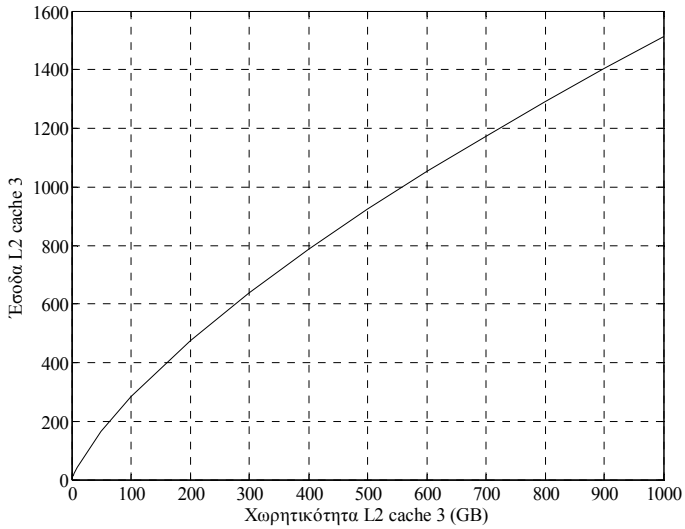
Στο Σχήμα 6.2, φαίνονται τα έσοδα για την L2 cache 1. Μπορεί να παρατηρηθεί ότι τα έσοδα της L2 cache 1 αυξάνονται όσο αυξάνεται κι η χωρητικότητα της.



Σχήμα 6.2. Έσοδα της L2 cache 1



Σχήμα 6.3. Έσοδα της L2 cache 2



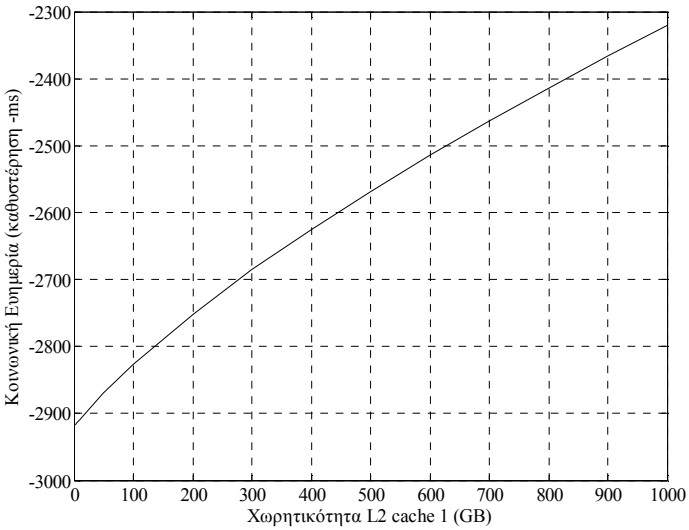
Σχήμα 6.4. Έσοδα της L2 cache 3

Τα έσοδα των L2 cache 2 και L2 cache 3 φαίνονται στο Σχήμα 6.3 και στο Σχήμα 6.4, αντίστοιχα, με τα ίδια συμπεράσματα.

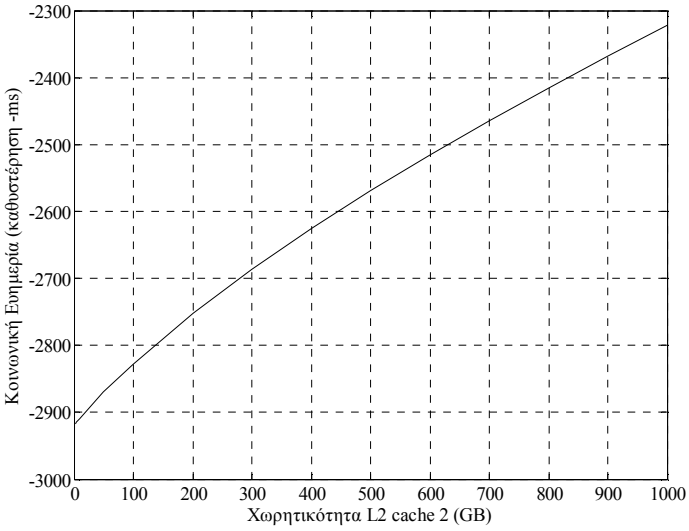
Στα Σχήμα 6.5, Σχήμα 6.6 και Σχήμα 6.7, φαίνεται η Κοινωνική Ευημερία (Social Welfare) που επιτυγχάνεται στο εξεταζόμενο σύστημα ανάλογα με το ποια L2 cache μεταβάλλει τη χωρητικότητα της. Ως Κοινωνική Ευημερία, ορίζεται το άθροισμα των καθαρών ωφελειών των αγοραστών (L1 cache) του συστήματος, δηλαδή

$$SW\left(\boldsymbol{x}^{(i)}, \boldsymbol{p}\right) = \sum_{i=1}^{n}\left(u_i\left(\boldsymbol{x}^{(i)}\right) - \sum_{j=1}^{k} p_j x_j^{(i)}\right)$$

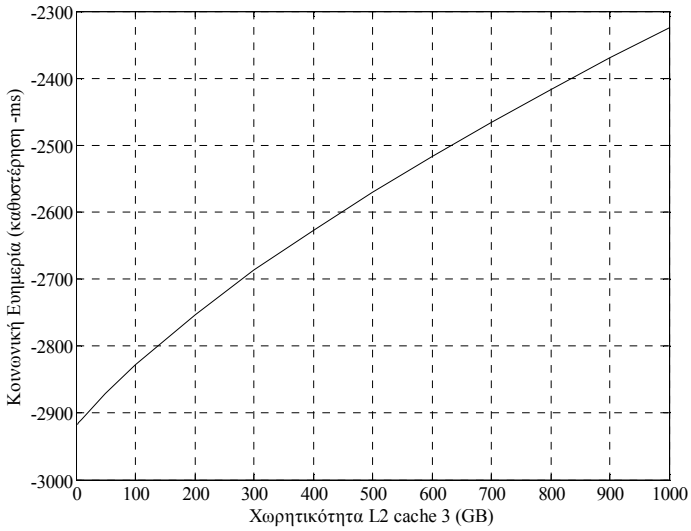
Οι αρνητικές τιμές που απεικονίζονται οφείλονται στο ότι οι συναρτήσεις ωφέλειας των αγοραστών έχουν οριστεί ως το αρνητικό του αντίστοιχου χρόνου εξυπηρέτησης. Από τη κλίση της καμπύλης, μπορεί να παρατηρηθεί η επίτευξη ικανοποιητικής Κοινωνικής Ευημερίας.



Σχήμα 6.5. Κοινωνική Ευημερία

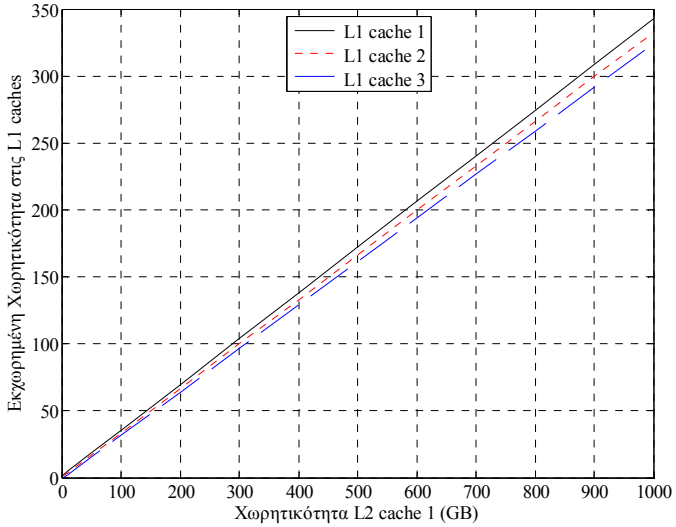


Σχήμα 6.6. Κοινωνική Ευημερία



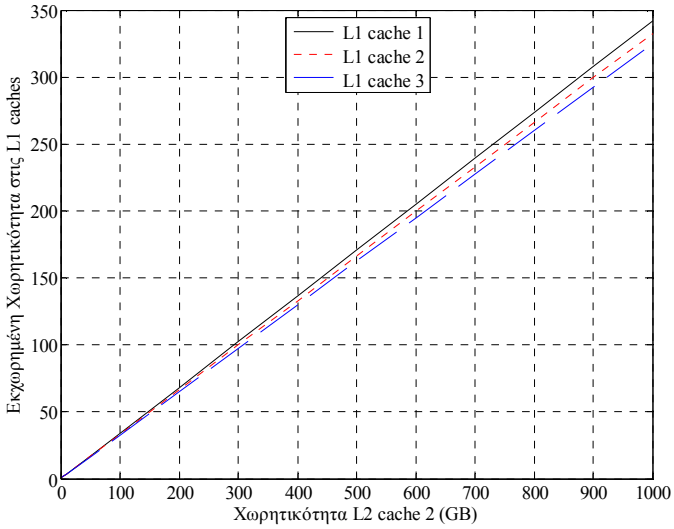
Σχήμα 6.7. Κοινωνική Ευημερία

Στο Σχήμα 6.8, φαίνεται η χωρητικότητα που εκχωρείται σε κάθε μια από τις τρεις L1 cache, όταν η χωρητικότητα της L2 cache 1, μεταβάλλεται. Σύμφωνα με τα αποτελέσματα, η L1 cache 1 παίρνει περισσότερη χωρητικότητα από την L1 cache 2, η οποία παίρνει περισσότερη χωρητικότητα από την L1 cache 3. Το συγκεκριμένο αποτέλεσμα οφείλεται στο ότι, καθώς $C_1 < C_2 < C_3$, η L1 cache 1 έχει περισσότερη ανάγκη για χωρητικότητα από την L1 cache 2, η οποία έχει περισσότερη ανάγκη από την L1 cache 3, γεγονός που εκμεταλλεύεται η L2 cache 1, ώστε να μεγιστοποιήσει τα κέρδη της.

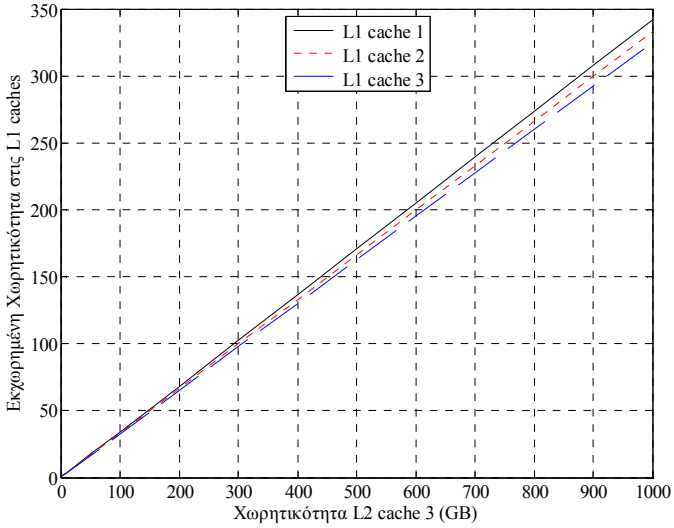


Σχήμα 6.8. Εκχωρημένη Χωρητικότητα στις L1 cache

Παρόμοια εκχώρηση χωρητικότητας στις L1 cache, παρατηρείται στο Σχήμα 6.9 (μεταβολή χωρητικότητας της L2 cache 2) και στο Σχήμα 6.10 (μεταβολή χωρητικότητας της L2 cache 3).



Σχήμα 6.9. Εκχωρημένη Χωρητικότητα στις L1 cache

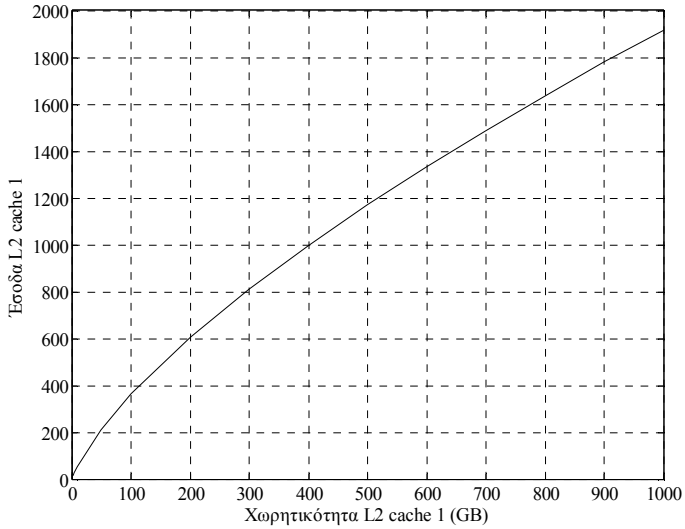


Σχήμα 6.10. Εκχωρημένη Χωρητικότητα στις L1 cache

6.2 Σενάριο 2 (Μεταβολή της συνολικής χωρητικότητας αντικειμένων αναφερόμενα από τις L1 cache)

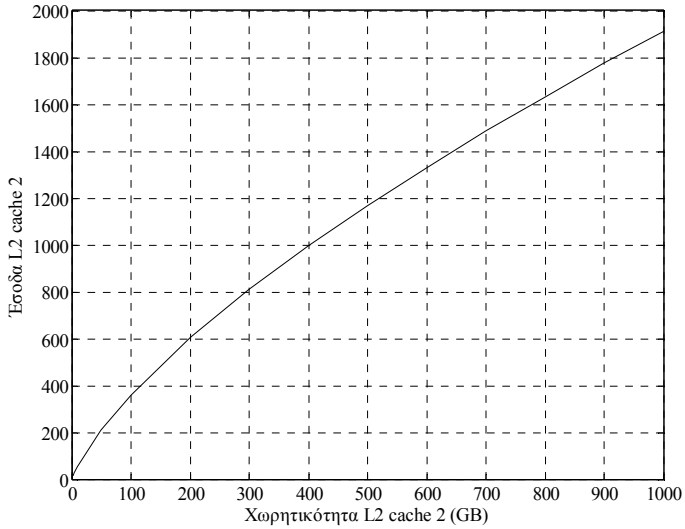
Εκτός από το σενάριο στο οποίο οι L1 cache διαφέρουν στη χωρητικότητα που διαθέτουν (C_i), μελετάται και το σενάριο, στο οποίο διαφέρουν στο συνολικό χώρο των αντικειμένων στα οποία αναφέρονται (O_i). Συγκεκριμένα, οι χωρητικότητες C_i τίθενται στις αρχικές τους τιμές, δηλαδή $(C_1, C_2, C_3) = (1\text{GB}, 1\text{GB}, 1\text{GB})$, ενώ για το συνολικό χώρο των αντικειμένων χρησιμοποιείται το διάνυσμα $(O_1, O_2, O_3) = (250\text{GB}, 500\text{GB}, 750\text{GB})$.

Στο Σχήμα 6.11, φαίνονται τα έσοδα για την L2 cache 1. Όπως και για το προηγούμενο σενάριο (διαφορετικές χωρητικότητες των L1 cache), παρατηρείται ότι τα έσοδα της L2 cache 1 αυξάνονται όσο αυξάνεται κι η χωρητικότητα της. Επίσης, μπορεί να παρατηρηθεί ότι τα έσοδα της L2 cache 1, σε αυτό το σενάριο, είναι αυξημένα σε σχέση με το προηγούμενο. Αυτό είναι αναμενόμενο, επειδή οι L1 cache έχουν (συνολικά) αυξημένες ανάγκες για αποθηκευτικό χώρο, καθώς έχουν μικρότερη χωρητικότητα (διάνυσμα χωρητικοτήτων $(C_1, C_2, C_3) = (1\text{GB}, 1\text{GB}, 1\text{GB})$), εδώ σε σχέση με το σενάριο 1 (διάνυσμα χωρητικοτήτων $(C_1, C_2, C_3) = (1\text{GB}, 2\text{GB}, 3\text{GB})$).

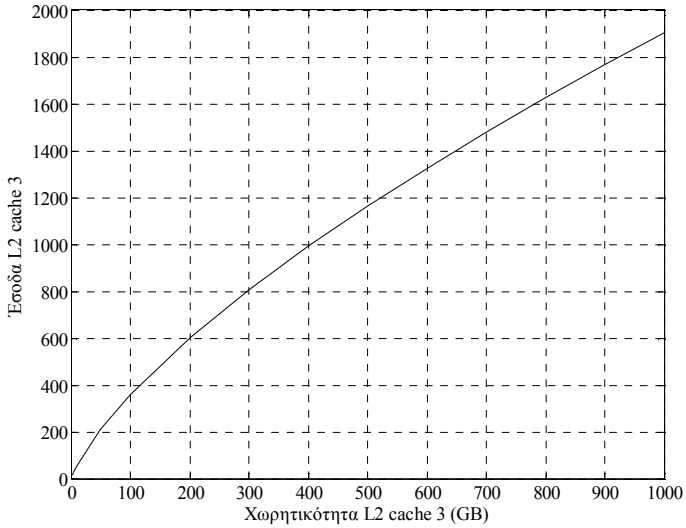


Σχήμα 6.11. Έσοδα της L2 cache 1

Τα έσοδα των L2 cache 2 και L2 cache 3 φαίνονται στο Σχήμα 6.12 και στο Σχήμα 6.13, αντίστοιχα, με τα ίδια συμπεράσματα.

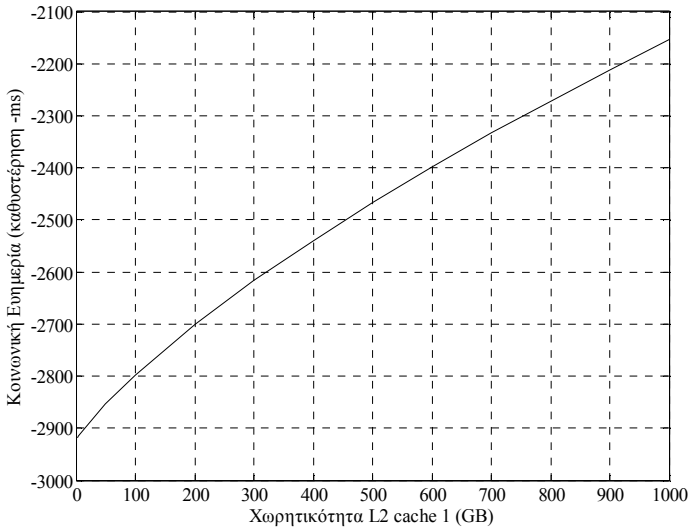


Σχήμα 6.12. Έσοδα της L2 cache 2

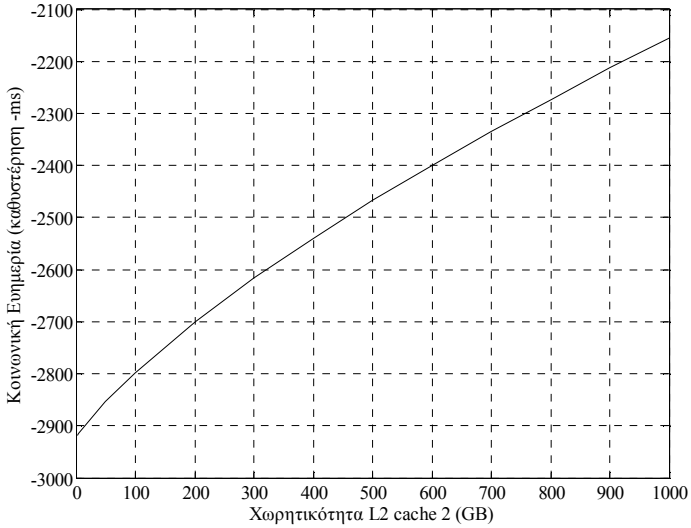


Σχήμα 6.13. Έσοδα της L2 cache 3

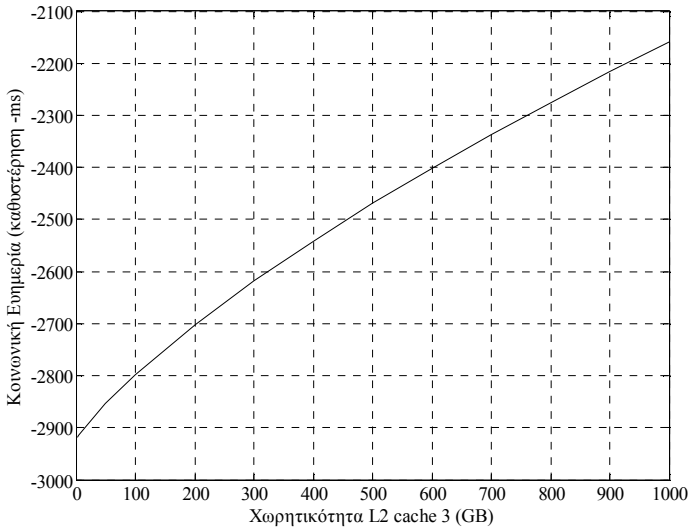
Στα Σχήμα 6.14, Σχήμα 6.15 και Σχήμα 6.16, φαίνεται η Κοινωνική Ευημερία (Social Welfare) που επιτυγχάνεται ανάλογα με το ποια L2 cache μεταβάλλει τη χωρητικότητα της. Από τη κλίση της καμπύλης, μπορεί να παρατηρηθεί η επίτευξη ικανοποιητικής Κοινωνικής Ευημερίας.



Σχήμα 6.14. Κοινωνική Ευημερία

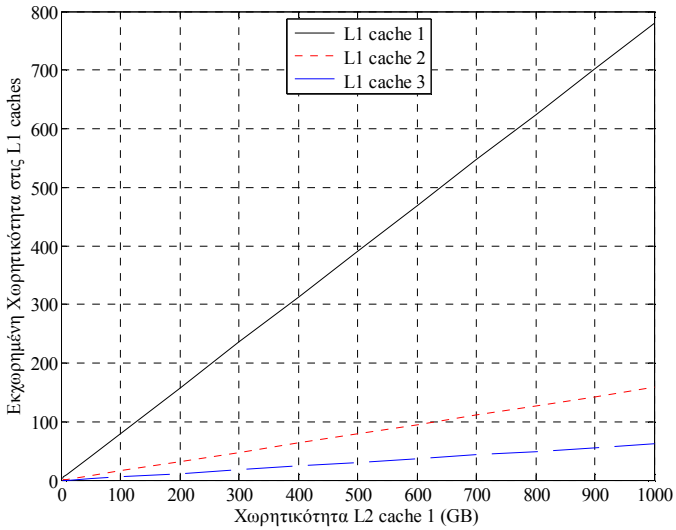


Σχήμα 6.15. Κοινωνική Ευημερία



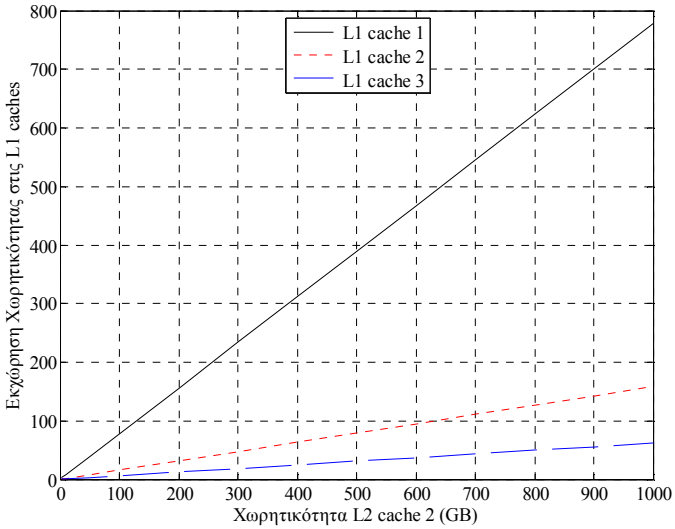
Σχήμα 6.16. Κοινωνική Ευημερία

Στο Σχήμα 6.17, φαίνεται η χωρητικότητα που εκχωρείται σε κάθε μια από τις τρεις L1 cache, όταν η χωρητικότητα της L2 cache 1, μεταβάλλεται. Σύμφωνα με τα αποτελέσματα, η L1 cache 1 παίρνει περισσότερη χωρητικότητα από την L1 cache 2, η οποία παίρνει περισσότερη χωρητικότητα από την L1 cache 3. Το συγκεκριμένο αποτέλεσμα οφείλεται στο ότι $O_1 < O_2 < O_3$. Όσο πιο μεγάλος είναι ο χώρος των αντικειμένων στα οποία αναφέρεται μια L1 cache, τόσο μικρότερη ωφέλεια της προσφέρει ένα συγκεκριμένο ποσό χωρητικότητας. Έτσι, είναι πιο αποτελεσματικό να εκχωρείται περισσότερη χωρητικότητα στις L1 cache που έχουν μικρότερο χώρο αναφερόμενων αντικειμένων.

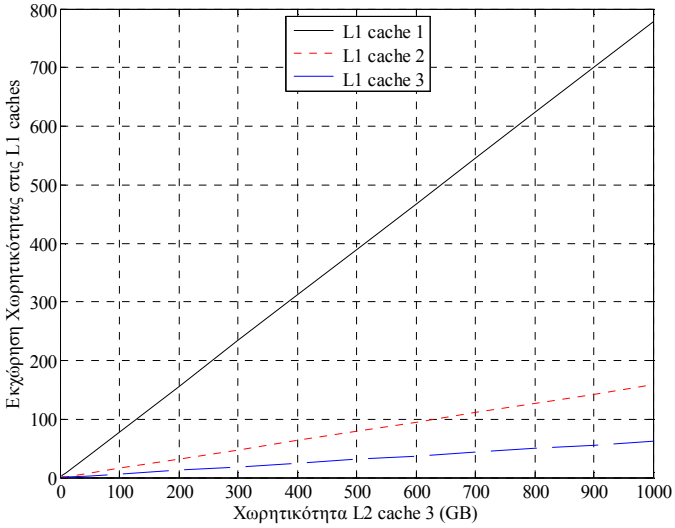


Σχήμα 6.17. Εκχωρημένη Χωρητικότητα στις L1 cache

Παρόμοια εκχώρηση χωρητικότητας στις L1 cache, παρατηρείται στο Σχήμα 6.18 (μεταβολή χωρητικότητας της L2 cache 2) και στο Σχήμα 6.19 (μεταβολή χωρητικότητας της L2 cache 3).



Σχήμα 6.18. Εκχωρημένη Χωρητικότητα στις L1 cache



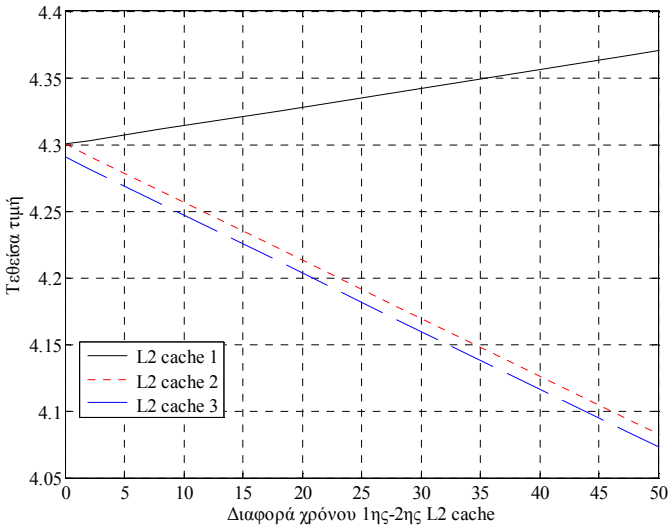
Σχήμα 6.19. Εκχωρημένη Χωρητικότητα στις L1 cache

6.3 Σενάριο 3 (Μεταβολή της διαφοράς του χρόνου εξυπηρέτησης από τις L2 cache)

Κατά τη διαδικασία ανάπτυξης της εργασίας (δες Κεφάλαιο 5), έγινε φανερό ότι ο χρόνος εξυπηρέτησης, και συγκεκριμένα η διαφορά του χρόνου εξυπηρέτησης από δύο διαδοχικές L2 cache, παίζει σημαντικό ρόλο στη διαμόρφωση συγκεκριμένων χαρακτηριστικών του εξεταζόμενου μοντέλου.

Στο Σχήμα 6.20, φαίνεται η μεταβολή που υφίσταται η τιμή που θέτει κάθε L2 cache, καθώς η διαφορά του χρόνου εξυπηρέτησης $\left(\tau_{r_2}^{(i)} - \tau_{r_1}^{(i)}\right)$, από τις L2 cache 2 και L2 cache 1, αυξάνεται. Η τιμή που τέθηκε για την διαφορά $\tau_{r_3}^{(i)} - \tau_{r_2}^{(i)}$ είναι αυτή του βασικού μοντέλου του Πίνακα 6.1, δηλαδή 3ms. Η παράμετρος που είχε αλλάξει στο προηγούμενο μοντέλο, έχει τεθεί στην αρχική τιμή της.

Στο Σχήμα 6.20, μπορούν να παρατηρηθούν τα εξής: 1) Η τιμή που θέτει η L2 cache 1 αυξάνεται όσο αυξάνεται η διαφορά του χρόνου και 2) η τιμή που θέτει η L2 cache 2 μειώνεται, όσο αυξάνεται η διαφορά του χρόνου, παρασύροντας στη φθίνουσα αυτή πορεία και την τιμή που θέτει η L2 cache 3.

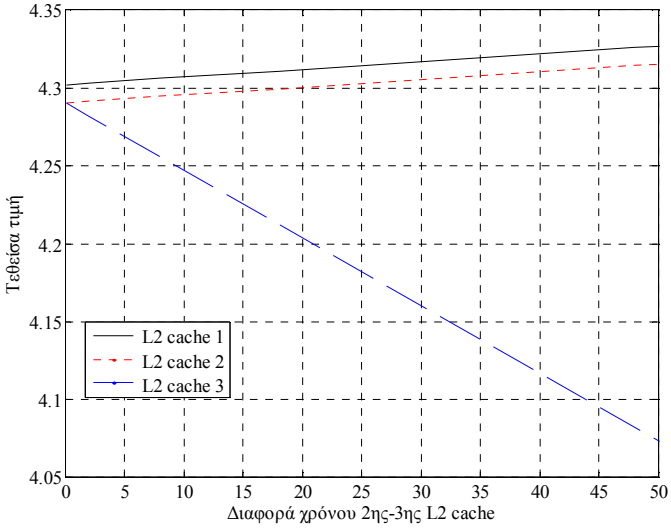


Σχήμα 6.20. Τεθείσα τιμή από τις L2 cache

Αναλόγως, στο Σχήμα 6.21, φαίνεται η μεταβολή που υφίσταται η τιμή που θέτει κάθε L2 cache, καθώς η διαφορά του χρόνου εξυπηρέτησης $\left(\tau_{r_3}^{(i)} - \tau_{r_2}^{(i)}\right)$, από τις L2

cache 3 και L2 cache 2, αυξάνει. Η τιμή που τέθηκε για την διαφορά $\tau_{r_2}^{(i)} - \tau_{r_1}^{(i)}$ είναι αυτή του βασικού μοντέλου του Πίνακα 6.1, δηλαδή 3ms.

Μπορούν να παρατηρηθούν τα εξής: 1) Η τιμή που θέτει η L2 cache 1, όπως και η τιμή που θέτει η L2 cache 2, αυξάνονται όσο αυξάνεται η διαφορά του χρόνου και 2) η τιμή που θέτει η L2 cache 3 μειώνεται, όσο αυξάνεται η διαφορά του χρόνου.



Σχήμα 6.21. Τεθείσα τιμή από τις L2 cache

7 Συμπεράσματα

Στην εργασία αυτή, προτάθηκε ένα Οικονομικό σχήμα για τη διαχείριση της υπηρεσίας caching σε μια ιεραρχία. Στα πλαίσια της μελέτης, προτάθηκε ένα μοντέλο βασισμένο σε Οικονομικές και Παιγνιοθεωρητικές αρχές, βάσει του οποίου ορίστηκαν οι συμπεριφορές που διέπουν τις εμπλεκόμενες οντότητες. Σύμφωνα με το μοντέλο αυτό, οι L1 cache, με σκοπό την επέκταση της χωρητικότητας τους, μπορούσαν να αγοράσουν επιπλέον χωρητικότητα από τις L2 cache. Σε αντίθεση με το τυπικό ιεραρχικό μοντέλο caching, στο οποίο η χωρητικότητα μιας cache διατίθεται προς κοινή χρήση, στο προτεινόμενο μοντέλο, οι αποθηκευτικοί χώροι των L2 cache χωρίζονται σε διακριτά τμήματα που χρησιμοποιούνται αποκλειστικά από κάθε L1 cache.

Για να αποφευχθούν προβλήματα μονοπώλησης της χωρητικότητας από "επιθετικές" L1 cache, ορίστηκε μια Ολιγοπωλιακή αγορά με χρέωση χωρίς διάκριση (ομοιόμορφη χρέωση) και οικονομικό αγαθό τον αποθηκευτικό χώρο. Κάθε L2 cache έθετε μια τιμή ανά μονάδα χωρητικότητας, την ίδια για κάθε μια L1 cache. Οι L2 cache ανταγωνίζονταν, μέσω των τιμών που έθεταν, για την προσέλκυση πελατών (L1 cache). Με τη βοήθεια της Θεωρίας Παιγνίων, το ανταγωνιστικό ως προς τις τιμές Ολιγοπωλιακό μοντέλο, αντιμετωπίστηκε ως μη συνεργατικό παίγνιο, και αποδείχθηκε η ύπαρξη Ισορροπίας Nash του παιγνίου. Ο προσδιορισμός του βέλτιστου διανύσματος τιμών χρέωσης ανά μονάδα αποθηκευτικού χώρου που έθεταν οι L2 cache, έγινε με τη χρήση διχοτομικού αλγόριθμου. Ο προσδιορισμός των διανυσμάτων εκχώρησης χωρητικότητας στις L1 cache, έγινε με χρήση αρχών της Θεωρίας Βελτιστοποίησης, και δόθηκε η αναλυτική έκφραση αυτών.

Η μελέτη των αποτελεσμάτων, μέσω συγκεκριμένων αριθμητικών παραδειγμάτων, έδειξε ότι σε κάθε περίπτωση, το σχήμα που εξετάσθηκε μπορεί να χαρακτηριστεί "δίκαιο", καθώς ακόμη και όταν οι χωρητικότητες των L2 cache ήταν σχετικά μικρές, όλες οι L1 cache λάμβαναν μη μηδενική χωρητικότητα. Καθώς οι αποθηκευτικοί χώροι των L2 cache αυξάνονταν, η εκχώρηση χωρητικοτήτων γίνονταν με βάση τις ανάγκες για χωρητικότητα κάθε μιας L1 cache. Αυτή με τον μικρότερο

αποθηκευτικό χώρο, λάμβανε και περισσότερη χωρητικότητα, γεγονός που οι L2 cache εκμεταλλεύονταν για μεγιστοποίηση των κερδών τους.

Επίσης, βρέθηκε η σχέση μεταξύ διαφόρων χαρακτηριστικών της δομής. Το ότι η εκχώρηση όλης της χωρητικότητας μιας L2 cache είναι απαραίτητη για την μεγιστοποίηση των κερδών της, ήταν γνωστό από την επίλυση του προβλήματος ΠΩΛΗΤΗΣ $_j$. Η χρήση των αριθμητικών παραδειγμάτων το επιβεβαίωσε και έδειξε ότι η αύξηση της χωρητικότητας των L2 cache είχε σαν αποτέλεσμα τη μείωση των τιμών που έθεταν, αλλά και την αύξηση των κερδών που είχαν. Ως εκ τούτου, οι L2 cache έχουν κίνητρο να επενδύσουν σε χωρητικότητα, ώστε να αυξήσουν τα κέρδη τους.

Ένα ακόμη, άξιο να αναφερθεί, συμπέρασμα είναι η επίδραση του χρόνου εξυπηρέτησης στην τεθείσα τιμή. Συγκεκριμένα, η αύξηση ή η μείωση της διαφοράς του χρόνου εξυπηρέτησης από δύο διαδοχικές L2 cache είναι καθοριστικός παράγοντας στη διαμόρφωση των βέλτιστων τιμών που αυτές θέτουν. Παρατηρήθηκε ότι, όταν ο χρόνος εξυπηρέτησης από μια L2 cache αυξάνεται, η τιμή που αυτή η cache ανακοινώνει, μειώνεται.

Κλείνοντας πρέπει να αναφέρουμε την ανάδειξη της Θεωρίας Παιγνίων ως βασικό εργαλείο για την μοντελοποίηση προβλημάτων αναλόγων με αυτό που εξετάστηκε στην παρούσα εργασία.

8 Αναφορές

[1] Γ. Αλυφαντής, "Διαχείριση Πόρων σε Συστήματα Κινητών Επικοινωνιών και Κατανεμημένα Υπολογιστικά Συστήματα", Διδακτορική Διατριβή, Αθήνα, 2008

[2] Χ. Αλιπράντης, S. Chakrabati, "Παίγνια και Λήψη Αποφάσεων", E.M.E., 2004

[3] D.P. Bertsekas, "Nonlinear Programming", Belmont, MA Athena Scientific, 1995

[4] L. Breslau, P. Cao, L. Fan, G. Phillips, S. Shenker, "Web Caching and Zipf-like Distributions: Evidence and Implications", Proc. IEEE INFOCOM 1999

[5] S.C. Chapra and R.P. Canale, "Numerical Methods for Engineers", McGraw-Hill, 1989

[6] C. Courcoubetis, and R. Weber, "Pricing Telecommunication Networks", Wiley, 2003

[7] F. Cowell, " Microeconomics: Principles and Analysis", Oxford, 2006

[8] Ö. Erçetin and L. Tassiulas, "Market-Based Resource Allocation for Content Delivery in the Internet", IEEE Transactions on Computers 52(12), σελ. 1573-1585, 2003

[9] J. Feigenbaum, C.Papadimitriou, and S.Shenker, "Sharing the Cost of Multicast Transmissions", Proceedings of the 32nd Annual ACM Symposium on Theory of Computing (STOC00), May 2000

[10] D. Fudenberg and J. Tirole, "Game Theory", MIT Press, Cambridge (MA), 1991

[11] S. Hadjiefthymiades, Y. Georgiadis and L. Merakos, "A Game Theoretic Approach to Web Caching", Proc. 3rd International IFIP-TC6 Networking Conference, Athens, Greece, 2004

[12] S.P. Hargreaves Heap, Y. Varoufakis, "Game Theory: A Critical Introduction", Routledge, 1995

[13] L. Kandiller, "Principles of Mathematics in Operations Research", Springer, 2007

[14] P. Konstanty and M. Kozipski, "Web Cache charging policies", position paper in NLANR Web Caching Workshop, Boulder, USA, 1997

[15] Y.A. Korilis, A.A. Lazar, and A. Orda, "Architecting Noncooperative Networks", IEEE JSAC, Vol. 13, No. 8, 1995

[16] A. Lazar, A. Orda and D. Pendarakis, "Virtual Path Bandwidth Allocation in Multiuser Networks", IEEE/ACM Transactions on Networking, Volume 5, Issue 6, December 1997, σελ. 861 – 871

[17] N. Laoutaris, G. Smaragdakis, A. Bestavros, I. Stavrakakis, "Mistreatment in Distributed Caching Groups: Causes and Implications", Proc. IEEE INFOCOM 2006, Barcelona, Spain

[18] D.G. Luenberger, "Linear and Nonlinear Programming", Addison Wesley, 1984

[19] A.B. MacKenzie, and S.B.Wicker, "Game Theory and the Design of Self-Configuring, Adaptive Wireless Networks", IEEE Communications Magazine, November 2001

[20] M.J. Osborne and A. Rubinstein, "A course in game theory", MIT Press, 1994

[21] M. Rabinovich and O. Spatscheck, "Web Caching and Replication", Addison Wesley, 2001

[22] J.B. Rosen, "Existence and Uniqueness of Equilibrium Points for Concave N-Person Games", Econometrica, Vol.33, No.3, 1965

[23] S.J. Shenker, "Making Greed Work in Networks: A Game Theoretic Analysis of Switch Service Disciplines", IEEE/ACM Transactions on Networking, Vol.3, No.6, December, 1995

[24] J. Tirole, "The Theory of Industrial Organization", MIT Press, 1998

[25] H.R. Varian, "Microeconomic Analysis", W.W. Norton, 1992

9 Παράρτημα: Αποδείξεις

Πρόταση 5.1. *Ο αναμενόμενος χρόνος εξυπηρέτησης* $E\{\tau_i\}$ *που δίνεται από την (5.10), παίρνει τη μορφή:*

$$E\{\tau_i\} = v_0^{(i)}\hat{\tau}_0^{(i)} + \left(1-v_0^{(i)}\right)\hat{\tau}_{k+1}^{(i)} + v_0^{(i)}\left(1-v_0^{(i)}\right)\left(\hat{\tau}_{k+1}^{(i)} - \hat{\tau}_1^{(i)}\right) - \left(1-v_0^{(i)}\right)\sum_{m=1}^{k}\left(v_m^{(i)}\left(\hat{\tau}_{m+1}^{(i)} - \hat{\tau}_m^{(i)}\right)\right)$$

Απόδειξη: Ο αναμενόμενος χρόνος εξυπηρέτησης $E\{\tau_i\}$ που δίνεται από την (5.10), γράφεται:

$$\begin{aligned}
E\{\tau_i\} &= v_0^{(i)}\hat{\tau}_0^{(i)} + \left(1-v_0^{(i)}\right)\sum_{m=1}^{k}\left(\left(v_m^{(i)} - v_{m-1}^{(i)}\right)\hat{\tau}_m^{(i)}\right) + \left(1-v_0^{(i)}\right)\left(1-v_k^{(i)} + v_0^{(i)}\right)\hat{\tau}_{k+1}^{(i)} \\
&= v_0^{(i)}\hat{\tau}_0^{(i)} + \left(1-v_0^{(i)}\right)\hat{\tau}_{k+1}^{(i)} + \left(1-v_0^{(i)}\right)\sum_{m=1}^{k}\left(\left(v_m^{(i)} - v_{m-1}^{(i)}\right)\hat{\tau}_m^{(i)}\right) - \left(1-v_0^{(i)}\right)\left(v_k^{(i)} - v_0^{(i)}\right)\hat{\tau}_{k+1}^{(i)} \\
&= v_0^{(i)}\hat{\tau}_0^{(i)} + \left(1-v_0^{(i)}\right)\hat{\tau}_{k+1}^{(i)} + \left(1-v_0^{(i)}\right)\sum_{m=1}^{k}\left(\left(v_m^{(i)} - v_{m-1}^{(i)}\right)\hat{\tau}_m^{(i)}\right) - \left(1-v_0^{(i)}\right)\sum_{m=1}^{k}\left(\left(v_m^{(i)} - v_{m-1}^{(i)}\right)\hat{\tau}_{k+1}^{(i)}\right) \\
&= v_0^{(i)}\hat{\tau}_0^{(i)} + \left(1-v_0^{(i)}\right)\hat{\tau}_{k+1}^{(i)} + \left(1-v_0^{(i)}\right)\sum_{m=1}^{k}\left(\left(v_m^{(i)} - v_{m-1}^{(i)}\right)\left(\hat{\tau}_m^{(i)} - \hat{\tau}_{k+1}^{(i)}\right)\right) \\
&= v_0^{(i)}\hat{\tau}_0^{(i)} + \left(1-v_0^{(i)}\right)\hat{\tau}_{k+1}^{(i)} + \left(1-v_0^{(i)}\right)\left(\sum_{m=1}^{k}\left(v_m^{(i)}\left(\hat{\tau}_m^{(i)} - \hat{\tau}_{k+1}^{(i)}\right)\right) - \sum_{m=1}^{k}\left(v_{m-1}^{(i)}\left(\hat{\tau}_m^{(i)} - \hat{\tau}_{k+1}^{(i)}\right)\right)\right) \\
&= v_0^{(i)}\hat{\tau}_0^{(i)} + \left(1-v_0^{(i)}\right)\hat{\tau}_{k+1}^{(i)} + \left(1-v_0^{(i)}\right)\left(\sum_{m=1}^{k}\left(v_m^{(i)}\left(\hat{\tau}_m^{(i)} - \hat{\tau}_{k+1}^{(i)}\right)\right) - \sum_{m=0}^{k-1}\left(v_m^{(i)}\left(\hat{\tau}_{m+1}^{(i)} - \hat{\tau}_{k+1}^{(i)}\right)\right)\right)
\end{aligned}$$

Βγάζοντας τον όρο με $m=0$ του δεύτερου αθροίσματος εκτός της παρένθεσης και παρατηρώντας ότι αυτό το δεύτερο άθροισμα μπορεί να επεκταθεί και για $m=k$ (επειδή $\hat{\tau}_{m+1}^{(i)} - \hat{\tau}_{k+1}^{(i)} = 0$ όταν $m=k$), παίρνουμε

$$\begin{aligned}
E\{\tau_i\} &= v_0^{(i)}\hat{\tau}_0^{(i)} + \left(1-v_0^{(i)}\right)\hat{\tau}_{k+1}^{(i)} - v_0^{(i)}\left(1-v_0^{(i)}\right)\left(\hat{\tau}_1^{(i)} - \hat{\tau}_{k+1}^{(i)}\right) \\
&\quad + \left(1-v_0^{(i)}\right)\left(\sum_{m=1}^{k}\left(v_m^{(i)}\left(\hat{\tau}_m^{(i)} - \hat{\tau}_{k+1}^{(i)}\right)\right) - \sum_{m=1}^{k}\left(v_m^{(i)}\left(\hat{\tau}_{m+1}^{(i)} - \hat{\tau}_{k+1}^{(i)}\right)\right)\right) \\
&= v_0^{(i)}\hat{\tau}_0^{(i)} + \left(1-v_0^{(i)}\right)\hat{\tau}_{k+1}^{(i)} - v_0^{(i)}\left(1-v_0^{(i)}\right)\left(\hat{\tau}_1^{(i)} - \hat{\tau}_{k+1}^{(i)}\right) + \left(1-v_0^{(i)}\right)\sum_{m=1}^{k}\left(v_m^{(i)}\left(\hat{\tau}_m^{(i)} - \hat{\tau}_{m+1}^{(i)}\right)\right)
\end{aligned}$$

$$= v_0^{(i)}\hat{\tau}_0^{(i)} + \left(1 - v_0^{(i)}\right)\hat{\tau}_{k+1}^{(i)} + v_0^{(i)}\left(1 - v_0^{(i)}\right)\left(\hat{\tau}_{k+1}^{(i)} - \hat{\tau}_1^{(i)}\right) - \left(1 - v_0^{(i)}\right)\sum_{m=1}^{k}\left(v_m^{(i)}\left(\hat{\tau}_{m+1}^{(i)} - \hat{\tau}_m^{(i)}\right)\right) \blacksquare$$

Πρόταση 5.2. *Η πραγματική συνάρτηση* $g(x) = x^k$ *με* $x > 0$ *και* $0 < k < 1$, *είναι γνησίως κοίλη.*

Απόδειξη: Είναι $g'(x) = k\,x^{k-1}$ και $g''(x) = k(k-1)x^{k-2} < 0$. ■

Πρόταση 5.3. *Η συνάρτηση* $\hat{u}_i(\boldsymbol{x}) = \left(1 - v_0^{(i)}\right)\sum_{m=1}^{k}\left(\Delta\hat{\tau}_m^{(i)} g_i\left(\frac{C_i + \sum_{j=1}^{m} x_j}{O_i}\right)\right)$, $i = 1,\ldots,n$, *είναι γνησίως κοίλη.*

Απόδειξη: Για κάθε $\boldsymbol{x} \neq \boldsymbol{y}$ και για κάθε $0 < h < 1$, έχουμε:

$$\hat{u}_i\left((1-h)\boldsymbol{x} + h\boldsymbol{y}\right) = \left(1 - v_0^{(i)}\right)\sum_{m=1}^{k}\left(\Delta\hat{\tau}_m^{(i)} g_i\left(\frac{C_i + \sum_{j=1}^{m}\left((1-h)x_j + hy_j\right)}{O_i}\right)\right)$$

$$= \left(1 - v_0^{(i)}\right)\sum_{m=1}^{k}\left(\Delta\hat{\tau}_m^{(i)} g_i\left((1-h)\frac{C_i + \sum_{j=1}^{m} x_j}{O_i} + h\frac{C_i + \sum_{j=1}^{m} y_j}{O_i}\right)\right)$$

$$> \left(1 - v_0^{(i)}\right)\sum_{m=1}^{k}\left(\Delta\hat{\tau}_m^{(i)}\left[(1-h)g_i\left(\frac{C_i + \sum_{j=1}^{m} x_j}{O_i}\right) + hg_i\left(\frac{C_i + \sum_{j=1}^{m} y_j}{O_i}\right)\right]\right)$$

$$= (1-h)\left(1-v_0^{(i)}\right)\sum_{m=1}^{k}\left(\Delta\hat{\tau}_m^{(i)} g_i\left(\frac{C_i+\sum_{j=1}^{m} x_j}{O_i}\right)\right)$$

$$+ h\left(1-v_0^{(i)}\right)\sum_{m=1}^{k}\left(\Delta\hat{\tau}_m^{(i)} g_i\left(\frac{C_i+\sum_{j=1}^{m} y_j}{O_i}\right)\right)$$

$$= (1-h)\,\hat{u}_i(\boldsymbol{x}) + h\,\hat{u}_i(\boldsymbol{y})$$

Η ανισότητα στο τρίτο βήμα προκύπτει από την γνήσια κοιλότητα της συνάρτησης $g_i(\cdot)$ και το γεγονός ότι όλοι οι άλλοι συντελεστές που υπάρχουν είναι θετικές ποσότητες. ■

Θεώρημα 5.1. *Η συνάρτηση καθαρού οφέλους της L1 cache i*

$$b_i\left(\boldsymbol{x}^{(i)}\right) = u_i\left(\boldsymbol{x}^{(i)}\right) - \boldsymbol{p}^T\boldsymbol{x}^{(i)} = \hat{u}_i\left(\boldsymbol{x}^{(i)}\right) - B_i - \boldsymbol{p}^T\boldsymbol{x}^{(i)}, \quad i=1,\ldots,n$$

είναι γνησίως κοίλη.

Απόδειξη: Με δεδομένη την κοιλότητα της $\hat{u}_i(\boldsymbol{x})$ προκύπτει άμεσα η γνήσια κοιλότητα της συνάρτησης καθαρού οφέλους $b_i\left(\boldsymbol{x}^{(i)}\right) = u_i\left(\boldsymbol{x}^{(i)}\right) - \boldsymbol{p}^T\boldsymbol{x}^{(i)} = \hat{u}_i\left(\boldsymbol{x}^{(i)}\right) - B_i - \boldsymbol{p}^T\boldsymbol{x}^{(i)}$, ως άθροισμα μιας κοίλης, μιας σταθερής και μιας γραμμικής συνάρτησης. Πράγματι για κάθε $\boldsymbol{x} \neq \boldsymbol{y}$ και για κάθε $0<h<1$, έχουμε:

$$b_i\left((1-h)\boldsymbol{x}+h\boldsymbol{y}\right) = \hat{u}_i\left((1-h)\boldsymbol{x}+h\boldsymbol{y}\right) - B_i + \boldsymbol{p}^{\mathrm{T}}\left((1-h)\boldsymbol{x}+h\boldsymbol{y}\right)$$

$$> (1-h)\,\hat{u}_i(\boldsymbol{x}) + h\,\hat{u}_i(\boldsymbol{y}) - B_i + (1-h)\boldsymbol{p}^{\mathrm{T}}(\boldsymbol{x}) + h\boldsymbol{p}^{\mathrm{T}}(\boldsymbol{y})$$

$$= (1-h)\,\hat{u}_i(\boldsymbol{x}) + h\,\hat{u}_i(\boldsymbol{y}) - B_i + (1-h)\boldsymbol{p}^{\mathrm{T}}(\boldsymbol{x}) + h\boldsymbol{p}^{\mathrm{T}}(\boldsymbol{y})$$

$$= (1-h)\,\hat{u}_i(\boldsymbol{x}) + h\,\hat{u}_i(\boldsymbol{y}) - (1-h)\,B_i - h\,B_i + (1-h)\boldsymbol{p}^{\mathrm{T}}(\boldsymbol{x}) + h\boldsymbol{p}^{\mathrm{T}}(\boldsymbol{y})$$

$$= (1-h)\,b_i(\boldsymbol{x}) + h\,b_i(\boldsymbol{y})\ \blacksquare$$

Πρόταση 5.4. *Υπάρχει μια μοναδική τιμή χρέωσης, έστω* p_j^*, *ώστε* $y_j\left(p_j^*, \boldsymbol{p}_{-j}\right) = C_{2,j}$, *για κάθε* $j = 1,\ldots,k$ *και για κάθε* $\boldsymbol{p}_{-j} \in S_{-j}$.

Απόδειξη: Η συνεχής και παραγωγίσιμη συνάρτηση $y_j\left(p_j, \boldsymbol{p}_{-j}\right)$, $j = 1,...,k$, είναι γνησίως φθίνουσα. Πράγματι

$$\frac{\partial y_j\left(p_j, \boldsymbol{p}_{-j}\right)}{\partial p_j} = \sum_{i=1}^{n}\left(-\frac{1}{a_i}\right)A_j^{(i)}\left(p_j - p_{j+1}\right)^{\frac{1}{a_i}-1} + \sum_{i=1}^{n}\left(-\frac{1}{a_i}\right)A_{j-1}^{(i)}\left(p_{j-1} - p_j\right)^{\frac{1}{a_i}-1}$$

$$= -\left[\sum_{i=1}^{n}\frac{1}{a_i}A_j^{(i)}\left(p_j - p_{j+1}\right)^{\frac{1}{a_i}-1} + \sum_{i=1}^{n}\frac{1}{a_i}A_{j-1}^{(i)}\left(p_{j-1} - p_j\right)^{\frac{1}{a_i}-1}\right] < 0$$

Το ζητούμενο της Πρότασης είναι άμεση συνέπεια της γνήσιας μονοτονίας της $y_j\left(p_j, \boldsymbol{p}_{-j}\right)$. ■

Πρόταση 5.5. *Η συνάρτηση κέρδους* $\pi_j\left(p_j, \boldsymbol{p}_{-j}\right)$ *της L2 cache* j, *παίρνει την μέγιστη τιμή της στο* p_j^*, *για κάθε* $j = 1,\ldots,k$ *και για κάθε* $\boldsymbol{p}_{-j} \in S_{-j}$.

Απόδειξη: Θα ξεκινήσουμε εξετάζοντας την μονοτονία της συνάρτησης κέρδους $\pi_j\left(p_j, \boldsymbol{p}_{-j}\right)$ της L2 cache j.

- Όταν $p_j < p_j^*$ και με δεδομένο ότι η συνάρτηση $y_j\left(p_j, \boldsymbol{p}_{-j}\right)$ είναι γνησίως φθίνουσα, έχουμε $y_j\left(p_j, \boldsymbol{p}_{-j}\right) > y_j\left(p_j^*, \boldsymbol{p}_{-j}\right) = C_{2,j}$, οπότε η συνάρτηση κέρδους δίνεται από την

  $$\pi_j\left(p_j, \boldsymbol{p}_{-j}\right) = p_j \cdot C_{2,j}$$

  και

  $$\frac{\partial \pi_j\left(p_j, \boldsymbol{p}_{-j}\right)}{\partial p_j} = C_{2,j} > 0$$

  δηλαδή η $\pi_j\left(p_j, \boldsymbol{p}_{-j}\right)$ είναι γνησίως αύξουσα για $p_j < p_j^*$.

- Όταν $p_j > p_j^*$, και με δεδομένο ότι η συνάρτηση $y_j\left(p_j, \boldsymbol{p}_{-j}\right)$ είναι γνησίως φθίνουσα, έχουμε $y_j\left(p_j, \boldsymbol{p}_{-j}\right) < y_j\left(p_j^*, \boldsymbol{p}_{-j}\right) = C_{2,j}$, οπότε η συνάρτηση κέρδους δίνεται από την

$$\pi_j\left(p_j,\boldsymbol{p}_{-j}\right)=p_j\cdot y_j\left(p_j,\boldsymbol{p}_{-j}\right)=p_j\left[\sum_{i=1}^{n}A_j^{(i)}\left(p_j-p_{j+1}\right)^{-\frac{1}{a_i}}-\sum_{i=1}^{n}A_{j-1}^{(i)}\left(p_{j-1}-p_j\right)^{-\frac{1}{a_i}}\right]$$

και

$$\frac{\partial\pi_j\left(p_j,\boldsymbol{p}_{-j}\right)}{\partial p_j}=y_j\left(p_j,\boldsymbol{p}_{-j}\right)+p_j\cdot\frac{\partial y_j\left(p_j,\boldsymbol{p}_{-j}\right)}{\partial p_j}$$

$$=\left[\sum_{i=1}^{n}A_j^{(i)}\left(p_j-p_{j+1}\right)^{-\frac{1}{a_i}}-\sum_{i=1}^{n}A_{j-1}^{(i)}\left(p_{j-1}-p_j\right)^{-\frac{1}{a_i}}\right]$$

$$+p_j\left[\sum_{i=1}^{n}\left(-\frac{1}{a_i}\right)A_j^{(i)}\left(p_j-p_{j+1}\right)^{-\frac{1}{a_i}-1}+\sum_{i=1}^{n}\left(-\frac{1}{a_i}\right)A_{j-1}^{(i)}\left(p_{j-1}-p_j\right)^{-\frac{1}{a_i}-1}\right]$$

$$=\sum_{i=1}^{n}A_j^{(i)}\left(p_j-p_{j+1}\right)^{-\frac{1}{a_i}}-\sum_{i=1}^{n}A_{j-1}^{(i)}\left(p_{j-1}-p_j\right)^{-\frac{1}{a_i}}$$

$$-\sum_{i=1}^{n}\frac{1}{a_i}A_j^{(i)}\frac{p_j}{p_j-p_{j+1}}\left(p_j-p_{j+1}\right)^{-\frac{1}{a_i}}-\sum_{i=1}^{n}\frac{1}{a_i}A_{j-1}^{(i)}\frac{p_j}{p_{j-1}-p_j}\left(p_{j-1}-p_j\right)^{-\frac{1}{a_i}}$$

$$=\sum_{i=1}^{n}\left(A_j^{(i)}\left(p_j-p_{j+1}\right)^{-\frac{1}{a_i}}\left(1-\frac{1}{a_i}\cdot\frac{p_j}{p_j-p_{j+1}}\right)\right)$$

$$-\sum_{i=1}^{n}\left(A_{j-1}^{(i)}\left(p_{j-1}-p_j\right)^{-\frac{1}{a_i}}\left(1+\frac{1}{a_i}\cdot\frac{p_j}{p_{j-1}-p_j}\right)\right)$$

όμως

$p_j>0$, για κάθε $j=1,\ldots,k$

$$p_j-p_{j+1}=\left(1-v_0^{(i)}\right)\left(\frac{1}{O_i}\right)^{1-a_i}\left(1-a_i\right)\Delta\hat{\tau}_j^{(i)}\left(C_i+\sum_{s=1}^{j}x_s^{(i)}\right)^{-a_i}>0,\ \text{για κάθε } j=1,\ldots,k$$

οπότε

$$p_j>p_j-p_{j+1}\Leftrightarrow\frac{p_j}{p_j-p_{j+1}}>1$$

επίσης

$$0<a_i<1\Leftrightarrow\frac{1}{a_i}>1$$

άρα

$$\frac{1}{a_i}\cdot\frac{p_j}{p_j-p_{j+1}}>1\Leftrightarrow-\frac{1}{a_i}\cdot\frac{p_j}{p_j-p_{j+1}}<-1\Leftrightarrow1-\frac{1}{a_i}\cdot\frac{p_j}{p_j-p_{j+1}}<0$$

και

$$1+\frac{1}{a_i}\cdot\frac{p_j}{p_{j-1}-p_j}>0$$

οπότε

$$\frac{\partial\pi_j\left(p_j,\boldsymbol{p}_{-j}\right)}{\partial p_j}=\sum_{i=1}^{n}\left(A_j^{(i)}\left(p_j-p_{j+1}\right)^{-\frac{1}{a_i}}\left(1-\frac{1}{a_i}\cdot\frac{p_j}{p_j-p_{j+1}}\right)\right)$$

$$-\sum_{i=1}^{n}\left(A_{j-1}^{(i)}\left(p_{j-1}-p_j\right)^{-\frac{1}{a_i}}\left(1+\frac{1}{a_i}\cdot\frac{p_j}{p_{j-1}-p_j}\right)\right)<0$$

δηλαδή η $\pi_j\left(p_j,\boldsymbol{p}_{-j}\right)$ είναι γνησίως φθίνουσα για $p_j>p_j^*$.

Από την παραπάνω μελέτη μονοτονίας της συνάρτησης κέρδους $\pi_j\left(p_j,\boldsymbol{p}_{-j}\right)$ της L2 cache j, και παρατηρώντας ότι αυτή είναι μια συνεχής συνάρτηση, προκύπτει ότι η $\pi_j\left(p_j,\boldsymbol{p}_{-j}\right)$ παίρνει την μέγιστη τιμή της στο p_j^*, $j=1,\ldots,k$. ■