系统设计第三课

小憩一下，马上开课

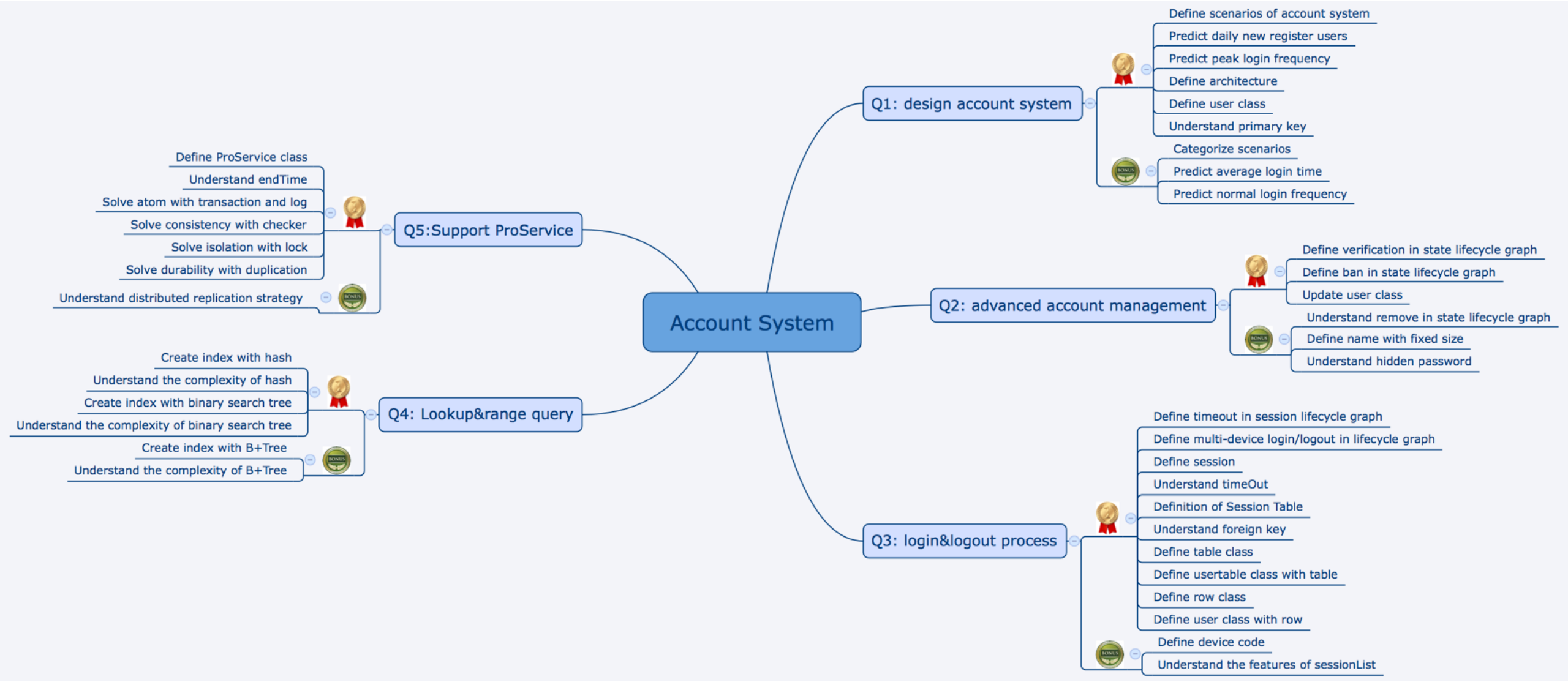
关注微信/微博，获取最新面试题及权威解答

微信: ninechapter

微博: http://www.weibo.com/ninechapter

官网: www.jiuzhang.com

Summary of class



System_Design_3

Web Crawler & Tiny URL

张无忌

2015-10-11

V 3.01

After class_3, you can answer

- Design crawler
  - Dropbox, Google, Turn, Alibaba
- Design thread-safe consumer and producer
  - Google, Amazon, TripAdvisor, Microsoft, Pure Storage, Dropbox, LinkedIn, Palantir, Intel, Bloomberg, …
- Design TinyURL
  - LinkedIn, Uber, Bloomberg, Hulu

Interviewer

- Let's implement "News on Singers"

LATEST MUSIC NEWS

> EXCLUSIVE: Rita Ora Explains Why It Couldn't Work Between Her And Zayn Malik

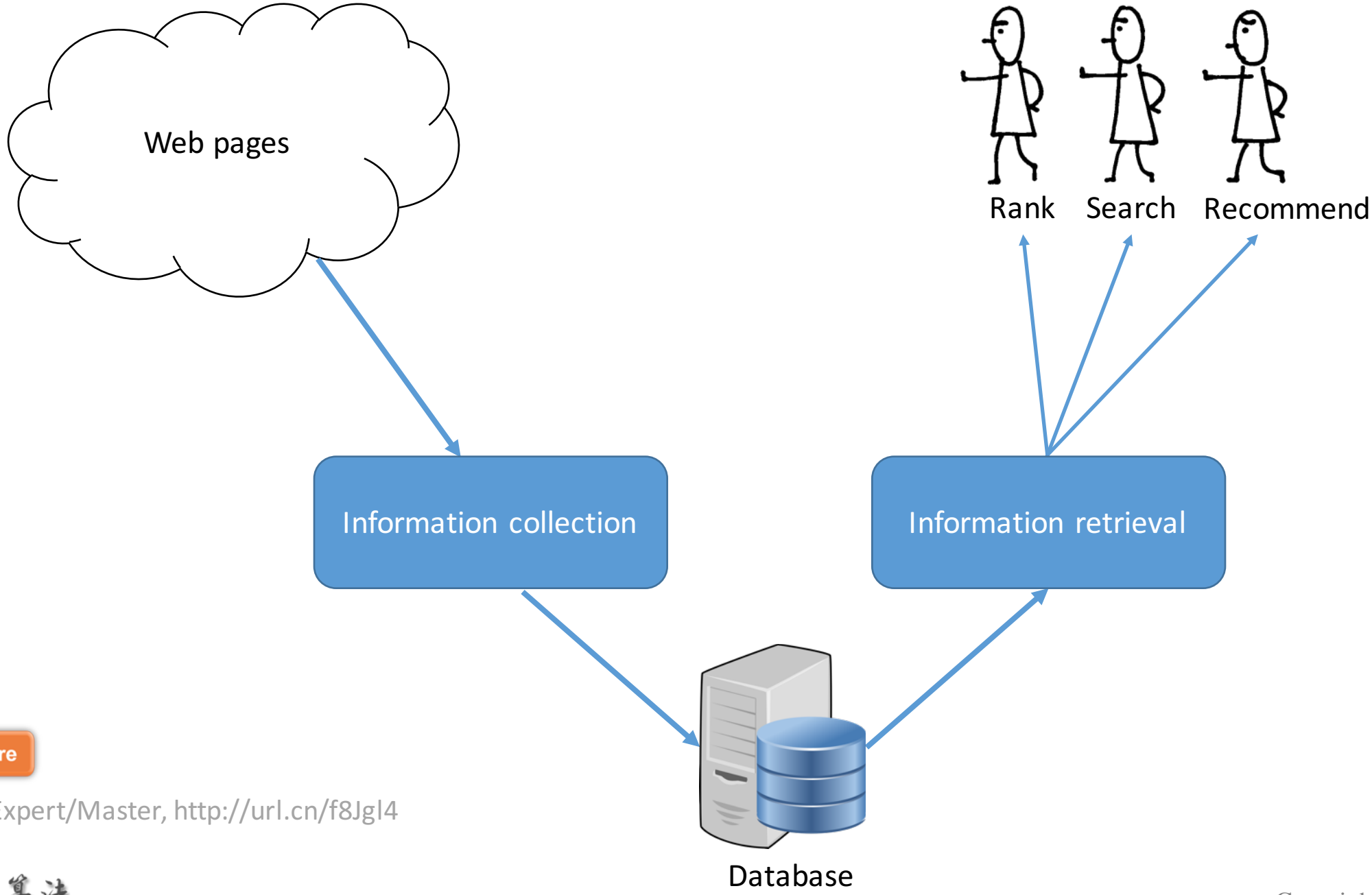
> LISTEN: Rita Ora Tells Us What To Expect From Her 'Body On Me' Video Feat Chris Brown

> LISTEN: David Guetta Previews New Song 'Sun Goes Down' With Showtek & Magic!

> WATCH: Miley Cyrus Floats Through Space As She Teases Hosting The 2015 MTV VMA Awards

> Behind The Scenes - Was One Direction's 'Drag Me Down' Vid Shot At NASA Space Station?

More Music News ➔



Read More

Novice/Expert/Master, http://url.cn/f8Jgl4

Interviewer: crawl a page

- Crawl a page from
  http://yue.ifeng.com/a/20150717/39747647_0.shtml

Crawl a page in Python 

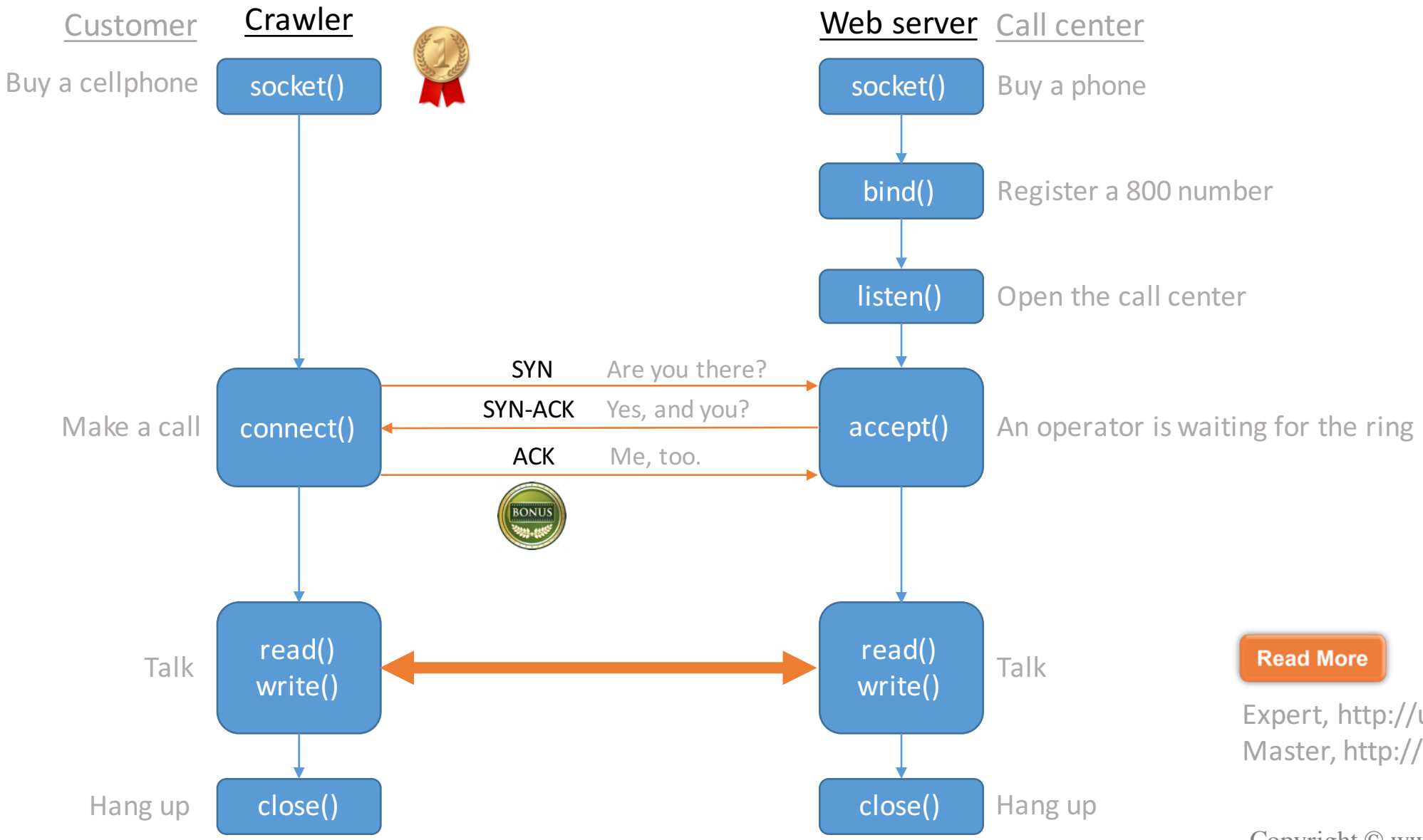
```
import urllib2


# Request source file
url = 'http://yue.ifeng.com/a/20150717/39747647_0.shtml'
request = urllib2.Request(url)          #Write a letter
response = urllib2.urlopen(request)     #Send the letter and get the reply
page = response.read()                  #Read the reply


# Save source file
webFile = open('webPage.html', 'wb')
webFile.write(page)
webFile.close()
```

Read More

Novice/Expert/Master, http://url.cn/NEHw7J
Novice/Expert, http://url.cn/ZANYur
Expert, http://url.cn/agT8mg
Expert, http://url.cn/SOBO8x
Expert/Master, http://url.cn/1eHONm

Interviewer: What is the network process when you are crawling a webpage?

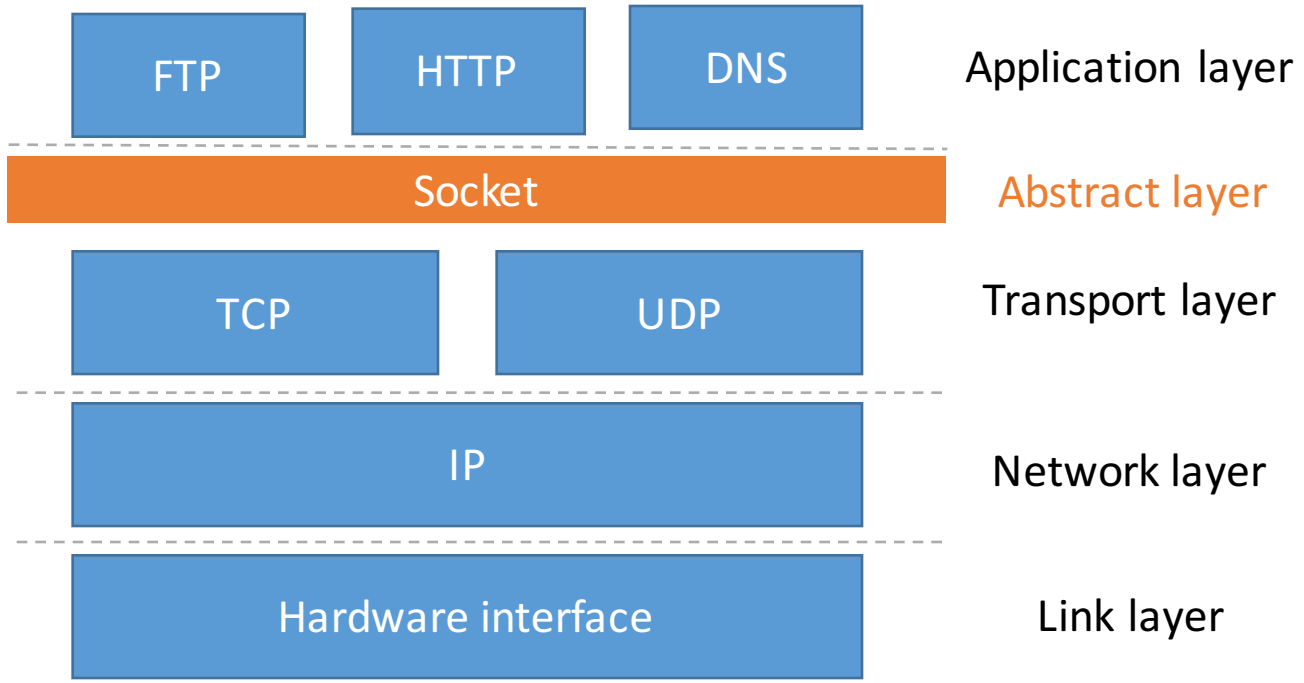
Call center VS web server



Read More

Expert, http://url.cn/PPdzXg
Master, http://url.cn/UqYdJV

Layers



Read More

Expert, http://url.cn/ZX9hMs
Master, http://url.cn/RTPIgV

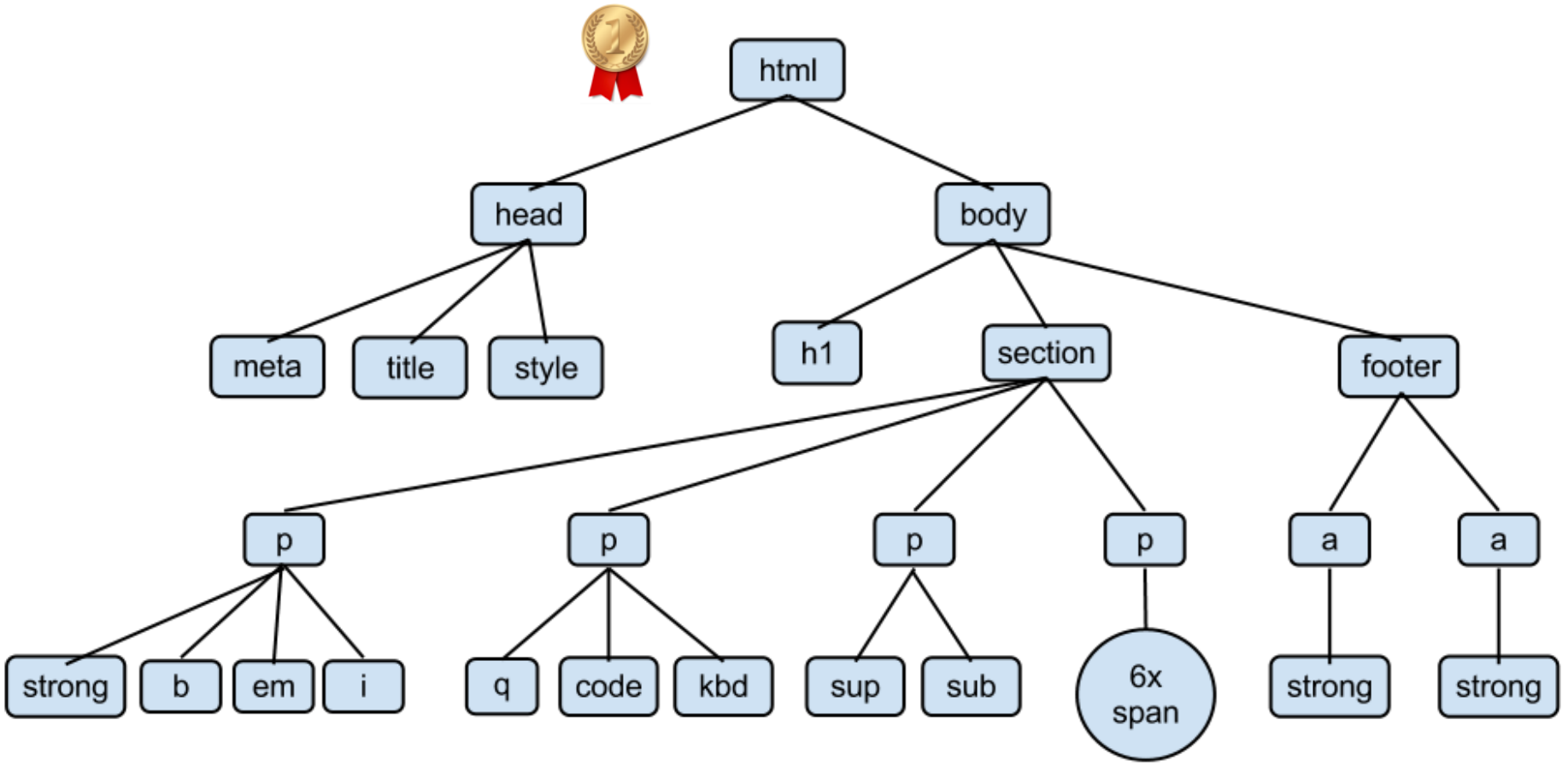
Interviewer: What is HTML?

```
<!DOCTYPE html>
<html>
<head>
<!--cmpp:start-->
    <meta charset="utf-8" />
    <title>音乐家聂耳逝世80周年 曾为国歌作曲|聂耳_凤凰音乐</title>
    <meta name="keywords" content="聂耳" />
    <meta name="description" content="1935年的7月17日，音乐家聂耳溺水逝世。他一生只经历了23个春秋，却创作出包括国歌《义勇军进行曲》、《卖报歌》、《金蛇狂舞》等许多优秀作品。 聂耳资料： 聂耳(1912－1935)，原名聂守信，字" />
        <meta name="og: webtype" content="news" />
        <meta property="og:url" content="http://yue.ifeng.com/a/20150717/39747647_0.shtml" />
        <meta property="og:title" content="音乐家聂耳逝世80周年 曾为国歌作曲" />
        <meta property="og:description" content="1935年的7月17日，音乐家聂耳溺水逝世。他一生只经历了23个春秋，却创作出包括国歌《义勇军进行曲》、《卖报歌》、《金蛇狂舞》等许多优秀作品。 聂耳资料： 聂耳(1912－1935)，原名聂守信，字"/>
        <meta name="og:time" content="2015年07月17日 10:02" />
        <meta name="og:category " content="凤凰音乐" />
        <meta property="og:image" content="http://y3.ifengimg.com/a/2015_29/dfed5a6d8dd010f_size12_w300_h409.jpg" />
        <meta name="og: img_slide" content=""/>
        <meta name="og: img_video" content=""/>
    <meta name="viewport" content="width=device-width, initial-scale=1.0, minimum-scale=1.0, maximum-scale=1.0, user-scalable=0" />
    <meta http-equiv="mobile-agent" content="format=html5;url=http://i.ifeng.com/ifengurl.f?
vt=5&ch=rj_bd_me&url=http://yue.ifeng.com/a/20150717/39747647_0.shtml" />
    <meta http-equiv="mobile-agent" content="format=xhtml;url=http://i.ifeng.com/ifengurl.f?
vt=2&ch=rj_bd_me&url=http://yue.ifeng.com/a/20150717/39747647_0.shtml" />
        <!--广告代码--><script src="http://m1.ifengimg.com/ifeng/sources/inice_v1.js"></script>
    <script>
window['recommendOpinion'] = function (selector){

    var selector = selector.split('_');
    if(selector.length >1){
        var num = jQuery('#' + selector[1]).html() /1;
        jQuery('#' + selector[1]).html(num +1);
    }

};
```

Read More

Novice/Expert/Master, http://url.cn/3Tdi9B

HTML uses a tree to represent a human



Interviewer: crawl all the news of a website

Crawl all the news from

- http://yue.ifeng.com

Identify the list page

- http://yue.ifeng.com/news/list_0/0.shtml

Identify the links of news

```
import re
```

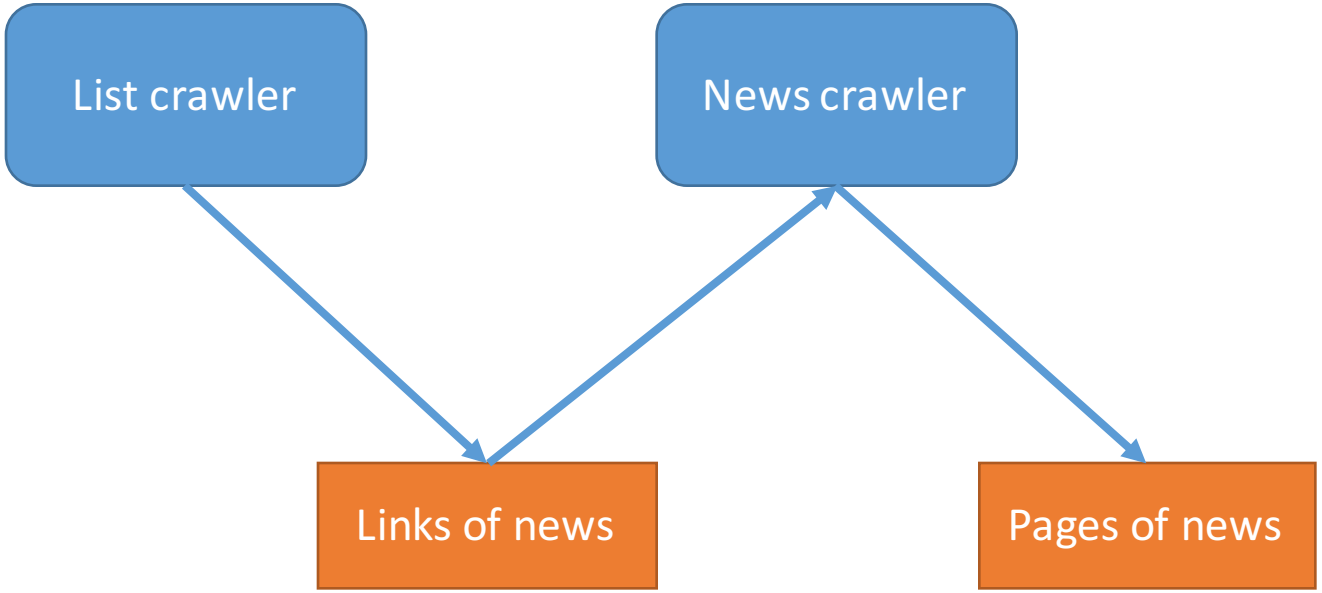
```
pattern = re.compile('<a href="http://yue.ifeng.com/news/detail_(.*?)"  target="_blank">(.*?)</a>', re.S)
items = re.findall(pattern, page)

for item in items:
    print item[0],item[1]
```

```
<li><a href="http://yue.ifeng.com/news/detail_2015_01/10/40163401_0.shtml"
target="_blank">·“歌手3”陈洁仪遭淘汰 歌迷力挺求开个唱</a><span>[01-10 03:21]</span></li>

<li><a href="http://yue.ifeng.com/news/detail_2015_01/09/40155354_0.shtml"
target="_blank">·唱作歌手吴梓涵—新专辑暖心发声</a><span>[01-09 40:29]</span></li>

<li><a href="http://yue.ifeng.com/news/detail_2015_01/09/40155188_0.shtml"
target="_blank">·罗中旭疑释前嫌献唱旧爱 好歌曲第二期看点全面开扒</a><span>[01-09 26:52]</span></li>

<li><a href="http://yue.ifeng.com/news/detail_2015_01/09/40154764_0.shtml"
target="_blank">·叶一茜发EP同届超女齐祝福 田亮惊喜现身深情献吻</a><span>[01-09 46:42]</span></li>
                    <div class="space10"></div>
<li><a href="http://yue.ifeng.com/news/detail_2015_01/08/40141951_0.shtml"
target="_blank">·“九公子”创始人荆纯：创业不能只靠粉丝经济(图)</a><span>[01-08 52:48]</span></li>
```

Read More

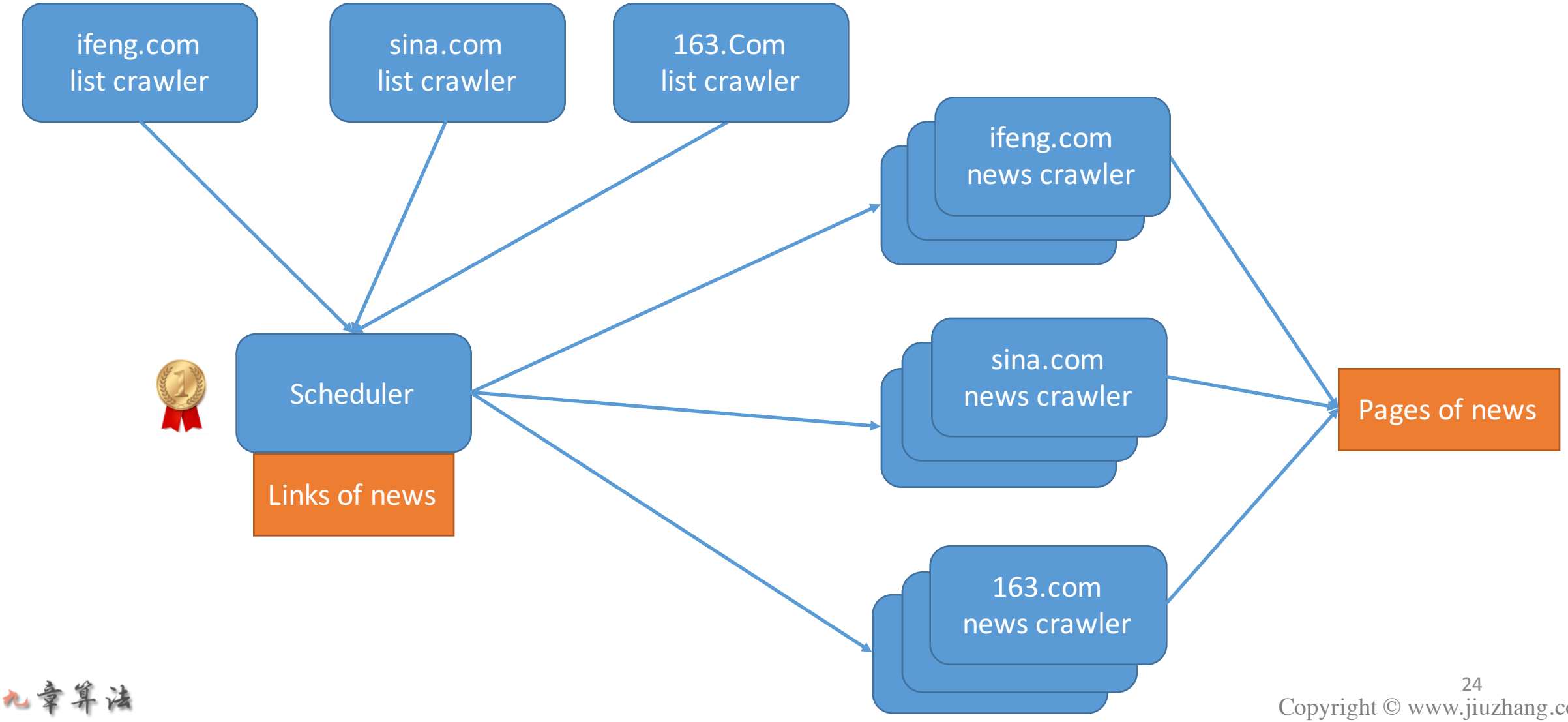
Novice/Expert, http://url.cn/TWdTap
Novice/Expert, http://url.cn/a8U9VH
Expert/Master, http://url.cn/dI93as

Architecture (v1)



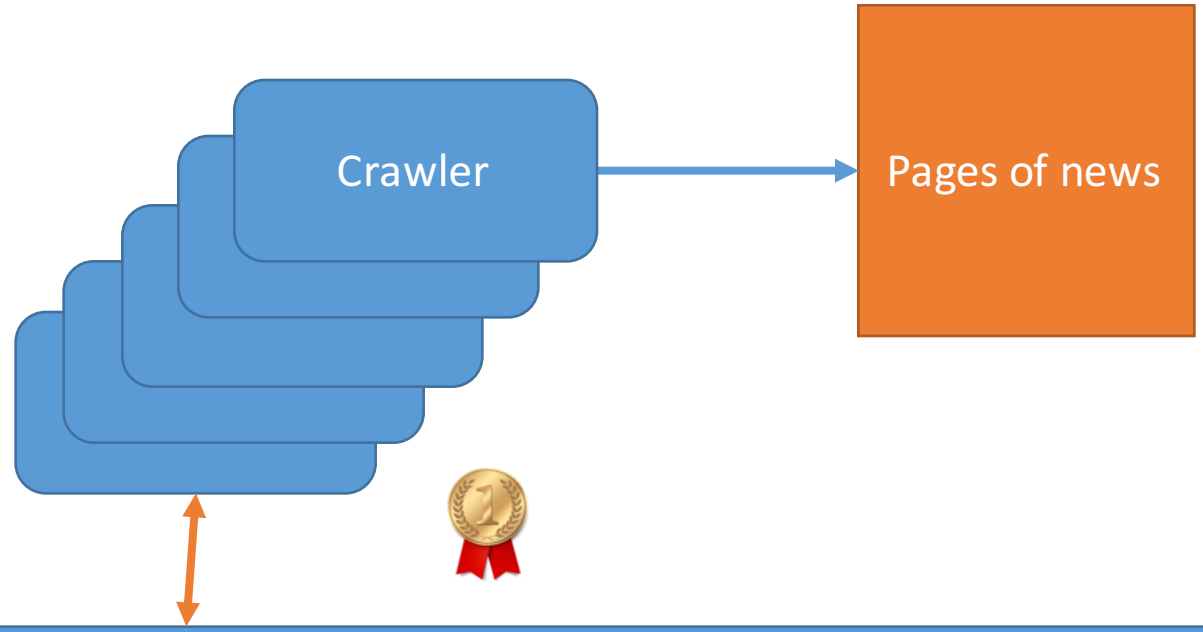
Interviewer: crawl from more websites

Architecture (v2)



Interviewer: Your crawlers are wasted!

Architecture (v3)

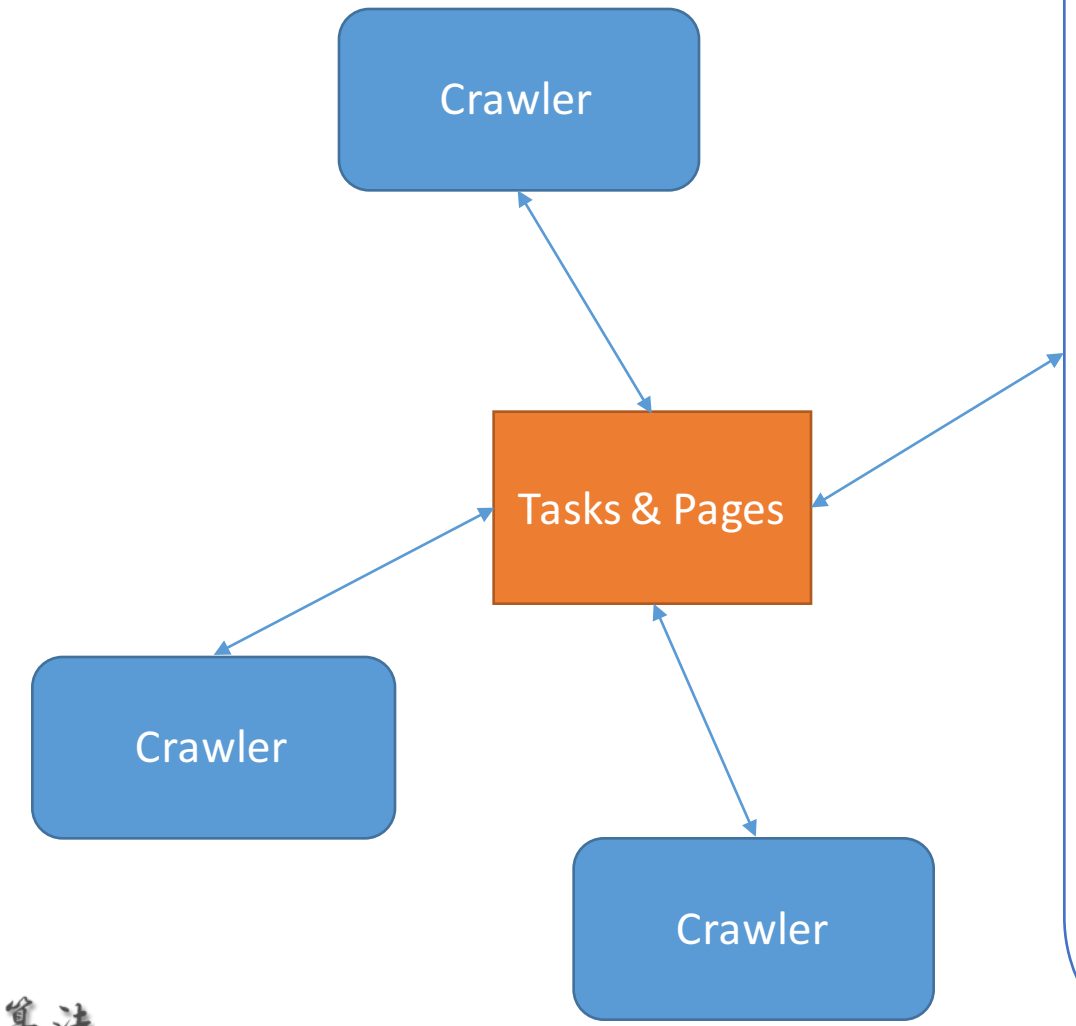


taskID	priority	type	state	link	availableTime	endTime
...	...	...	...	...	...	...
1122	0	page_ifeng	done	http://yue.ifeng.com/a/20150717/39747647_0.shtml	142113_0721_2015	142113_0721_2015
1123	0	page_ifeng	working	http://yue.ifeng.com/a/20150716/39747577_0.shtml	142114_0721_2015	
1124	0	page_sina	new	http://ent.sina.com.cn/m/c/2015-07-21/doc-ifxfaswm0925126.shtml	142115_0721_2015	
1125	1	list_ifeng	new	http://yue.ifeng.com/news/list_0/0.shtml	142216_0721_2015	
...	...	...	...	...	...	...

taskTable

Interviewer: design scheduler?

Solution with sleep

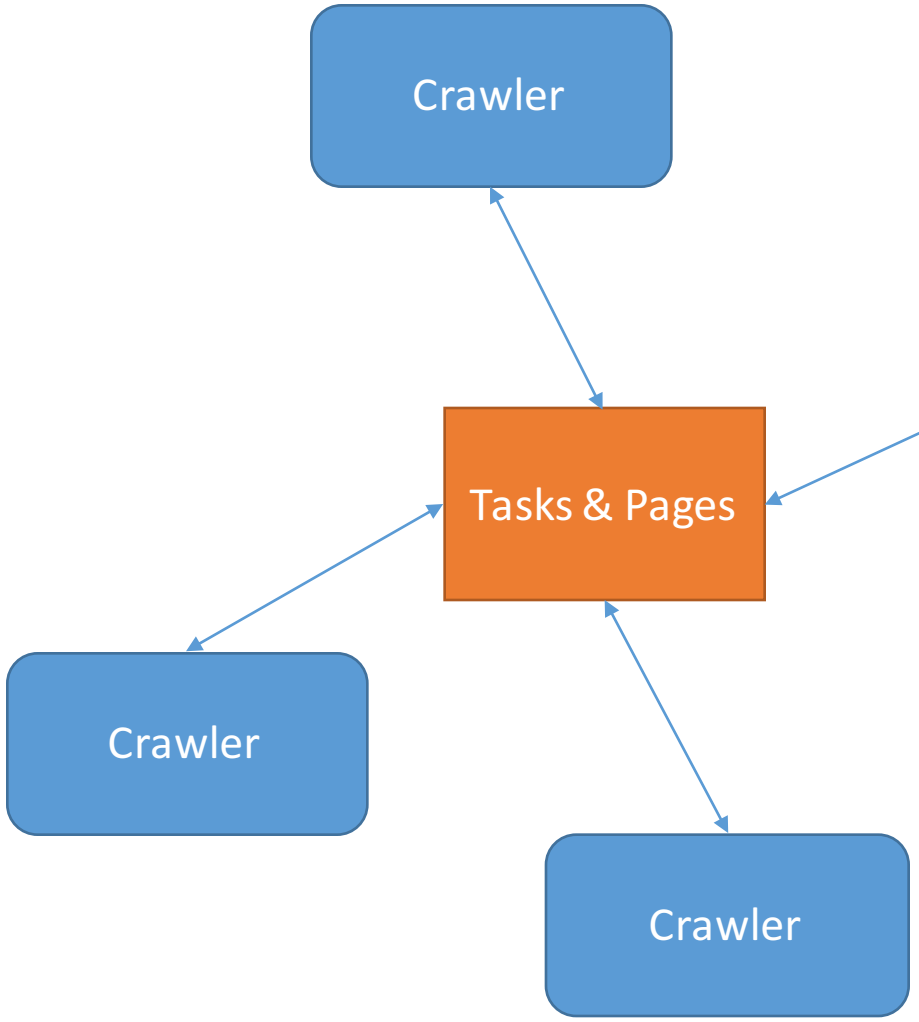


Crawler

```
While(true)
  Lock(taskTable)
  While taskTable.Find("state=new") == NULL
    Release(taskTable)
    Sleep(1s)
    Lock(taskTable)

  task = taskTable.FindOne("state=new")
  task.state = "working"
  Release(taskTable)

  page = Crawl(task.url)

  If task.type == "list"
    Lock(taskTable)
    For newTask In page:
      taskTable.Add(newTask)
    task.state = "done"
    Release(taskTable)
  Else
    Lock(pageTable)
    pageTable.Add(page)
    Release(pageTable)
    Lock(taskTable)
    task.state = "done"
    Release(taskTable)
```

Interviewer: design scheduler with conditional variable

Solution with conditional variable



Crawler

```
While(true)
  Lock(taskTable)
  While( taskTable.Find("state=new") == NULL )
    Cond_Wait (cond, taskTable)
  task = taskTable.FineOne("state=new")
  task.state = "working"
  Release(taskTable)

  page = Crawl(task.url)

  If task.type == "list"
    Lock(taskTable)
    For newTask In page:
      taskTable.Add(newTask)
      Cond_Signal(cond)
    task.state = "done"
    Release(taskTable)
  Else
    Lock(pageTable)
    pageTable.Add(page)
    Release(pageTable)
    Lock(taskTable)
    task.state = "done"
    Release(taskTable)
```

Conditional variable

```
Cond_Wait(cond, mutex)
    Lock(cond.threadWaitList);
    cond.threadWaitList.Add(this.thread);
    Release(cond.threadWaitList);

    Release(mutex) ;
    Block(this.thread);
    Lock(mutex);
```

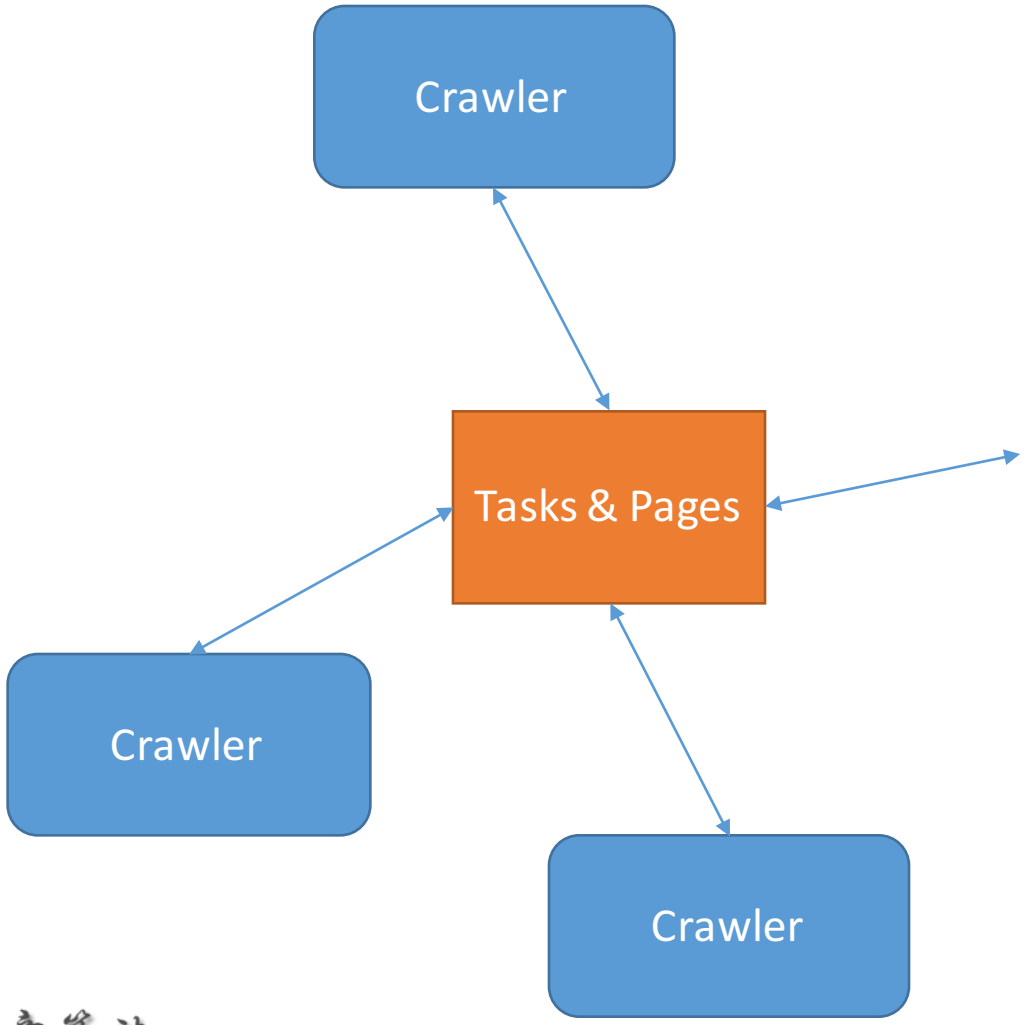
```
Cond_Signal(cond)
    Lock(cond.threadWaitList);
    If( cond.threadWaitList.size()>0 );
        thread = cond.threadWaitList.Pop();
        Wakeup(thread);
    Release(cond.threadWaitList);
```

Read More

Novice, http://url.cn/44qusn
Novice, http://url.cn/cYsLZ3
Expert, http://url.cn/coAAeW
Expert, http://url.cn/TNVxRu
Expert, http://url.cn/edly1C

Interviewer: design scheduler with semaphore

Solution with semaphore



Crawler

```
While(true)
    Wait(numberOfNewTask)
    Lock(taskTable)
    task = taskTable.FindOne("state=new")
    task.state = "working"
    Release(taskTable)

    page = Crawl(task.url)

    If task.type == "list"
        Lock(taskTable)
        For newTask In page:
            taskTable.Add(newTask)
            Signal(numberOfNewTask)
        task.state = "done"
        Release(taskTable)
    Else
        Lock(pageTable)
        pageTable.Add(page)
        Release(pageTable)
        Lock(taskTable)
        task.state = "done"
        Release(taskTable)
```

Semaphore

```
Wait(semaphore)
    Lock(semaphore);
    semaphore.value--;
    If( semaphore.value<0 )
        semaphore.processWaitList.Add(this.process);
        Release(semaphore);
        Block(this.process);
    Else
        Release(semaphore);
```

```
Signal(semaphore)
    Lock(semaphore);
    semaphore.value++;
    If( semaphore.value<=0 )
        process = semaphore.processWaitList.Pop();
        Wakeup(process);
    Release(semaphore)
```

Read More

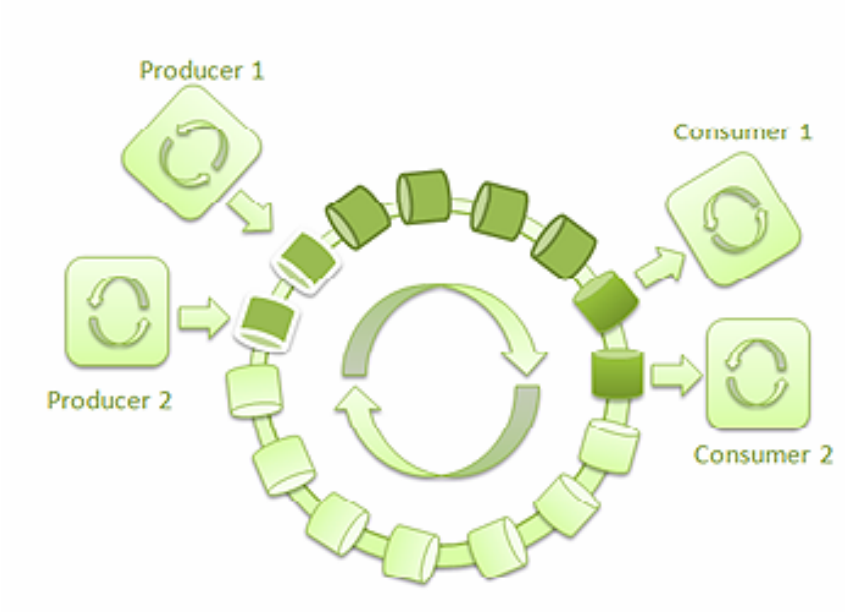
Novice, http://url.cn/XWQJWo
Novice, http://url.cn/g7xO6G
Expert, http://url.cn/daQTsi
Expert, http://url.cn/YUG6Nd
Expert/Master, http://url.cn/a9NI3i

Interviewer: design the fastest consumers and producers

Thread-safe consumer and producer

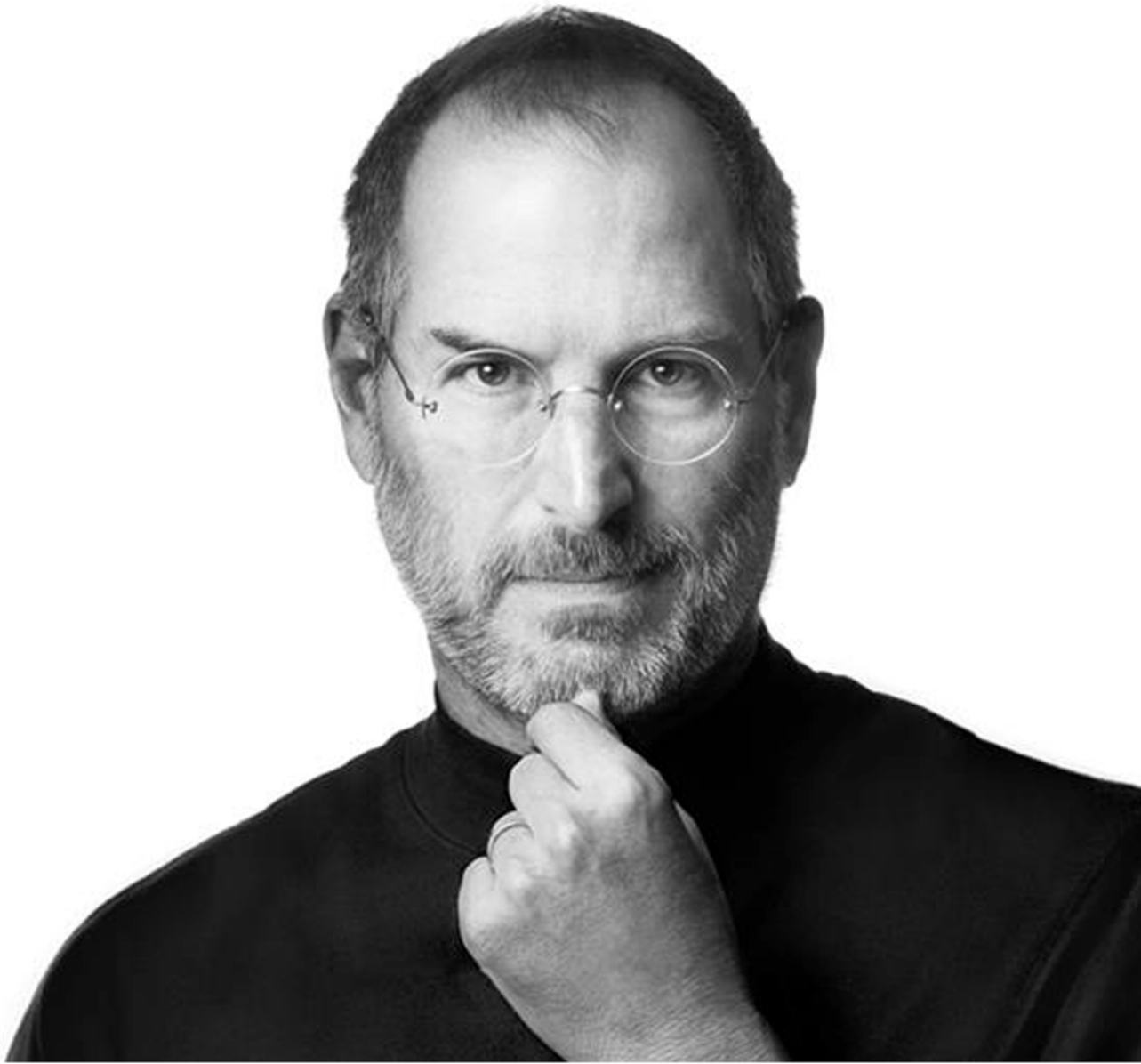
Read More

Novice, http://url.cn/cNG8WK
Expert, http://url.cn/cFPpaN
Expert, http://url.cn/C1lIpq
Expert/Master, http://url.cn/6CCw4i
Master, http://url.cn/9kdFF2
Master, http://url.cn/ORvrlX
Master, http://url.cn/dYRtFj



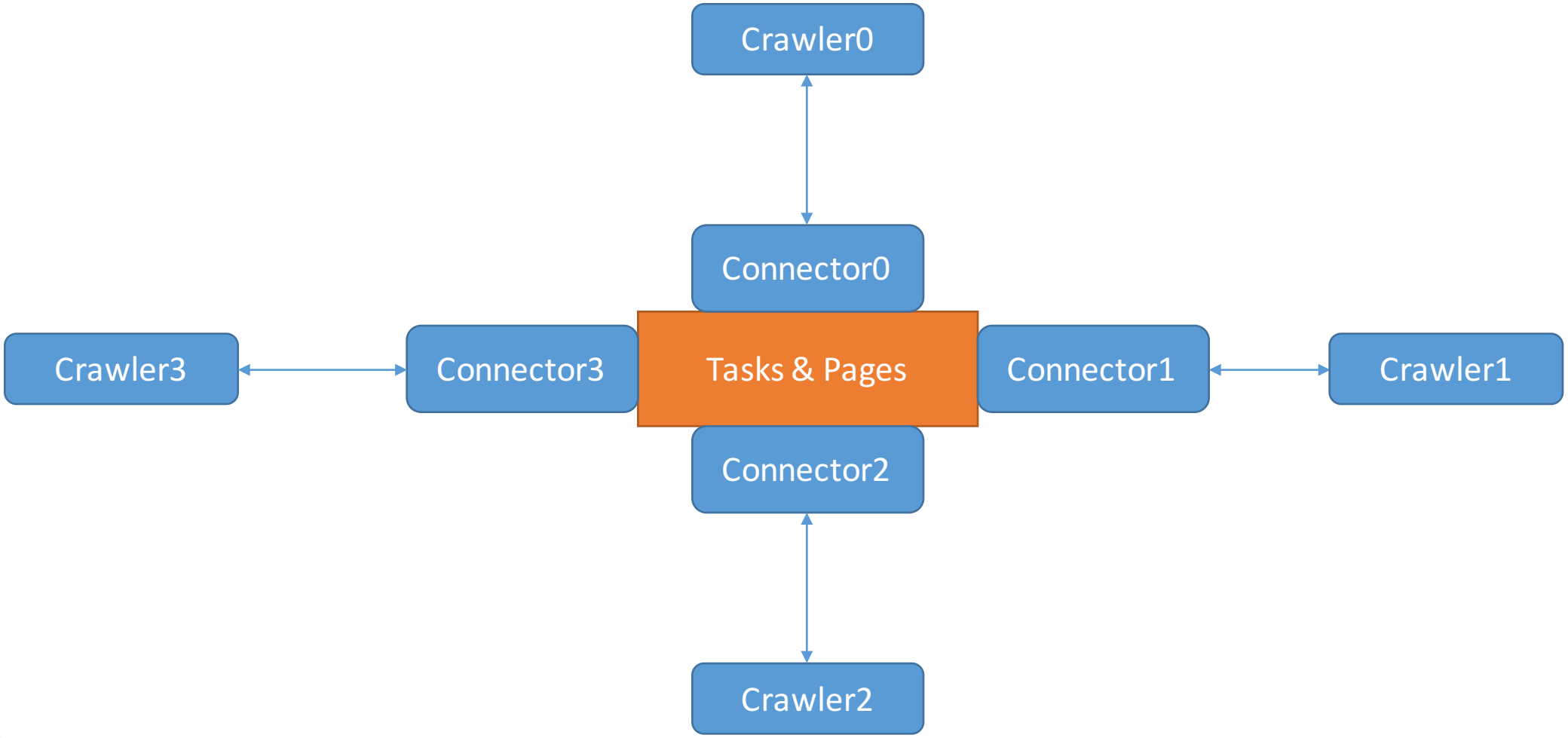
LMAX Disruptor

Stay Hungry,
Stay Foolish.



Interviewer: distribute crawlers in multiple machines

Solution with one-to-one



Connector0

```
crawler = WaitConnection()

While(true)
   Wait(numberOfNewTask)
   Lock(taskTable)
   task = taskTable.FindOne("state=new")
   task.state = "working"
   Release(taskTable)

   Send(crawler0, task)
   page = Receive(crawler0)

   ...
   ...
   ...
```

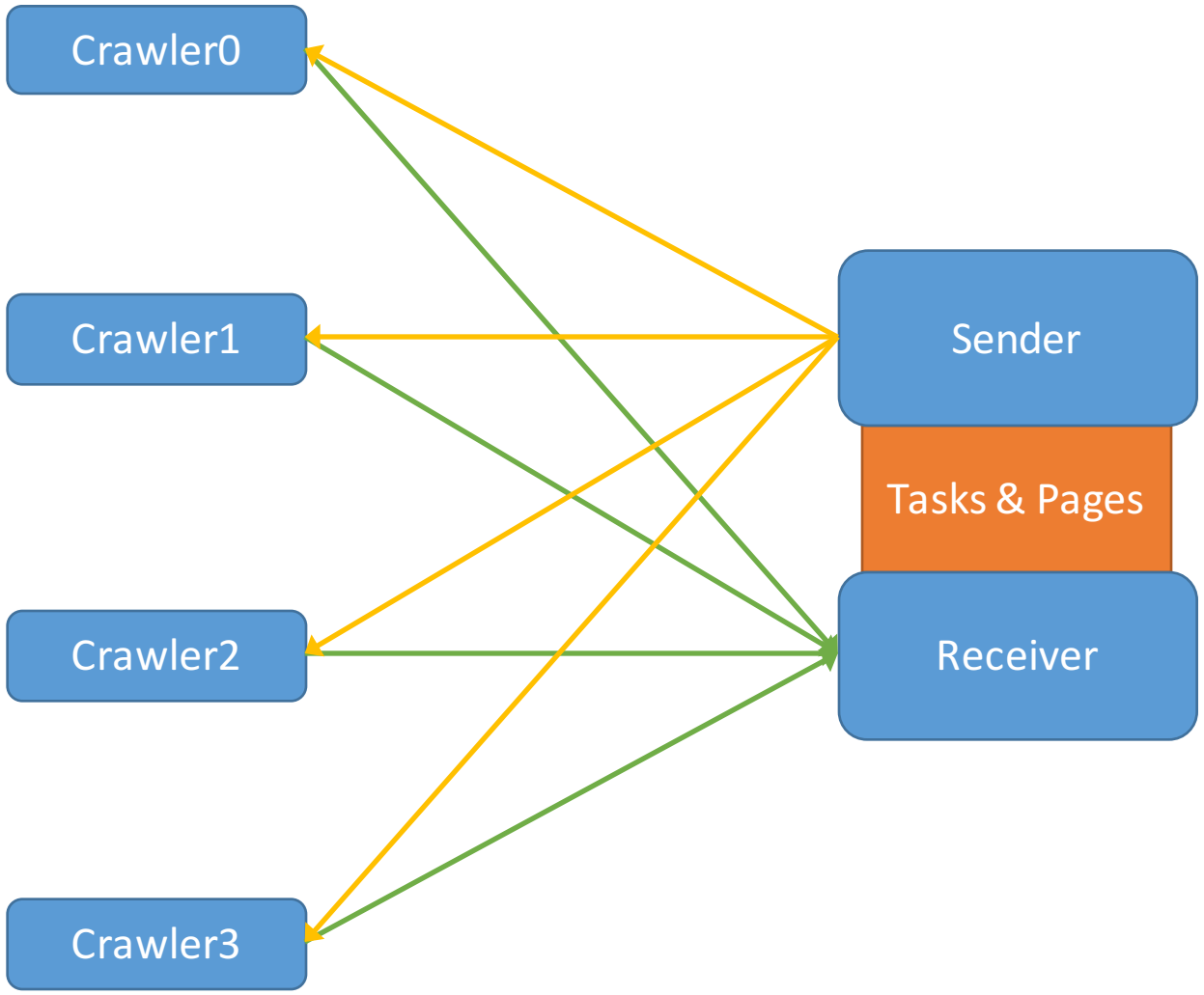
Crawler0

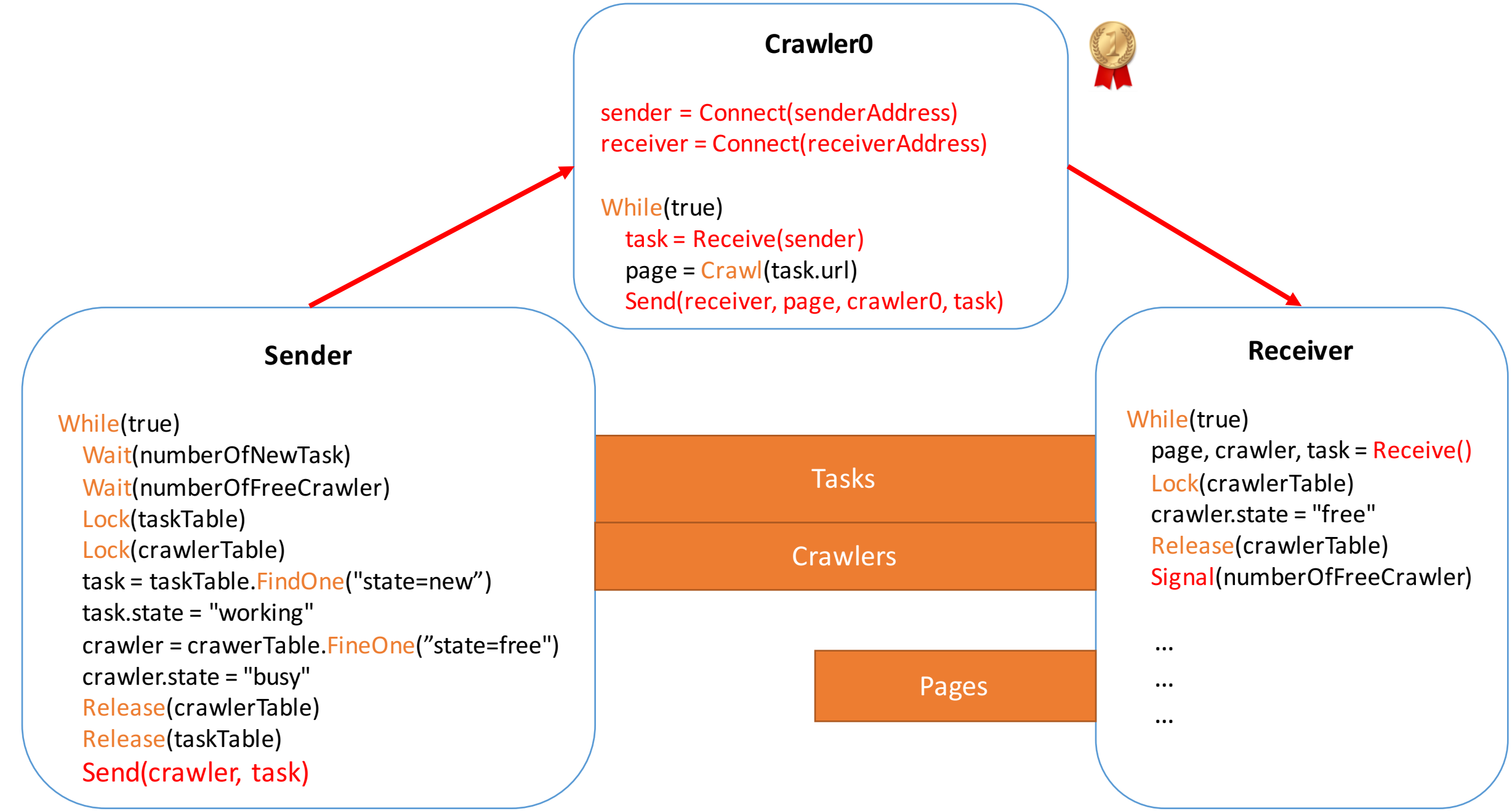
```
connector  = Connect(connector0Address)

While(true)
   task = Receive(connector)
   page = Crawl(task.url)
   Send(connector, page)
```

Tasks & Pages

Interviewer: Can you reduce the number of connectors?

Solution with one-to-many





Interviewer: Can you use database?

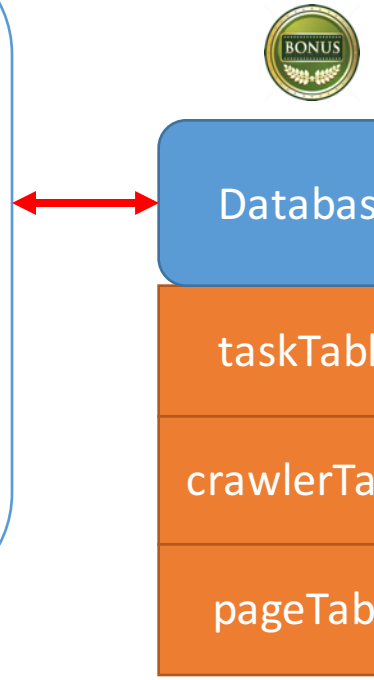
Solution with database

Crawler0

```
sender = Connect(senderAddress)
receiver = Connect(receiverAddress)

While(true)
    task = Receive(sender)
    page = Crawl(task.url)
    Send(receiver, page, crawler0, task)
```

Sender

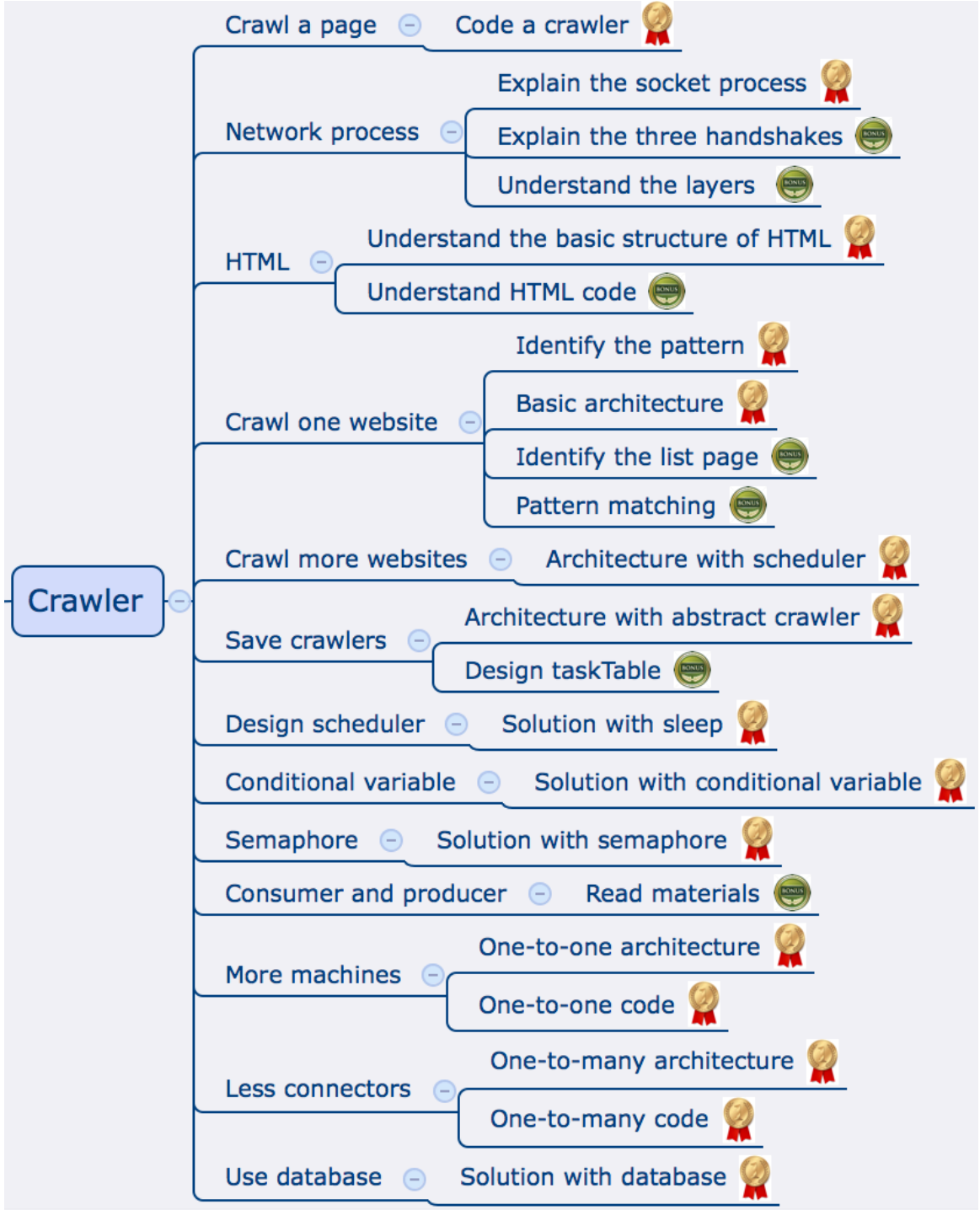
```
While(true)
    Wait(numberOfNewTask)
    Wait(numberOfFreeCrawler)
    task = Database.FindOne(task.state="new")
                   .Update(task.state="working")
    crawler = Database.FindOne(crawler.state="free")
                      .Update(crawler.state ="busy")
    Send(crawler, task)
```



Receiver

```
While(true)
    page, crawler, task = Receive()
    Database.Update(crawler.state="free")
    Signal(numberOfFreeCrawler)

    If task.type == "list"
        For newTask In page
            Database.Add(newTask)
            Signal(numberOfNewTask)
    Else
        Database.Add(page)
    Database.Update(task.state="done")
```

Crawler Summary (15+7)



Interviewer: design Tiny URL

Read More

Novice, https://goo.gl/
Novice/Expert, http://url.cn/bycYvq
Novice/Expert/Master, http://url.cn/aXRa8e
Expert, http://url.cn/euBds7

Let’s do it together ☺

Google url shortener

Paste your long URL here:

Google

Shorten URL

All goo.gl URLs and click analytics are public and can be accessed by anyone.

Clicks for the past: two hours | day | week | month | **all time**

LONG URL	CREATED	SHORT URL		CLICKS
www.jiuzhang.com/course/2	3 days ago	goo.gl/9OTcuV	Details	563

Crack a design in 5 steps

- Scenario: case/interface
- Necessary: constrain/hypothesis

- Application: service/algorithm
- Kilobit: data

- Evolve

Interviewer: What is the Scenario?

Scenario 

short_url Insert(long_url)
long_url Lookup(short_url)

Interviewer: What is the Necessary?

Necessary for 1million daily users 

- Daily active users
  - 1,000,000
- Insert
  - Per day: 1,000,000 * 1%(function usage) * 10(function frequency) = 100,000
  - Per year: 100,000*365 = 36,500,000
  - Per second: 100,000/86400 = 1.2
- Lookup
  - Per day: 1,000,000* 100%(function usage) * 3(function frequency) = 3,000,000
  - Per second: 3,000,000/86400 = 35

Interviewer: What is the Algorithm?

Insert 

LongToShort

longURL	shortURL
www.jiuzhang.com/course/2/	…
www.inoreader.com/all_articles	…
blog.csdn.net/longyulu/article/details/9159589	…
…	…

ShortToLong

shortURL	shortURL
…	www.jiuzhang.com/course/2/
…	www.inoreader.com/all_articles
…	blog.csdn.net/longyulu/article/details/9159589
…	…

```
class Shortener{
    map<string,string> mLongToShort
    map<string,string> mShortToLong

    string Insert( string longURL )
        If( mLongToShort.Find( longURL ) == NULL )
            string shortURL = GenerateShortURL();
            mLongToShort[ longURL ] = shortURL;
            mShortToLong[ shortURL ] = longURL;
        Return shortURL;
};
```

GenerateShortURL 

LongToShort

longURL	shortURL
www.jiuzhang.com/course/2/	0
www.inoreader.com/all_articles	1
http://blog.csdn.net/longyulu/article/details/9159589	2
…	…

ShortToLong

shortURL	shortURL
0	www.jiuzhang.com/course/2/
1	www.inoreader.com/all_articles
2	http://blog.csdn.net/longyulu/article/details/9159589
…	…

```
string GenerateShortURL()
    Return string( mLongToShort.size() );
```

Interviewer: How to reduce the size of shortURL?

Selection of characters 

	Before	After
Yearly URL	36,500,000	36,500,000
Usable characters	[0-9] = 10	[0-9a-zA-Z] = 62
Encoding length	$Log_{10}(36{,}500{,}000) = 7.6 = 8$	$Log_{62}(36{,}500{,}000) = 4.2 = 5$
Example	goo.gl/36500000	goo.gl/2t9jG

GenerateShorterURL 

```
string GenerateShortURL( string longURL )
  Return ConvertTo62( LongToShort.size() );

string ConvertTo62(int number)
    char Encode[62] = {‘0’,…,’9’, ’a’, …,’z’, …, ’A’,…,’Z’};
    string ret = "";
    While(number)
        ret = Encode[ number%62 ] + ret;
        number/=62;
    Return ret;
```

Interviewer: What is the Kilobit?

Data 

Average size of longURL = 100 bytes
Average size of shortURL = 4 bytes (int)
State = 4 byte
Daily new URL = 100,000 * 108 = 10.8MB
Yearly new URL = 10.8*365 = 4GB

URL table (MySQL or NoSQL)

longURL (varchar)	shortURL (int)	state (int)
www.jiuzhang.com/course/2/	0	0
www.inoreader.com/all_articles	1	0
http://blog.csdn.net/longyulu/article/details/9159589	2	1
...	...	0

Follow up 

- How to support random?
  - Random(0,range)
- How to avoid conflicting?
  - Try again
- How to implement time-limited service?
  - Expire/state
- How to cache?
  - Pre-load
  - Replacement

Read More

Master, http://url.cn/dr4uux

Interviewer: How to support analysis?

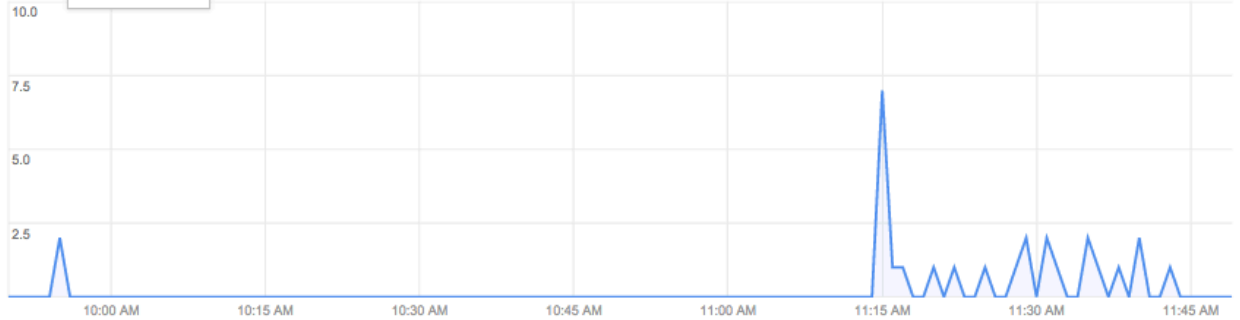
http://goo.gl/9OTcuV

http://www.jiuzhang.com/course/2/

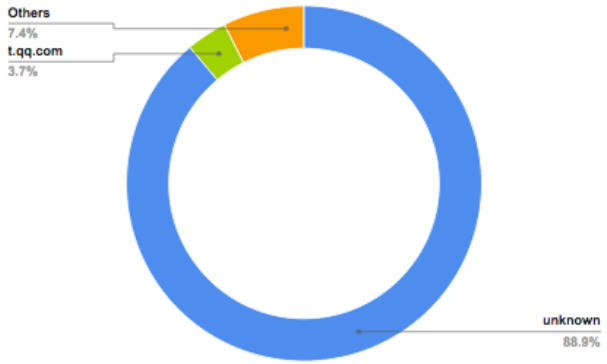
Created: 2 days ago

Total Clicks
27

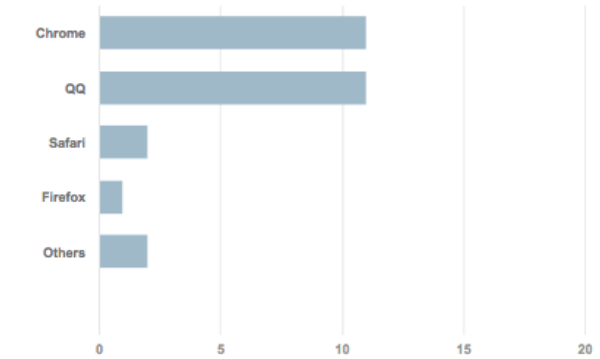
Clicks for the past: **two hours** | day | week | month | all time



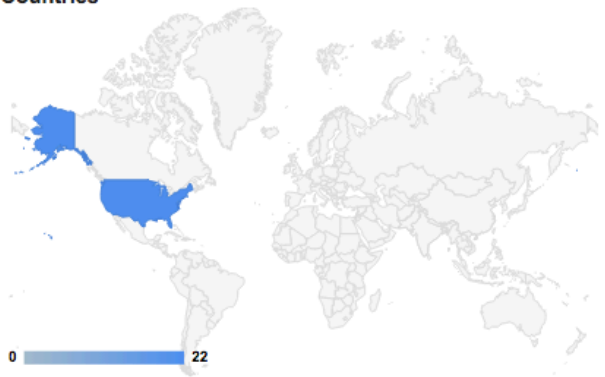
Referrers



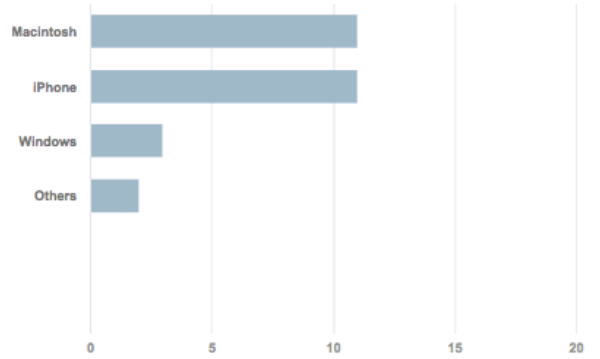
Browsers



Countries

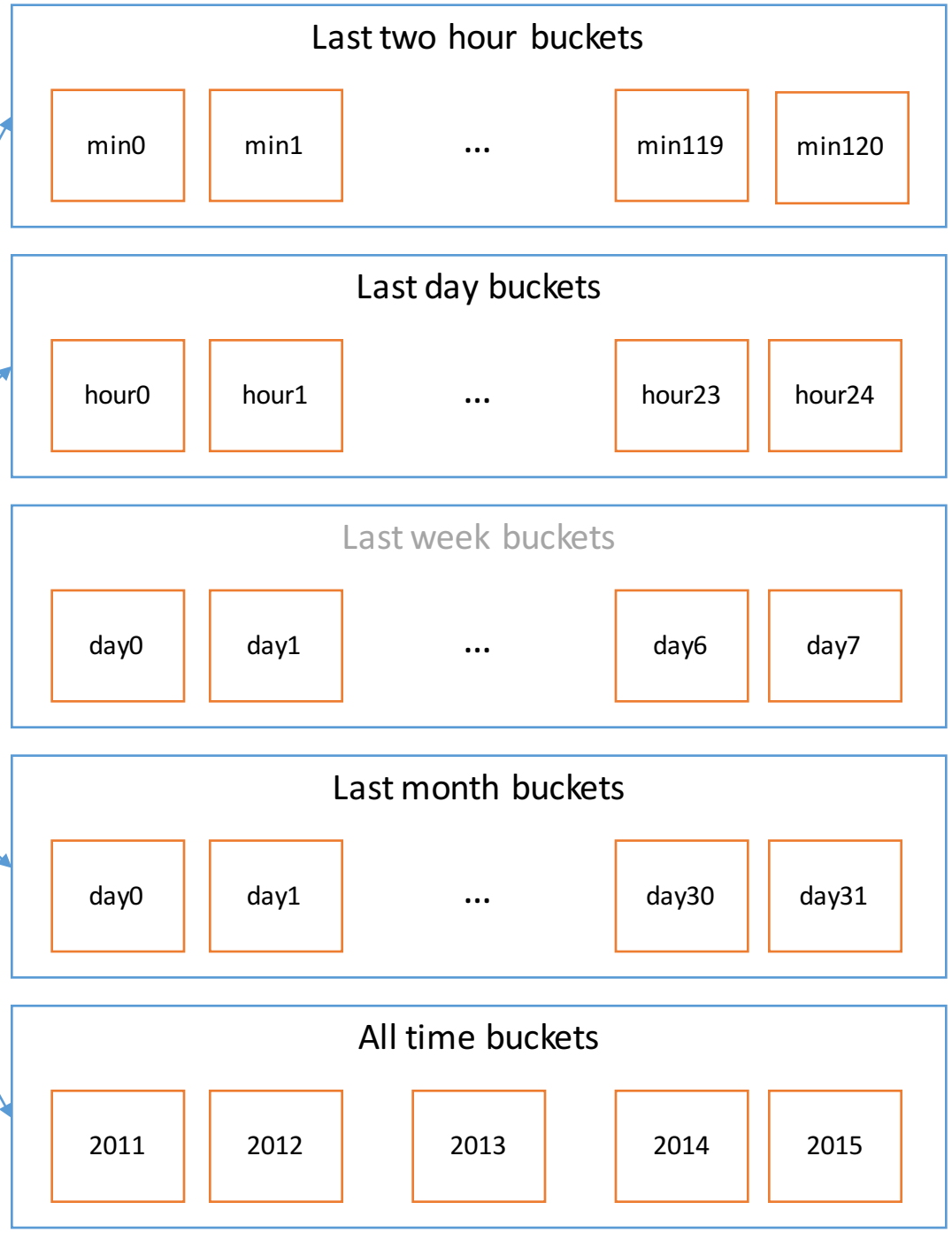


Platforms

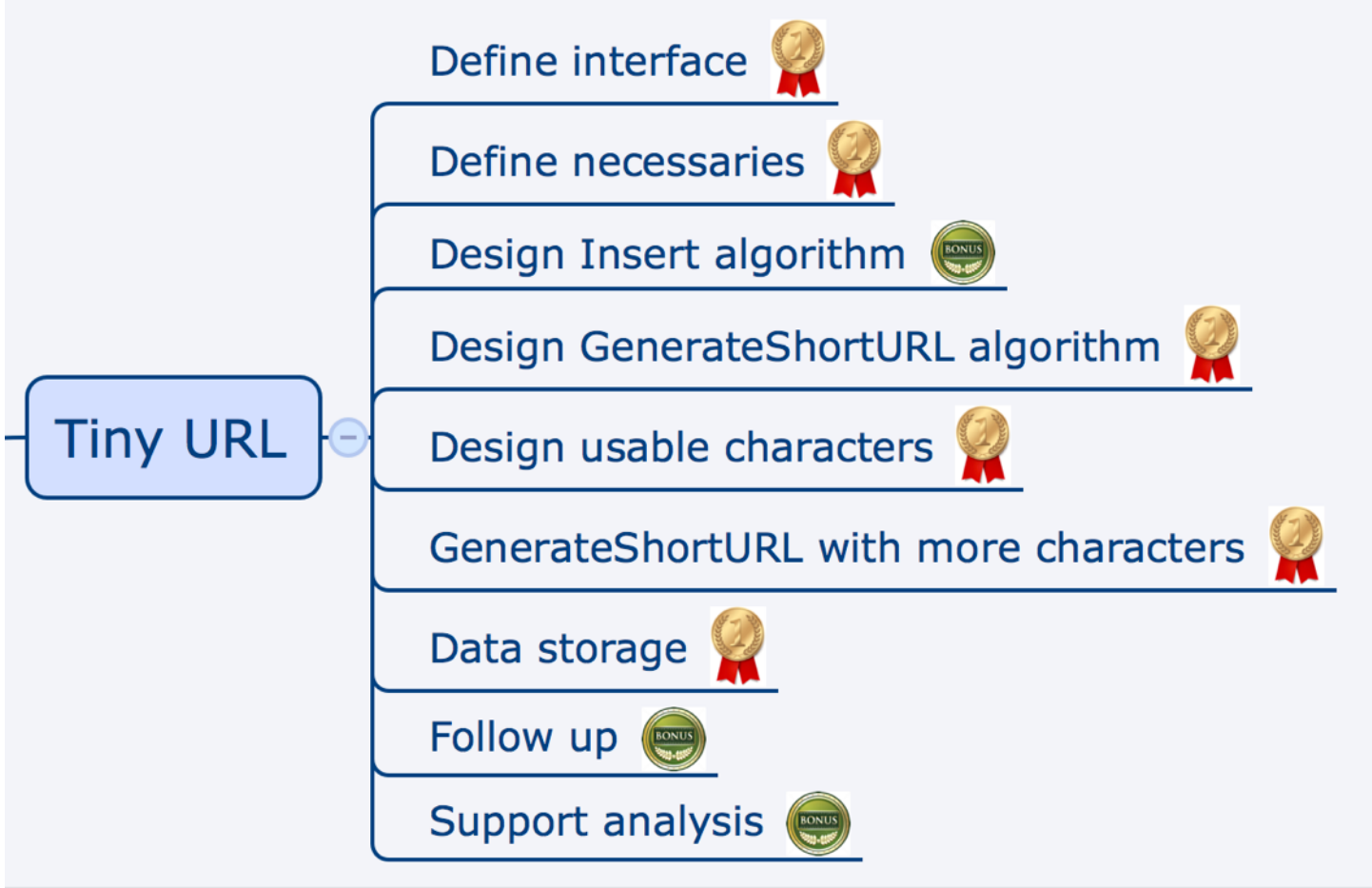


A log

- Date: 2015-08-18
- Time: 11:40
- Referrer: t.qq.com
- IP: 103.235.3.4
- Country: United states
- To: 9OTcuV
- Browser: Chrome
- Platform: Macintosh

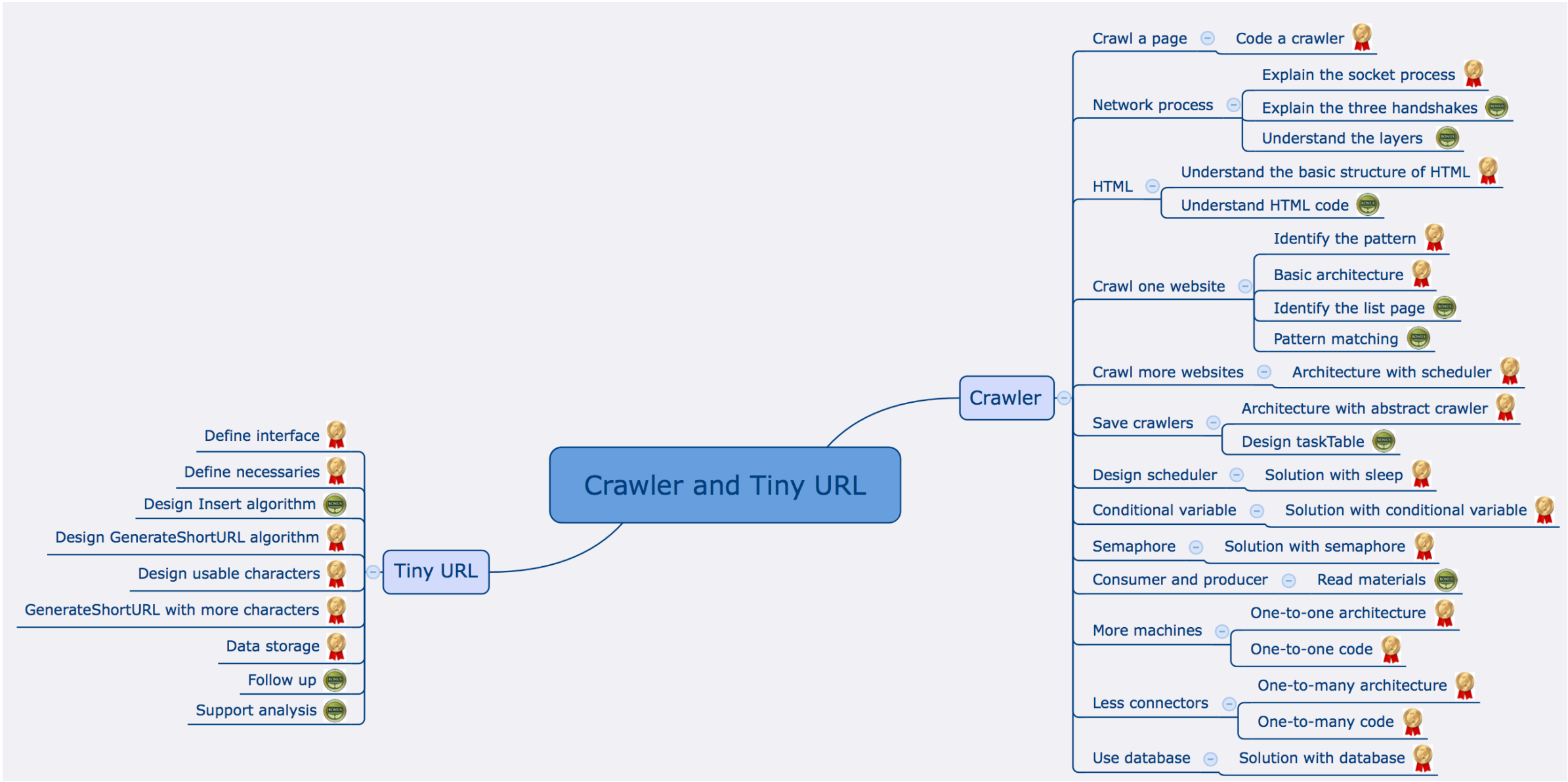


Summary of Tiny URL



Read More Novice, http://www.xn--vi8hiv.ws/

Class3 summary



Keyword

- Crawler
- HTTP
- Socket
- HTML
- TCP
- Thread
- Lock
- Conditional variable
- Semaphore
- URL
- Insert
- Lookup
- Encoding

QA

关注微信/微博，获取最新面试题及权威解答

微信: ninechapter

微博: http://www.weibo.com/ninechapter

官网: www.jiuzhang.com

One more thing

- How to ask?
  1. Answer the question by yourself
  2. Google solutions/unknown concepts
  3. Summarize your solution in details in QA
  4. Ask your question directly and clearly

能工摹其形，巧匠摄其魂