Ανάκληση Πληροφορίας

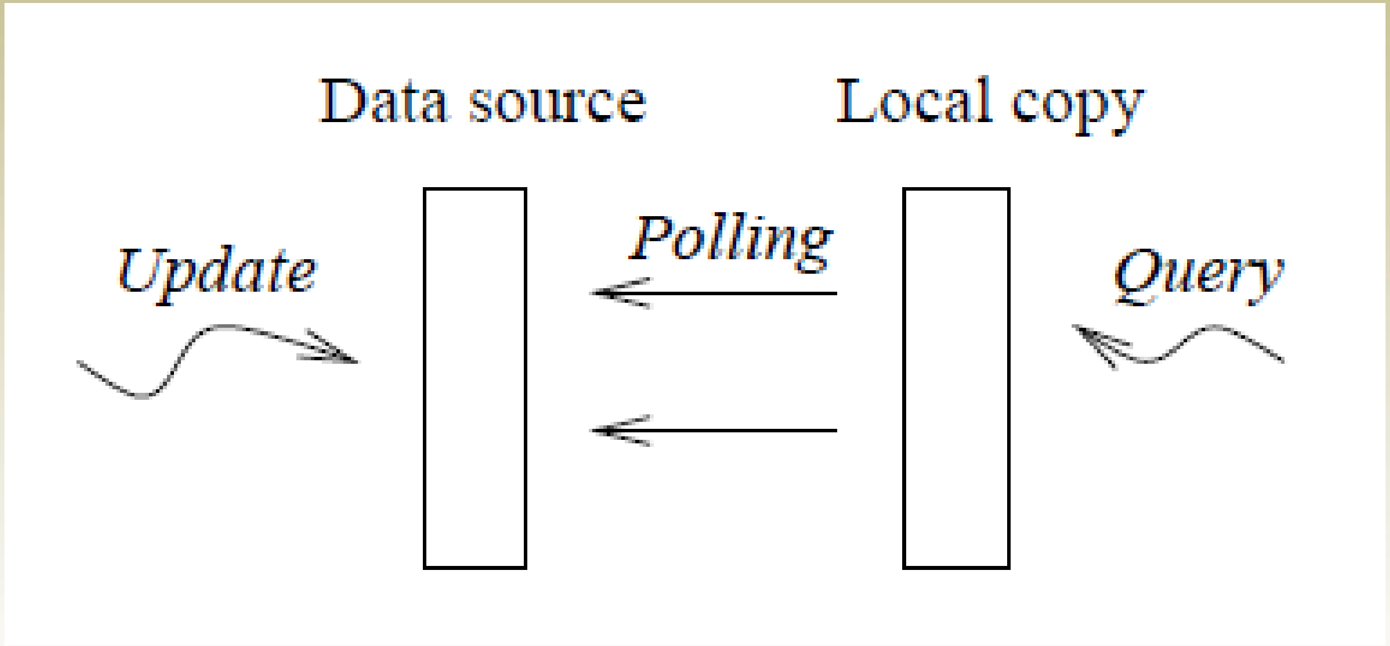
Information Retrieval

Διδάσκων –
Δημήτριος Κατσαρός

Ερπυστές στον Παγκόσμιο Ιστό

Το πρόβλημα της ανενέωσης σελίδων στον index από τους ερπυστές

Θυμηθείτε: Το πρόβλημα



Σύμβολα

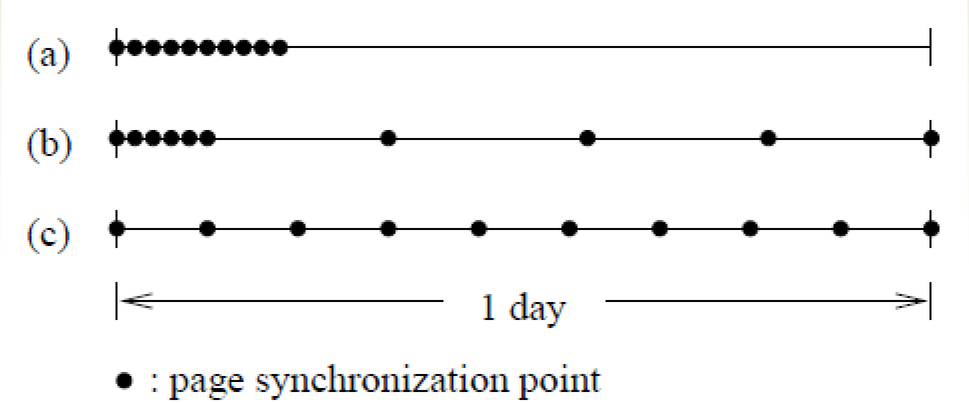
symbol	meaning
(a) $\bar{F}(S)$, $\bar{F}(e_i)$	Freshness of database S (and element e_i) averaged over time
(b) $\bar{A}(S)$, $\bar{A}(e_i)$	Age of database S (and element e_i) averaged over time
(c) $\bar{F}(\lambda_i, f_i)$, $\bar{A}(\lambda_i, f_i)$	Freshness (and age) of element e_i averaged over time, when the element changes at the rate λ_i and is synchronized at the frequency f_i
(i) λ_i	Change frequency of element e_i
(j) f_i $(= 1/I_i)$	Synchronization frequency of element e_i
(k) λ	Average change frequency of database elements
(l) f $(= 1/I)$	Average synchronization frequency of database elements

Πολιτικές συγχρονισμού

- **Synchronization frequency** (Συχνότητα συγχρονισμού)
  - Συγχρονίζουμε N στοιχεία ανά I μονάδες χρόνου
  - Μεταβάλλοντας το I, προσαρμόζουμε το πόσο συχνά συγχονίζουμε τον index (της μηχανής αναζήτησης)
- **Resource allocation** (Ανάθεση πόρων)
  - Uniform-allocation policy
    - Συγχρονίζουμε όλα τα στοιχεία με τον ίδιο ρυθμό (συχνότητα) f, ανεξάρτητα από τον ρυθμό αλλαγής τους, δηλ., $f=f_i=f_j \; \forall \; i,j$
  - Non-uniform-allocation policy
    - Συγχονίζουμε τα στοιχεία με διαφορετική συχνότητα. Με μια *proportional-allocation* πολιτική policy συγρονίζουμε το e_i με συχνότητα f_i που είναι αναλογη της συχνότητας αλλαγής του λ_i, δηλαδή, $\lambda_i/f_i = \lambda_j/f_j \; \forall \; i,j$

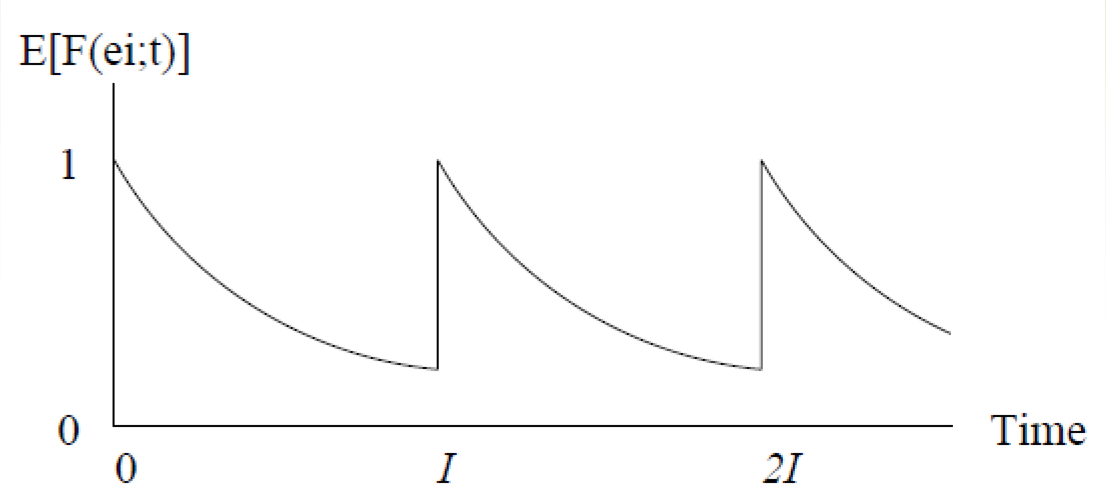
Πολιτικές συγχρονισμού

- **Synchronization Order** (Σειρά συγχρονισμού)
  - Fixed order: δηλ., ένα συγκεκριμένο στοιχείο συγχρονίζεται σε ένα σταθερό διάστημα
  - Random order: τυχαία επιλογή χωρίς επανατοποθέτηση
  - Purely random order: τυχαία επιλογή με επανατοποθέτηση
- **Synchronization Points** (Σημεία συγχρονισμού)
  - Όλα στην αρχή
  - Τα περισσότερα στην αρχή, τα υπόλοιπα ομοιόμορφα
  - Ομοιόμορφα



Ανάλυση πολιτικών sync order

- Υποθέτουμε
  - Uniform change-frequency των στοιχείων
  - Άρα, uniform allocation policy
- **<u>Ανάλυση fixed-order policy</u>**
  - Προφανώς η χρονική εξέλιξη της expected freshness ενός στοιχείου θα είναι όπως στο σχήμα, εάν υποθέσουμε ότι απαιτεί *I* secs για να συγχρονίζουμε όλα τα στοιχεία της βάσης



Ανάλυση της fixed-order policy

Θεώρημα. Όταν το στοιχείο e_i συγχρονίζεται στο σταθερό διάστημα I secs, ο χρονικός μέσος της freshness του e_i είναι ίσος με τον χρονικό μέσο της $E[F(e_i;t)]$ στο διάστημα $(0,I)$, δηλαδή:

$$\overline{F}(e_i) = \frac{1}{I}\int_0^I E[F(e_i;t)]dt$$

Απόδειξη.

$$\overline{F}(e_i) = \lim_{t\to\infty} \frac{\int_0^t F(e_i;t)dt}{t} =$$

$$= \lim_{n\to\infty} \frac{\sum_{j=0}^{n-1} \int_{jI}^{(j+1)I} F(e_i;t)dt}{\sum_{j=0}^{n-1} I}$$

$$= \lim_{n\to\infty} \frac{\sum_{j=0}^{n-1} \int_0^I F(e_i;t+jI)dt}{nI} = \frac{1}{I}\int_0^I \left[\lim_{n\to\infty}\frac{1}{n}\sum_{j=0}^{n-1} F(e_i;t+jI)\right]dt \quad \textbf{(A)}$$

Ανάλυση της fixed-order policy

- Επειδή συγχρονίζουμε το e_i κάθε I secs από t=0, $F(e_i; t+jI)$ είναι η freshness του t secs μετά από κάθε συγχρονισμό
- Επομένως $\frac{1}{n}\sum_{j=0}^{n-1} F(e_i; t+jI)$ δηλ., η μέση freshness σε t secs μετά από κάθε συγχρονισμό, θα συγκλίνει στην expected value $E[F(e_i;t)]$, καθώς $n \rightarrow \infty$
- Δηλαδή $$\lim_{n\to\infty} \frac{1}{n}\sum_{j=0}^{n-1} F(e_i; t+jI) = E[F(e_i;t)]$$
- Τότε, $$\frac{1}{I}\int_0^I \left[\lim_{n\to\infty} \frac{1}{n}\sum_{j=0}^{n-1} F(e_i; t+jI)\right] dt = \frac{1}{I}\int_0^I E[F(e_i;t)]dt \qquad \textbf{(B)}$$
- Από τις **(A)** και **(B)**, προκύπτει ότι: $F(e_i) = \frac{1}{I}\int_0^I E[F(e_i;t)]dt$

Τέλος απόδειξης

Ανάλυση της fixed-order policy

- Με βάση το Θεώρημα, μπορούμε να υπολογίζουμε την freshness του e_i ως εξής:

$$\overline{F}(e_i) = \frac{1}{I}\int_0^I E[F(e_i;t)]dt = \frac{1}{I}\int_0^I e^{-\lambda t}dt = \frac{1-e^{-\lambda I}}{\lambda I} = \frac{1-e^{-\lambda/f}}{\lambda/f}$$

- Υποθέσαμε ότι όλα τα στοιχεία αλλάζουν με την ίδια συχνότητα λ και ότι όλα συγχρονίζονται στο ίδιο διάστημα I, και άρα η παραπάνω εξίσωση ισχύει για κάθε στοιχείο e_i. Επομένως η freshness του index S είναι:

$$\overline{F}(S) = \frac{1}{N}\sum_{i=1}^{N}\overline{F}(e_i) = \frac{1-e^{-\lambda/f}}{\lambda/f}$$

- Επίσης, βρίσκουμε (δείτε επόμενη διαφάνεια) ότι η age θα είναι:

$$\overline{A}(S) = I\left(\frac{1}{2} - \frac{1}{\lambda/f} + \frac{1-e^{-\lambda/f}}{(\lambda/f)^2}\right)$$

Εύρεση της age στην fixed-order πολιτική

$$\overline{A}(S) = \frac{1}{N}\sum_{i=1}^{N}\overline{A}(e_i) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{I}\int_0^I E[A(e_i;t)]dt\right) = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{I}\int_0^I t\Big(1-\frac{1-e^{-\lambda_i t}}{\lambda_i t}\Big)dt\right]$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{1}{I}\left[\int_0^I t dt - \int_0^I \frac{t}{\lambda_i t}dt + \int_0^I \frac{te^{-\lambda_i t}}{\lambda_i t}dt\right]\right\}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{1}{I}\left[\frac{t^2}{2}\Big|_0^I - \frac{t}{\lambda_i}\Big|_0^I - \frac{e^{-\lambda_i t}}{\lambda_i^2}\Big|_0^I\right]\right\}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{1}{I}\left[\frac{I^2}{2} - \frac{I}{\lambda_i} - \frac{e^{-\lambda_i I}}{\lambda_i^2} + \frac{1}{\lambda_i^2}\right]\right\}$$

$$\overset{\lambda_i=\lambda}{=} I\left[\frac{1}{2} - \frac{1}{\lambda I} + \frac{1-e^{\lambda I}}{\lambda^2 I^2}\right]$$

$$\overset{I=1/f}{=} I\left[\frac{1}{2} - \frac{1}{\lambda/f} + \frac{1-e^{\lambda/f}}{\lambda^2/f^2}\right]$$

Αφού υποθέσαμε uniform allocation policy

Ανάλυση της random-order policy

- Αφού συγχρονίζουμε το e_i σε οποιαδήποτε στιγμή κατά το διάστημα I, το διάστημα συγχρονισμού του e_i δεν είναι σταθερό. Στην μια ακραία περίπτωση θα είναι σχεδόν $2I$, και στην άλλη ακραία περίπτωση θα είναι σχεδόν 0
- Επομένως το synchronization interval του e_i, δηλ. το W, δεν είναι σταθερός αριθμός, αλλά ακολουθεί κάποια κατανομή $f_W(t)$. Άρα, η εξίσωση του Θεωρήματος θα πρέπει να γίνει:

$$\overline{F}(e_i) = \lim_{t\to\infty} \frac{1}{t} \int_0^t F(e_i;t)dt = \frac{\int_0^{2I} f_W(s)\Big(\int_0^s E[F(e_i;t)]dt\Big)ds}{\int_0^{2I} f_W(s)\ s\ ds} \quad \textbf{(Γ)}$$

- Για να βρούμε το ολοκλήρωμα, πρέπει να βρούμε την κλειστή μορφή της $f_W(t)$

Ανάλυση της random-order policy

Λήμμα. Έστω ότι $T_1(T_2)$ είναι η χρονική στιγμή όταν το στοιχείο e_i συγχρονίζεται την πρώτη (δεύτερη) φορά υπό την random-order πολιτική. Τότε, η p.d.f. του $W=T_1-T_2$, δηλαδή το διάστημα συγχρονισμού του e_i είναι:

$$f_W(t) = \begin{cases} \frac{t}{I^2} & 0 \le t \le I \\ \frac{2I-t}{I^2} & I \le t \le 2I \\ 0 & \text{otherwise} \end{cases}$$

Απόδειξη.

Οι p.d.f. των T_1 και T_2 είναι οι κάτωθι:

$$f_{T_1} = \begin{cases} \frac{1}{I} & 0 \le t \le I \\ 0 & \text{otherwise} \end{cases} \qquad f_{T_2} = \begin{cases} \frac{1}{I} & I \le t \le 2I \\ 0 & \text{otherwise} \end{cases}$$

Ανάλυση της random-order policy

Τότε:

$$f_W(t) = f(T_2 - T_1 = t)$$

$$= \int_0^I f(T_1 = s) f(T_2 - T_1 = t | T_1 = s) ds$$

$$= \int_0^I f(T_1 = s) f(T_2 = s + t) ds$$

$$= \frac{1}{I} \int_0^I f(T_2 = s + t) ds$$

Όταν $t < 0$ ή t > 2I, $f(T_2 = s + t) = 0$ για κάθε $s \in (0, I)$. Επομένως:

$$f_W(t) = \frac{1}{I} \int_0^I f(T_2 = s + t) ds = 0$$

Ανάλυση της random-order policy

Όταν $0 \le t \le I$, $f(T_2=s+t) = 1/I$ για κάθε $s \in (I\text{-}t, I)$. Επομένως:

$$f_W(t) = \frac{1}{I}\int_{I-t}^{I}\frac{1}{I}ds = \frac{t}{I^2}$$

Όταν $I \le t \le 2I$, $f(T_2=s+t) = 1/I$ για κάθε $s \in (0, 2I\text{-}t)$. Επομένως:

$$f_W(t) = \frac{1}{I}\int_{0}^{2I-t}\frac{1}{I}ds = \frac{2I-t}{I^2}$$

Τέλος απόδειξης

Ανάλυση της random-order policy

- Με βάση το Λήμμα και την εξίσωση **(Γ)**, μπορούμε να υπολογίσουμε την freshness της random-order policy (και για το στοιχείο e_i αλλά και για τον index S) ως εξής:

$$\overline{F}(e_i) = \overline{F}(S) = \frac{1}{\lambda/f}\left[1 - \left(\frac{1-e^{-\lambda/f}}{\lambda/f}\right)^2\right]$$

- Με ανάλογη διαδικασία, βρίσκουμε ότι η age είναι:

$$\overline{A}(e_i) = \overline{A}(S) = I\left[\frac{1}{3} + \left(\frac{1}{2} - \frac{1}{\lambda/f}\right)^2 - \left(\frac{1-e^{-\lambda/f}}{(\lambda/f)^2}\right)^2\right]$$

Ανάλυση της purely random-order policy

- Ακολουθούμε την ίδια συλλογιστική με την random-order policy, αλλά πλέον το διάστημα συγχρονισμού είναι unbounded, συνεπώς η εξίσωση **(Γ)** γίνεται:

$$\overline{F}(e_i) = \lim_{t \to \infty} \frac{1}{t} \int_0^t F(e_i;t)dt = \frac{\int_0^\infty f_W(s)\left(\int_0^s E[F(e_i;t)]dt\right)ds}{\int_0^\infty f_W(s)\ s\ ds}$$

- Από τον νόμο των σπάνιων γεγονότων, συνάγουμε ότι:

$$f_W(t) = \begin{cases} fe^{-ft} & t \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

- Συνεπώς (επόμενη διαφάνεια):

$$\overline{F}(S) = \frac{1}{1 + \lambda/f} \qquad \overline{A}(S) = I\left(\frac{\lambda/f}{1 + \lambda/f}\right)$$

Εύρεση της μέσης freshness στοιχείου και index στην purely random-order policy

$$\int_0^\infty f_W(s)s\,ds = \int_0^\infty f e^{-fs} s\,ds = f\int_0^\infty s e^{-fs} ds = -f\left[\frac{fs+1}{f^2}e^{-fs}\Big|_0^\infty\right] = \frac{-f}{f^2} = \frac{-1}{f}(0-1)$$
$$= \frac{1}{f}$$

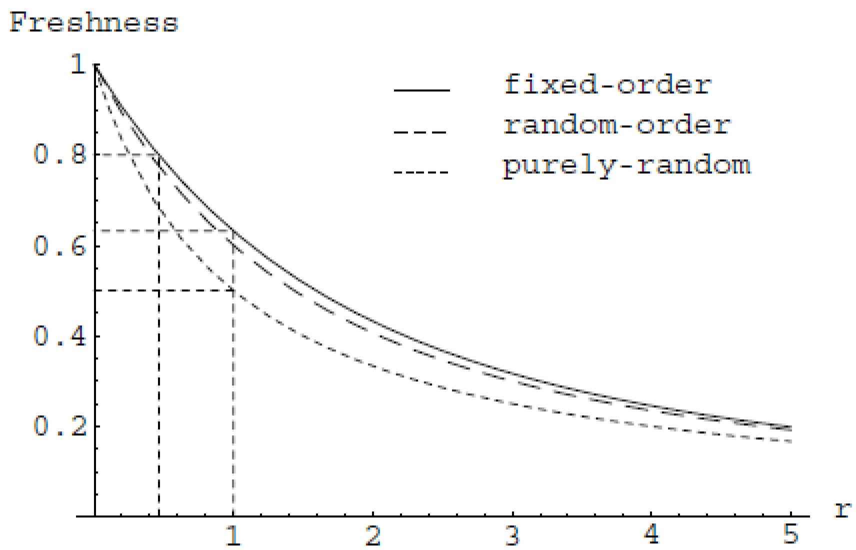
$$\int_0^s E[F(e_i;t)]dt = \int_0^s e^{-\lambda t}dt = \int_0^s \frac{-1}{\lambda}\left(e^{-\lambda t}\right)' dt = \frac{-1}{\lambda}e^{-\lambda t}\Big|_0^s = \frac{-1}{\lambda}(e^{-\lambda s}-1)$$
$$= \frac{1}{\lambda}(1-e^{-\lambda s})$$

$$\int_0^\infty f_W(s)\Big(\int_0^s E[F(e_i;t)]dt\Big)ds = \int_0^\infty f e^{-fs}\frac{1}{\lambda}(1-e^{-\lambda s})ds = \frac{f}{\lambda}\int_0^\infty e^{-fs}(1-e^{-\lambda s})ds$$
$$= \frac{f}{\lambda}\int_0^\infty e^{-fs}ds - \frac{f}{\lambda}\int_0^\infty e^{-(f+\lambda)s}ds = \frac{f}{\lambda}\Big(\frac{-1}{f}[0-1]\Big) - \frac{f}{\lambda}\frac{-1}{f+\lambda}[0-1] = \frac{1}{\lambda} - \frac{f}{\lambda(f+\lambda)}$$
$$= \frac{(f+\lambda)-f}{\lambda(f+\lambda)} = \frac{1}{f+\lambda}$$

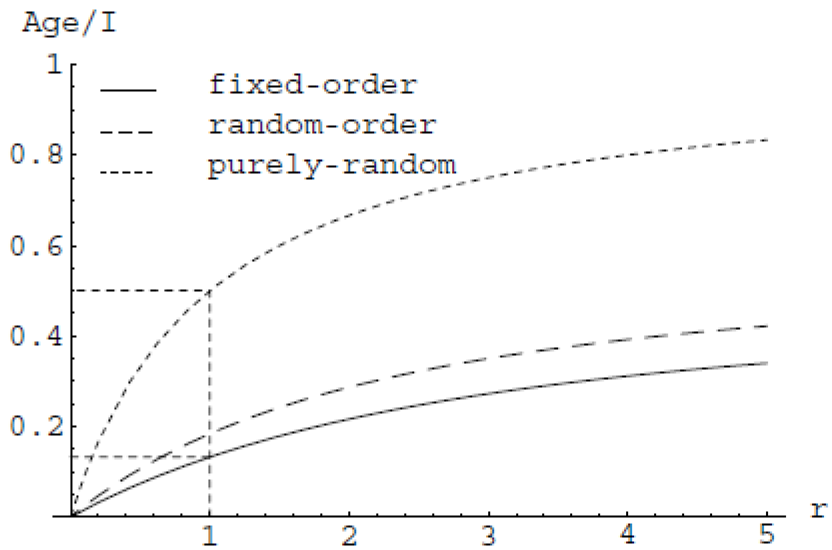
Άρα: $\overline{F}(e_i) = \dfrac{1/(f+\lambda)}{1/f} = \dfrac{f}{f+\lambda} = \dfrac{1}{1+\lambda/f} \implies \overline{F}(S) = \dfrac{1}{1+\lambda/f}$

Ανάλυση πολιτικών sync order: Σύνοψη

policy	*Freshness* $\bar{F}(S)$	*Age* $\bar{A}(S)$
Fixed-order	$\frac{1-e^{-r}}{r}$	$I(\frac{1}{2} - \frac{1}{r} + \frac{1-e^{-r}}{r^2})$
Random-order	$\frac{1}{r}(1 - (\frac{1-e^{-r}}{r})^2)$	$I(\frac{1}{3} + (\frac{1}{2} - \frac{1}{r})^2 - (\frac{1-e^{-r}}{r^2})^2)$
Purely-random	$\frac{1}{1+r}$	$I(\frac{r}{1+r})$



(a) Freshness graph over $r = \lambda/f$



(b) Age graph over $r = \lambda/f$

Ανάλυση πολιτικών resource allocation

- Υποθέτουμε
  - Non uniform change-frequency των στοιχείων
  - Fixed-order sync policy (αφού είναι η καλύτερη)
    - Συγχρονίζουμε στο fixed interval I_i (=$1/f_i$) για το στοιχείο e_i
  - Οι change frequencies των στοιχείων ακολουθούν την γάμμα κατανομή
    - Γενίκευση των: εκθετική, χ^2, (πλησιάζει την) κανονική (όταν είναι μικρή η variance)
- Δείξαμε ήδη ότι:

$$\overline{F}(\lambda_i, f_i) = \frac{1 - e^{-\lambda_i/f_i}}{\lambda_i/f_i} \qquad \textbf{(Δ)}$$

$$\overline{A}(\lambda_i, f_i) = \frac{1}{f_i}\left(\frac{1}{2} - \frac{1}{\lambda_i/f_i} + \frac{1 - e^{-\lambda_i/f_i}}{(\lambda_i/f_i)^2}\right) \qquad \textbf{(E)}$$

Ανάλυση πολιτικών resource allocation

- Η γάμμα κατανομή με παραμέτρους α>0 και μ>0

$$g(x) = \frac{\mu}{\Gamma(\alpha)}(\mu x)^{\alpha-1}e^{-\mu x}, x > 0 \qquad \textbf{(ΣΤ)}$$

με μέση τιμή και διακύμανση

$$E[X] = \frac{\alpha}{\mu}, \qquad Var[X] = \frac{\alpha}{\mu^2} \qquad \textbf{(Ζ)}$$

Ανάλυση της uniform allocation policy

- Δεδομένου ότι $f = f_i \ \forall \ i$, και από την σχετική σχέση στην προηγούμενη διάλεξη, έχουμε:

$$\overline{F}(S)_u = \frac{1}{N}\sum_{i=1}^{N}\overline{F}(e_i) = \frac{1}{N}\sum_{i=1}^{N}\overline{F}(\lambda_i, f)$$

Ανάλυση της uniform allocation policy

- Όταν το N είναι μεγάλο, μπορούμε να προσεγγίσουμε το προηγούμενο μέσο όρο με το weighted ολοκλήρωμα

$$\overline{F}(S)_u = \frac{1}{N}\sum_{i=1}^{N}\overline{F}(\lambda_i, f) = \int_0^{\infty} g(\lambda)\overline{F}(\lambda, f)d\lambda$$

- Αντικαθιστώντας τα *g(λ)* και *F(λ,f)* με την βοήθεια των σχέσεων **(ΣΤ)** και **(Δ)** έχουμε:

$$\overline{F}(S)_u = \frac{\mu f}{1-\alpha}\left(\left(1+\frac{1}{\mu f}\right)^{1-\alpha} - 1\right)$$

- Θέτοντας λ = α/μ (δηλ., ο μέσος) και $\delta^2 = \frac{\alpha/\mu^2}{(\alpha/\mu)^2}$ (δηλ., var/μέσος²), η $F(S)_u$ γίνεται:

$$\overline{F}(S)_u = \frac{1-(1+r\delta^2)^{1-\frac{1}{\delta^2}}}{r(1-\delta^2)} \quad (r = \lambda/f)$$

Ανάλυση της uniform allocation policy

- Υπολογίζοντας την age $A(S)_u$ του index, βρίσκουμε:

$$\overline{A}(S)_u = \frac{1}{N}\sum_{i=1}^{N}\overline{A}(\lambda_i, f) = \int_0^\infty g(\lambda)\overline{A}(\lambda, f)d\lambda$$

$$= \frac{I}{(1-\delta^2)}\left[\frac{1-\delta^2}{2} - \frac{1}{r} + \frac{1-(1+r\delta^2)^{2-\frac{1}{\delta^2}}}{r^2(1-2\delta^2)}\right]$$

<u>Ανάλυση της proportional allocation policy</u>

- Από τον ορισμό της proportional allocation policy, ισχύει ότι $\lambda_i/f_i = \lambda/f\ \forall\ i$, και από τις εξισώσεις **(Δ)** και **(Ε)** μπορούμε να βρούμε ότι:

$$\overline{F}(\lambda_i, f_i) = \frac{1-e^{-\lambda/f}}{\lambda/f} \qquad \overline{A}(\lambda_i, f_i) = \frac{1}{\lambda}\left[\frac{\lambda}{f}\left(\frac{1}{2} - \frac{1}{\lambda/f} + \frac{1-e^{-\lambda/f}}{(\lambda/f)^2}\right)\right]$$

Ανάλυση της proportional allocation policy

- Επομένως οι $F(S)_p$ και $A(S)_p$ γίνονται:

$$\overline{F}(S)_p = \frac{1}{N}\sum_{i=1}^{N}\overline{F}(\lambda_i, f_i) = \frac{1-e^{-\lambda/f}}{\lambda/f}$$

$$\overline{A}(S)_p = \frac{1}{N}\sum_{i=1}^{N}\overline{A}(\lambda_i, f_i)$$

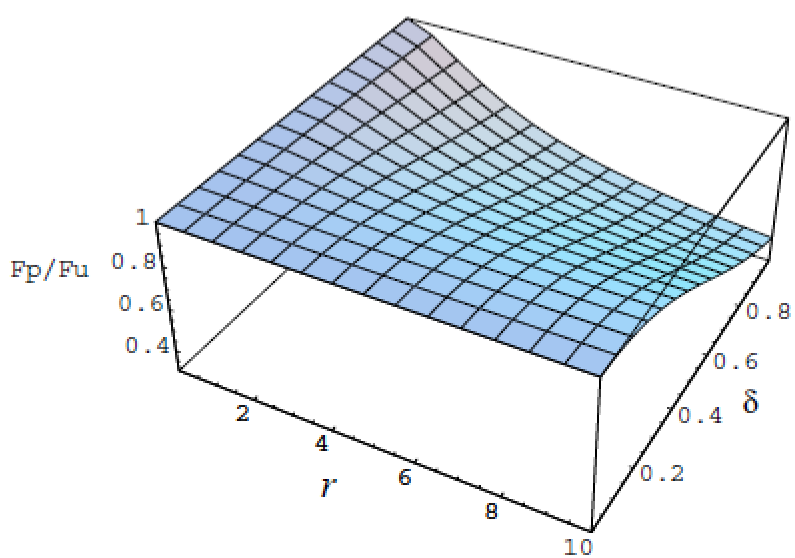
$$= \left[\frac{\lambda}{f}\left(\frac{1}{2} - \frac{1}{\lambda/f} + \frac{1-e^{-\lambda/f}}{(\lambda/f)^2}\right)\right]\left(\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\lambda_i}\right)$$

$$= \left[\frac{\lambda}{f}\left(\frac{1}{2} - \frac{1}{\lambda/f} + \frac{1-e^{-\lambda/f}}{(\lambda/f)^2}\right)\right]\int_0^{\infty} g(\lambda)\frac{1}{\lambda}d\lambda$$

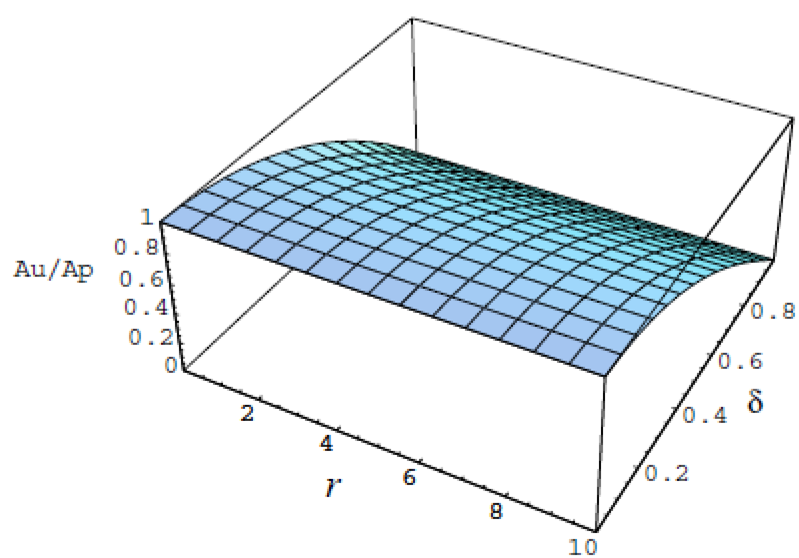
$$= \left[\frac{1}{2} - \frac{1}{\lambda/f} + \frac{1-e^{-\lambda/f}}{(\lambda/f)^2}\right]\frac{I}{(1-\delta^2)}$$

Ανάλυση πολιτικών res alloc: Σύνοψη

policy	Freshness $\bar{F}(S)$	Age $\bar{A}(S)$
Uniform allocation	$\frac{1-(1+r\delta^2)^{1-\frac{1}{\delta^2}}}{r(1-\delta^2)}$	$\frac{I}{(1-\delta^2)}[\frac{1-\delta^2}{2} - \frac{1}{r} + \frac{1-(1+r\delta^2)^{2-\frac{1}{\delta^2}}}{r^2(1-2\delta^2)}]$
Proportional allocation	$\frac{1-e^{-r}}{r}$	$\frac{I}{(1-\delta^2)}[\frac{1}{2} - \frac{1}{r} + \frac{1-e^{-r}}{r^2}]$



(a) $\bar{F}(S)_p/\bar{F}(S)_u$ graph over r and δ



(b) $\bar{A}(S)_u/\bar{A}(S)_p$ graph over r and δ

Υπεροχή της uniform allocation έναντι της proportional allocation

- **Ισχύει πάντα:**

$$\overline{F}(S)_u \geq \overline{F}(S)_p$$

- **Ισχύει πάντα:**

$$\overline{A}(S)_u \leq \overline{A}(S)_p$$

Η απόδειξη των παραπάνω σχέσεων, καθώς και εύρεση της optimal resource allocation policy είναι πέρα από τους σκοπούς της διάλεξης/μαθήματος