Clouds and Big Data: Between Efficient Datacentres and Demanding Users

Alexandru Iosup
Delft University of Technology
The Netherlands

Team: **Undergrad** Tim Hegeman, ... **Grad** Yong Guo, Mihai Capota, Bogdan Ghit
Researchers Marcin Biczak, Otto Visser **Staff** Henk Sips, Dick Epema
Collaborators* Ana Lucia Varbanescu (UvA, Ams), Claudio Martella (VU, Giraph), KIT, Intel Research Labs, IBM TJ Watson, SAP, Google Inc. MV, Salesforce SF, ...

* Not their fault for any mistakes in this presentation. Or so they wish.

August 19, 2014

5th Cloud Control Workshop, Moelle, Sweden, Aug 2014

Delft University of Technology

The Parallel and Distributed Systems Group at TU Delft

VENI

Alexandru Iosup

Grids/Clouds
P2P systems
Big Data
Online gaming

Dick Epema

Grids/Clouds
P2P systems
Video-on-demand
e-Science

VENI

Ana Lucia Varbanescu (now UvA)

HPC systems
Multi-cores
Big Data
e-Science

Henk Sips

HPC systems
Multi-cores
P2P systems

VENI

Johan Pouwelse

P2P systems
File-sharing
Video-on-demand

Home page

- www.pds.ewi.tudelft.nl

Publications

- see PDS publication database at publications.st.ewi.tudelft.nl

tribler
The Koala Grid Scheduler
Open TTD @large
GrenchMark
The Grid Workloads Archive

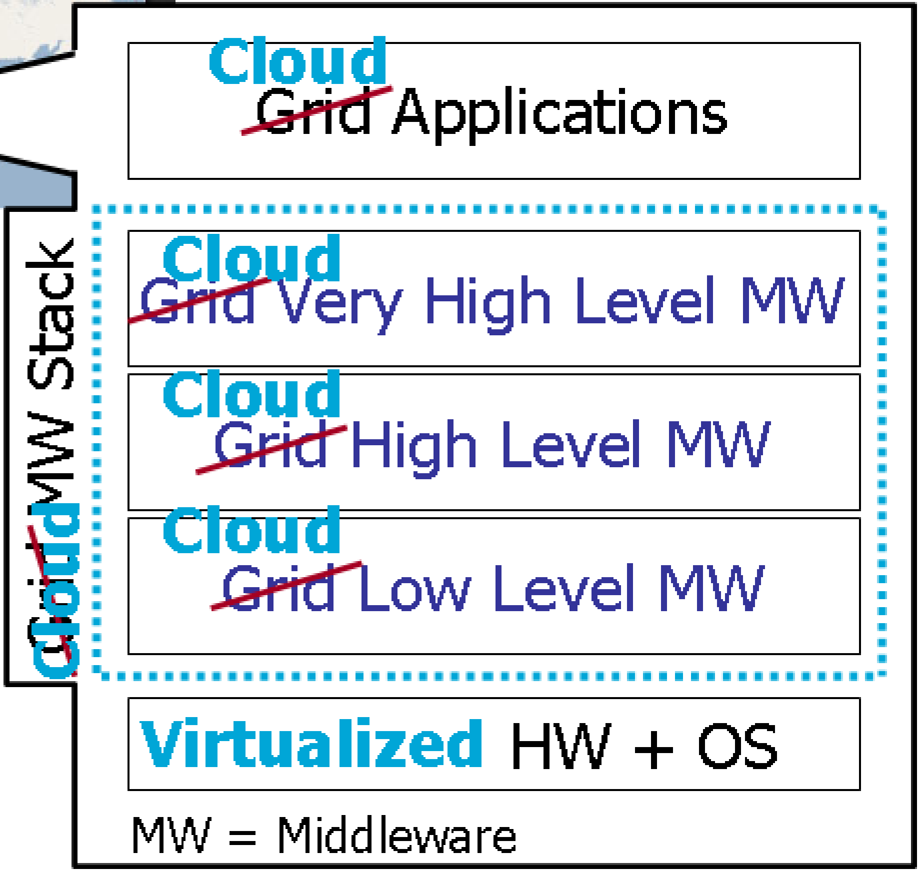
Winners IEEE TCSC Scale Challenge 2014

TU Delft

What is Cloud Computing? A Descendant* of the Grid Idea

Source: http://royal.pingdom.com/2008/04/11/map-of-all-google-data-center-locations/

* Subset.



"A computational grid is a hardware and software infrastructure that provides dependable, consistent, pervasive, and inexpensive access to high-end computational capabilities [+ for] nontrivial QoS." I. Foster, 1998 + 1999

TU Delft

Lessons From Grids, via a Detour

The Overwhelming Growth of Knowledge

“When 12 men founded the Royal Society in 1660, it was possible for an educated person to encompass all of scientific knowledge. In the last 50 years, such has been the pace of scientific advance that even the best scientists cannot keep up with discoveries at frontiers outside their own field.”
Tony Blair,
PM Speech, May 2002

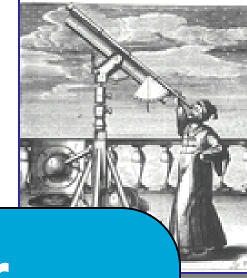
Professionals already know they don’t know [it all]

Number of Publications	1993 1997	1997 2001
[illegible]	[illegible],733	1,265,808
[illegible]	[illegible]730	1,347,985
[illegible]	[illegible]33	342,535
[illegible]	[illegible]93	318,286
[illegible]	[illegible]51	336,858
France	203,814	232,058
Canada	168,331	166,216
Italy	122,398	147,023
Switzerland	57,664	66,761
Netherlands	83,600	92,526

Data: King,The scientific impact of nations,Nature’04.

Lessons From Grids

From Hypothesis to Data

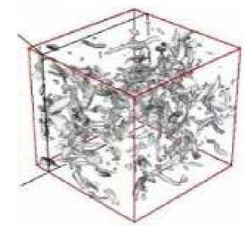


The Fourth Paradigm is suitable for professionals who already know they don't know [enough to formulate good hypotheses], yet need to deliver quickly

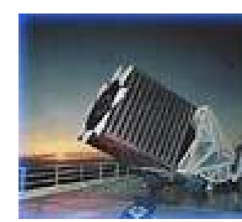
ena

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - K\frac{c^2}{a^2}$$

- Last few decades:
  a **computational** branch simulating complex phenomena



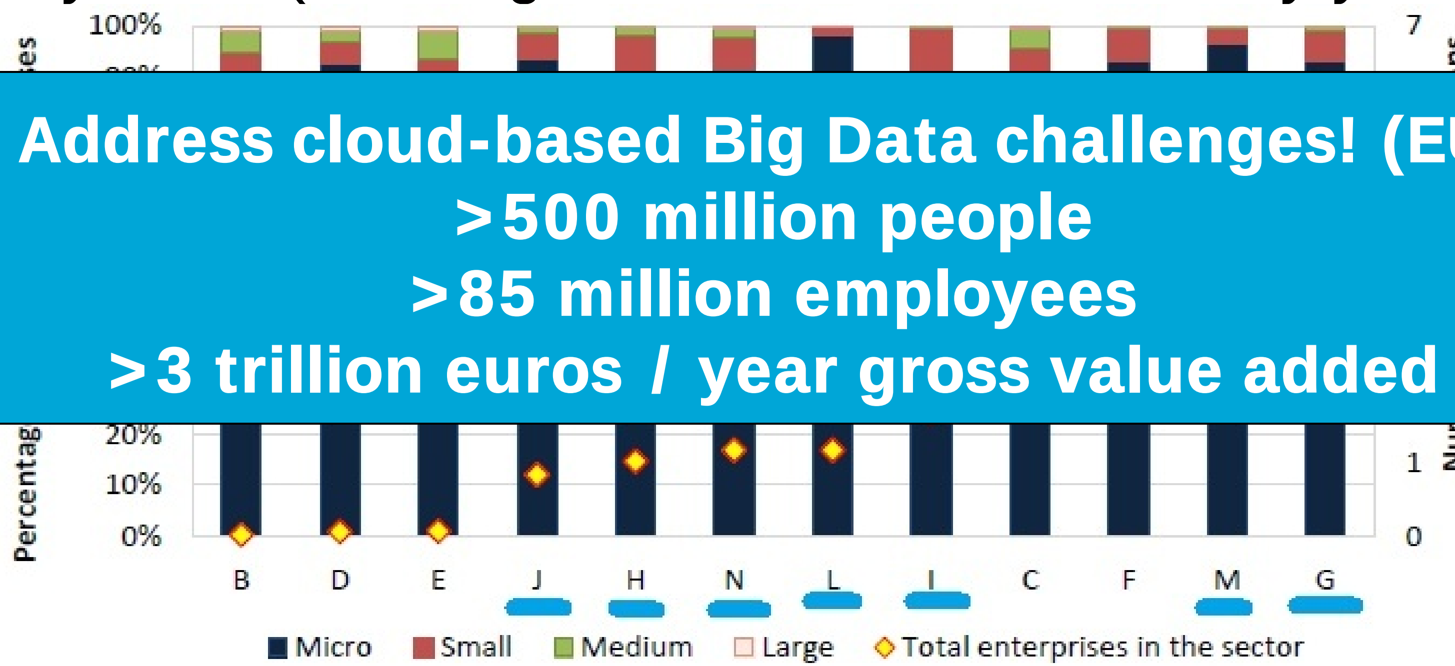
- Today (**the Fourth Paradigm**):
  data exploration
  unify theory, experiment, and simulation
  - Data captured by instruments or generated by simulator
  - Processed by software
  - Information/Knowledge stored in computer
  - Scientist analyzes results using data management and statistics



Source: Jim Gray and "The Fourth Paradigm" (Jan 2007 and, posthumously, 2011), http://research.microsoft.com/en-us/collaboration/fourthparadigm/

The Vision: Everyone Is a Scientist! (the Fourth Paradigm)

- Data as individual right, enabling high-quality **lifestyle of individuals** and modern **societal services**
- Data as workhorse in creating **commercial services** by SMEs (~60% gross value added, for many years)

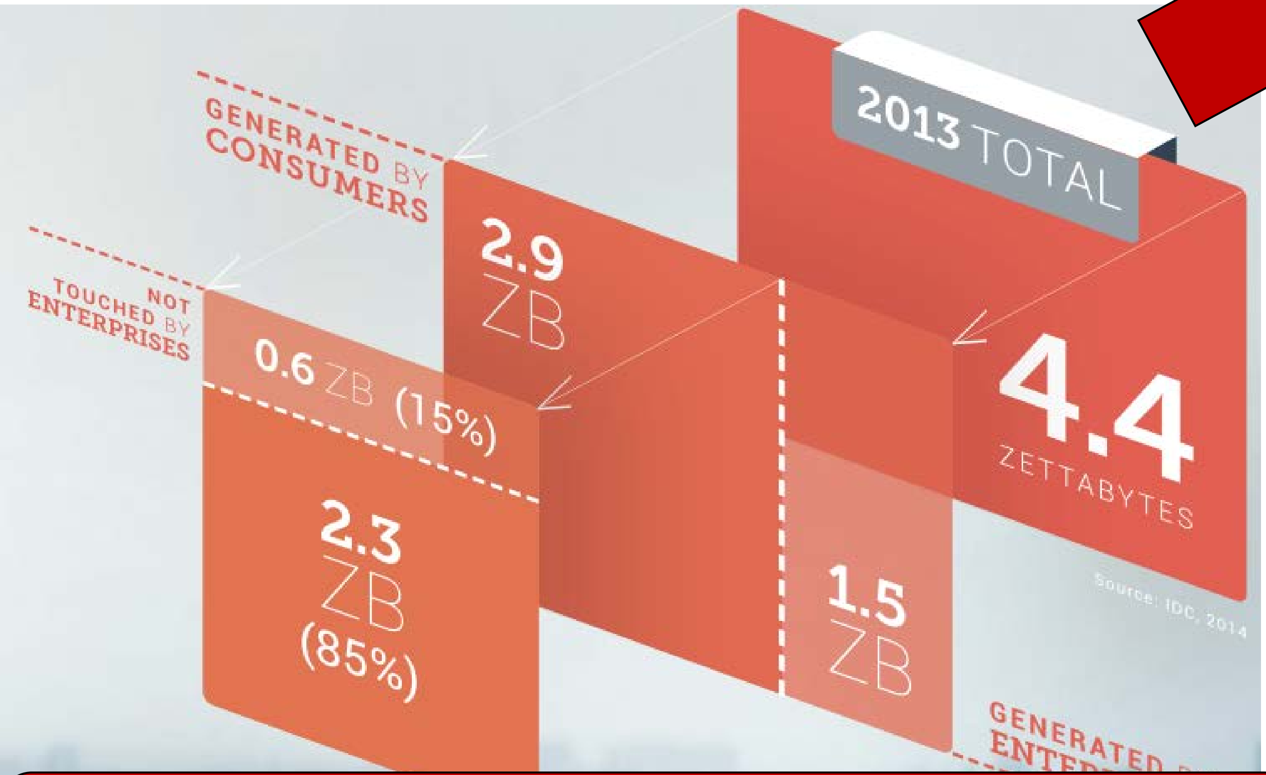


Address cloud-based Big Data challenges! (EU)
>500 million people
>85 million employees
>3 trillion euros / year gross value added

Sources: European Commission Annual Reports 2012 & 2013, ECORYS, Eurostat, National Statistical Offices, DIW, DIW econ, London Economics.

TUDelft

Can We Afford This Vision?
The "Data Deluge"



Data Deluge = data generated by humans and devices (IoT)

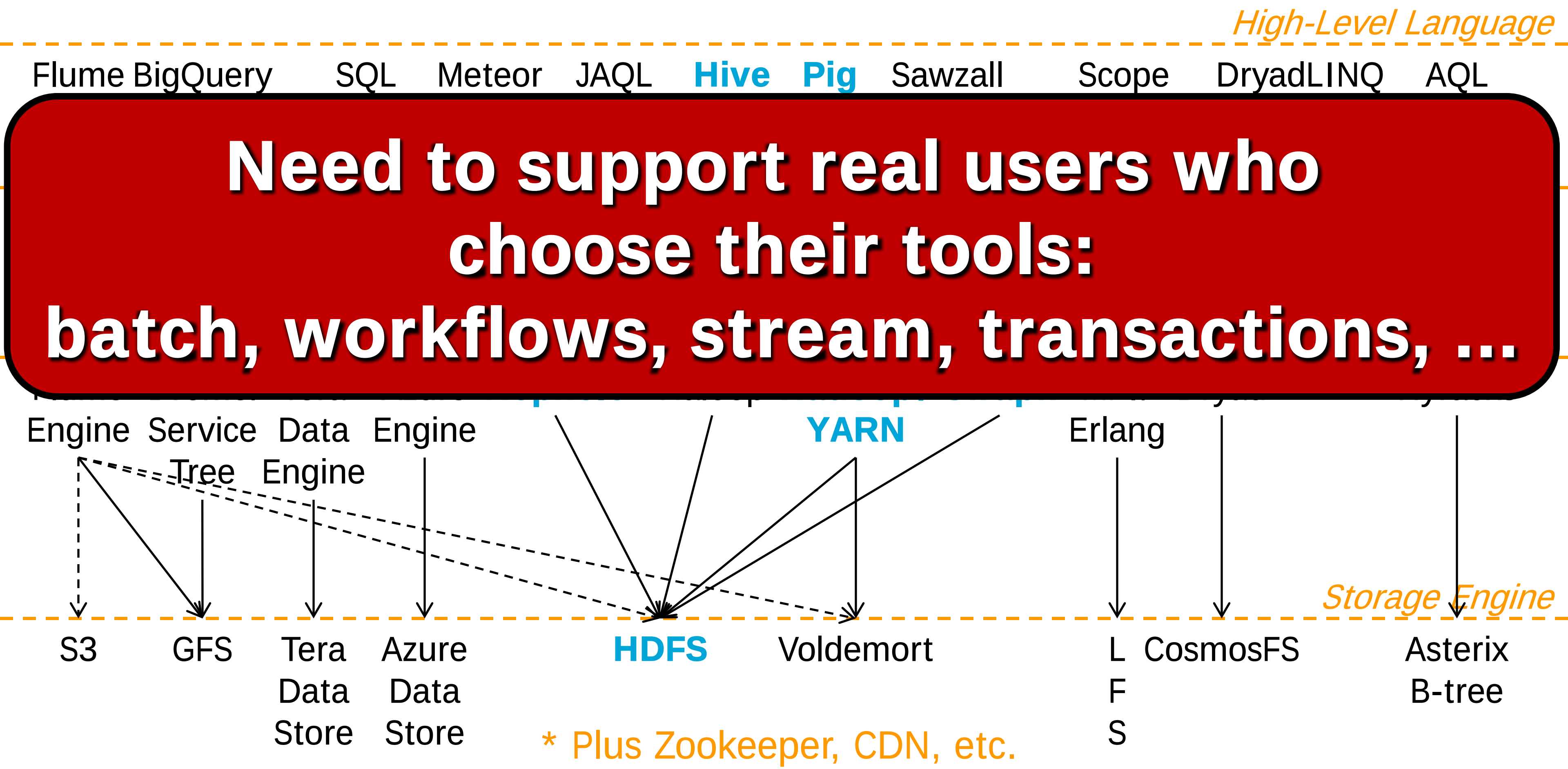
- Interacting
- Understanding
- Deciding
- Creating

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

Sources: IDC, EMC.

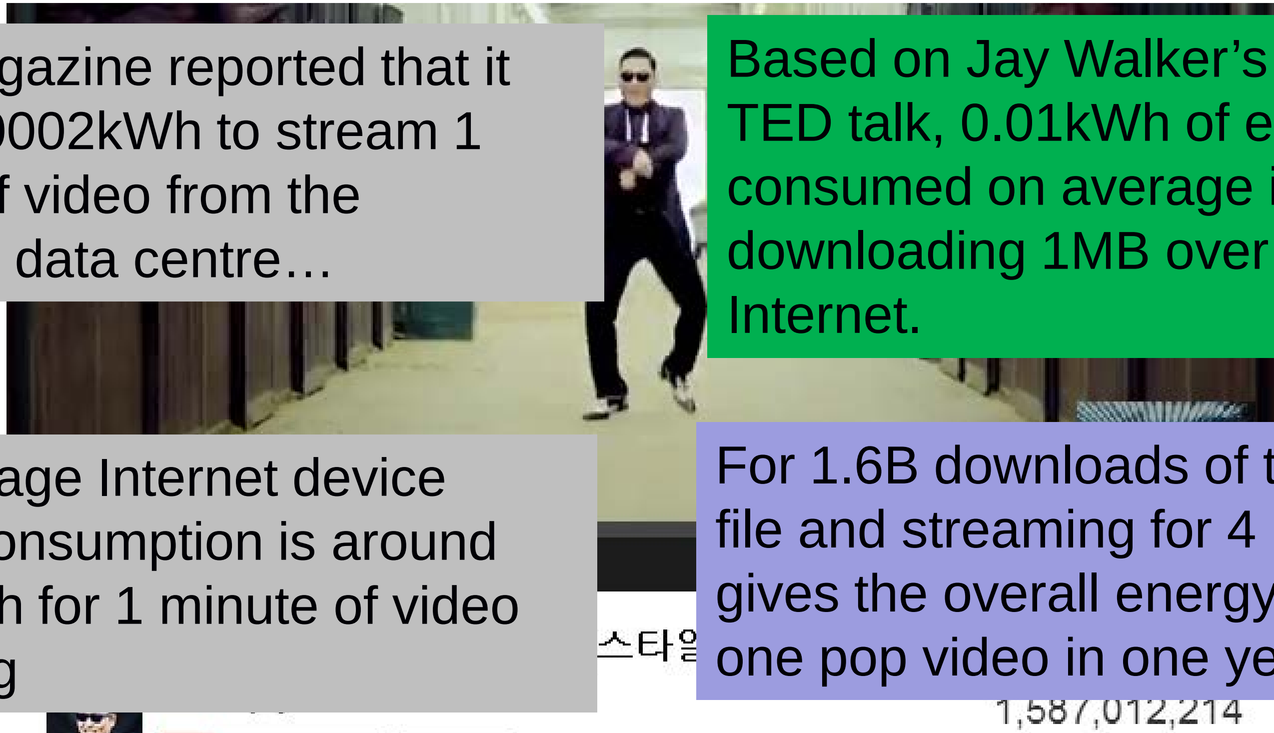
* New Vs later: ours is "vicissitude"

Can We Afford This Vision? The Current Tech
Big Data = Systems of Systems



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

The Challenge: Can We Afford This Vision? Not with the Current Resources (An Anecdote)



Time magazine reported that it takes 0.0002kWh to stream 1 minute of video from the YouTube data centre…

Based on Jay Walker's recent TED talk, 0.01kWh of energy is consumed on average in downloading 1MB over the Internet.

The average Internet device energy consumption is around 0.001kWh for 1 minute of video streaming

For 1.6B downloads of this 17MB file and streaming for 4 minutes gives the overall energy for this one pop video in one year…

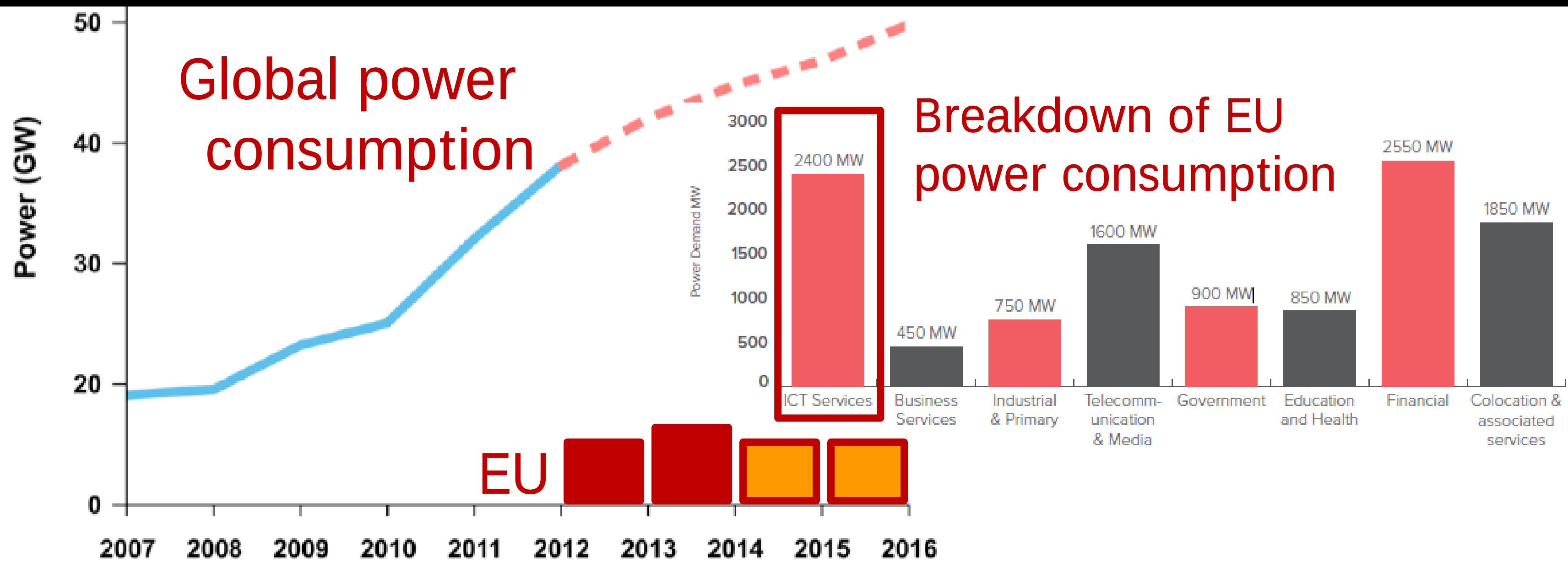
>300GWh = more than some countries in a year, >35MW of 24/7/365 diesel, >100M liters of oil, 80,000 cars running for a year, ...

Source: Ian Bitterlin and Jon Summers, UoL, UK, Jul 2013.
Note: Psy has now >2.75 billion views, so roughly 450GWh (Jun 2014).

Can We Afford This Vision? Not with the Current Resources

- Energy resources

Need efficient datacentres



Data Source: Powering the Datacenter, DatacenterDynamics, 2013

One-third of global data center energy use is in U.S., but growth rates are fastest in emerging economies.

Sources: DatacenterDynamics and Jon Summers, UoL, UK.

Scheduling in IaaS Clouds
An Overview

Cloud operator:

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

Need usage and user-aware scheduling policies

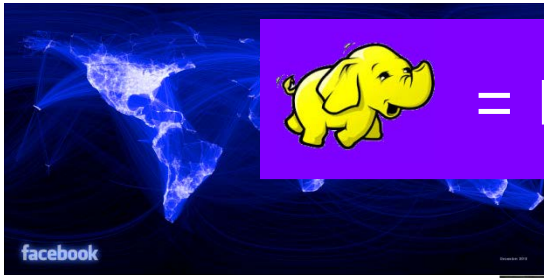
Cloud customer:

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

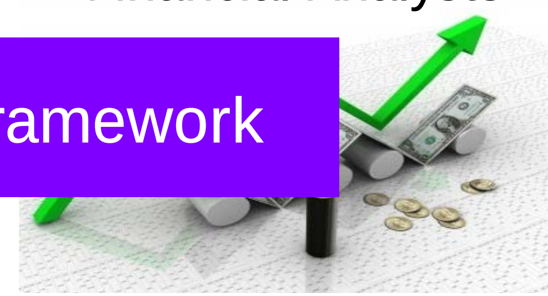
Delft University of Technology

The “Big Data cake” in the Data Center

Online Social Networks



Financial Analysts

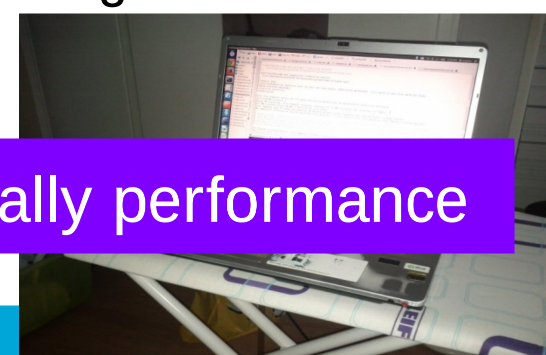


= Hadoop / MapReduce framework

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

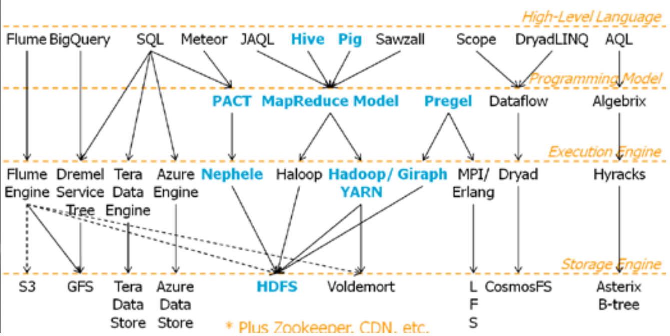
Universe Explorers

Big Data Enthusiast



Multiple frameworks = Isolation, especially performance

Everyone is a Scientist!
Can We Afford This Vision?

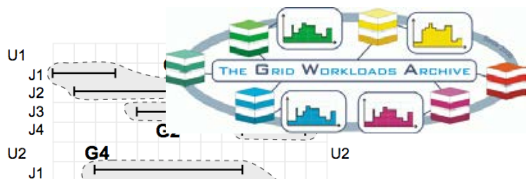


We need to build a cloud ecosystem that is very efficient, very user-friendly. For this, we need to combine sw.eng., distr.sys., parallel sys., DB, …

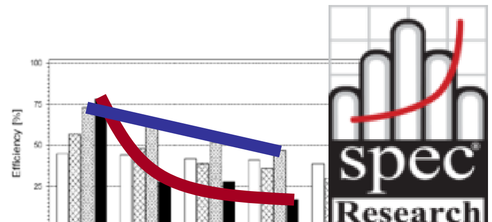
Agenda

Everyone is a Scientist

Can we afford it?



Workloads



Benchmarking

Scheduling

Graph Analytics

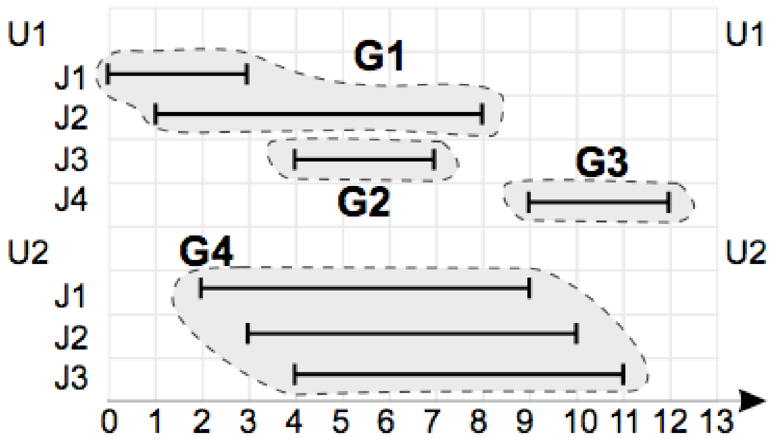
BTWorld

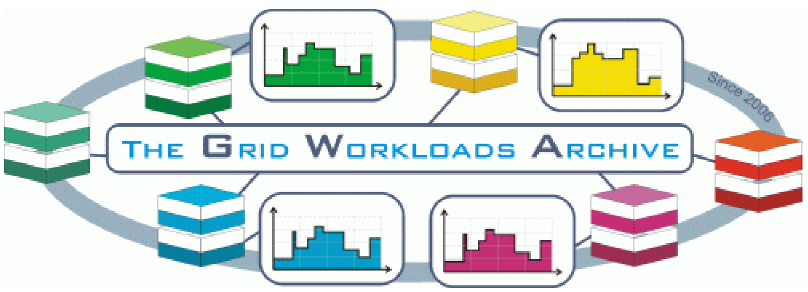
Elastic MR

Conclusion

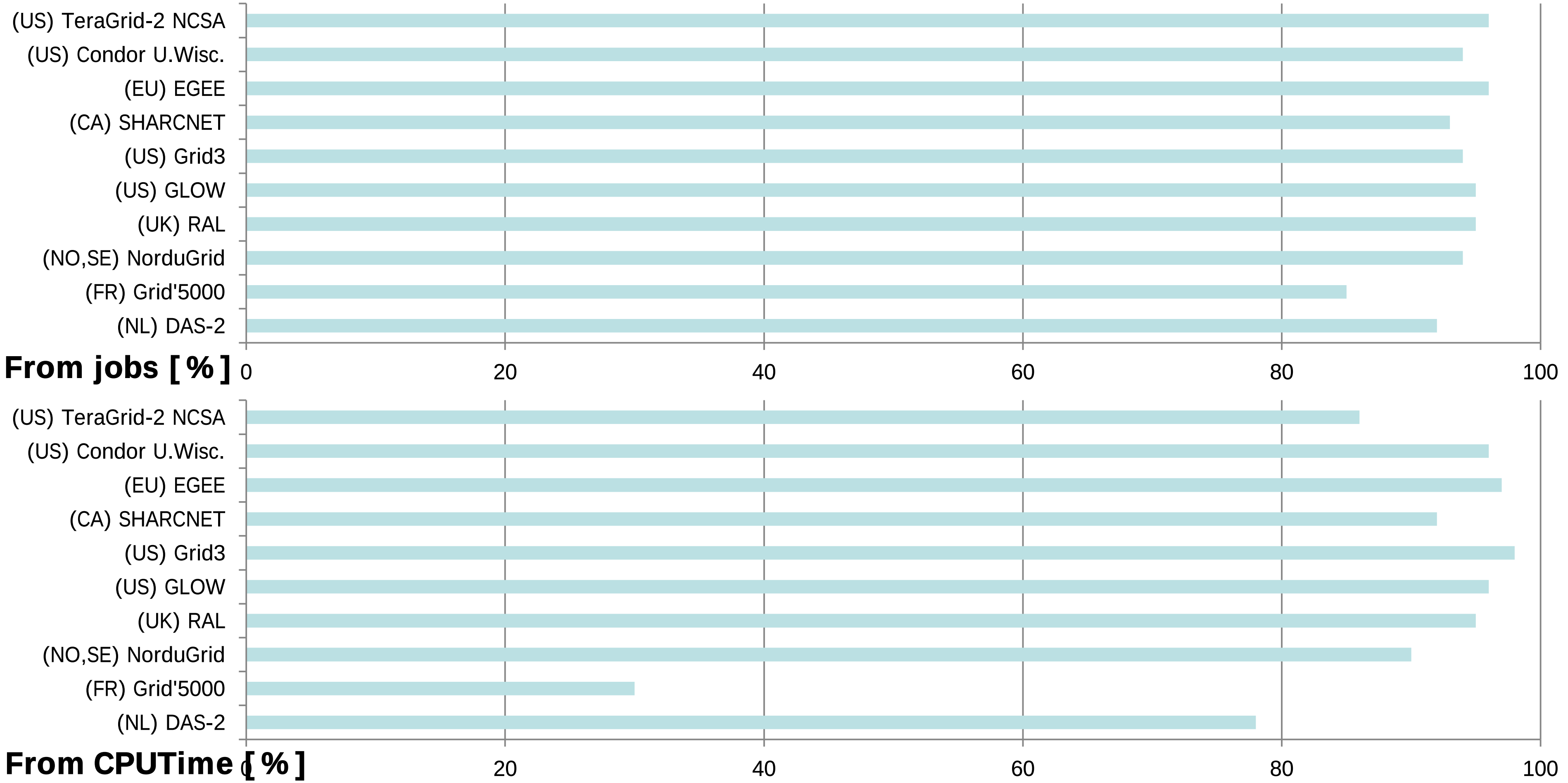
TU Delft

Workloads


U1
G1
J1
J2
J3
J4
G3
G2
U2
G4
J1
J2
J3
0 1 2 3 4 5 6 7 8 9 10 11 12 13


The Grid Workloads Archive
Since 2006

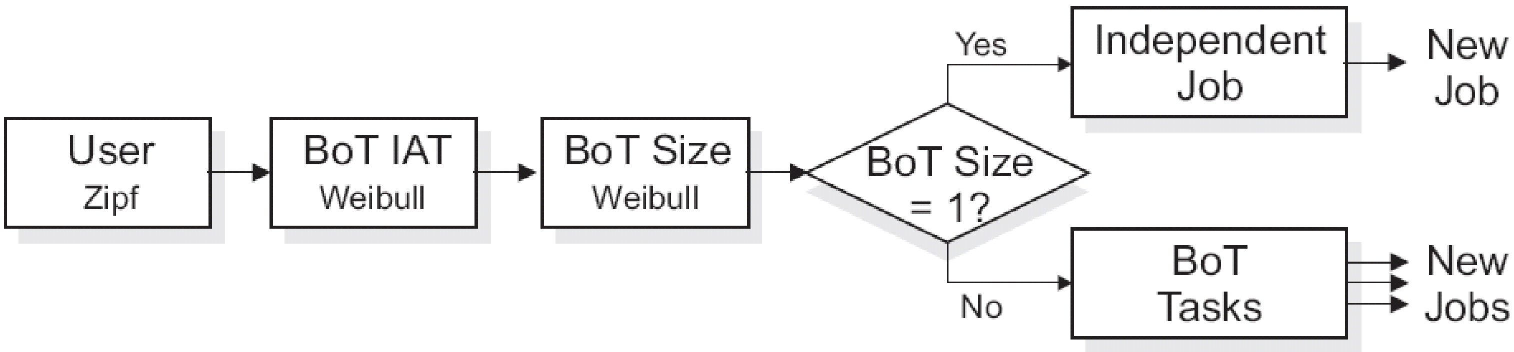
BoTs Are the Dominant Programming Model for Grid Computing (Many Tasks)



Iosup and Epema: Grid Computing Workloads. IEEE Internet Computing 15(2): 19-26 (2011)

Q0

Statistical BoT Workload Model



- Single arrival process for both BoTs and parallel jobs
- Validated with 7 grid workloads

A. Iosup, O. Sonmez, S. Anoep, and D.H.J. Epema. The Performance of Bags-of-Tasks in Large-Scale Distributed Systems, HPDC, pp. 97-108, 2008.

Workflows Exist in Grids, but Did Not Find Evidence of a Dominant Programming Model

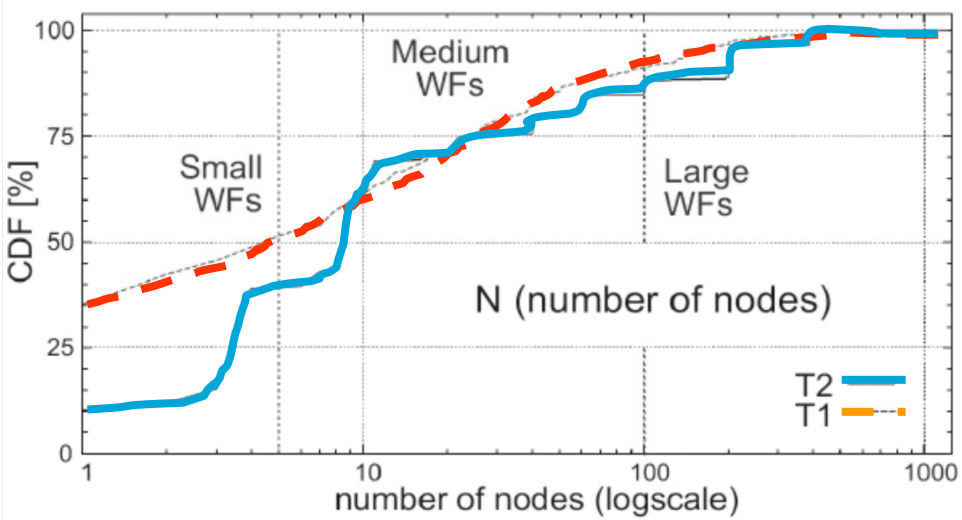
- Traces

Trace	Source	Duration	Number of WFs	Number of Tasks	CPUdays
T1	DEE	09/06-10/07	4,113	122k	152
T2	EE2	05/07-11/07	1,030	46k	41

- Selected Findings

Loose coupling

- Graph with 3-4 levels
- Average WF ~10s of jobs
- 75% WFs are <=40 jobs
  95% are <=200 jobs
- 85% WFs take <10 mins

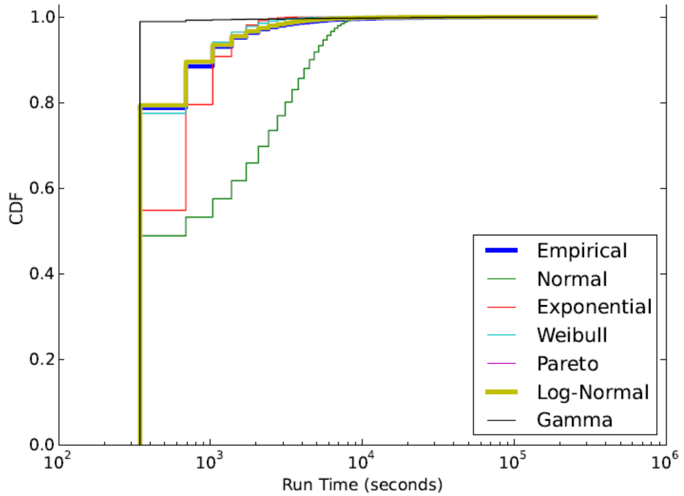


Ostermann et al., On the Characteristics of Grid Workflows, CoreGRID Integrated Research in Grid Computing (CGIW), 2008.

Delft

Statistical MapReduce Models From Long-Term Usage Traces

- Started 2010,
  excellent studies now exist
- Real traces
  - Yahoo
  - Google
  - 2 x Social Network Provider
  - (currently looking at 2 SME traces)



Model	Tasks	Correlation	Map/Reduce Modeled	Sign. Level	Indirect Distr. Sel.
Complex Model	Indirect	Run time – Disk	Separately	0.05	Best fits
Relaxed Complex Model	Indirect	Run time – Disk	Separately	0.02	All fits
Safe Complex Model	Direct	Run time – Disk	Separately	0.05	–
Simple Model	Direct	–	Together	0.05	–

de Ruiter and Iosup. A workload model for MapReduce. MSc thesis at TU Delft. Jun 2012. Available online via TU Delft Library, http://library.tudelft.nl .

Survey of *Used* Graph-Processing Algorithms

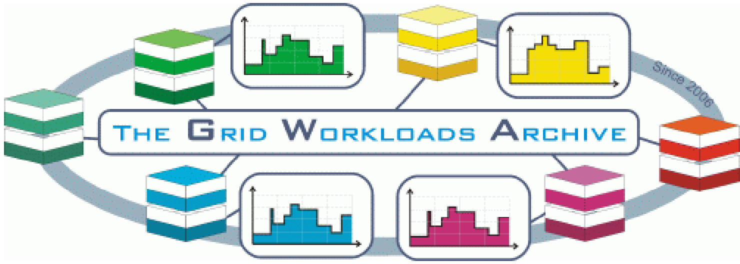
- Literature survey of of metrics, datasets, and algorithms
  - 10 top research conferences: SIGMOD, VLDB, HPDC ...
  - Key word: graph processing, social network
  - 2009–2013, 124 articles

Class	Examples	%
Graph Statistics	Diameter, PageRank	16.1
Graph Traversal	BFS, SSSP, DFS	46.3
Connected Component	Reachability, BiCC	13.4
Community Detection	Clustering, Nearest Neighbor	5.4
Graph Evolution	Forest Fire Model, PAM	4.0
Other	Sampling, Partitioning	14.8

Y. Guo, M. Biczak, A. L. Varbanescu, A. Iosup, C. Martella, and T. L. Willke. How Well do Graph-Processing Platforms Perform? An Empirical Performance Evaluation and Analysis, IPDPS'14.

Take-Home Message

- Data available, some trace archives



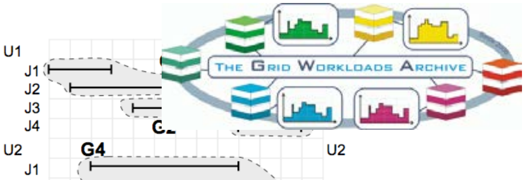
- Compute-intensive workloads
  - Bags of Tasks
  - Workflows
- Data-intensive workloads
  - Still much to do to understand
  - Survey of graph analytics algorithms
  - MapReduce workflow for time-based analytics

TU Delft

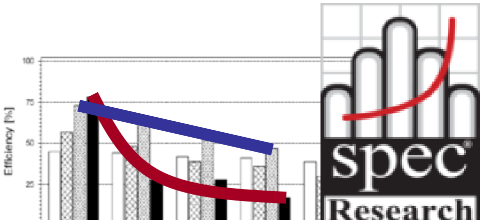
Agenda

Everyone is a Scientist

Can we afford it?



Workloads



Benchmarking

Scheduling

Graph Analytics

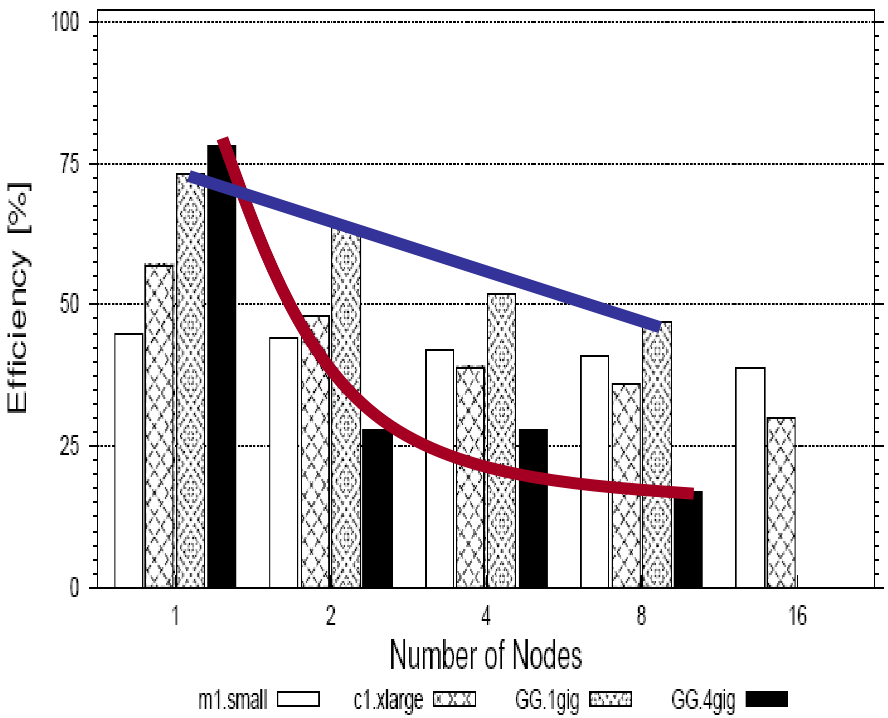
BTWorld

Elastic MR

Conclusion

TU Delft

Benchmarking



spec Research

TU Delft

SPEC Research Group (RG)

The Research Group of the
Standard Performance Evaluation Corporation

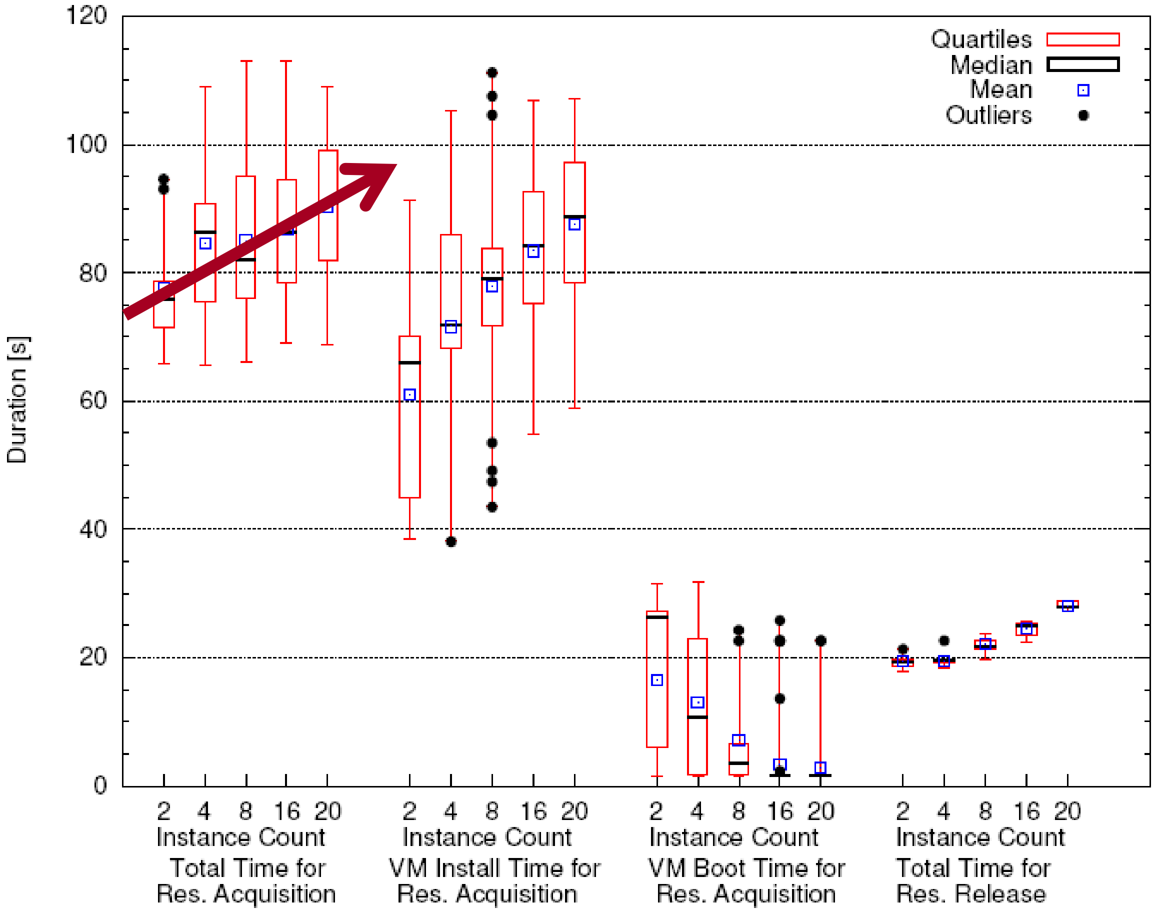
Mission Statement

- Provide a platform for collaborative research efforts in the areas of computer benchmarking and quantitative system analysis
- Provide metrics, tools and benchmarks for evaluating early prototypes and research results as well as full-blown implementations
- Foster interactions and collaborations btw. industry and academia

Join us! *http://research.spec.org*

Delft University of Technology

Multi-Resource Provisioning Time Can Be High

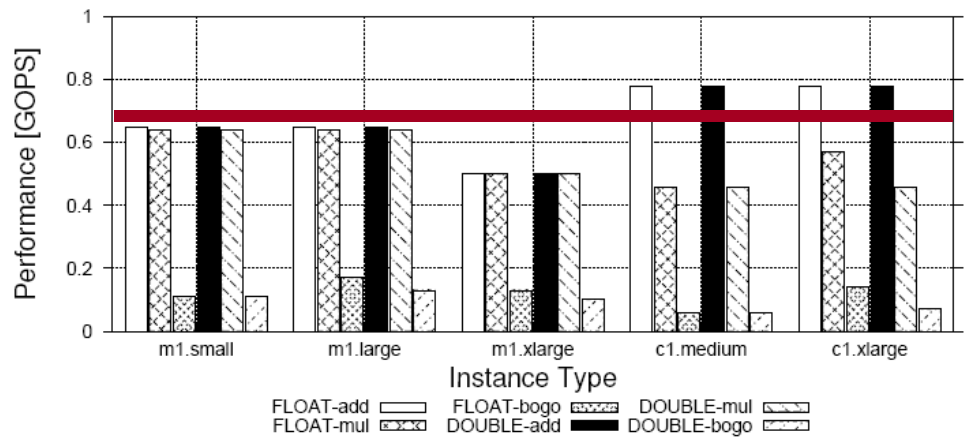


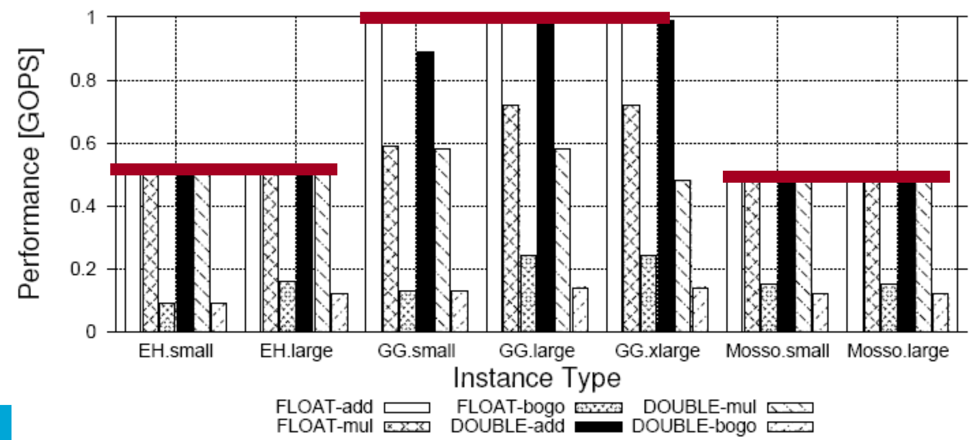
- Time for *multi*-resource increases with number of resources

Iosup et al., Performance Analysis of Cloud Computing Services for Many Tasks Scientific Computing, (IEEE TPDS 2011).

Performance of Cloud Services Can Be Low

- ECU definition: “a 1.1 GHz 2007 Opteron” ~ 4 flops per cycle at full pipeline, which means at peak performance one ECU equals 4.4 gigaflops per second (GFLOPS)
- Real performance 0.6..0.1 GFLOPS = ~1/4..1/7 theoretical peak
- Parallel performance low

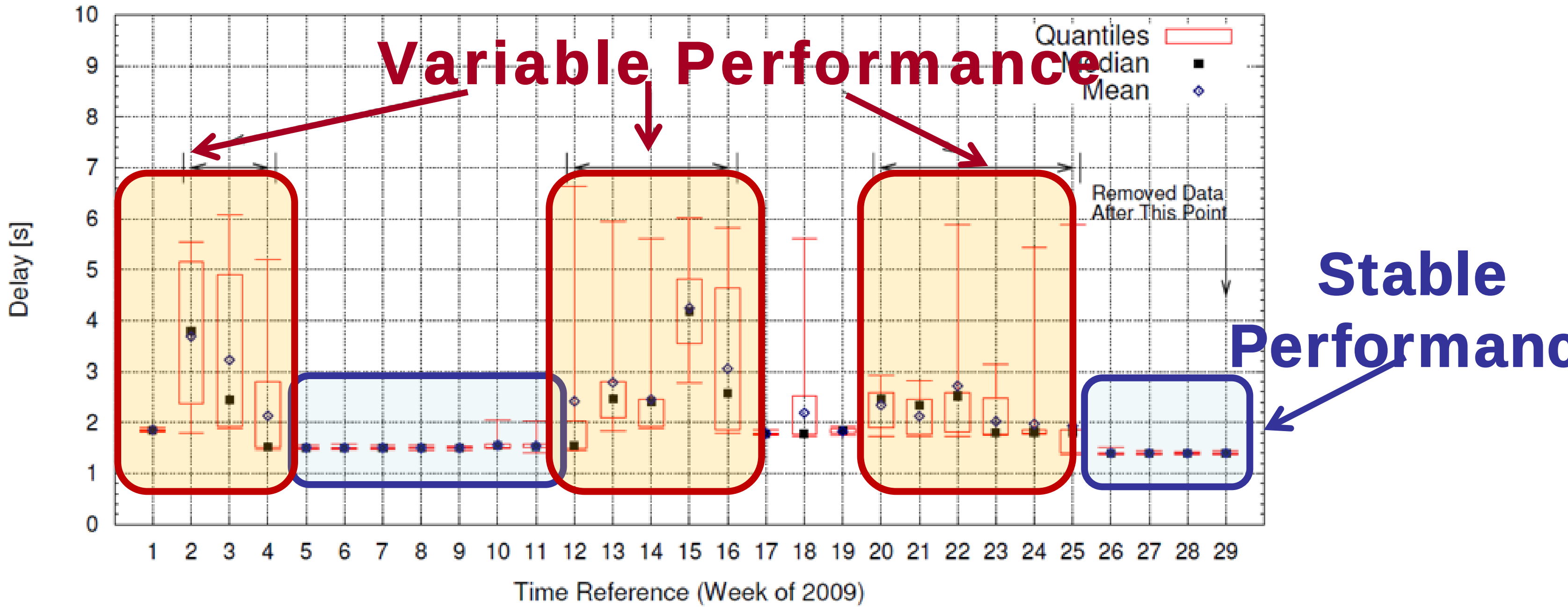




Iosup et al., Performance Analysis of Cloud Computing Services for Many Tasks Scientific Computing, (IEEE TPDS 2011).

Performance of Production Cloud Services Can Vary Short- and Long-Term

Q2



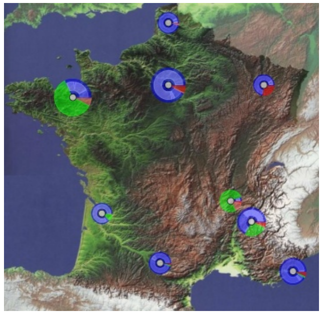
- **Average Lag Time [s]:** Time it takes for a posted message to become available to read. Average over multiple queues.
- Long periods of stability (low IQR and range)
- Periods of high performance variability also exist

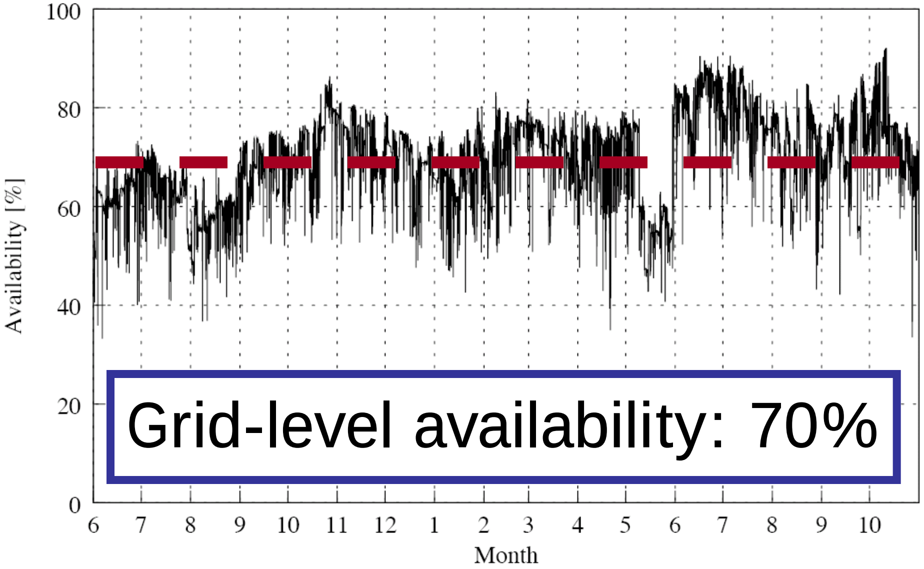
August 19, 2014

Iosup, Yigitbasi, Epema. On the Performance Variability of Production Cloud Services, (IEEE CCgrid 2011).

Resource Availability in Multi-Clusters is a Challenge

- Environment: Grid'5000 traces
  - jobs 05/2004-11/2006 (30 mo., 950K jobs)
  - resource availability traces 05/2005-11/2006 (18 mo., 600K events)
- **Resource availability model for multi-cluster grids**





- **Correlated failures time-correlated, space-correlated**

Iosup, Jan, Sonmez, and Epema, On the Dynamic Resource Availability in Grids, Grid 2007.
Yigitbasi, Gallet, Kondo, Iosup, Epema: Analysis and modeling of time-correlated failures in large-scale distributed systems. GRID 2010: 65-72

Take-Home Message

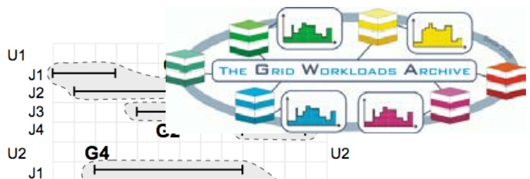
- Towards Self-Benchmarking Systems...
- Performance evaluation is difficult in clouds
- Reveals interesting patterns of operation
  - Multi-resource performance issues
  - Peak-performance issues
  - Variability in performance, perhaps due to multi-tenancy
  - High availability issues, correlated failures, etc.

- **Join SPEC Research!**

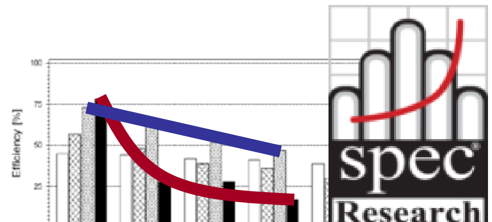
Agenda

Everyone is a Scientist

Can we afford it?



Workloads



Benchmarking

Scheduling

Graph Analytics

BTWorld

Elastic MR

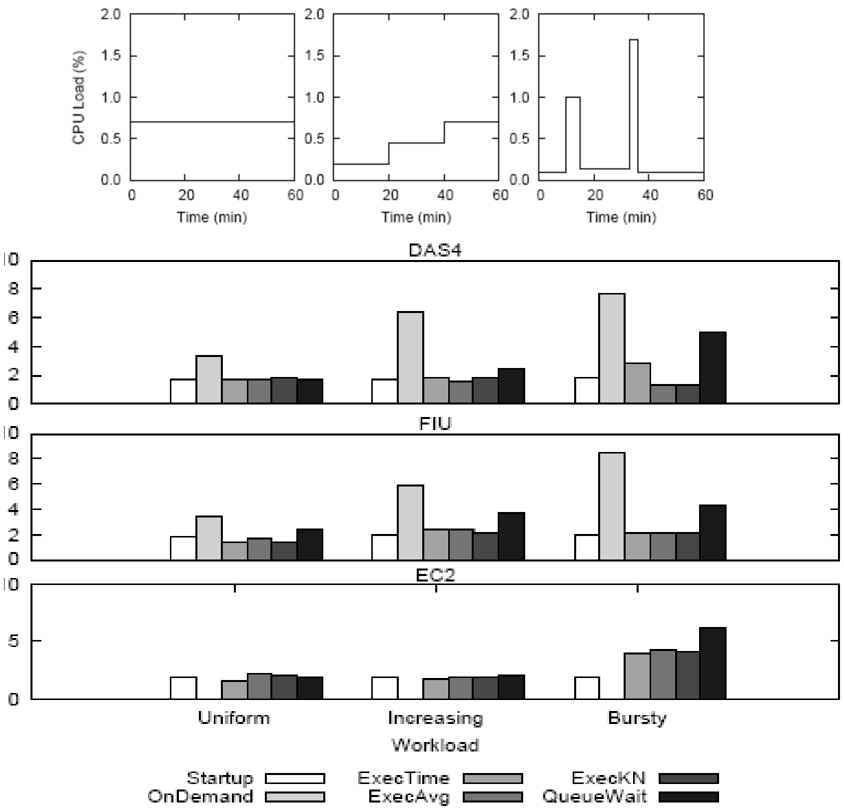
Conclusion

TU Delft

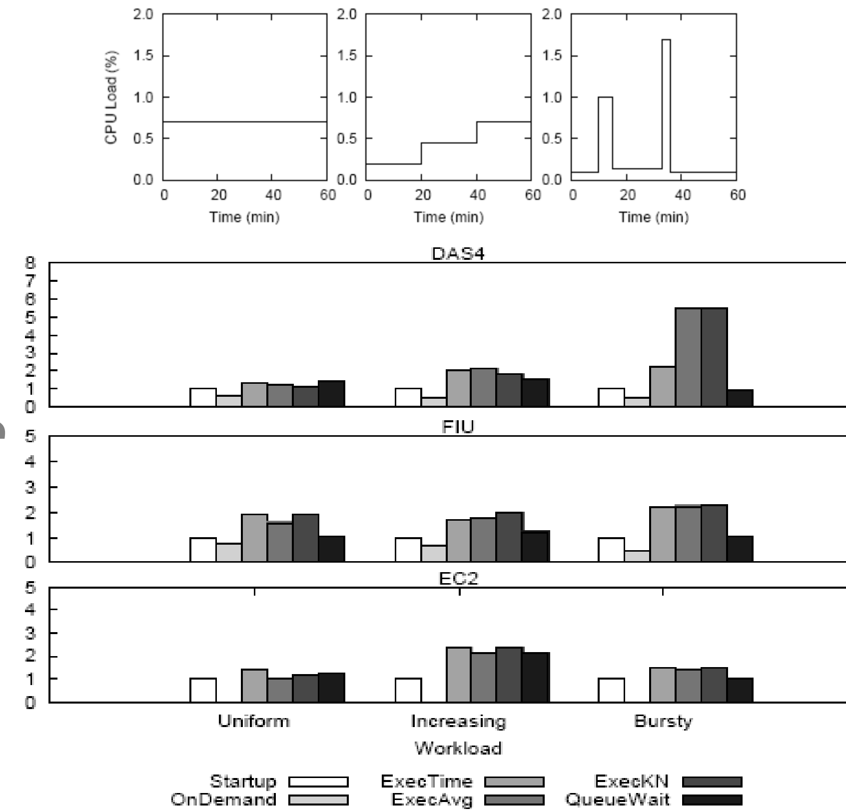
Scheduling

Trade-offs in Scheduling Policies for IaaS Clouds

Cost Efficiency



Utility



- Trade-off Utility-Cost needs further investigation
- Performance or Cost, not both:
  the policies we have studied improve one, but not both

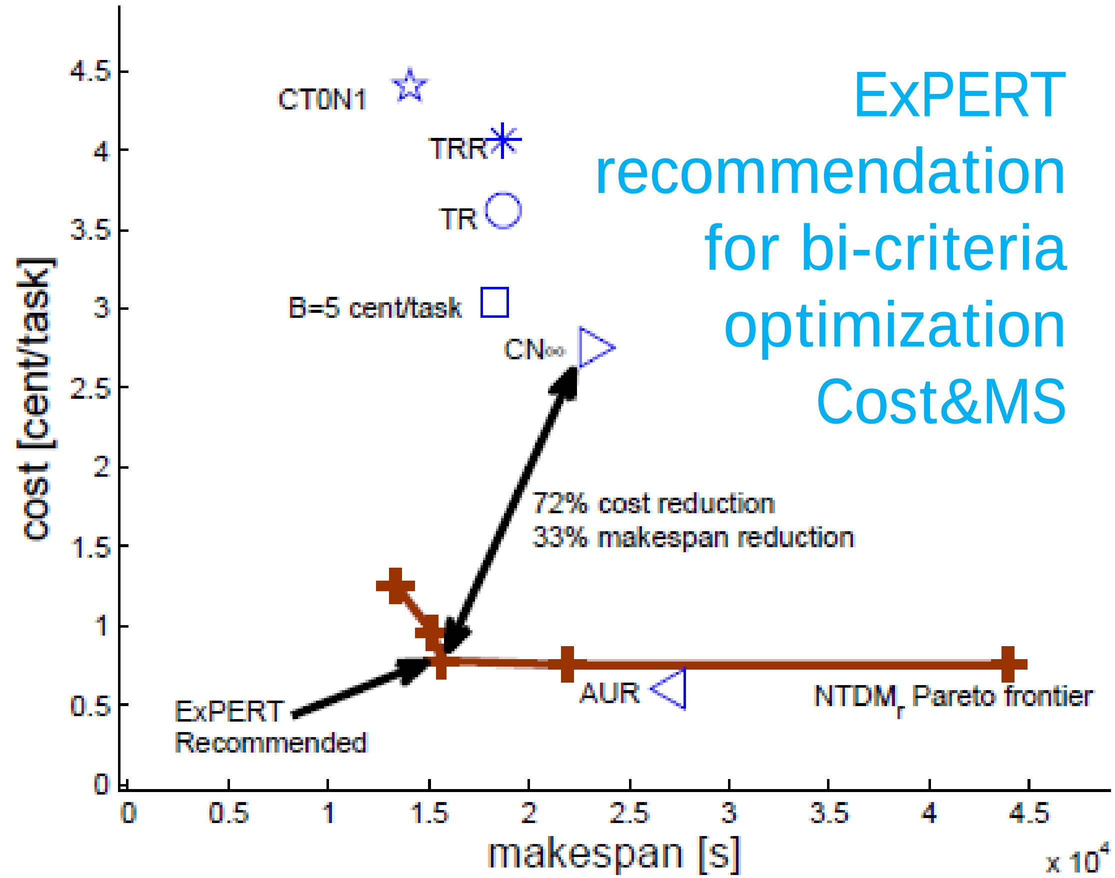
Villegas, Antoniou, Sadjadi, Iosup. An Analysis of Provisioning and Allocation Policies for Infrastructure-as-a-Service Clouds, CCGrid 2012

ExPERT

- D—task instance deadline
- T—when to replicate?
- N—how many times to replicate on unreliable?
- Nr—max ratio reliable:unreliable

Vs Common Policies

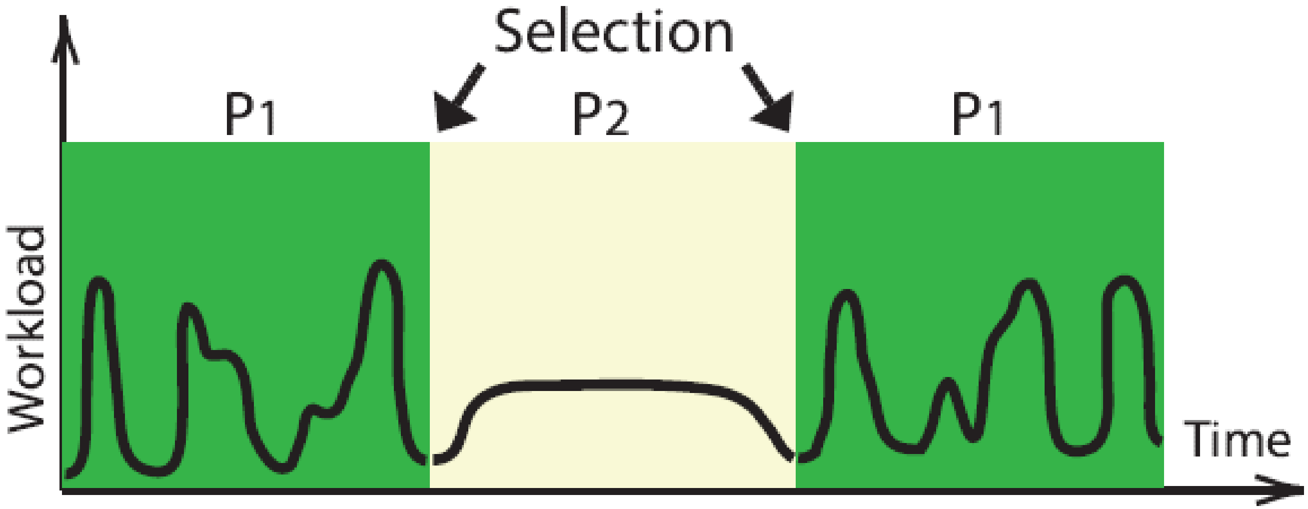
- AR—all to reliable
- AUR—all to unreliable, no replication
- TRR—Tail Replicate immediately to Reliable (N=0,T=0)
- TR—Tail to Reliable (N=0,T=D)
- CNinf—combine resources, no replication
- CT0N1—combine resources, replicate immediately at tail, N=1
- B=*cents/task—budget



ExPERT recommendation for bi-criteria optimization Cost&MS

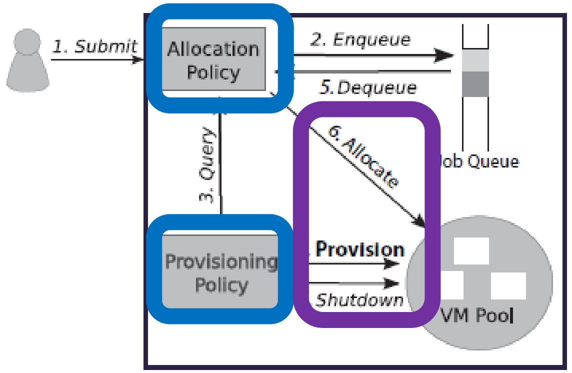
Agmon Ben-Yehuda, Schuster, Sharov, Silberstein, Iosup. ExPERT: pareto-efficient task replication on grids and a cloud. IPDPS’12.

What is Portfolio Scheduling? In a Nutshell, for Data Centers



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

Portfolio Scheduling
The Process



Which policies to include?

Creation

Which policy to activate?

Selection

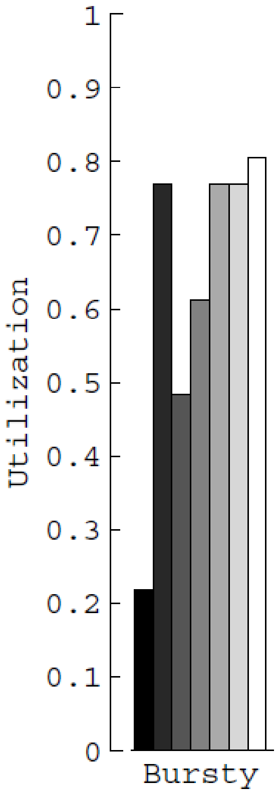
Application

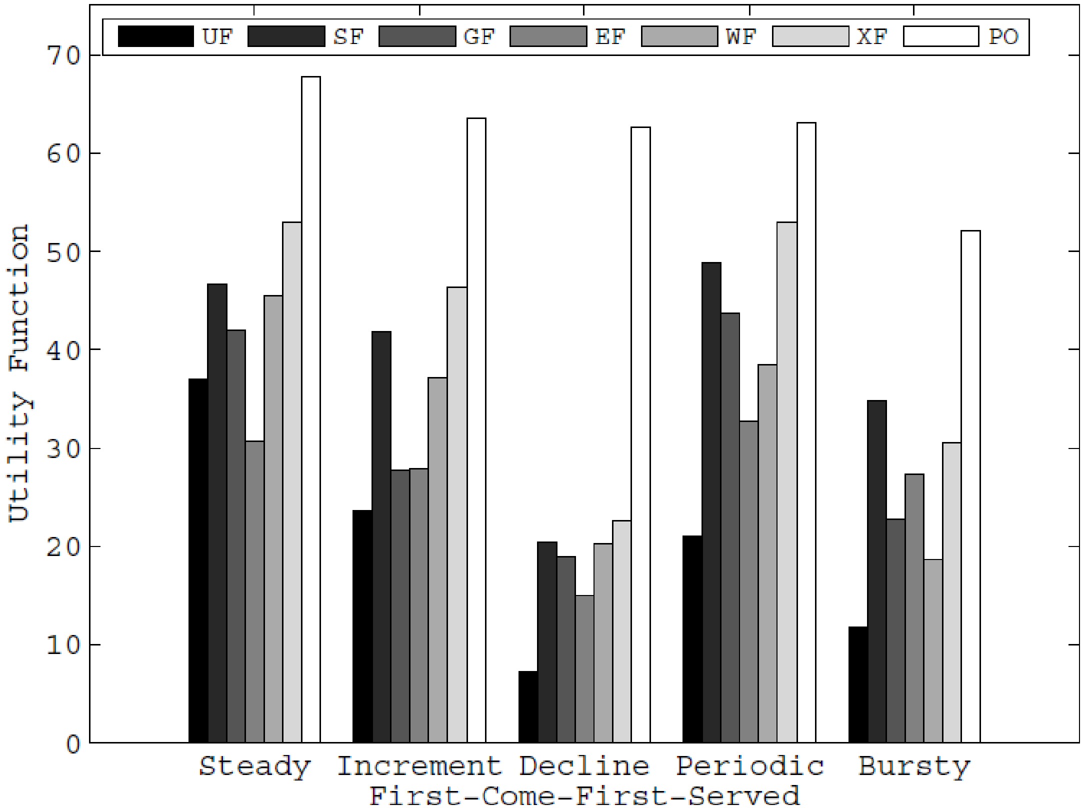
Which resources? What to log?

Reflection

Which changes to the portfolio?

Experimental Results, Synthetic Workloads
Resource Utilization + Workload Utility





- POrtfolio leads to high utilization
- Start-Up leads to poor utilization

- POrtfolio leads to better utility
- Start-Up leads to poor utility

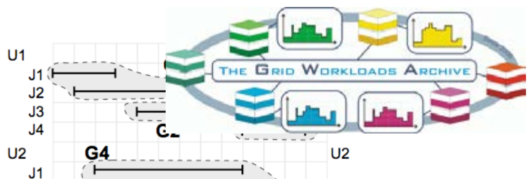
Deng, Song, Ren, Iosup: Exploring portfolio scheduling for long-term execution of scientific workloads in IaaS clouds. SC 2013: 55

TU Delft

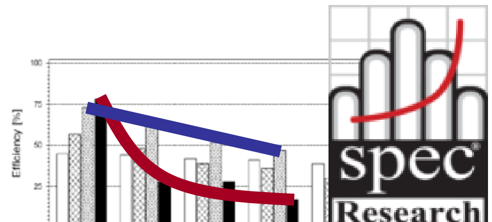
Agenda

Everyone is a Scientist

Can we afford it?



Workloads



Benchmarking

Scheduling

Graph Analytics

BTWorld

Elastic MR

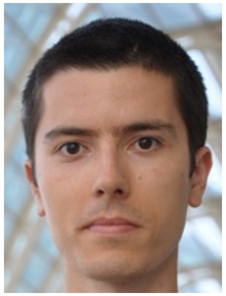
Conclusion

TU Delft

Graph Analytics: Our Team

Alexandru Iosup
TU Delft
Big Data & Clouds
Res. management
Systems, Benchmarking

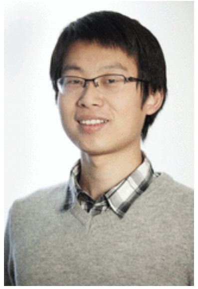
Dick Epema
TU Delft
Big Data & Clouds
Res. management
Systems



Mihai Capota
TU Delft
Big Data apps
Benchmarking

Ana Lucia Varbanescu
U. Amsterdam
Graph processing
Benchmarking

Claudio Martella
VU Amsterdam
Graph processing

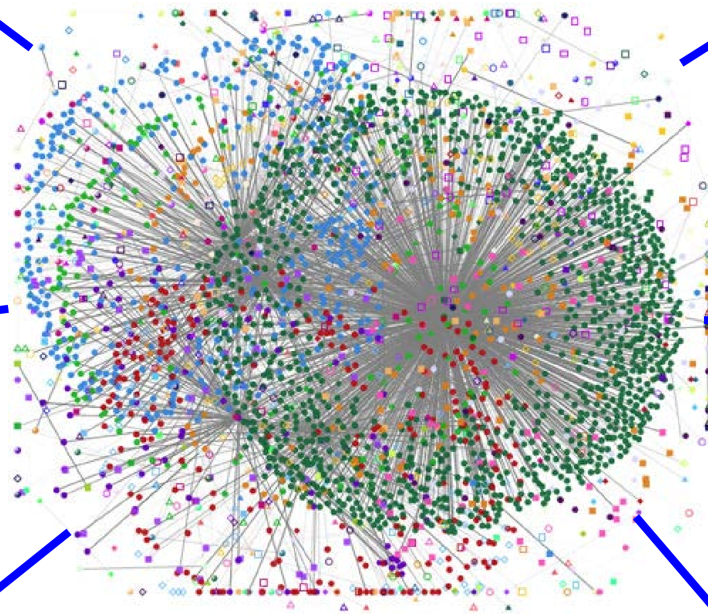


Yong Guo
TU Delft
Graph processing
Benchmarking

Marcin Biczak
TU Delft
Big Data & Clouds
Performance & Development

Graph Analytics

The data deluge: large-scale graphs



Platform diversity

- Platform: the combined hardware, software, and programming system that is being used to complete a graph processing task.

YAR
N

Trini
ty

What is the performance of these platforms?

Performance Metrics

Graph Diversity

Algorithm Diversity

- Graph500
  - Single application (BFS), Single class of synthetic datasets
- Few existing platform-centric comparative studies
  - Prove the superiority of a given system, limited set of metrics

Our vision: a benchmarking suite for graph processing across all platforms

Our Method

A benchmark suite for
performance evaluation of graph-processing platforms

1. Multiple Metrics, e.g.,
   - Execution time
   - Normalized: EPS, VPS
   - Utilization
2. Representative graphs with various characteristics, e.g.,
   - Size
   - Directivity
   - Density
3. Typical graph algorithms, e.g.,
   - BFS
   - Connected components

http://bit.ly/10hYdIU

Guo, Biczak, Varbanescu, Iosup, Martella, Willke.
How Well do Graph-Processing Platforms Perform?
An Empirical Performance Evaluation and Analysis

Graphitti

Platforms we have evaluated

Portability

- Distributed or non-distributed
- Graph-specific or generic

YAR
N

Distributed (Generic)	Distributed (Graph-specific)	Non-distributed (Graph-specific)

Y. Guo, M. Biczak, A. L. Varbanescu, A. Iosup, C. Martella, and T. L. Willke. How Well do Graph-Processing Platforms Perform? An Empirical Performance Evaluation and Analysis,IPDPS'14.

Key Findings From the Study of 6 Platforms

- Performance is function of (Dataset, Algorithm+Data Structure, Platform, Deployment)
  - Previous performance studies may lead to tunnel vision
  - Also looked at data structure, for CPU/GPU (submitted to ICPE'15)

- Platforms have their own drawbacks (crashes, long execution time, tuning, etc.)

- Some platforms can scale up reasonably with cluster size (horizontally) or number of cores (vertically)
  - Strong vs weak scaling still a challenge—workload scaling tricky

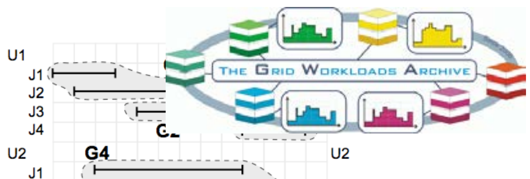
Y. Guo, M. Biczak, A. L. Varbanescu, A. Iosup, C. Martella, and T. L. Willke. How Well do Graph-Processing Platforms Perform? An Empirical Performance Evaluation and Analysis,IPDPS'14.

TUDelft

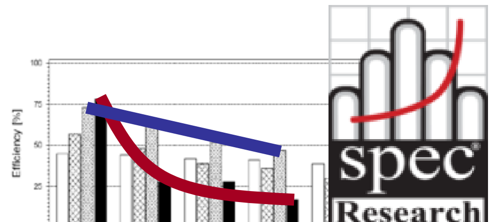
Agenda

Everyone is a Scientist

Can we afford it?



Workloads



Benchmarking

Scheduling

Graph Analytics

BTWorld

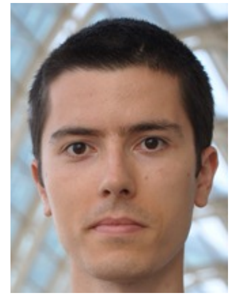
Elastic MR

Conclusion

BTWorld: Our Team

Alexandru Iosup
TU Delft
Big Data & Clouds
Res. management
Systems, Benchmarking

Dick Epema
TU Delft
Big Data & Clouds
Res. management
Systems



Mihai Capota
TU Delft
Big Data apps
Benchmarking

Jan Hidders

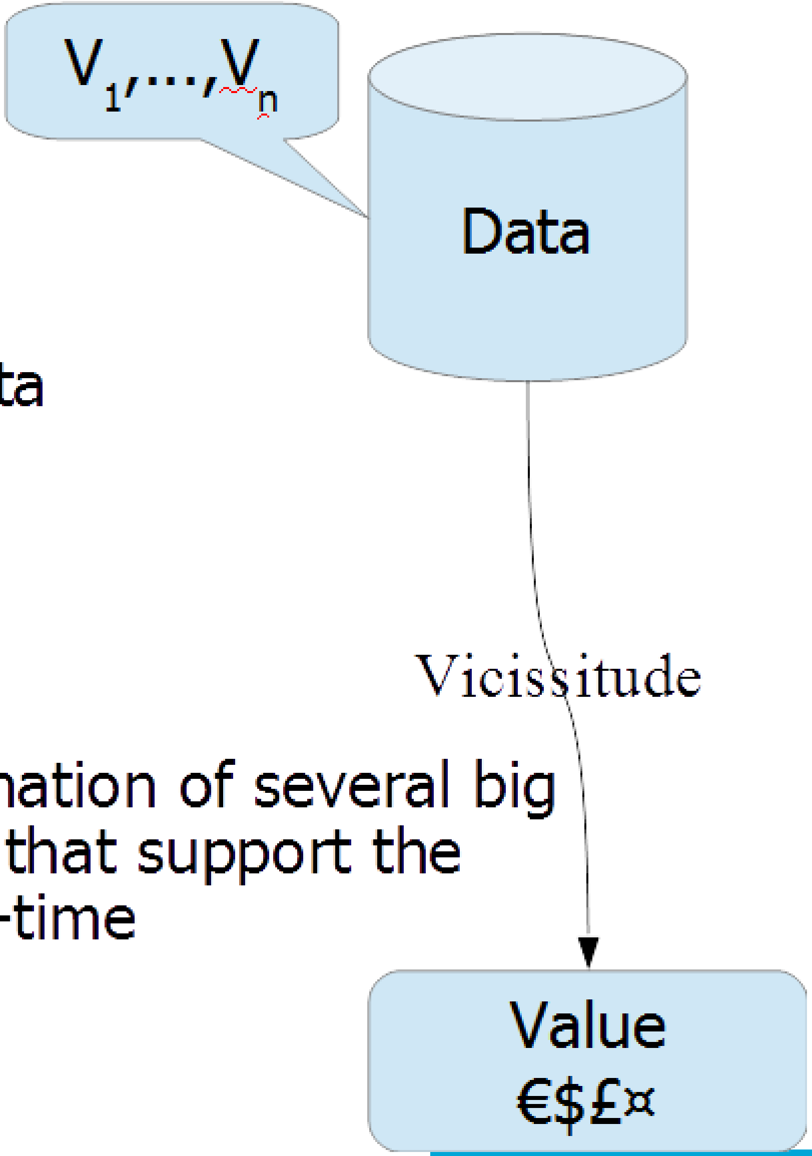
Tim Hegeman

BTWorld

tribler

Winners IEEE TCSC Scale Challenge 2014

Vs of big data

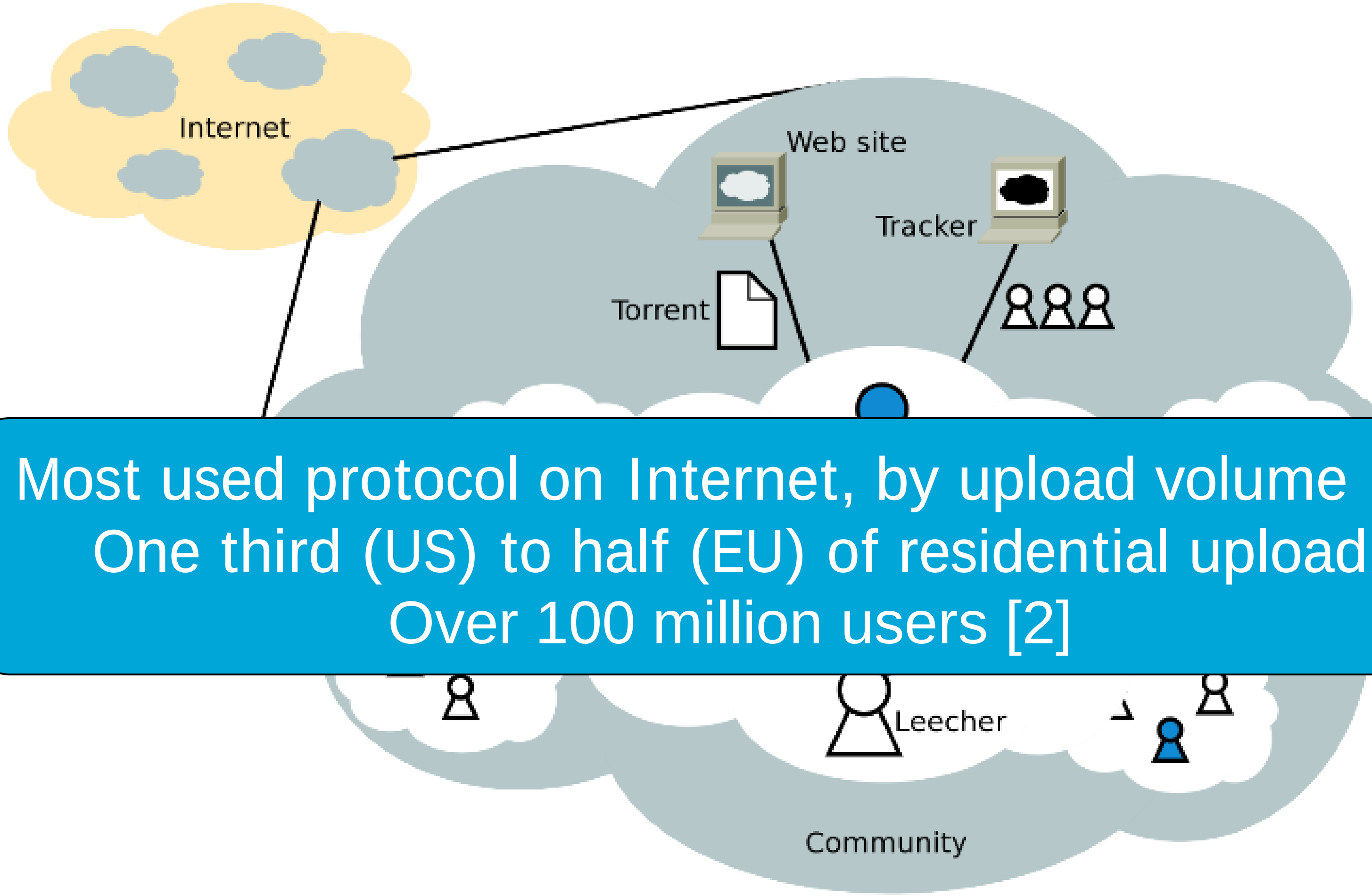


- Volume – large scale of data
- Variety – different forms of data
- Velocity – timeliness of data
- Veracity – uncertainty of data
- **Vicissitude** – dynamic combination of several big data Vs in processing systems that support the addition of new queries at run-time

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

TU Delft

Observing BitTorrent: Managing A Typical Global Distributed System

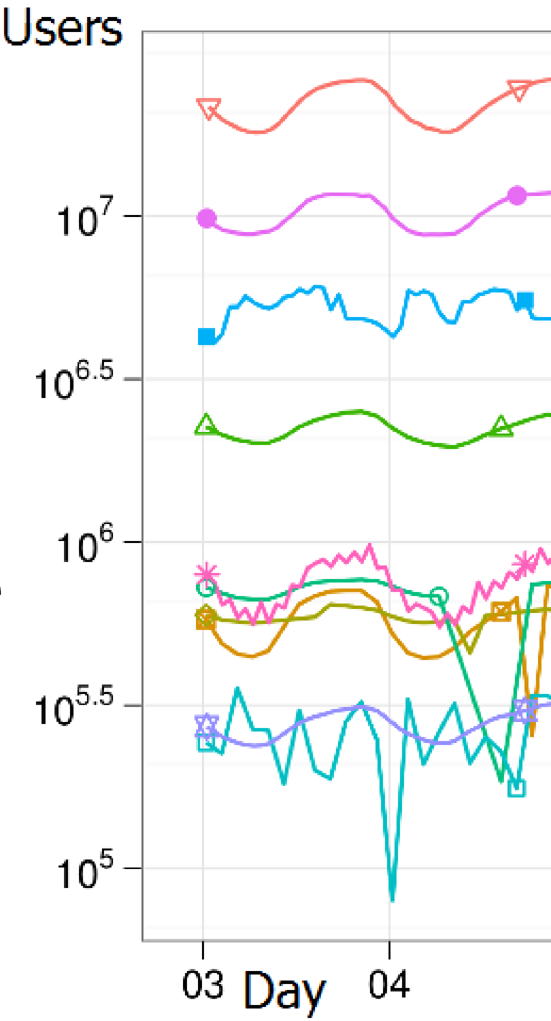


Most used protocol on Internet, by upload volume [1]
One third (US) to half (EU) of residential upload
Over 100 million users [2]

[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

Collected Data

- Ongoing longitudinal study, since 2009
- Data-driven project:
  data first, ask questions later
- Over 15TB of data, 1 file/tracker/sample
- Timestamped, multi-record files
  - Hash: unique id for file
  - Tracker: unique id for tracker
  - Information per file: seeders, leechers
  - Structured and semi-structured data



Wojciechowski, Capota, Pouwelse, and Iosup. BTWorld: Towards observing the global BitTorrent file-sharing network. HPDC 2010

The BTWorld Use Case (When Long-Term Traces Do Not Exist)

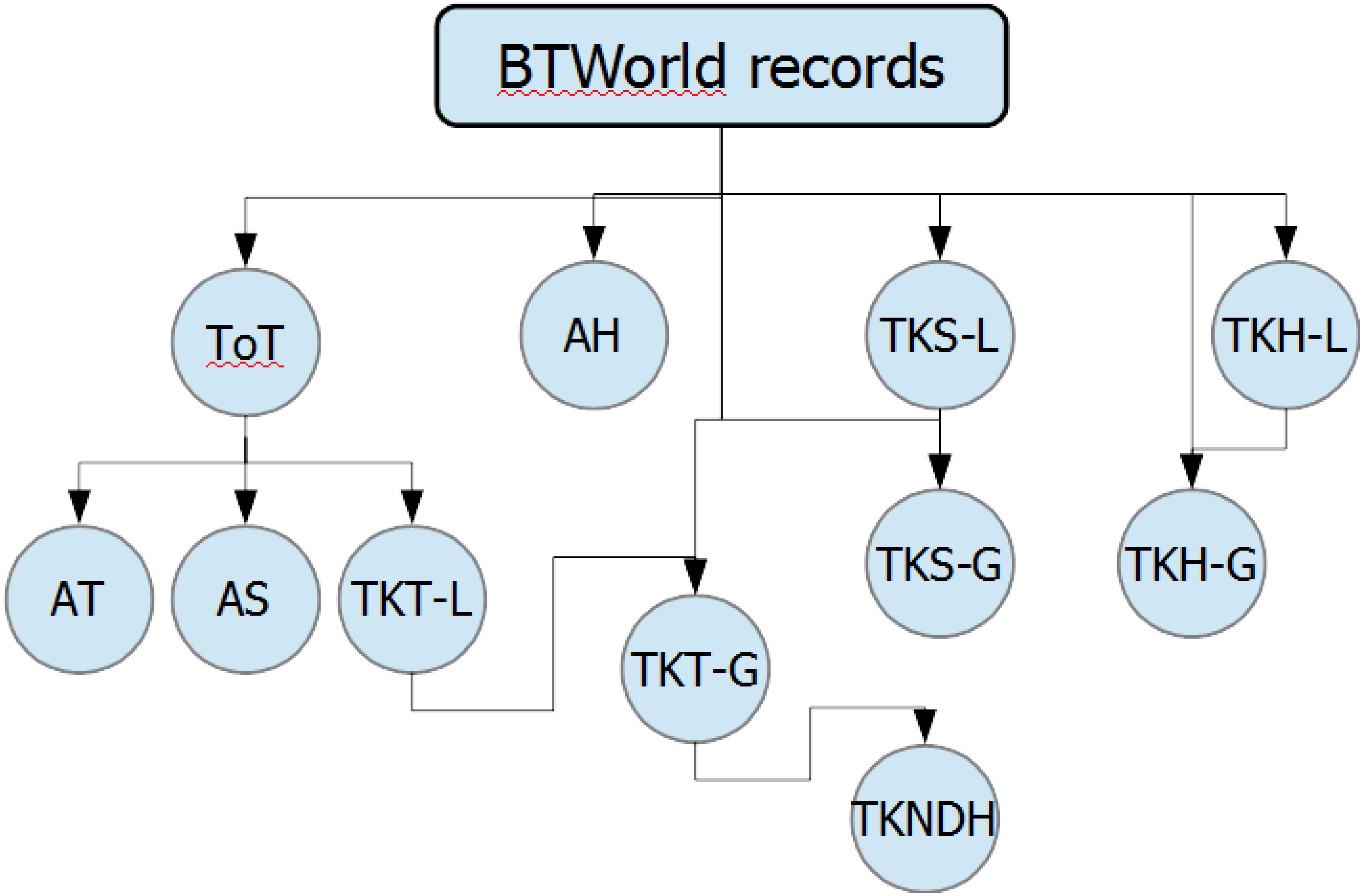
Analyst Questions

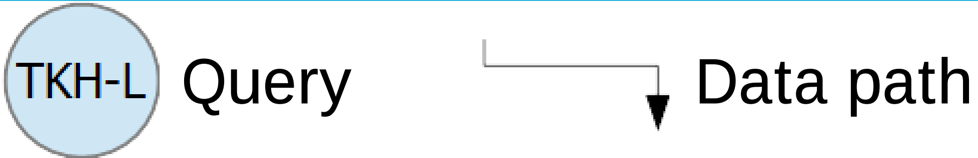
- How does the number of peers evolve over time?
- How long are files available?
- Did the legal bans and tracker take-downs impact BT?
- How does the location of trackers evolve over time?
- Etc.

These questions need to be translated into queries

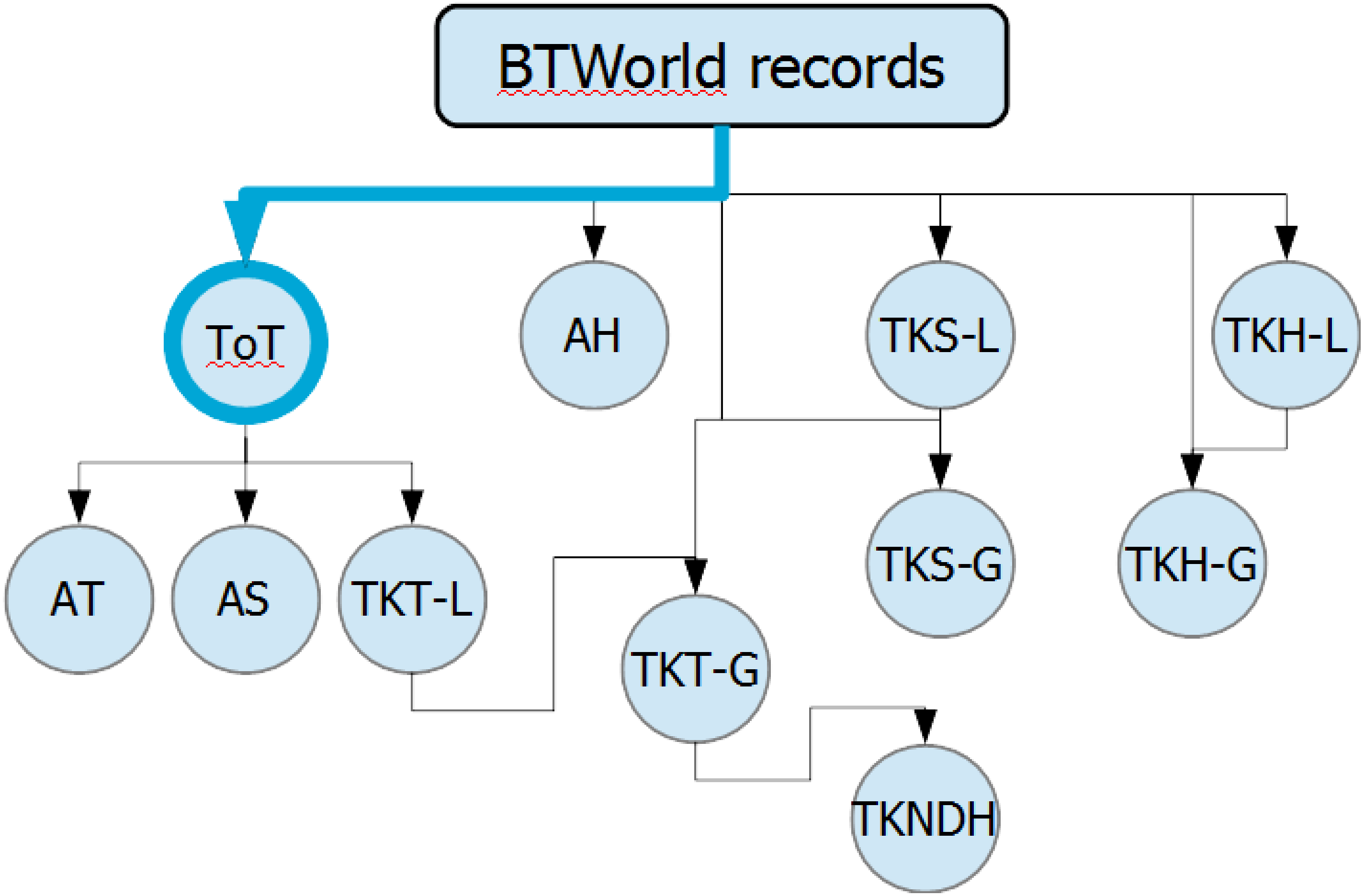
Hegeman, Ghit, Capotã, Hidders, Epema, Iosup. The BTWorld Use Case for Big Data Analytics: Description, MapReduce Logical Workflow, and Empirical Evaluation.IEEE BigData’13

The BTWorld Workflow



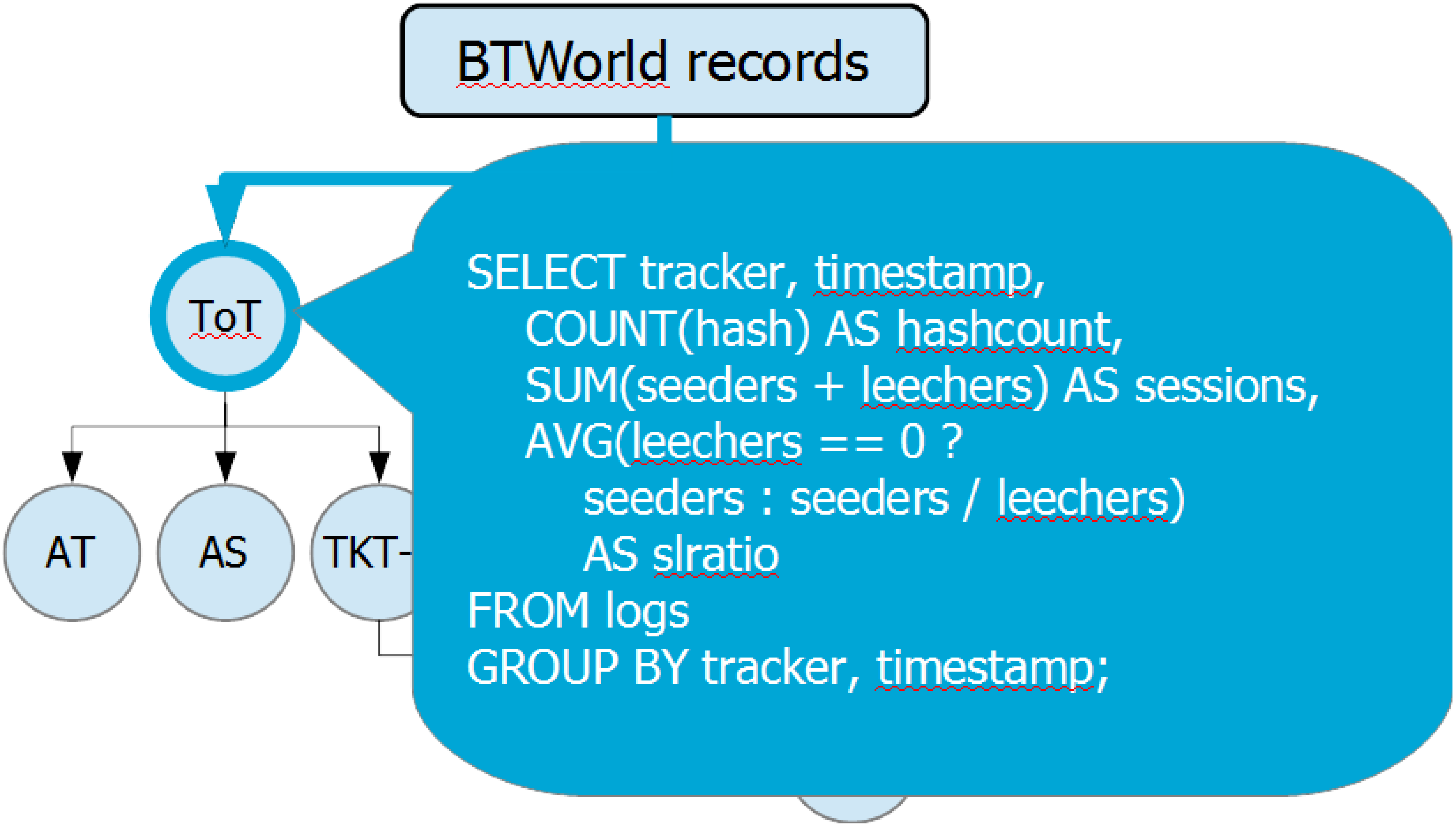


The BTWorld Workflow

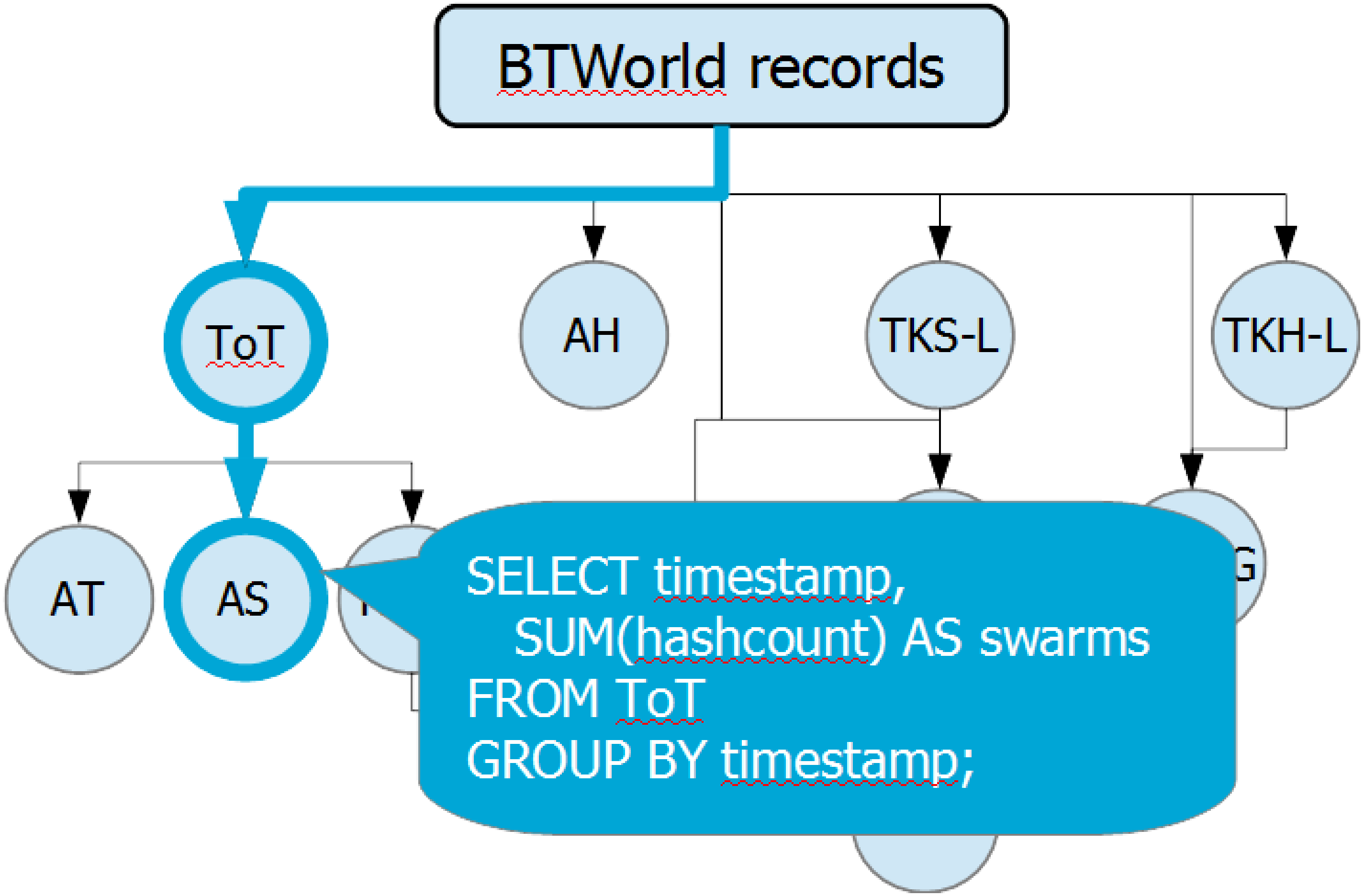


TU Delft

The BTWorld Workload



The BTWorld Workload



MapReduce-based Workflow for the BTWorld Use Case

Query Diversity

- Queries use different operators, stress different parts of system
- This kind of workflow is **not** modeled well by single-application benchmarks

Global Top K Trackers (TKT-G):

```
SELECT *
FROM logs
NATURAL JOIN (
    SELECT tracker
    FROM TKTL
    GROUP BY tracker
    ORDER BY MAX(sessions) DESC
    LIMIT k);
```

Active Hashes (AH):

```
SELECT timestamp, COUNT(DISTINCT(hash))
FROM logs
GROUP BY timestamp;
```

Hegeman, Ghit, Capotã, Hidders, Epema, Iosup. The BTWorld Use Case for Big Data Analytics: Description, MapReduce Logical Workflow, and Empirical Evaluation.IEEE BigData’13

Beyond BTWorld

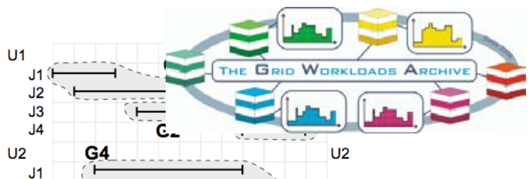
BitTorrent	Trackers	Swarms	Hashes
Finance	Stock markets	Stock listings	Stocks
Tourism	Travel agents	Vacation packages	Venues

- Monitoring large scale distributed computer systems
- Benchmarking

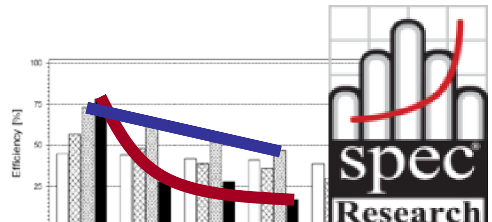
Agenda

Everyone is a Scientist

Can we afford it?



Workloads



Benchmarking

Scheduling

Graph Analytics

BTWorld

Elastic MR

Conclusion

TU Delft

Elastic MapReduce: Our Team

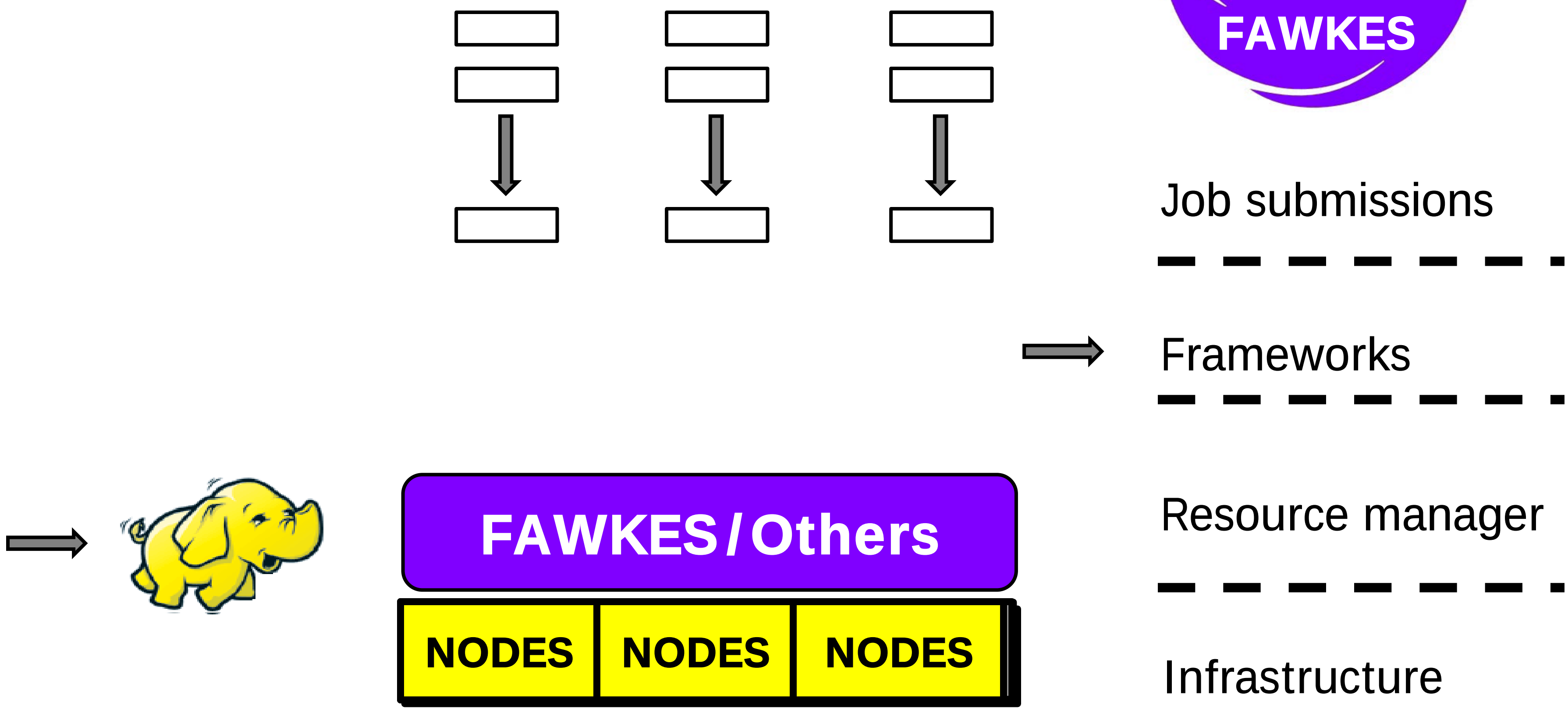
Bogdan Ghit
TU Delft
Systems
Workloads

Dick Epema
TU Delft
Big Data & Clouds
Res. management
Systems

Alexandru Iosup
TU Delft
Big Data & Clouds
Res. management
Systems, Benchmarking

Dynamic Big Data Processing

Fawkes = Elastic MapReduce via Two-level scheduling architecture

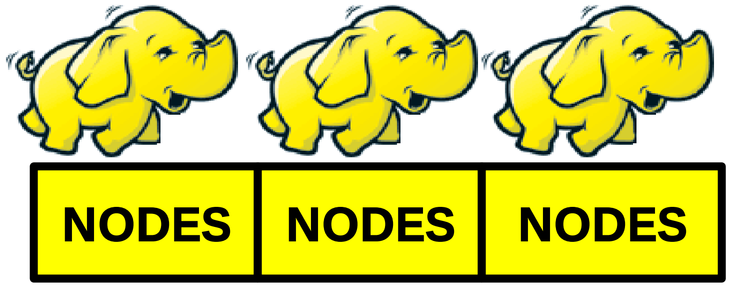


Ghit, Yigitbasi, Iosup, Epema, Iosup. Balanced Resource Allocations Across Multiple Dynamic MapReduce Clusters. ACM SIGMETRICS 2014.

Elastic MapReduce

MapReduce framework

- Distributed file system
- Execution engine
- Data locality constraints

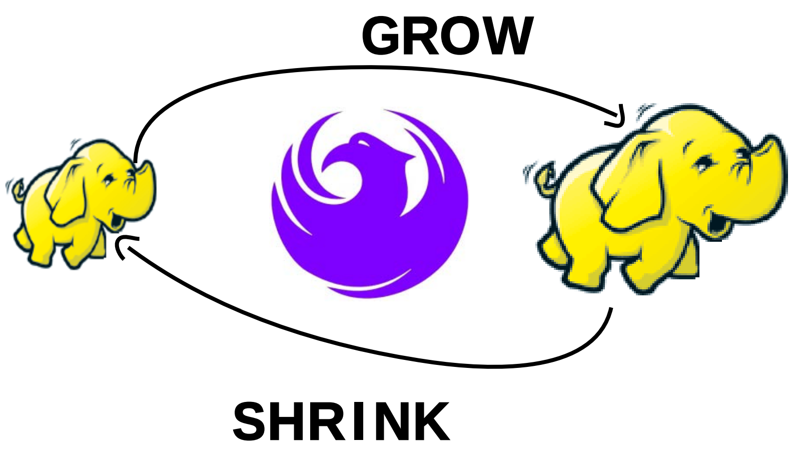


Because workloads may be time-varying:

- Poor resource utilization
- Imbalanced service levels

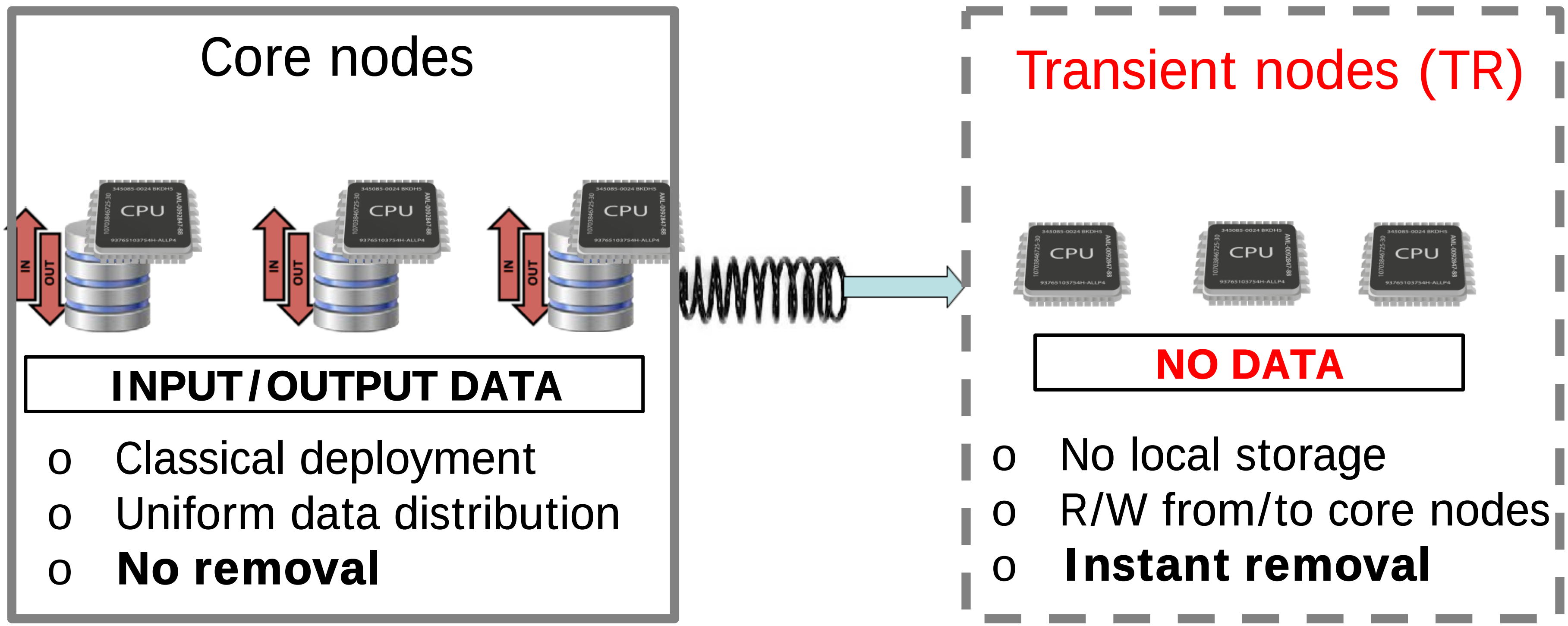
Grow and shrink MapReduce

- High resource utilization
- Reconfiguration for balanced service levels
- Break data locality



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

No data locality



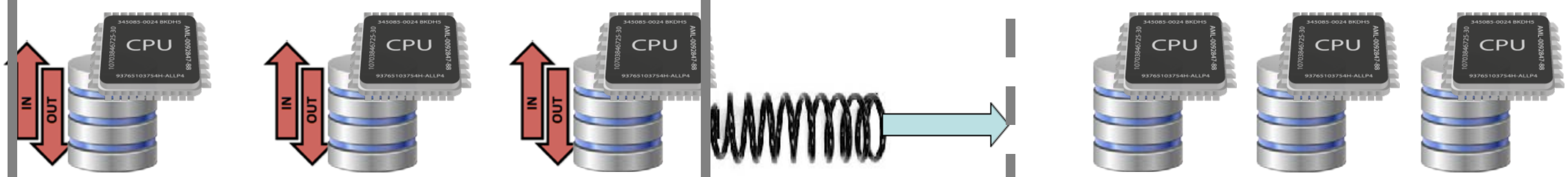
Performance?

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

Relaxed data locality

Core nodes

Trans-core nodes (TC)



INPUT/OUTPUT DATA

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

OUTPUT DATA

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

Better performance?

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

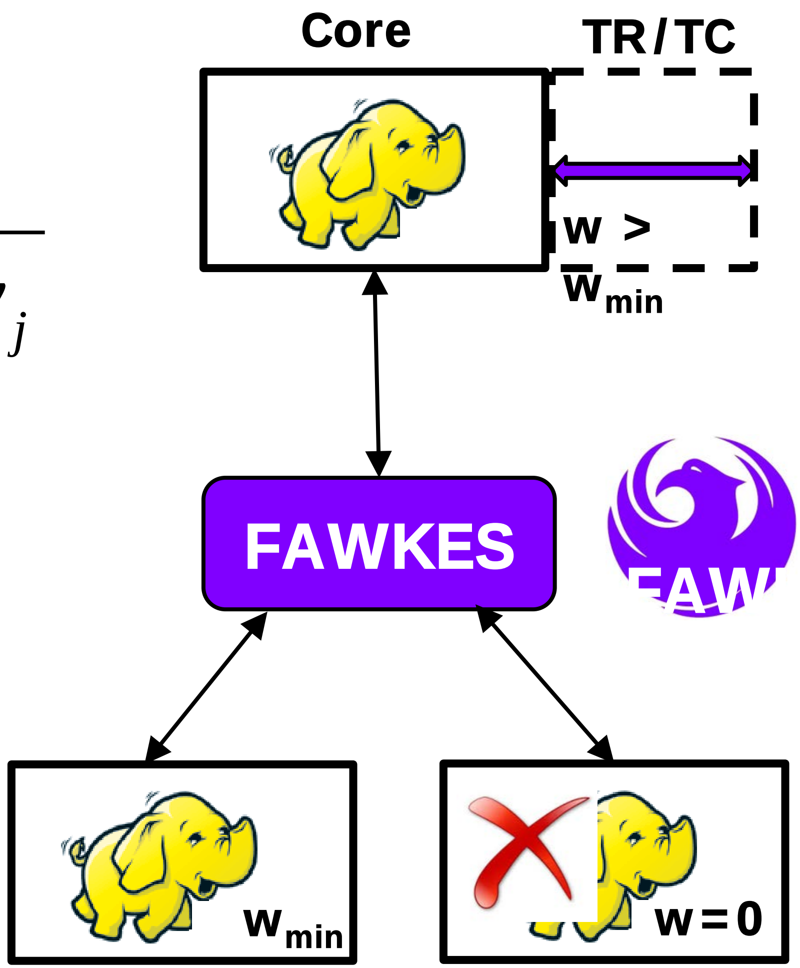
FAWKES in a nutshell

1. Size of MapReduce cluster
   - Changes dynamically
   - Balanced by weight

$$s_i = \frac{w_i}{\sum w_j}$$

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

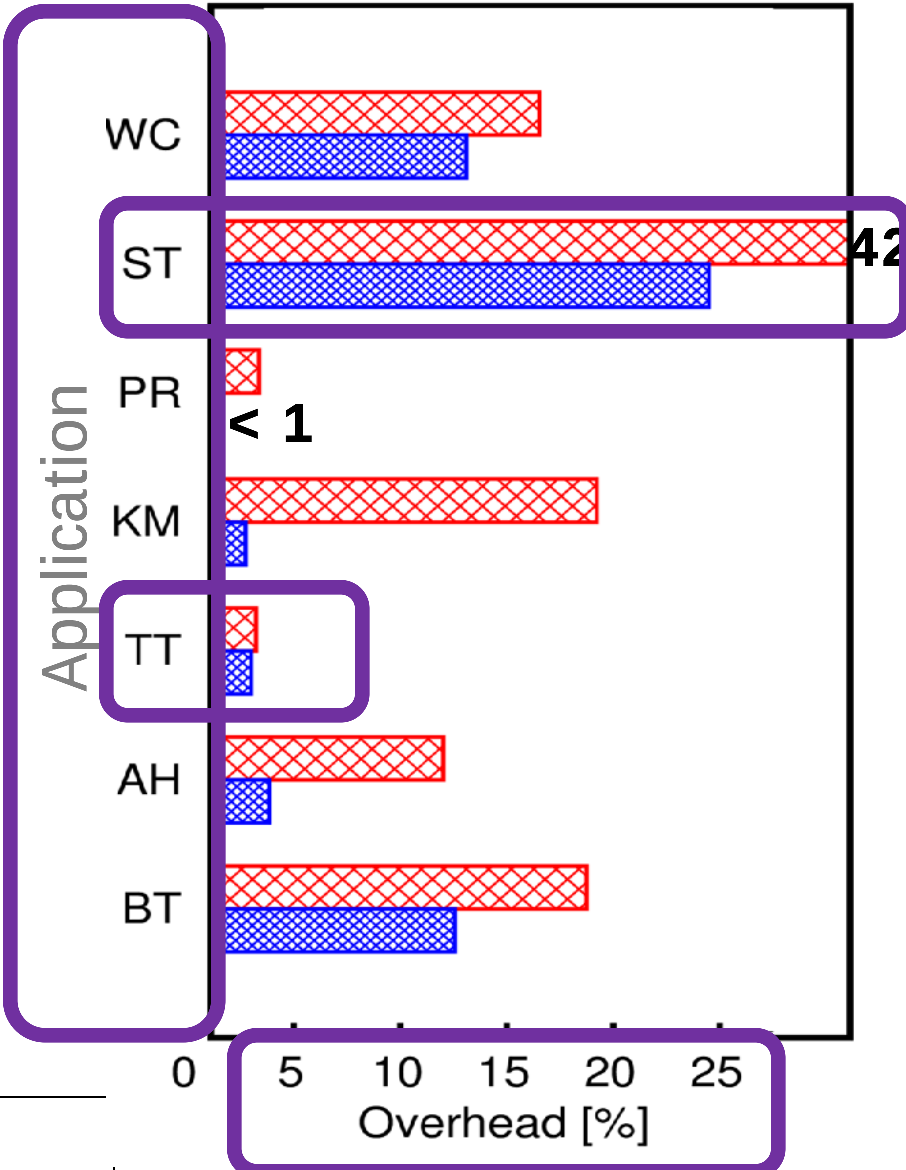
3. Shrinks and grows frameworks to
   - Allocate new frameworks (min. shares)
   - Give fair shares to existing ones



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

Performance of dynamic MapReduce

10 core +10xTR
10 core +10xTC
vs.
20 core nodes



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

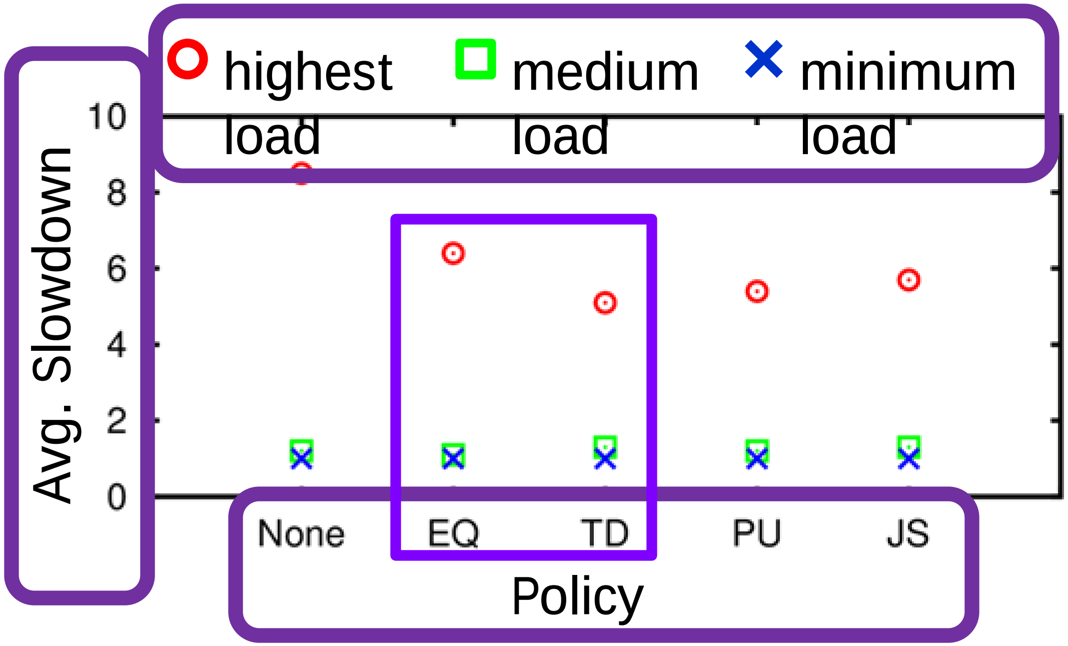
TC - **needed** for disk-intensive workloads.

Dynamic MapReduce: < 25% overhead

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

Performance of FAWKES

Nodes	45
Frameworks	3
Min. shares	10
Datasets	300 GB
Jobs submitted	900



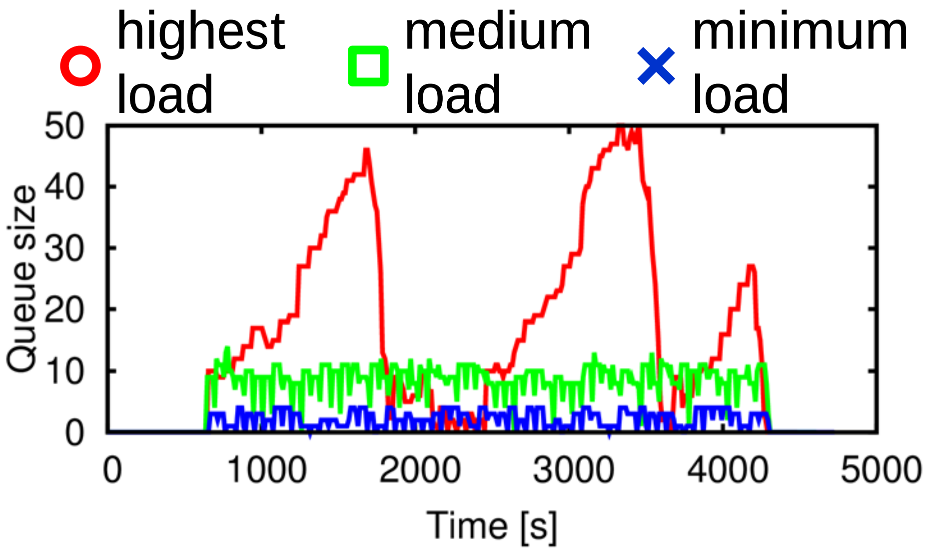
Up to 20% lower slowdown

None – Minimum shares
EQ – EQual shares
TD – Task Demand
PU – Processor Usage
JS – Job Slowdown

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

FAWKES: behind the scenes

EQ

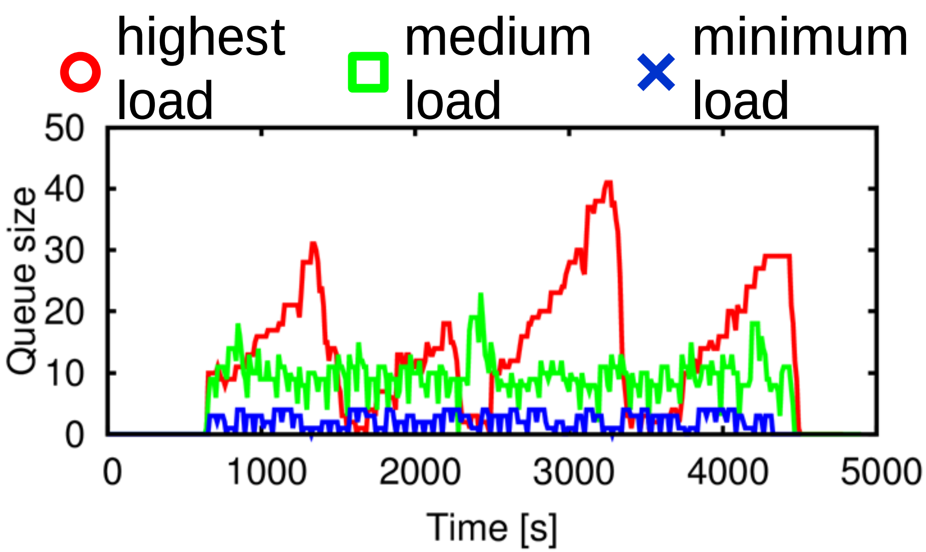


Utilizations: 60% / 23% / 5%

Imbalanced

EQ – EQual shares
TD – Task Demand

TD

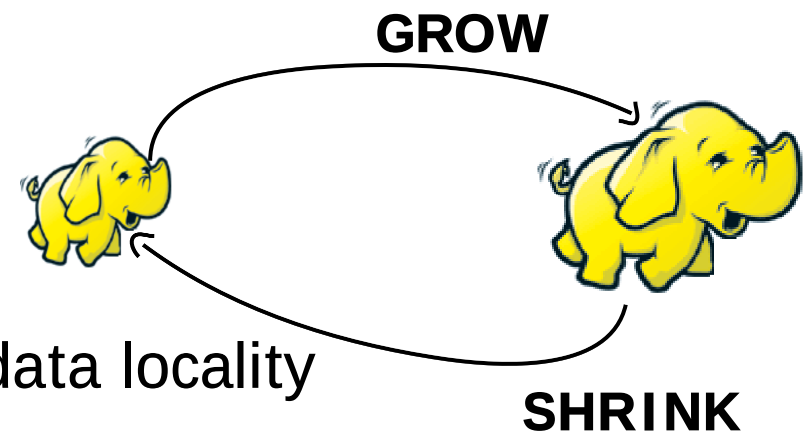


Utilizations: 50% / 30% / 8%

More balanced

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

Take-home message



1. Dynamic MapReduce relaxes data locality
2. FAWKES policies can reduce imbalance between frameworks
3. More aggressive policies?

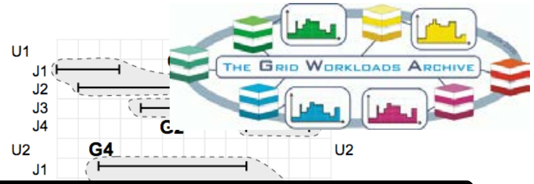
FAWKES

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

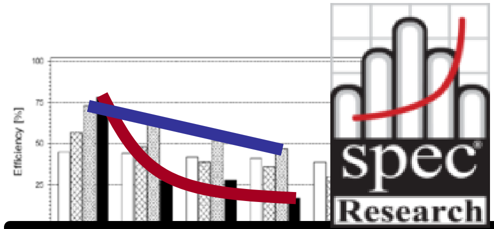
Agenda

Everyone is a Scientist

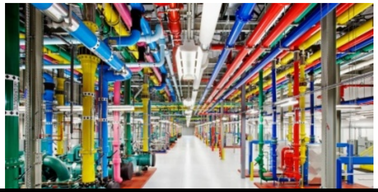
Can we afford it?



Workloads



Benchmarking



Scheduling

Graph Analytics

BTWorld

Elastic MR

Conclusion

TU Delft

~~Conclusion~~ Take-Home Message

- **"Everyone is a Scientist!"**
  Our vision of a growing, leading Europe
- **Cloud-based Big Data is a grand challenge**
  - Managing the datacentre
  - Helping demanding users

- **In this talk**
  - Understanding workloads
  - Benchmarking
  - Scheduling
  - Cloud-based big data:
    graph processing, data processing workflows, elastic MapReduce

Thank you for your attention! Questions? Suggestions? Observations?

More Info:

- http://www.st.ewi.tudelft.nl/~iosup/
- http://www.pds.ewi.tudelft.nl/
- http://research.spec.org

Do not hesitate to contact me...

Alexandru Iosup

A.Iosup@tudelft.nl
(or google “iosup”)
Parallel and Distributed Systems Group
Delft University of Technology