Global-Scale Applications Rely on Datacenters, Datacenters Rely on Scalable Computer Systems

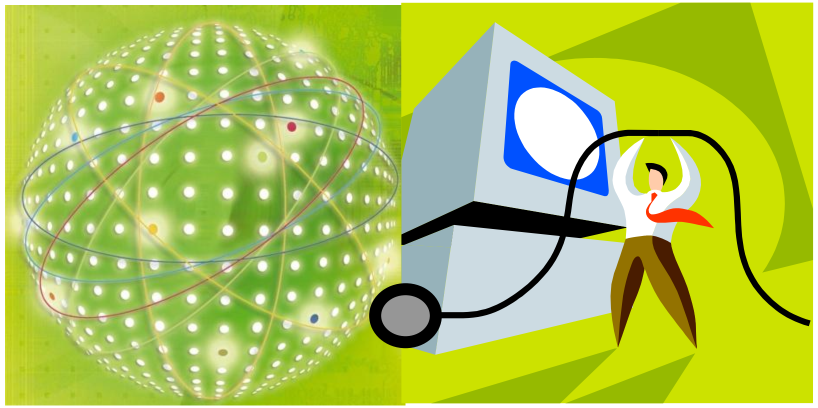
Alexandru Iosup
Parallel and Distributed Systems Group

@AIosup

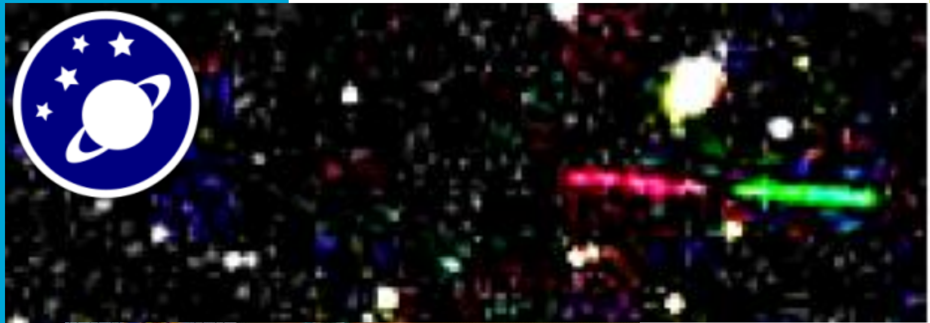
TUDelft

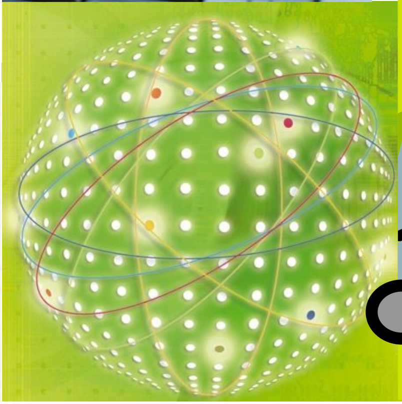
PDS Group

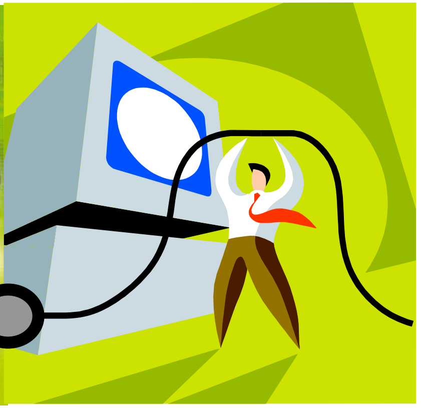
This Is the Golden Age of Scalable Computing

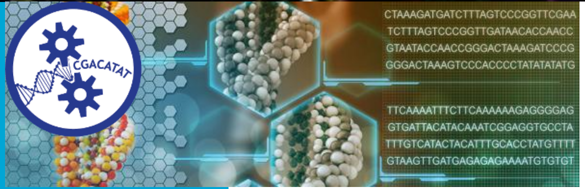


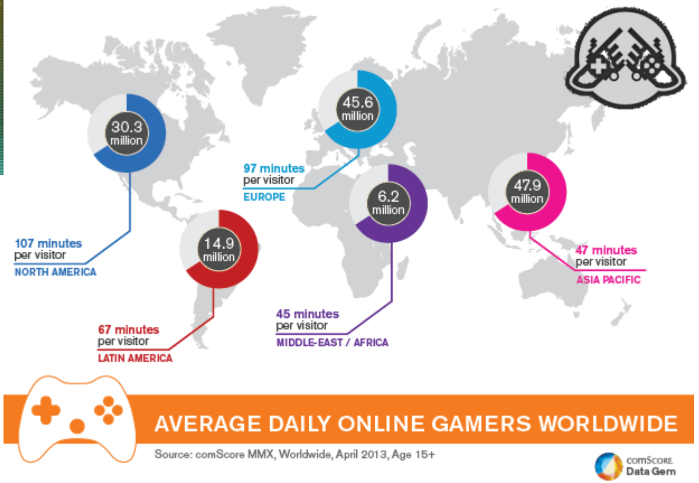
This Is the Golden Age of Scalable Computing







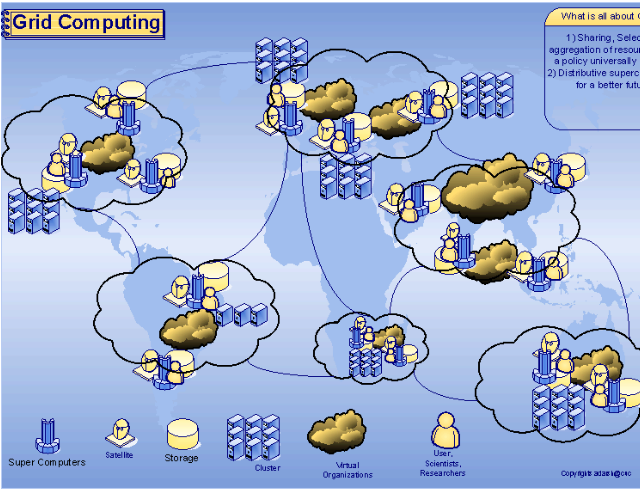

CTAAAGATGATCTTTAGTCCCGGTTCGAA
TCTTTAGTCCCGGTTGATAACACCAACC
GTAATACCAACCGGGACTAAAGATCCCG
GGGACTAAAGTCCCACCCCTATATATATG
TTCAAAATTTCTTCAAAAAAGAGGGGGAG
GTGATTACATACAAATCGGAGGTGCCTA
TTTGTCATACTACATTTGCACCTATGTTTT
GTAAGTTGATGAGAGAGAAAATGTGTGT
CGACATAT

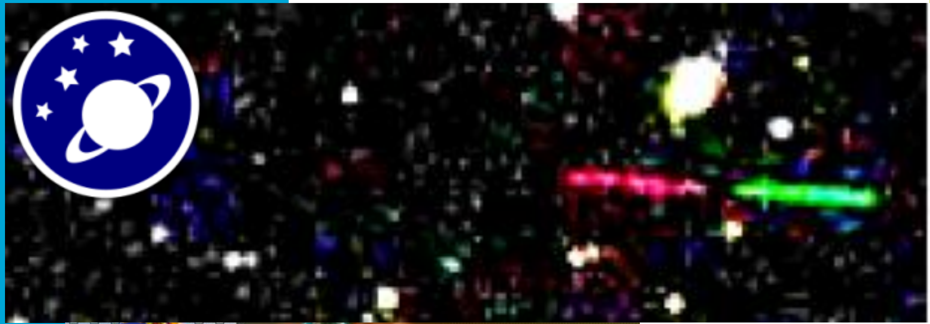


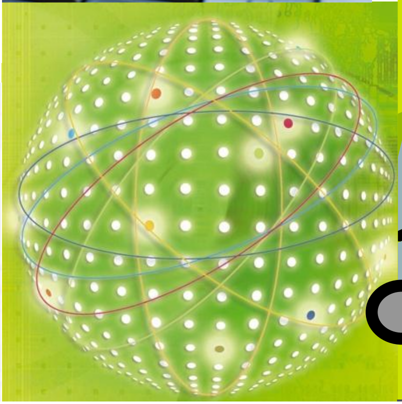
AVERAGE DAILY ONLINE GAMERS WORLDWIDE

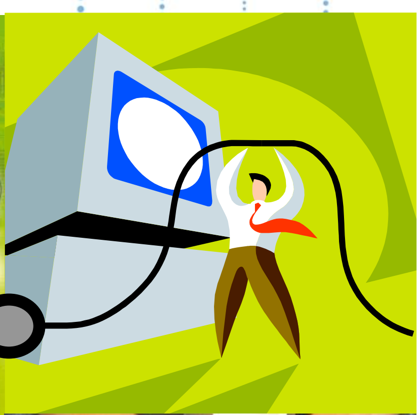
Source: comScore MMX, Worldwide, April 2013, Age 15+

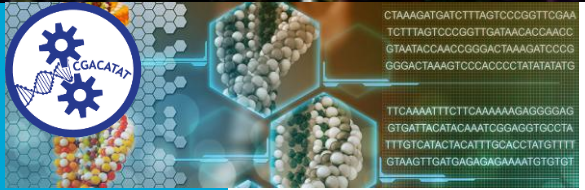
This Is the Golden Age of Scalable Computing

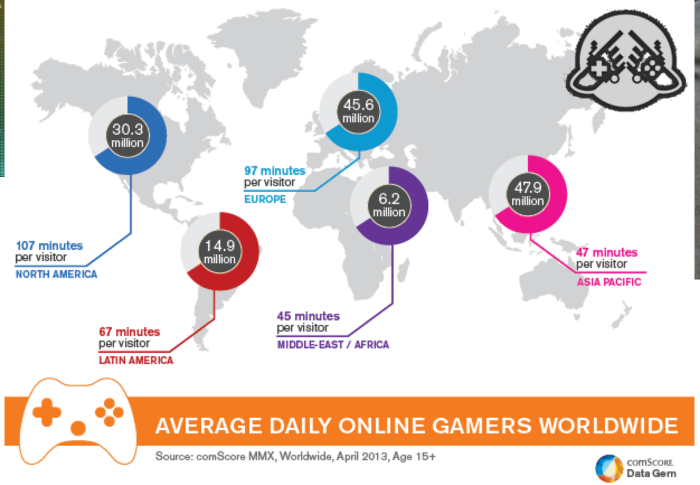


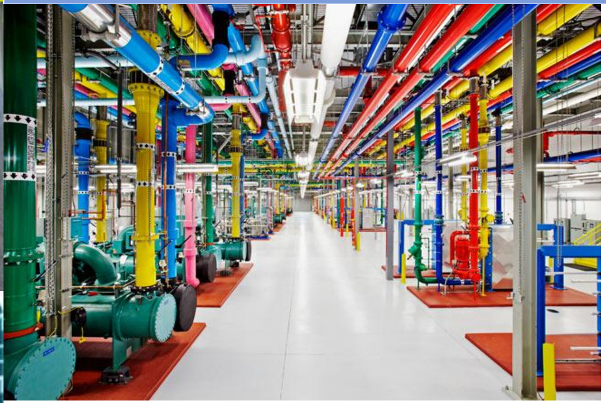







CGACATAT
CTAAAGATGATCTTTAGTCCCGGTTCGAA
TCTTTAGTCCCGGTTGATAACACCAACC
GTAATACCAACCGGGACTAAAGATCCCG
GGGACTAAAGTCCCACCCCTATATATATG
TTCAAAATTTCTTCAAAAAAGAGGGGAG
GTGATTACATACAAATCGGAGGTGCCTA
TTTGTCATACTACATTTGCACCTATGTTTT
GTAAGTTGATGAGAGAGAAAATGTGTGT





Agenda

1. The Golden Age of scalable computing
2. The core idea of cloud computing

3. Enabling technologies (homework)

4. The scheduling challenge
5. The Ecosystem Navigation challenge
6. The Big Cake challenge
7. Jevon's Effect challenge (IEEE Scale Challenge Award)

8. Take-home message

TUDelft

PDS Group

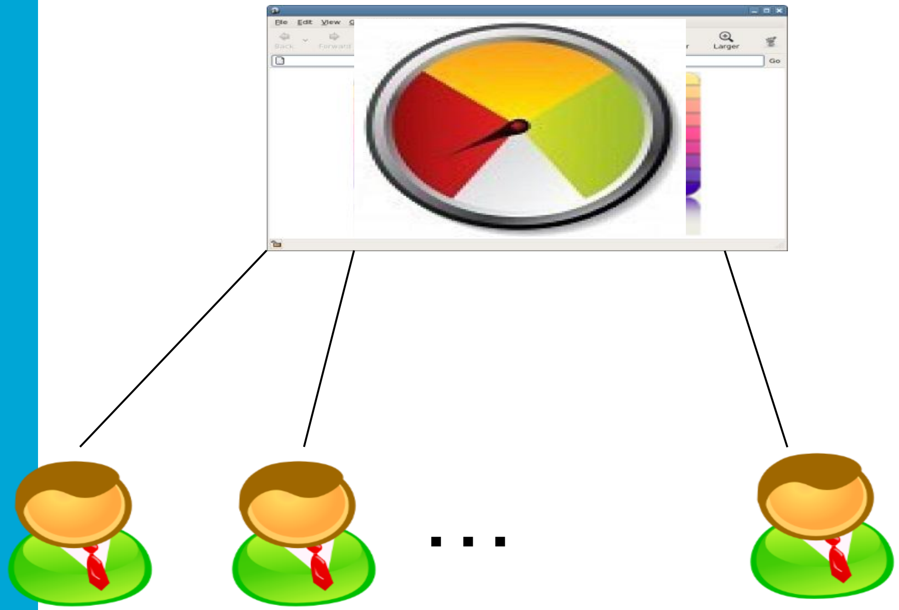
Joe Has an Idea ($$$)

(Source: A. Antoniou, MSc Defense, TU Delft, 2012. Original idea: A. Iosup, 2011.)

Solution #1

Buy then Maintain

- Big up-front commitment
- Load variability: NOT supported

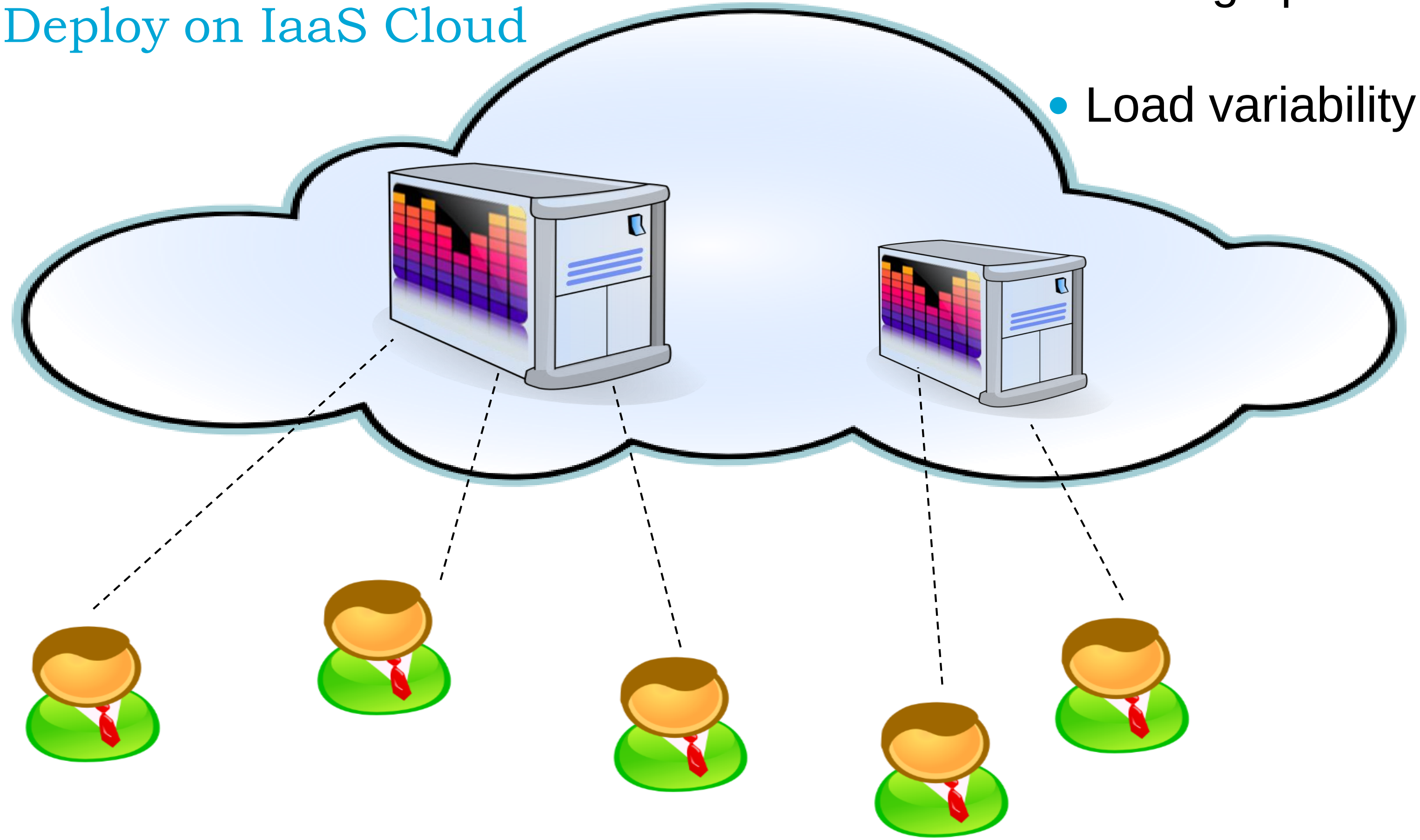


(Source: A. Antoniou, MSc Defense, TU Delft, 2012. Original idea: A. Iosup, 2011.)

Solution #2

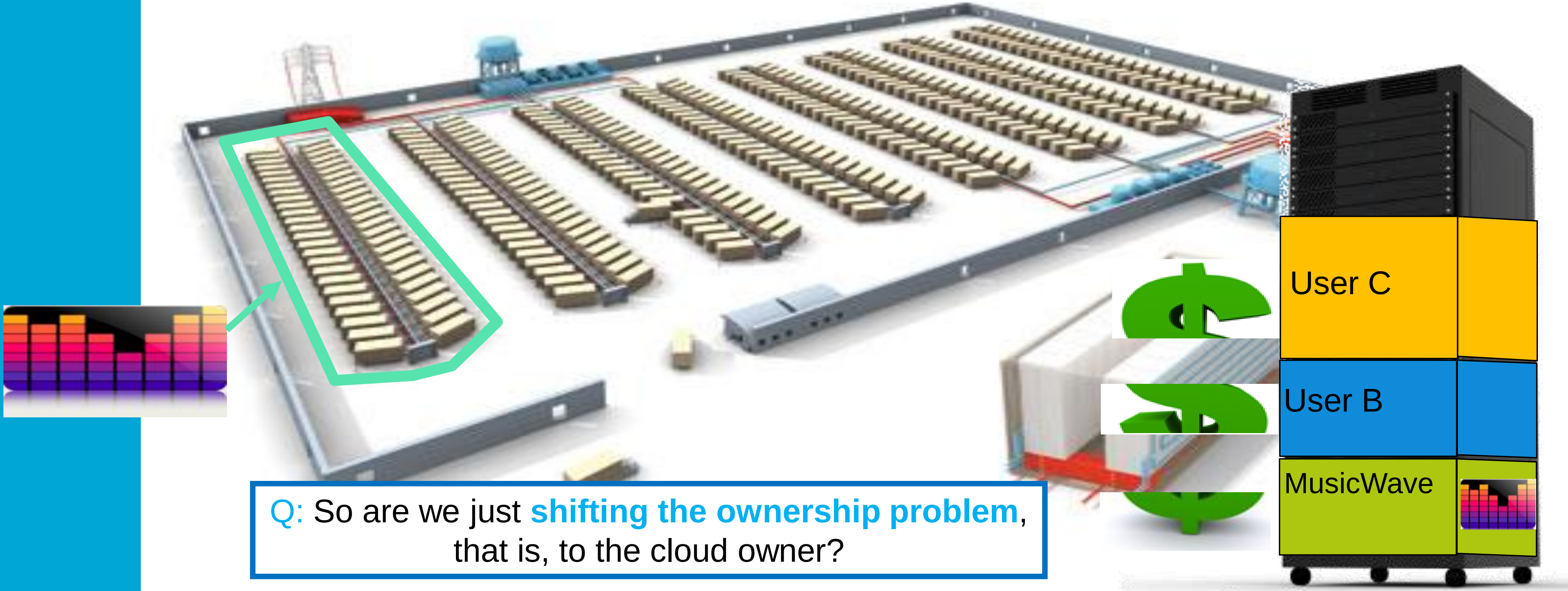
Deploy on IaaS Cloud

- NO big up-front commitment
- Load variability: supported



(Source: A. Antoniou, MSc Defense, TU Delft, 2012. Original idea: V. Nae, 2008.)

Inside a Cloud Datacenter: Infrastructure as a Service



Q: So are we just **shifting the ownership problem**, that is, to the cloud owner?

(Source: A. Antoniou, MSc Defense, TU Delft, 2012. Original idea: A. Iosup, 2011.)

Agenda

1. The Golden Age of scalable computing
2. The core idea of cloud computing

3. Enabling technologies (homework)

4. The scheduling challenge
5. The Ecosystem Navigation challenge
6. The Big Cake challenge
7. Jevon's Effect challenge (IEEE Scale Challenge Award)

8. Take-home message

THE PIZZA-BOX STACK

- The 1U server

Image source: http://www.avadirect.com/images/showroom/346739_1.jpg

THE PIZZA-BOX STACK

- The 1U server

Image source: http://www.avadirect.com/images/showroom/346739_1.jpg

THE PIZZA-BOX STACK

- The 1U server

Image source: http://www.avadirect.com/images/showroom/346739_1.jpg

THE PIZZA-BOX STACK

- The 1U server
- The 19” server rack (42U is now standard)

Image source: http://www.avadirect.com/images/showroom/346739_1.jpg

Image source: http://www.xrackpro.com/v/vspfiles/images/optimized/19_inch_rack.jpg

THE DATA CENTER NETWORK

- Network bandwidth per rack
  - 1 x 48-port GigE switch = 40 UP-, 8 DOWN-links

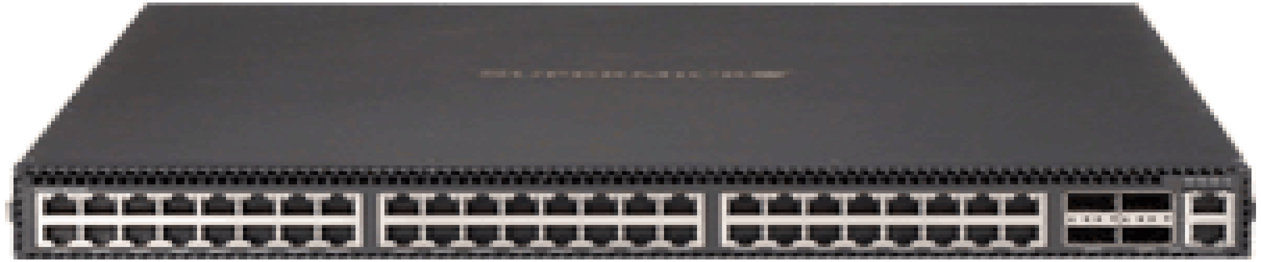


Image source: http://www.supermicro.com/a_images/products/Accessories/SSE-X3348T.gif

- Network bandwidth per socket
  - (fast) 1 Gbps for 10 GigE rack switch
  - (slow) 100 Mbps for 1 GigE rack switch
  - (exorbitant) 10 GBps for ncHT3 (supercomputing class)

Source: Dennis Abts (Google, Inc.) and John Kim (KAIST), High Performance Data Center Networks, 2011

SERVERS + SERVER RACKS +
INTRA-RACK NETWORK + INTER-RACK NETWORK

AN ENTIRE FLOOR IN A
GOOGLE DATACENTER

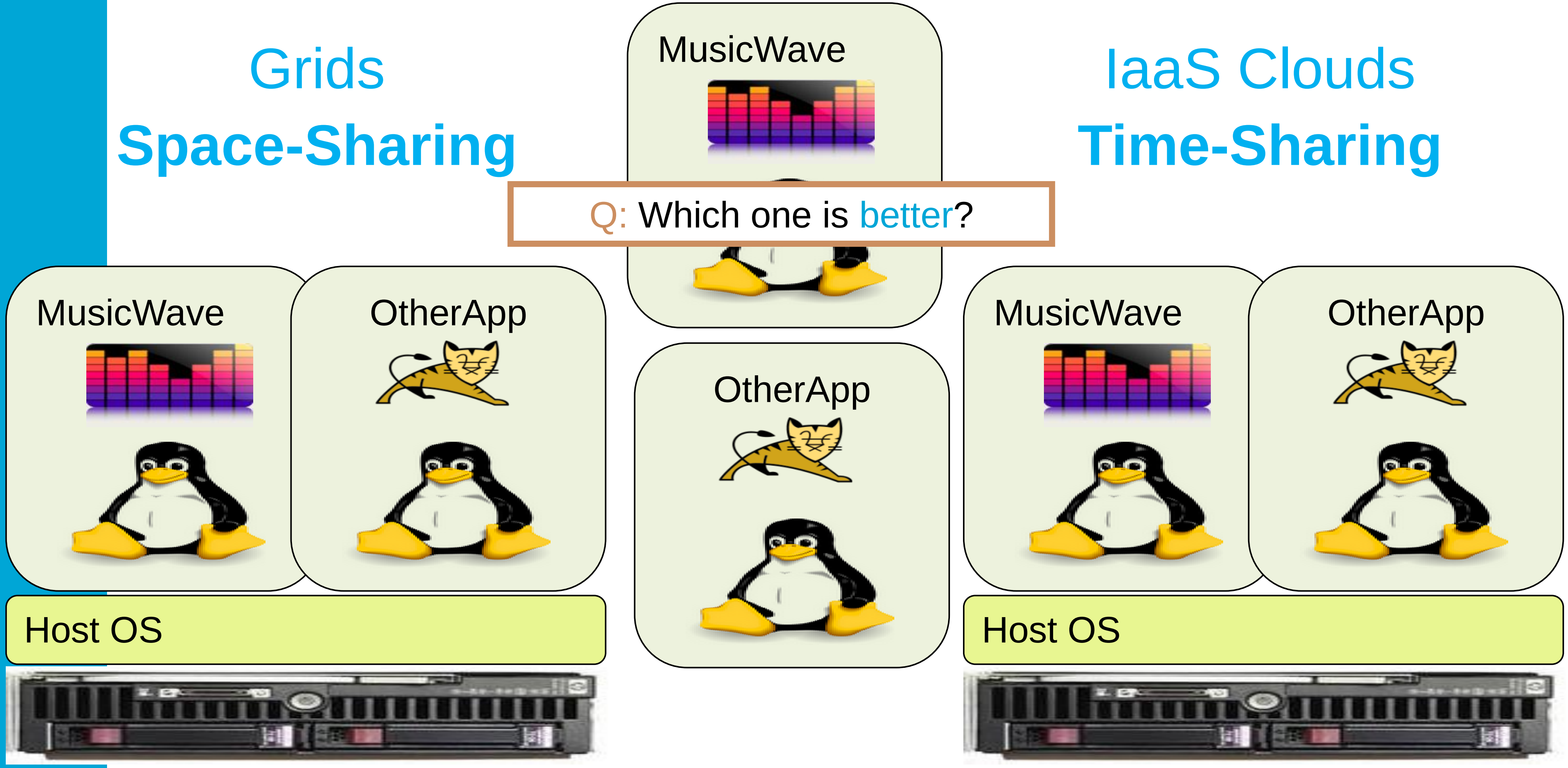
Resource Sharing Models

Grids

Space-Sharing

IaaS Clouds

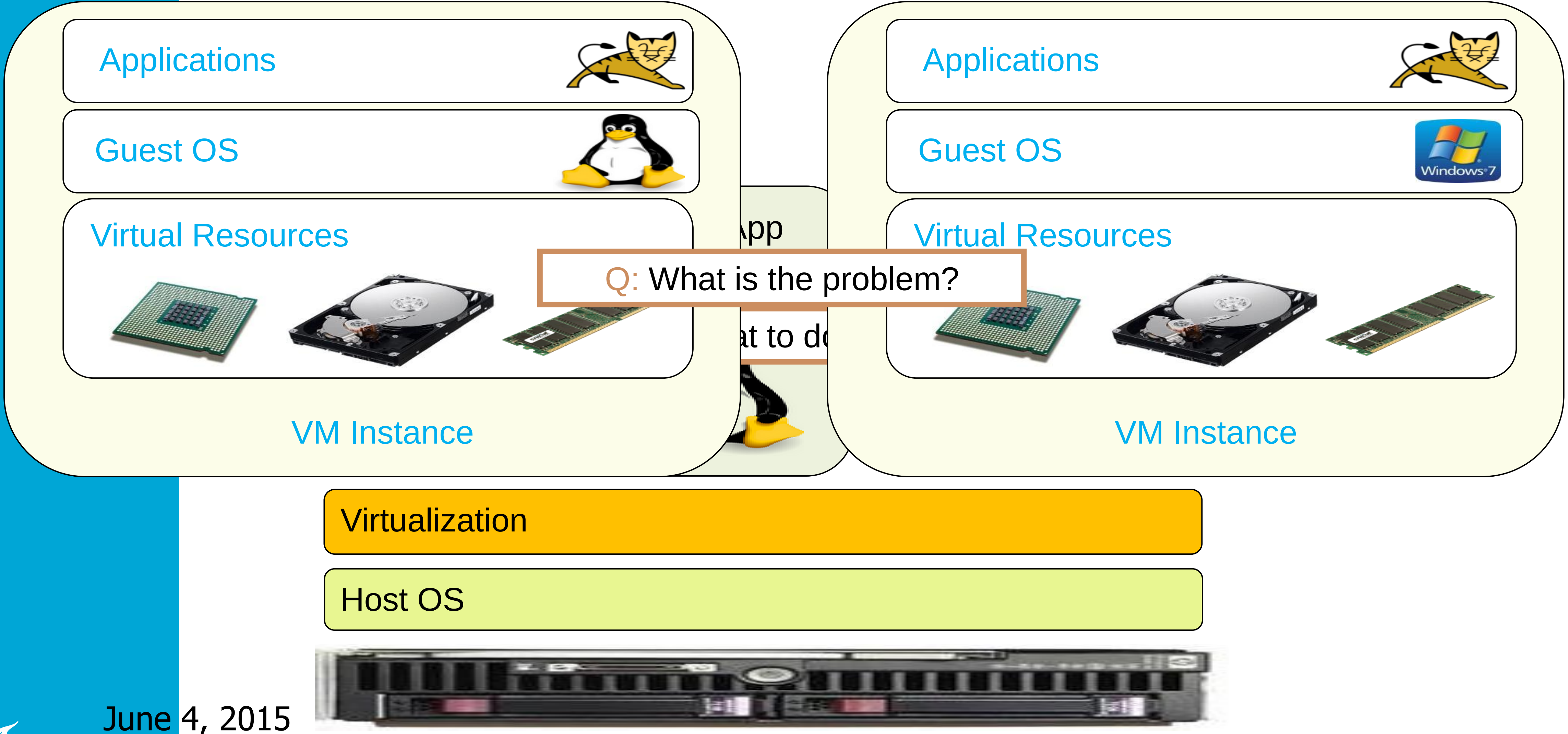
Time-Sharing



Q: Which one is better?

(Source: A. Iosup, Course IN4392, TU Delft, 2014)

Virtualization



Applications

Guest OS

Virtual Resources

VM Instance

Applications

Guest OS

Virtual Resources

VM Instance

Virtualization

Host OS

Q: What is the problem?

(Source: A. Iosup, Course IN4392, TU Delft, 2014)

Agenda

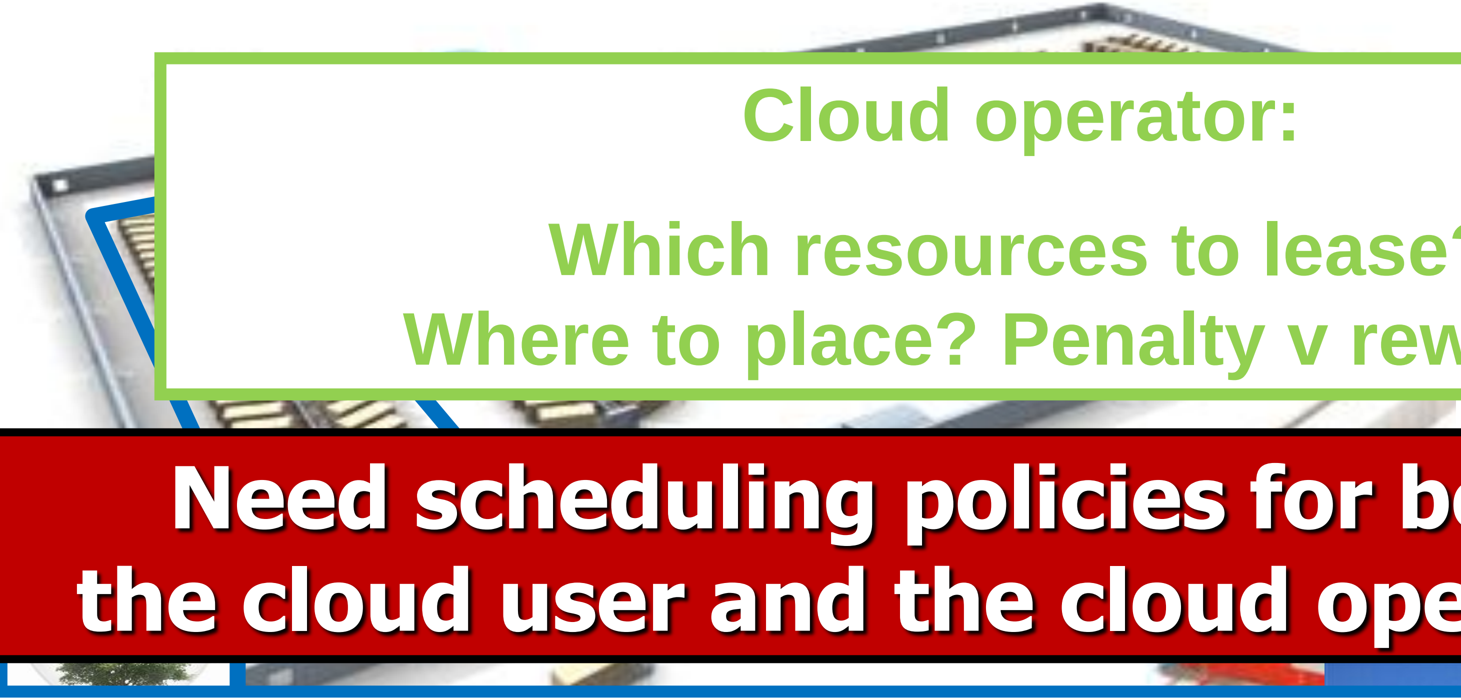
1. The Golden Age of scalable computing
2. The core idea of cloud computing

3. Enabling technologies (homework)

4. The scheduling challenge
5. The Ecosystem Navigation challenge
6. The Big Cake challenge
7. Jevon's Effect challenge (IEEE Scale Challenge Award)

8. Take-home message

The Scheduling Challenge



Cloud operator:

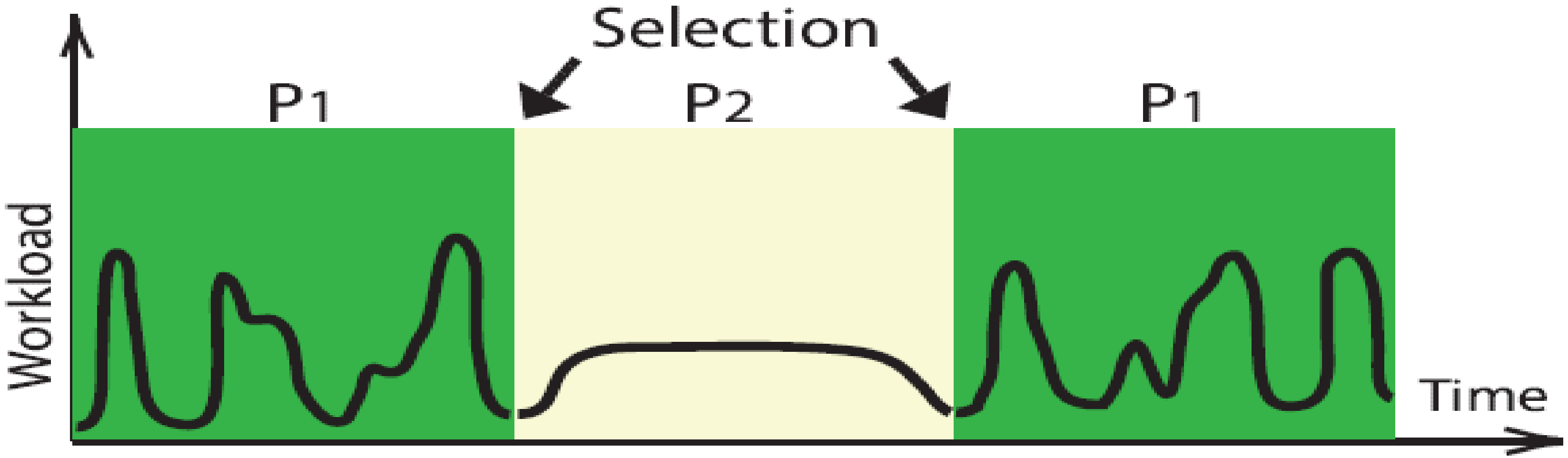
Which resources to lease?
Where to place? Penalty v reward?

Need scheduling policies for both
the cloud user and the cloud operator

Cloud customer:

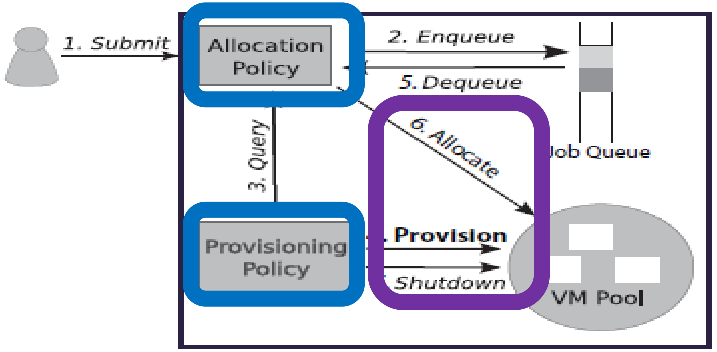
Which resources to lease?
When? How many? When stop?
Utility functions?

Portfolio Scheduling, In A Nutshell



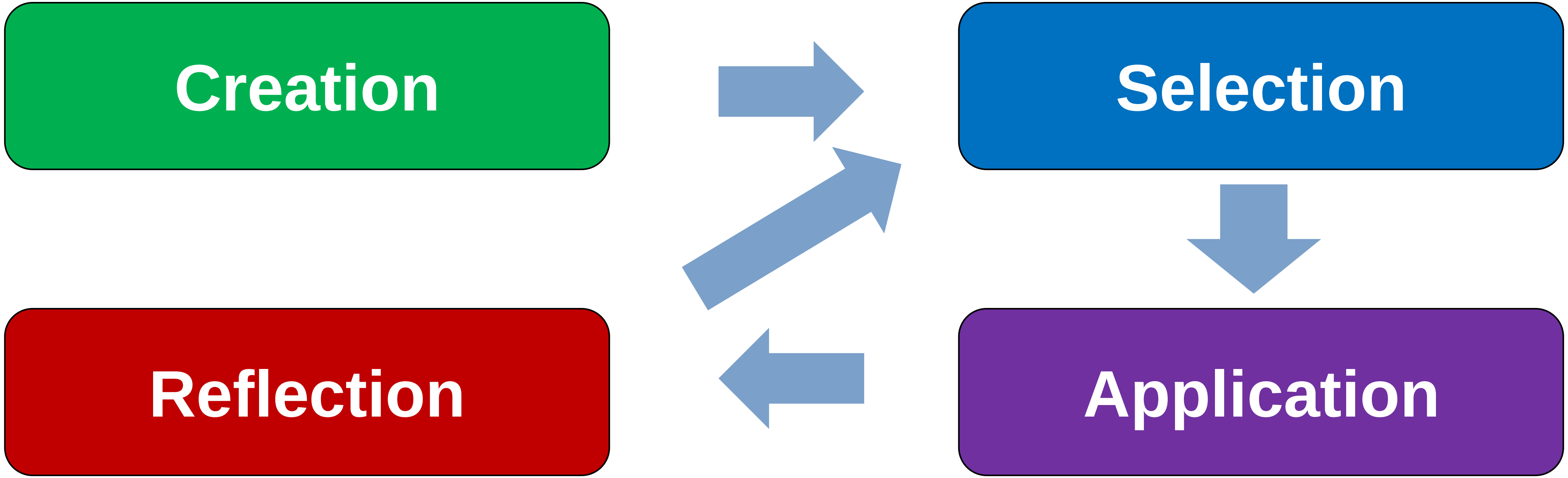
- Create a set of scheduling policies
  - Resource provisioning and allocation policies, in this work
- Online selection of the active policy, at important moments
- Same principle for other changes: pricing model, system, …

Portfolio Scheduling: Process



Which policies to include?

Which policy to activate?



Creation

Selection

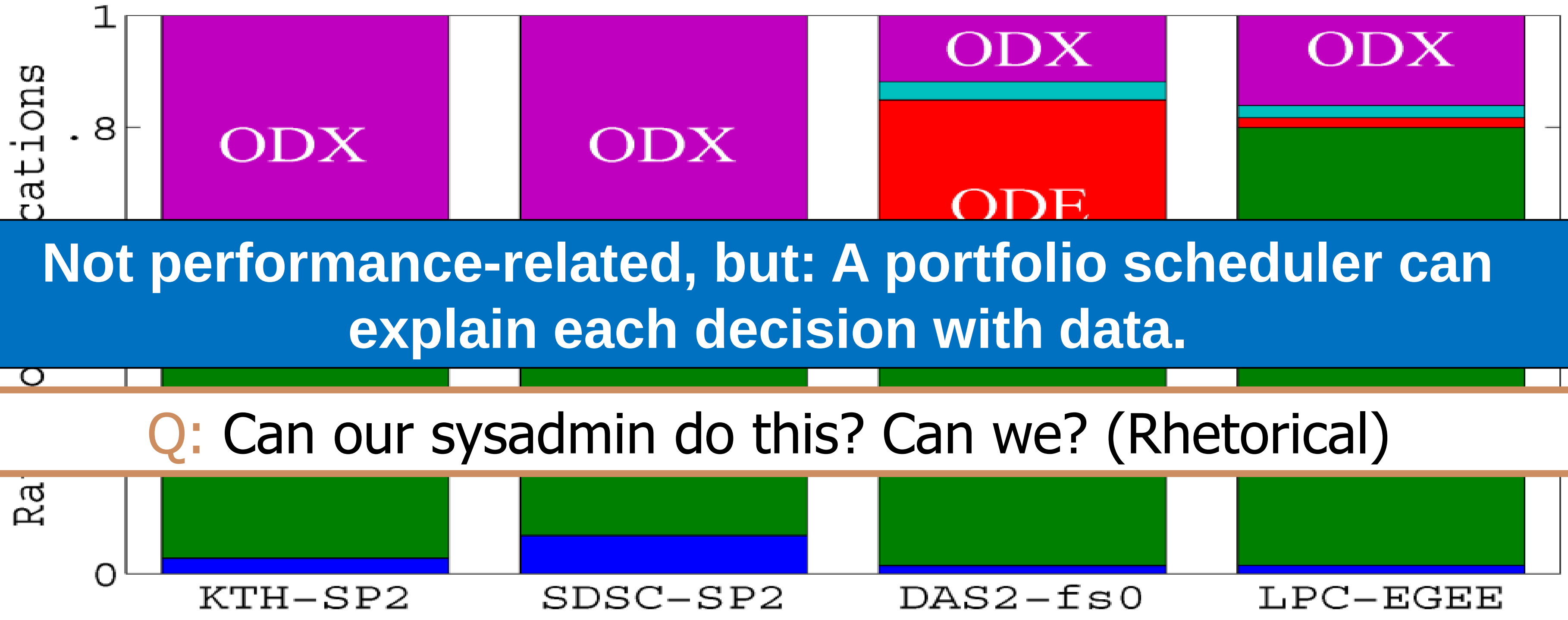
Application

Reflection

Which changes to the portfolio?

Which resources? What to log?

Good Results for Scientific Computing, Business-Critical, and Online Gaming Workloads



Not performance-related, but: A portfolio scheduler can explain each decision with data.

Q: Can our sysadmin do this? Can we? (Rhetorical)

- No single dominant policy

Agenda

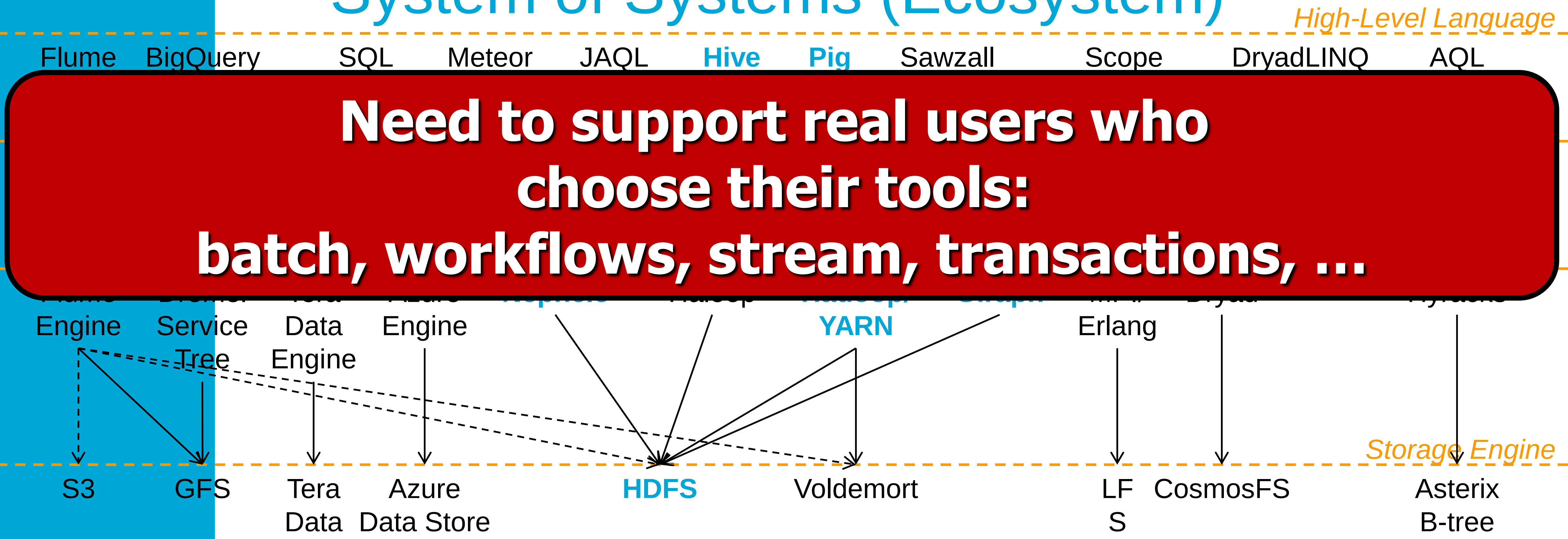
1. The Golden Age of scalable computing
2. The core idea of cloud computing

3. Enabling technologies (homework)

4. The scheduling challenge
5. The Ecosystem Navigation challenge
6. The Big Cake challenge
7. Jevon's Effect challenge (IEEE Scale Challenge Award)

8. Take-home message

Big Data Processing = System of Systems (Ecosystem)

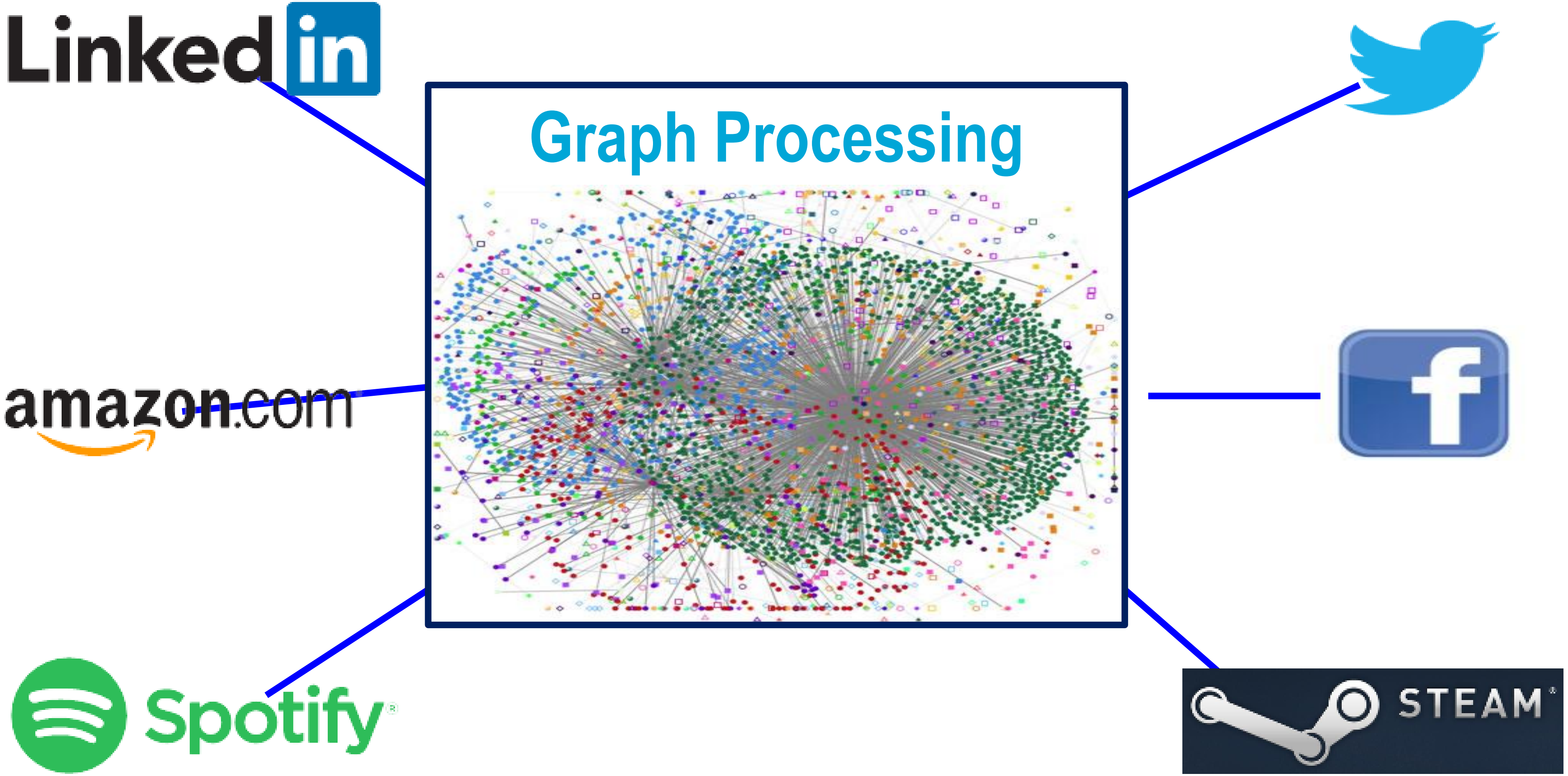


Need to support real users who choose their tools: batch, workflows, stream, transactions, …

* Plus Zookeeper, CDN, etc.

Adapted from: Dagstuhl Seminar on Information Management in the Cloud, http://www.dagstuhl.de/program/calendar/partlist/?semnr=11321&SUOG

THE DATA DELUGE: LARGE-SCALE GRAPHS TENS OF BILLIONS OF EDGES



Graph Processing

PLATFORM DIVERSITY

Neo4j

Oracle Labs
PGX

APACHE
GIRAPH

YARN

Spark

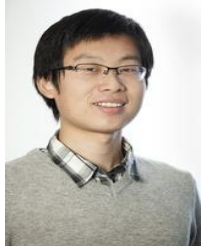
GraphX

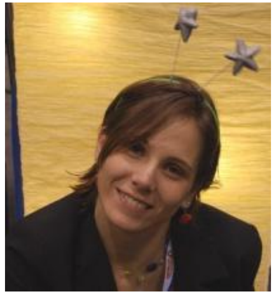
hadoop

GraphLab

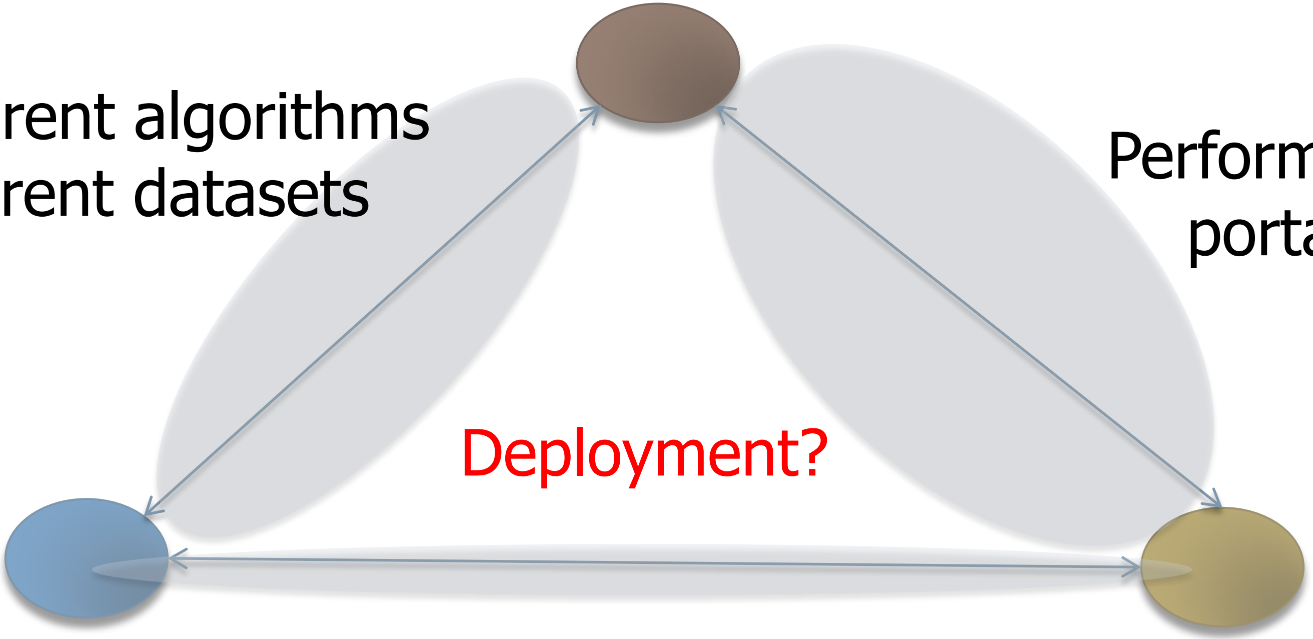
StratoSphere
Above the Clouds

P-A-D TRIANGLE





Algorithm



Different algorithms
for different datasets

Performance enabled,
portability disabled

Deployment?

Dataset

No systematic findings yet

Platform

GRAPHALYTICS: THE FIRST COMPREHENSIVE BENCHMARK FOR BIG DATA GRAPH PROCESSING

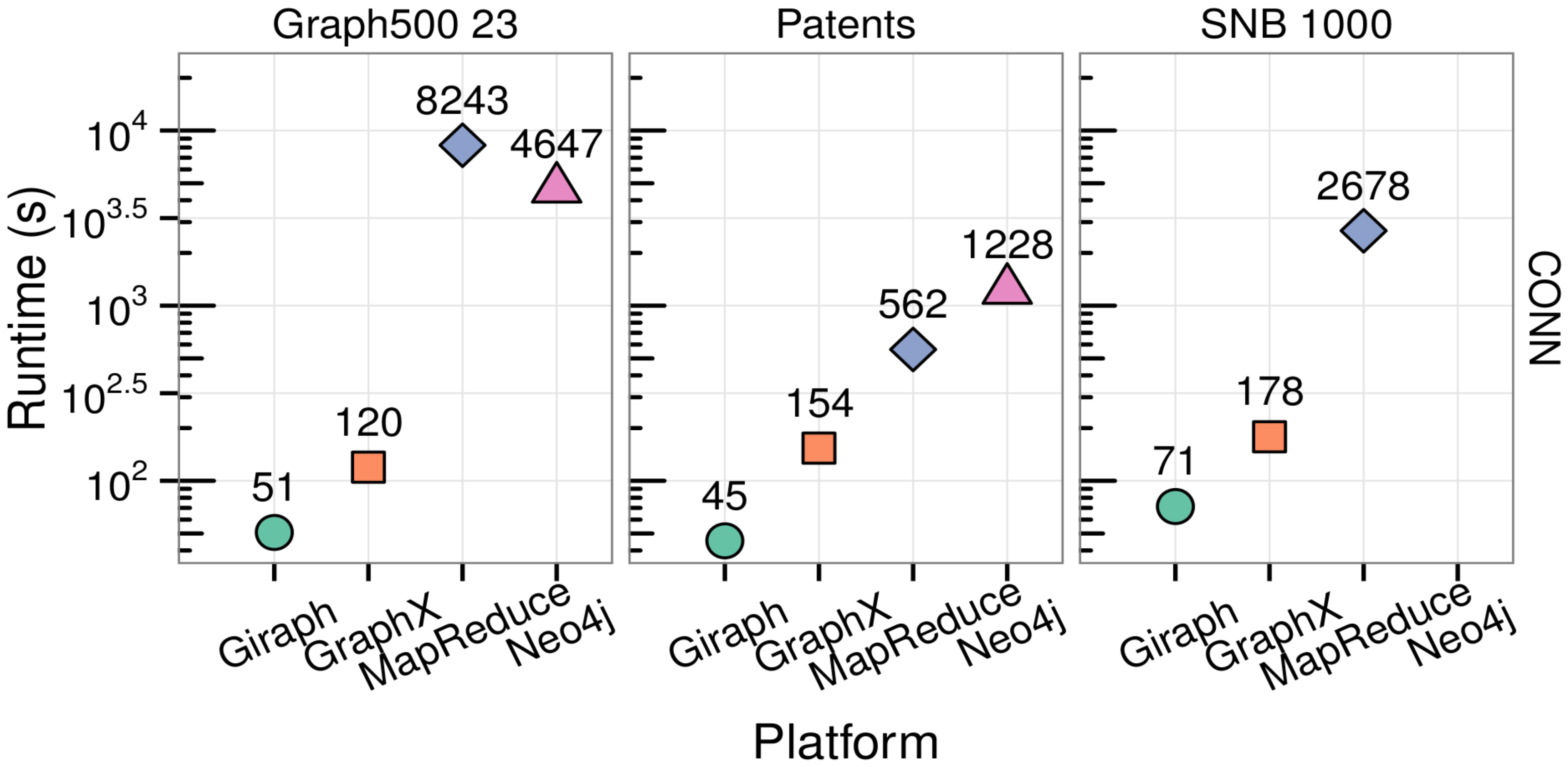
- Advanced benchmarking harness
- Choke-point analysis
- Realistic graph generator

- Co-sponsored by Oracle
- Supported by LDBC, partially developed through SPEC RG

LDBC *The graph & RDF benchmark reference*

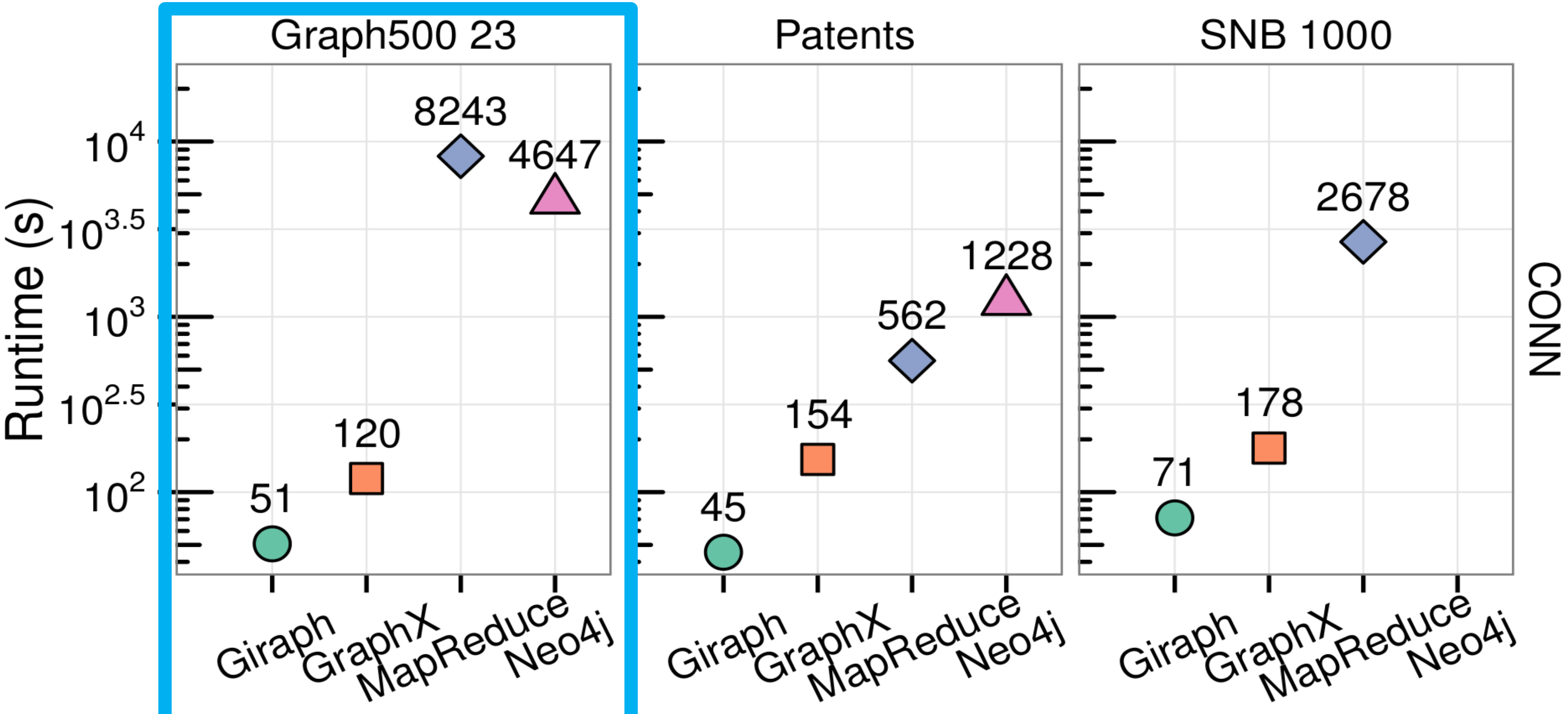
spec Research

RUNTIME



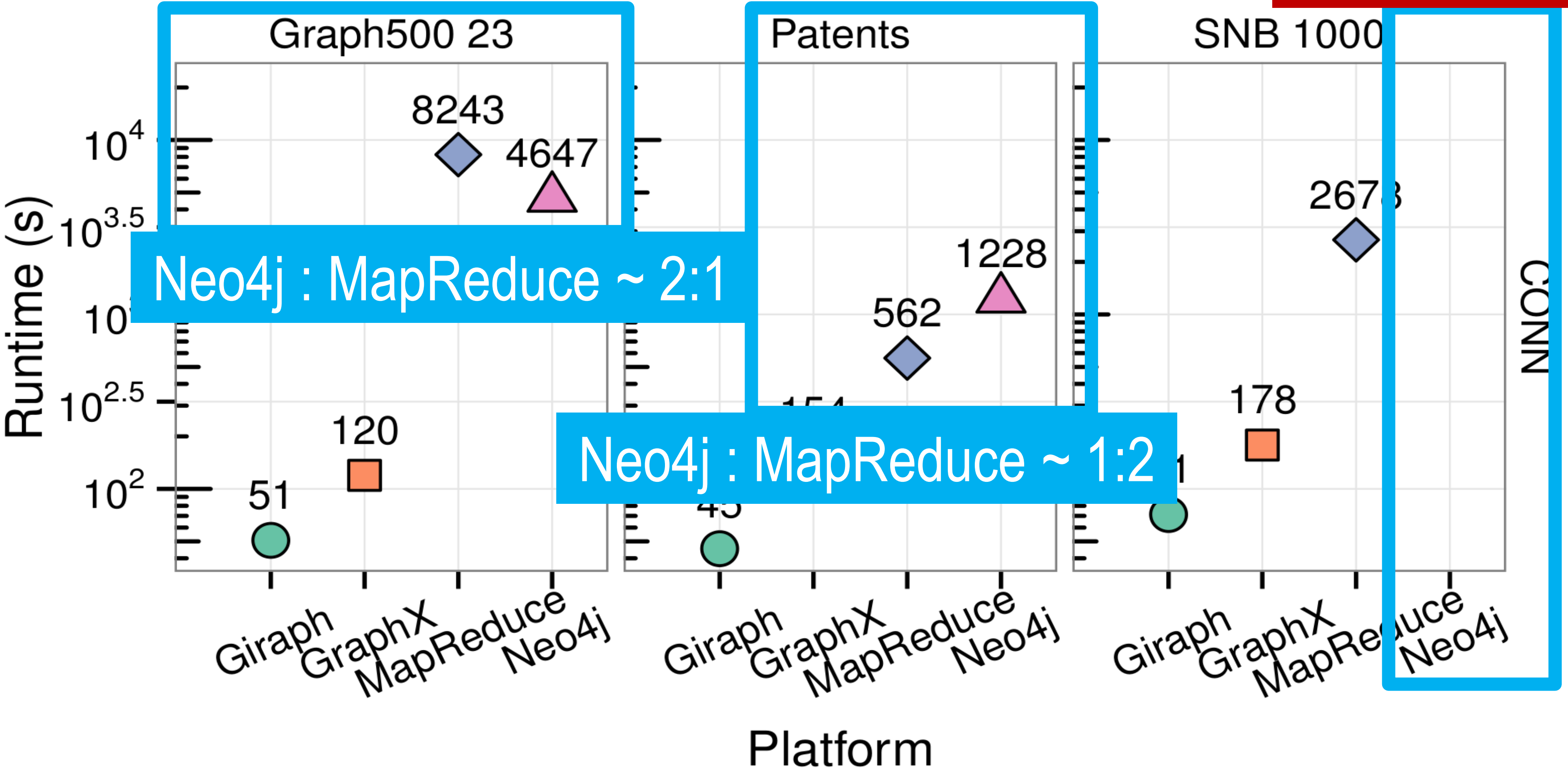
Graph500 23
Patents
SNB 1000
CONN
Runtime (s)
10^4
$10^{3.5}$
10^3
$10^{2.5}$
10^2
8243
4647
120
51
1228
562
154
45
2678
178
71
Giraph
GraphX
MapReduce
Neo4j
Platform

RUNTIME: THE PLATFORM HAS LARGE IMPACT



2 orders of magnitude difference due to platform

RUNTIME: THE DATASET HAS LARGE IMPACT



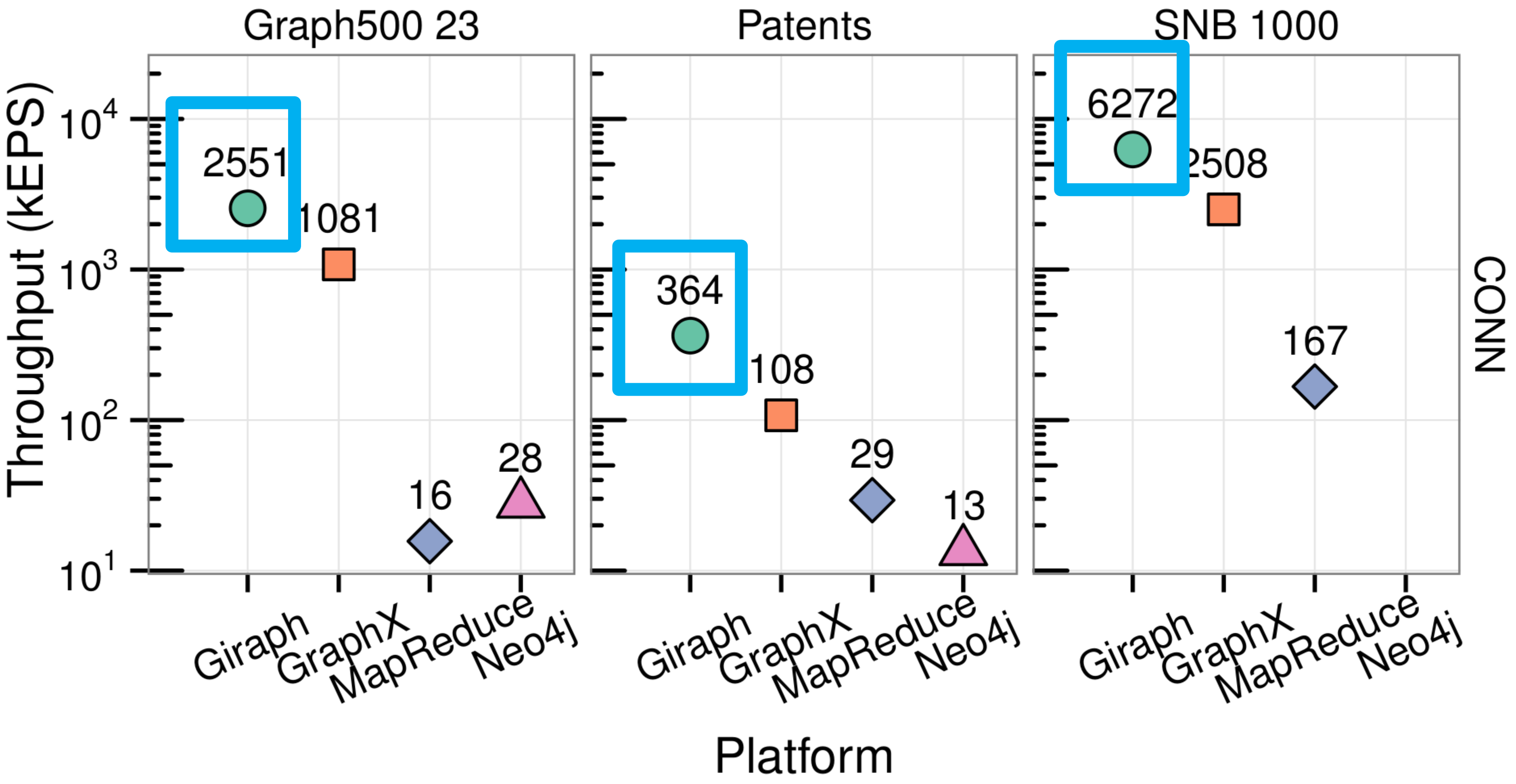
Neo4j fails

Neo4j : MapReduce ~ 2:1

Neo4j : MapReduce ~ 1:2

THROUGHPUT: THE DATASET STRUCTURE MATTERS!

20x difference



Agenda

1. The Golden Age of scalable computing
2. The core idea of cloud computing

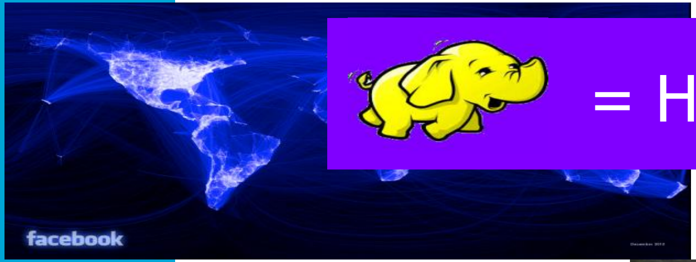
3. Enabling technologies (homework)

4. The scheduling challenge
5. The Ecosystem Navigation challenge
6. The Big Cake challenge
7. Jevon's Effect challenge (IEEE Scale Challenge Award)

8. Take-home message

The “Big Cake” In the Datacenter

Online Social Networks



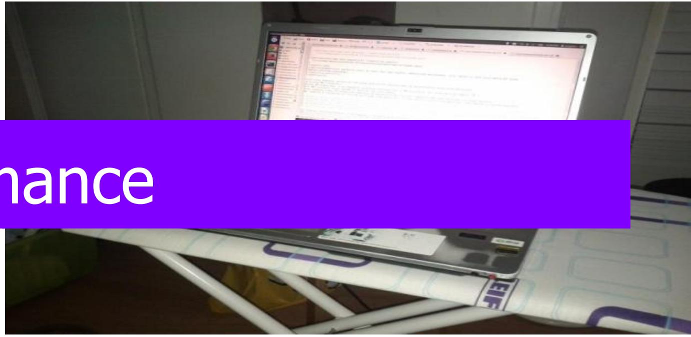
= Hadoop / MapReduce framework

Financial Analysts

Need multi-tenant, self-aware schedulers and resource managers

Universe Explorers

Big Data Enthusiast

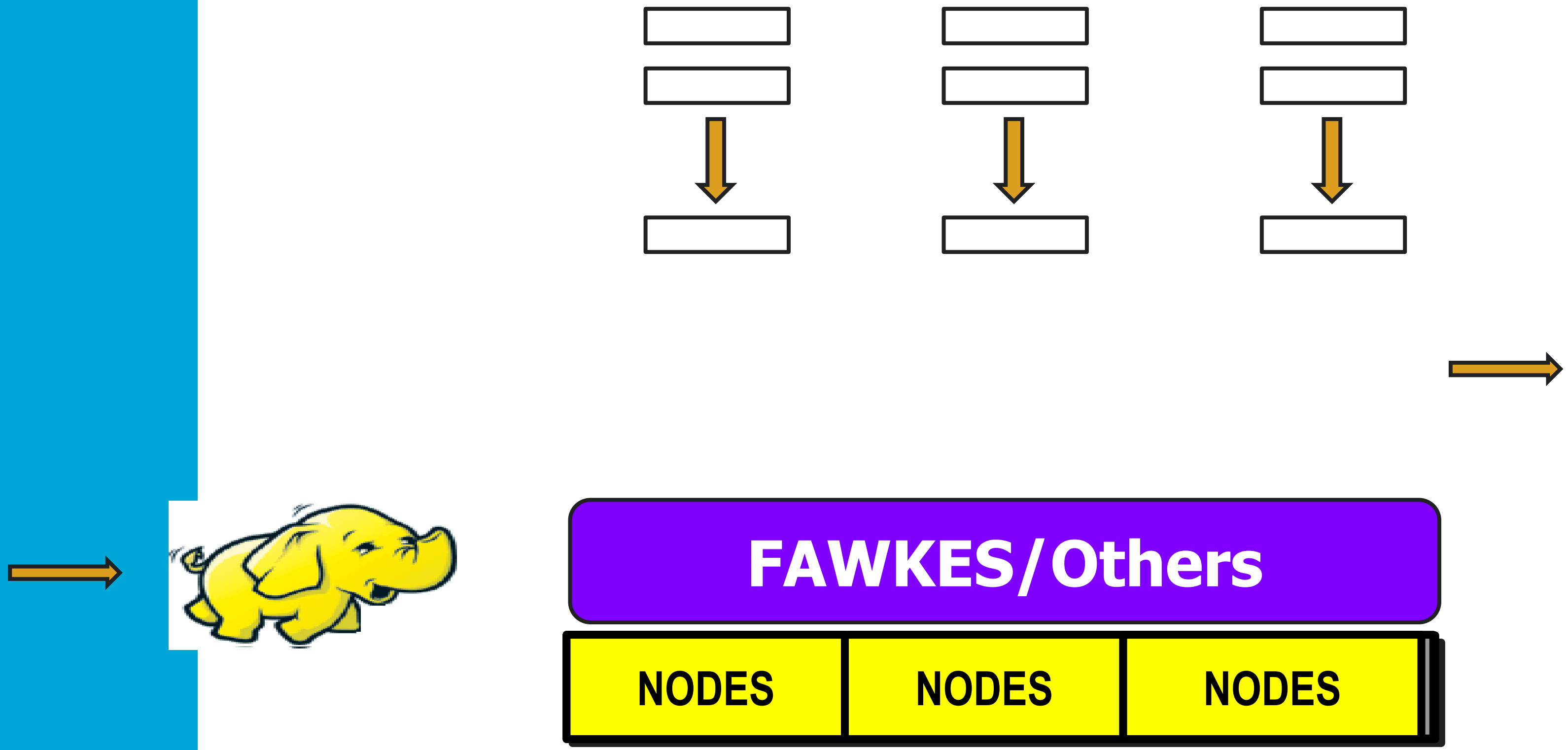


Multiple frameworks = Isolation, especially performance

(Source: B. Ghit et al., SIGMETRICS 2014)

DYNAMIC BIG DATA PROCESSING

Fawkes = Elastic MapReduce



FAWKES

Job submissions

Frameworks

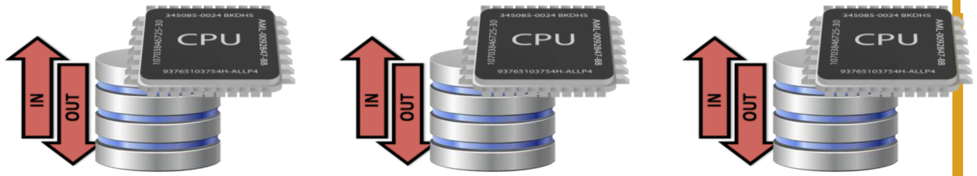
Resource manager

Infrastructure

Ghit, Yigitbasi, Iosup, Epema, Iosup. Balanced Resource Allocations Across Multiple Dynamic MapReduce Clusters.
ACM SIGMETRICS 2014.

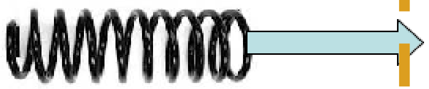
ELASTICITY FOR MAPREDUCE FRAMEWORRKS

Core nodes



INPUT/OUTPUT DATA

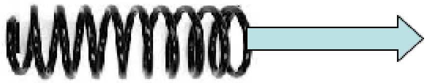
- ○ Classical deployment
- ○ Uniform data distribution
- ○ **No removal**



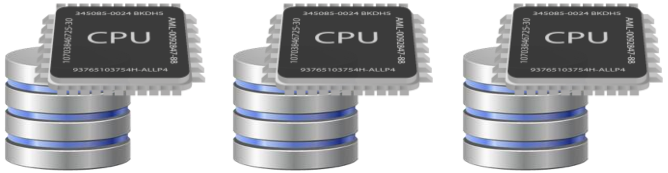
Transient nodes (TR)

NO DATA

- ○ No local storage
- ○ R/W from/to core nodes
- ○ **Instant removal**

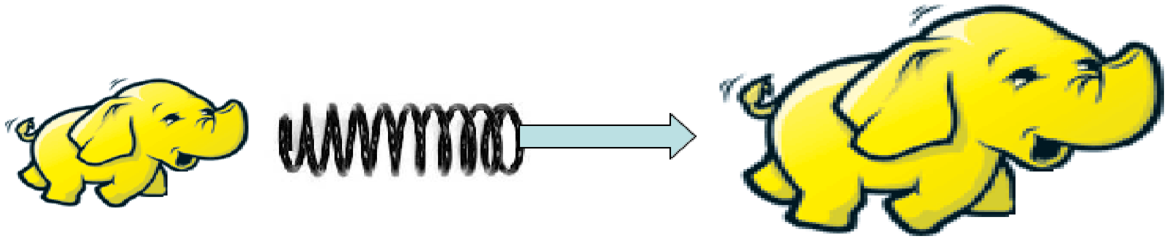


Trans-core nodes (TC)



OUTPUT DATA

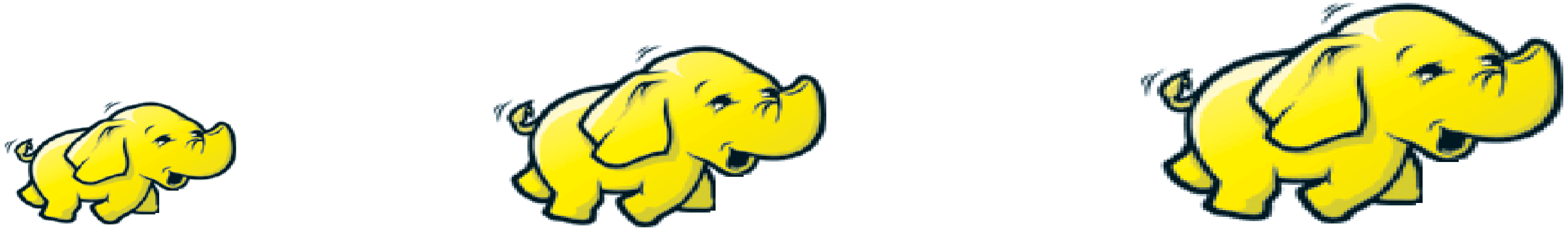
- ○ Local storage, no input
- ○ Only R from core nodes
- ○ **Delayed removal**



FAWKES IN A NUTSHELL [1/2]

Because workloads may be time-varying:

- Poor resource utilization
- Imbalanced service levels



$w_1 < w_2 < w_3$

1. Fair framework size:

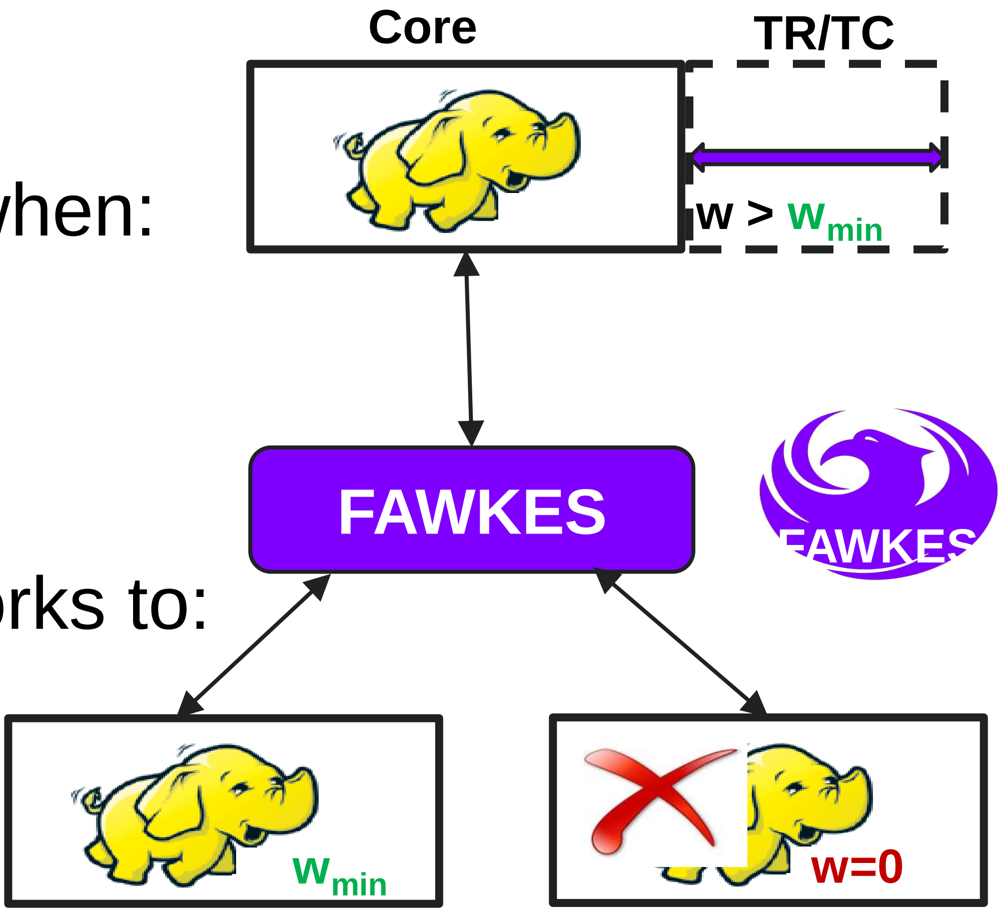
$$s_i = \frac{w_i}{w_1 + w_2 + w_3}, \quad i = 1,2,3$$

(Source: B. Ghit et al., SIGMETRICS 2014)

FAWKES IN A NUTSHELL [2/2]

2. Updates dynamic weights when:
   - New frameworks arrive
   - Framework states change

3. Shrinks and grows frameworks to:
   - Allocate new frameworks
   - Give fair shares to existing frameworks
   - Eliminate unused frameworks



(Source: B. Ghit et al., SIGMETRICS 2014)

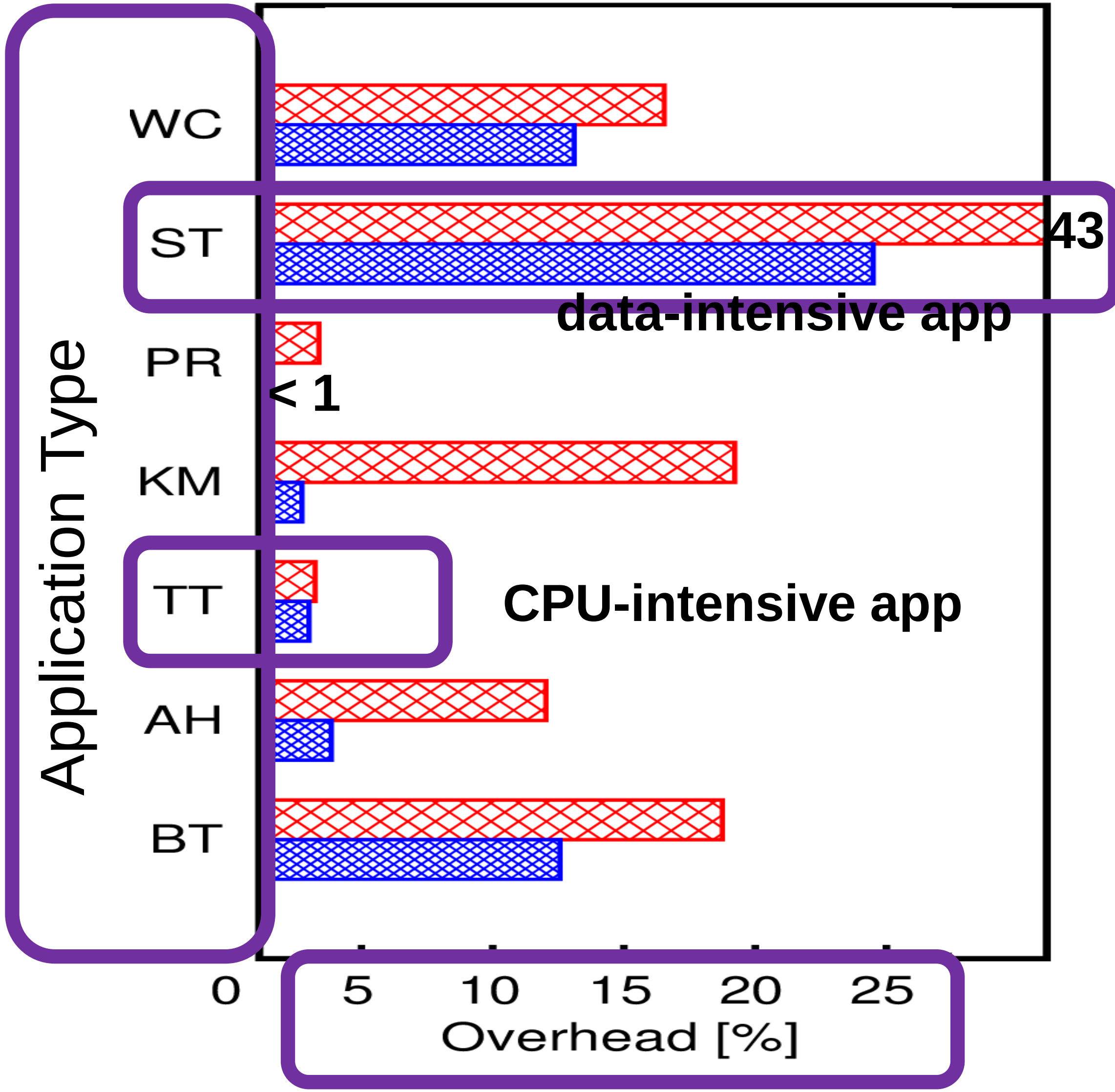
PERFORMANCE OF DYNAMIC MAPREDUCE

10 core +10xTR ▨

10 core +10xTC ▨

vs.

20 core nodes (baseline)



43

data-intensive app

< 1

CPU-intensive app

TR - **good** for compute-intensive workloads.

TC - **needed** for disk-intensive workloads.

Dynamic MapReduce:
< 25% overhead

Fawkes also reduces imbalance

(Source: B. Ghit et al., SIGMETRICS 2014)

Agenda

1. The Golden Age of scalable computing
2. The core idea of cloud computing

3. Enabling technologies (homework)

4. The scheduling challenge
5. The Ecosystem Navigation challenge
6. The Big Cake challenge
7. Jevon's Effect challenge (IEEE Scale Challenge Award)

8. Take-home message

Jevon’s Effect: More Efficient, Less Capable

Over 500 YouTube videos have at least 100,000,000 viewers each.

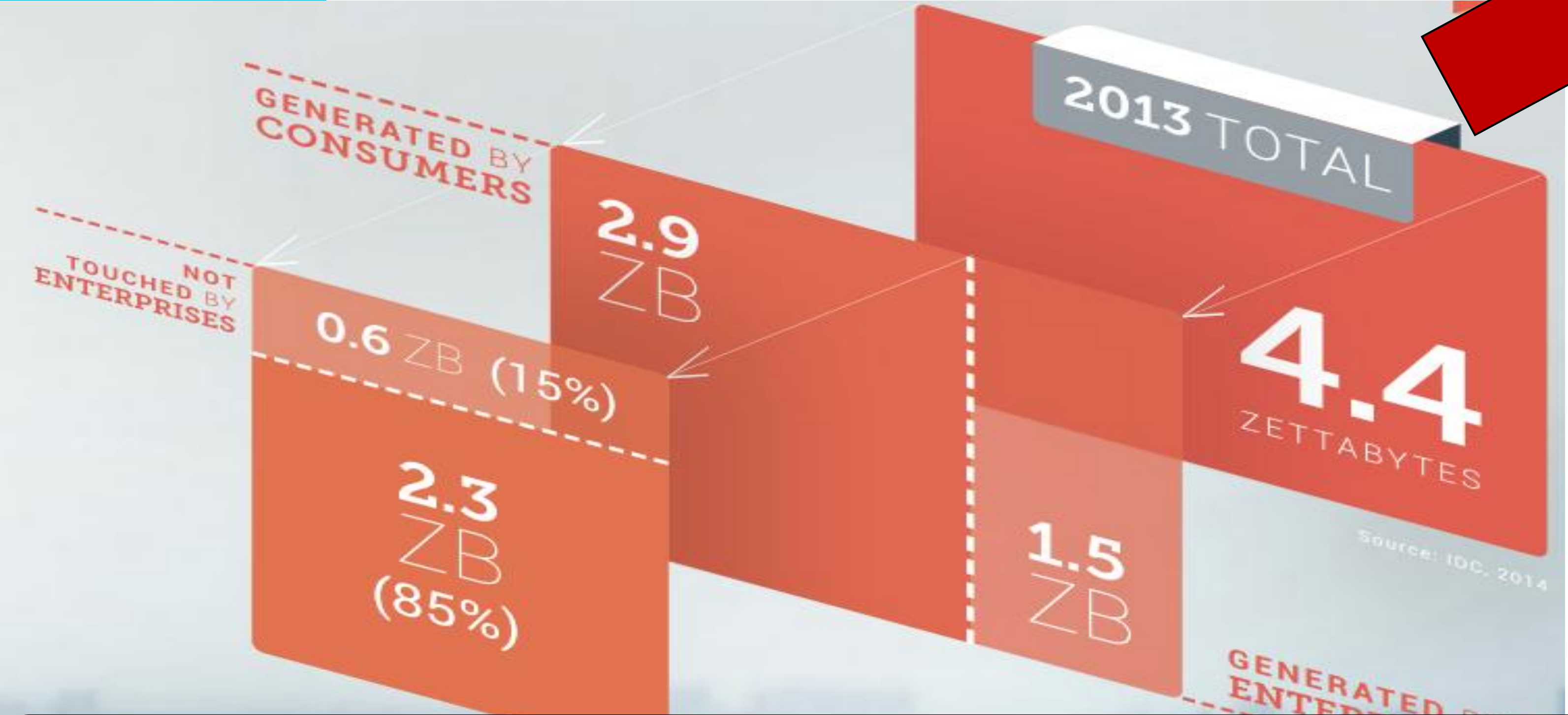
If you want to help kill the planet:
https://www.youtube.com/playlist?list=PLirAqAtl_h2r5g8xGajEwdXd3x1sZh8hC

PSY Gangnam consumed ~500GWh

= more than entire countries* in a year (*41 countries),
= over 50MW of 24/7/365 diesel, 135M liters of oil,
= 100,000 cars running for a year, ...

Source: Ian Bitterlin and Jon Summers, UoL, UK, Jul 2013.
Note: Psy has now >3 billion views (Jun 2015).

The New "Jevon's Effect": The "Data Deluge"



44
ZETTABYTES
2020

Data Deluge =

data generated by humans and devices (IoT)

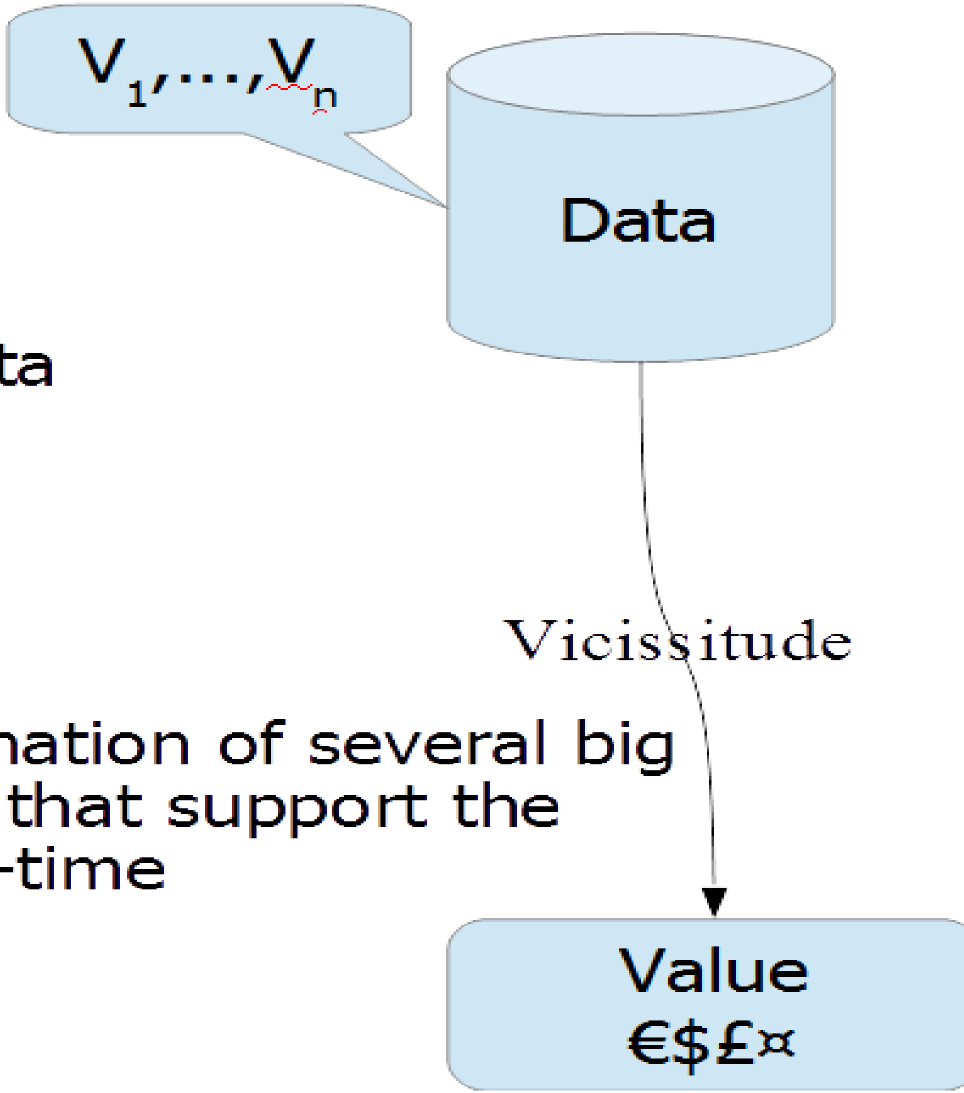
- Interacting
- Understanding
- Deciding
- Creating

Need to address
Volume, Velocity, Variety of Big Data*

* New Vs later: ours is "vicissitude"

Sources: IDC, EMC.

Vs of big data

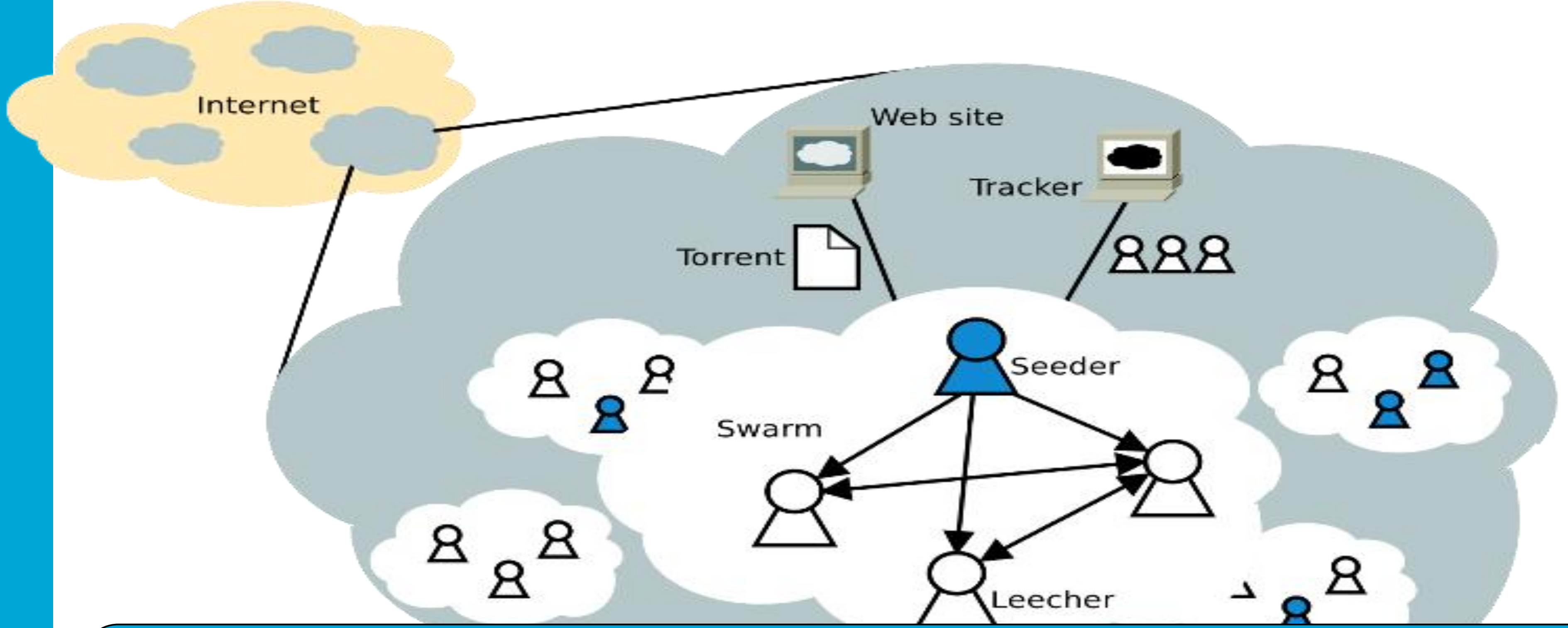


- Volume – large scale of data
- Variety – different forms of data
- Velocity – timeliness of data
- Veracity – uncertainty of data
- **Vicissitude** – dynamic combination of several big data Vs in processing systems that support the addition of new queries at run-time

vicissitude *noun* [vɨˈsɪsɨˌtu()d]:
a favorable or unfavorable event or situation that occurs by chance; a fluctuation of state or condition

http://merriam-webster.com/dictionary/vicissitude

OBSERVING BITTORRENT: MANAGING A TYPICAL GLOBALLY DISTRIBUTED SYSTEM

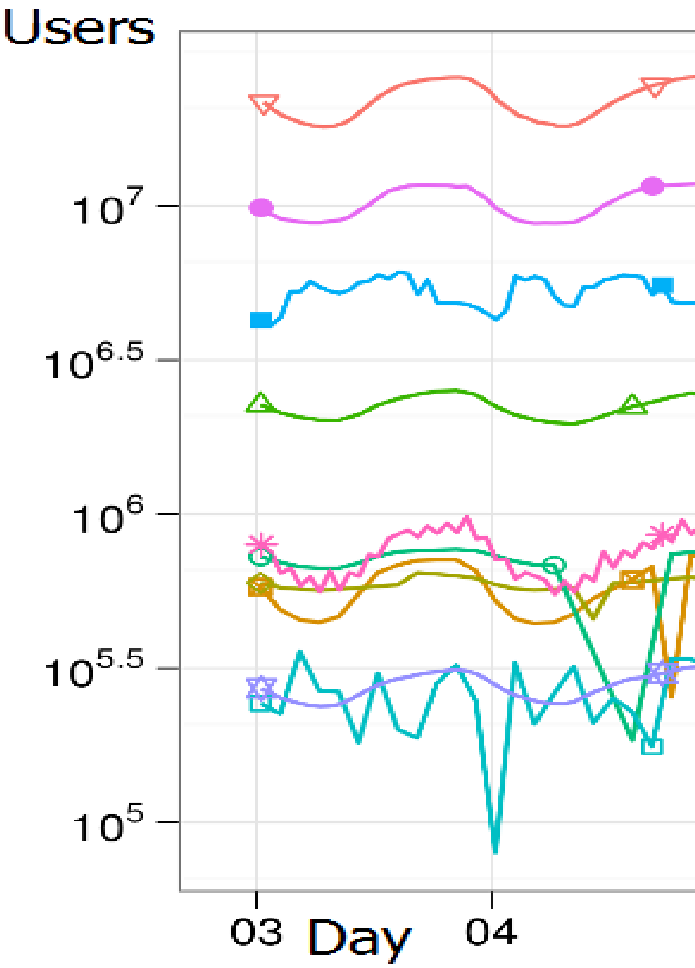


Most used protocol on Internet, by upload volume [1]
One third (US) to half (EU) of residential upload
Over 100 million users [2]

[1] https://sandvine.com/downloads/general/global-internet-phenomena/2013/2h-2013-global-internet-phenomena-report.pdf
[2] http://www.bittorrent.com/company/about/ces_2012_150m_users

BTWORLD: A TYPICAL BIG DATA PROJECT

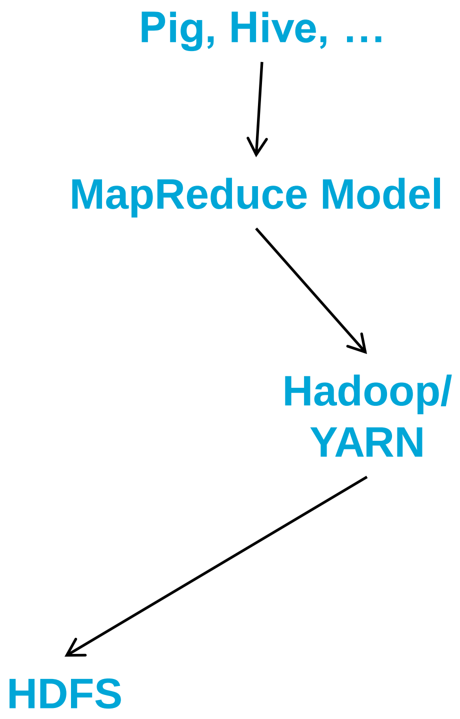
- Ongoing longitudinal study, 5 YEARS
- Data-driven project to understand BitTorrent: data first, ask questions later
    - Over 15 TB of structured and semi-structured data added during the project
    - Queries added during project, e.g.,
      How does the BitTorrent population vary?
      How does BitTorrent change over time?



Wojciechowski et al. Towards observing the global BitTorrent file-sharing network. HPDC 2010

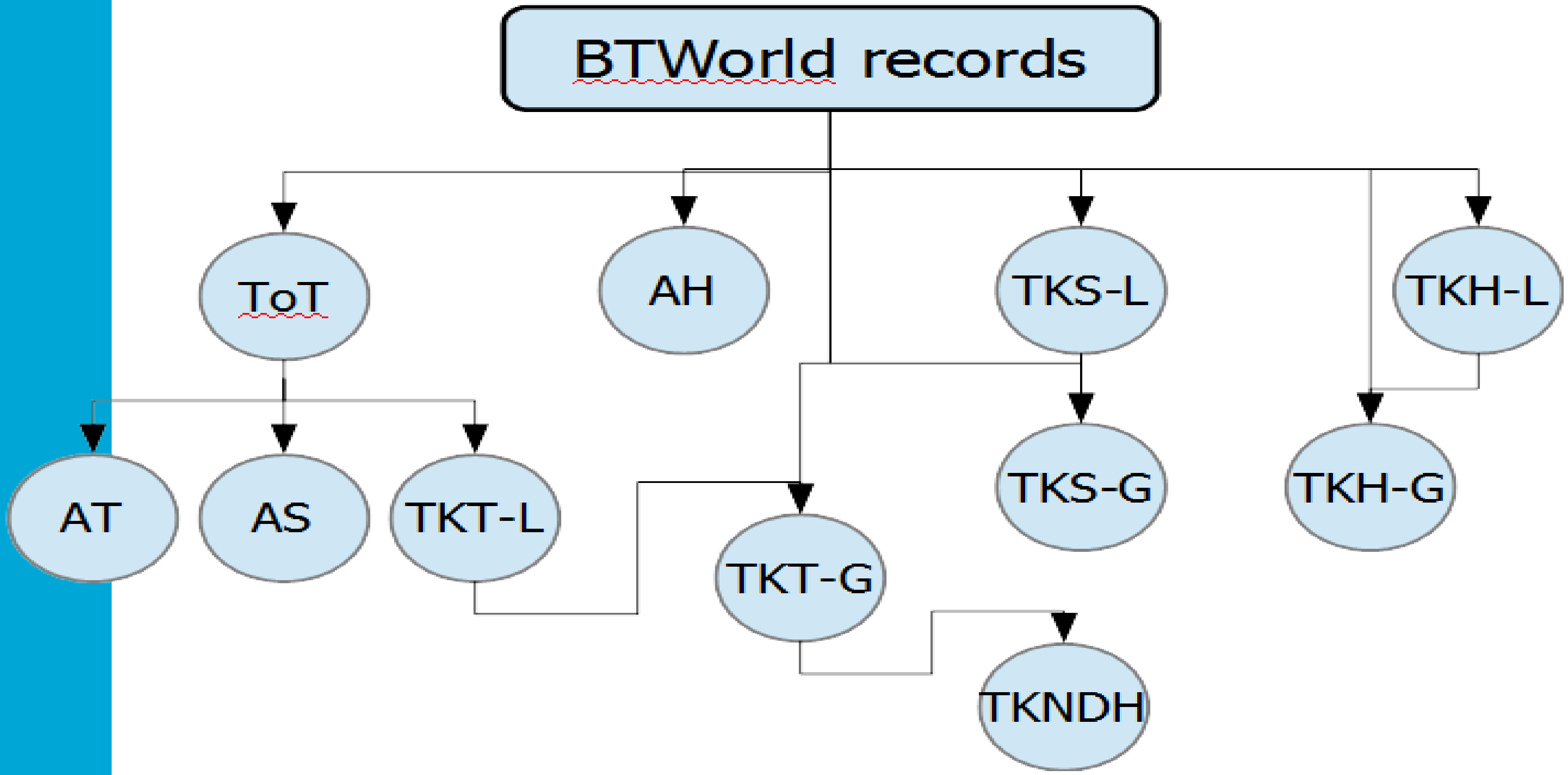
THE MAPREDUCE ECOSYSTEM (A BIG PROBLEM IN BIG DATA)

- Widely used in industry and academia
  - Similar to other big data stacks
- Complex software to tune
  - 100s of parameters
  - Non-linear effects common
- Lots of issues cause crashes [1]
- Focus on Small and Medium Enterprises (60% GPD)
  - No resources or even competence to fix issues
  - Difficult to make stack work for own problems

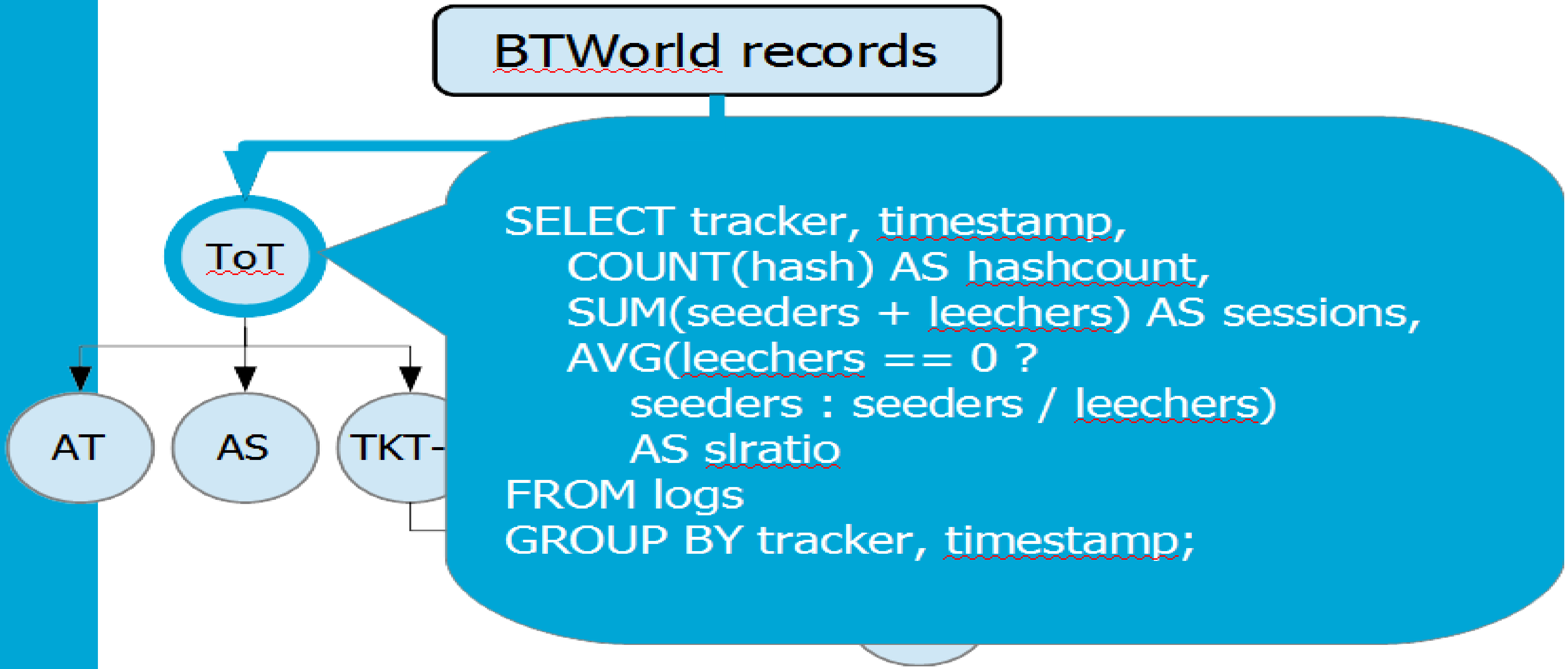


[1] Ewen et al., “Spinning Fast Iterative Data Flows”, PVLDB 2012

THE BTWORLD WORKFLOW

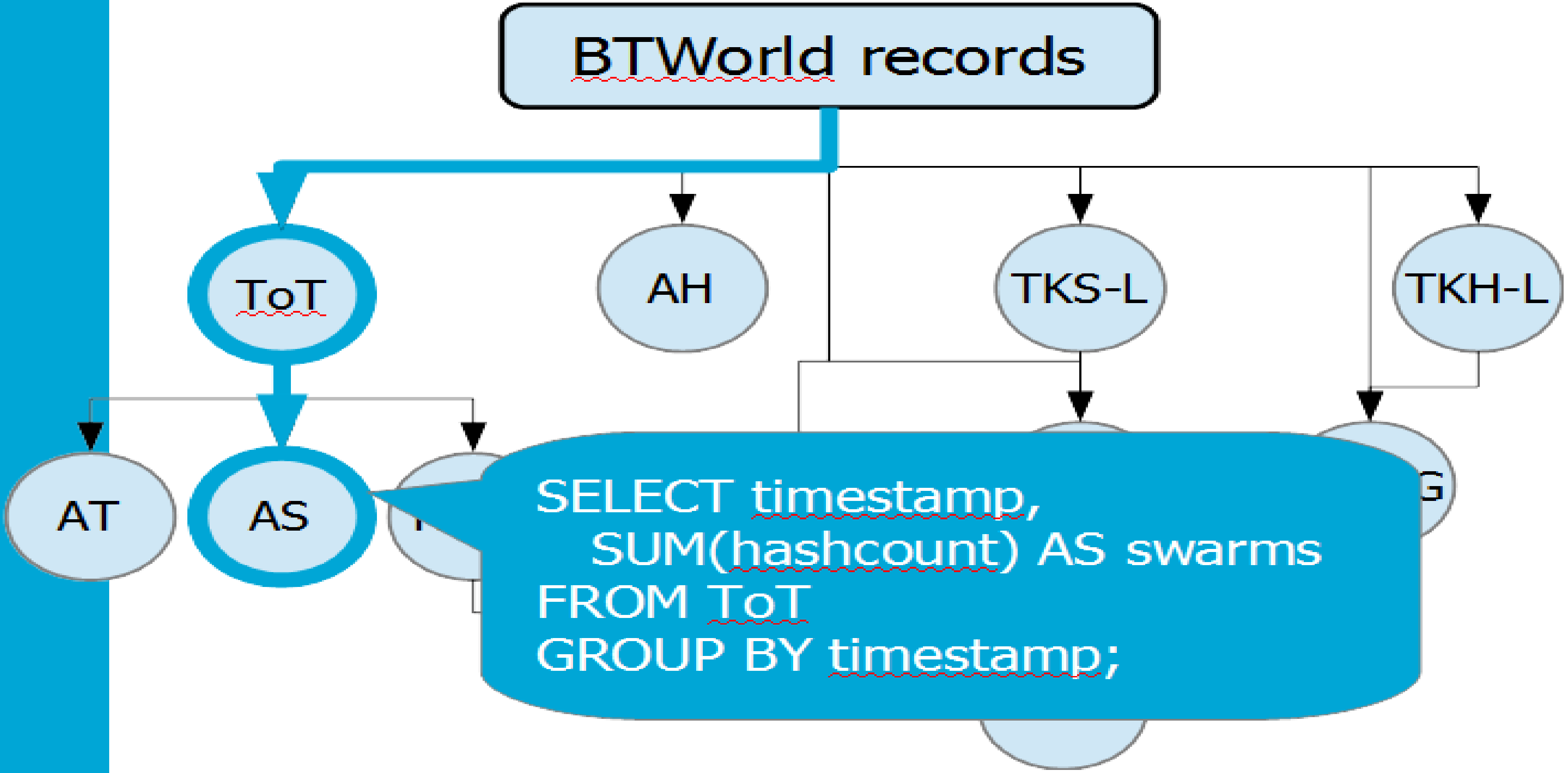


THE BTWORLD WORKLOAD



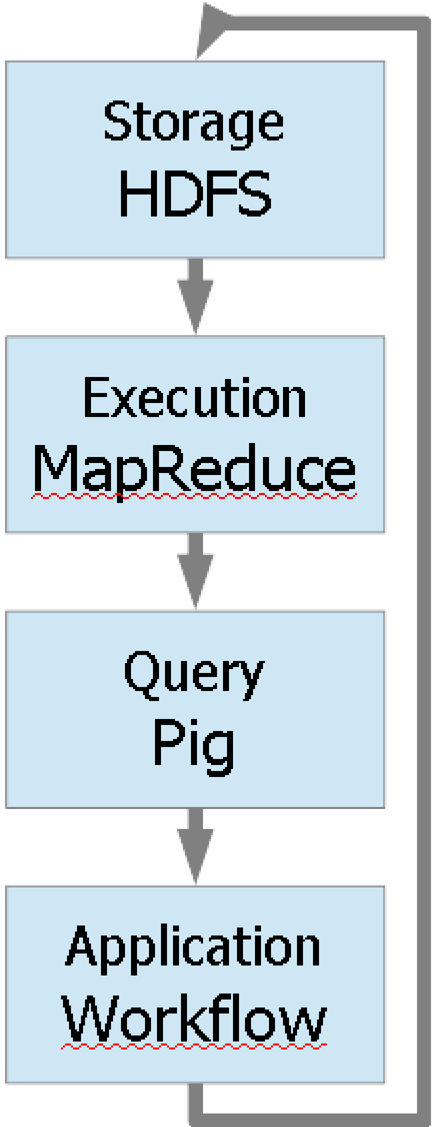
```
SELECT tracker, timestamp,
    COUNT(hash) AS hashcount,
    SUM(seeders + leechers) AS sessions,
    AVG(leechers == 0 ?
        seeders : seeders / leechers)
        AS slratio
FROM logs
GROUP BY tracker, timestamp;
```

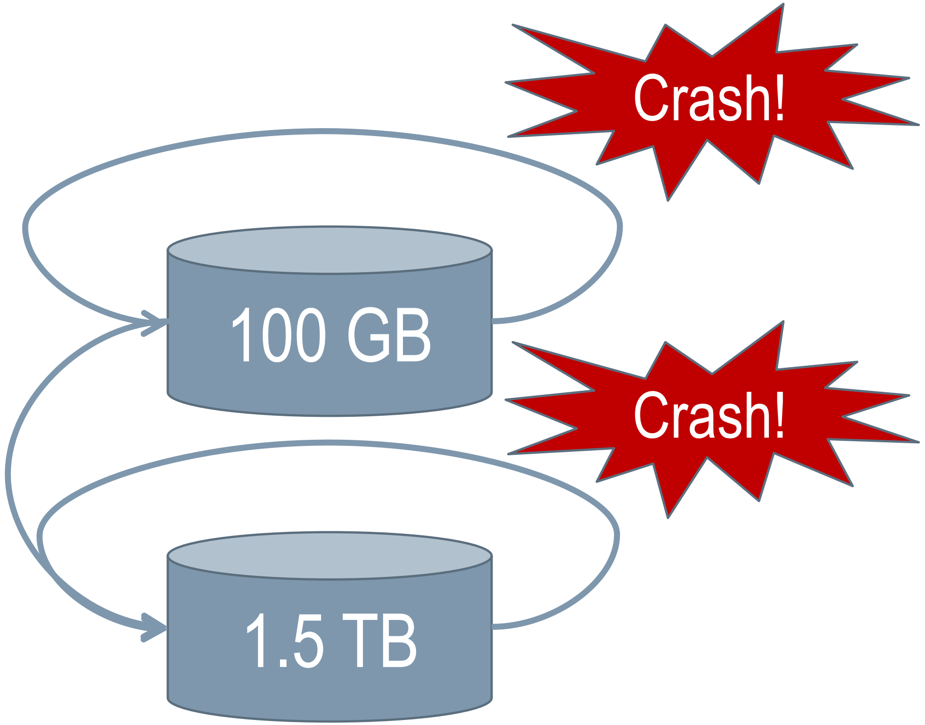
THE BTWORLD WORKLOAD



```
SELECT timestamp,
   SUM(hashcount) AS swarms
FROM ToT
GROUP BY timestamp;
```

OPTIMIZATION CYCLE

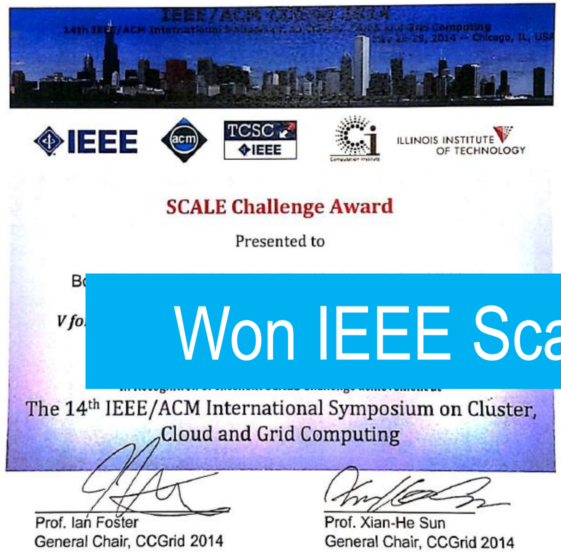




- HDFS: reduced replication, concatenate small files
- MapReduce: memory per task vs number of tasks, mappers then reducers
- Pig: specialized joins, multistage adaptive joins
- Workflow: reuse data between stages, common queries

GENERAL PROBLEM

Domain	Data Collection	Entities	Identifiers
BitTorrent	Trackers	Swarms	Hashes
Finance	Stock markets	Stock listings	Stocks
Tourism	Travel agents	Vacation packages	Venues



IEEE/ACM CCGrid 2014

14th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing

May 26-29, 2014 -- Chicago, IL, USA

IEEE acm TCSC IEEE Ci ILLINOIS INSTITUTE OF TECHNOLOGY

SCALE Challenge Award

Presented to

Bo

V fo

The 14th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing

Prof. Ian Foster
General Chair, CCGrid 2014

Prof. Xian-He Sun
General Chair, CCGrid 2014

Won IEEE Scale Challenge 2014!

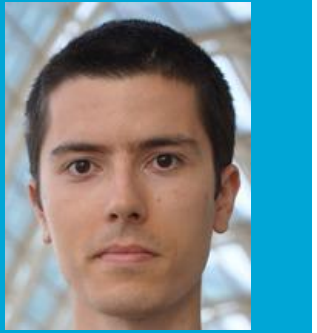
Agenda

1. The Golden Age of scalable computing
2. The core idea of cloud computing

3. Enabling technologies (homework)

4. The scheduling challenge
5. The Ecosystem Navigation challenge
6. The Big Cake challenge
7. Jevon's Effect challenge (IEEE Scale Challenge Award)

8. Take-home message



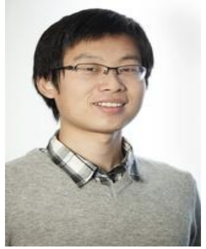
Take-Home Message

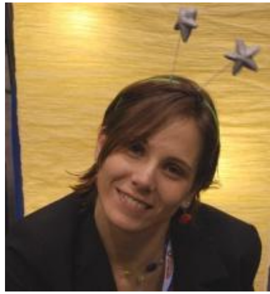
The Golden Age of scalable computing

Cloud computing + Big Data

Important New Challenges

1. *The scheduling challenge*
2. *The ecosystem navigation challenge*
3. *The big cake challenge*
4. *Jevon’s effect challenge*





Recommended Reading

Elastic Big Data and Computing

- B. Ghit, N. Yigitbasi (Intel Research Labs, Portland), A. Iosup, and D. Epema. Balanced Resource Allocations Across Multiple Dynamic MapReduce Clusters. SIGMETRICS 2014
- L. Fei, B. Ghit, A. Iosup, D. H. J. Epema: KOALA-C: A task allocator for integrated multicluster and multicloud environments. CLUSTER 2014: 57-65
- K. Deng, J. Song, K. Ren, A. Iosup: Exploring portfolio scheduling for long-term execution of scientific workloads in IaaS clouds. SC 2013: 55

Time-Based Analytics

- B. Ghit, M. Capota, T. Hegeman, J. Hidders, D. Epema, and A. Iosup. V for Vicissitude: The Challenge of Scaling Complex Big Data Workflows. Winners IEEE Scale Challenge 2014

Graph Processing

- Y. Guo, M. Biczak, A. L. Varbanescu, A. Iosup, C. Martella, T. L. Willke: How Well Do Graph-Processing Platforms Perform? An Empirical Performance Evaluation and Analysis. IPDPS 2014: 395-404
- A. L. Varbanescu, M. Verstraaten, C. de Laat, A. Penders, A. Iosup, H. J. Sips: Can Portability Improve Performance?: An Empirical Study of Parallel Graph Analytics. ICPE 2015: 277-287