Anatomie d'un moteur de recherche

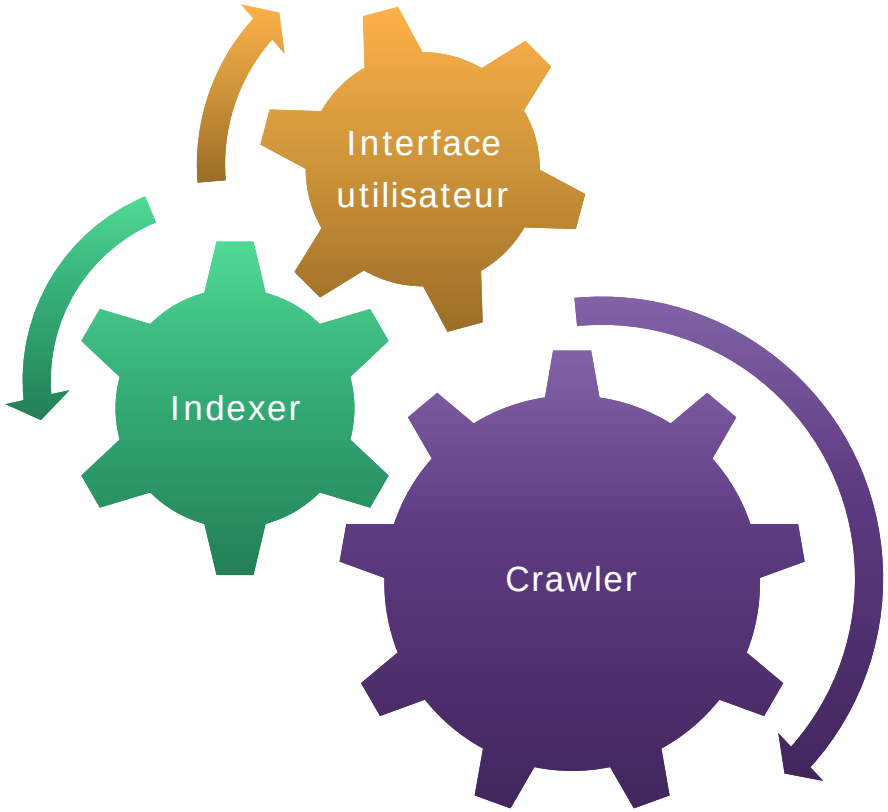
SEO Camp'us -4 et 5 février 2009

Philippe YONNET

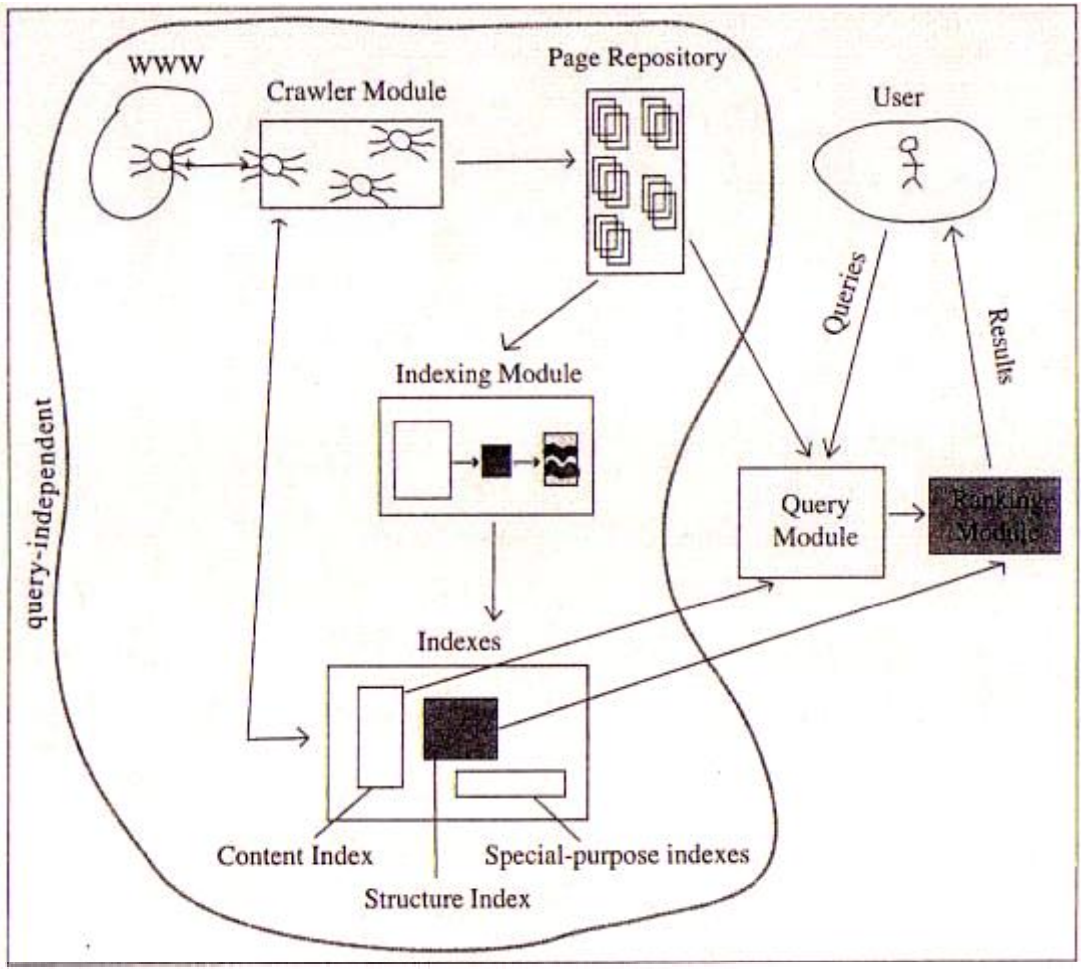
Directeur du pôle métiers – Aposition

Président de l'association SEOCamp

Les grands composants d'un moteur



L'architecture classique d'un moteur



Crawler le web

Combien de pages web ?

- Impossible à savoir précisément (voir plus loin)
- Google déclare connaître 1000 milliards d'urls différentes

Combien de pages utiles sur le web

- 120 / 130 milliards

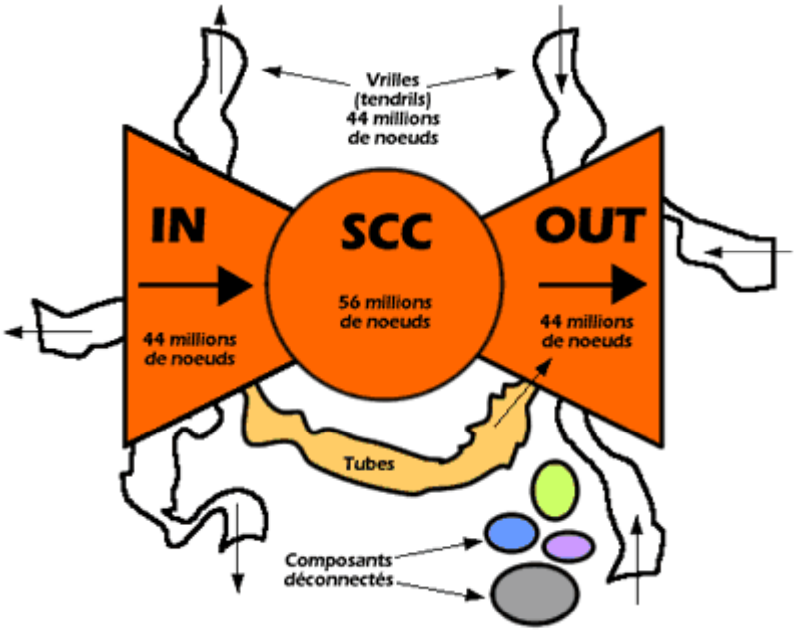
Combien de pages dans les moteurs ?

- Quelques dizaines de milliards
- Pas toutes placées dans le même type d'index
  - Index primaire / index secondaire
  - La révolution « Teragoogle » (alias Bigdaddy)

Crawler le web : un exercice difficile

Kleinberg

La topologie du web est particulière

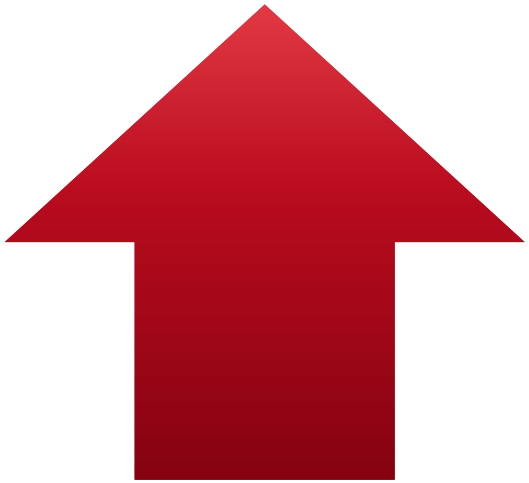


Jon Kleinberg (labo IBM Almaden) : « the web as a graph »

L’importance des urls seeds (urls "semence")

Urls découvertes dans les pages découvertes à partir des urls de niveau 1	
Urls découvertes dans les pages découvertes à partir des urls de niveau 1	
Url découvertes dans les pages correspondant aux urls seeds	• Ne pas les crawler au fur et à mesure => ordonnanceur
Url « seeds »	• Choix déterminant • Deux jeux d’urls = deux index différents

Evolution des techniques de crawl



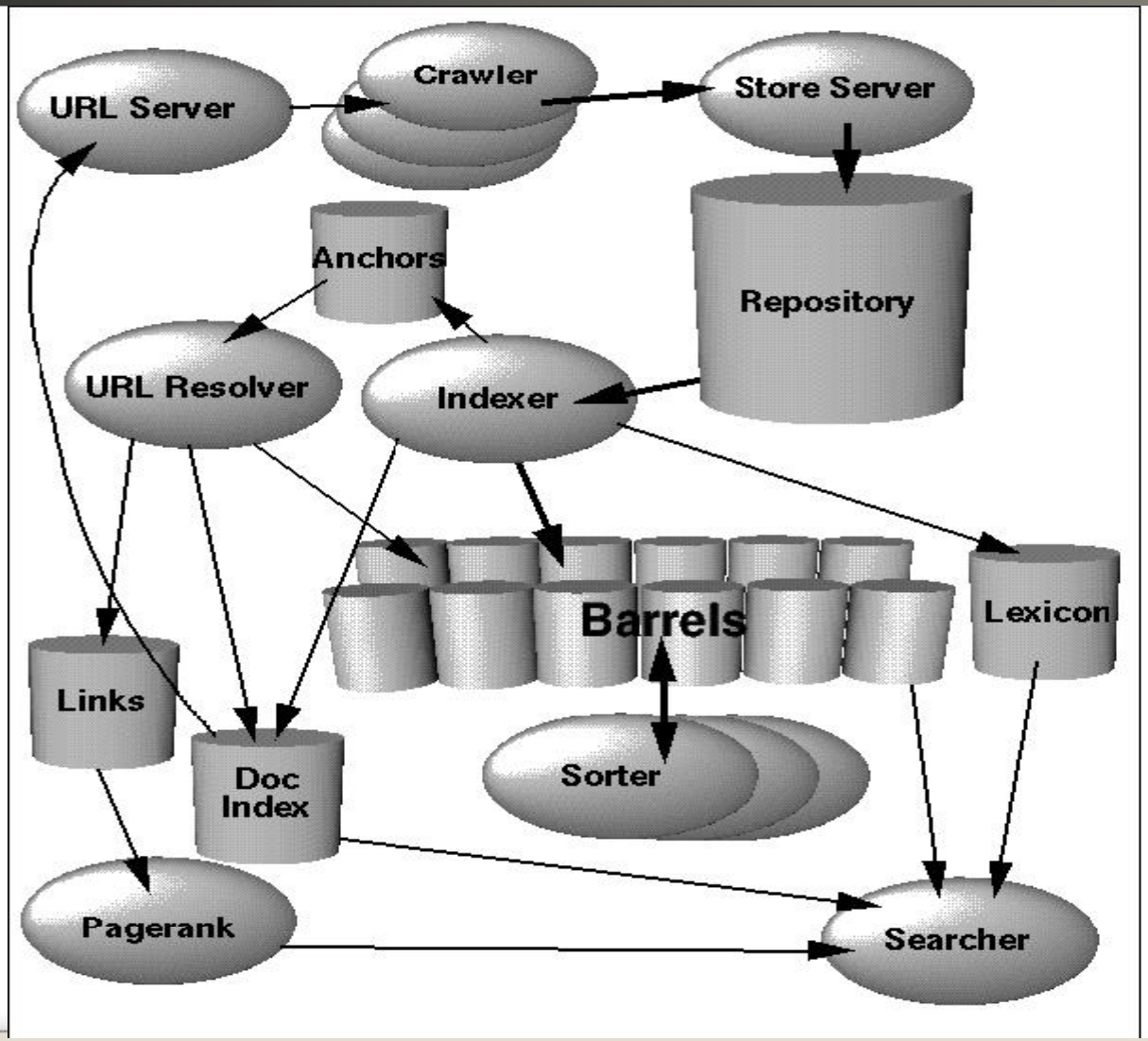
Crawl fermé

- Ancien système de Google
- On part du principe que l'on va crawler un nombre déterminé d'url (celles découvertes au crawl précédent + un niveau de nouvelles par exemple)
- Cycle de 28 jours
  - 1 : deep crawl
  - 2 : indexation + calcul du PR
  - 3 : mise à jour des datacenters

Crawl ouvert

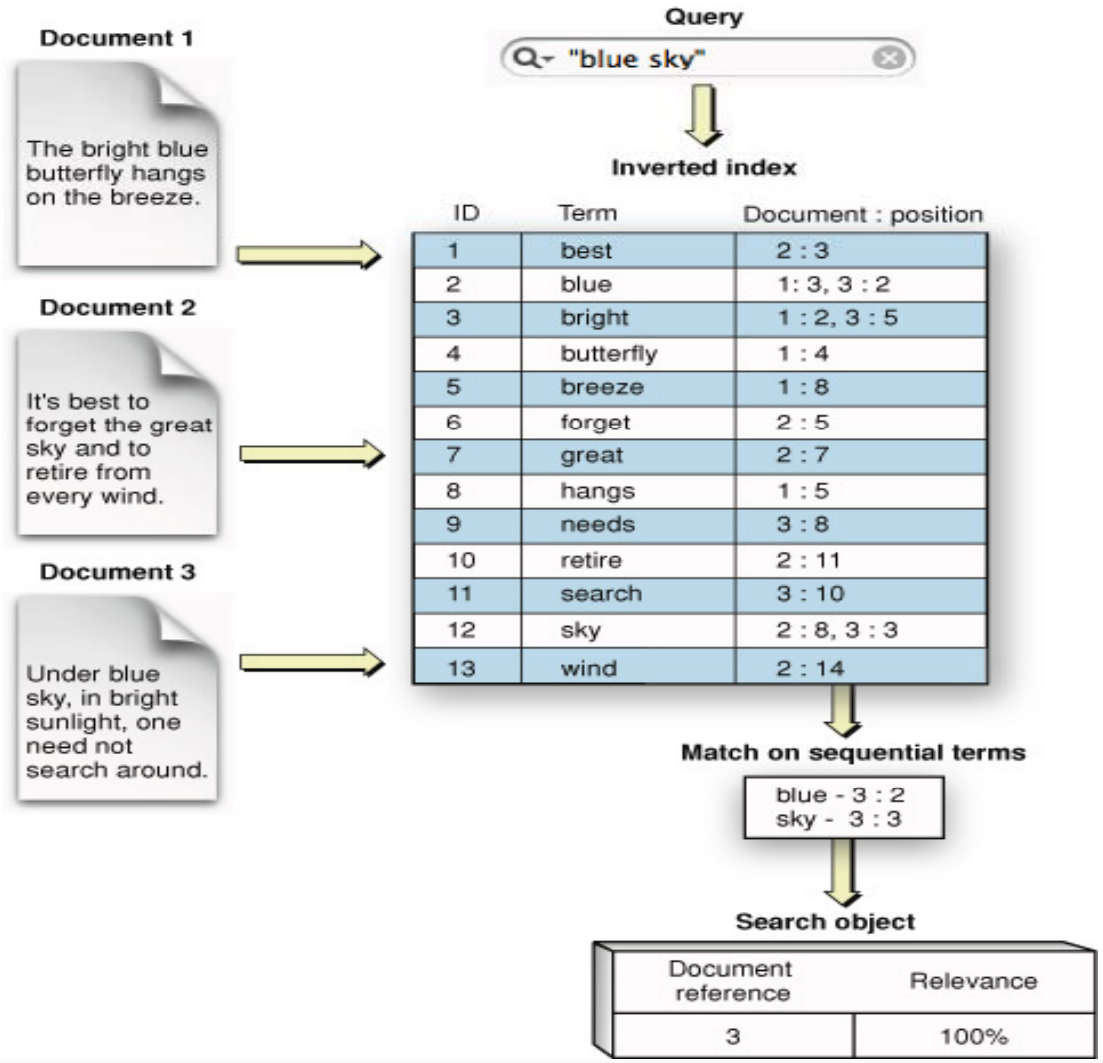
- Si on découvre des urls, on les crawle
- On recrawle les urls connus à différents intervalles de temps
- On ne cherche pas à crawler tout le web, et on considère que le crawl n'est jamais fini

Architecture de Google



L'architecture de Google selon l'article fondateur de Page et Brin

L’indexation : page + position

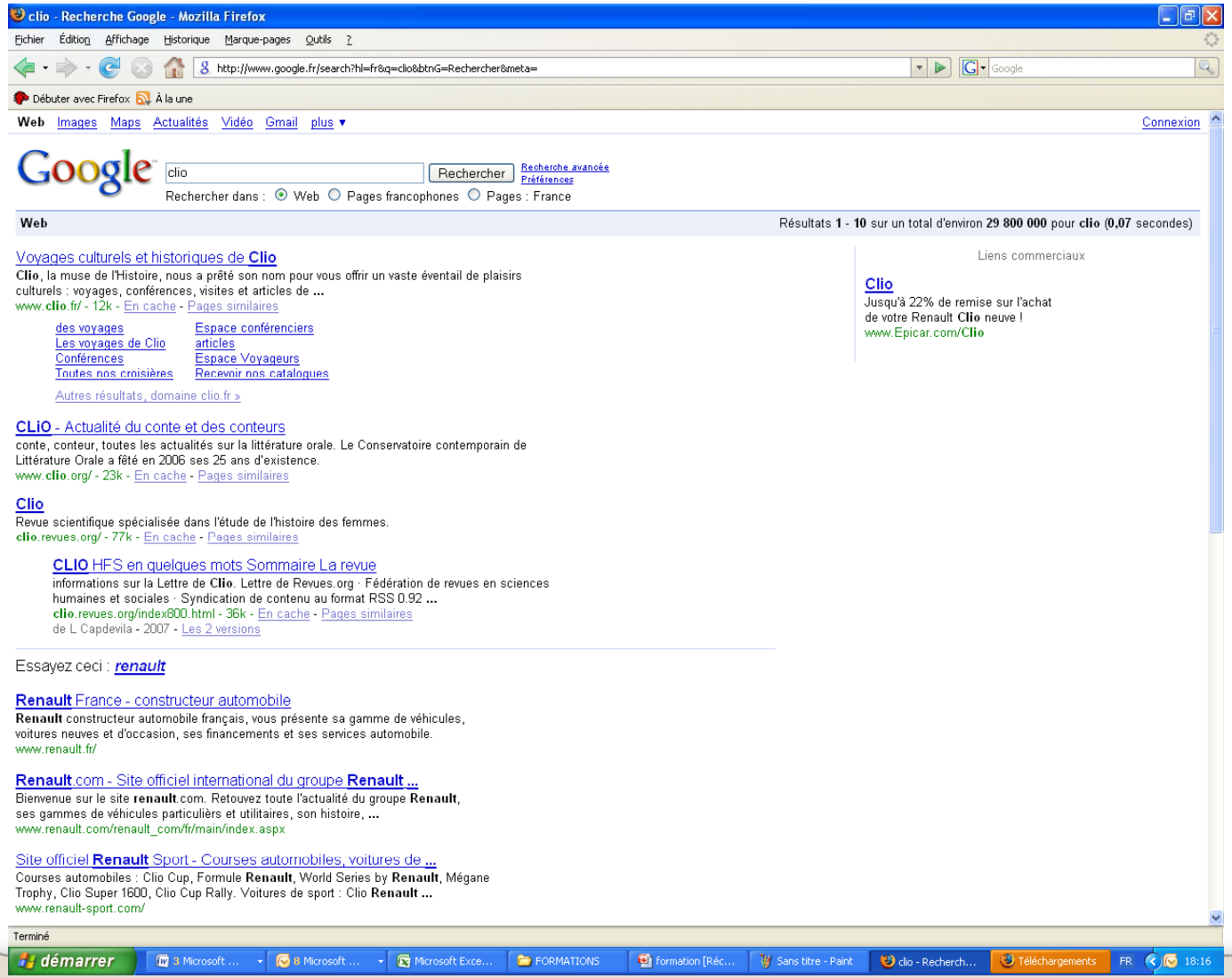


L’interface utilisateur

Le snippet


Sign in
Google
Web Images Video News Maps Desktop more »
blake Field
Search
Advanced Search
Preferences
Web
Results 101 - 110 of about 1,820,000 for blake Field. (0.07 seconds)
Blake completes Masters field | TENNIS | SPORT | tvnz.co.nz
Title
ASB Classic & Heineken Open. Blake completes Masters field ... American James Blake on Sunday secured the eighth and last spot at the season-ending Masters ...
tvnz.co.nz/view/page/488120/879943 - 61k - Cached - Similar pages
New Low-Complexity Bit-Parallel Finite Field Multipliers Using ...
21 Ian F. Blake , Shuhong Gao , Robert Lambert, Constructive problems for irreducible polynomials over finite fields, Proceedings of the third Canadian ...
Snippet
portal.acm.org/citation.cfm?coll=GUIDE&dl=GUIDE&id=305091 - Similar pages
Contributors and Sources
Blake Ferris lives in New York where he experiments with exotic ... Force Field www.wondermagnet.com is a fun site with all kinds of silly magnet stuff, ...
www.theodoregray.com/PeriodicTable/Contributors.html - 316k - Cached - Similar pages
Size
Talisman Announces First Oil From Blake and Production Restart ...
wholly owned subsidiary, Talisman Energy (UK) Limited, has resumed production in the Ross oil field and has commenced production in the Blake oil field over ...
www.encyclopedia.com/doc/1G1-75993928.html - 32k - Cached - Similar pages
Cache
How Not to be a Racist by Blake A. Field
By: Blake A. Field. Copyright 2003 Blake A. Field. Of course, I have always consistently denied that I am a racist. Repeatedly, I made the case that I love ...
www.blakeswritings.com/Racist.html - Similar pages
Similar
FOR THE LOTTO by Blake A. Field
by: Blake A. Field. Copyright 2001 Blake A. Field. Back when, back before the State of Indiana decided that gambling was OK ...
www.blakeswritings.com/ForTheLotto.htm - Similar pages
URL

Un exemple d'évolution de l'Interface
L'expansion de requête à base de thésaurus



Recherche sur clio
Renvoie sur renault