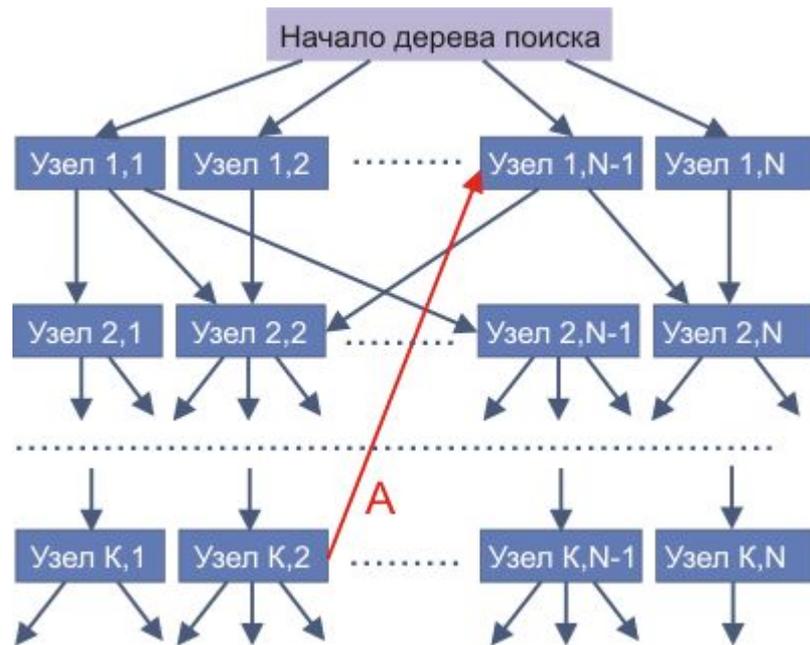


Петли в URL



Поиск

- Стоп-списки
- <meta> создана для предоставления информации о содержании документа для поисковых роботов, браузеров и других приложений. Структура метки:
<meta http-equiv=response content=description name=description URL=url>.
- Тезаурусы
- Локальный поиск (*grep = globally search a regular expression and print*)
- Кластерный поиск
- Структура индекса (базы данных):

Структура поисковой системы

- **web-агенты.** Осуществляют поиск серверов, извлекают оттуда документы и передают их системе обработки.
- **Система обработки.** Индексирует полученные документы, используя синтаксический разбор и стоп-листы (где, помимо прочего, содержатся все стандартные операторы и атрибуты HTML).
- **Система поиска.** Воспринимает запрос от системы обслуживания, осуществляет поиск в индексных файлах, формирует список найденных ссылок на документы.
- **Система обслуживания**

STOP-LIST

- Артикли
- Предлоги, междометия
- Модальные глаголы, вводные слова
- Проблема склонения и спряжения (языковые особенности)
- Проблемы комбинаций слов: Витамин А, “To be, or not to be?”
- Проблемы со словами-командами в Winword? PostScript, HTML и т.д.

robots.txt

- `user-agent: * # *` соответствует любому имени робота
- `disallow: /cgi-bin/ #` не допускает робот в каталог `cgi-bin`
- `disallow: /tmp/ #` не следует индексировать временные файлы
- `disallow: /private/ #` не следует заходить в частные каталоги

Релевантность

Пусть f - частота, с которой встречаются различные слова в тексте, а u - относительное значение полезности (важности).

$$f(u) = C \frac{1}{u}$$

$$W = \log \frac{N}{n} + 1$$

N - число документов и n - число документов, в которых встречается данный индексный термин

$$L = \sqrt{w_1^2 + w_2^2 + \dots + w_n^2}, \text{ где } w_i = \mathbf{tf} \times \mathbf{idf}$$

tf - (term frequency), частота, с которой встречается данный индексный термин;

IDF (Inverted Document Frequency) - величина, обратная частоте, с которой данный термин встречается во всей совокупности документов.

$$W = \frac{\mathbf{tf} \times \mathbf{idf}}{\sqrt{w_1^2 + w_2^2 + \dots + w_n^2}}$$

Кластерные методы поиска

- Кластерные методы, выбираемые для использования в экспериментальных поисковых системах должны удовлетворять некоторым определенным требованиям. Это:
- методы, создающие кластеры, не должны существенно их изменять при добавлении новых объектов. То есть, должны быть устойчивы по отношению к объему коллекции.
- методы должны быть устойчивы и в том отношении, что небольшие ошибки в описании объектов приводят к небольшим изменениям результата процесса кластеризации.
- результат, производимый методами кластеризации, должен быть независимым от начального порядка объектов.

Сетевая надежность

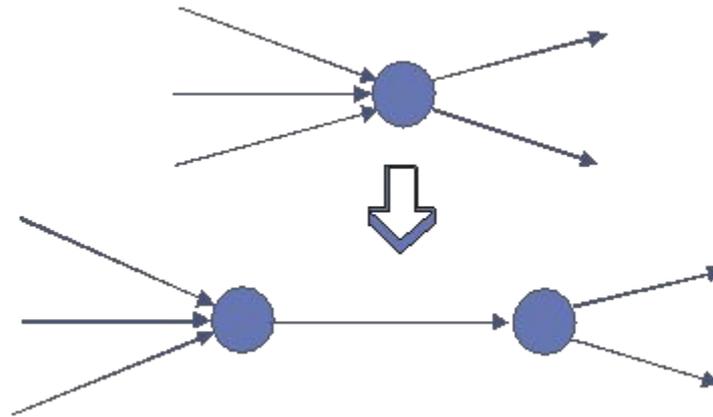
- Надежность всякой системы определяется надежностью составляющих ее элементов. А надежность элементов задается временем наработки на отказ или вероятностью отказа за оговоренный период времени
- Надежность – вероятность отказа за оговоренный отрезок времени (что такое отказ? – Выход из строя оборудования, канала, программы, перегрузка канала или сервера)
- Сложность оценки надежности сети определяется множественностью связей в сети. При N узлах она имеет порядок $N!$
- Прежде чем писать и запускать программу расчета надежности сети надо научиться оценивать - а хватит ли имеющихся вычислительных ресурсов для решения поставленной задачи в текущем тысячелетии

Формальное описание

- Сеть $G = (V, E)$, в которой V - набор узлов или вершин графа сети, а E - набор неориентированных ребер или набор ориентированных дуг. Большинство исследований по сетевой надежности посвящены k -терминальным мерам. Пусть имеется набор из K узлов и узел $s \in K$ ($k = |K|$).
- Задана сеть G , и все дуги графа, описывающего сеть, имеют вероятность надежности p . Тогда k -терминальная мера надежности определяется как (Pr - вероятность):
- $Rel(G, s, K, p) = Pr[\text{существует хотя бы один работающий путь от } s \text{ до каждого узла из набора } K]$

- Существует два важных частных случая мер: 2-терминальная мера с $|K|=2$ и всекритериальная мера, где $K=V$. Эти меры принято обозначать $Rel_2(G,s,p)$ и $Rel_A(G,s,p)$, соответственно (Rel - надежность).

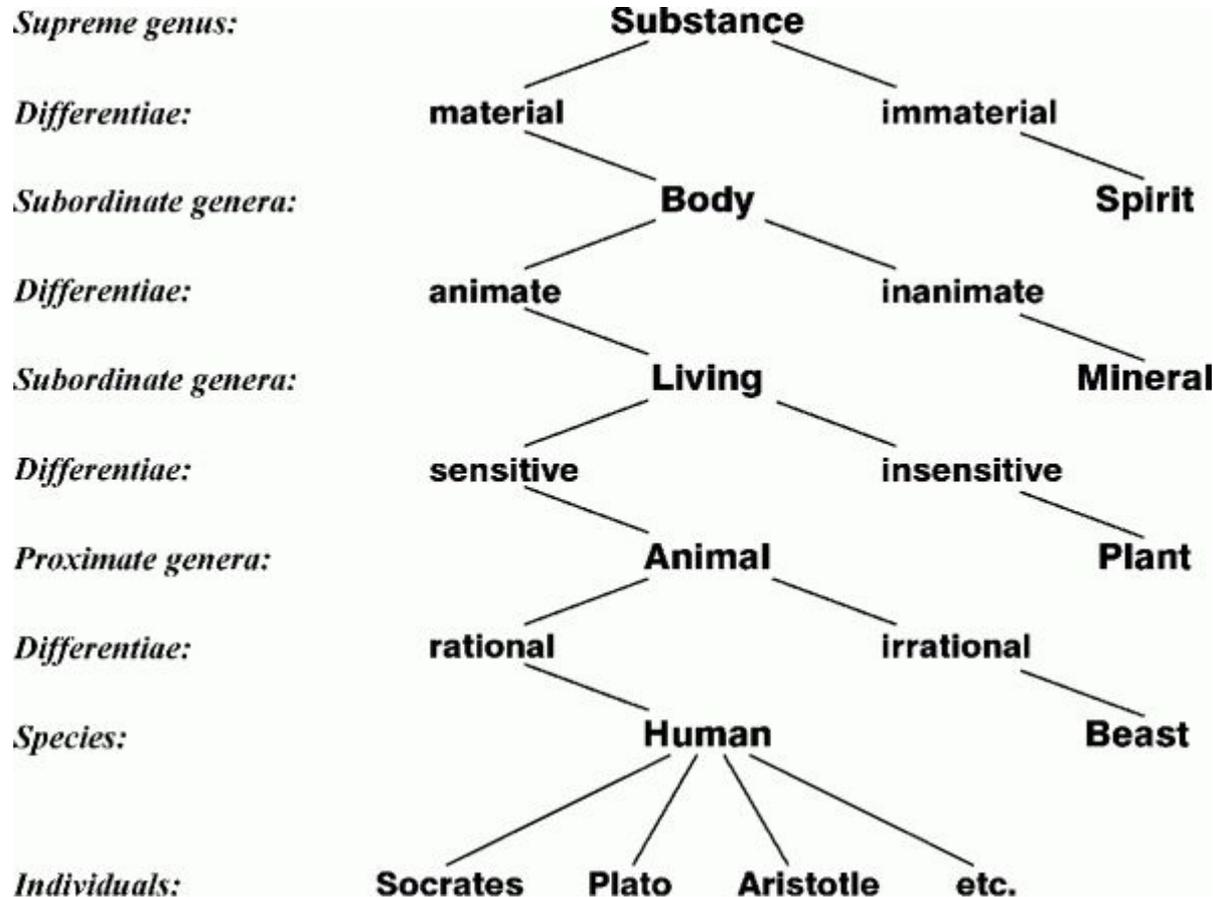
Эквивалентные преобразования



$$\text{Rel}(\text{SBS}, p) = \sum_{i=0}^m F_i p^{m-i} (1-p)^i$$

Общий член в полиноме надежности $F_i p^{m-i} (1-p)^i$ представляет собой вероятность того, что работает ровно $m-i$ компонентов сети и функционирует система в целом. **SBS** (stochastic binary system) - представляет собой систему, которая отказывает случайным образом в результате случайного выхода из строя ее компонента.

Семантические сети (Питер Испанский (1329))



10 аристотелевых категорий

- Содержание
- количество
- качество
- отношение
- где (место)
- когда (время)
- находящийся в положении (позиция)
- имеющий (состояние)
-

Семантические отношения

- Антонимия
- Ассоциации
- Причина-следствие
- Ононимы
- Гипонимные отношения
- Отношения положения
- Владение
- Меронимия

Семантика, синтактика,

прагматика

- **Знак** - это материальный объект, который служит для обозначения другого объекта и используется для передачи информации о последнем
- Все проявления и применения знаков и знаковых систем изучает **семиотика**. Предметом семиотики является связь знаков друг с другом с обозначаемыми ими объектами и явлениями, а также с субъектами их использующими для целей коммуникаций. Семиотика содержит в себе три раздела: *семантика, синтактика и прагматика*

Семантика, синтактика, прагматика

- **Семантика** изучает отношения между знаком и тем, что он обозначает или замещает
- **Синтактика** знаковых систем изучает их структуру и правила соединения знаков.
- **Прагматика** изучает законы функционирования знаковой системы, как средства коммуникации субъектов. С прагматикой связаны такие понятия как ценность и цель.

- Строя такие графы и присваивая ребрам графа определенные веса (метрики) в соответствии со степенью сродства понятий, можно подойти к решению проблемы контекста. При этом предполагается, что существует множество семантических деревьев, характеризующих общие языковые связи. Тогда при выявлении контекстного значения определенного слова или группы слов в пределах конкретного текстового фрагмента, строятся семантические деревья для этого фрагмента. В этих деревьях веса берутся из

Telnet (RFC-854; порт 23)

- TELNET предлагает три услуги:
- Определяет сетевой виртуальный терминал (**NVT** - network virtual terminal), который обеспечивает стандартный интерфейс к удаленной системе.
- Включает механизм, который позволяет клиенту и серверу согласовать опции обмена
- Обеспечивает симметрию соединения, допуская любой программе (например FTP) выступать в качестве клиента

- Уязвимость протокола Telnet для хакеров привела к тому, что в последнее время эта утилита часто заменяется **SSH** (Secure Shell) или другими программами, обеспечивающими безопасный удаленный доступ. Но следует иметь в виду, что SSH сохраняет алгоритмическую логику telnet, хотя и использует порт 22 вместо 23. К числу недостатков Telnet следует отнести невозможность эффективного использования мышки.

Формат блока данных Telnet

Первый октет	Второй октет	3-ий октет (опционно)
Интерпретируется как команда IAC	Код команды	Опция

IAC WILL ECHO (255 251 1) (Interpret As Command)

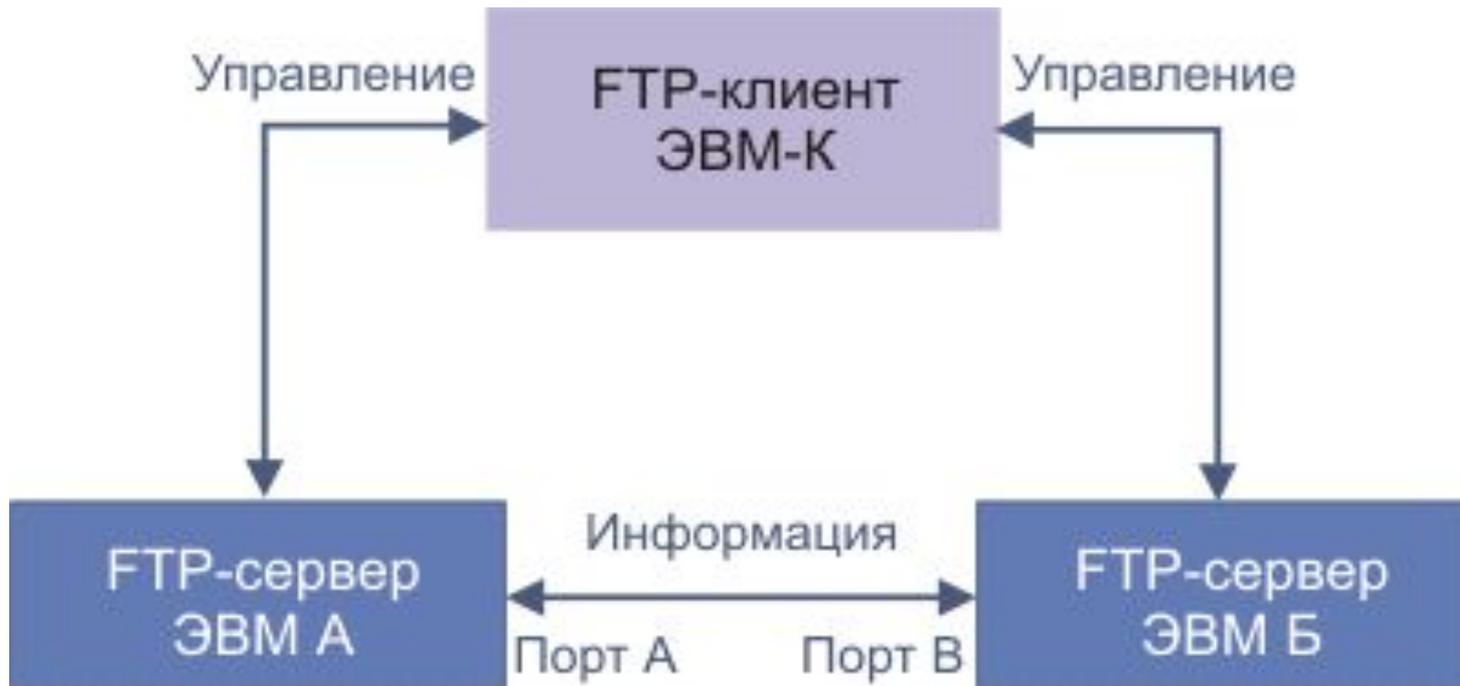
IAC код=0xff

FTP (RFC-959)

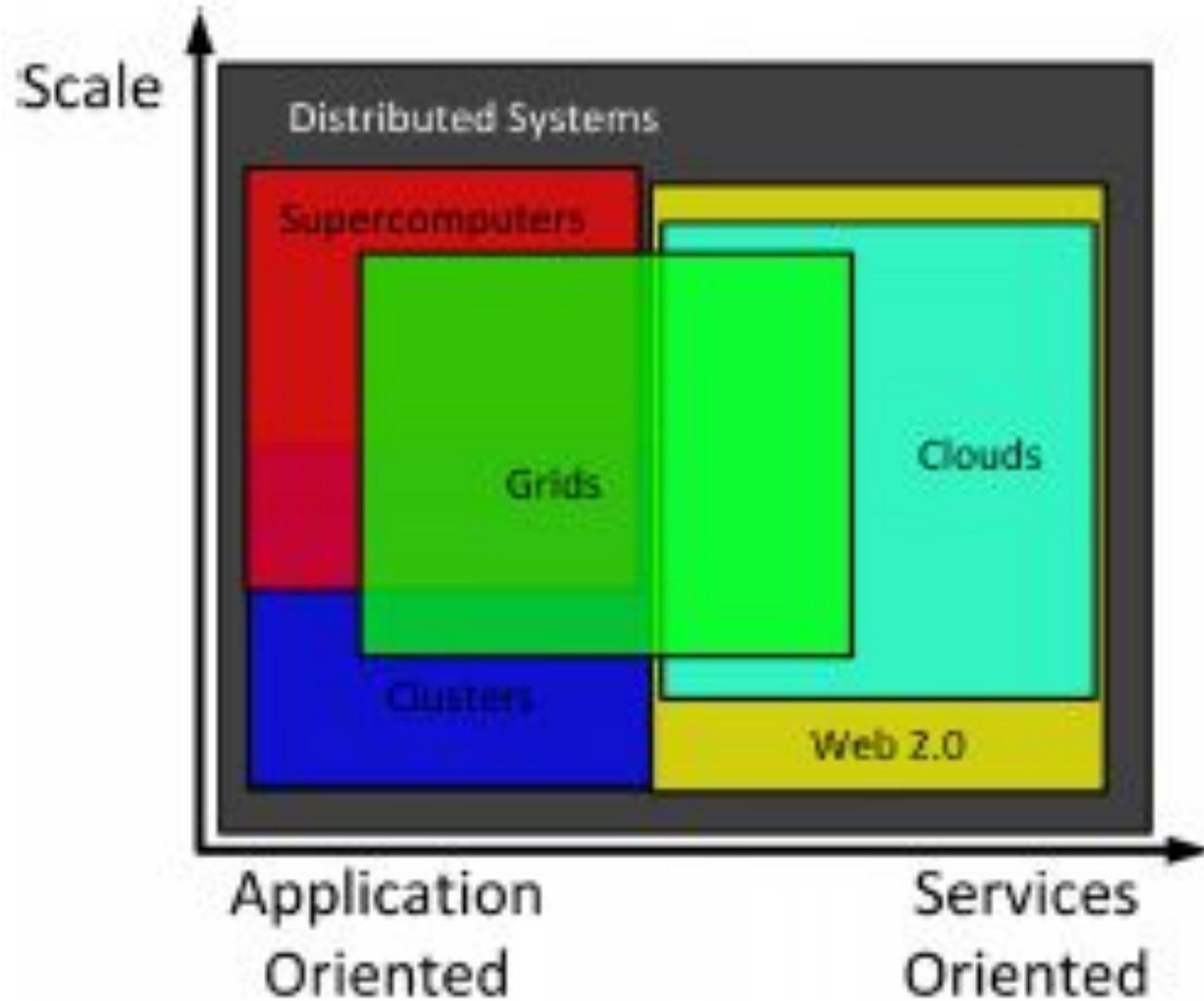


Анонимное FTP

FTP



CLOUD

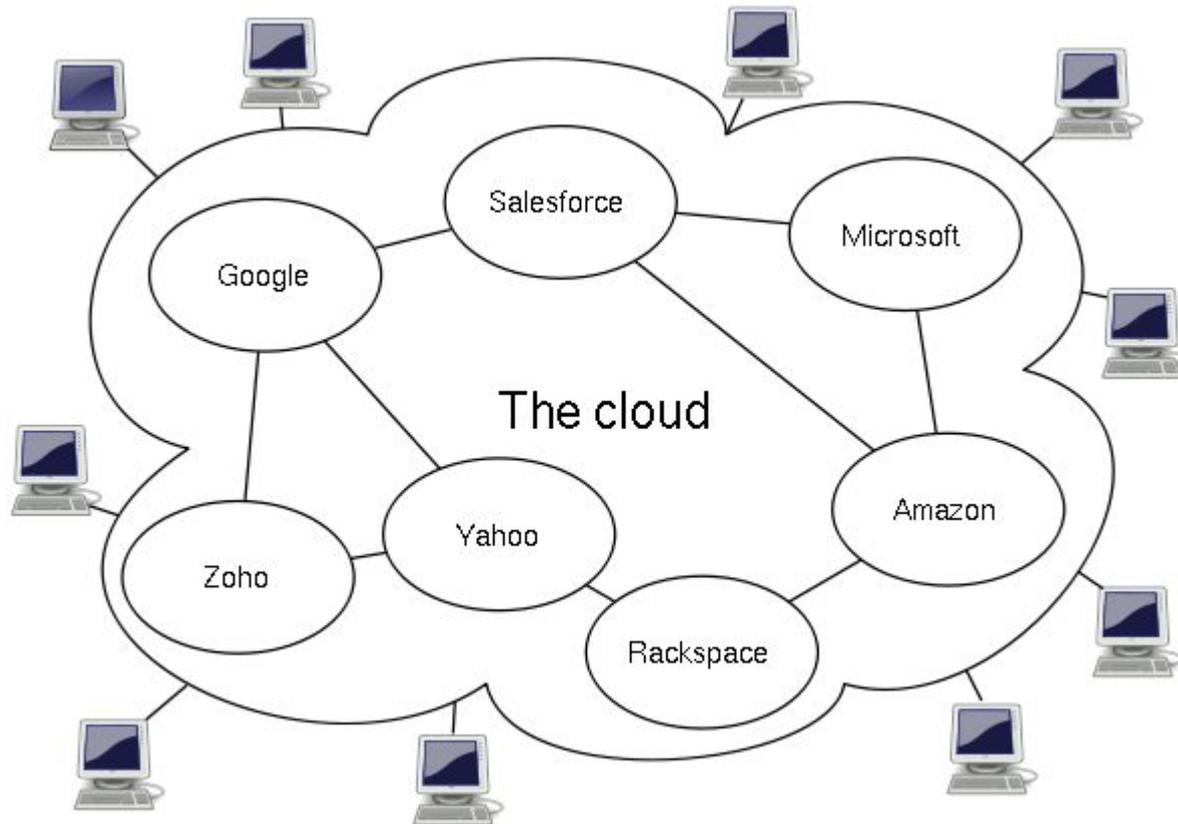


- **SaaS Software as a service**
- **storage-as-a-service**
- **database-as-a-service**
- **information-as-a-service**
- **process-as-a-service**
- **application-as-a-service = software-as-a-service**
- **platform-as-a-service**
- **integration-as-a-service**
- **IT-as-a-service**
- **Hardware Infrastructure-as-a-service**
- **Software Infrastructure-as-a-service**
- **security-as-a-service**
- **management/governance-as-a-service)**
- **testing-as-a-service**
- **UC-as-a-service**

Категории сервис-провайдеров

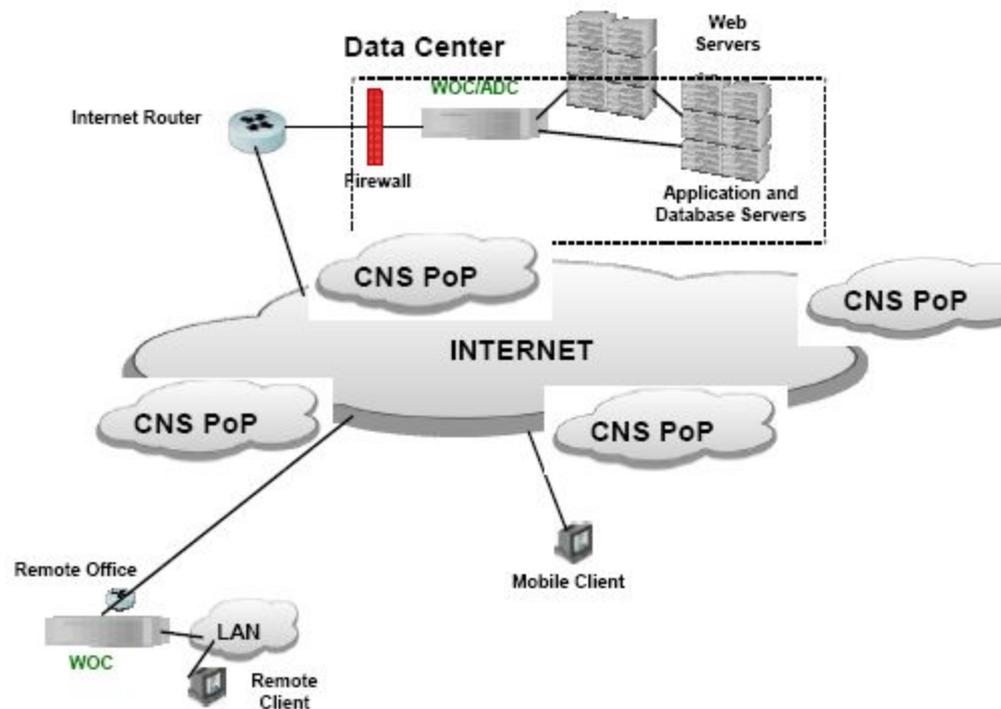


Логическая схема Cloud computing



Интернет-структура для CNS - Cloud Network Service.

(POP - Point of Presence; WOC - WAN Optimization Controller)



WHOIS (RFC-0954, -3912; порта 43)

- Сейчас существуют региональные серверы для Европы
- [RIPE](#) - - [фр.](#) - фр. [Reseaux IP Europeens](#) + - фр. Reseaux IP Europeens + [англ.](#) - фр. Reseaux IP Europeens + англ. [Network Coordination Centre](#),
- [RIPN](#) - - [Российский НИИ Развития общественных сетей](#),
- [ARIN](#) - Северная Америка,
- [LacNIC](#) - Южная Америка,
- [AfrinIC](#) - Африка и
- [APNIC](#) - Азия ([Asia Pacific NIC](#)).
- В системах UNIX имеется аналог этой службы - **rwho**

WHOIS

- Следует учитывать, что длина отклика в случае запроса WHOIS в разы больше длительности запроса, и это может использоваться для атак типа отказа обслуживания (DoS). Такая особенность является причиной того, что некоторые whois-серверы не откликаются, если запросы поступают слишком часто. По этой причине можно рекомендовать создавать свою базу данных для часто запрашиваемых имен и адресов.