Benchmarking Big Data in the Data Center: A TU Delft Perspective

Alexandru Iosup
Delft University of Technology
The Netherlands

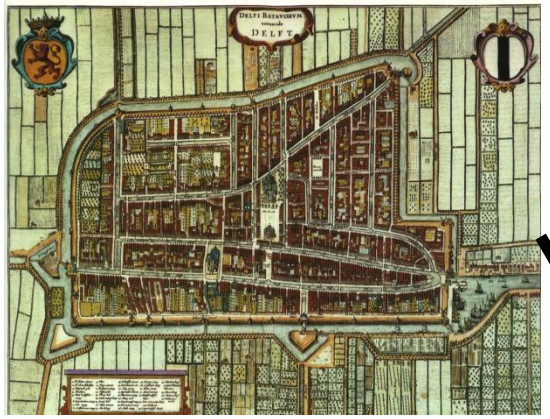
Team: **Undergrad** Tim Hegeman, … **Grad** Yong Guo, Mihai Capota, Bogdan Ghit
Researchers Marcin Biczak, Otto Visser **Staff** Henk Sips, Dick Epema
Collaborators* Ana Lucia Varbanescu (UvA, Ams), Claudio Martella (VU, Giraph), KIT, Intel Research Labs, IBM TJ Watson, SAP, Google Inc. MV, Salesforce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

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

spec Research

TU Delft
Delft University of Technology

(TU) Delft – the Netherlands – Europe



founded 13th century
pop: 100,000

founded 1842
pop: 13,000

pop: 16.5 M

Potsdam (We are here)

TU Delft

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

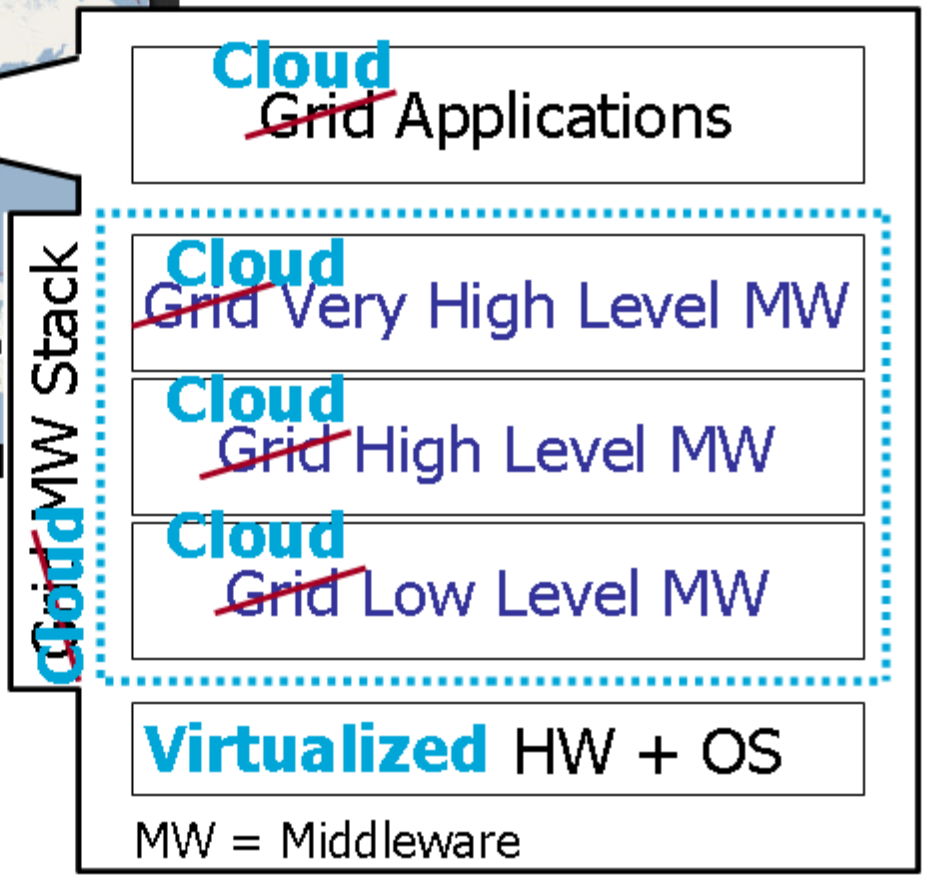
- see PDS publication database at publications.st.ewi.tudelft.nl

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/

"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

* Subset.



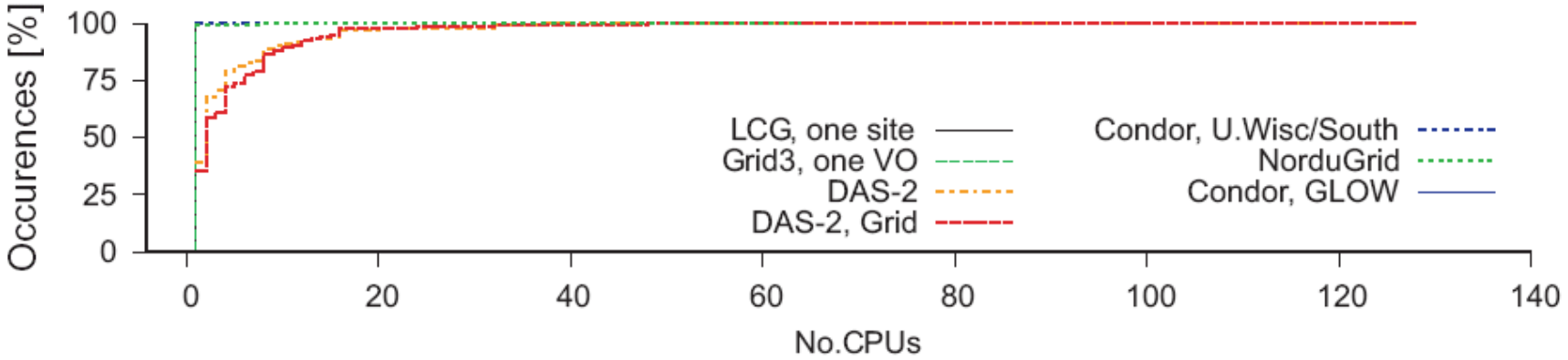
TU Delft

Lessons From Grids

From Parallel to Many-Task Computing (users, not designers, find what works)

Show the numbers, please!

- Average job size is 1 (that is, there are **no [!] tightly-coupled, only conveniently parallel jobs**)



A. Iosup, C. Dumitrescu, D.H.J. Epema, H. Li, L. Wolters, How are Real Grids Used? The Analysis of Four Grid Traces and Its Implications, Grid 2006.

A. Iosup and D.H.J. Epema, Grid Computing Workloads, IEEE Internet Computing 15(2): 19-26 (2011)

Large-scale = unreliable infrastructure

Resource management is key

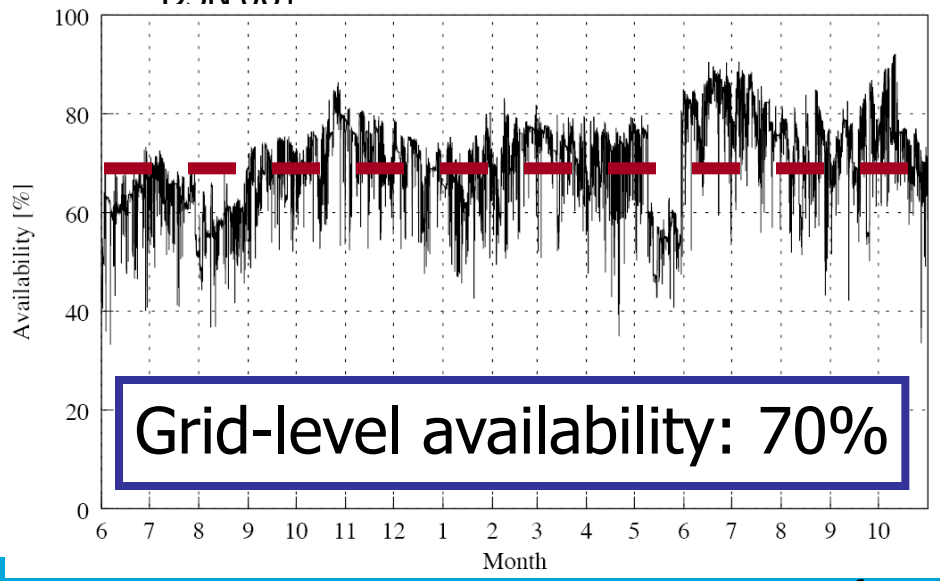
- Server
- Cluster
- Production

- 99.9(…)9% reliable
- 5x decrease in failure rate after first year [Schroeder and Gibson, DSN'06]

SiteName	TOTAL
BNL-LCG2	0 vs 1 (0.00%)
INFN-T1	19066 vs 6042 (75.94%)
NIKHEF-ELPROD	5994 vs 22270 (21.21%)
RAL-LCG2	21631 vs 22391 (49.14%)
Taiwan-LCG2	18254 vs 9246 (66.38%)
USCMS-FNAL-WC1	101542 vs 8623 (92.17%)
pic	12851 vs 6627 (65.98%)
TOTAL	495281 vs 167668 (74.71%)

CERN LCG jobs
74.71% successful
25.29% unsuccessful

Source: dboard-gr.cern.ch, May'07.



A. Iosup, M. Jan, O. Sonmez, and D.H.J. Epema, On the Dynamic Resource Availability in Grids, Grid 2007, Sep 2007.

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 [illegible] person to enc[illegible] scientific kno[illegible] the last 50 ye[illegible] been the pace [illegible] 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.

TU Delft

Lessons from Grids

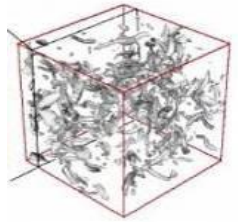
From Hypothesis to Data

ena

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

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

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

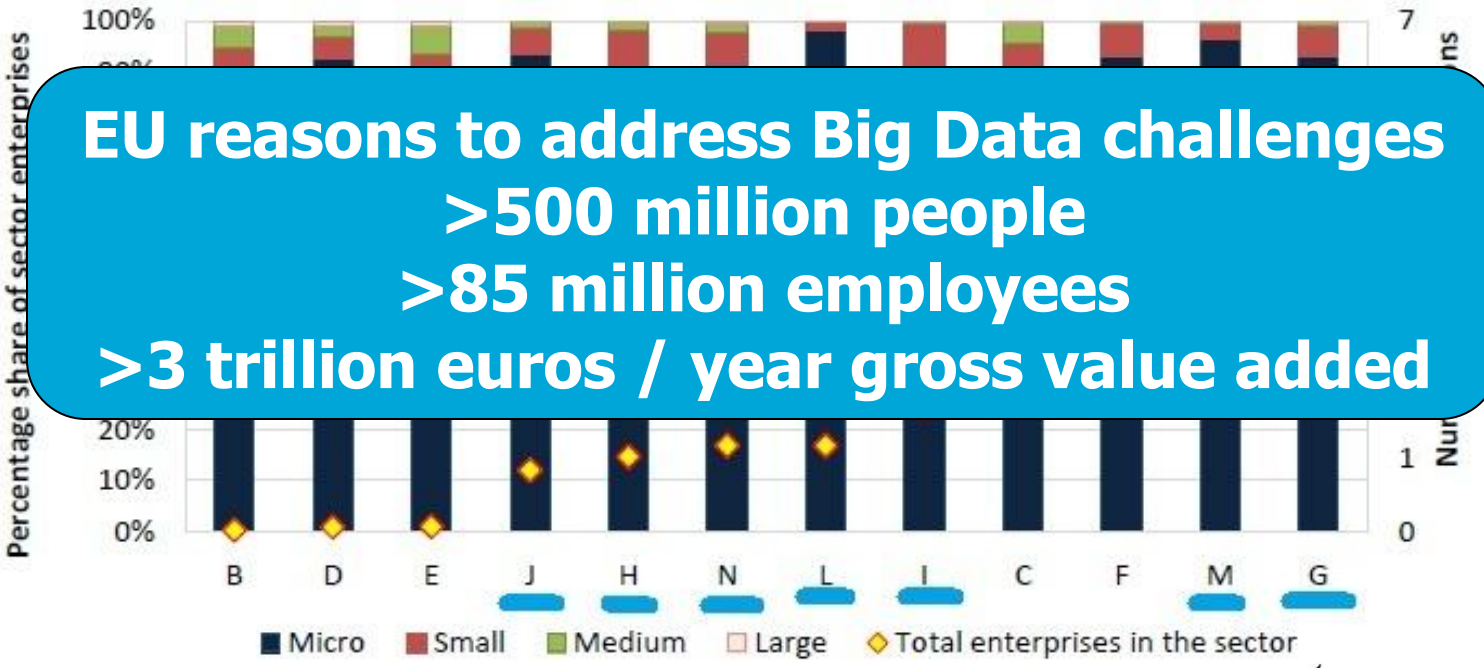


- Today (**the Fourth Paradigm**):
  data exploration
  unify theory, experiment, and simulation
  - Data captured by instruments or generated by simulator
  - Processed by software
  - Information/Knowledge stored in computer
  - Scientist analyzes results using data management and statistics

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)

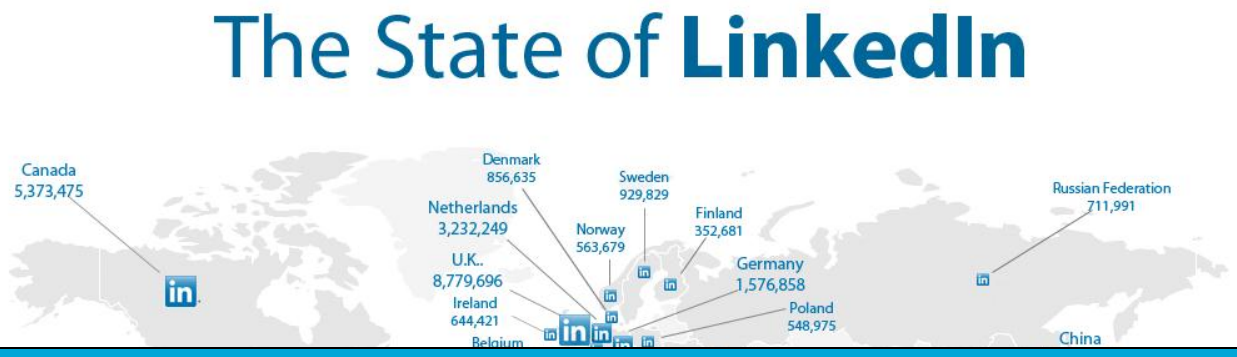


EU reasons to address Big Data challenges
>500 million people
>85 million employees
>3 trillion euros / year gross value added

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

TU Delft

Data at the Core of Our Society: The LinkedIn Example



A very good resource for matchmaking workforce and prospective employers

Vital for your company's life, as your Head of HR would tell you

Vital for the prospective employees

2

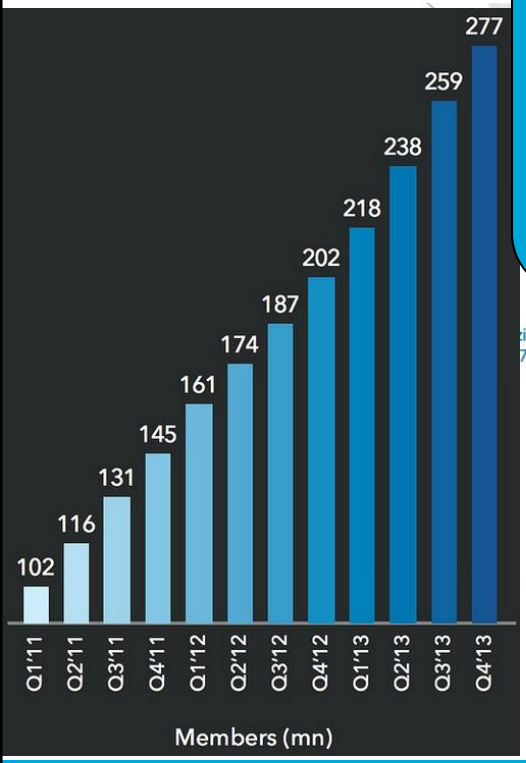
100M Mar 2011, 69M May 2010

Sources: Vincenzo Cosenza, The State of LinkedIn, http://vincos.it/the-state-of-linkedin/
via Christopher Penn, http://www.shiftcomm.com/2014/02/state-linkedin-social-media-dark-horse/

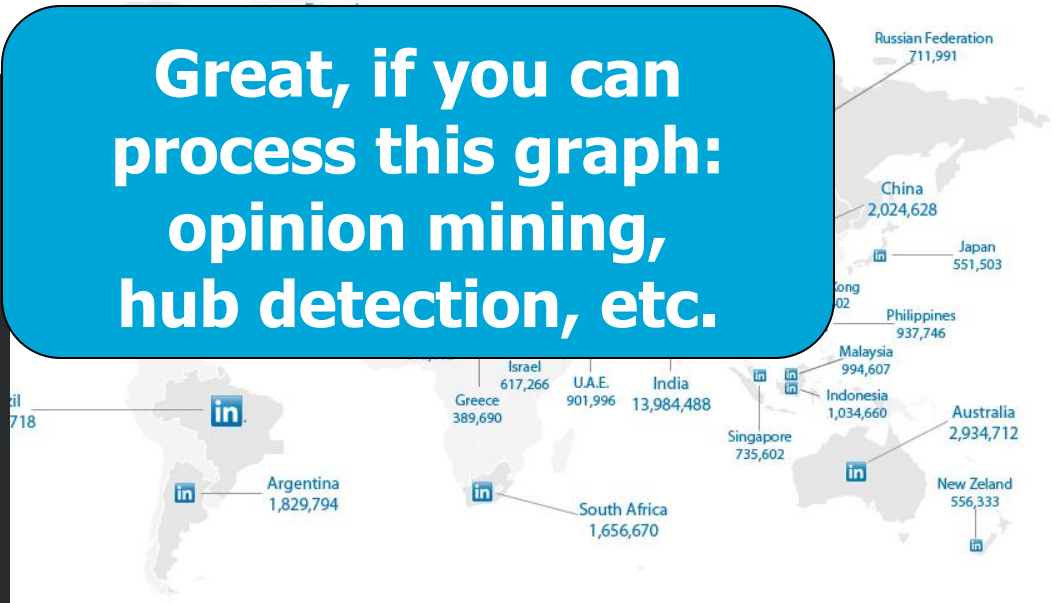
TU Delft

Data at the Core of Our Society: The LinkedIn Example

3-4 new users every second



Great, if you can process this graph: opinion mining, hub detection, etc.



150,000,000 registered members — **Feb 2012**

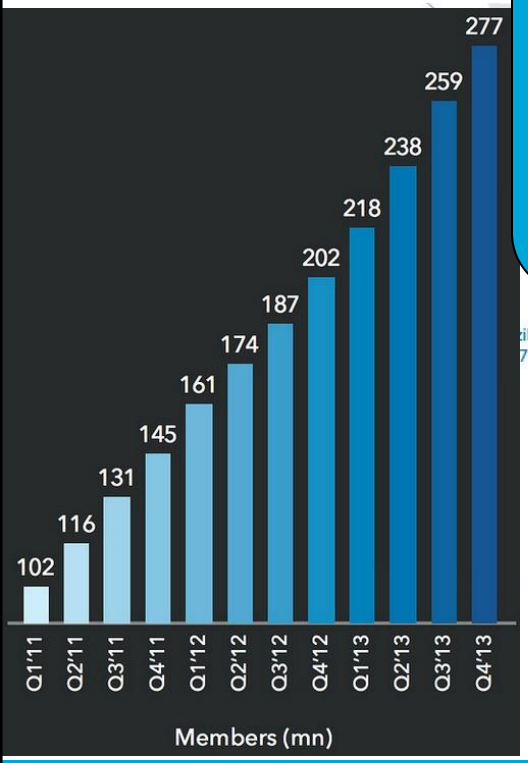
100M Mar 2011, 69M May 2010

Sources: Vincenzo Cosenza, The State of LinkedIn, http://vincos.it/the-state-of-linkedin/
via Christopher Penn, http://www.shiftcomm.com/2014/02/state-linkedin-social-media-dark-horse/

TU Delft

Data at the Core of Our Society: The LinkedIn Example

3-4 new users every second



Great, if you can process this graph: opinion mining, hub detection, etc.

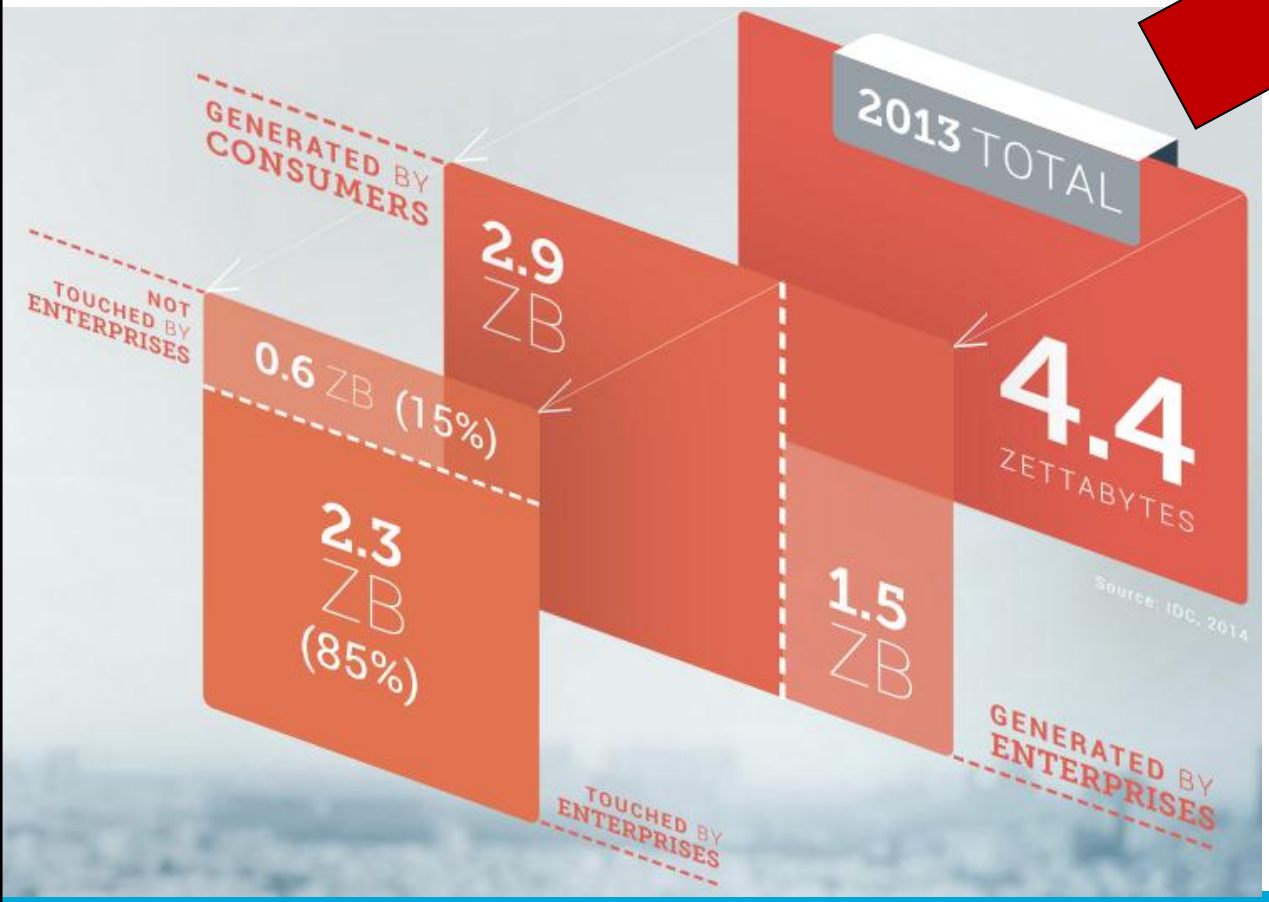
139/277 million questions of customer retention, so time-based analytics

but fewer visitors (and page views)

Sources: Vincenzo Cosenza, The State of LinkedIn, http://vincos.it/the-state-of-linkedin/
via Christopher Penn, http://www.shiftcomm.com/2014/02/state-linkedin-social-media-dark-horse/

TU Delft

LinkedIn Is Part of the "Data Deluge"



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

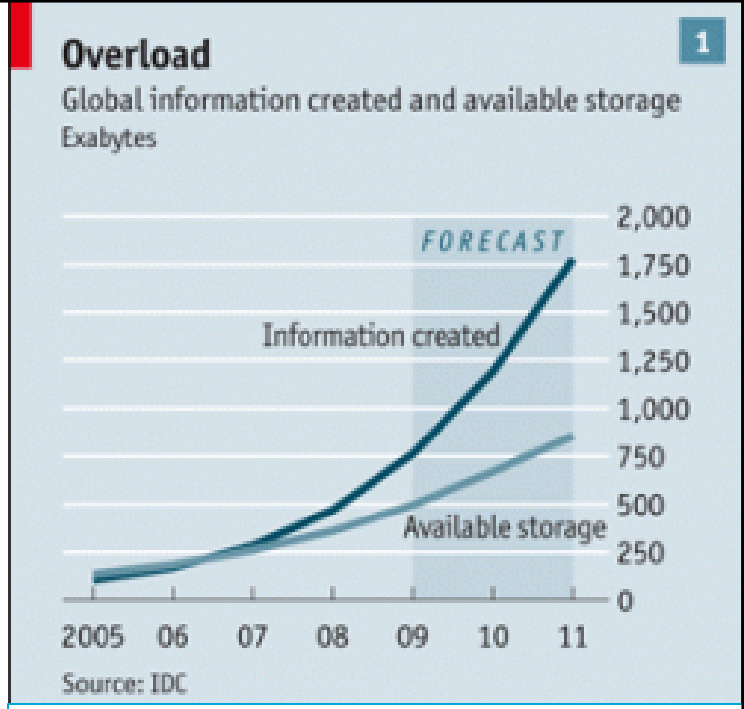
- Interacting
- Understanding
- Deciding
- Creating

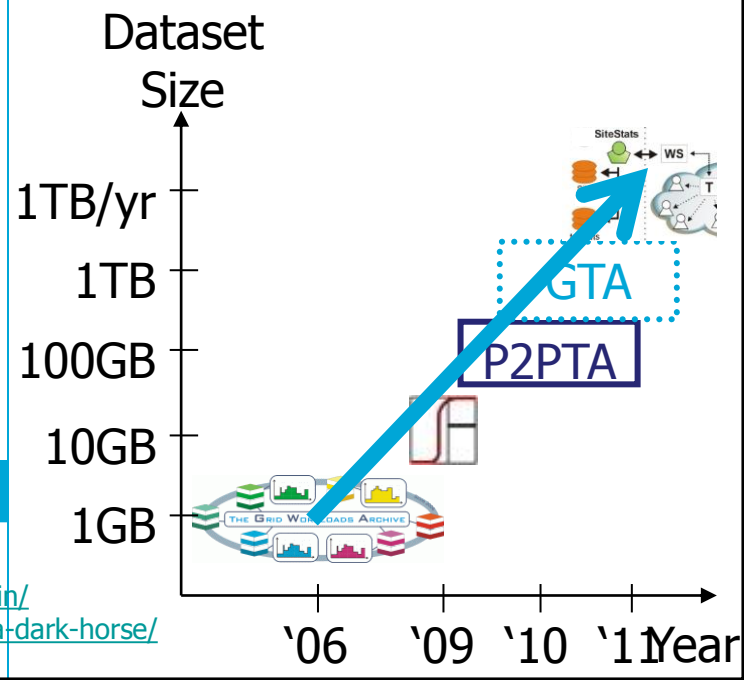
Sources: IDC, EMC.

TU Delft

The Data Deluge Is A Challenge for Tech But Good for Us[ers]

- All human knowledge
  - Until 2005: 150 Exa-Bytes
  - 2010: 1,200 Exa-Bytes
- Online gaming (Consumer)
  - 2002: 20TB/year/game
  - 2008: 1.4PB/year/game (only stats)
- Public archives (Science)
  - 2006: GBs/archive
  - 2011: TBs/year/archive





Sources: Vincenzo Cosenza, The State of LinkedIn, http://vincos.it/the-state-of-linkedin/
via Christopher Penn, http://www.shiftcomm.com/2014/02/state-linkedin-social-media-dark-horse/

The Challenge: The Three "V"s of Big Data* When You Can, Keep *and* Process Everything

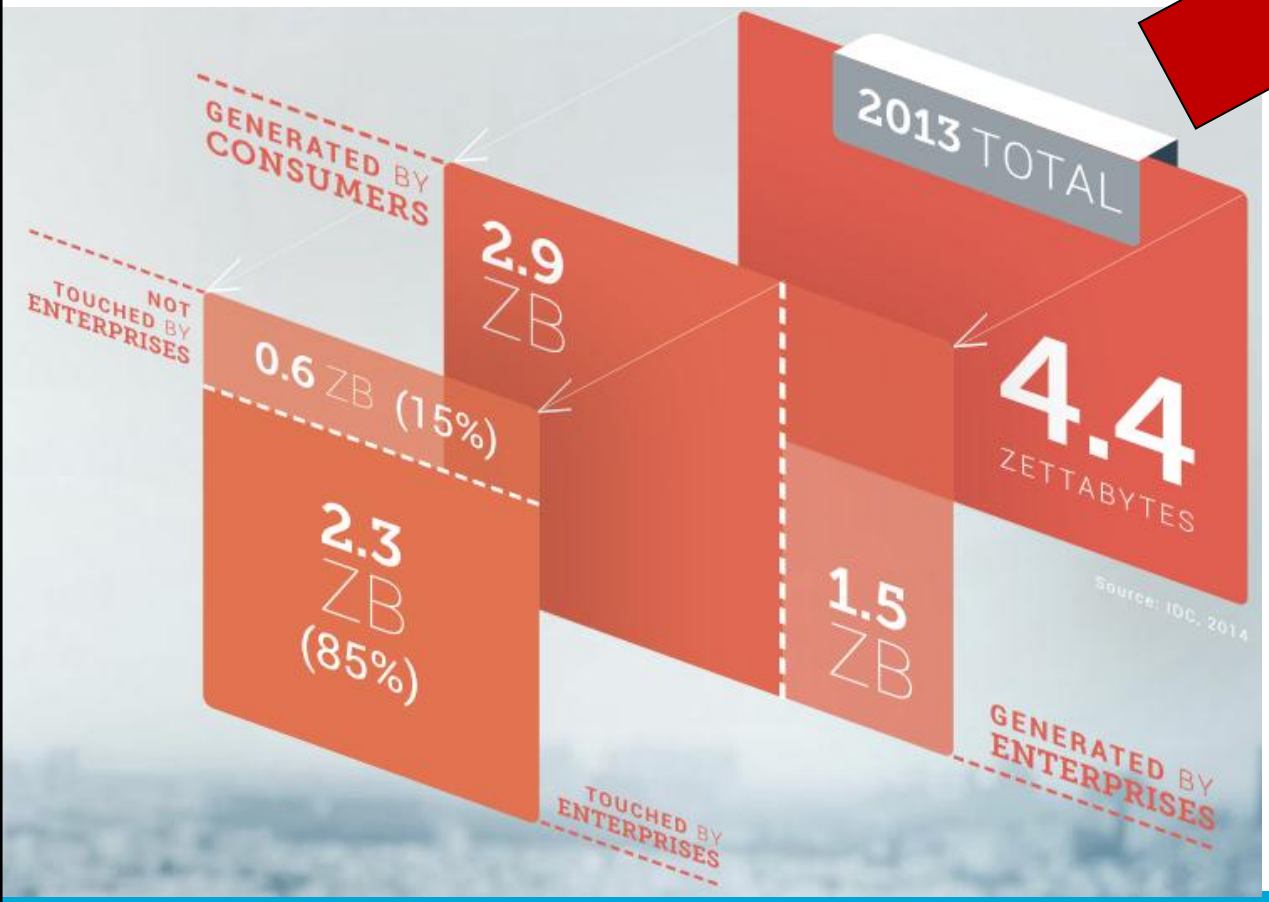
* New Vs later: ours is "vicissitude"

- Volume
  - More data vs. better models
  - Exponential growth + iterative models
  - Scalable storage and distributed queries
- Velocity
  - Speed of the feedback loop
  - Gain competitive advantage: fast recommendations
  - Analysis in near-real time to extract value
- Variety
  - The data can become messy: text, video, audio, etc.
  - Difficult to integrate into applications

Too big, too fast, does not comply with traditional DB

Adapted from: Doug Laney, "3D data management", META Group/Gartner report, Feb 2001. http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

The "Data Deluge"



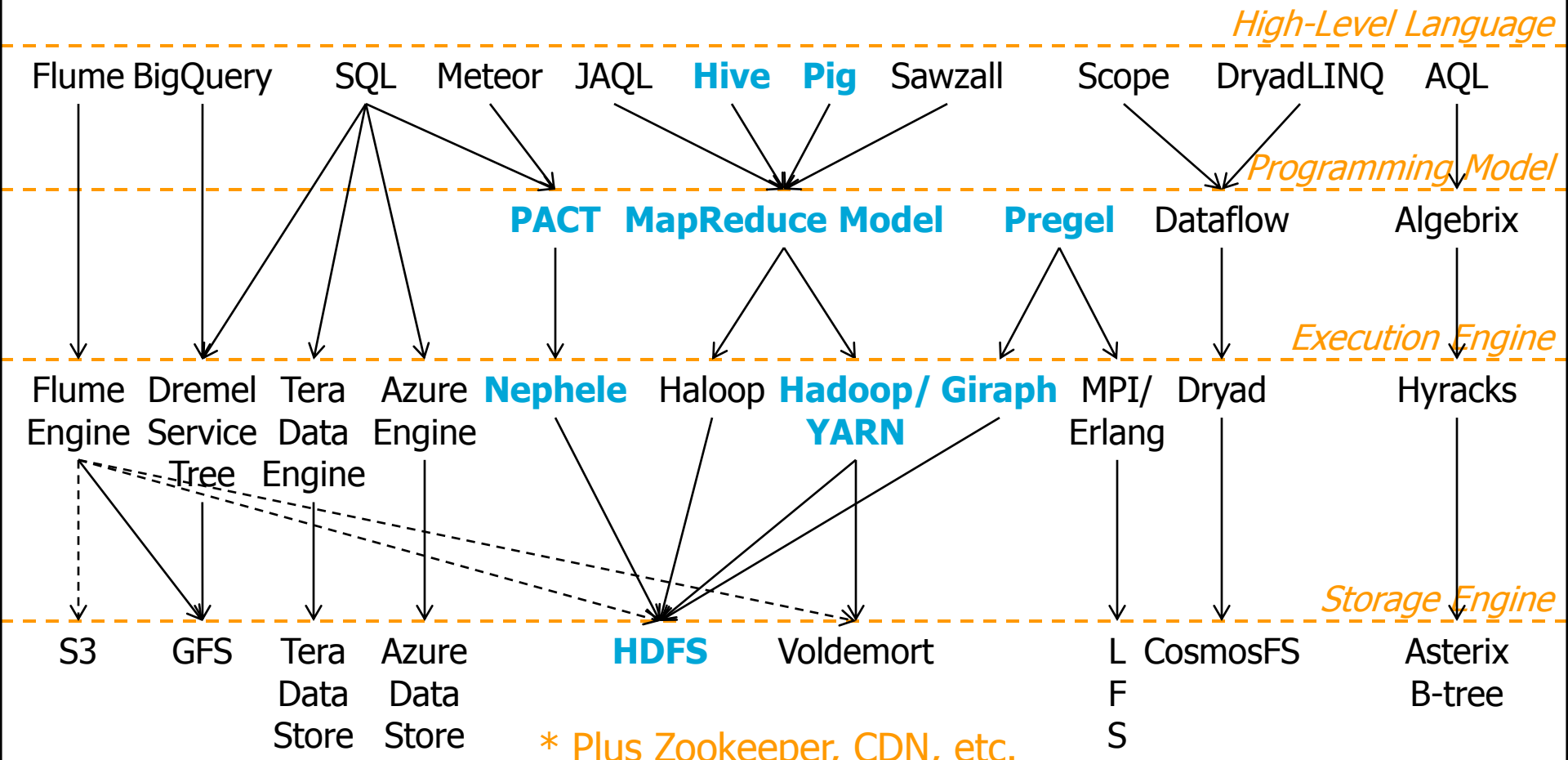
Data Deluge = data generated by humans and devices (IoT)

- Interacting
- Understanding
- Deciding
- Creating

Sources: IDC, EMC.

TU Delft

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

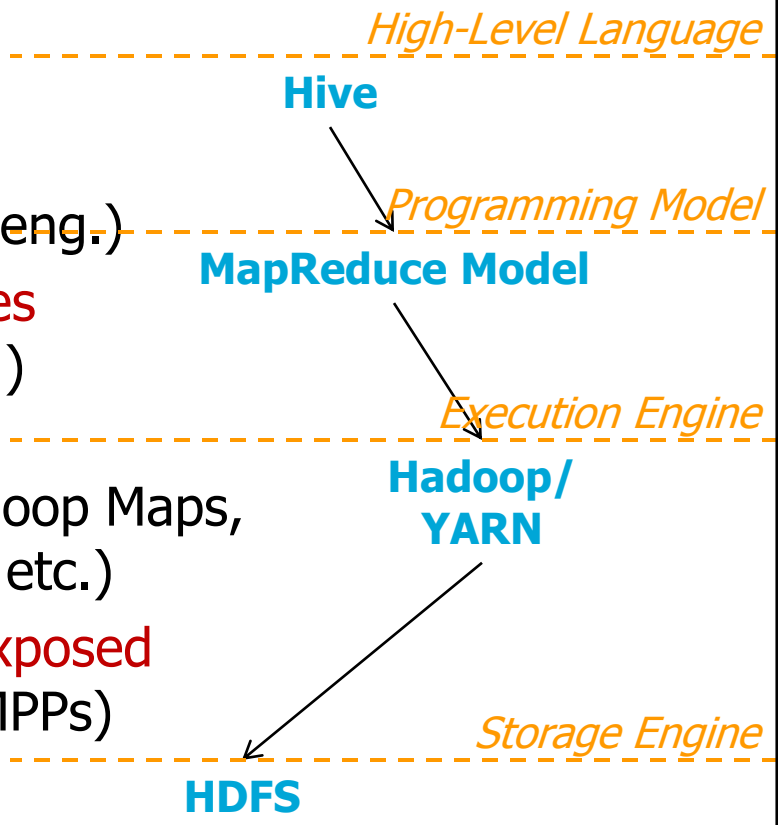


Adapted from: Dagstuhl Seminar on Information Management in the Cloud, http://www.dagstuhl.de/program/calendar/partlist/?semnr=11321&SUOG

TU Delft

Can We Afford This Vision? The Current Tech Monolithic Systems

- Monolithic
  - Integrated stack
    (can still learn from decades of sw.eng.)
  - Fixed set of homogeneous resources
    (we forgot 2 decades of distrib.sys.)
  - Execution engines do not coexist
    (we're running now MPI inside Hadoop Maps, Hadoop jobs inside MPI processes, etc.)
  - Little performance information is exposed
    (we forgot 4 decades of par.sys., MPPs)
  - …

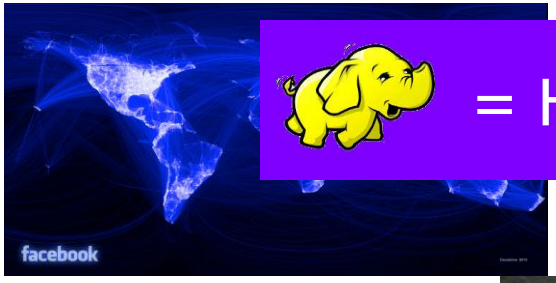


Stuck in stacks!

A. L. Varbanescu and A. Iosup, On Many-Task Big Data Processing: from GPUs to Clouds. Proc. of SC|12 (MTAGS).
http://www.pds.ewi.tudelft.nl/~iosup/many-tasks-big-data-vision13mtags_v100.pdf

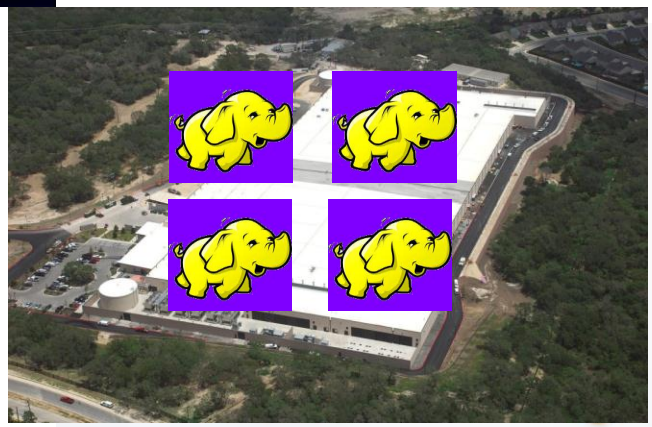
The "Big Data cake" in the Data Center

Online Social Networks



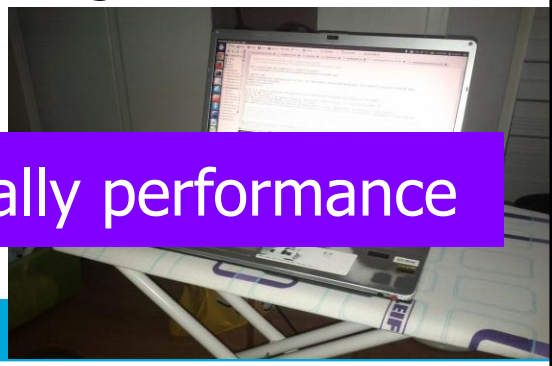
Financial Analysts

= Hadoop / MapReduce framework



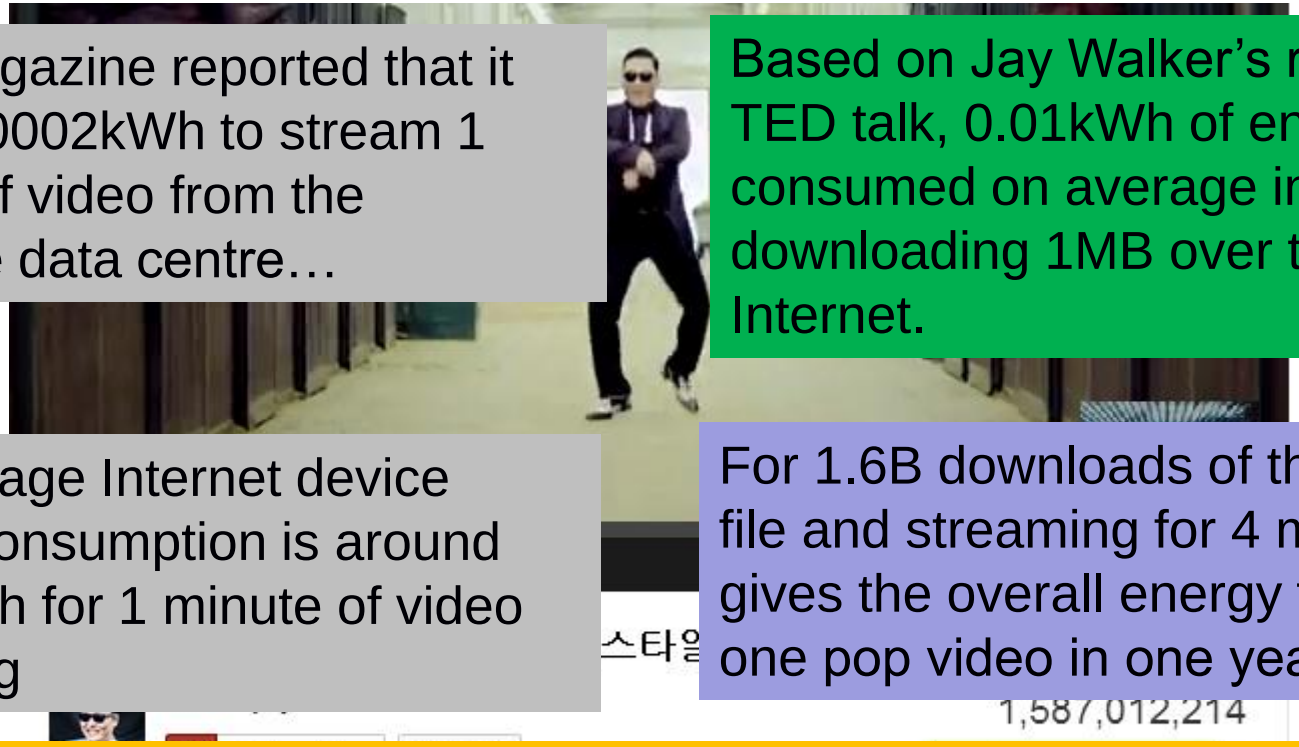
Universe Explorers

Big Data Enthusiast



Multiple frameworks = Isolation, especially performance

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…

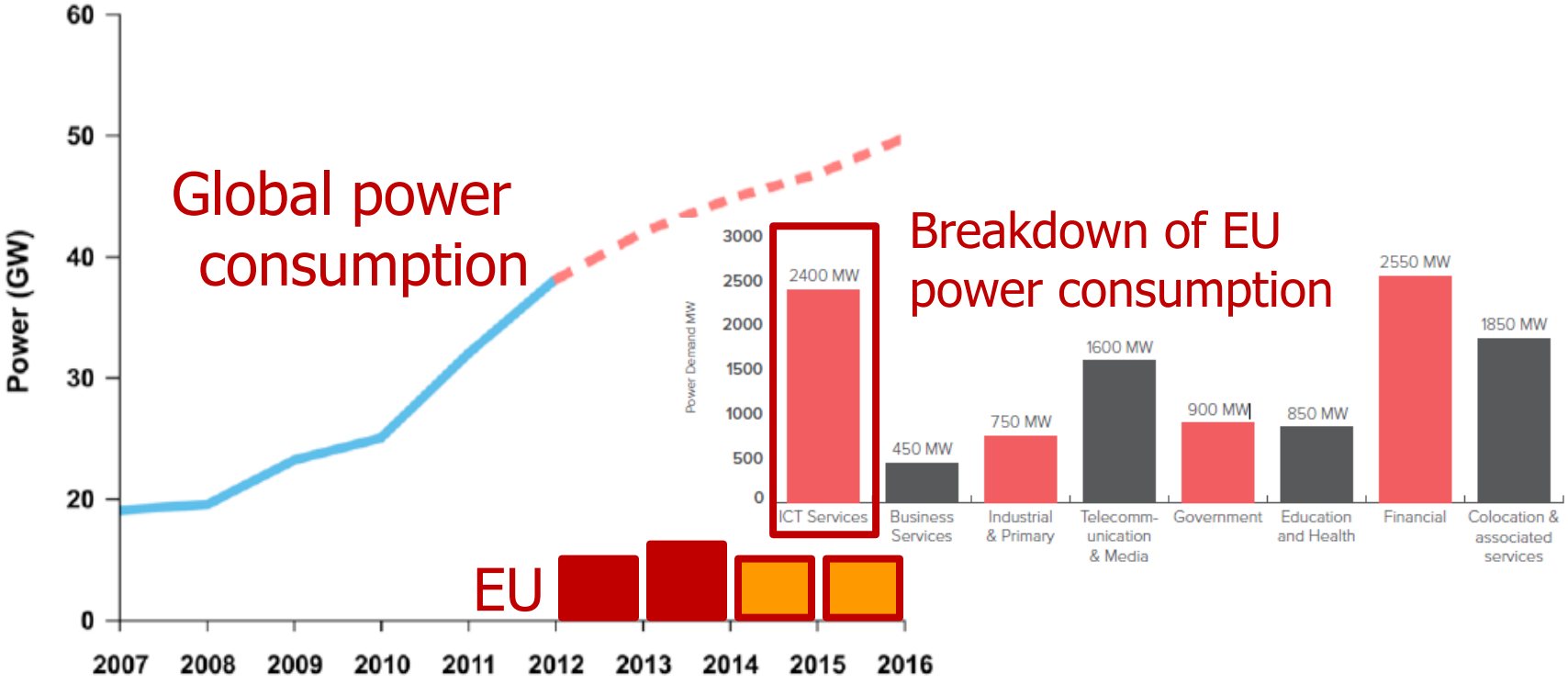
>300GWh = more than some countries in a year, >35MW of 24/7/365 diesel, >100M liters of oil, 80,000 cars running for a year, ...

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

TU Delft

Can We Afford This Vision? Not with the Current Resources

- Energy resources



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.

TU Delft

Everyone is a Scientist!
Can We Afford This Vision?

I don't know.
But we need to become very efficient.
For this, we need to combine sw.eng., distr.sys., parallel sys., DB.
Then, we need to show numbers!

Why Big Data Benchmarking?

- Establish and share best-practices in giving quantitative answers to important questions about Big Data

- Use in procurement
- Use in system design
- Use in system tuning and operation
- Use in performance management

- **Use in training**

Big Data in the Data Center: 10 Main Challenges in 4 Categories*

* List not exhaustive

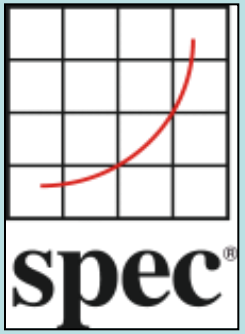
- **Methodological**
  1. Experiment compression, both design and actual evaluation
  2. Beyond black-box testing through testing short-term dynamics and long-term evolution
  3. Impact of middleware
- **System-Related**
  1. Reliability, availability, and system-related properties
  2. Massive-scale, multi-site benchmarking
  3. Performance isolation, multi-tenancy models
- **Workload-related**
  1. Workload = Dataset + Activity
  2. Statistical workload models + analysis of coverage
  3. Benchmarking performance isolation under various multi-tenancy workloads
- **Metric-Related**
  1. Beyond traditional performance: variability, elasticity, cost, etc.
  2. Uniform reporting

Iosup et al., IaaS Cloud Benchmarking: Approaches, Challenges, and Experience, MTAGS 2012.

Guo, Varbanescu, Iosup, Martella, Willke. Benchmarking Graph-Processing Performs: A Vision, ICPE WiP 2014.

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

TU Delft
Delft University of Technology

Find more information on: *http://research.spec.org*

A Call to Arms

- Defining workloads
- Understanding the metrics, datasets, and algorithms used in practice: fill in our survey http://goo.gl/TJwkTg

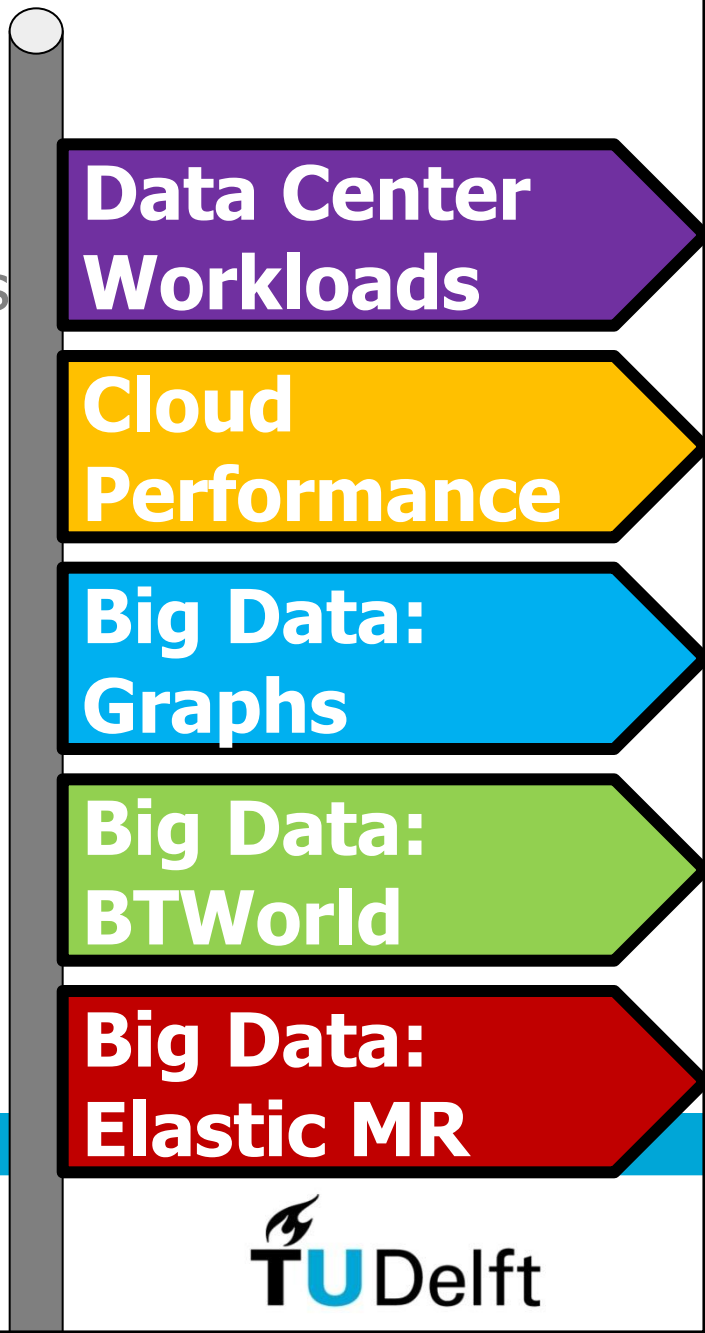
- Evaluating and reporting on various platforms

Join us within the SPEC RG Cloud Working Group

http://research.spec.org/working-groups/rg-cloud-working-group.html

Agenda

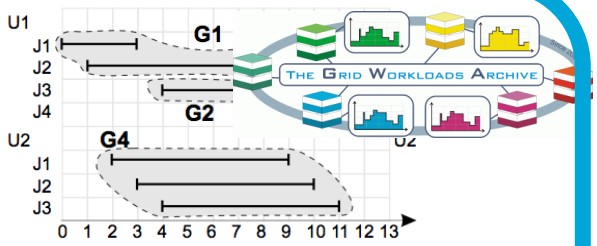
1. Everyone is a scientist!
2. Benchmarking: let's show the numbers
3. **Datacenter Workloads**
4. **Cloud Performance & Perf. Variability**
5. **Performance of Graph-Processing Platforms (Giraph, GraphLab, …)**
6. **BitTorrent World: A MapReduce Workflow**
7. **Elastic MapReduce Performance**
8. Conclusion



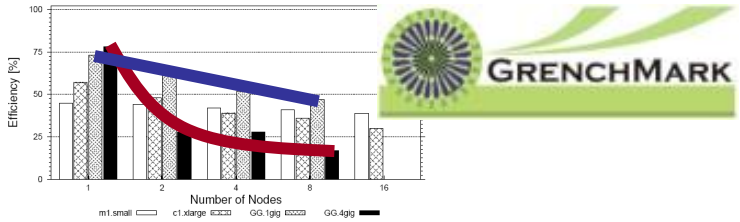
TU Delft

Agenda

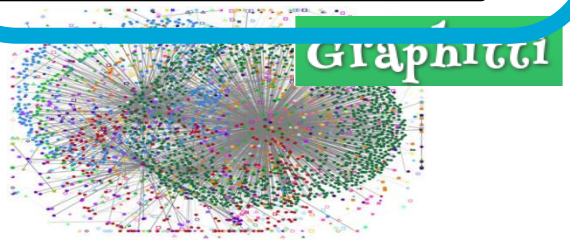
1. Everyone is a scientist!
2. Benchmarking: let's show the numbers



Data Center Workloads



Cloud Performance



Big Data: Graphs

Big Data: BTWorld

Big Data: Elastic MR

3. Conclusion

TU Delft

Data Center Workloads: Our Team

Alexandru Iosup
TU Delft

BoTs
Workflows
Big Data
Statistical modeling

Dick Epema
TU Delft

BoTs
Grids

Mathieu Jan
TU Delft/INRIA

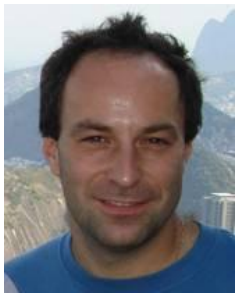
BoTs
Statistical modeling

Ozan Sonmez
TU Delft

BoTs

Thomas de Ruiter
TU Delft

MapReduce
Big Data
Statistical modeling



Radu Prodan
U.Isbk.

Workflows

Thomas Fahringer
U.Isbk.

Workflows

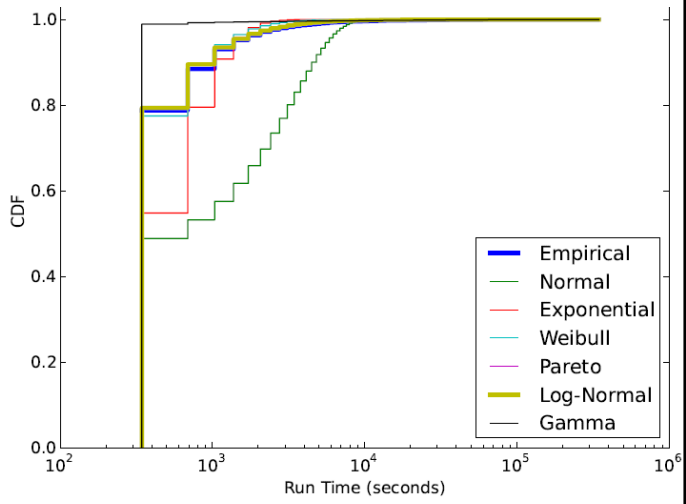
Simon Ostermann
U.Isbk.

Workflows

TU 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)

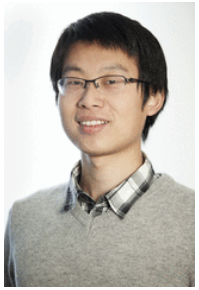


Model	Tasks	Correlation	Map/Reduce Modeled	Sign. Level	Indirect Distr. Sel.
Complex Model	Indirect	Run time – Disk	Separately	0.05	Best fits
Relaxed Complex Model	Indirect	Run time – Disk	Separately	0.02	All fits
Safe Complex Model	Direct	Run time – Disk	Separately	0.05	–
Simple Model	Direct	–	Together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

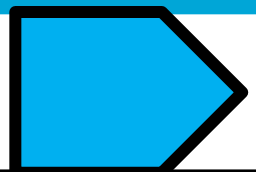
de Ruiter and Iosup. A workload model for MapReduce. MSc thesis at TU Delft. Jun 2012. Available online via TU Delft Library, http://library.tudelft.nl .

Graph Processing Workloads

- No representative workloads, perhaps even algorithm coverage is difficult to analyze
- See work on graph processing



Yong Guo
TU Delft
Graph processing
Benchmarking



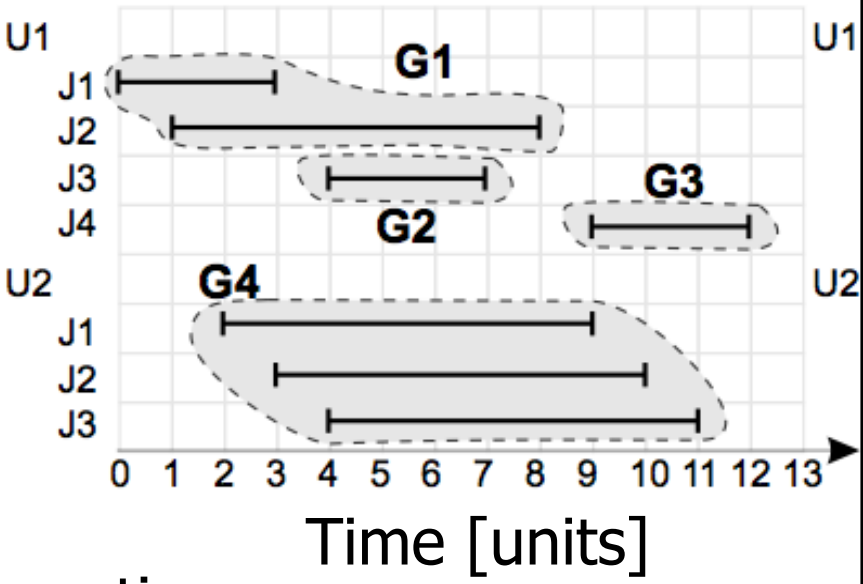
What is a Bag of Tasks (BoT)? A System View

BoT = set of jobs sent by a user…

$$W_u = \{J_i | user(J_i) = u\}$$

…that is submitted at most Δs after the first job

$$ST(J') \leq ST(J) + \Delta$$



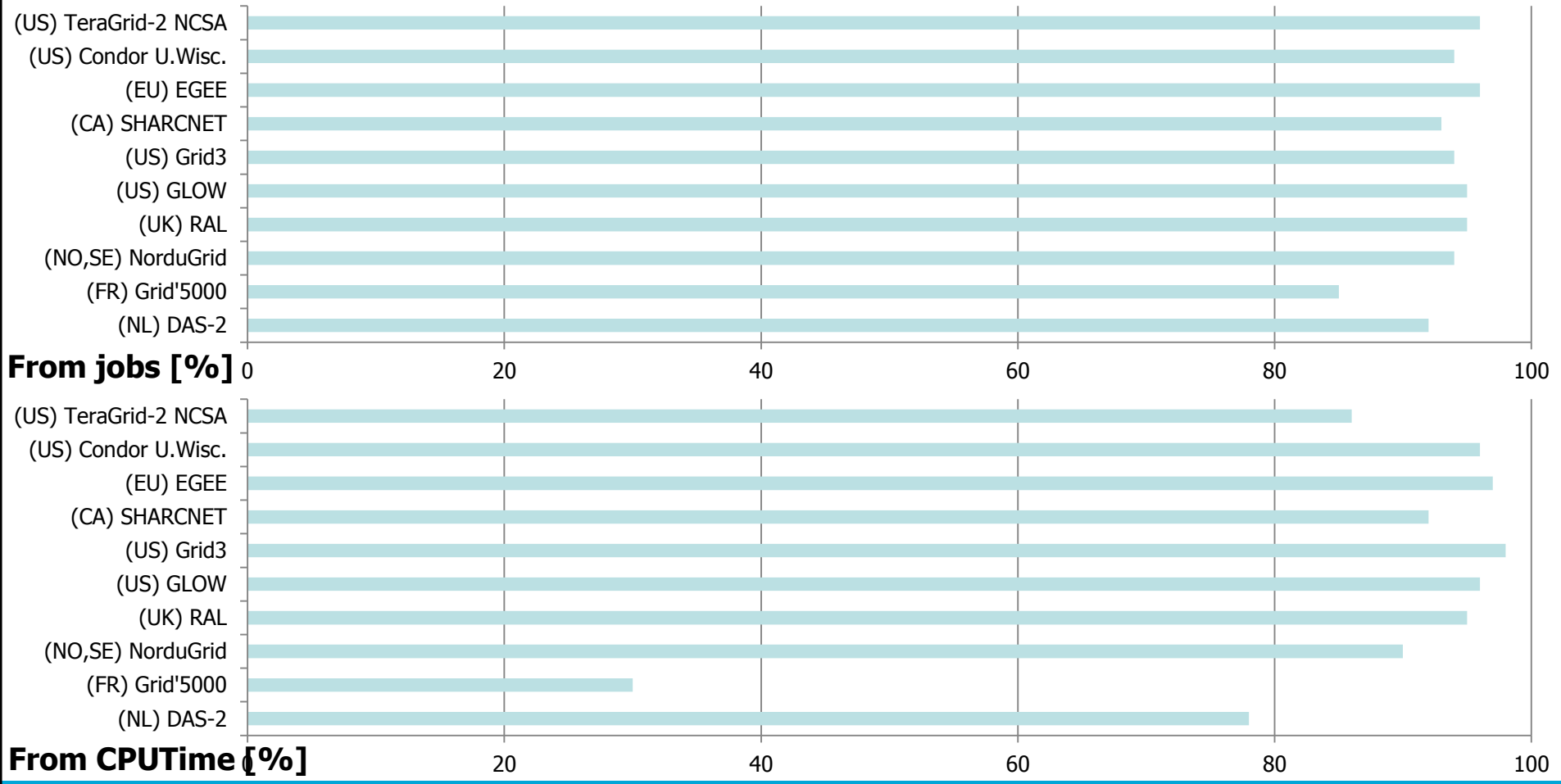
- Why Bag of *Tasks*? From the perspective of the user, jobs in set are just tasks of a larger job
- A single useful result from the complete BoT
- Result can be combination of all tasks, or a selection of the results of most or even a single task

Iosup et al., The Characteristics and Performance of Groups of Jobs in Grids, Euro-Par, LNCS, vol.4641, pp. 382-393, 2007.

Applications of the BoT Programming Model

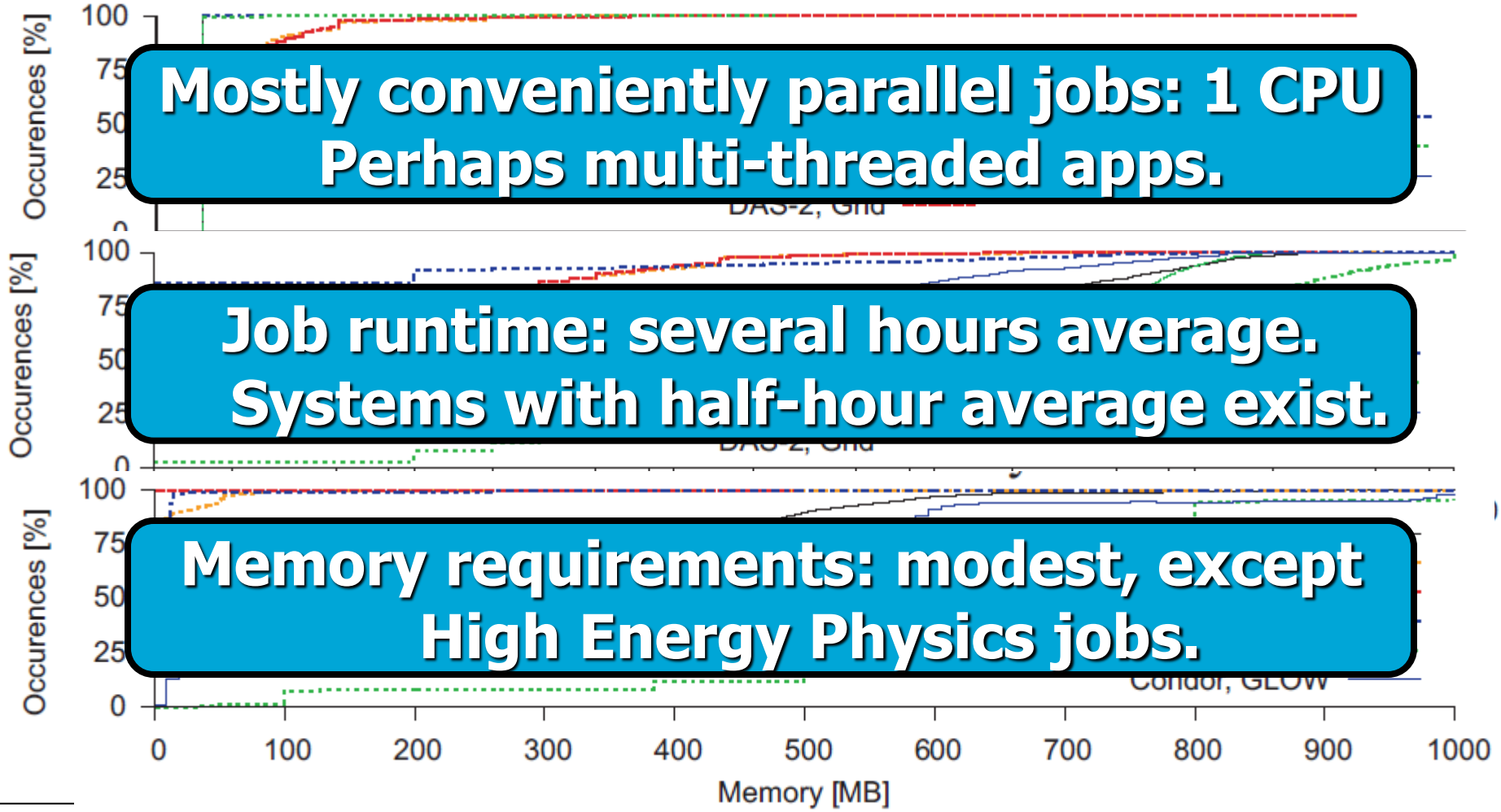
- Parameter sweeps
  - Comprehensive, possibly exhaustive investigation of a model
  - Very useful in engineering and simulation-based science
- Monte Carlo simulations
  - Simulation with random elements: fixed time yet limited inaccuracy
  - Very useful in engineering and simulation-based science
- Many other types of batch processing
  - Periodic computation, Cycle scavenging
  - Very useful to automate operations and reduce waste

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)

BoTs by Numbers: CPUs, Runtime, Mem



Iosup et al., The Grid Workloads Archive, FGCS, 2008.

Iosup and Epema, Grid Computing Workloads, IEEE Internet Computing, 2011.

Actual numbers.

BoTs by numbers: I/O, Files, Remote Sys

T-12	I/O [KOps]				I/O Traffic [MB]		
	Total	Rd	Wr	Wr %	Total	Rd	Wr %
				20%	469	174	63%
				20%	144	114	21%
				3%	161	130	19%
					389	33	92%
					330	31	91%

I/O: modest, except HEP

Rd:Wr varies widely

I/O,HEP: 65MBps/experiment

Upper bound for typical sci.apps.

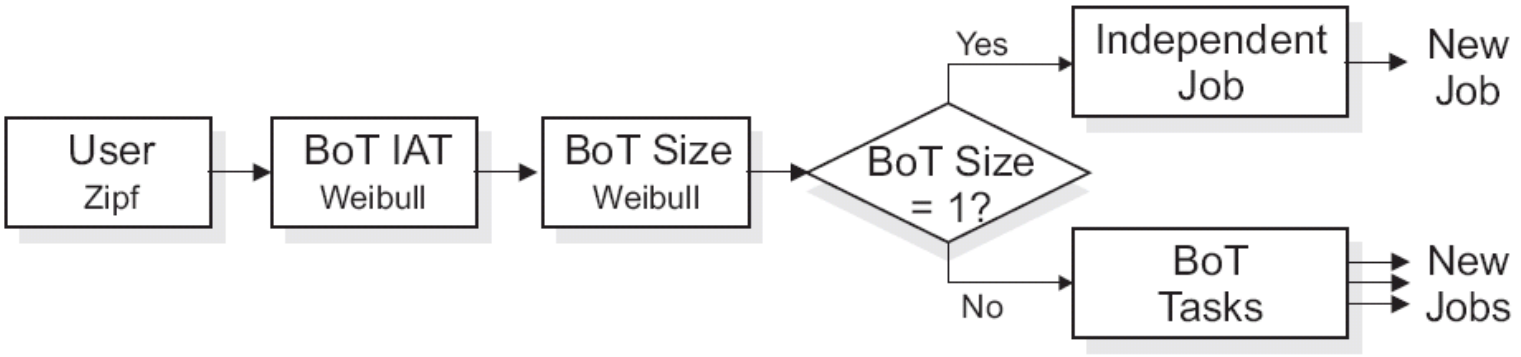
T-12	File Transfer [MB]				Remote Sys. Calls [MB]			
part	Total	In	In / Out %		Total	In	In / Out %	
t1	10,865	8,259	76%	24%	28	16	59%	41%
t2	1,736	1,542	89%	11%	71	28	40%	60%
								58%
								0%
					44	40	91%	9%

Remote Sys.: small Xfers, latency important

Netw: 2-10GB, input mostly

Iosup and Epema, Grid Computing Workloads, IEEE Internet Computing, 2011.

Statistical BoT Workload Model



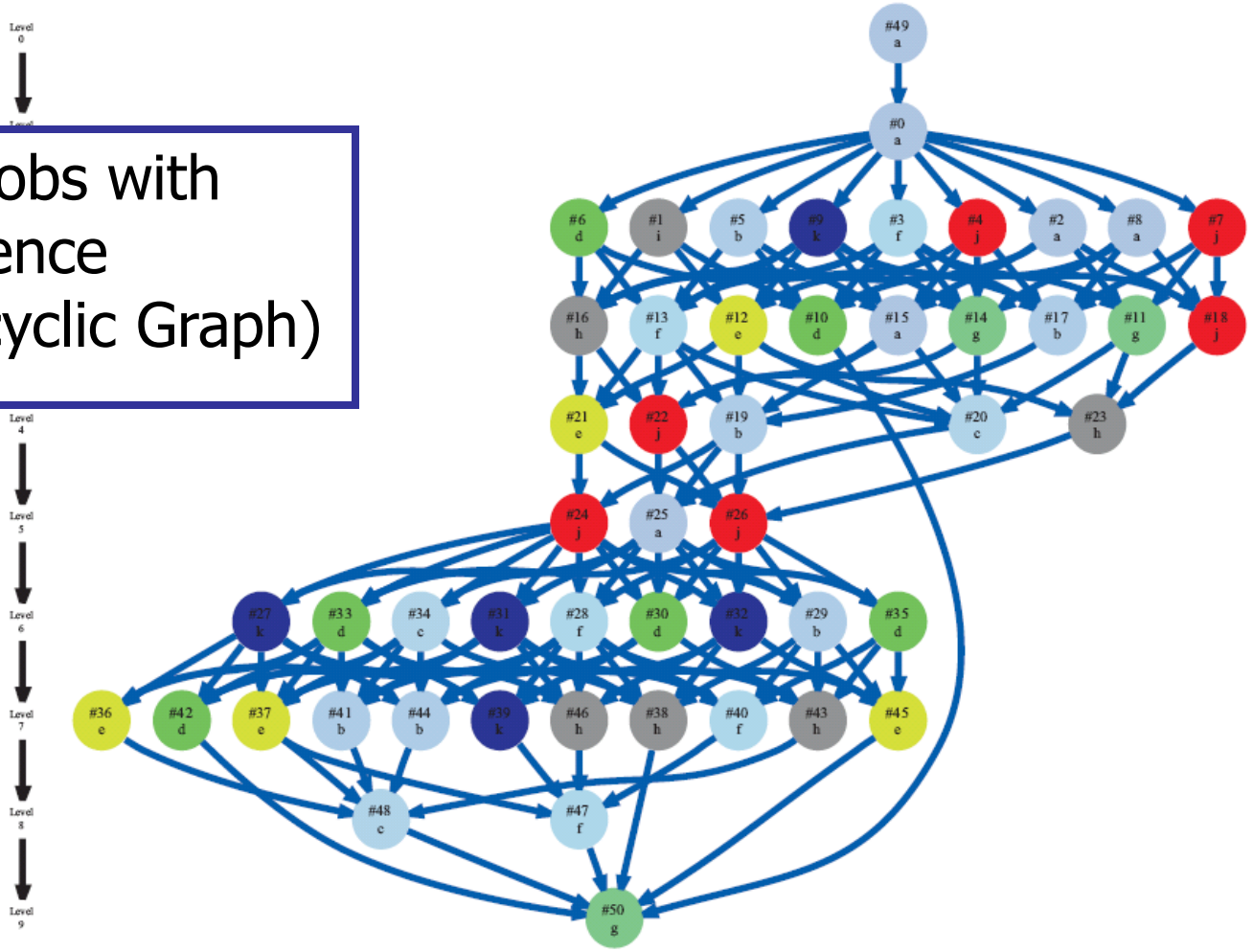
- Single arrival process for both BoTs and parallel jobs
- Validated with 7 grid workloads

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.

TU Delft

What is a Wokflow?

WF = set of jobs with precedence
(think Direct Acyclic Graph)



TU Delft Q0

Applications of the Workflow Programming Model

- Complex applications
  - Complex filtering of data
  - Complex analysis of instrument measurements
- Applications created by non-CS scientists*
  - Workflows have a natural correspondence in the real-world, as descriptions of a scientific procedure
  - Visual model of a graph sometimes easier to program
- Precursor of the MapReduce Programming Model (next slides)

*Adapted from: Carole Goble and David de Roure, Chapter in "The Fourth Paradigm", http://research.microsoft.com/en-us/collaboration/fourthparadigm/

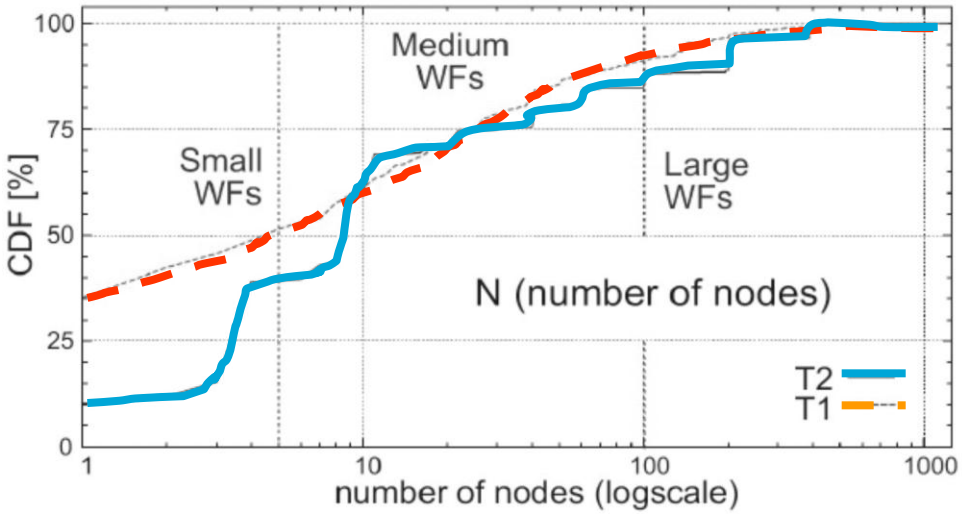
Workflows Exist in Grids, but Did No Evidence of a Dominant Programming Model

- 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

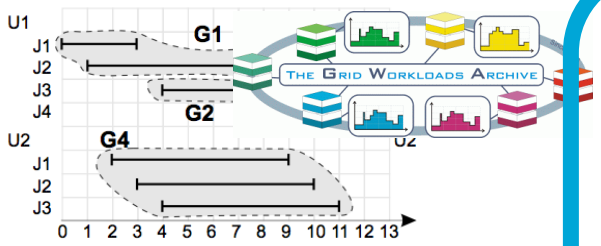
Loose coupling
- Graph with 3-4 levels
- Average WF ~10s of jobs
- 75% WFs are <=40 jobs
  95% are <=200 jobs
- 85% WFs take <10 mins

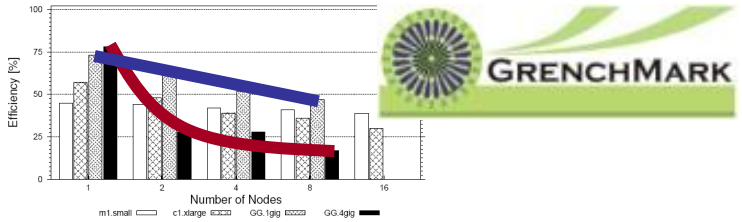


Ostermann et al., On the Characteristics of Grid Workflows, CoreGRID Integrated Research in Grid Computing (CGIW), 2008.

Agenda

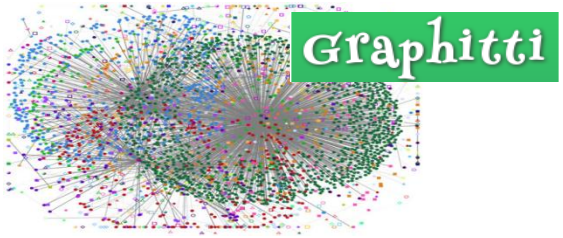
1. Everyone is a scientist!
2. Benchmarking: let's show the numbers

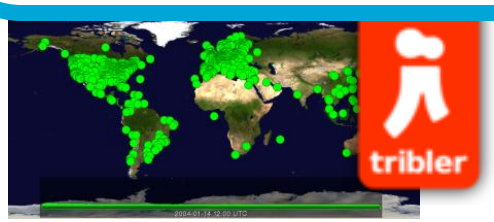




Data Center Workloads

Cloud Performance





Big Data: Graphs

Big Data: BTWorld

Big Data: Elastic MR

3. Conclusion

TU Delft

Cloud Performance and Performance Variability: Our Team

Alexandru Iosup
TU Delft

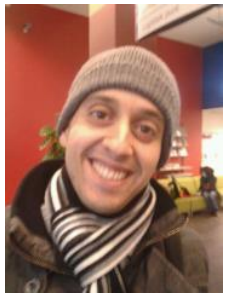
Performance
Variability
Isolation
Multi-tenancy
Benchmarking

Dick Epema
TU Delft

Performance
IaaS clouds

Nezih Yigitbasi
TU Delft

Performance
Variability



Athanasios Antoniou
TU Delft

Performance
Isolation

Radu Prodan
U.Isbk.

Benchmarking

Thomas Fahringer
U.Isbk.

Benchmarking

Simon Ostermann
U.Isbk.

Benchmarking

TU Delft

Some Previous Work (>50 important references across our studies)

Virtualization Overhead

- Loss below 5% for computation [Barham03] [Clark04]
- Loss below 15% for networking [Barham03] [Menon05]
- Loss below 30% for parallel I/O [Vetter08]
- Negligible for compute-intensive HPC kernels [You06] [Panda06]

Cloud Performance Evaluation

- Performance and cost of executing a sci. workflows [Dee08]
- Study of Amazon S3 [Palankar08]
- Amazon EC2 for the NPB benchmark suite [Walker08] or selected HPC benchmarks [Hill08]
- CloudCmp [Li10]
- Kosmann et al.

August 6, 2014

Q1

Production IaaS Cloud Services

- **Production IaaS cloud:** lease resources (infrastructure) to users, operate on the market and have active customers

Name	Cores (ECUs)	RAM [GB]	Archi. [bit]	Disk [GB]	Cost [$/h]
Amazon EC2					
m1.small	1 (1)	1.7	32	160	0.1
m1.large	2 (4)	7.5	64	850	0.4
m1.xlarge	4 (8)	15.0	64	1,690	0.8
c1.medium	2 (5)	1.7	32	350	0.2
c1.xlarge	8 (20)	7.0	64	1,690	0.8
GoGrid (GG)					
GG.small	1	1.0	32	60	0.19
GG.large	1	1.0	64	60	0.19
GG.xlarge	3	4.0	64	240	0.76
Elastic Hosts (EH)					
EH.small	1	1.0	32	30	£0.042
EH.large	1	4.0	64	30	£0.09
Mosso					
Mosso.small	4	1.0	64	40	0.06
Mosso.large	4	4.0	64	160	0.24

August 6, 2014

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

Our Method

Q1

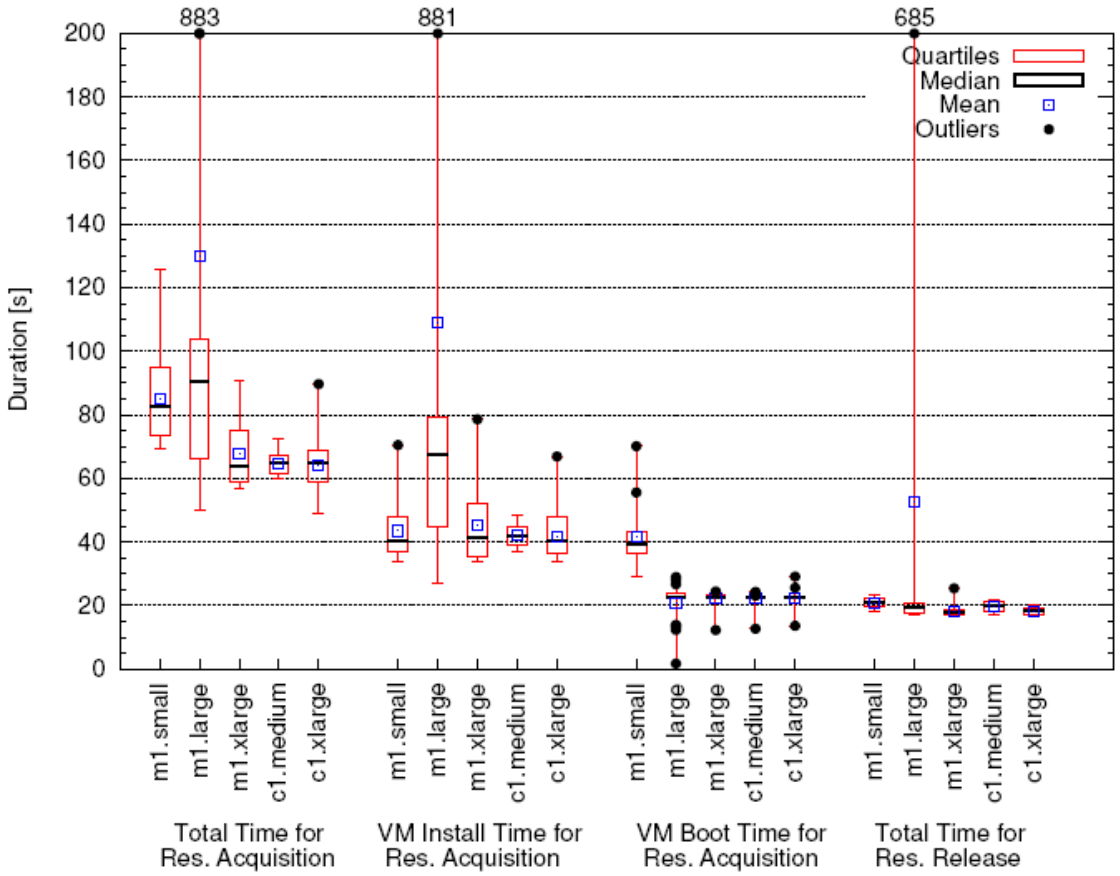
- Based on general performance technique: model performance of individual components; system performance is performance of workload + model [Saavedra and Smith, ACM TOCS'96]
- Adapt to clouds:
  1. Cloud-specific elements: resource provisioning and allocation
  2. Benchmarks for single- and multi-machine jobs
  3. Benchmark CPU, memory, I/O, etc.:

Type	Suite/Benchmark	Resource	Unit
SI	LMbench/all [24]	Many	Many
SI	Bonnie/all [25], [26]	Disk	MBps
SI	CacheBench/all [27]	Memory	MBps
MI	HPCC/HPL [28], [29]	CPU	GFLOPS
MI	HPCC/DGEMM [30]	CPU	GFLOPS
MI	HPCC/STREAM [30]	Memory	GBps
MI	HPCC/RandomAccess [31]	Network	MUPS
MI	HPCC/b_{eff}(lat.,bw.) [32]	Comm.	μs, GBps

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

Single Resource Provisioning/Release

Q1

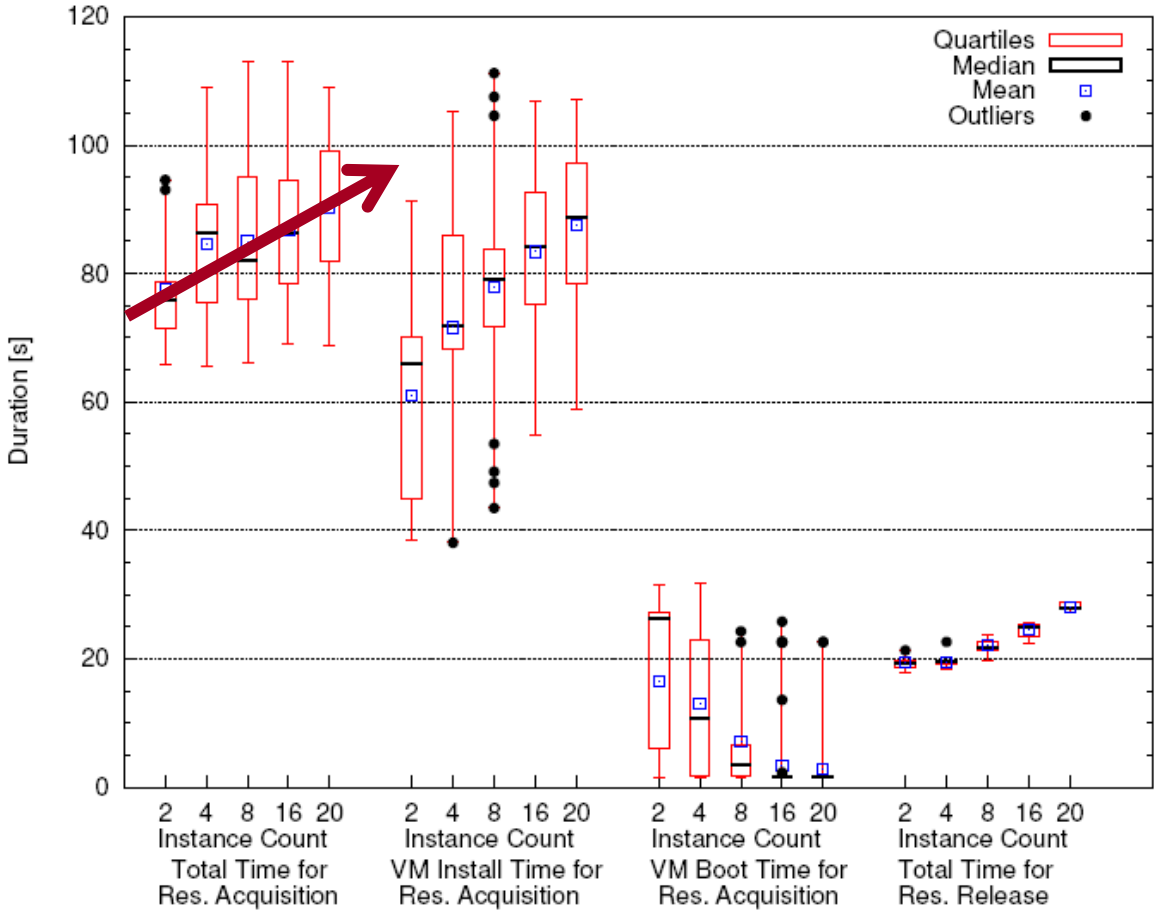


- Time depends on instance type
- Boot time non-negligible

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

Multi-Resource Provisioning/Release

Q1



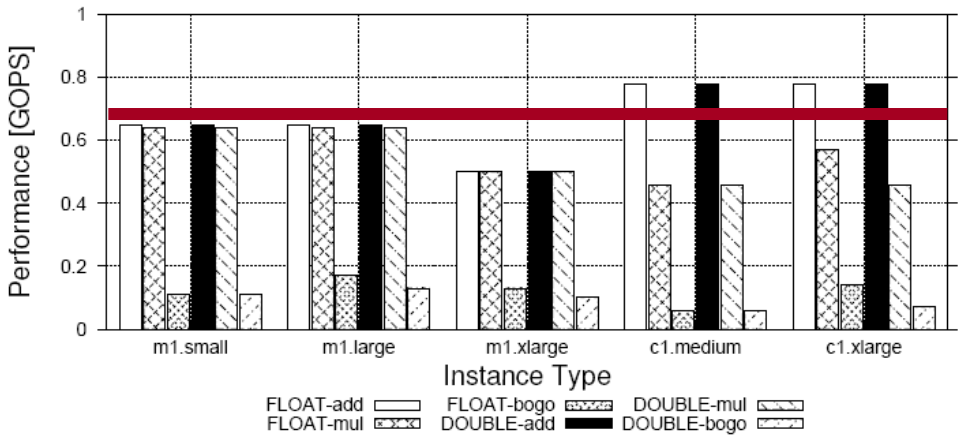
- Time for ***multi***-resource increases with number of resources

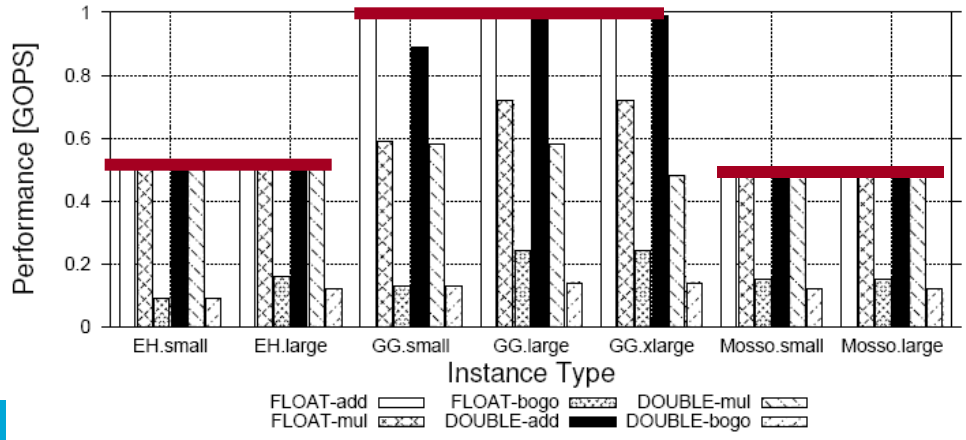
Iosup et al., Performance Analysis of Cloud Computing Services for Many Tasks Scientific Computing, (IEEE TPDS 2011).

CPU Performance of Single Resource

Q1

- 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

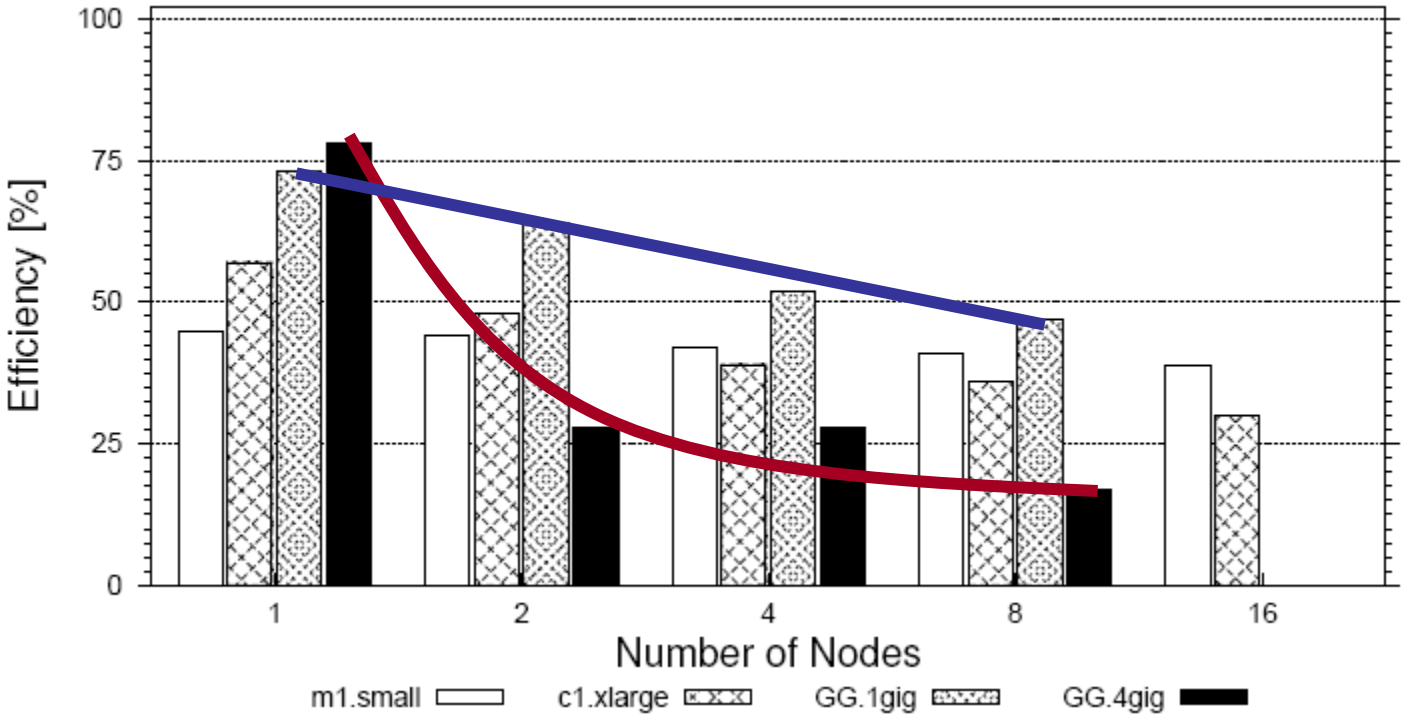




Iosup et al., Performance Analysis of Cloud Computing Services for Many Tasks Scientific Computing, (IEEE TPDS 2011).

HPLinpack Performance (Parallel)

Q1



- Low efficiency for parallel compute-intensive applications
- Low performance vs cluster computing and supercomputing

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

Q2

Production Cloud Services

- **Production cloud:** operate on the market and have active customers

- **IaaS/PaaS:**
  Amazon Web Services (AWS)
  - EC2 (Elastic Compute Cloud)
  - S3 (Simple Storage Service)
  - SQS (Simple Queueing Service)
  - SDB (Simple Database)
  - FPS (Flexible Payment Service)

- **PaaS:**
  Google App Engine (GAE)
  - Run (Python/Java runtime)
  - Datastore (Database) ~ SDB
  - Memcache (Caching)
  - URL Fetch (Web crawling)

August 6, 2014

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

Q2

Our Method Performance Traces [1/3]

- CloudStatus*
  - Real-time values and weekly averages for most of the AWS and GAE services
- Periodic performance probes
  - Sampling rate is under 2 minutes

* www.cloudstatus.com

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

Our Method Analysis [2/3]

Q2

1. Find out whether variability is present
   - Investigate several months whether the performance metric is highly variable

2. Find out the characteristics of variability
   - Basic statistics: the five quartiles (Q_0-Q_4) including the median (Q_2), the mean, the standard deviation
   - Derivative statistic: the IQR (Q_3-Q_1)
   - CoV > 1.1 indicate high variability

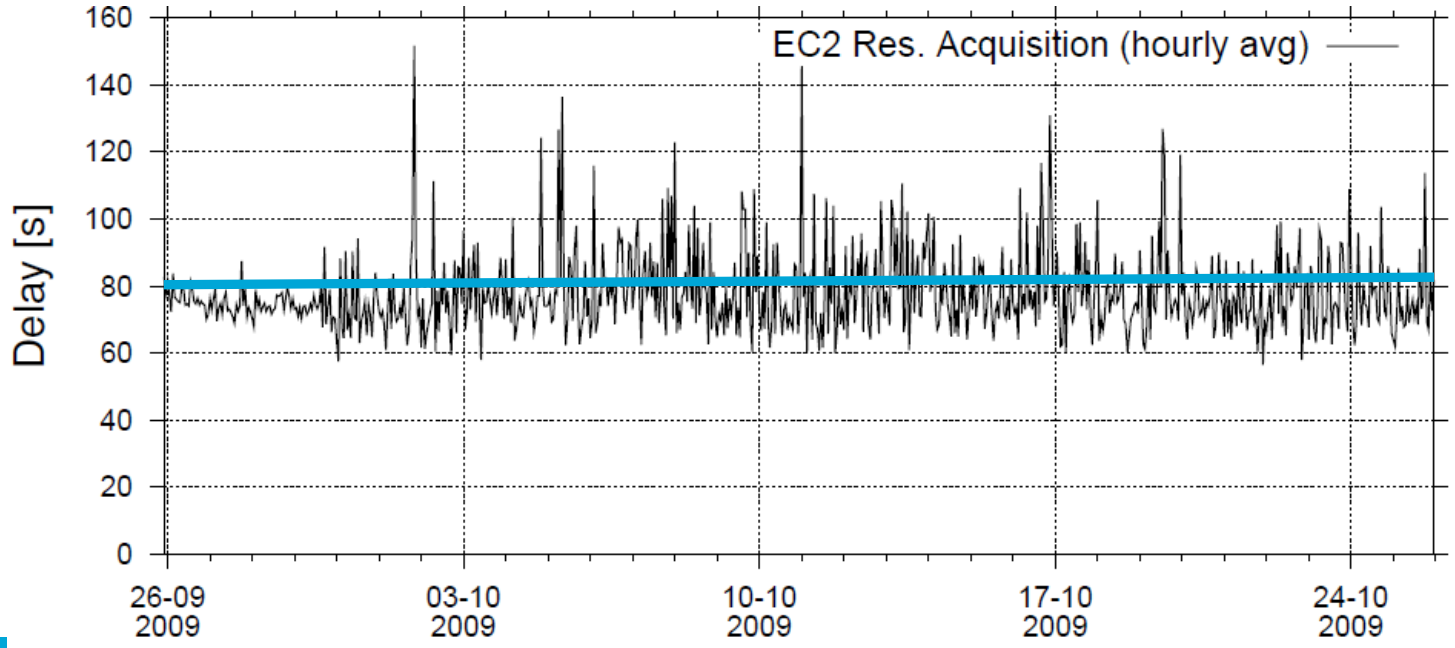
3. Analyze the performance variability time patterns
   - Investigate for each performance metric the presence of daily/monthly/weekly/yearly time patterns
   - E.g., for monthly patterns divide the dataset into twelve subsets and for each subset compute the statistics and plot for visual inspection

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

Q2

Our Method [3/3]
Is Variability Present?

- **Validated Assumption:** The performance delivered by production services is variable.

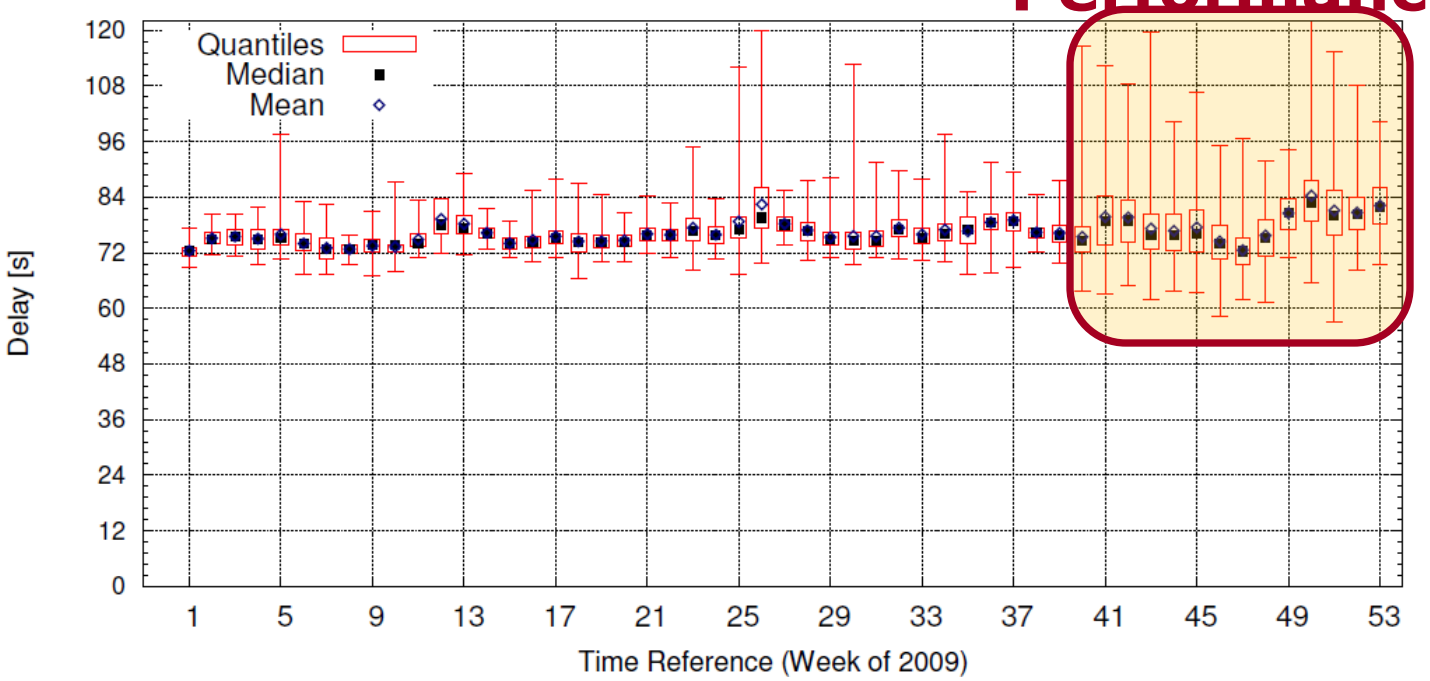


Iosup, Yigitbasi, Epema. On the Performance Variability of Production Cloud Services, (IEEE CCgrid 2011).

AWS Dataset (1/4): EC2

Q2

Variable Performance

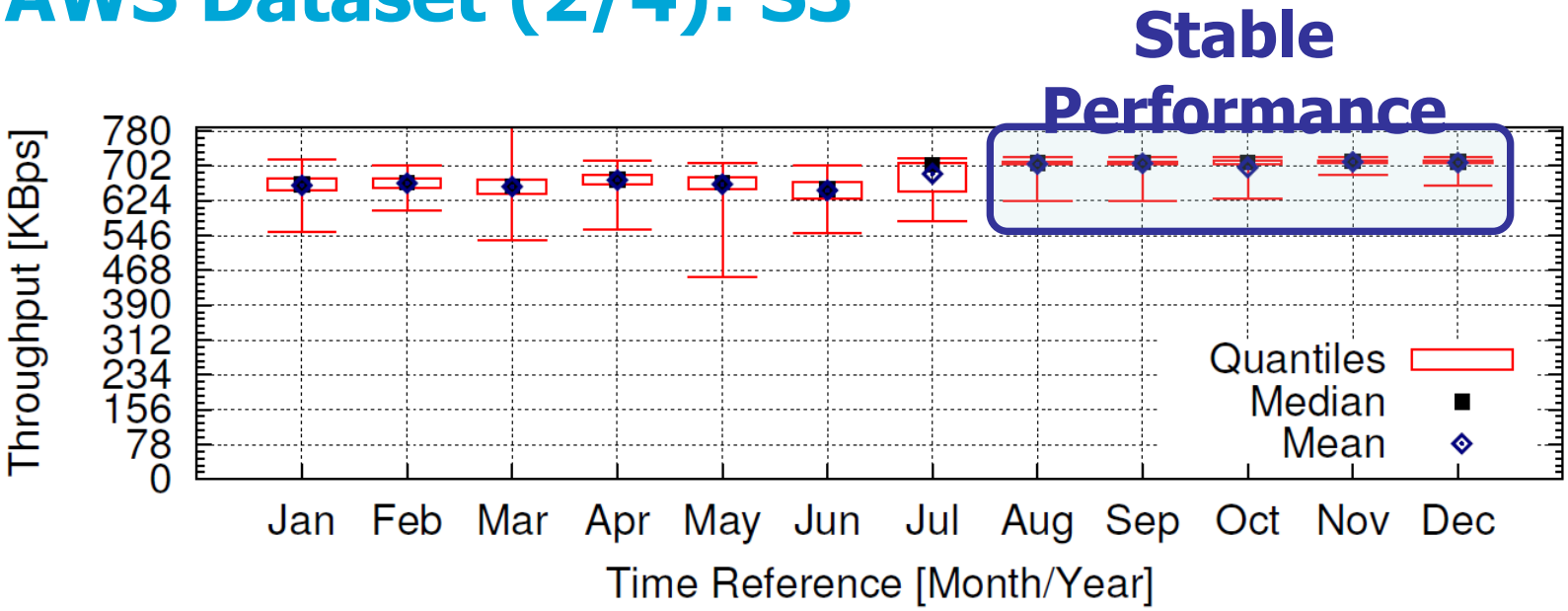


- **Deployment Latency [s]:** Time it takes to start a small instance, from the startup to the time the instance is available
- Higher IQR and range from week 41 to the end of the year; possible reasons:
  - Increasing EC2 user base
  - Impact on applications using EC2 for auto-scaling

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

AWS Dataset (2/4): S3

Q2

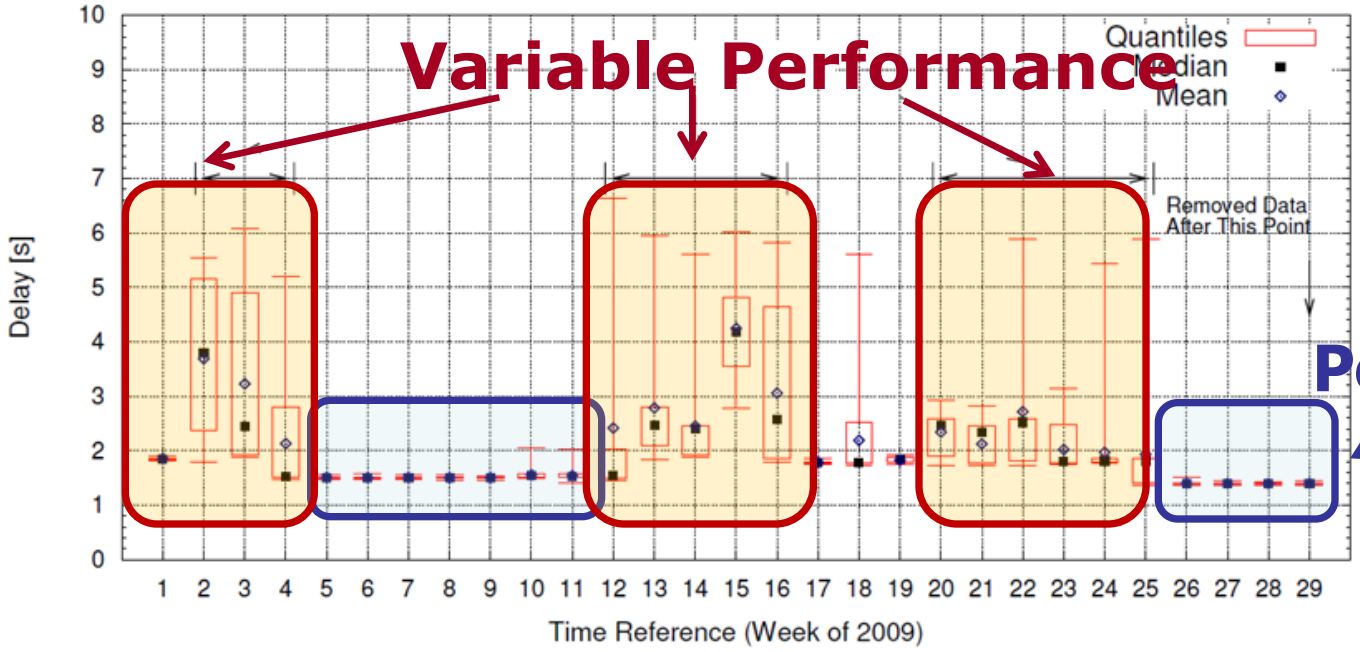


- **Get Throughput [bytes/s]:** Estimated rate at which an object in a bucket is read
- The last five months of the year exhibit much lower IQR and range
  - More stable performance for the last five months
  - Probably due to software/infrastructure upgrades

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

AWS Dataset (3/4): SQS

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

AWS Dataset (4/4): Summary

Q2

- **All services exhibit time patterns in performance**
- EC2: periods of special behavior
- SDB and S3: daily, monthly and yearly patterns
- SQS and FPS: periods of special behavior

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

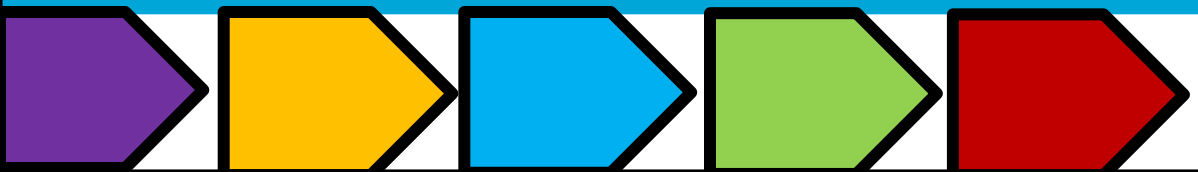
Summary

Q1

- Evaluated several commercial alternatives
- IaaS clouds: lower performance than theoretical peak
  - Especially CPU (GFLOPS)
  - Some users have started to lease opportunistically: lease many machines, retain only machines with best performance
- IaaS and PaaS clouds: high performance variability
  - Difficult to enforce performance guarantees

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

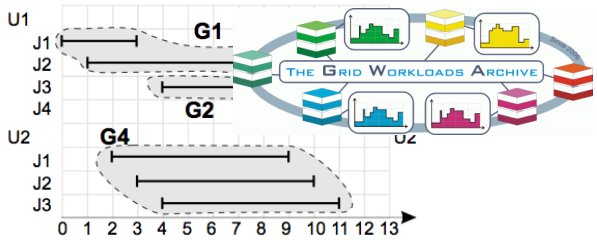
Iosup, Yigitbasi, Epema. On the Performance Variability of Production Cloud Services, (IEEE CCgrid 2011).

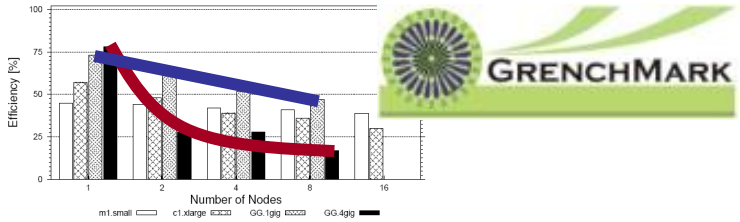


TU Delft

Agenda

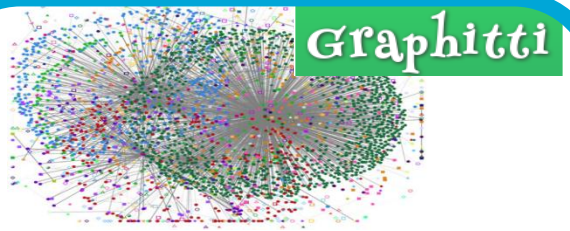
1. Everyone is a scientist!
2. Benchmarking: let's show the numbers





Data Center Workloads

Cloud Performance



Big Data: Graphs

Big Data: BTWorld

Big Data: Elastic MR

3. Conclusion

TU Delft

Graph Processing: Our Team

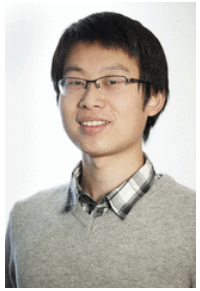
Alexandru Iosup
TU Delft
Big Data & Clouds
Res. management
Systems, Benchmarking

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

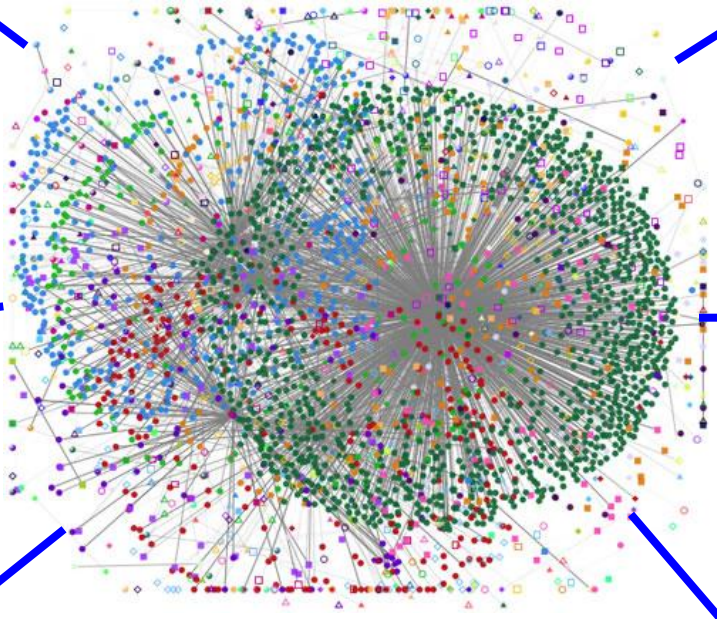
Claudio Martella
VU Amsterdam
Graph processing



Yong Guo
TU Delft
Graph processing
Benchmarking

Marcin Biczak
TU Delft
Big Data & Clouds
Performance & Development

The data deluge: large-scale graphs



LinkedIn

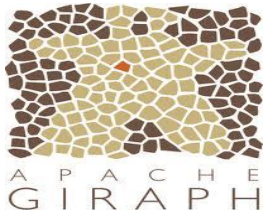
YAHOO!

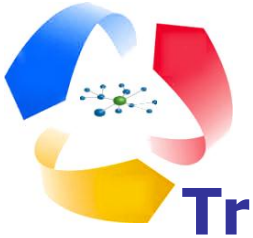
friendster

XFIRE

Platform diversity

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





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

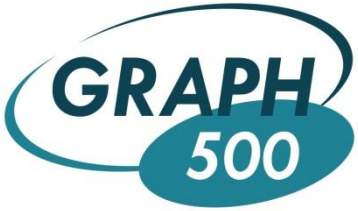
Selection and Design of Performance Metrics for Graph Processing

- Raw processing power
  - Execution time
  - Actual computation time
  - Edges/Vertices per second
- Resource utilization (sys)
  - CPU, memory, network
- Scalability
  - Horizontal vs. vertical
  - Strong vs. weak
- Overhead
  - Data ingestion time
  - Overhead time
- Elasticity (?)

Dataset Selection: Application Domains

- Number of vertices, edges, link density, size, directivity, etc.

	Graphs	#V	#E	d	$\bar{D}$	Directivity
G1	Amazon	262,111	1,234,877	1.8	4.7	directed
G2	WikiTalk	2,388,953	5,018,445	0.1	2.1	directed
G3	KGS	293,290	16,558,839	38.5	112.9	undirected
G4	Citation	3,764,117	16,511,742	0.1	4.4	directed
G5	DotaLeague	61,171	50,870,316	2,719.0	1,663.2	undirected
G6	Synth	2,394,536	64,152,015	2.2	53.6	undirected
G7	Friendster	65,608,366	1,806,067,135	0.1	55.1	undirected



The Game Trace Archive

https://snap.stanford.edu/ http://www.graph500.org/ http://gta.st.ewi.tudelft.nl/

TU Delft

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

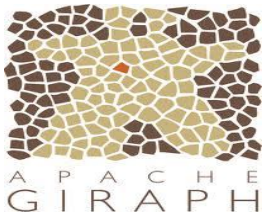
Y. Guo, M. Biczak, A. L. Varbanescu, A. Iosup, C. Martella, and T. L. Willke. How Well do Graph-Processing Platforms Perform? An Empirical Performance Evaluation and Analysis,IPDPS'14.

Platforms we have evaluated

Portability

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

Distributed (Generic)

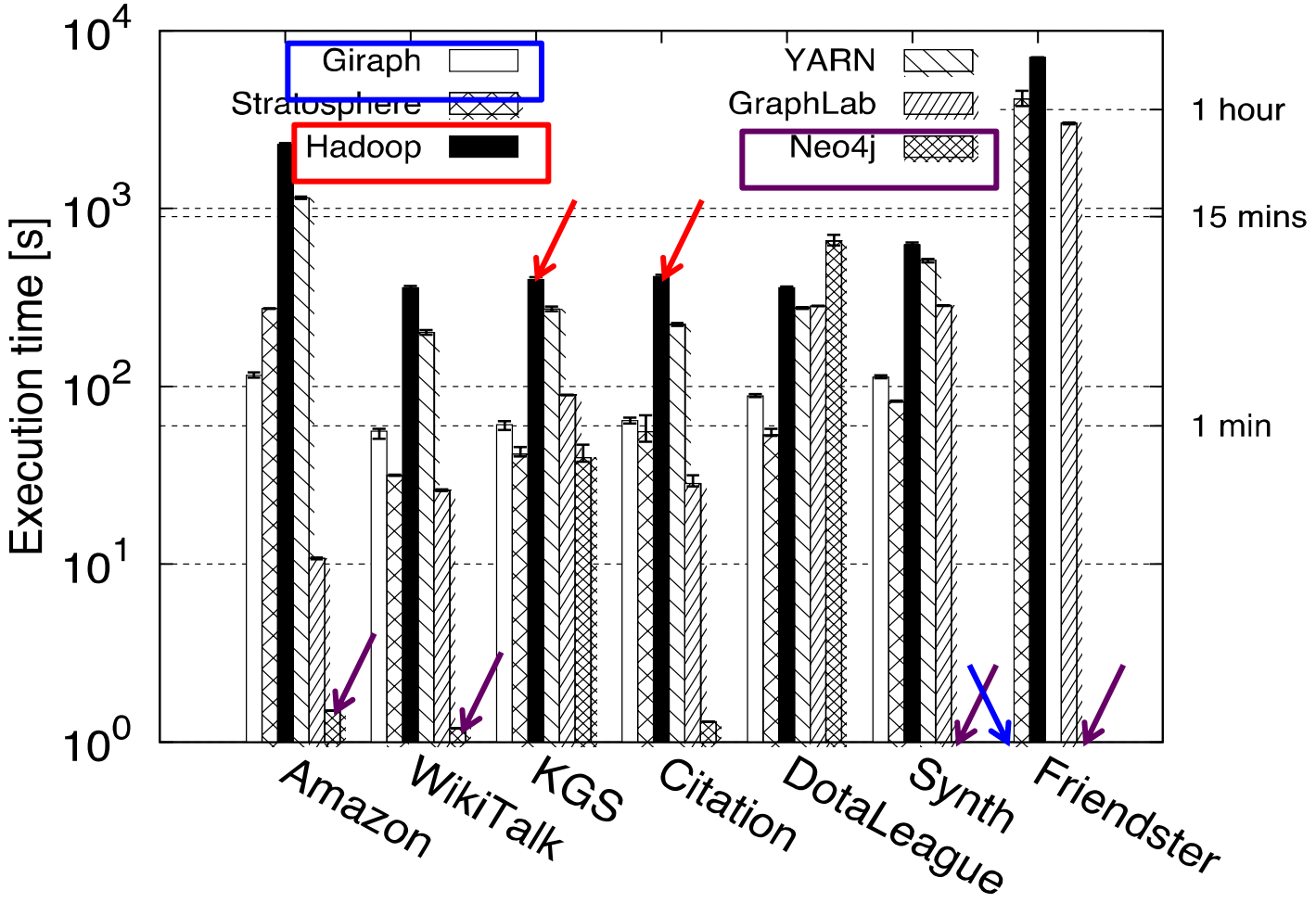


Distributed (Graph-specific)

Non-distributed (Graph-specific)

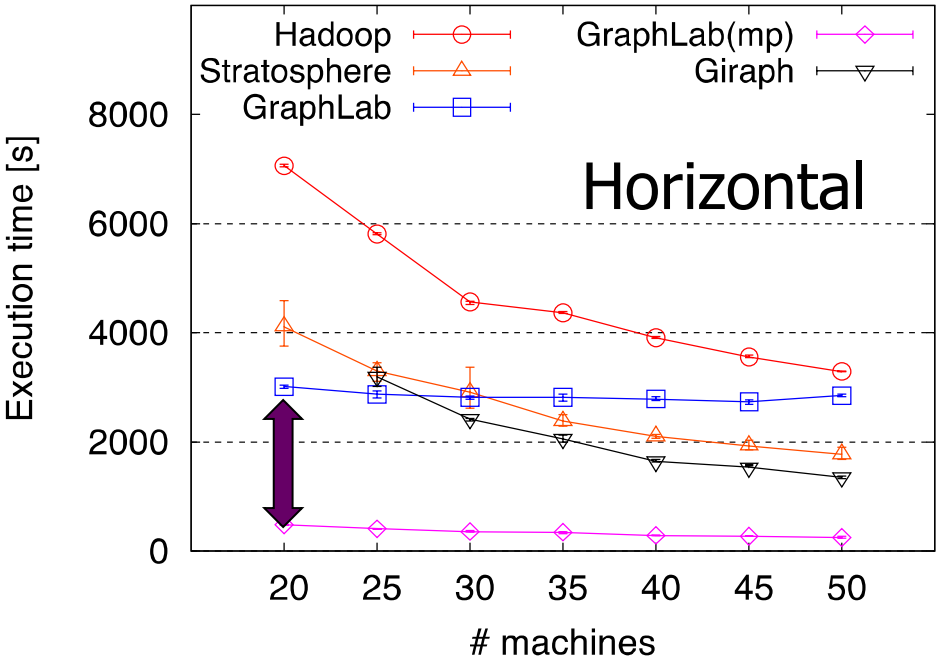
Y. Guo, M. Biczak, A. L. Varbanescu, A. Iosup, C. Martella, and T. L. Willke. How Well do Graph-Processing Platforms Perform? An Empirical Performance Evaluation and Analysis,IPDPS'14.

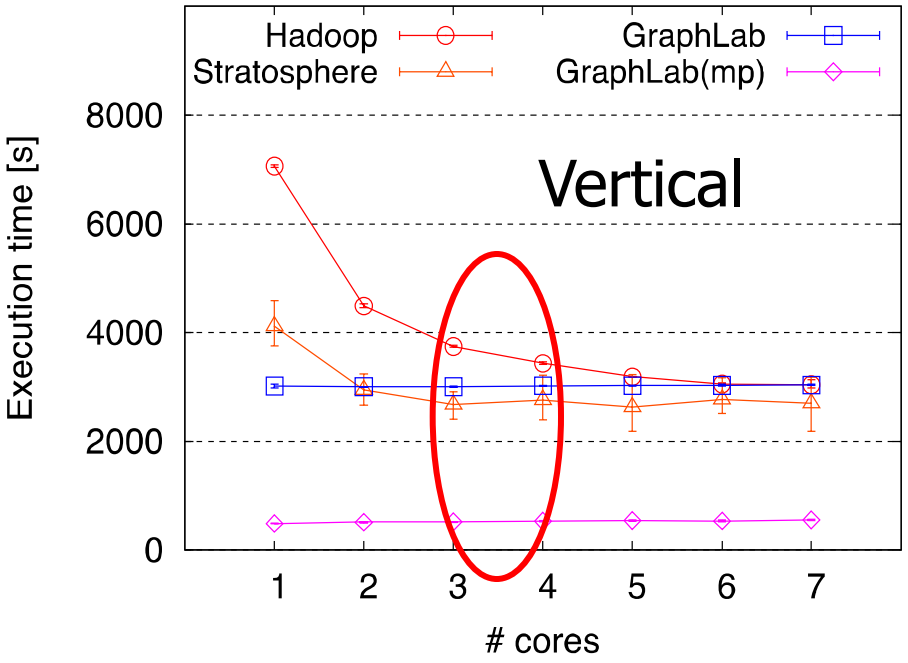
BFS: results for all platforms, all graphs



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.

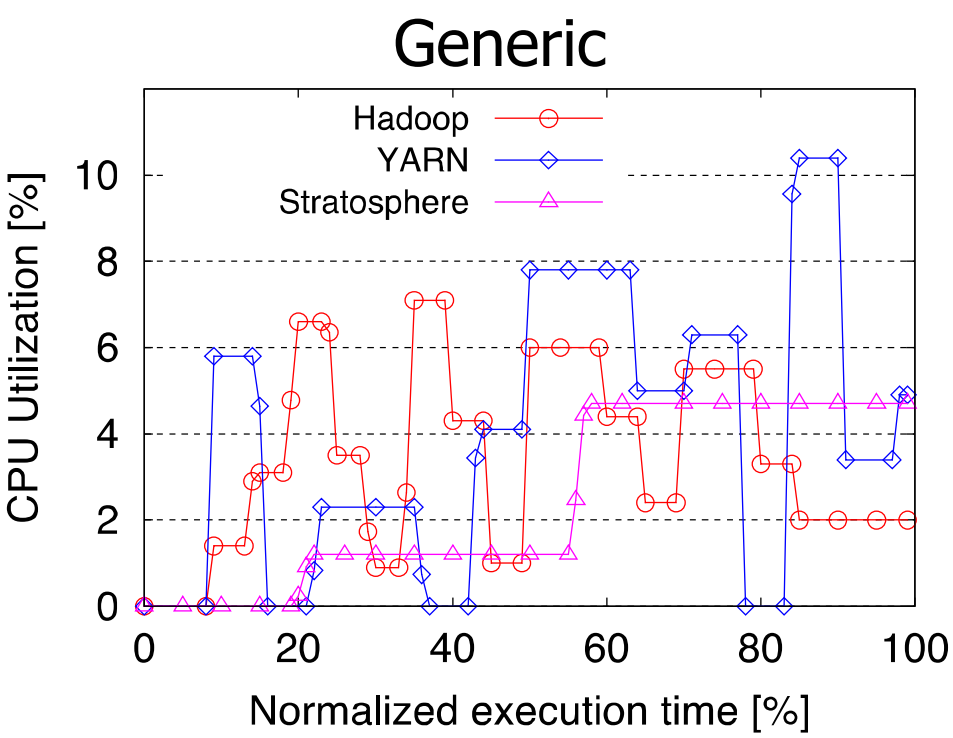
Scalability: BFS on Friendster

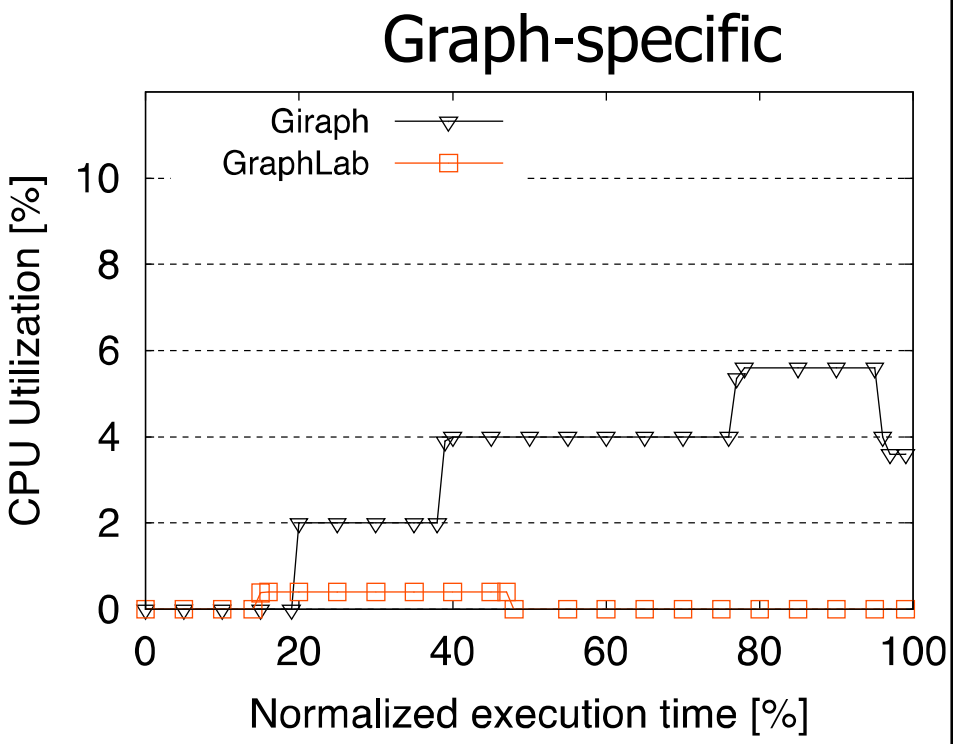




- Using more computing machines/cores can reduce execution time
- Tuning needed, e.g., for GraphLab, split large input file into number of chunks equal to the number of machines

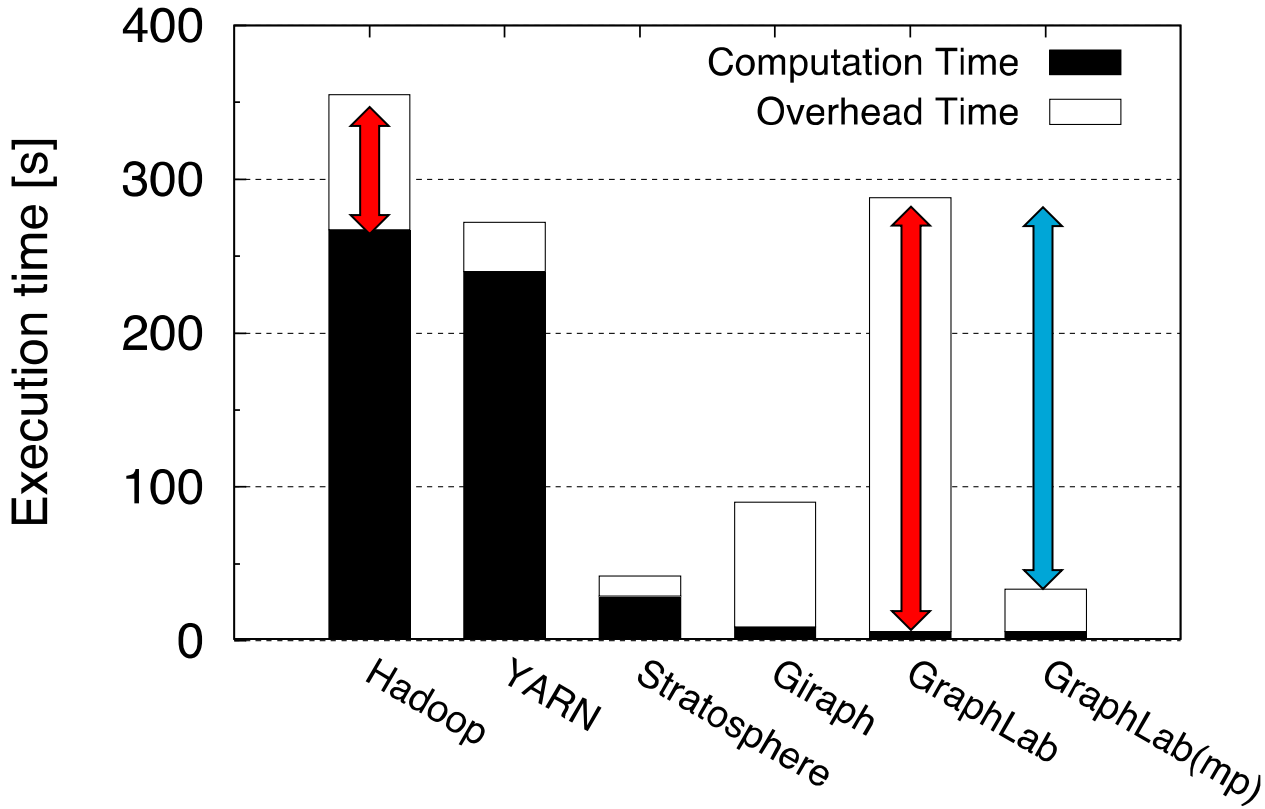
The CPU utilization: computing node





- YARN and Hadoop exhibit obvious volatility
- The CPU utilization of graph-specific platforms is lower

Overhead: BFS on DotaLeague



- The percentage of overhead time is diverse across the platforms, algorithms, and graphs—tuning is only sometimes an option

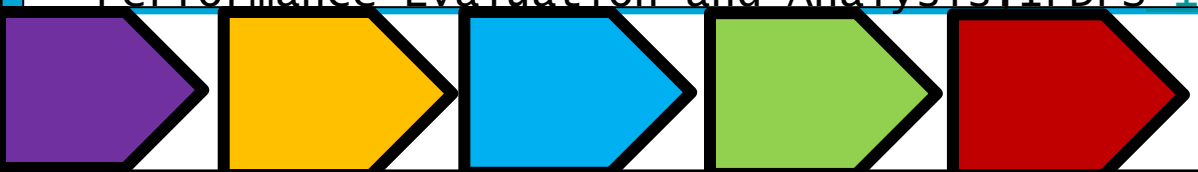
Key Findings From the Study of 6 Platforms

- Performance is function of (Dataset, Algorithm, 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.)

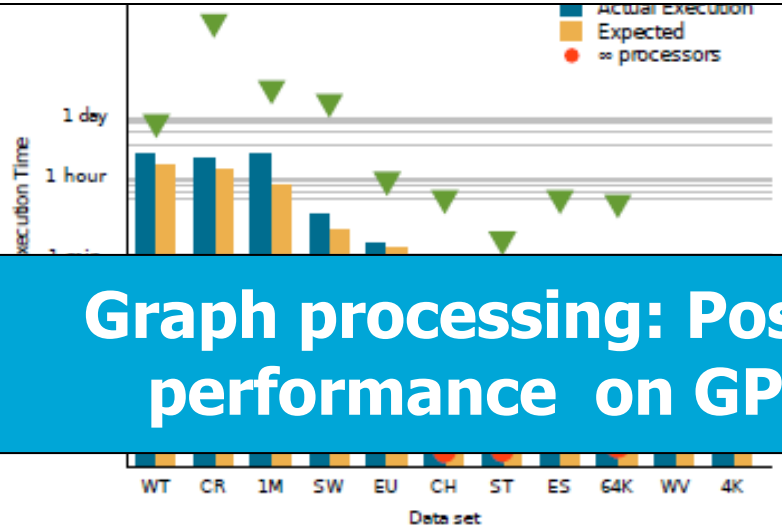
- Some platforms can scale up reasonably with cluster size (horizontally) or number of cores (vertically)
  - Strong vs weak scaling still a challenge—workload scaling tricky

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.



GPUs vs CPUs: All-Pairs Shortest Path

Pender and Varbanescu. MSc thesis at TU Delft. Jun 2012. TU Delft Library, http://library.tudelft.nl .

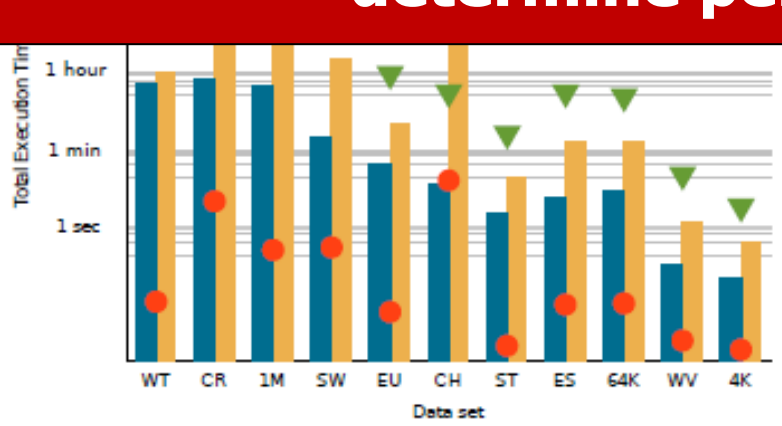


(a) Intel Xeon E5620

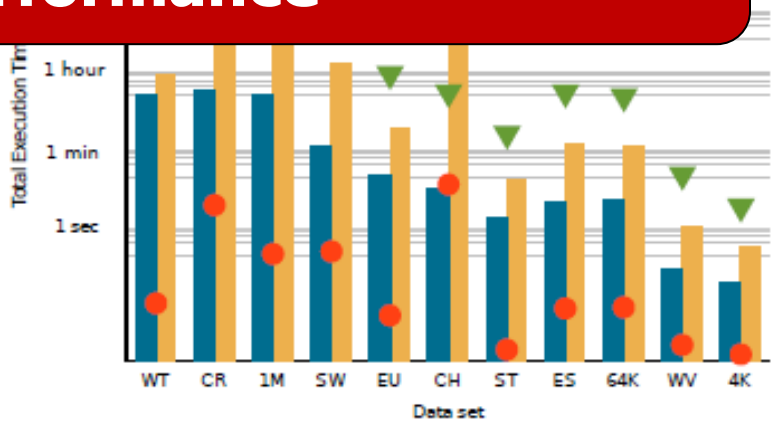
	Dataset
WT	Wikipedia Talk Network
CR	California Road Network
1M	Graph 1M
WV	Wikipedia Vote
4K	Graph 4K

Graph processing: Possible to get better performance on GPUs than on CPUs

However, Algorithm and Dataset also determine performance



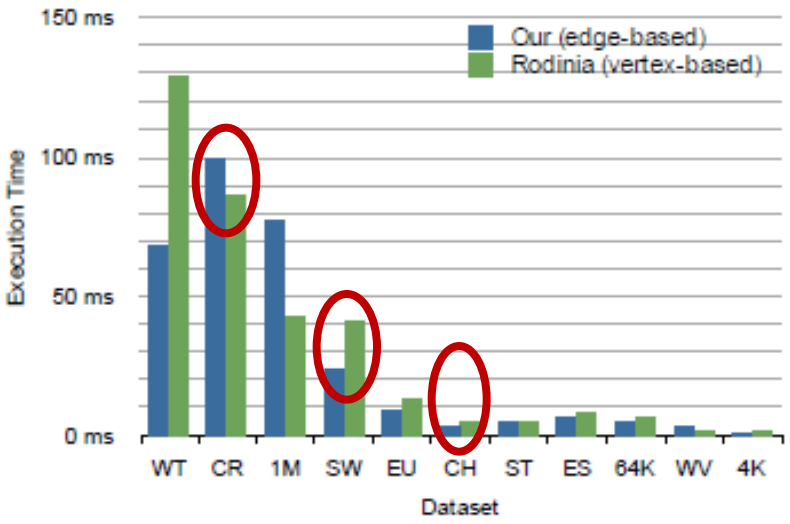
(c) Nvidia Tesla C2050/ C2070



(d) Nvidia GeForce GTX480

GPUs vs CPUs: BFS vs Data Format, E/V-based

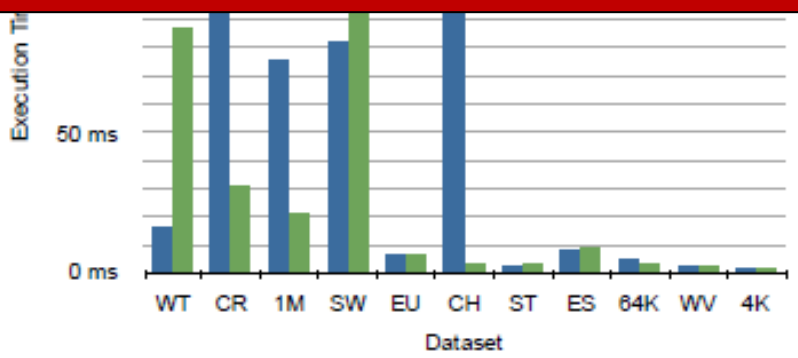
Pender and Varbanescu. MSc thesis at TU Delft. Jun 2012. TU Delft Library, http://library.tudelft.nl .



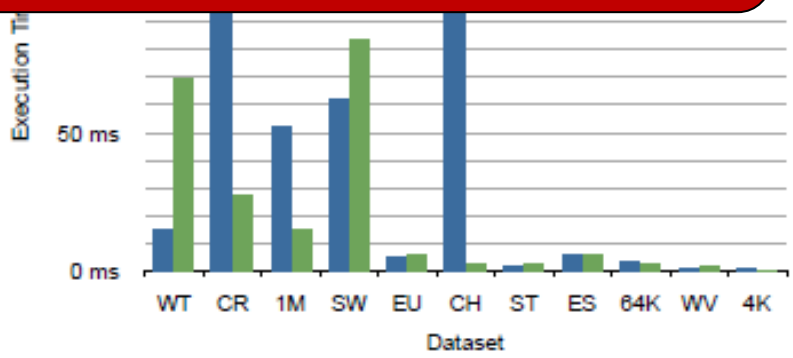
(a) Intel Xeon E5620

	Dataset
WT	Wikipedia Talk Network
CR	California Road Network
1M	Graph 1M
SW	Stanford Web Graph
EU	EU Email Communication Network
CH	Chain 100K
ST	Star 100K
ES	Epinions Social Network
64K	Graph 64K
WV	Wikipedia Vote
4K	Graph 4K

However, data format can also determine performance



(c) Nvidia Tesla C2050/ C2070

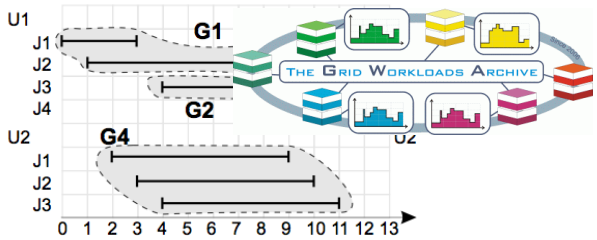


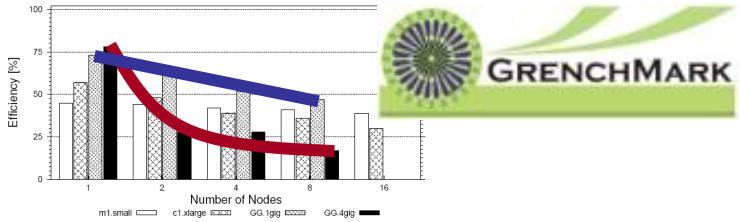
(d) Nvidia GeForce GTX480

TU Delft

Agenda

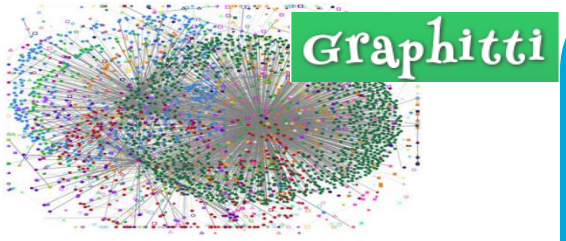
1. Everyone is a scientist!
2. Benchmarking: let's show the numbers





Data Center Workloads

Cloud Performance



Big Data: Graphs

Big Data: BTWorld

Big Data: Elastic MR

3. Conclusion

TU Delft

BTWorld: Our Team

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

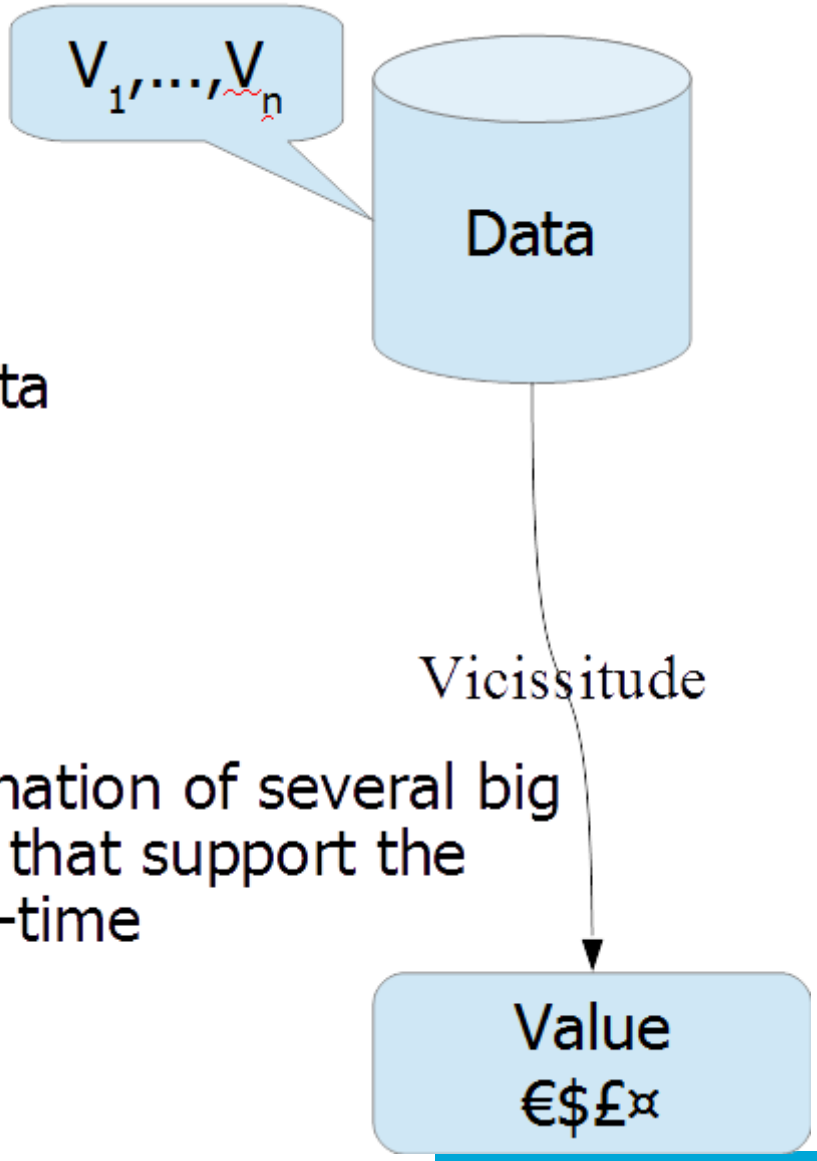
Dick Epema
TU Delft
Big Data & Clouds
Res. management
Systems

Mihai Capota
TU Delft
Big Data apps
Benchmarking

Jan Hidders

Tim Hegeman

Vs of big data

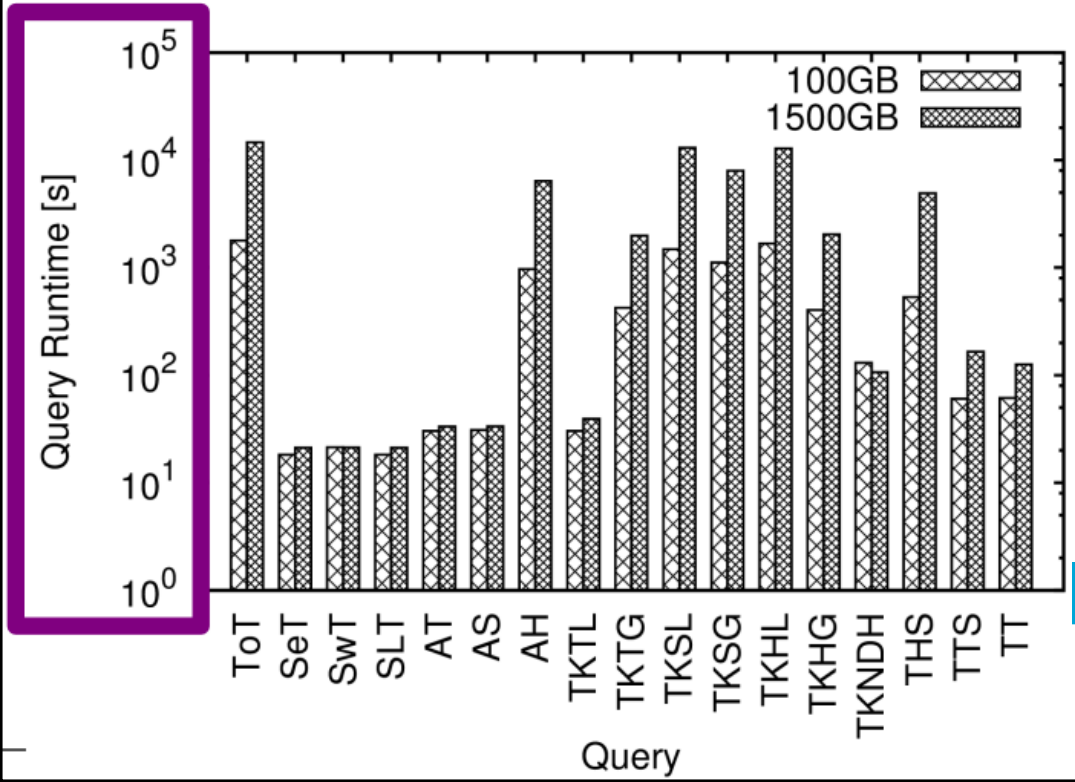


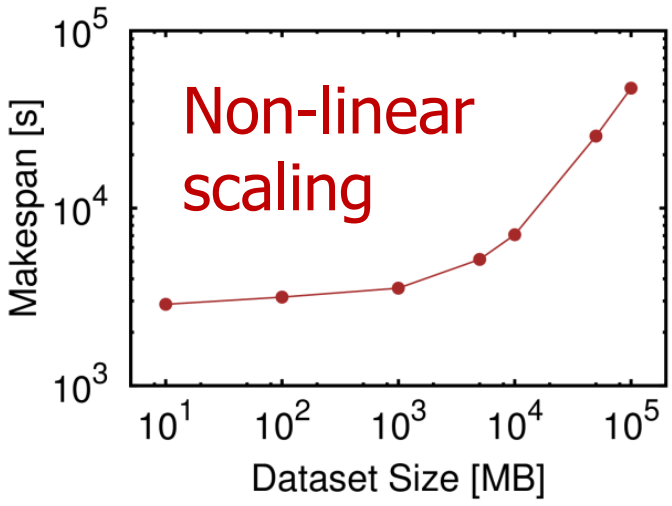
- Volume – large scale of data
- Variety – different forms of data
- Velocity – timeliness of data
- Veracity – uncertainty of data
- **Vicissitude** – dynamic combination of several big data Vs in processing systems that support the addition of new queries at run-time

vicissitude *noun* [vɪ'sɪsɪ tu(:)d]:
a favorable or unfavorable event or situation that occurs by chance; a fluctuation of state or condition
http://merriam-webster.com/dictionary/vicissitude

Benchmarking MapReduce Systems

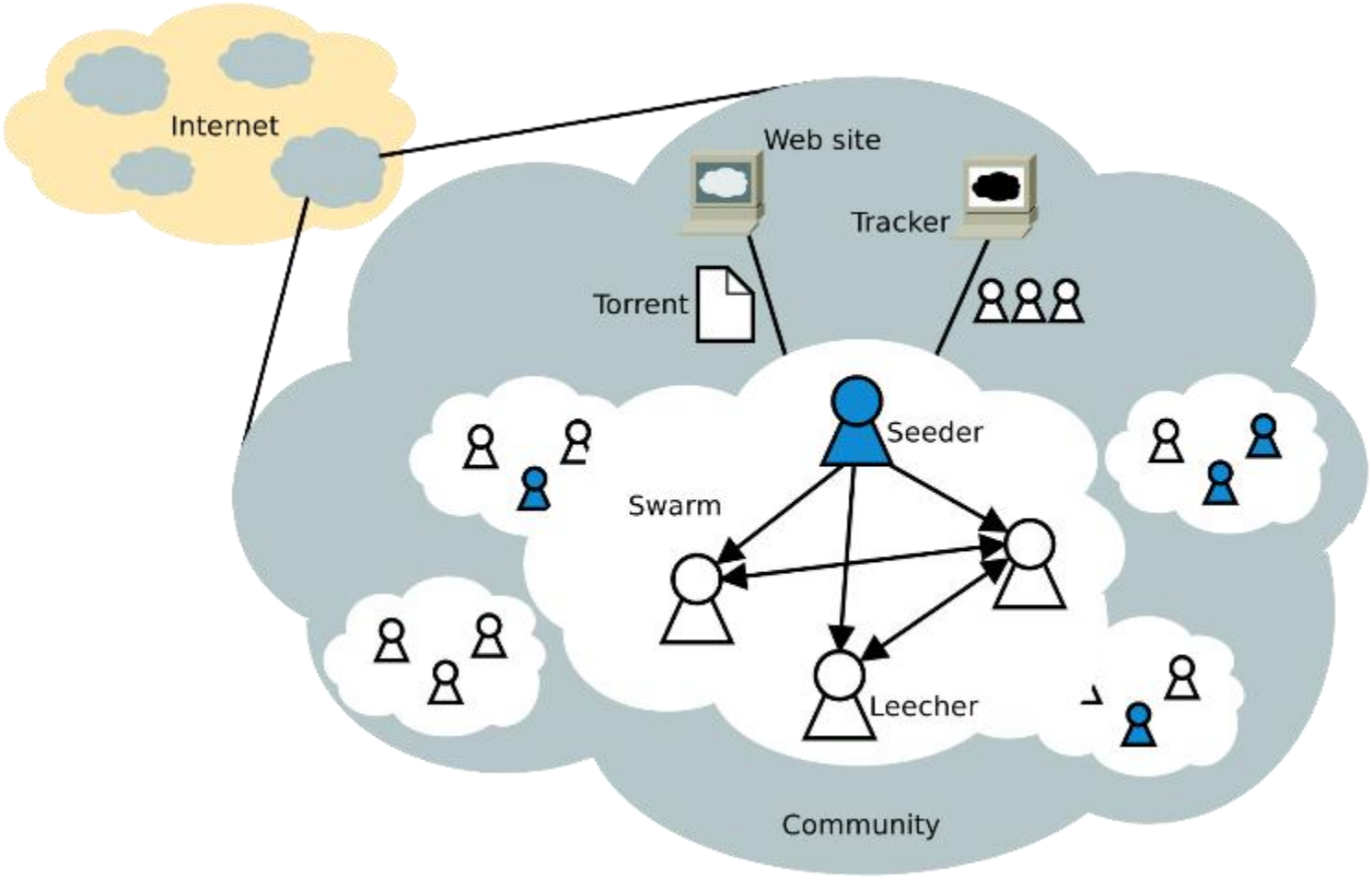
	Queries/Jobs	Workload Diversity	Data Set	Data Layout	Data Volume
MRBench [15]	business queries	high	TPC-H	relational data	3 GB
N-body Shop [14]	filter and correlate data	reduced	N-body simulations	relational data	50 TB
DisCo [6]	co-clustering	reduced	Netflix [29]	adjacency matrix	100 GB
MadLINQ [7]	matrix algorithms	reduced	Netflix [29]	matrix	2 GB
ClueWeb09 [30]	web search	reduced	Wikipedia	html	25 TB
GridMix [16], PigMix [17]	artificial	reduced	random	binary/text	variable
HiBench [31], PUMA [32]	text/web analysis	high	Wikipedia	binary/text/html	variable
WL Suites [12]	production traces	high	-	-	-
BTWorld	**P2P analysis**	**high**	**BitTorrent logs**	**relational data**	**14 TB**





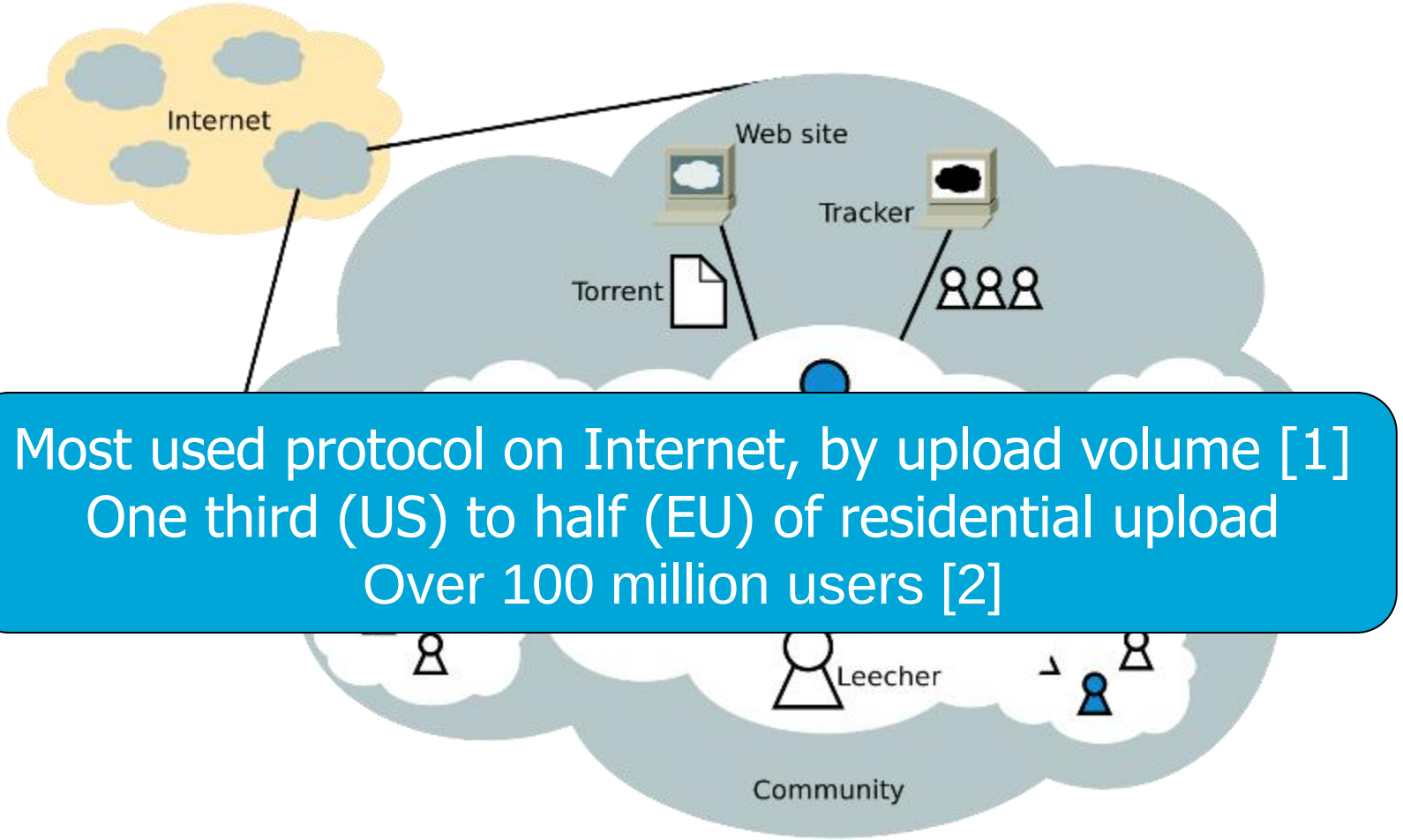
TU Delft

Observing BitTorrent: Managing A Typical Global Distributed System



TU Delft

Observing BitTorrent: Managing A Typical Global Distributed System

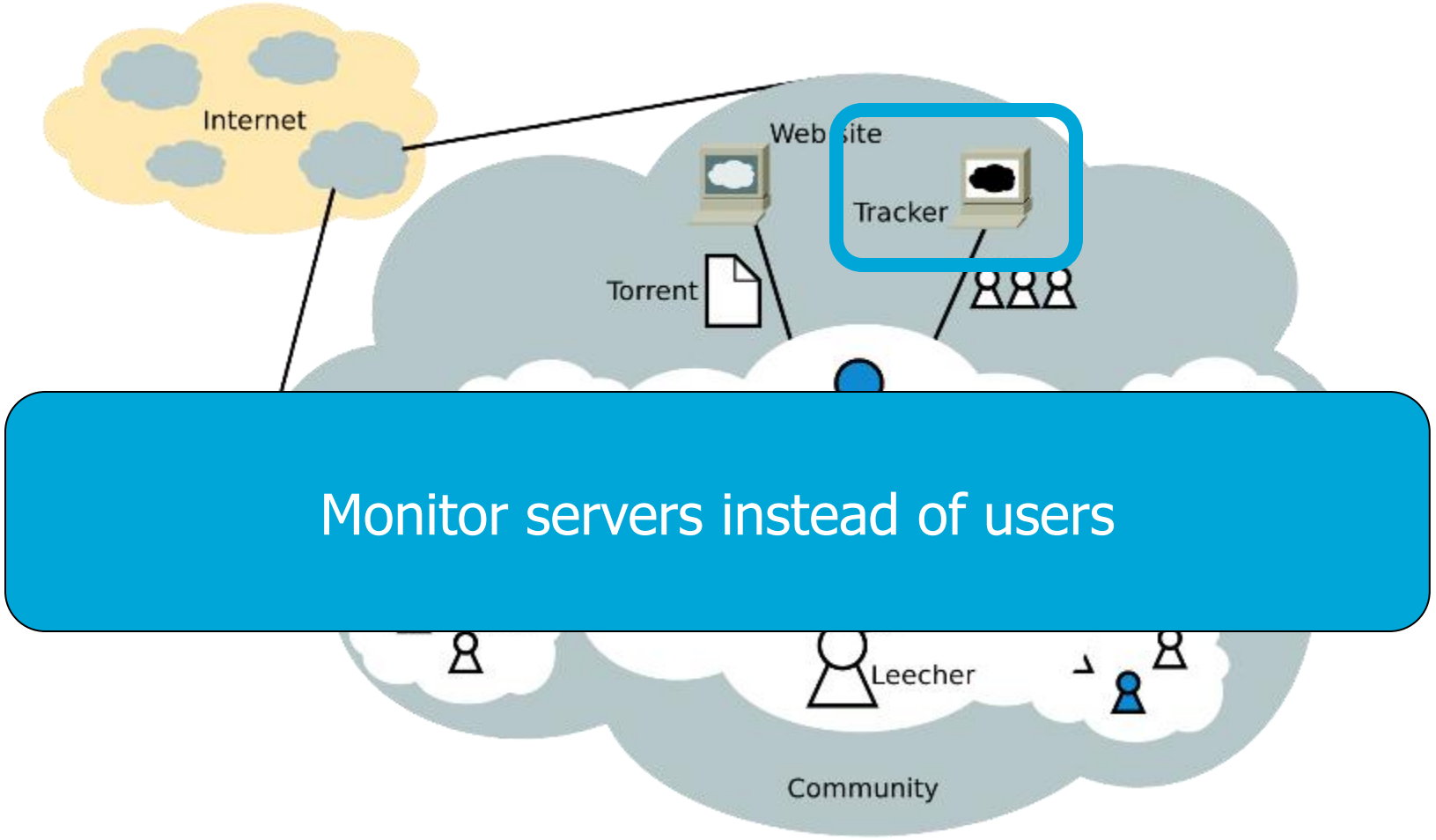


Most used protocol on Internet, by upload volume [1]
One third (US) to half (EU) of residential upload
Over 100 million users [2]

[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

TU Delft

Observing BitTorrent: Managing A Typical Global Distributed System

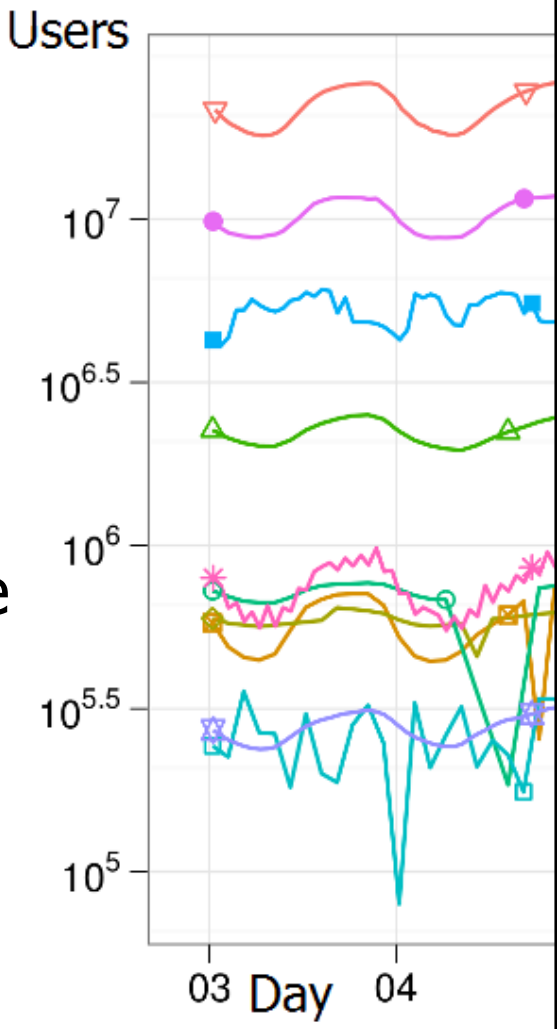


TU Delft

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

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)

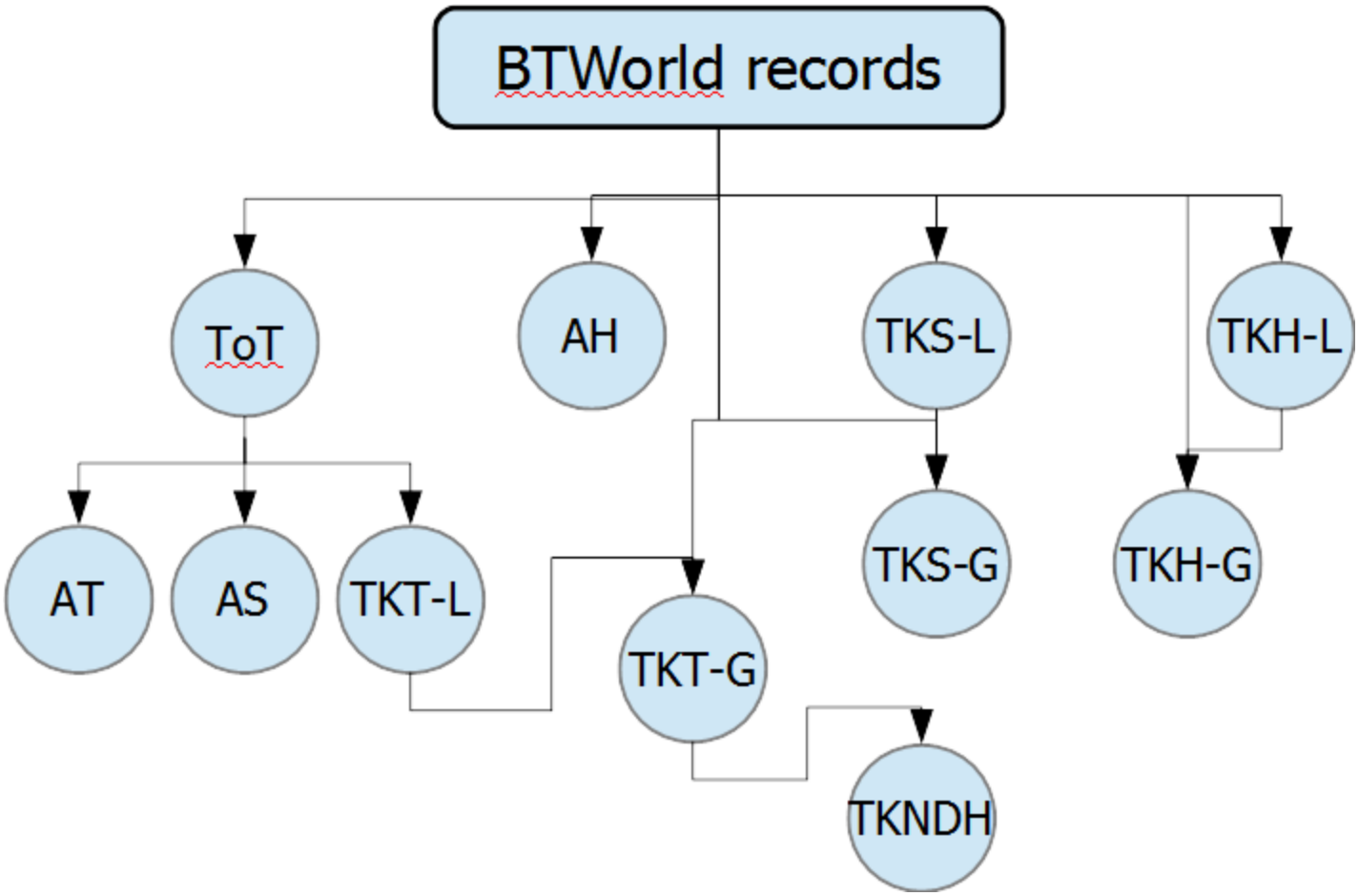
Analyst Questions

- 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

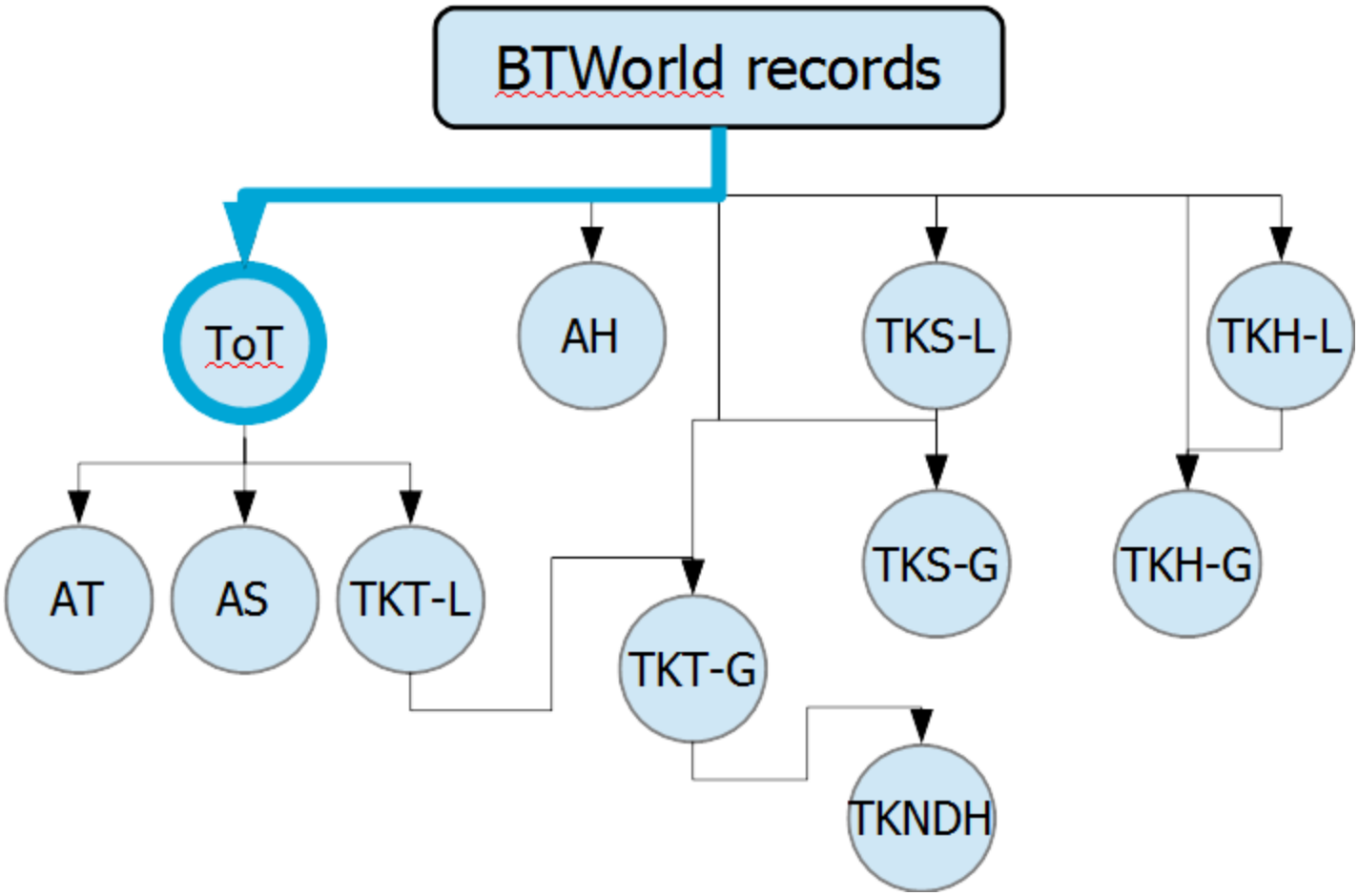


TKH-L Query

Data path

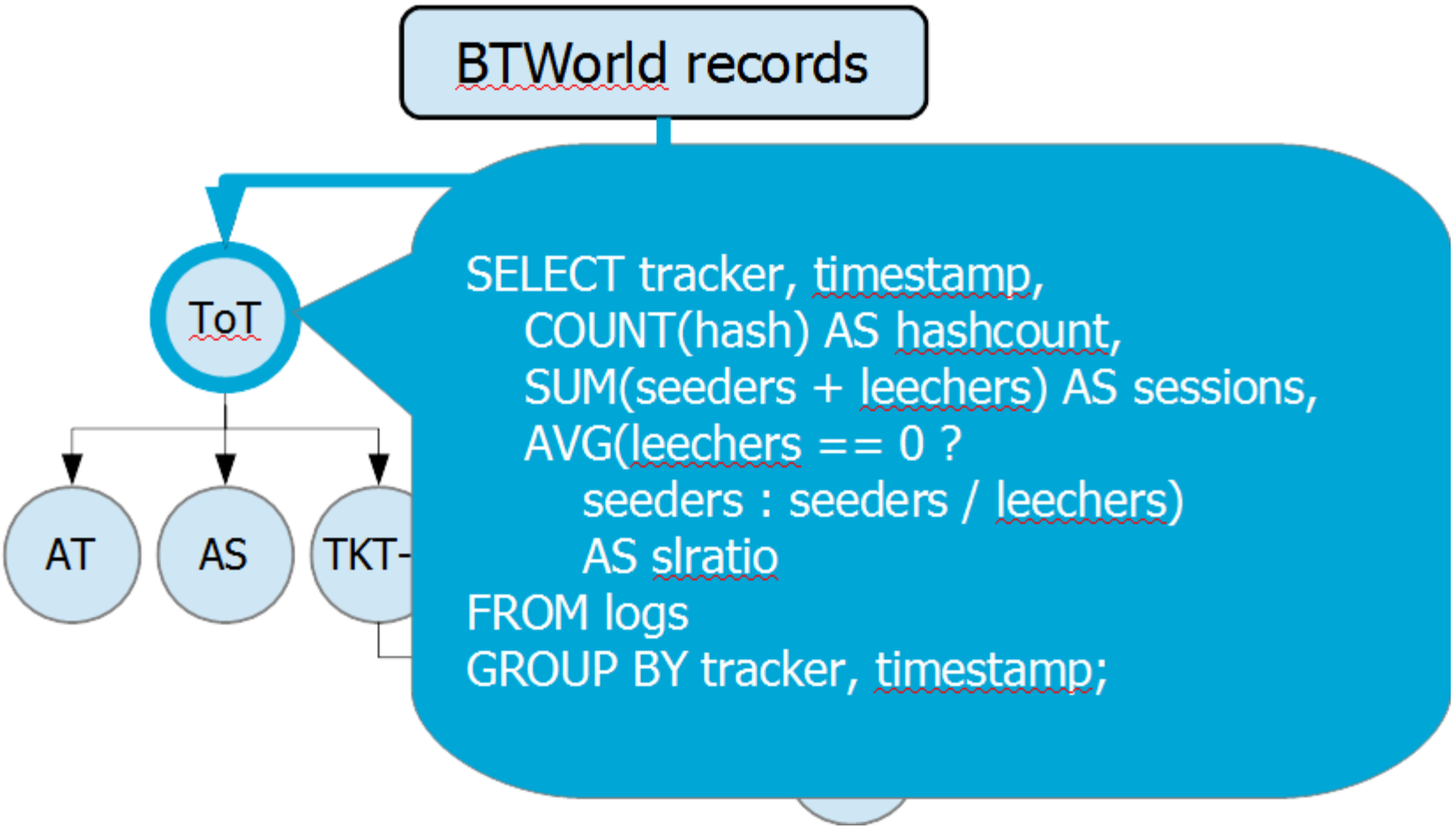
TU Delft

The BTWorld Workflow



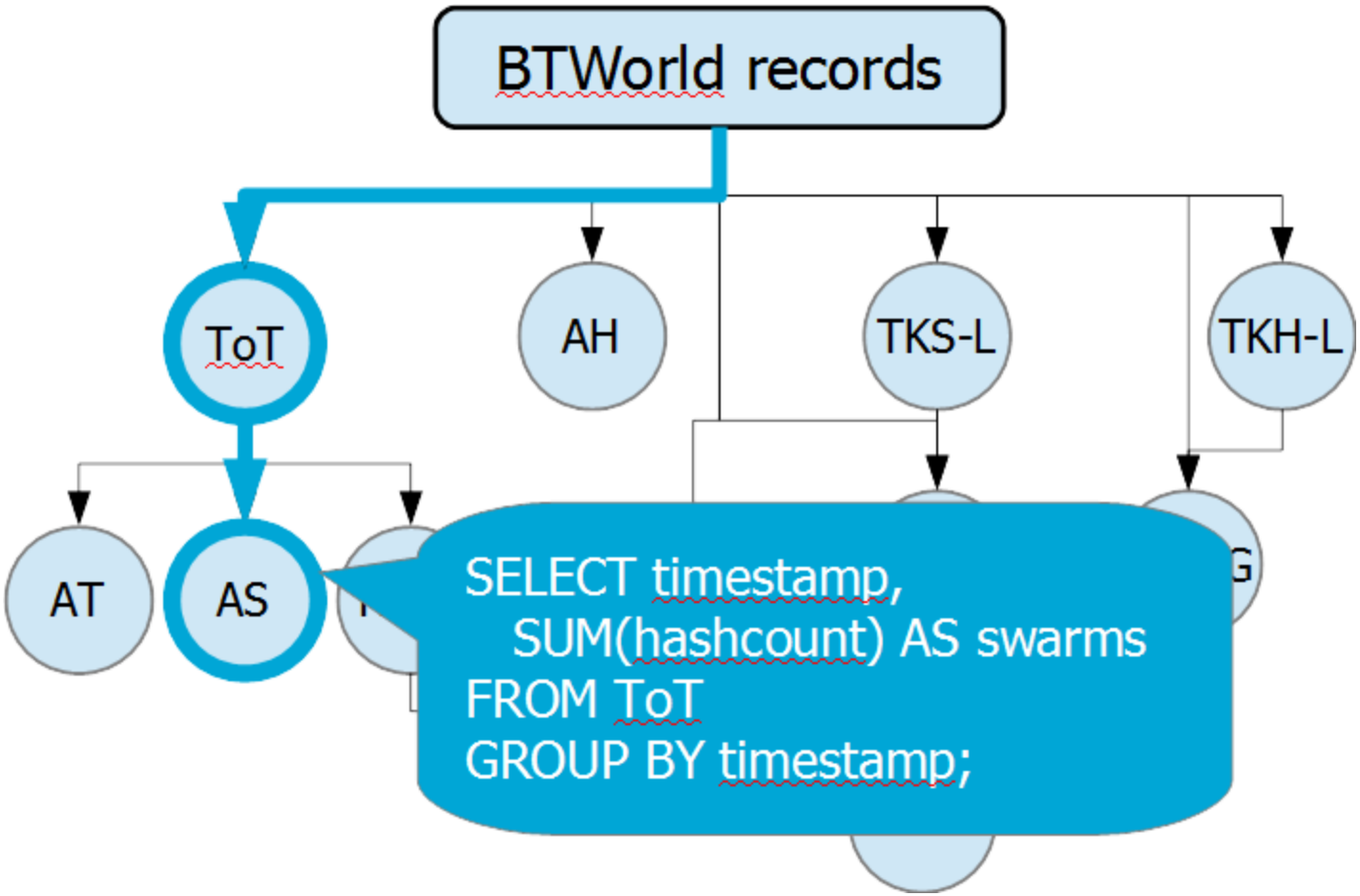
TU Delft

The BTWorld Workload



TU Delft

The BTWorld Workload



TU Delft

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

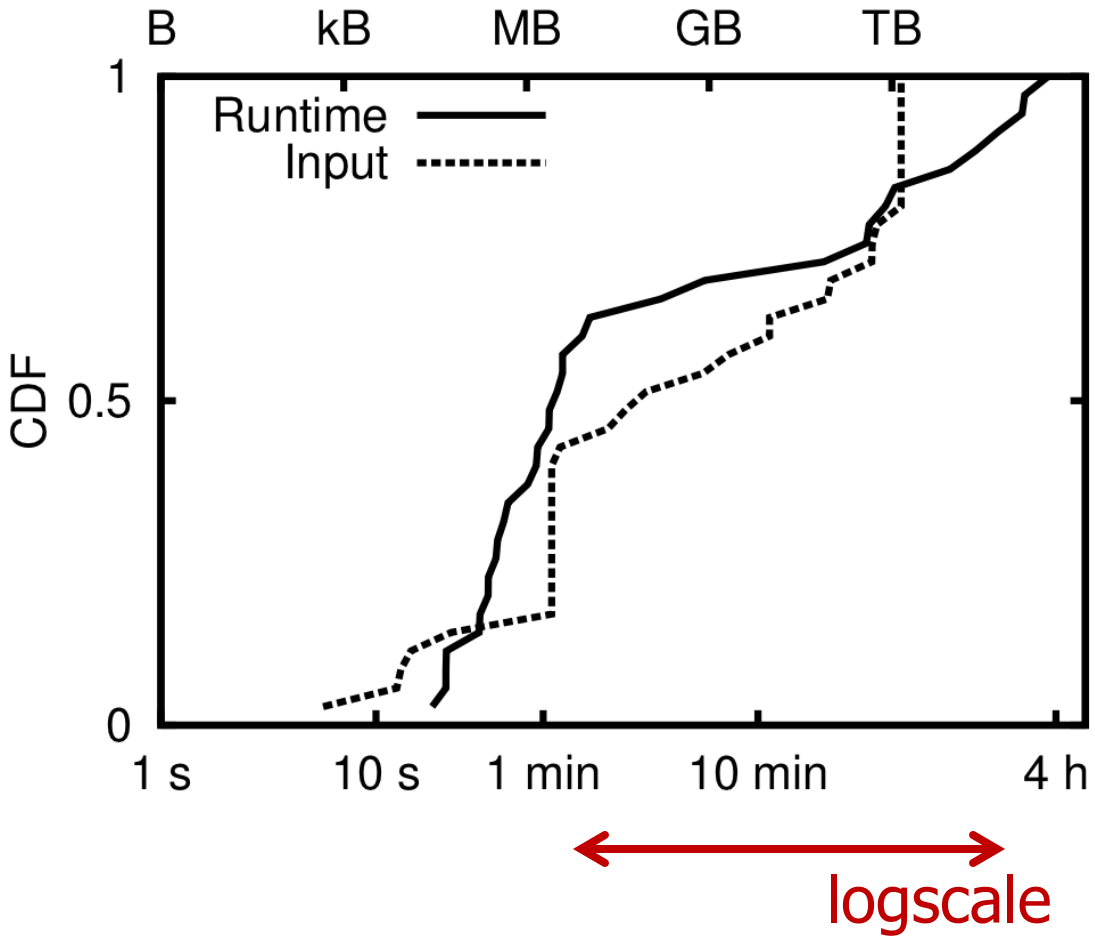
Cluster configuration—DAS4

Cluster size	24 nodes
Map slots	92
Reduce slots	92
Memory per task	6 GiB
Total cluster memory	552 GiB
Scheduler	FIFO
HDFS replication	2

Representative for SMEs

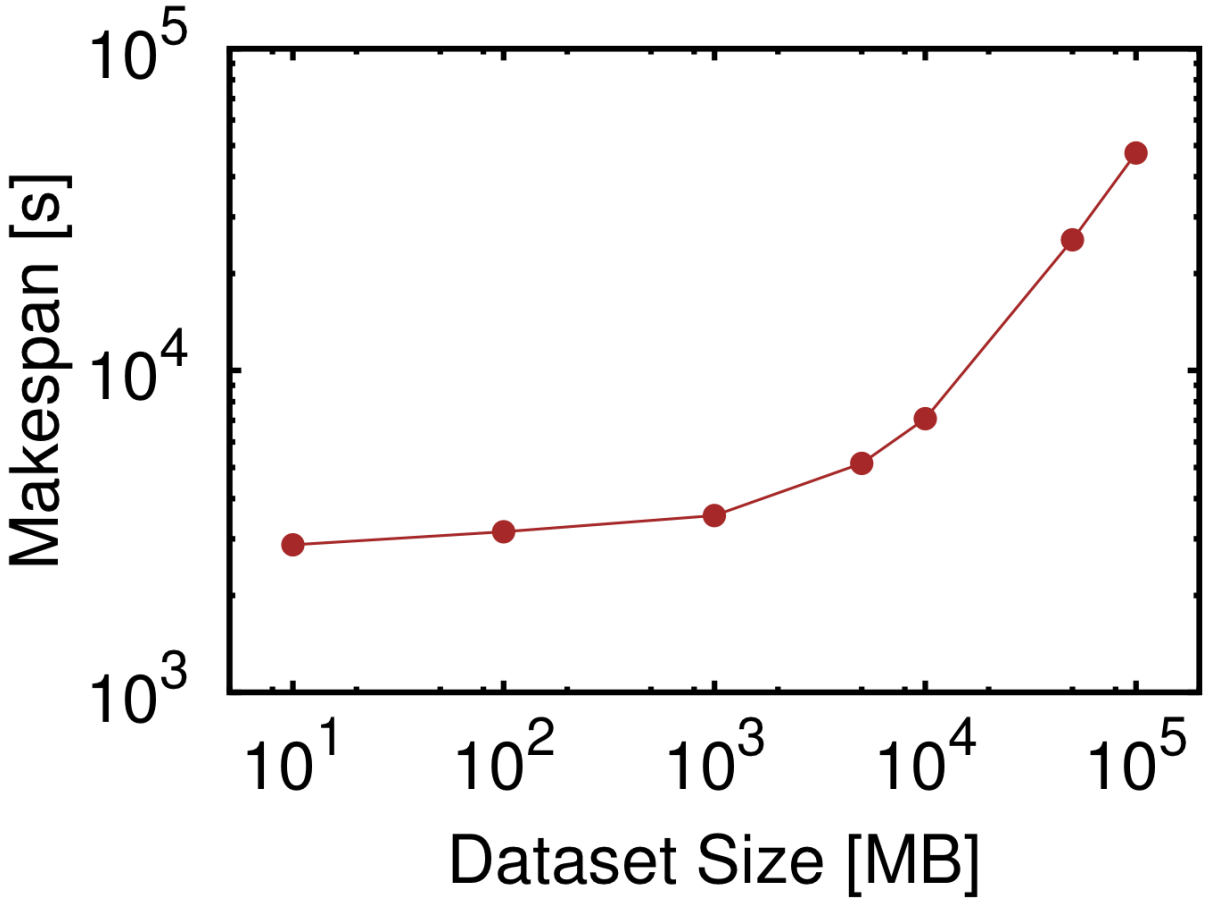
M. Capota, B. Ghit, T. Hegeman, D. Epema, and A. Iosup. V for Vicissitude: The Challenge of Scaling Complex Big Data Workflows, In the IEEE/ACM CCGRID (SCALE Challenge 2014). 2014. Winner of Challenge.

Variety



M. Capota, B. Ghit, T. Hegeman, D. Epema, and A. Iosup. V for Vicissitude: The Challenge of Scaling Complex Big Data Workflows, In the IEEE/ACM CCGRID (SCALE Challenge 2014). 2014. Winner of Challenge.

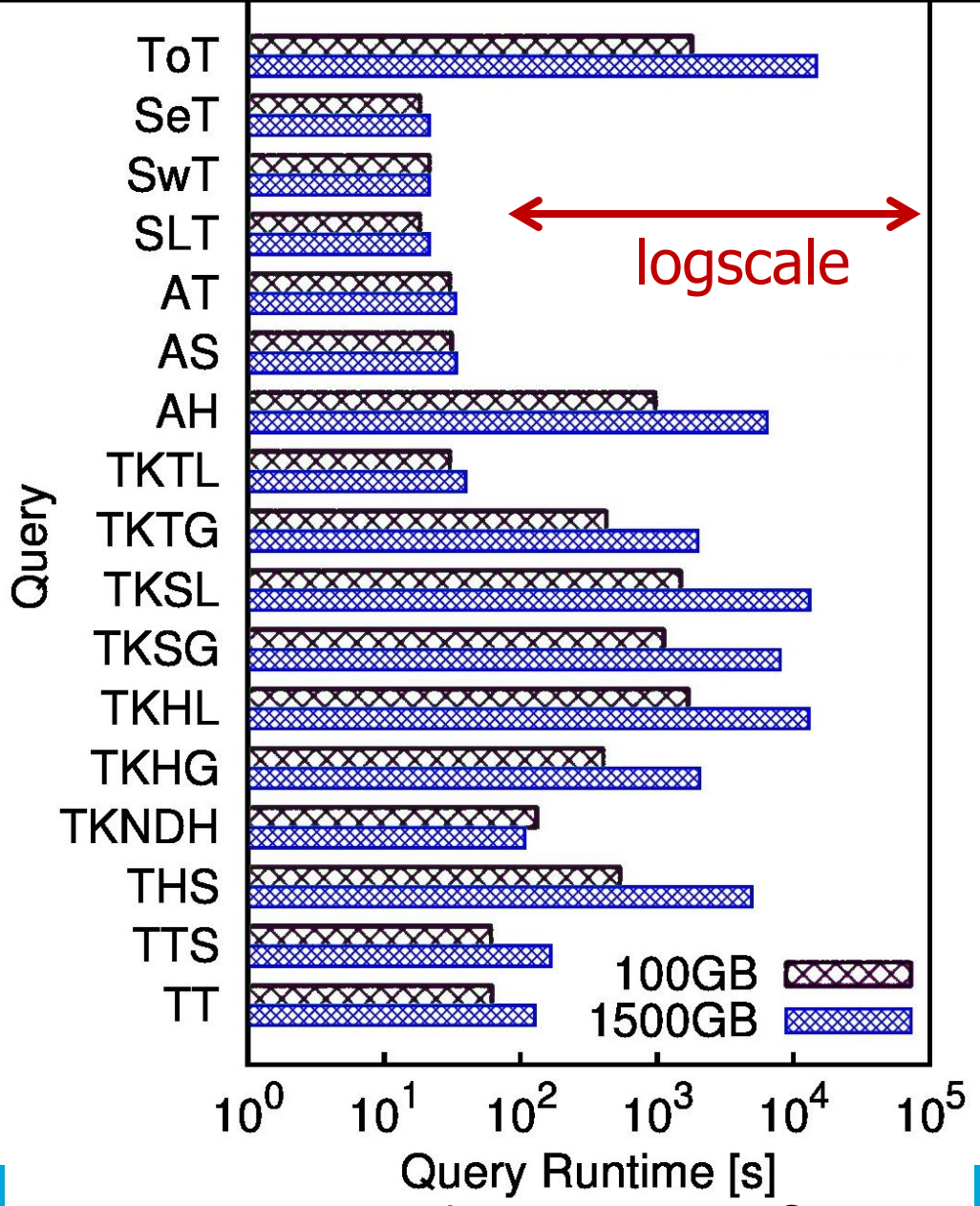
Workload Does Not Scale Linearly



M. Capota, B. Ghit, T. Hegeman, D. Epema, and A. Iosup. V for Vicissitude: The Challenge of Scaling Complex Big Data Workflows, In the IEEE/ACM CCGRID (SCALE Challenge 2014). 2014. Winner of Challenge.

Results

- Long vs Short queries
  - Short relatively scale-free
  - Long do not scale super-linearly
- Possible to tune systems to avoid effects of vicissitude



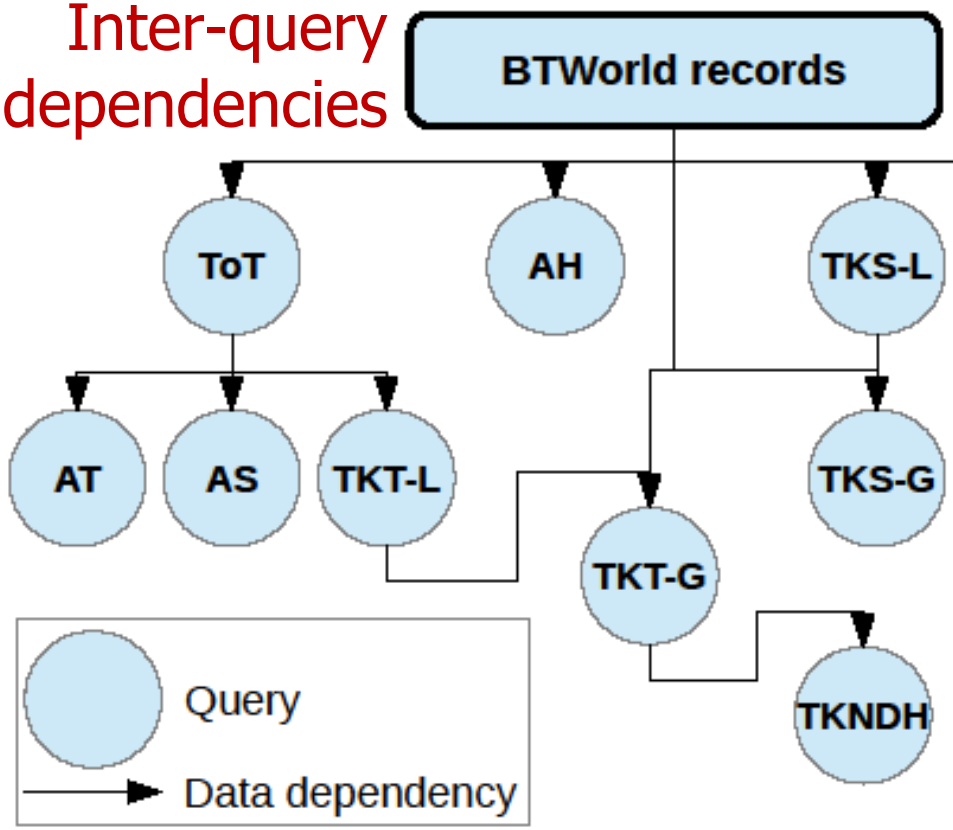
M. Capota, B. Ghit, T. Hegeman, D. Epema, and A. Iosup. V for Vicissitude: The Challenge of Scaling Complex Big Data Workflows, In the IEEE/ACM CCGRID (SCALE Challenge 2014). 2014. Winner of Challenge.

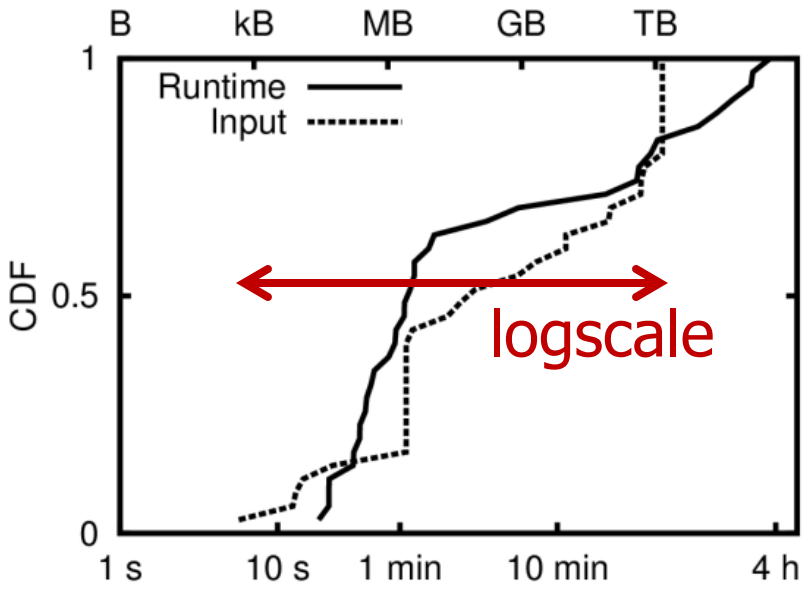
Workflows = Data Vicissitude

Use Case: Monitoring Large-Scale Distributed Computing System with 160M users

Inter-query dependencies

Diverse queries
New queries during project





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

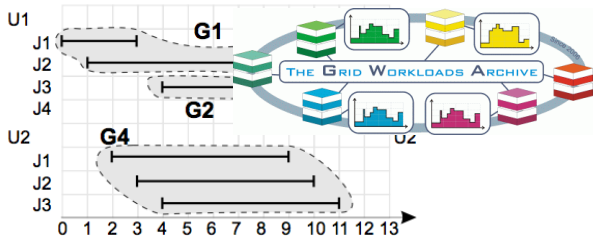
BitTorrent	Trackers	Swarms	Hashes
Finance	Stock markets	Stock listings	Stocks
Tourism	Travel agents	Vacation packages	Venues

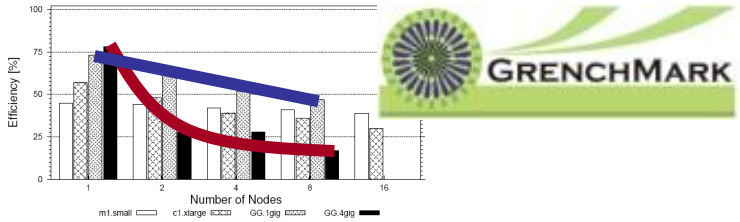
- Monitoring large scale distributed computer systems
- Benchmarking

TU Delft

Agenda

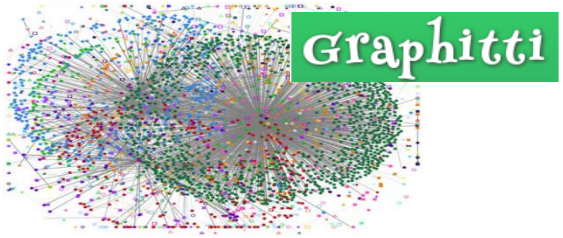
1. Everyone is a scientist!
2. Benchmarking: let's show the numbers





Data Center Workloads

Cloud Performance



Big Data: Graphs

Big Data: BTWorld

Big Data: Elastic MR

3. Conclusion

TU Delft

Elastic MapReduce: Our Team

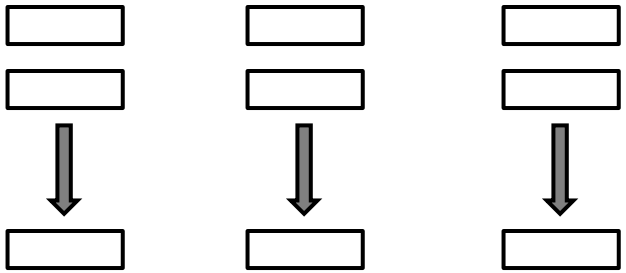
Bogdan Ghit
TU Delft
Systems
Workloads

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



Job submissions

Frameworks

Resource manager

Infrastructure

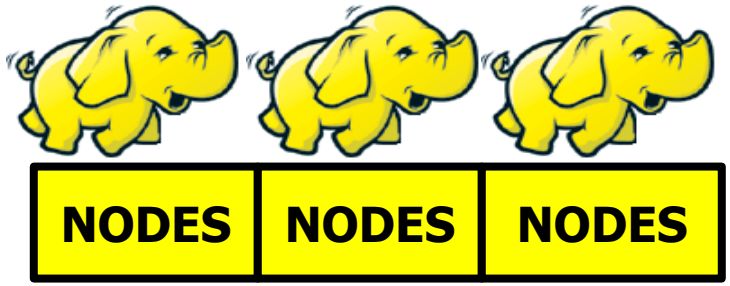
FAWKES/Others

NODES | NODES | NODES

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

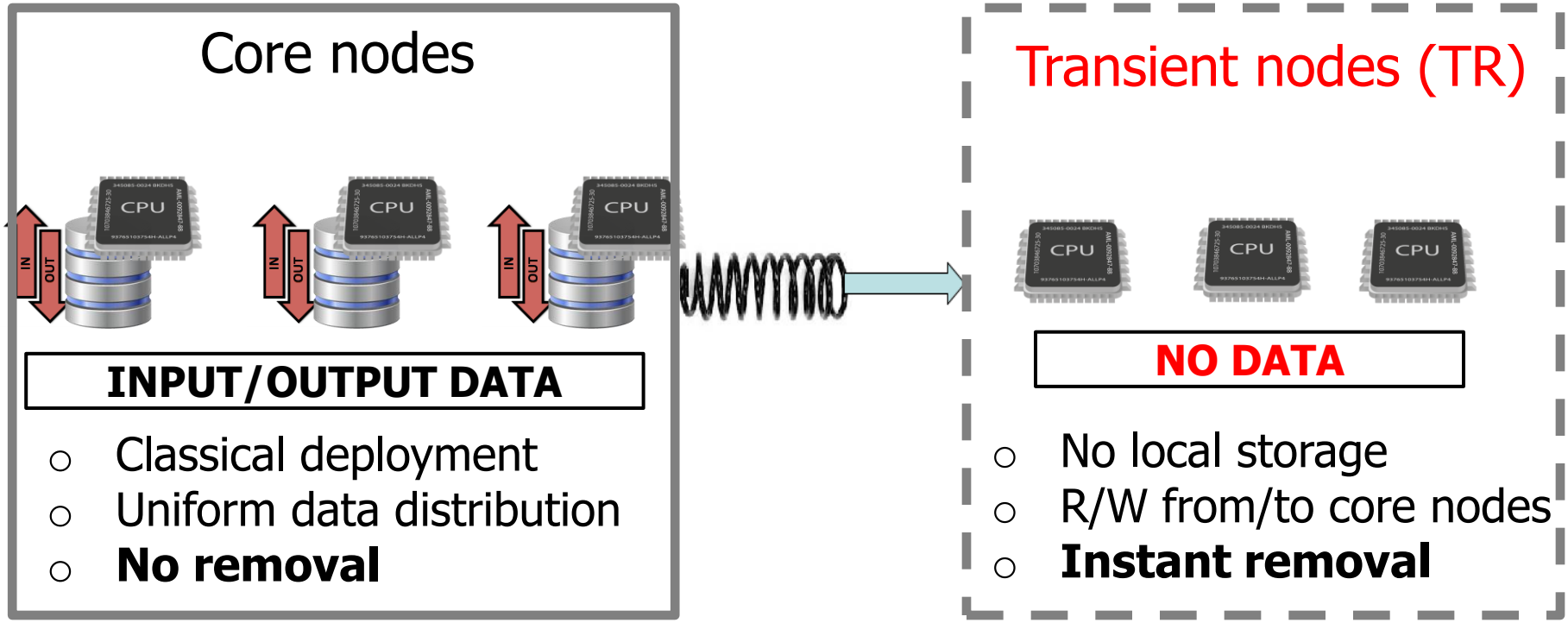


Because workloads may be time-varying:
- Poor resource utilization
- Imbalanced service levels

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.

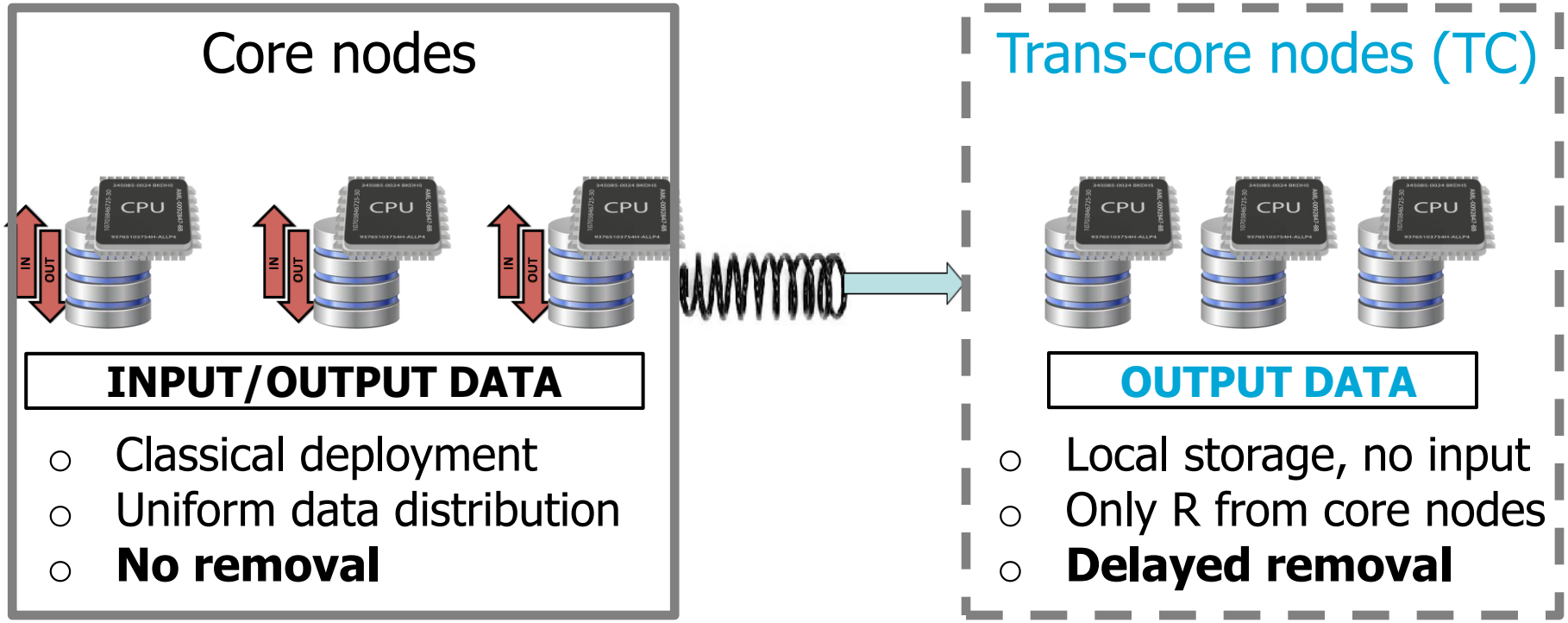
No data locality



Performance?

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

Relaxed data locality



Better performance?

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

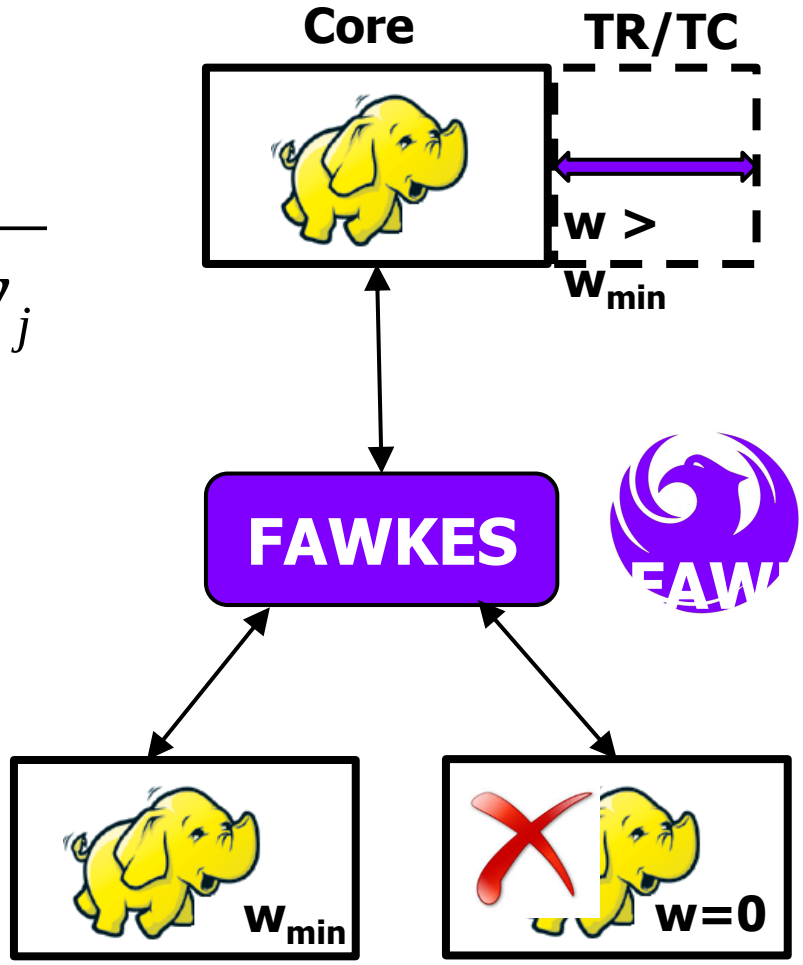
FAWKES in a nutshell

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

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

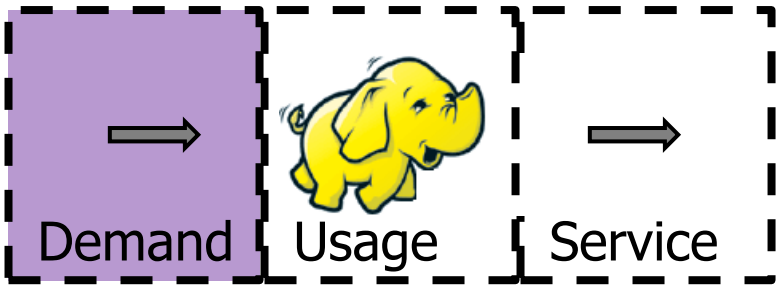
2. Updates dynamic weights when
   - New frameworks arrive
   - Framework states change

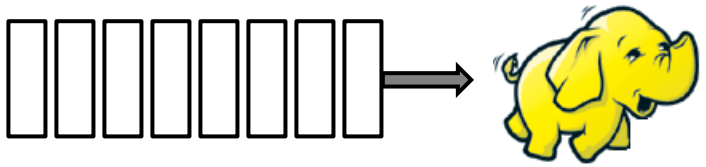
3. Shrinks and grows frameworks to
   - Allocate new frameworks (min. shares)
   - Give fair shares to existing ones



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

How to differentiate frameworks? (1/3)

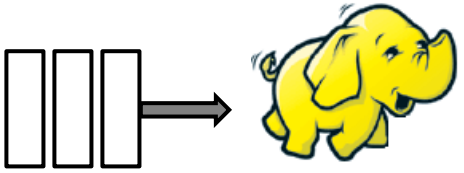




By demand – 3 policies:

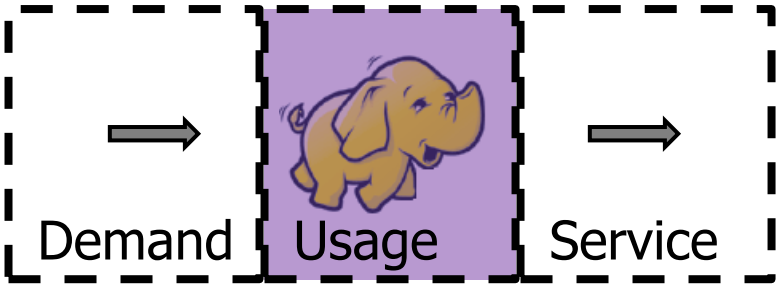
- Job Demand (JD)
- Data Demand (DD)
- Task Demand (TD)

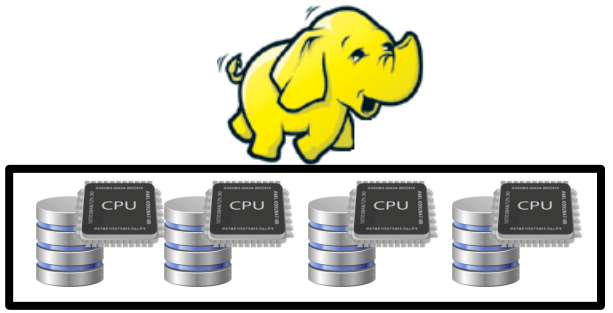
VS.



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

How to differentiate frameworks? (2/3)





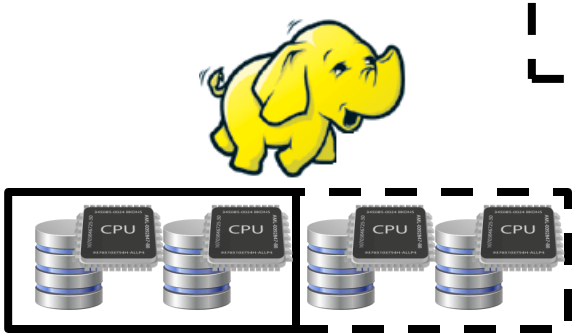
By usage – 3 policies:

- Processor Usage (PU)
- Disk Usage (DU)
- Resource Usage (RU)

VS.

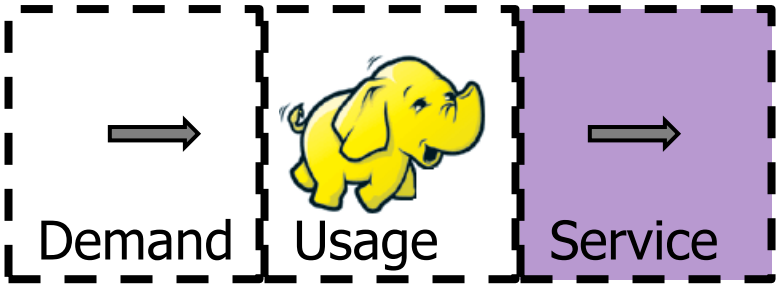
USED

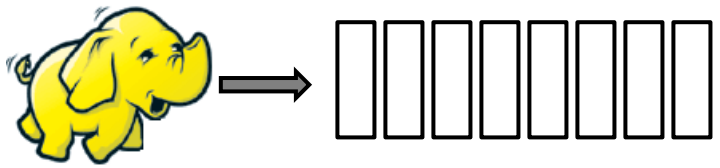
IDLE



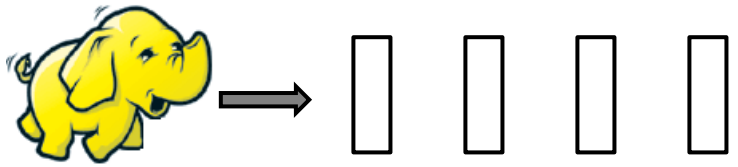
Ghit, Yigitbasi, Iosup, Epema, Iosup. Balanced Resource Allocations Across Multiple Dynamic MapReduce Clusters. ACM SIGMETRICS 2014.

How to differentiate frameworks? (3/3)





VS.

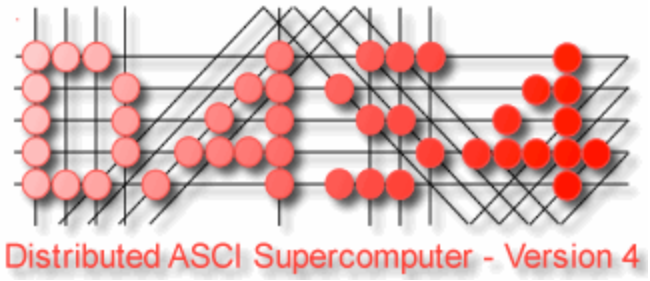


By service – 3 policies:

- Job Slowdown (JS)
- Job Throughput (JT)
- Task Throughput (TT)

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

Experimental setup



DAS-4 multicluster system:

- 200 dual-quad-core compute nodes
- 24 GB memory per node
- 150 TB total storage
- 20 Gbps InfiniBand

Hadoop deployment:

- Hadoop-1.0 over InfiniBand
- 6 map + 2 reduce slots per node
- 128 MB block size

Overview of experiments:

- Most experiments on 20 nodes
- Up to 60 working nodes
- More than 3 months system time

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

MapReduce applications

Application	Type	Input	Output
Wordcount (WC)	CPU	200 GB	5.5 MB
Sort (ST)	Disk	200 GB	200 GB
PageRank (PR)	CPU	50 GB	1.5 MB
K-Means (KM)	Both	70 GB	72 GB
TrackerOverTime (TT)	CPU	100 GB	3.9 MB
ActiveHashes (AH)	Both	100 GB	90 KB
BTWorld (BT)	Both	100 GB	73 GB

Synthetic benchmarks:
- HiBench suite
- Single applications
- Random datasets

Real-world applications:
- BTWorld workflow
- 14 Pig queries
- BitTorrent monitoring data

Ghit, Capota, Hegeman, Hidders, Iosup, Epema, "The Challenge of Scaling Complex Big Data Workflows", CCGrid 2014. SCALE Challenge Winner.

Performance of dynamic MapReduce

10 core +10xTR

10 core +10xTC

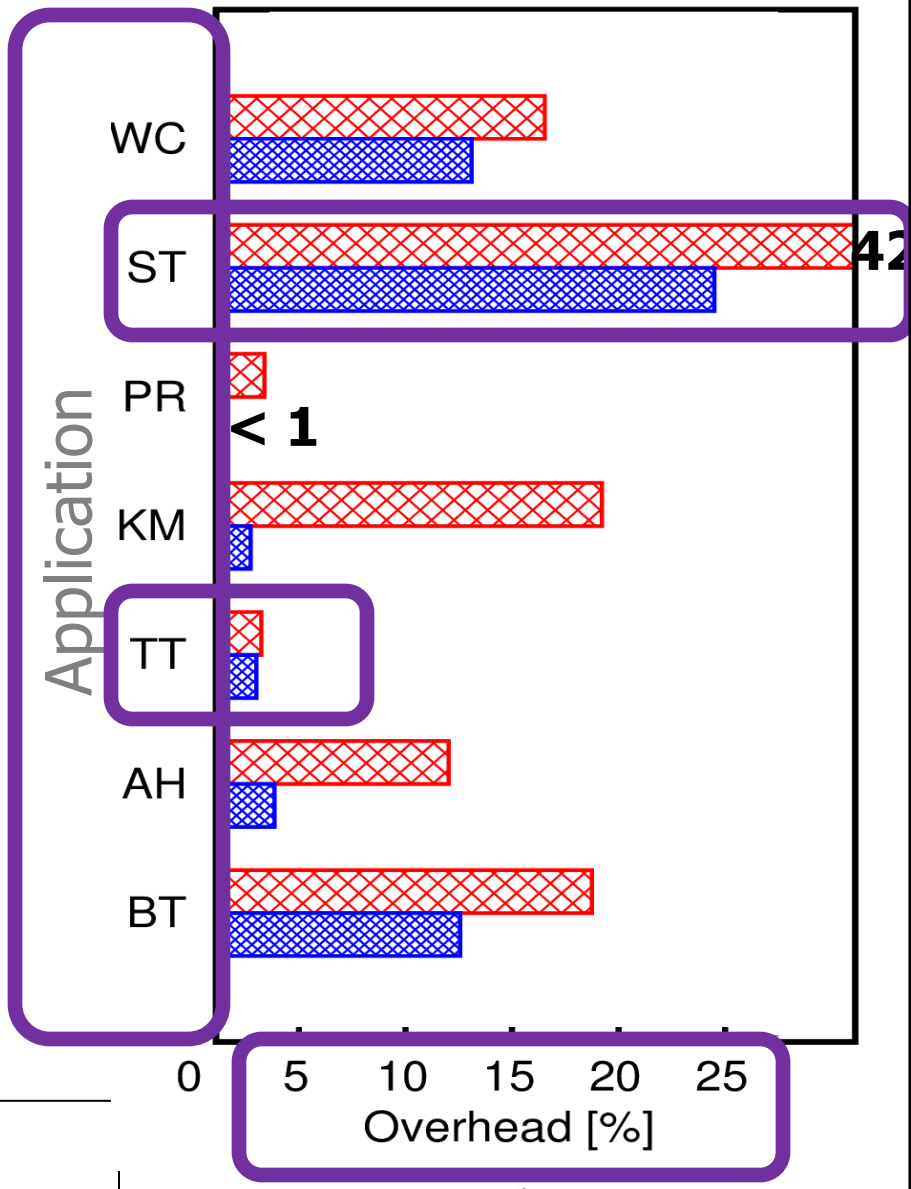
vs.

20 core nodes

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

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

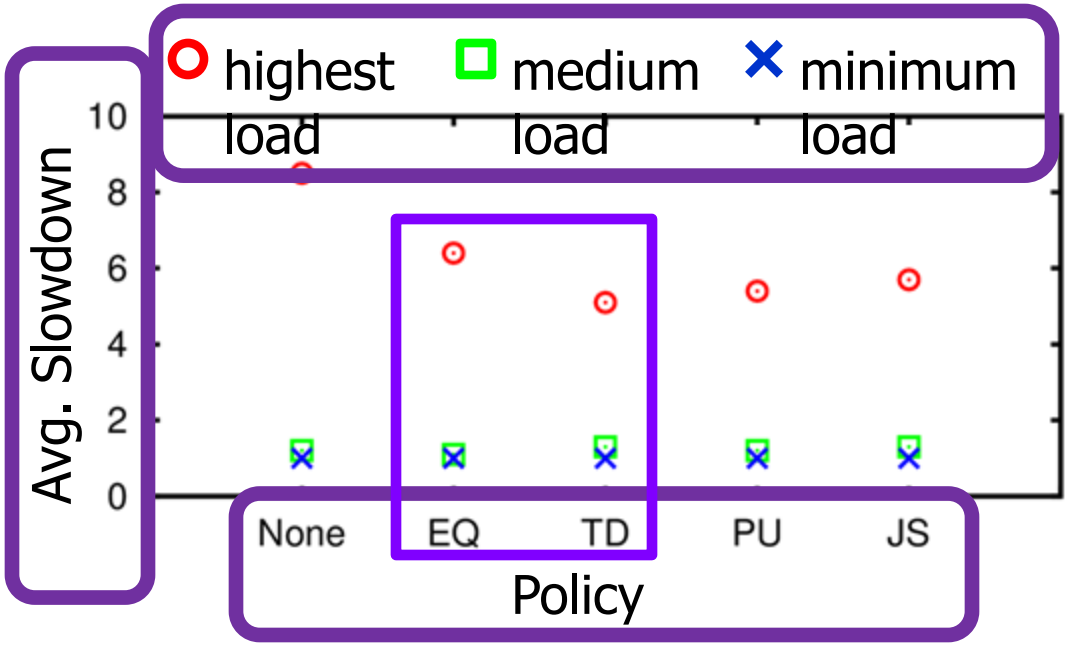
Dynamic MapReduce:
< 25% overhead



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



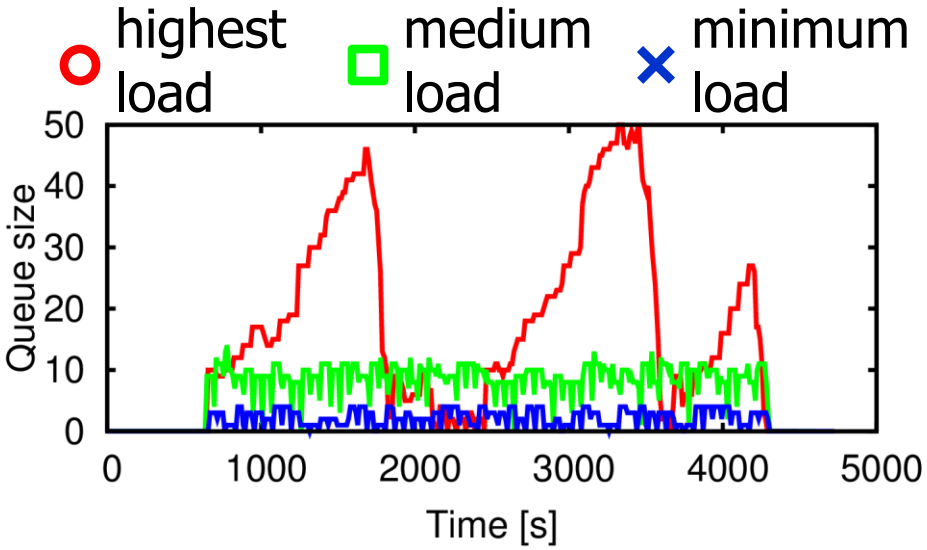
None – Minimum shares
EQ – EQual shares
TD – Task Demand
PU – Processor Usage
JS – Job Slowdown

Up to 20% lower slowdown

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

FAWKES: behind the scenes

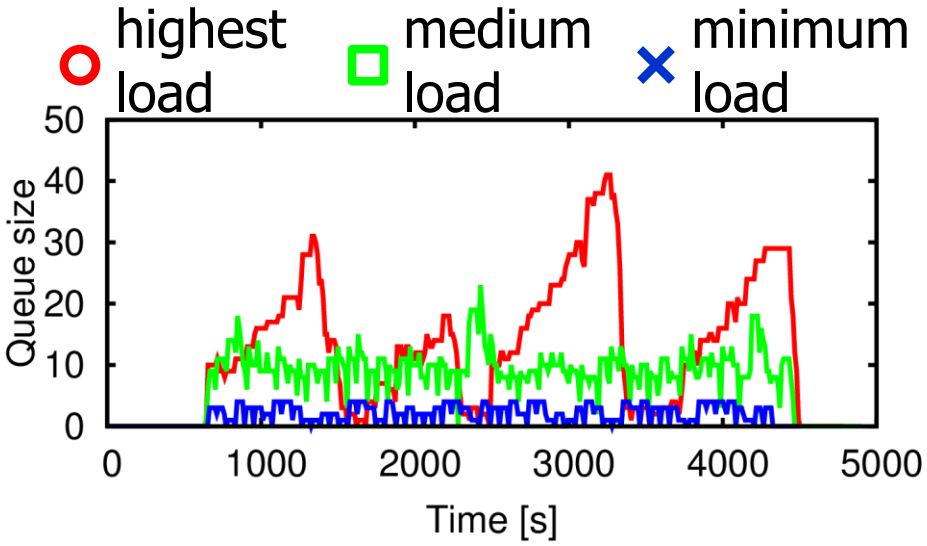
EQ



Utilizations: 60% / 23% / 5%

Imbalanced

TD



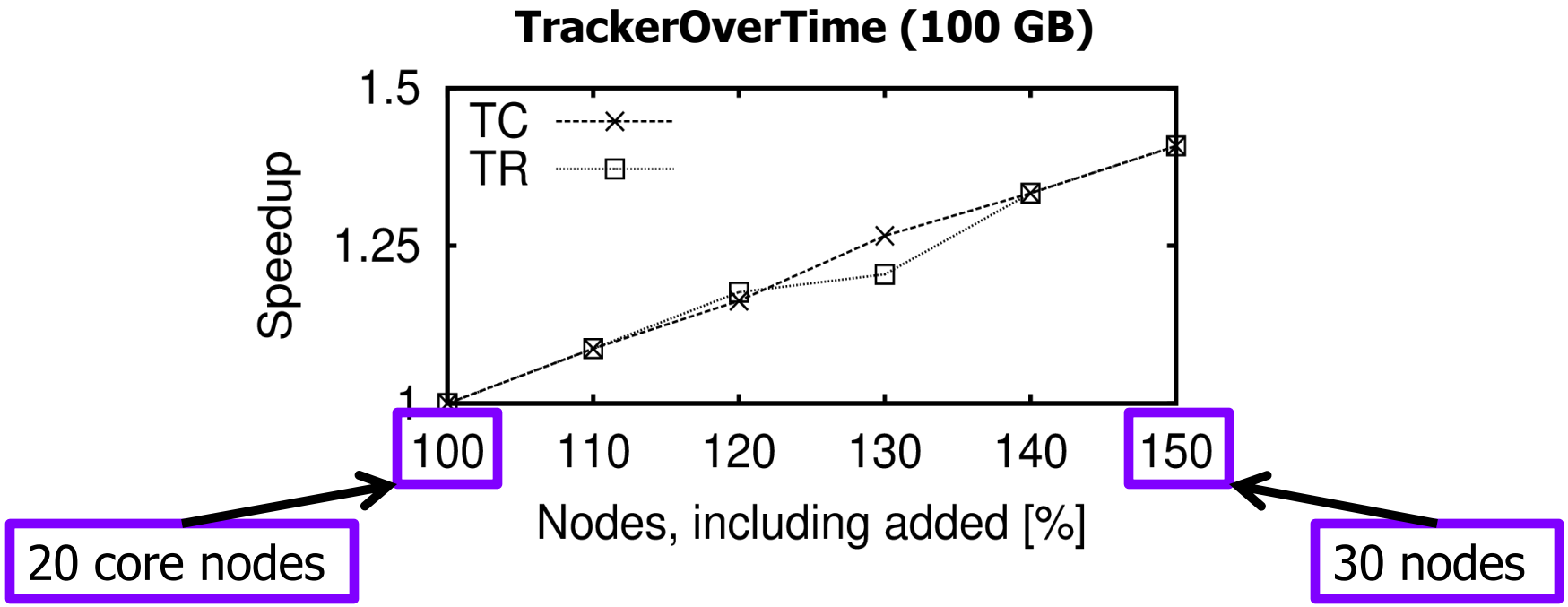
EQ – EQual shares
TD – Task Demand

Utilizations: 50% / 30% / 8%

More balanced

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

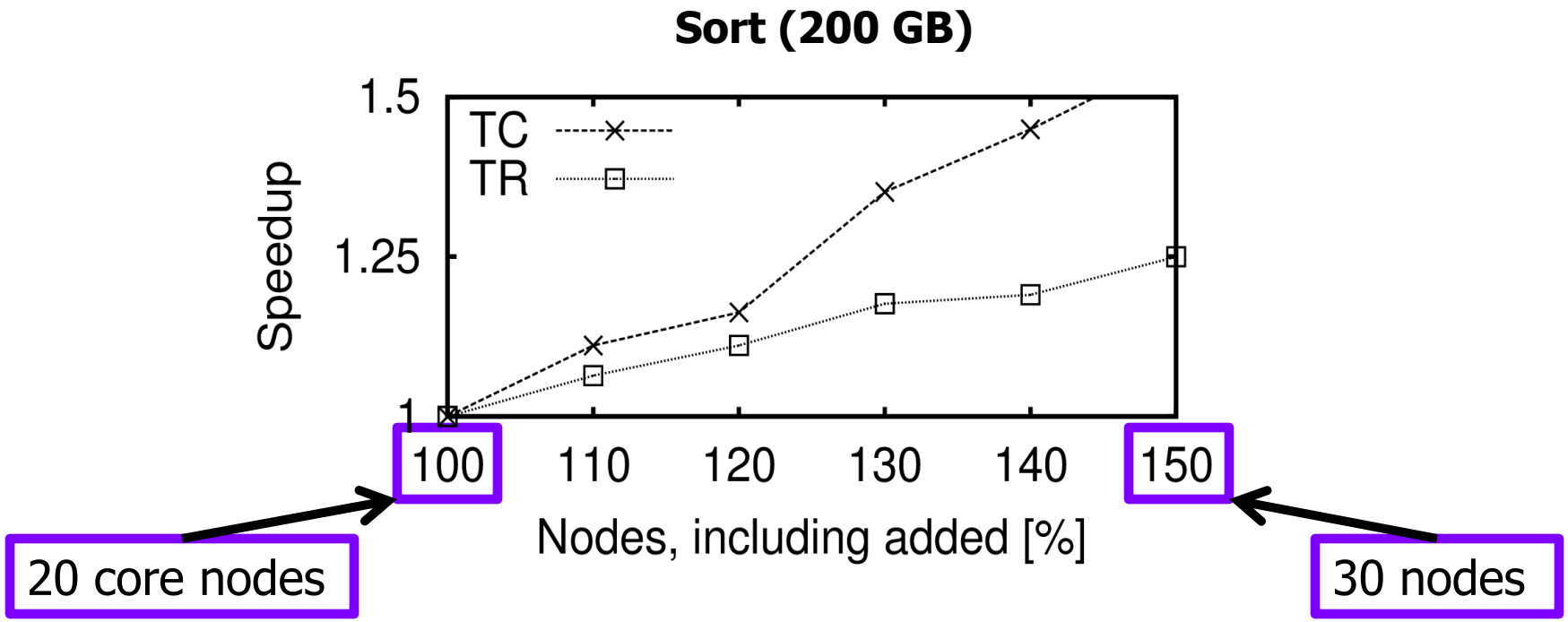
Speedup when growing (1/2)



TR nodes deliver good performance for CPU bound workloads

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

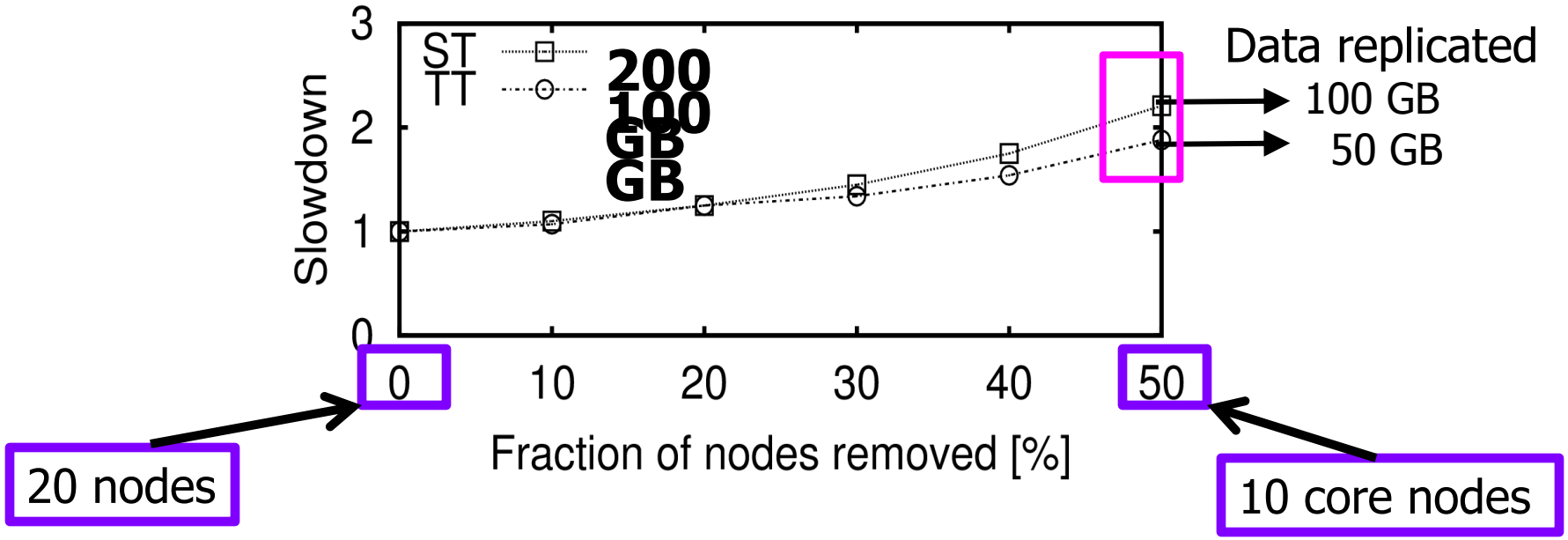
Speedup when growing (2/2)



(Only) TC nodes deliver good performance for disk-bound workloads

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

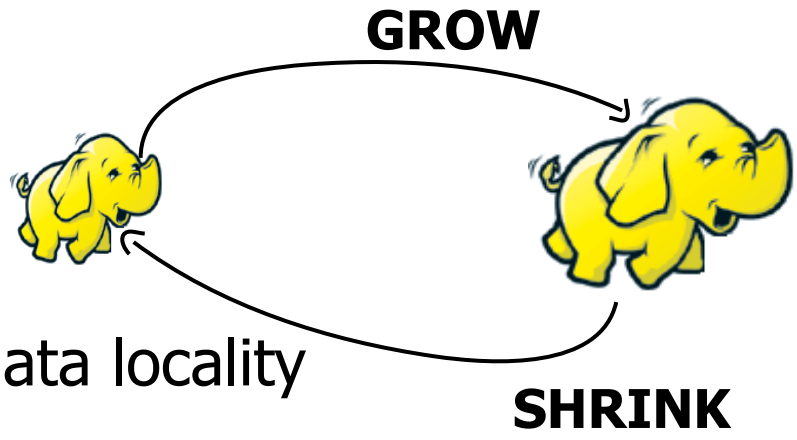
Slowdown when shrinking



Job slowdown increases linearly with the amount of replicated data

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

Take-home message



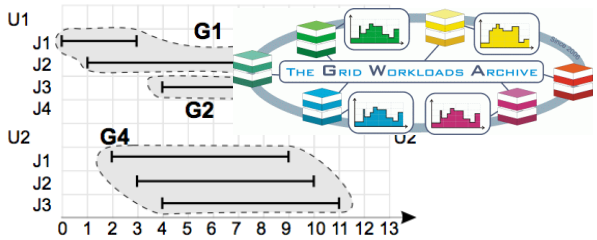
1. Dynamic MapReduce relaxes data locality
2. FAWKES policies can reduce imbalance between frameworks
3. More aggressive policies?

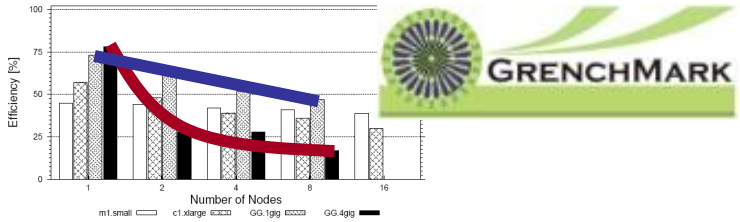
FAWKES

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

Agenda

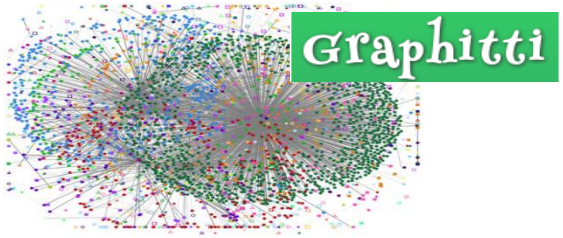
1. Everyone is a scientist!
2. Benchmarking: let's show the numbers





Data Center Workloads

Cloud Performance



Big Data: Graphs

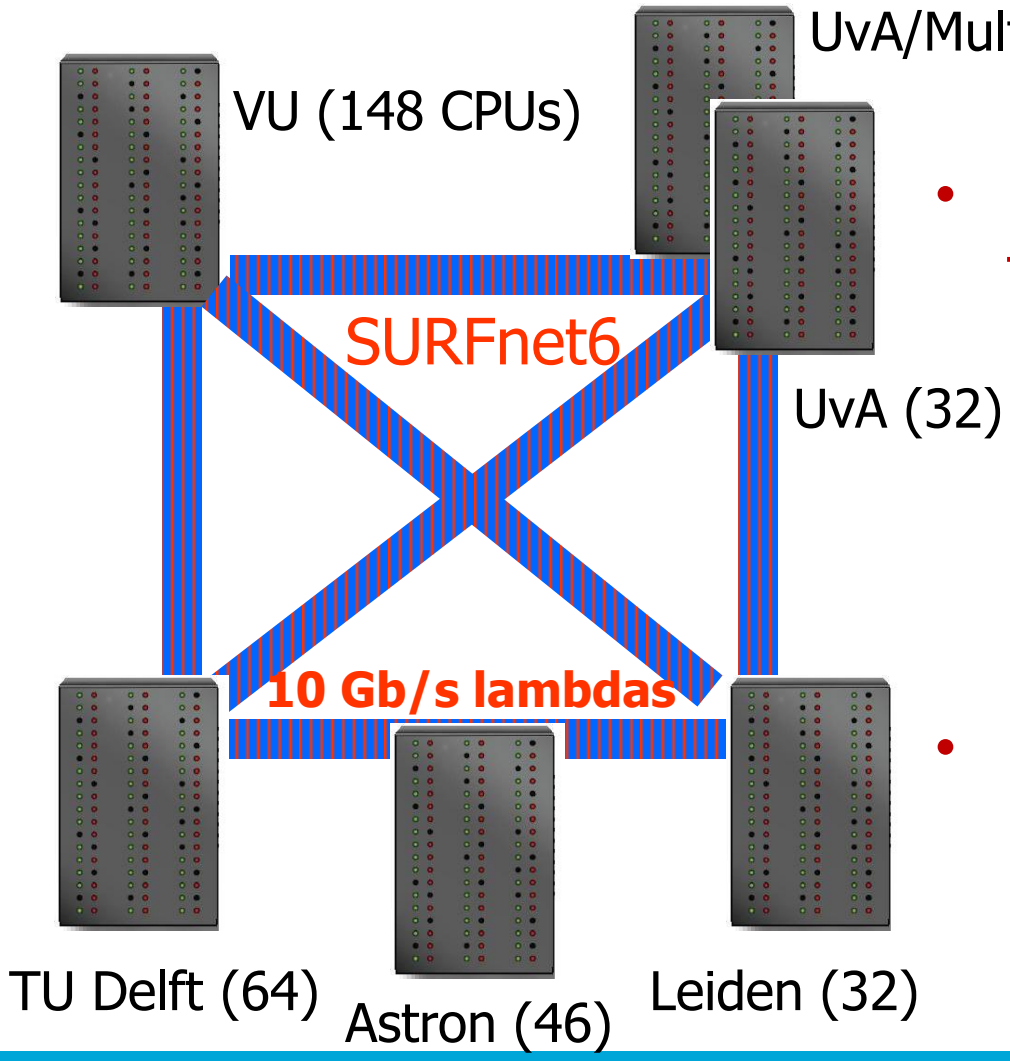
Big Data: BTWorld

Big Data: Elastic MR

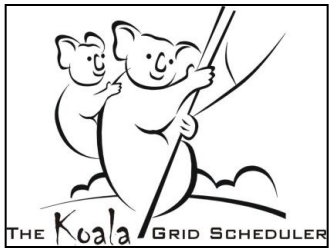
3. Conclusion

TU Delft

The DAS-4 Infrastructure

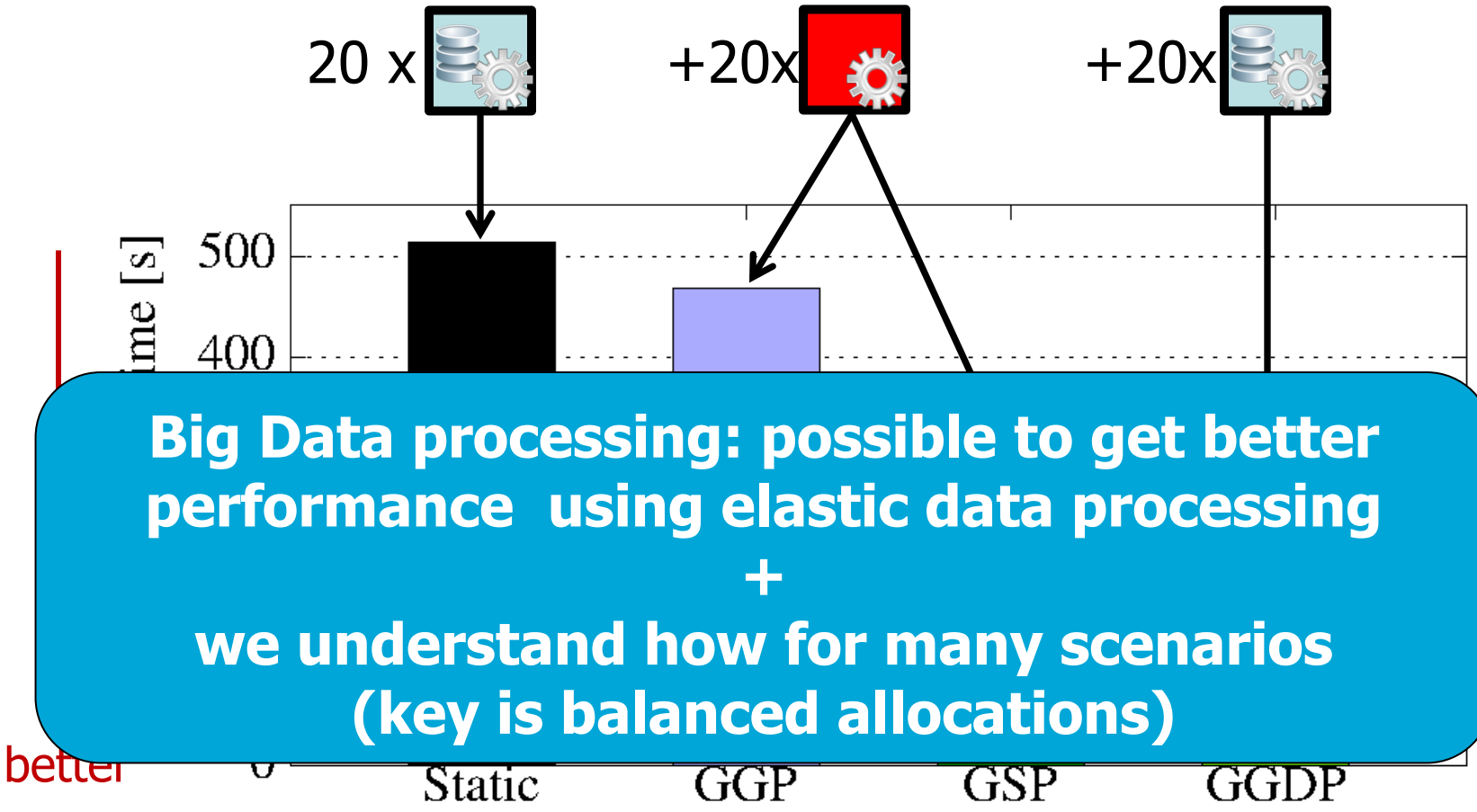


- Used for research in systems for over a decade
  - 1,600 cores (quad cores)
  - 2.4 GHz CPUs, GPUs
  - 180 TB storage
  - 10 Gbps Infiniband
  - 1 Gbps Ethernet
- Koala grid scheduler



TU Delft

Performance of Resizing using Static, Transient, and Core Nodes



Big Data processing: possible to get better performance using elastic data processing
+
we understand how for many scenarios (key is balanced allocations)

better

Sort + WordCount
(50 jobs, 1-50GB)

B. Ghit, N. Yigitbasi, A. Iosup, and D. Epema. Balanced Resource Allocations Across Multiple Dynamic MapReduce Clusters, SIGMETRICS 2014.

Conclusion: Take-Home Message

- **Big Data is necessary and grand challenge**
- **Big Data = Systems of Systems**
  - **Big data programming models have ecosystems**
  - **Stuck in stacks!**
  - Many trade-offs, many problems
- **In this talk**
  - Predictability challenges: we need to understand workload (modeling) and performance (benchmarking)
  - Early steps for benchmarking big data: graph processing, data processing workflows
  - Elasticity challenges: dynamic 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

Alexandru Iosup

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