Архитектура Поиска Яндекса

На основе лекции для Малого ШАДа

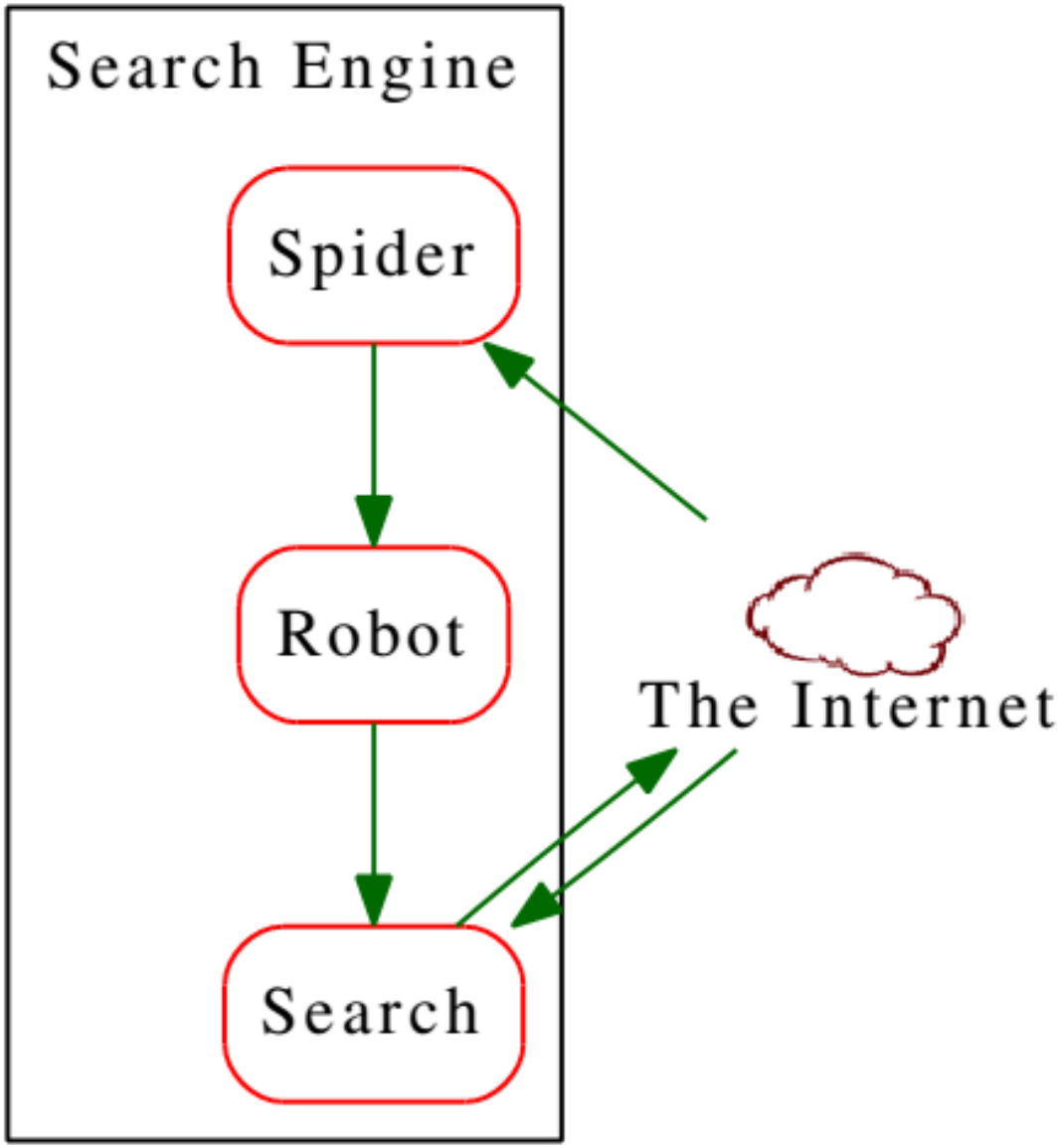
https://habrahabr.ru/company/yandex/blog/204282/

Поисковая машина

Паук
ходит по интернету и пытается выкачать как можно больше информации.

Робот
обрабатывает документы таким образом, чтобы по ним было проще искать.

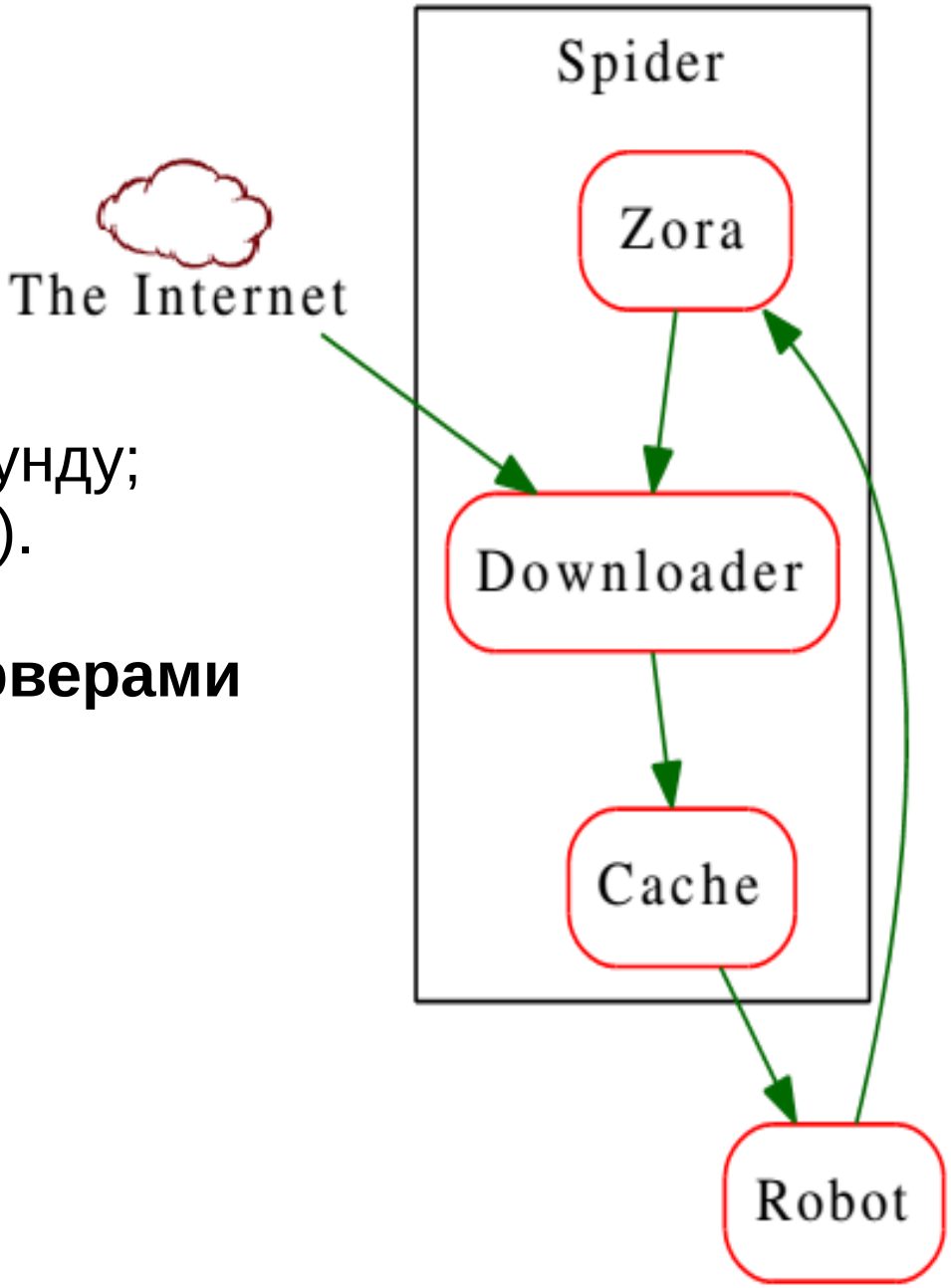
Непосредственно поиск
получает запросы и выдает ответы.



Паук

Рунет
Качающие сервера: 300;
Нагрузка: 20 000 документов в секунду;
Трафик: 400 МБайт/с (3200 Мбит/с).

Что если Паук всеми своими серверами начнет скачивать один сайт?

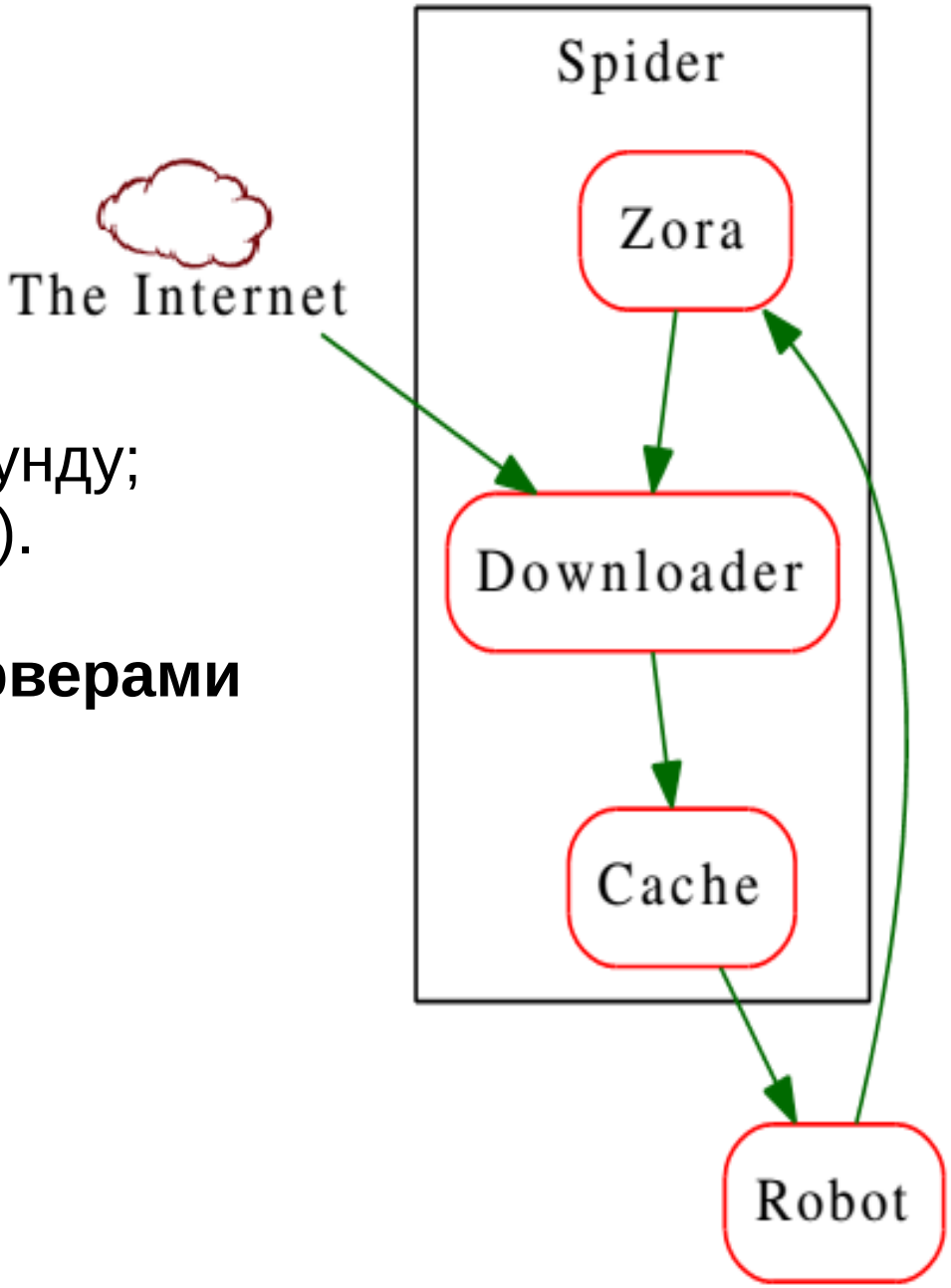


Паук

Рунет
Качающие сервера: 300;
Нагрузка: 20 000 документов в секунду;
Трафик: 400 МБайт/с (3200 Мбит/с).

Что если Паук всеми своими серверами начнет скачивать один сайт?
достаточно мощная DDoS-атака

Что тогда делать?

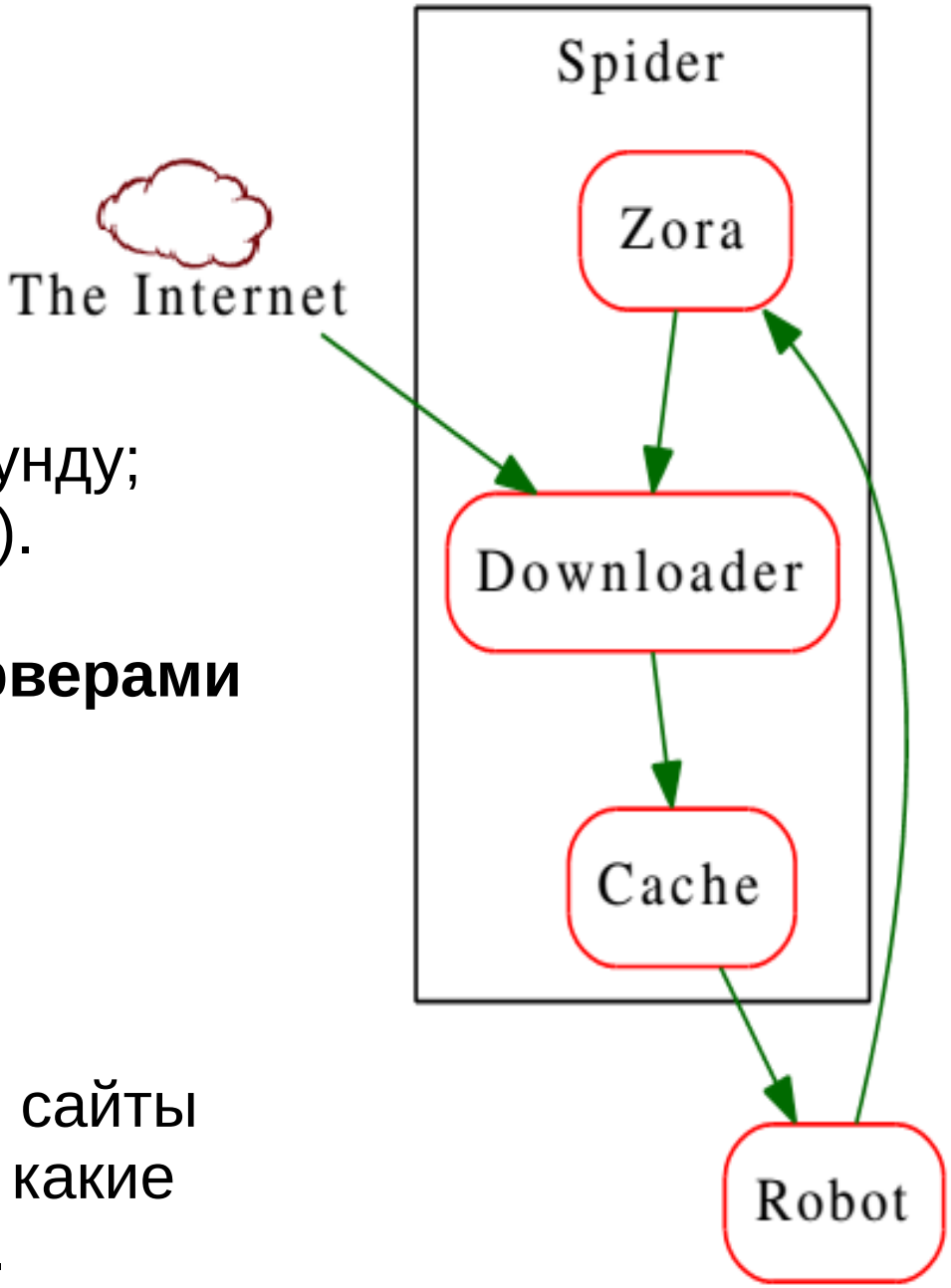


Паук

Рунет
Качающие сервера: 300;
Нагрузка: 20 000 документов в секунду;
Трафик: 400 МБайт/с (3200 Мбит/с).

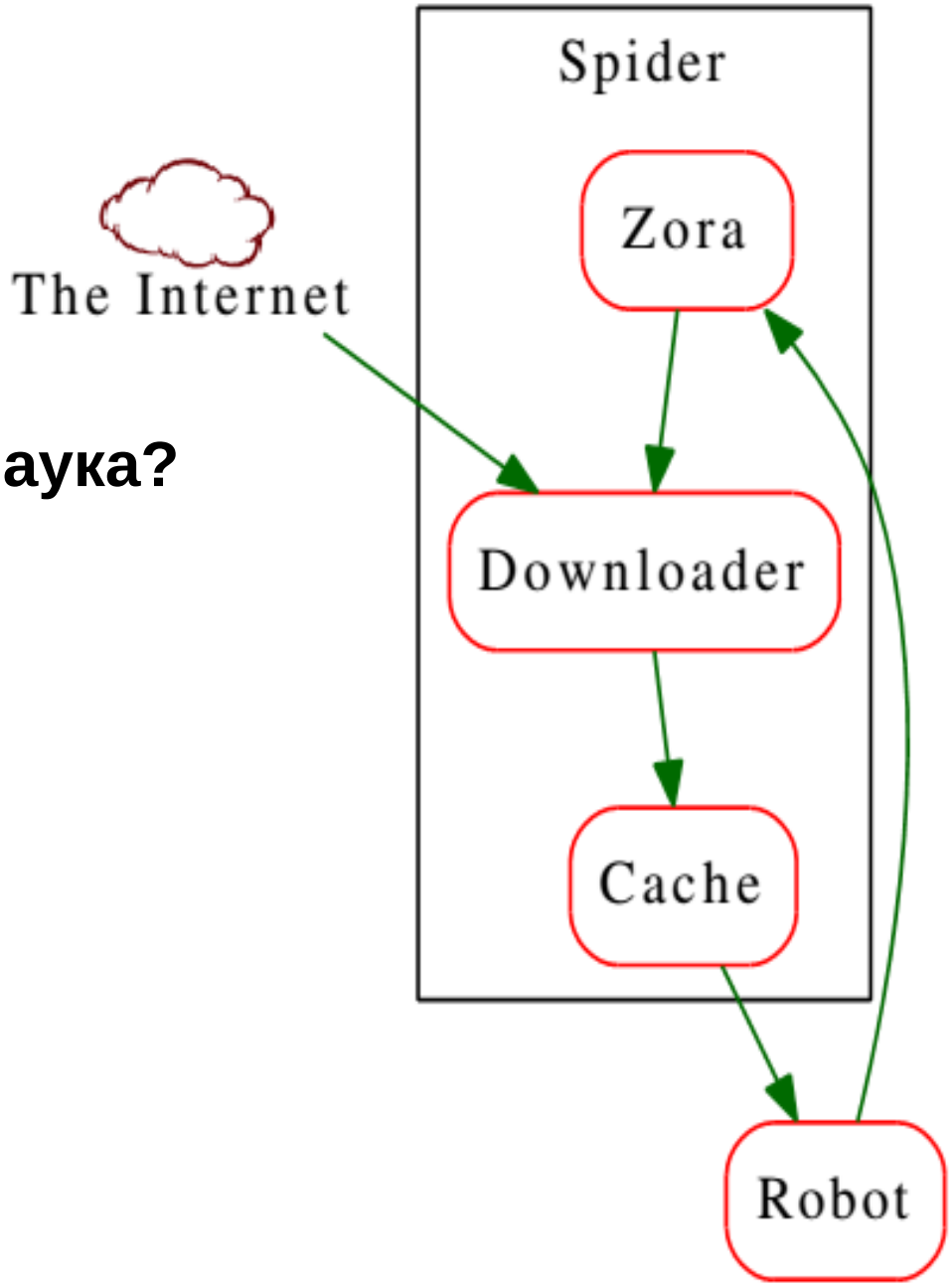
Что если Паук всеми своими серверами начнет скачивать один сайт?
достаточно мощная DDoS-атака

Что тогда делать?
Zora
координирует закачки, знает, какие сайты были закэшированы недавно, а на какие стоит сходить в ближайшее время.



Паук

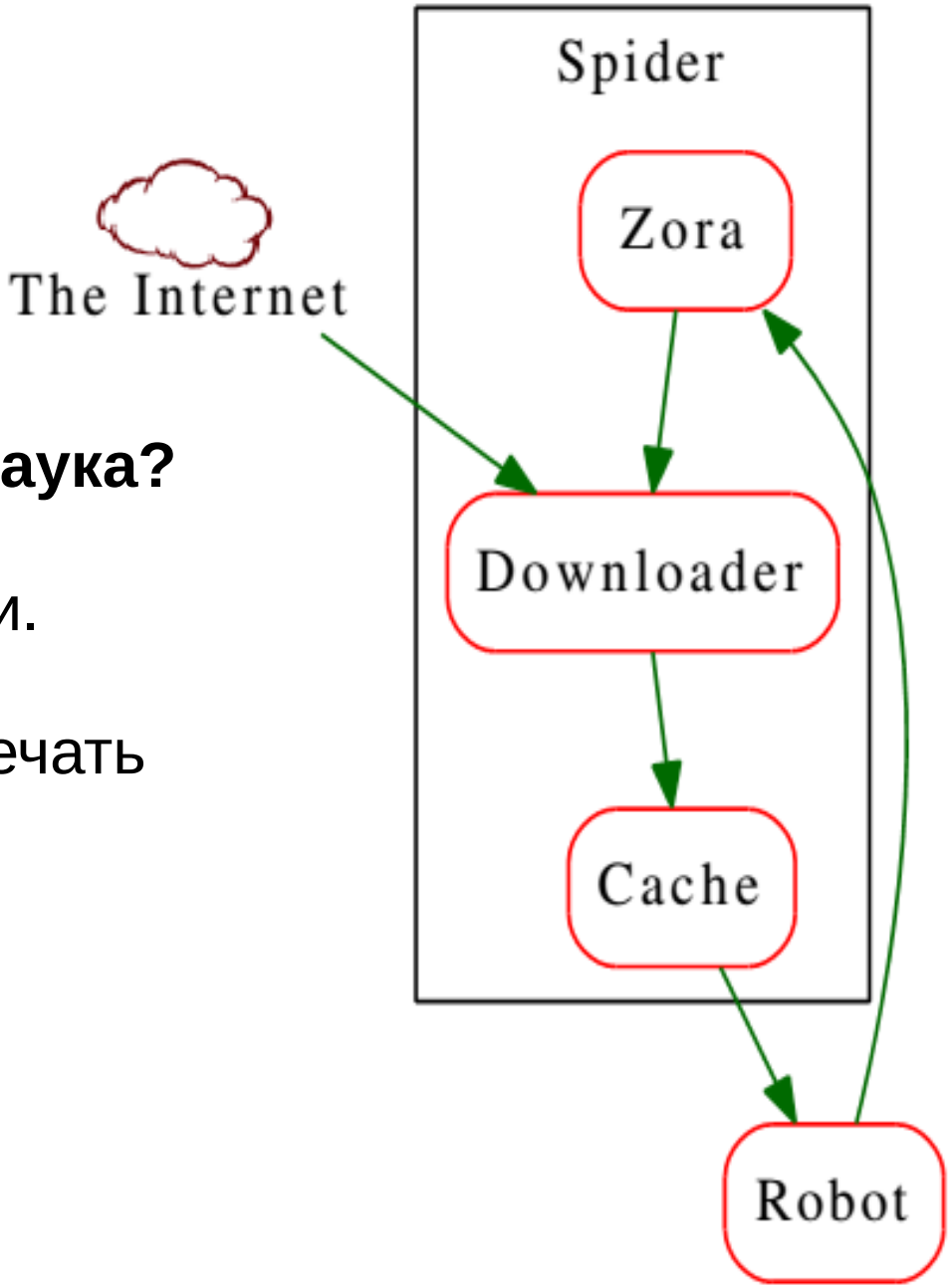
Как измерить качество работы паука?



Паук

Как измерить качество работы паука?

- какой процент сайтов мы увидели.
- насколько быстро мы умеем замечать изменения.



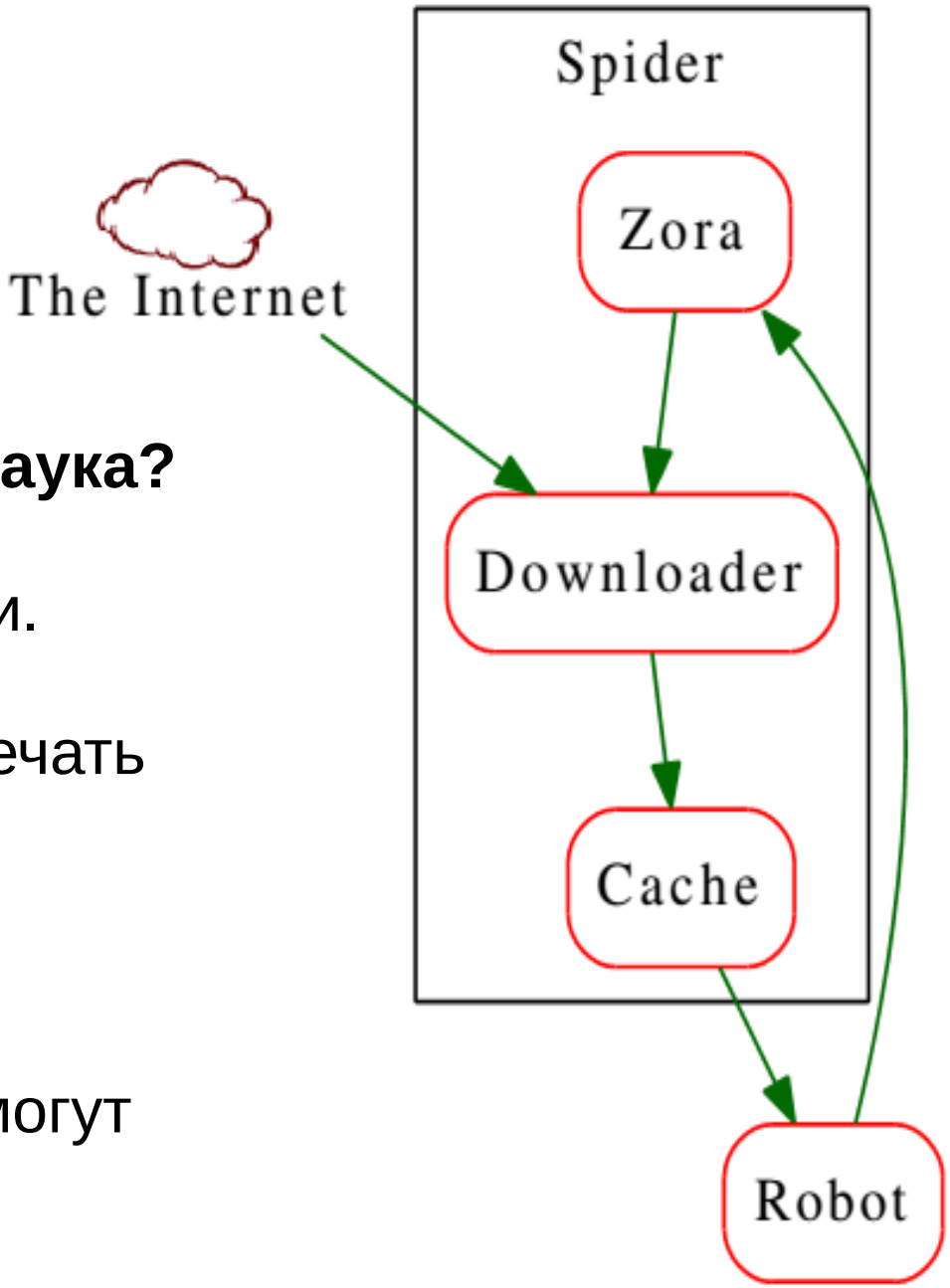
Паук

Как измерить качество работы паука?

какой процент сайтов мы увидели.

насколько быстро мы умеем замечать изменения.

+ страницы из разных регионов могут выглядеть по-разному.



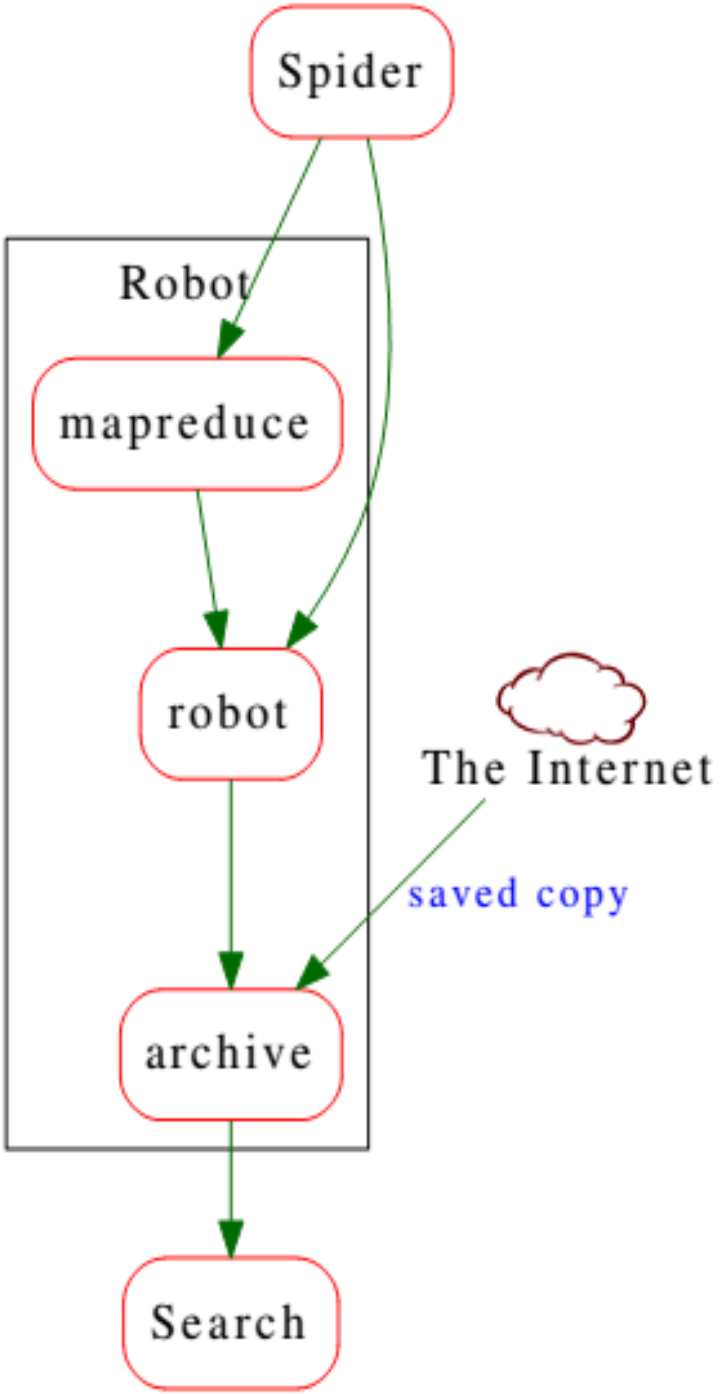
Робот

MapReduce
позволяет посчитать *признаки*
для каждого выкачанного
Пауком документа

Кластер серверов
сборка поискового индекса

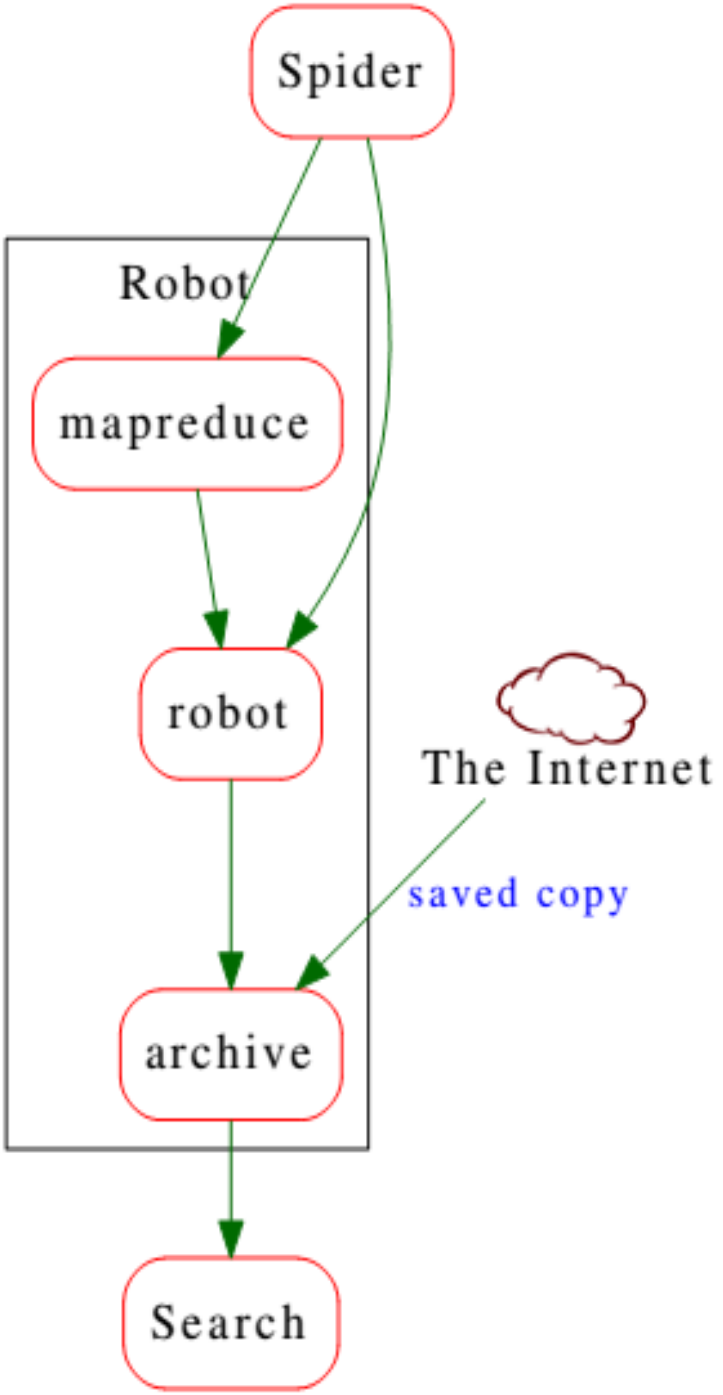
Архив
содержит несколько версий
поисковой базы

Робот представляет собой
более двух тысяч серверов.



Признаки

Какие типы признаков бывают?

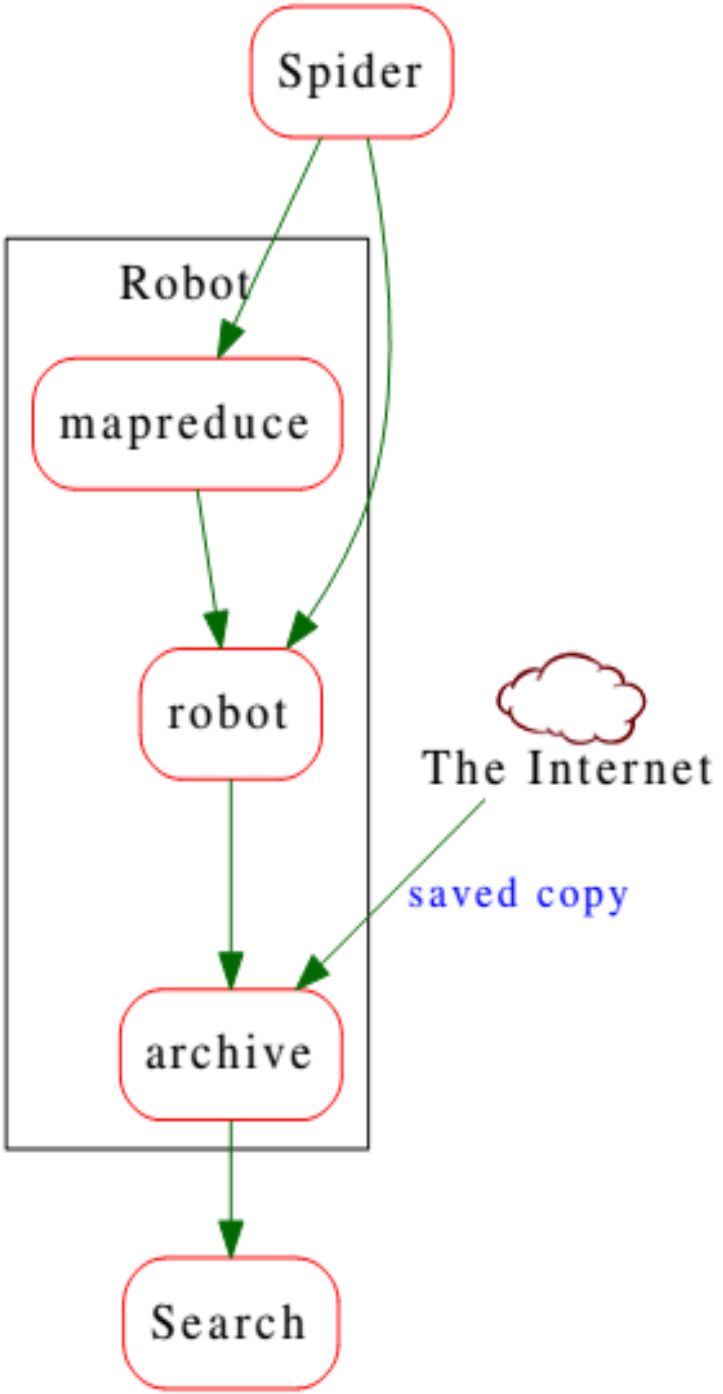


Признаки

Какие типы признаков бывают?

Быстрые
вычисляются для документа вместе с поисковым запросом

Медленные
считаются однократно и присваиваются исключительно документу



Устройство поискового индекса

Пусть есть три документа:

1. *Мама мыла раму*
2. *Рамы в Москве купить*
3. *Москва для мам*

Как построить поисковый индекс?

Устройство поискового индекса

Пусть есть три документа:

1. *Мама мыла раму*
2. *Рамы в Москве купить*
3. *Москва для мам*

Как построить поисковый индекс?

Соответствующий поисковый индекс:

Мама (1, 3)
Мыть (1)
Рама (1, 2)
Москва (2, 3)
Купить (2)
В (2)
Для (3)

Размер реального поискового индекса – 214 ТБ