Технологии копирования данных «большого» объёма

Докладчик: А.Е. Шевель

Краткий обзор

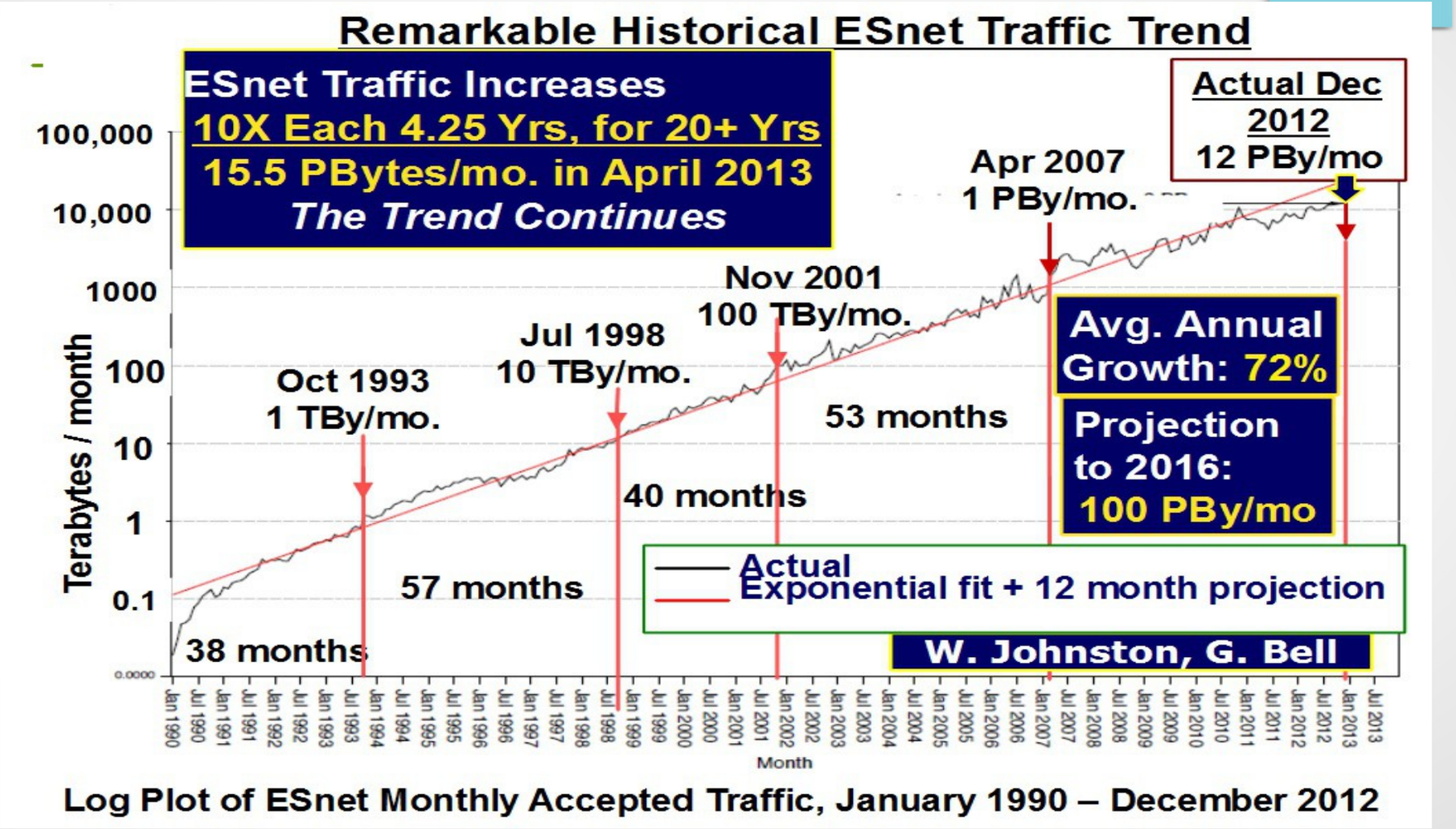
- Источники данных большого объёма.
- Описание и архитектура данных большого объёма.
- Технологии используемые при передаче данных большого объёма.
- Выполняемое исследование и области применения результатов.

Научные источники данных «большого» объёма

- **Научные экспериментальные установки**
  - http://www.lsst.org - Large Synoptic Survey Telescope
    - 15 TB per night (may be 10 PB/year)
  - https://www.skatelescope.org/ - Square Kilometre Array
    - 300-1500 PB/year
  - http://www.cern.ch — CERN
    - ~20PB/year (FAIR ~ same)
  - http://www.iter.org - International Thermonuclear Experimental Reactor
    - ~1 PB/year
  - http://www.cta-observatory.org/ - CTA - The Cherenkov Telescope Array
    - ~20 PB/year

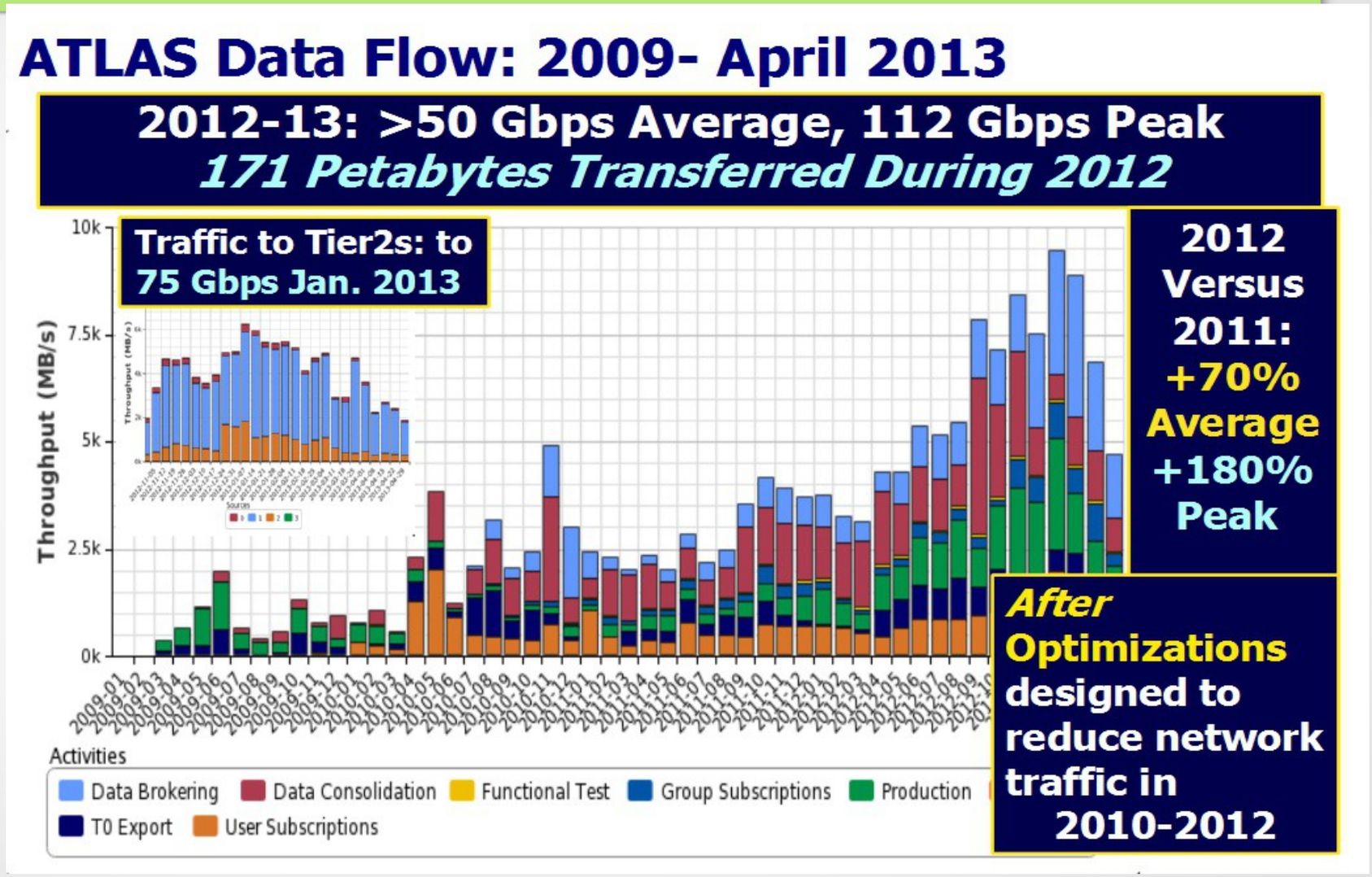
Рост сетевого трафика

Harvey B Newman



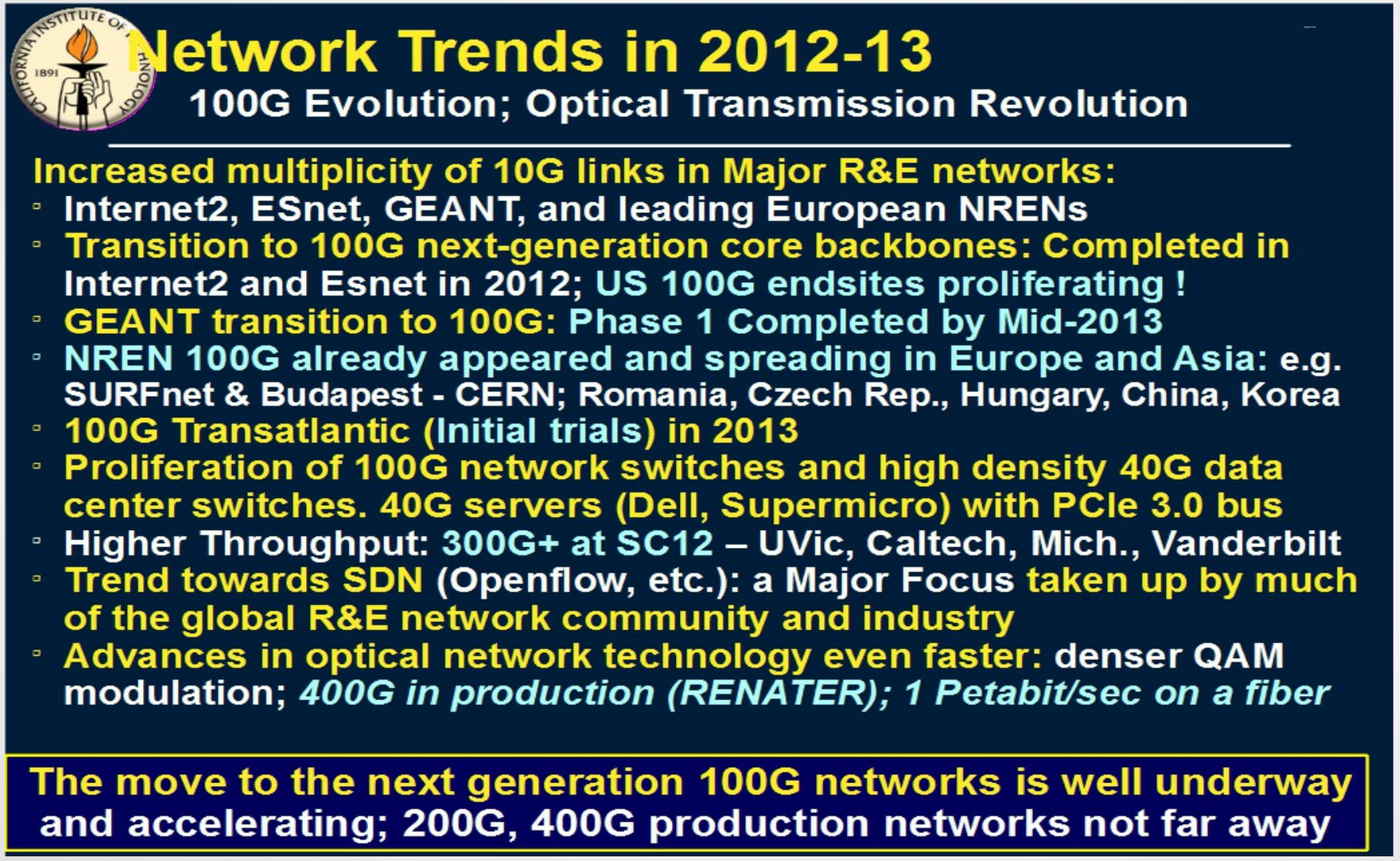
Сетевой трафик ATLAS

Harvey B Newman



Развитие сетевых технологий

Harvey B Newman



Как изменятся запросы на данные в будущем

Paul Sheldon & Alan Tackett
Vanderbilt University

10 Year Projections

- Requirements in 10 years? Look 10 years back...

Source: Ian Fisk and Jim Shank

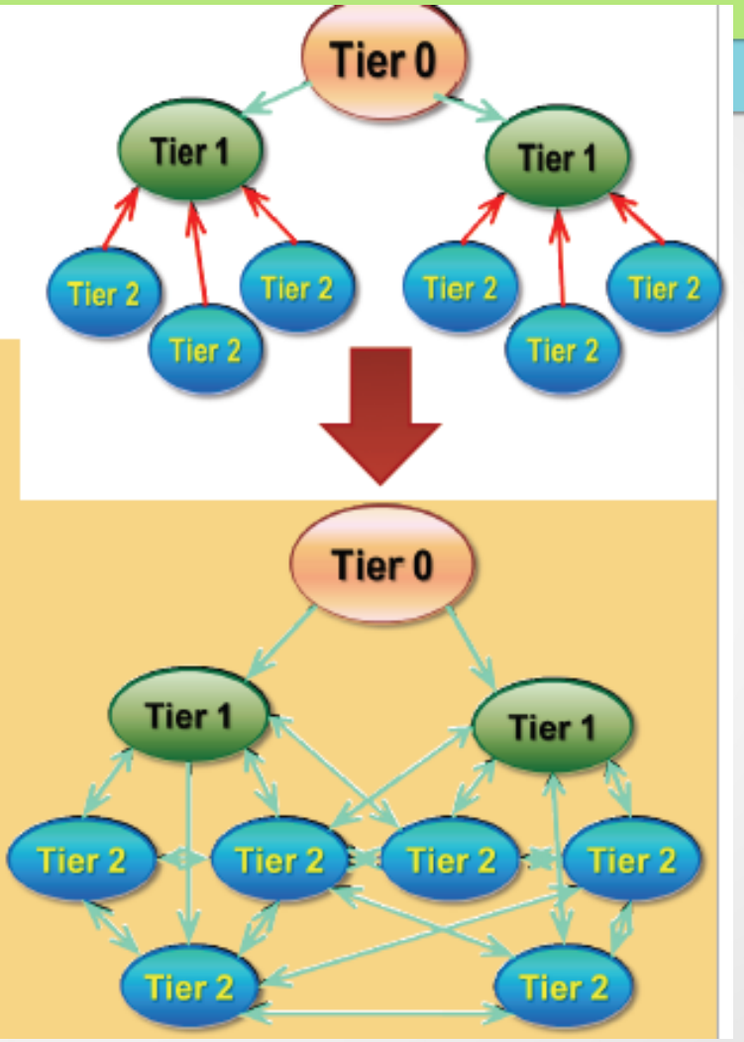
Metric	Tevatron(2003)	LHC(2012)
Remote Computing Capacity	15kHS06 (DZero Estimated)	450kHS06 (CMS)
User Jobs launched per day	10k per day	200-300k jobs per day
Disk Capacity per experiment in PB	0.5PB	60PB
Data on Tape per experiment	400TB	70PB
MC Processing Capacity per month for Full Simulation	3M	300M
Data Served from dCache at FNAL per day	25TB per day	10PB per day
Wide Area networking from host lab	200Mb/s	20000Mb/s
Inter VO transfer volume per day	6TB (DZero SAM)	546TB (ATLAS)

Emerging Data Logistics Needs of the LHC Expts 17 September 5, 2013

Paul Sheldon & Alan Tackett
Vanderbilt University

- Model must evolve
- Was: Hierarchical, strategic pre-placement of data sets…

- **Meshed data flows:** Any site can use any other site to get data
- **Dynamic data caching:** Analysis sites will pull datasets from other sites "on demand"
- **Remote data access:** Jobs executing locally access data at remote sites in quasi-real time



Сетевая инфраструктура LHC

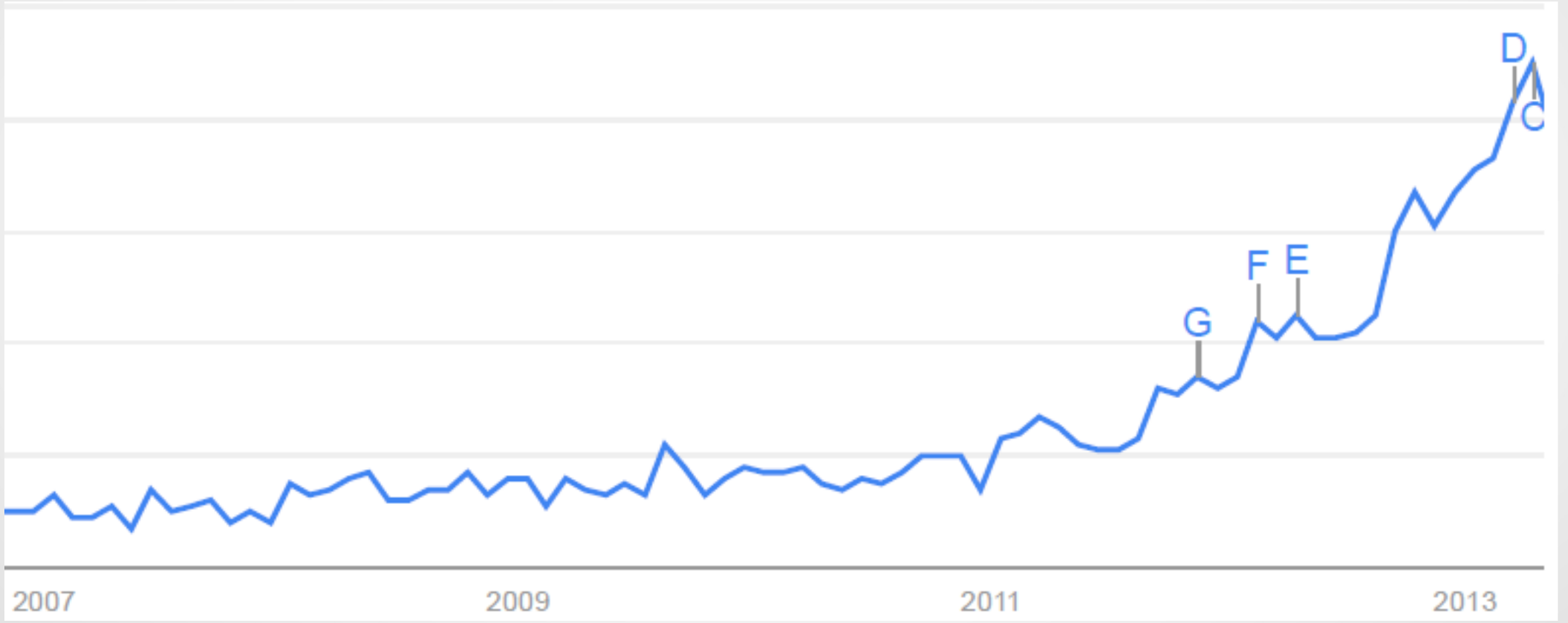
- LHC OPN — Optical Private Network
- LHC ONE — Open Networking Environment

  цель LHCONE является обеспечение множества точек входа в частную компьютерную сеть объединяющую кластеры LHC T1/2/3 .
- ANSE — Advanced Network Services for [LHC] Experiments

Другие генераторы «больших» данных и предпринятые шаги по изучению этого феномена

- NIST - http://bigdatawg.nist.gov/usecases.php - много примеров
  - Любая крупная компания (в особенности телекоммуникационная)
  - Копии видео потоков с тысяч камер наблюдения и т. д.
  - Последнее время много говорят про сохранение данных на длительный период (Data Preservation).
- http://bigdatawg.nist.gov/WhiteHouse_big_data_press_release.pdf - 29 марта 2012 Обама подписал выделение $200M на изучение феномена «большие данные» (Big Data).

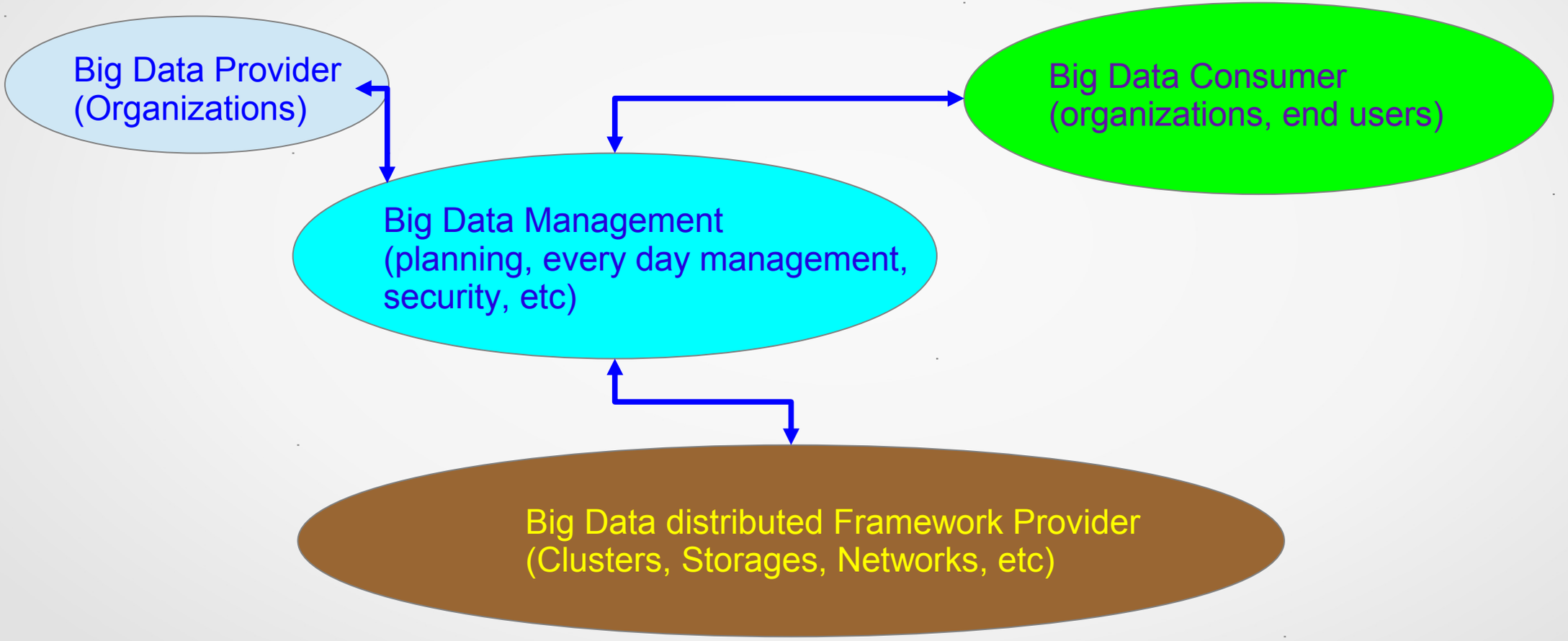
Google trend for «the big data»



Большие данные — Big Data

- «Большие» данные характеризуются комплексом признаков:
  - Скорость поступления (изменения) данных (Velocity)
  - Объём данных (Volume)
  - Разнообразие данных (сложность структуры) (Variety)

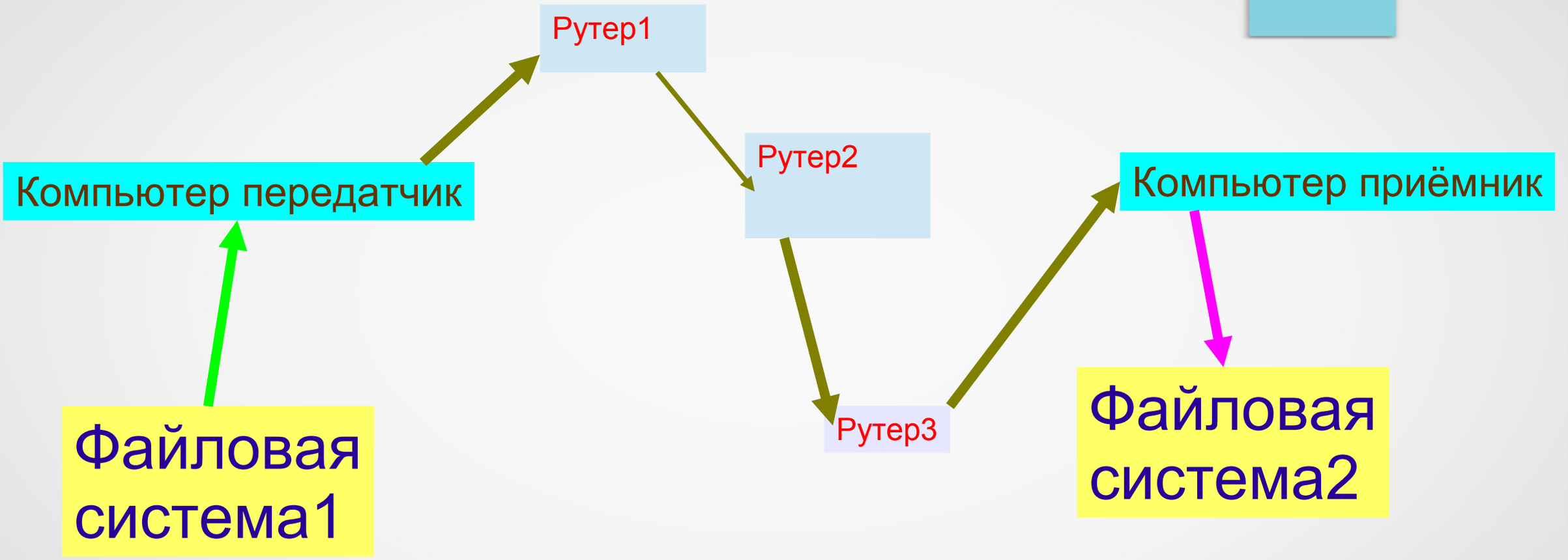
Архитектура “больших” данных



Особенности передачи больших данных

- Обычно время передачи больших данных составляет часы или дни.
- За такое время может сильно измениться обстановка в канале связи (параметры: RTT, процент потери сетевых пакетов, пропускная способность канала).
- Наконец, может возникнуть перерыв в канале связи (секунды, минуты, часы, дни).
- Полезно иметь возможность использовать два и более независимых канала связи (если каждый из двух каналов связи имеет вероятность выхода из строя во время передачи скажем 10%, вероятность одновременного выхода из строя двух каналов составит 1%).

Последовательность передачи данных



Тема: копирование/реплицирование «больших» данных

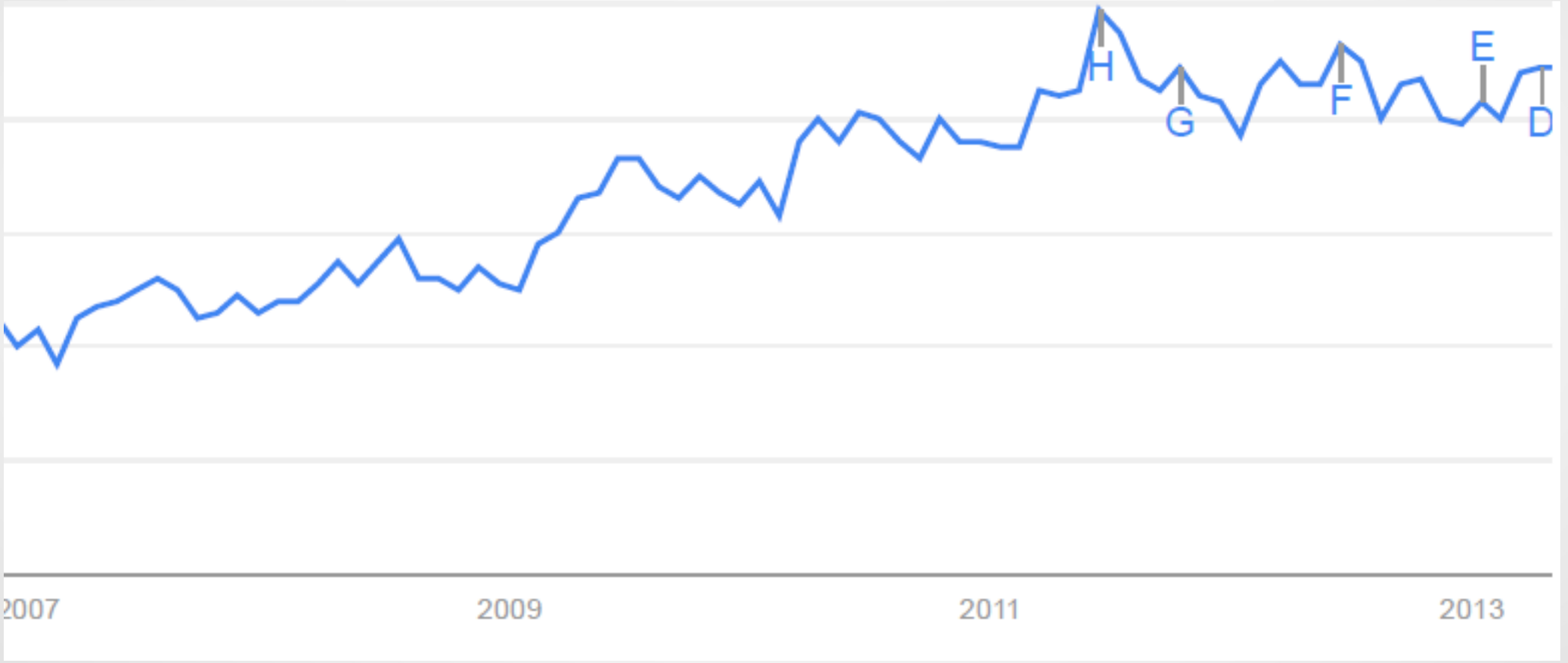
В лаборатория сетевых технологий http://sdn.ifmo.ru/ в НИУ ИТМО http://www.ifmo.ru/ сформировано направление исследований (среди других) «копирование/реплицирование данных большого объёма через Интернет».

- Создаётся распределённый Испытательный Стенд (100 ТБ дисковой памяти + сервер 96 GB оперативной памяти под OS RedHat/ScientificLinux с каждой стороны).
    - Сравнительное изучение доступных средств (включая CMS PhEDEx и ATLAS DQ2, Rucio) копирования/реплицирования данных большого объёма (тестирование и измерения).
    - Далее разработка прототипа системы копирования/реплицирования данных большого объёма, которая может быть использована в конкретных экспериментах.

Предполагаемые исследования

- В Российском сетевом сегменте немного доступных скоростных каналов (100 Gbit и выше). Так что в исследованиях планируется использовать то, что реально доступно (в основном 1 Gbit).
- При тестировании разных способов передачи больших данных предполагается использовать более, чем один канал.
- Предлагается использовать относительно новую технологию — Программно Конфигурируемые Сети (ПКС - SDN), поскольку с этой технологией уже проводили несколько серий тестовых измерений более года (аппаратные и программные реализации).
- Кооперация с GENI (http://www.geni.net/)

Google trend for SDN



Технологические особенности передачи данных

- Основные протоколы низкого уровня — стек TCP/IP.
  - Число настраиваемых параметров в Линукс немало (чуть больше 880).
  - Многое зависит от свойств конкретного канала связи.
  - Некоторые параметры TCP очень важны, например, размер блока, размер сегмента, размер окна.
  - Для увеличения скорости передачи данных (даже по одному каналу) одним из основных мероприятий является организация многопоточной передачи данных.

Тестирование на первом этапе (программные компоненты)

- GridFTP - http://www.globus.org/toolkit/data/gridftp/
- BBFTP - http://doc.in2p3.fr/bbftp/
- FDT - http://monalisa.cern.ch/FDT/
- Другие компоненты, в том числе системы наблюдения за состоянием канала, например, perfSONAR и другие.

Сравнительное изучение имеющихся систем (критерии сравнения)

- Доступность.
- Интерфейс с оператором.
- Производительность.
- Надёжность.
- Хранение истории передач.
- Выдача оценки времени выполнения передачи данных большого объёма.
- Объём потребляемых ресурсов (CPU, память, проч).
- Метод проверки правильности передачи.
- Другие

Основные цели работы

- Комбинируя уже разработанные современные компоненты и методы с нашими идеями/разработками/опытом достичь максимально возможной пропускной способности при передаче больших данных на существующих каналах связи.
- Применить … (у кого-то есть предложения?)
- Привлечь студентов ...(у кого-то имеются идеи?)

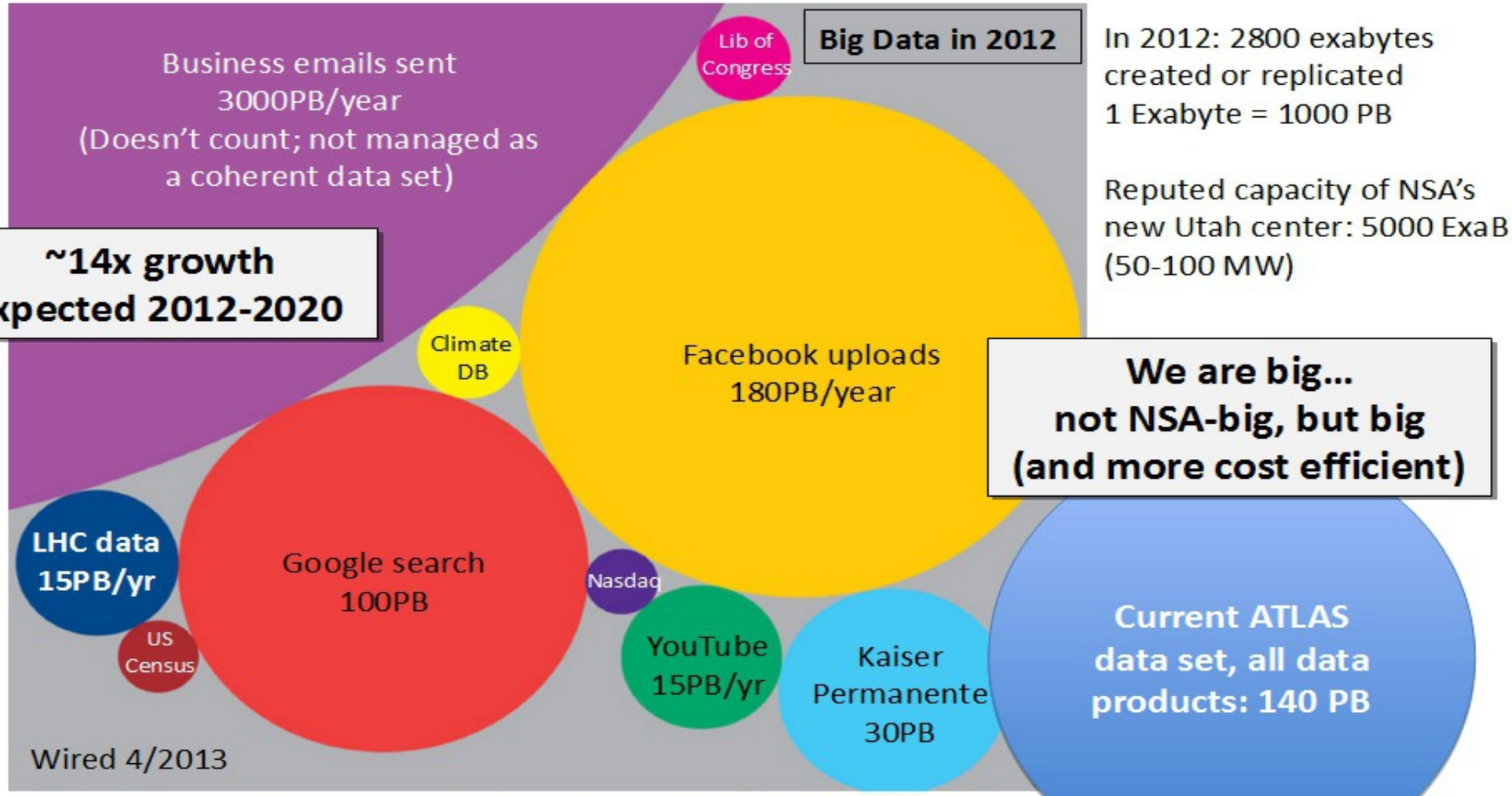
Вопросы ?

Дополнительная информация по Большим Данным

- В физике высоких энергий
  - http://icfa-scic.web.cern.ch/ICFA-SCIC/meetings.html
  - https://indico.cern.ch/contributionDisplay.py?sessionId=0&contribId=487&confId=214784
  - https://indico.cern.ch/contributionDisplay.py?sessionId=0&contribId=484&confId=214784
- В других областях:
  - https://www.dropbox.com/s/3j7x7nnl0vpkcfb/Sokolov_27_06.pdf (Рус)

Data Management

Where is LHC in Big Data Terms?



In 2012: 2800 exabytes created or replicated
1 Exabyte = 1000 PB

Reputed capacity of NSA's new Utah center: 5000 ExaB (50-100 MW)

http://www.wired.com/magazine/2013/04/bigdata/

Decimal

Value		Metric
1000	KB	kilobyte
1000^2	MB	megabyte
1000^3	GB	gigabyte
1000^4	TB	terabyte
1000^5	PB	petabyte
1000^6	EB	exabyte
1000^7	ZB	**zettabyte**
1000^8	YB	yottabyte