

Χαρακτηρισμός Παγκόσμιου Ιστού

Μ. Δικαιάκος

ΕΠΛ602

Μετρήσεις Διακίνησης Ιστού

Web Traffic Management

- ❖ Βήματα:
 - ❖ Παρακολούθηση (monitoring) των διακινήσεων Ιστού σε κάποια τοποθεσία.
 - ❖ Δημιουργία εγγραφών μετρήσεων σε κάποιο **μορφότυπο** και αποθήκευσή τους σε **αρχεία απογραφής (logs)**.
 - ❖ Προ-επεξεργασία των αρχείων απογραφής για περαιτέρω ανάλυση.
- ❖ Τεχνικές για παρακολούθηση διακίνησης Ιστού:
 - ❖ Παρακολούθηση **πελάτη** (client monitoring).
 - ❖ Παρακολούθηση **διαμεσολαβητή** (proxy monitoring).
 - ❖ Παρακολούθηση **εξυπηρετητή** (server monitoring).
 - ❖ Παρακολούθηση **πακέτων** (packet monitoring).
 - ❖ **Ενεργές μετρήσεις** (active measurements).

Κίνητρα Μετρήσεων Ιστού

- ❖ Κίνητρα παρόχων περιεχομένου:
 - ❖ Διερεύνηση της σημασιολογίας των επισκέψεων.
 - ❖ Βελτιστοποίηση διοργάνωσης περιεχομένου.
 - ❖ Επανασχεδιασμός ιστιακών τόπων.
- ❖ Κίνητρα εταιρειών φιλοξενίας ιστιακών τόπων:
 - ❖ Στατιστική ανάλυση διακίνησης.
 - ❖ Προγραμματισμός πόρων.
 - ❖ Διαρρύθμιση εξυπηρετητών.
- ❖ Κίνητρα για διαχειριστές δικτύων (net-ops):
 - ❖ Συγκέντρωση στατιστικών για σχεδιασμό και βελτιστοποίηση δικτύου.
 - ❖ Έγκαιρη ανακάλυψη εισβολών.
- ❖ Κίνητρα για ερευνητές:
 - ❖ Ανάλυση δυνατοτήτων και χρήσης πρωτοκόλλου HTTP.
 - ❖ Ανάλυση της δυναμικής της διακίνησης στο Διαδίκτυο.
 - ❖ Ανάπτυξη μοντέλων διακίνησης, προσομοιωτών κλπ.
 - ❖ Web Mining

Google Analytics

Site Usage



477 [Visits](#)



972 [Pageviews](#)



2.04 [Pages/Visit](#)



59.54% [Bounce Rate](#)



00:02:10 [Avg. Time on Site](#)



37.95% [% New Visits](#)

Visitors Overview



216 Visitors

[view report](#)

Map Overlay



[view report](#)

Traffic Sources Overview



■ **Search Engines**
245 (51.36%)

■ **Referring Sites**
153 (32.08%)

■ **Direct Traffic**
79 (16.56%)

[view report](#)

Content Overview

Pages	Pageviews	% Pageviews
/index.php	454	46.71%
/people.html	122	12.55%
/projects.html	83	8.54%
/activities.php	82	8.44%
/papers.html	63	6.48%

[view report](#)

Google Analytics

Visitors Overview

Oct 13, 2007 - Nov 12, 2007 ▾


Export ▾ Email Add to Dashboard




216 people visited this site

 **477** [Visits](#)


 **216** [Absolute Unique Visitors](#)

 **972** [Pageviews](#)

 **2.04** [Average Pageviews](#)

 **00:02:10** [Time on Site](#)

 **59.54%** [Bounce Rate](#)

 **37.95%** [New Visits](#)

Visitor Segmentation

 Visitors Profile: [languages](#), [network locations](#), [user defined](#)

☐ Browser Profile: [browsers](#), [operating systems](#), [browser and operating systems](#), [screen colors](#), [screen resolutions](#), [java support](#), [Flash](#)

 [Map Overlay](#)
Geolocation visualization

Google Analytics Information

Map Overlay

Oct 13, 2007 - Nov 12, 2007 ▾

Export ▾ Email Add to Dashboard

Visits ▾

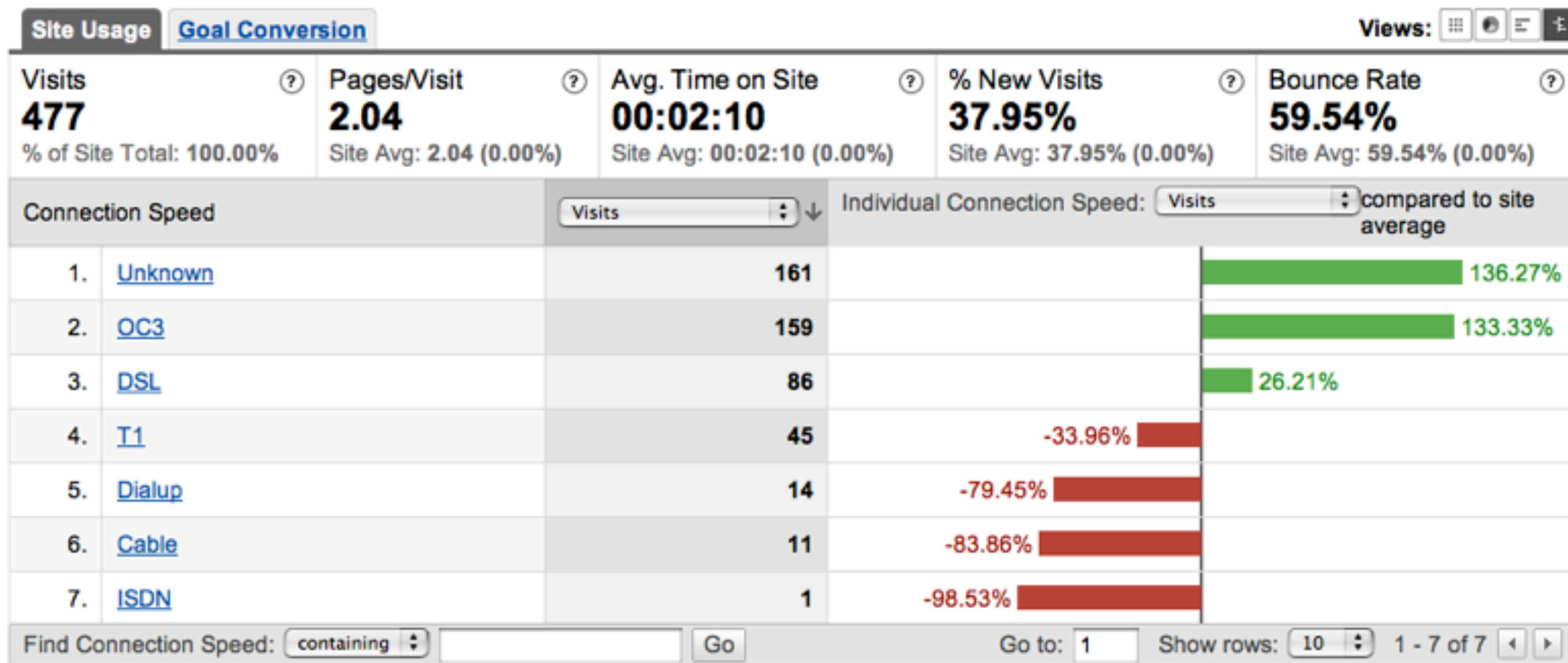


477 visits came from 104 cities

Detail Level: City | [Country/Territory](#) | [Sub Continent Region](#) | [Continent](#) Segment: [Choose...](#) ▾

Google Analytics Information

477 visits used 7 connection speeds



All traffic sources sent a total of 477 visits



16.56% [Direct Traffic](#)



32.08% [Referring Sites](#)



51.36% [Search Engines](#)



■ **Search Engines**
245 (51.36%)

■ **Referring Sites**
153 (32.08%)

■ **Direct Traffic**
79 (16.56%)

Top Traffic Sources

Referring sites sent 153 visits via 26 sources

Segment: [Source](#) ▾

Site Usage

[Goal Conversion](#)

Views:

Visits 153 % of Site Total: 32.08%	Pages/Visit 2.44 Site Avg: 2.04 (19.96%)	Avg. Time on Site 00:03:12 Site Avg: 00:02:10 (47.74%)	% New Visits 33.99% Site Avg: 37.95% (-10.43%)	Bounce Rate 47.06% Site Avg: 59.54% (-20.96%)
---	---	---	---	--

Source	Visits ↓	Pages/Visit	Avg. Time on Site	% New Visits	Bounce Rate
1. cs.ucy.ac.cy	54	2.43	00:01:20	14.81%	53.70%
2. cygrid.org.cy	33	3.27	00:06:41	18.18%	21.21%
3. www2.cs.ucy.ac.cy	20	1.90	00:03:46	5.00%	60.00%
4. grid.ucy.ac.cy	8	3.50	00:12:24	25.00%	25.00%
5. images.google.com	7	1.86	00:00:18	85.71%	28.57%
6. its.cs.ucy.ac.cy	4	2.25	00:01:11	75.00%	50.00%
7. images.google.co.uk	3	2.00	00:00:12	100.00%	66.67%

Google Analytics Setup

```
<script  
  src="http://www.google-analytics.com/urchin.js"  
  type="text/javascript">  
</script>
```

```
<script type="text/javascript">  
  _uacct="UA-xxxx-x";  
  urchinTracker();  
</script>
```

Τεχνικές Απογραφής

❖ Απογραφή Εξυπηρετητών Ιστού:

- ❖ Απογραφή αιτημάτων πελατών, που αποστέλλονται στον εξυπηρετητή.
- ❖ Απογράφονται: πληροφορίες για πελάτη (διεύθυνση IP), χρονοσφραγίδα αιτήματος, πληροφορίες για μήνυμα αιτήματος και απόκρισης.
- ❖ Για οικονομία χώρου: IP, Method, URI, Response code
- ❖ Ποιά(ές) χρονοσφραγίδα(ες);
- ❖ Ποιά είναι η ακρίβεια της χρονοσφραγίδας (1sec).

❖ Πως ανακαλύπτουμε τα **πρότυπα** (patterns) συμπεριφοράς των χρηστών μέσα από τα αρχεία απογραφής;

- ❖ Διαμεσολαβητές, απομνημόνευση στον πελάτη, κοινοχρησία μηχανών, δυναμικά IP.
- ❖ Αποφυγή απομνημόνευσης στο δίκτυο/πελάτες: **cache busting**

Τεχνικές Απογραφής

- ❖ **Απογραφή Διαμεσολαβητών Ιστού:**
 - ❖ Απογραφή των αιτημάτων που διακινούνται μέσω του διαμεσολαβητή.
 - ❖ Πιο λεπτομερείς πληροφορίες για τα πρότυπα συμπεριφοράς πελατών.
 - ❖ Τι συμπεράσματα μπορούμε να εξάγουμε από αναλύσεις αρχείων απογραφής διαμεσολαβητών;
 - ❖ Ποιοί παράγοντες επηρεάζουν την ακρίβεια των συμπερασμάτων μας;
- ❖ **Απογραφή Λειτουργίας Πελατών (φυλλομετρητών):**
 - ❖ Πλεονεκτήματα.
 - ❖ Δυσκολίες.

Τεχνικές Απογραφής

- ❖ Παρακολούθηση πακέτων:
 - ❖ Στο επίπεδο του TCP/IP.
 - ❖ Απογραφή πακέτων IP και ανασύσταση διακίνησης HTTP.
 - ❖ Πού και πώς;
 - ❖ Μεγαλύτερη ακρίβεια χρονοσφραγίδων.
 - ❖ Μειονεκτήματα.
- ❖ Μειονεκτήματα αναλύσεων βασισμένων σε αρχεία απογραφής: **έλλειψη επαρκούς πληροφορίας** για ανάλυση της επίδοσης που υφίστανται οι χρήστες.
 - ❖ Π.χ. πως αναλύεται η υστέρηση στην ανάκτηση μιας σύνθετης ιστοσελίδας;
 - ❖ Πως μπορούμε να κάνουμε μια συστηματική διερεύνηση της επίδοσης του Ιστού μέσα από τα αρχεία απογραφής;
 - ❖ Οι τεχνικές παθητικής παρακολούθησης δεν μας δίνουν απαντήσεις σε αυτές τις κατευθύνσεις...

Τεχνικές Απογραφής

- ❖ **Ενεργές Μετρήσεις** (active measurements): βασίζονται στη χρήση ενός **συνθετικού διεκπεραιωτή πελάτη** (user agent), ο οποίος εγείρει αιτήματα και απογράφει πληροφορίες για απαντήσεις (χρονοσφραγίδες, επικεφαλίδες HTTP).
- ❖ Βασικά ερωτήματα:
 - ❖ Πού θα τοποθετηθεί ο διεκπεραιωτής;
 - ❖ Τι αιτήματα πρέπει να δημιουργεί ο διεκπεραιωτής;
 - ❖ Ποιές μετρήσεις χρειάζεται να συλλέγονται;

Αρχεία Απογραφής

- ❖ Μορφότυποι Αρχείων Απογραφής Διαμεσολαβητών-Εξυπηρετητών.
- ❖ **Common Log Format (CLF):**
 - ❖ **Remote host:** IP or DNS (through reverse DNS lookup)
 - ❖ **Remote identity:** owner associated with TCP connection on client machine
 - ❖ **Authenticated user:** name provided by user for authentication and carried by the **Authorization** header of the HTTP request
 - ❖ **Time:** χρονοσφραγίδα λήψης αιτήματος.
 - ❖ **Request:** μέθοδος αιτήματος, αιτούμενο URI και έκδοση πρωτοκόλλου.
 - ❖ **Response code:** τριψήφιος κωδικός απάντησης HTTP.
 - ❖ **Content length:** αριθμός ψηφίων που αφορά στην απόκριση του εξυπηρετητή στο απογραφόμενο αίτημα

Αρχεία Απογραφής

- ❖ Το **Extended Common Log Format (ECLF)** περιλαμβάνει μια σειρά άλλων πεδίων, πέραν αυτών του CLF:
 - ❖ **User agent**: information on user agent software
 - ❖ **Referer**: URI from which request-URI was obtained
 - ❖ **Request** processing time
 - ❖ **Request header size**
 - ❖ κλπ

Προ-επεξεργασία (Pre-processing)

- ❖ Σάρωση δεδομένων για εντοπισμό λανθασμένων εγγραφών αρχείου απογραφής.
- ❖ Κοσκίνισμα δεδομένων για διαγραφή άχρηστων εγγραφών.
- ❖ Μετασχηματισμός των δεδομένων σε μορφότυπο που διευκολύνει την ανάλυση.

Χαρακτηρισμός Φορτίου Εργασίας Ιστού (Web Workload Characterization)

- ❖ **Workload** (φορτίο εργασίας): το σύνολο όλων των εισροών που δέχεται ένα σύστημα μέσα σε μια χρονική περίοδο.
- ❖ **Ποσοτικά Μοντέλα Φορτίου Εργασίας** (quantitative models): αποτελούνται από **συλλογές παραμέτρων** που αναπαριστούν βασικές ιδιότητες ενός φορτίου εργασίας, οι οποίες:
 - ❖ Επηρεάζουν τη δέσμευση πόρων.
 - ❖ Επηρεάζουν την επίδοση του συστήματος.
 - ❖ Μπορούν να κατευθύνουν μελέτες επίδοσης (benchmarking, προσομοιώσεις κλπ).
- ❖ **Ανάπτυξη μοντέλων φορτίου εργασίας:**
 - ❖ Καθορισμός σημαντικών παραμέτρων μοντέλου.
 - ❖ Ανάλυση μετρήσεων για ποσοτικοποίηση των παραμέτρων.
 - ❖ **Επικύρωση** του μοντέλου μέσω σύγκρισης με την πραγματικότητα.

Μοντέλα Φορτίου Εργασίας Ιστού

- ❖ Πολύ μεγάλη μεταβλητότητα χαρακτηριστικών.
- ❖ Μετρήσεις διαφέρουν από περίπτωση σε περίπτωση.
- ❖ ... ωστόσο κάποιες στατιστικές ιδιότητες παρατηρούνται σε διαφορετικές μελέτες και εκφράζουν γενικότερα χαρακτηριστικά της *δυναμικής του Ιστού*.
- ❖ Βασικές (στατιστικές) ιδιότητες ιστιακών φορτίων εργασίας αφορούν σε:
 - ❖ **Χαρακτηριστικά μηνυμάτων HTTP**: συχνότητα μεθόδων και αποκρίσεων, ρυθμός αφίξεων, κατανομή αφίξεων
 - ❖ **Χαρακτηριστικά πόρων**: μορφότυπος περιεχομένου, μέγεθος, δημοτικότητα, ρυθμός αλλαγής, αριθμός ενσωματωμένων πόρων.
 - ❖ **Συμπεριφορά χρηστών**: χρονικό διάστημα ανάμεσα σε διαδοχικά αιτήματα, διάρκεια συνόδου, πρότυπα επισκέψεων.

Εφαρμογές Μοντέλων Φορτίου Εργασίας

- ❖ Εντοπισμός προβλημάτων επίδοσης (performance analysis, bottlenecks' identification)
- ❖ Επιμέτρηση επίδοσης δομοστοιχείων λογισμικού (benchmarking).
- ❖ Προγραμματισμός χωρητικότητας (capacity planning).

Κριτήρια Επιλογής Παραμέτρων

- ❖ Αποσύνδεση από τις λεπτομέρειες υλοποίησης του συστήματος που θέλουμε να διερευνήσουμε. Έτσι, προτείνεται η επιλογή παραμέτρων όπως: User-perceived latency, server throughput, packet loss rate
- ❖ Ορθό επίπεδο λεπτομέρειας: π.χ. ο χρόνος μεταξύ δύο αιτημάτων αλλά όχι τα περιεχόμενα των επικεφαλίδων αιτημάτων.
- ❖ Ανεξαρτησία παραμέτρων μεταξύ τους: αλληλεξαρτήσεις μεταξύ παραμέτρων δυσχεραίνουν την προσπάθεια αποτύπωσης των χαρακτηριστικών εργασίας φορτίου μέσω απλών μοντέλων.

Παραδείγματα Παραμέτρων Φορτίων Ιστού

- ❖ Πρωτόκολλο:
 - ❖ Μέθοδος Αιτήματος
 - ❖ Κώδικας Απόκρισης
- ❖ Πόροι:
 - ❖ Τύπος περιεχομένου
 - ❖ Μέγεθος Πόρου
 - ❖ Μέγεθος Απόκρισης
 - ❖ Δημοτικότητα
 - ❖ Συχνότητα Αλλαγών
 - ❖ Χρονική Τοπικότητα
 - ❖ Αριθμός ενσωματωμένων πόρων
- ❖ Χρήστες:
 - ❖ Μεσοδιαστήματα συνόδων
 - ❖ Αριθμός κλικ ανά σύνοδο
 - ❖ Μεσοδιάστημα αφίξεως αιτημάτων

Στατιστική Περιγραφή Παραμέτρων Φορτίων Ιστού

- ❖ Μέση τιμή, ενδιάμεσος και μεταβλητότητα:
 - ❖ Η χρησιμότητα της μέσης τιμής.
 - ❖ Συσχέτιση μέσης τιμής και ενδιαμέσου.
 - ❖ Χρήση της μεταβλητότητας (τυπικής απόκλισης):
 - ❖ Μικρή μεταβλητότητα σημαίνει ότι οι παράμετροι κυμαίνονται κοντά στον μέσο όρο.
 - ❖ Μεγάλη μεταβλητότητα σημαίνει ότι υπάρχουν παράμετροι με μεγάλη διαφοροποίηση από τον μέσο όρο.
- ❖ Εξυπηρετητής σερβίρει μηνύματα μεγέθους:
 - ❖ 4100, 4700, 4200, 20000, 4000 bytes
 - ❖ Average: 7400 bytes
 - ❖ Median: 4200 bytes

Κατανομές Πιθανοτήτων Workloads

- ❖ Για την περιγραφή των παραμέτρων φορτίων εργασίας Ιστού, χρησιμοποιούμε εξισώσεις κατανομών. Π.χ.: $F(x) = \Pr[X > x] = e^{-\lambda x}$ για εκθετική κατανομή με μέσο όρο $1/\lambda$ και $x \geq 0$.
- ❖ Πως επιλέγουμε μια κατανομή;
- ❖ Πως ελέγχουμε την **εγκυρότητα** της επιλογής μας;
 1. Πρώτα προσαρμόζουμε (fitting) την κατανομή στα δεδομένα των μετρήσεών μας – π.χ το λ θα εξισωθεί με τη μέση τιμή των μετρήσεών μας.
 2. Εφαρμόζουμε στατιστικό έλεγχο για να συγκρίνουμε πόσο “κοντά” είναι τα δεδομένα μας με την κατανομή που υιοθετούμε (“goodness” of the fit test).
- ❖ Σε αρκετές περιπτώσεις δεν υπάρχει γνωστή κατανομή που να προσαρμόζεται στα δεδομένα μας – έτσι χρησιμοποιούμε διαφορετικές κατανομές για διαφορετικά τμήματα των δεδομένων.

Διερεύνηση Χαρακτηριστικών Πρωτοκόλλου

- ❖ Αιτήματα: συντριπτική πλειοψηφία από αιτήματα **GET**.
- ❖ Μικρή ανοδική τάση αύξησης των **POST**.
- ❖ 75-90% των αποκρίσεων έχουν κωδικό **200 OK**.
- ❖ 10-30% των αποκρίσεων έχουν κωδικό **304 Not Modified**. Η συχνότητα του 304 εξαρτάται από:
 - ❖ Το ποσοστό του απομνημονευόμενου περιεχομένου στον εξυπηρετητή.
 - ❖ Την συχνότητα των αλλαγών στο περιεχόμενο.
 - ❖ Την πιθανότητα αιτημάτων για απομνημονευμένο περιεχομένο.

Τοπικά αποτελέσματα (2001-2002)

www.cs.ucy.ac.cy

- ❖ GET: 99%
- ❖ 2xx: 60.25%
- ❖ 3xx: 30.55%
- ❖ 304 (only): 28.48%
- ❖ 4xx: 9.19%
- ❖ 5xx: 0.01%

www.ucy.ac.cy

- ❖ GET: 99%
- ❖ 2xx: 52,2%
- ❖ 3xx: 25,97%
- ❖ 304 (only): 23,77%
- ❖ 4xx: 21,82%
- ❖ 5xx: 0.12%

Χαρακτηριστικά Ιστιακών Πόρων

- ❖ Μορφότυποι περιεχομένου: στατιστικές σχετικά με τον τύπο των μεταφερόμενων πόρων μας δίνουν μια εικόνα του είδους των δεδομένων που δημοσιεύονται στον Ιστό.
- ❖ Συντριπτική πλειοψηφία μεταφερόμενου περιεχομένου:
 - ❖ Κείμενο (**text/html**, **text/plain**)
 - ❖ Εικόνες (**image/jpeg**, **image/gif**)
- ❖ Υπόλοιπο περιεχόμενο: **ps**, **pdf**, **multimedia**.
- ❖ Η εμφάνιση νέων εφαρμογών μπορεί να έχει δραστικές και απρόσμενες επιπτώσεις στην δημοτικότητα μορφοτύπων.

Διερεύνηση Χαρακτηριστικών Πόρων

- ❖ Ο συνδυασμός περιεχομένου που παρέχεται από έναν εξυπηρετητή εξαρτάται:
 - ❖ Από τους ιστιακούς τόπους που φιλοξενεί ο εξυπηρετητής.
 - ❖ Από τους χρήστες και τις συνδέσεις τους.

www.cs.ucy.ac.cy

- ❖ HTML, text: 58.29%
- ❖ Images: 35.2%
- ❖ Sound, Video: 0.14%
- ❖ PS, PDF: 4.74%
- ❖ Zipped: 0.22%

www.ucy.ac.cy

- ❖ HTML, text: 89.93%
- ❖ Images: 1.02%
- ❖ Sound, Video: 0.18%
- ❖ PS, PDF: 3.15%
- ❖ Zipped: 0.23%

Ποσοστά επί τοις εκατό των αιτημάτων

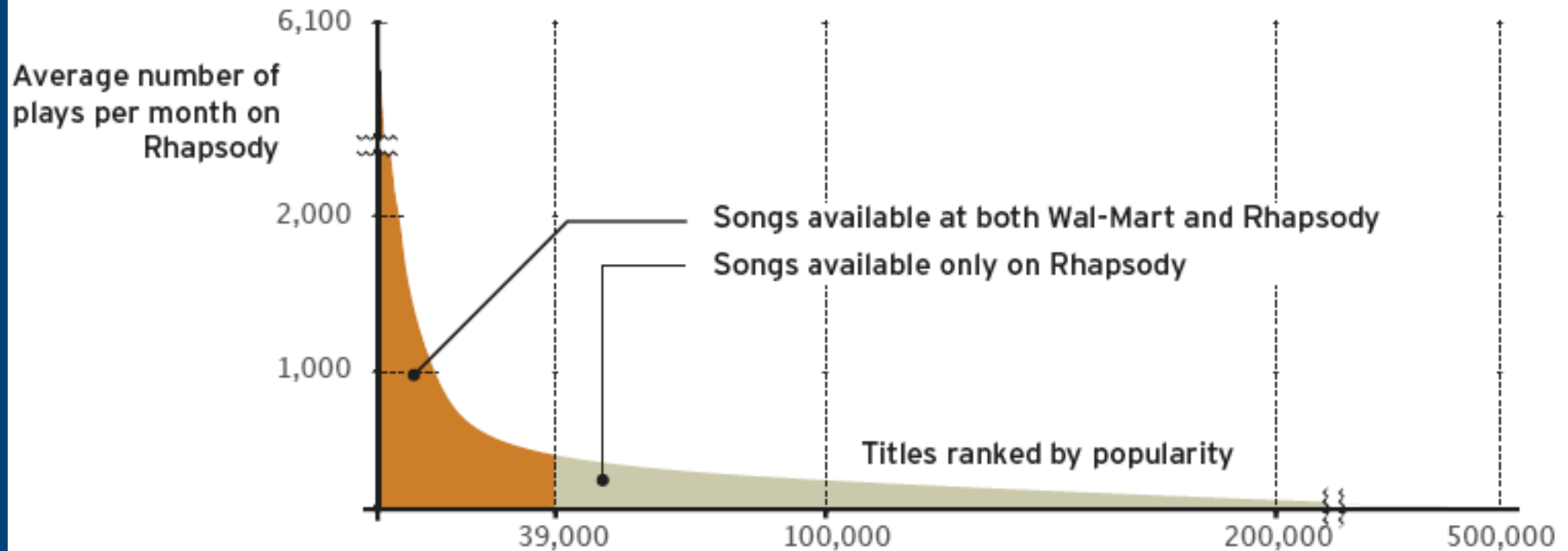
Μεγέθη Ιστιακών Πόρων

- ❖ Μέσο μέγεθος αρχείων HTML: 4-8 KB
- ❖ Μέσο μέγεθος αρχείων εικόνας: 14 KB
- ❖ Ενδιάμεσο μέγεθος αρχείων HTML: 2 KB
- ❖ Παρατηρήσεις:
 - ❖ Υπάρχει πολύ **μεγάλη μεταβλητότητα** στο μέγεθος των αρχείων HTML (και των εικόνων).
 - ❖ Δεδομένης της πολύ μεγάλης μεταβλητότητας, ο υπολογισμός και η χρήση παραμέτρων όπως ο μέρος όρος ή η μεταβλητότητα δεν μας δίνουν πολύ χρήσιμη πληροφορία.
 - ❖ Χρειαζόμαστε λοιπόν **συναρτήσεις κατανομής** και/ή **ενδιάμεσους** (medians) για να μοντελοποιήσουμε το μέγεθος των ιστιακών πόρων.

Μοντελοποίηση Μεγέθους Ιστιακών Πόρων

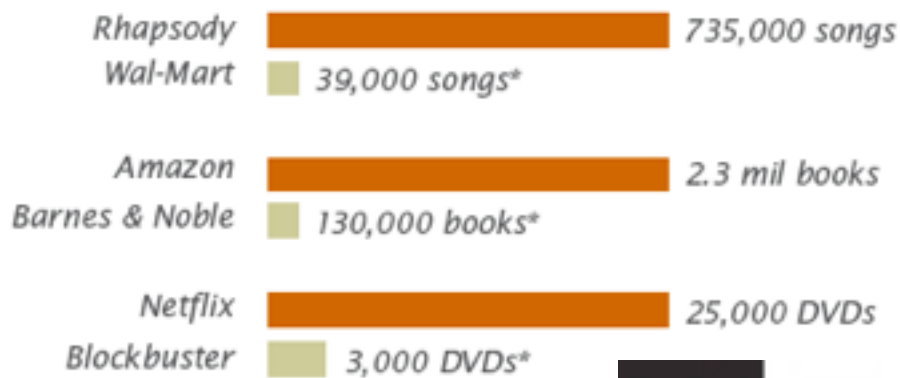
- ❖ Η υψηλή μεταβλητότητα στα μεγέθη των πόρων του Ιστού απεικονίζεται από την κατανομή Pareto:
$$F(x) = \Pr[X > x] = (k / x)^a, x \geq k$$
 - ❖ a : shape parameter ($2 > a > 0$)
 - ❖ k : scale parameter
 - ❖ Mean = $ka/(a-1)$, for $a > 1$
- ❖ Η Pareto είναι κατανομή **heavy-tailed**— ο όρος **ευτραφής ουρά** αναφέρεται στο πόσο αργά μειώνεται η $F(x)$ καθώς αυξάνεται το x .
- ❖ Μια κατανομή καλείται **heavy-tailed** (ή **long-tailed**) αν η ουρά της μειώνεται πιο αργά από την ουρά οποιασδήποτε εκθετικής κατανομής.
- ❖ Σε λογαριθμικό διάγραμμα, μια heavy-tailed κατανομή **μειώνεται γραμμικά** με κλίση η οποία σχετίζεται με την τιμή της a .
- ❖ Όσο πιο μικρή είναι η a , τόσο μεγαλύτερη μεταβλητότητα υπάρχει στα μεγέθη των ιστοικών πόρων.
- ❖ Οι περισσότερες μελέτες έχουν βρεί ότι **$1.5 > a > 1.0$**

Heavy-tailed distributions



TOTAL INVENTORY

* inventory in a typical store



THE NEW GROWTH MARKET

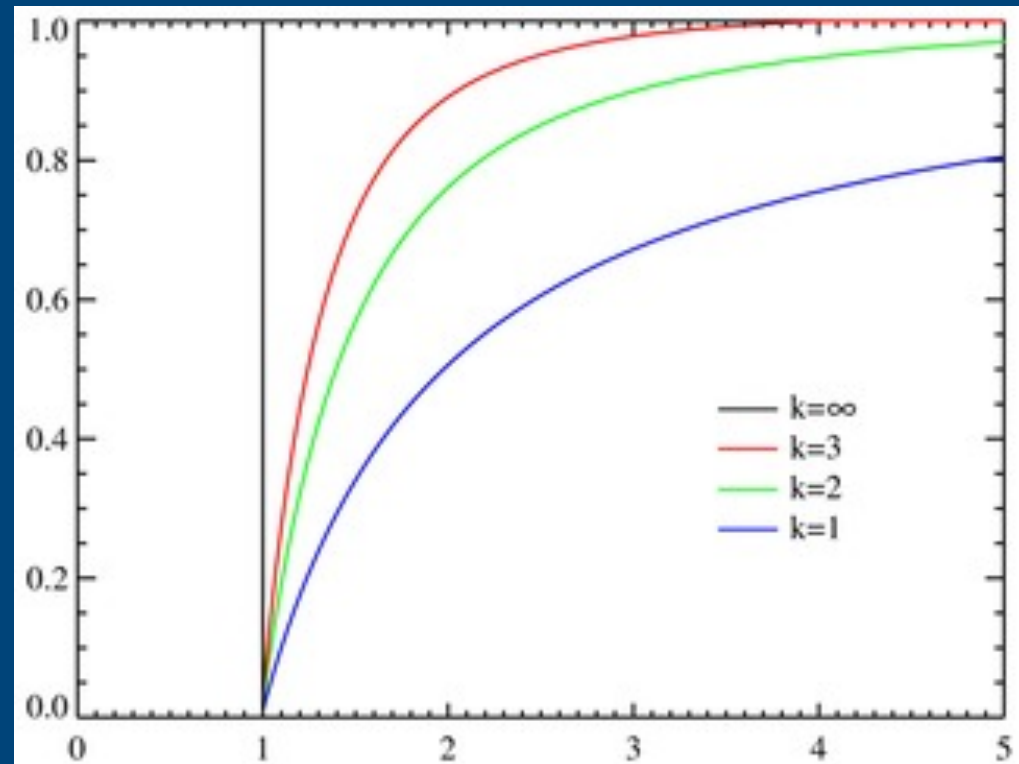
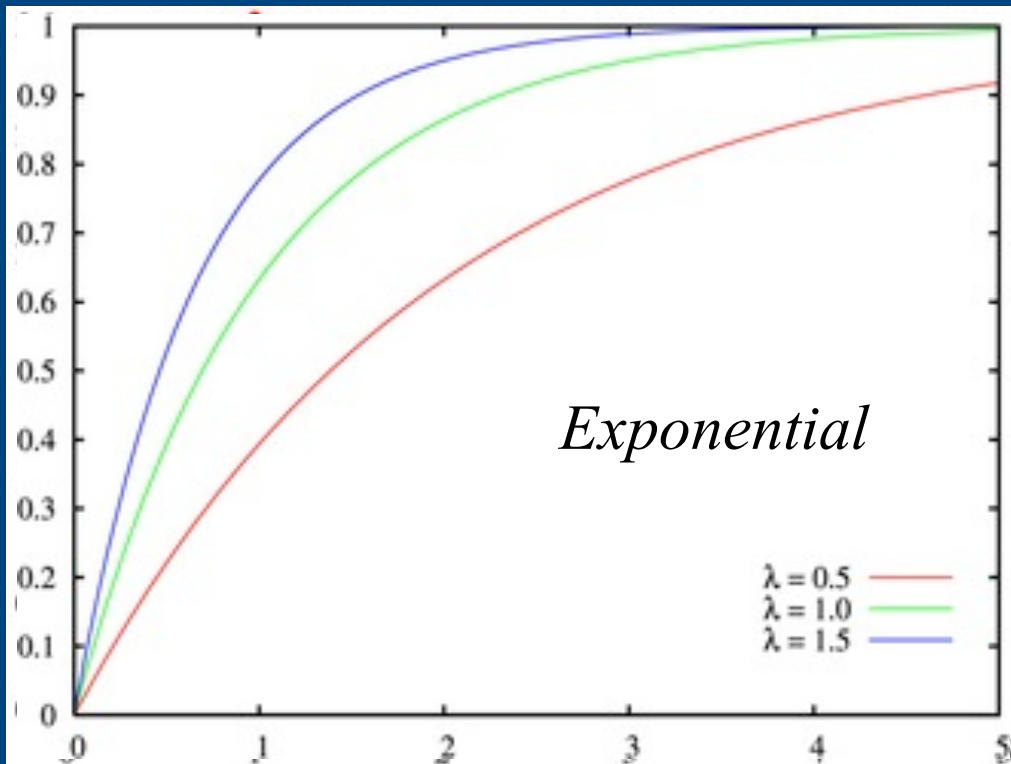
Obscure products you can't get anywhere but online

Orange — product not available in offline retail stores (% total sales)



WIRED

Pareto PMF και PDF



Μοντελοποίηση Μεγέθους Ιστιακών Πόρων

- ❖ Η Pareto δεν είναι πάντοτε αρκετή για να περιγράψει ολόκληρη την κατανομή των μεγεθών ιστιακών πόρων.
- ❖ Κυρίως χρησιμοποιείται με επιτυχία διότι η ουρά της περιγράφει την μεταβλητότητα των μεγεθών των **μεγάλων** ιστιακών πόρων.
- ❖ Τα μεγέθη μικροτέρων ιστιακών πόρων περιγράφονται καλύτερα από την **lognormal** κατανομή, μια κατανομή δηλαδή της οποίας ο λογάριθμος ακολουθεί την κανονική κατανομή.
- ❖ Επομένως, για την μοντελοποίηση του μεγέθους ιστιακών πόρων χρησιμοποιούμε **υβριδικές** κατανομές.

Μοντελοποίηση Μεγέθους Ιστιακών Πόρων

- ❖ Τυχαίες μεταβλητές οι οποίες ακολουθούν κατανομές heavy-tailed λαμβάνουν *πολλές μικρές τιμές* αναμεμιγμένες με *λίγες πολύ μεγάλες τιμές*.
- ❖ Οι μεγάλες τιμές τείνουν να καθορίζουν τις τιμές απλών στατιστικών μεγεθών (μέσους όρους κλπ).
- ❖ Απαλοιφή λίγων μεγάλων τιμών μπορεί να αλλάξει δραστικά τις τιμές των στατιστικών μεγεθών.
- ❖ Συνεπώς, υποσύνολα ενός δείγματος τιμών δεν δίνουν καλή εκτίμηση των στατιστικών μεγεθών ολόκληρου του δείγματος (το Κεντρικό Οριακό Θεώρημα δεν ισχύει – αφού ισχύει σε περιπτώσεις κατανομών πεπερασμένης μεταβλητότητας).

Κεντρικό Οριακό Θεώρημα

Probability and Statistics > Statistical Distributions > Limit Theorems

Central Limit Theorem



Let X_1, X_2, \dots, X_N be a set of N independent random variates and each X_i have an arbitrary probability distribution $P(x_1, \dots, x_N)$ with mean μ_i and a finite variance σ_i^2 . Then the normal form variate

$$X_{\text{norm}} \equiv \frac{\sum_{i=1}^N x_i - \sum_{i=1}^N \mu_i}{\sqrt{\sum_{i=1}^N \sigma_i^2}} \quad (1)$$

has a limiting cumulative distribution function which approaches a normal distribution.

Wolfram MathWorld the web's most extensive mathematics resource
Built with Mathematica Technology

Δημοτικότητα Ιστιακών Πόρων

- ❖ Η δημοτικότητα των ιστιακών πόρων μας ενδιαφέρει για πολλούς λόγους:
 - ❖ **Σημασιολογικά**: δέχονται όλοι οι πόροι τον ίδιο αριθμό αιτημάτων;
 - ❖ Για βελτίωση **επίδοσης** με χρήση απομνημόνευσης: αν λίγοι πόροι δέχονται τα περισσότερα αιτήματα, τότε συμφέρει να απομνημονεύουμε αυτούς τους δημοφιλείς πόρους, ώστε να βελτιώνουμε την επίδοση των πιο συχνών αιτημάτων.
 - ❖ Για βελτίωση **επίδοσης** με χρήση προ-ανάκτησης (**prefetching**): φροντίζουμε να μετακινήσουμε πλησιέστερα στους χρήστες τα πιο δημοφιλή αντικείμενα – ή να τα κρατήσουμε στην κεντρική μνήμη.

Δημοτικότητα Ιστιακών Πόρων: Ορισμός

- ❖ Η δημοτικότητα (popularity) ορίζεται σαν ποσοστό των αιτημάτων που δέχεται ένας ιστιακός πόρος επί των συνόλου των αιτημάτων του ιστιακού τόπου στον οποίο ανήκει.
 - ❖ Οι ιστιακοί πόροι ταξινομούνται σε φθίνουσα διάταξη δημοτικότητας.
 - ❖ Η σειρά (rank) ενός ιστιακού πόρου σε αυτή τη διάταξη χρησιμοποιείται για τον χαρακτηρισμό της κατανομής πιθανότητας της δημοτικότητας ιστιακών πόρων.
- ❖ Μελέτες έχουν δείξει ότι η probability distribution function του rank ακολουθεί την συνάρτηση:
$$P(r) = k \cdot r^{-1}, k=\text{proportionality constant}$$
- ❖ Το ποσοστό των αιτημάτων ενός πόρου είναι αντιστρόφως ανάλογο της σειράς του: Νόμος του Zipf.

Αλλα Θέματα Χαρακτηρισμού Ιστού

- ❖ **Ρυθμός αλλαγής** περιεχομένου ιστιακών πόρων (rate of change).
- ❖ **Χρονική τοπικότητα**: πιθανότητα ότι ένας αιτούμενος ιστιακός πόρος θα ζητηθεί πάλι σε σύντομο χρονικό διάστημα (temporal locality).
- ❖ Αριθμός **ενθυλακωμένων πόρων** (embedded resources).
- ❖ **Self-similarity** στη διακίνηση δικτύου: η διακίνηση εμφανίζει σοβαρότατες διακυμάνσεις σε μεγάλο εύρος χρονικών κλιμάκων, από μικροδευτερόλεπτα μέχρι αρκετά λεπτά και ώρες (burstiness).

Self-Similarity

Self-Similarity



An object is said to be self-similar if it looks "roughly" the same on any scale. [Fractals](#) are a particularly interesting class of self-similar objects. Self-similar objects with parameters N and s are described by a power law such as

$$N = s^d,$$

where

$$d = \frac{\ln N}{\ln s}$$

is the "[dimension](#)" of the scaling law, known as the [Hausdorff dimension](#).

Wolfram MathWorld the web's most extensive mathematics resource
Built with Mathematica Technology

Άλλα Θέματα Χαρακτηρισμού Ιστού

- ❖ Η συμπεριφορά ενός χρήστη και το φορτίο εργασίας που αυτός συνεπάγεται μπορεί να μοντελοποιηθεί σε τρία επίπεδα:
 - ❖ **Σύννοδος (session)**: η αλληλουχία των αιτημάτων ενός χρήστη στη διάρκεια μιας επίσκεψης του σε κάποιον ιστιακό τόπο. Ακολουθείται από περίοδο απραξίας (σιγής).
 - ❖ Στη διάρκεια μιας συνόδου, ο χρήστης πραγματοποιεί ένα ή περισσότερα **κλίκ** για να ζητήσει το άνοιγμα ενός υπερσυνδέσμου.
 - ❖ Ένα κλικ συνεπάγεται την αποστολή ενός **αιτήματος** HTTP, το οποίο ακολουθείται συνήθως από άλλα αυτομάτως εγχειρόμενα από τον φυλλομετρητή αιτήματα.

Χαρακτηρισμός Συμπεριφοράς Χρηστών

- ❖ Η δημιουργία μοντέλων ακριβείας φορτίου εργασίας για τον Ιστό πρέπει να λαμβάνει υπόψιν τα χαρακτηριστικά *αφίξεων* σε επίπεδο:
 - ❖ Συνόδου χρήστη.
 - ❖ Συνδέσεων TCP
 - ❖ Αιτημάτων HTTP
- ❖ *Αφίξεις συνόδων* (session arrivals): μετρήσεις δείχνουν ότι ο χρόνος που μεσολαβεί ανάμεσα στις αρχές δύο διαδοχικών συνόδων ακολουθεί *εκθετική κατανομή*.
- ❖ Η εκθετική κατανομή προκύπτει όταν έχουμε αφίξεις *Poisson* – δηλ. ο ένας χρήστης έρχεται ανεξάρτητα από τον άλλο.
- ❖ Ωστόσο, οι αφίξεις συνδέσεων TCP και αιτημάτων HTTP *δεν είναι Poisson* έχουν ριπαία συμπεριφορά (*bursty*)- άρα και με υψηλή μεταβλητότητα).
- ❖ Η υπόθεση της υψηλής μεταβλητότητας σημαίνει ότι δεν μπορούμε να κάνουμε προγραμματισμό χωρητικότητας και προβλέψη με βάση τους μέσους όρους...

Κλίκ ανά σύνοδο

- ❖ Οι περισσότερες σύνοδοι έχουν μικρό αριθμό από κλικ: 4-10.
- ❖ Ωστόσο οι μετρήσεις αλλάζουν ανάλογα με τον ιστιακό τόπο: πχ. eCommerce sites vs. Search Engines.
- ❖ Η κατανομή των κλικ είναι **Pareto** (heavy tailed): συνεπώς έχουμε ορισμένες συνόδους με πολύ μεγάλο αριθμό από κλικ σε σύγκριση με άλλες.

Χρόνος ανάμεσα στις αφίξεις HTTP

- ❖ Ο χρόνος ανάμεσα σε δύο διαδοχικά κλικ αναφέρεται ως **χρόνος σιγής** (*think time* ή *quiet time*).
- ❖ Η εκτίμηση του χρόνου αυτού είναι χρήσιμη, καθώς μπορεί να επηρεάσει την αποτελεσματικότητα πολιτικών για κλείσιμο επίμονων συνδέσεων.
- ❖ Συνήθως ο χρόνος ανάμεσα σε διαδοχικά κλικ κυμαίνεται σε **λιγότερα από 60 δευτερόλεπτα**. Ωστόσο, ένα μικρό ποσοστό χρόνων σιγής είναι πολύ μεγάλοι.
- ❖ Μετρήσεις έχουν δείξει ότι οι χρόνοι σιγής ακολουθούν κατανομή **Pareto** με heavy tail και α γύρω στο 1.5.

Χρόνος ανάμεσα στις αφίξεις HTTP

- ❖ Μια περίοδος στη ζωή ενός ιστιακού τόπου μπορεί να μοντελοποιηθεί σαν μια σειρά από περιόδους on/off.
- ❖ Μια on περίοδος αντιστοιχεί στην ανάκτηση μιας ιστοσελίδας και των ενθυλακωμένων αντικειμένων της.
- ❖ Μια off περίοδος αντιστοιχεί σε περίοδο σιγής των χρηστών.
- ❖ Οι διάρκειες των περιόδων on και off ακολουθούν heavy-tailed κατανομές Pareto.
- ❖ Η διακίνηση που «βλέπει» ένας ιστιακός τόπος είναι αποτέλεσμα επιστροφών πολλαπλών περιόδων on/off, καθεμιά εκ των οποίων οφείλεται σε διαφορετικούς χρήστες.
- ❖ Συνεπώς, το φορτίο στους εξυπηρετητές και στο δίκτυο εμφανίζει χαρακτηριστικά **self-similarity**.

Σύνοψη

Κατανομή	Παράμετρος Φορτίου Εργασίας
Εκθετική	Session interarrival times
Pareto	Response sizes (tail of distribution) Resource sizes (tail of distribution) Number of embedded images Request interarrival times
Lognormal	Response sizes (body of distribution) Resource sizes (body of distribution) Temporal locality
Zipf-like	Resource popularity