Big Data in the Cloud: Enabling the Fourth Paradigm by Matching SMEs with Datacenters

60km
35mi

founded 1842
pop: 13,000

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.

May 14, 2014

2nd ISO/IEC JTC 1 Study Group on Big Data, Amsterdam

spec Research

TU Delft
Delft University of Technology

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

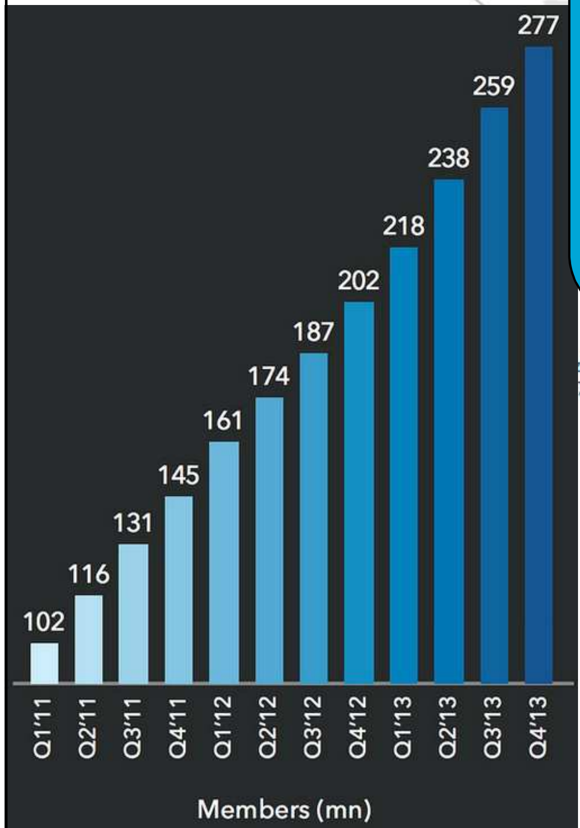
Vital for the prospective employees

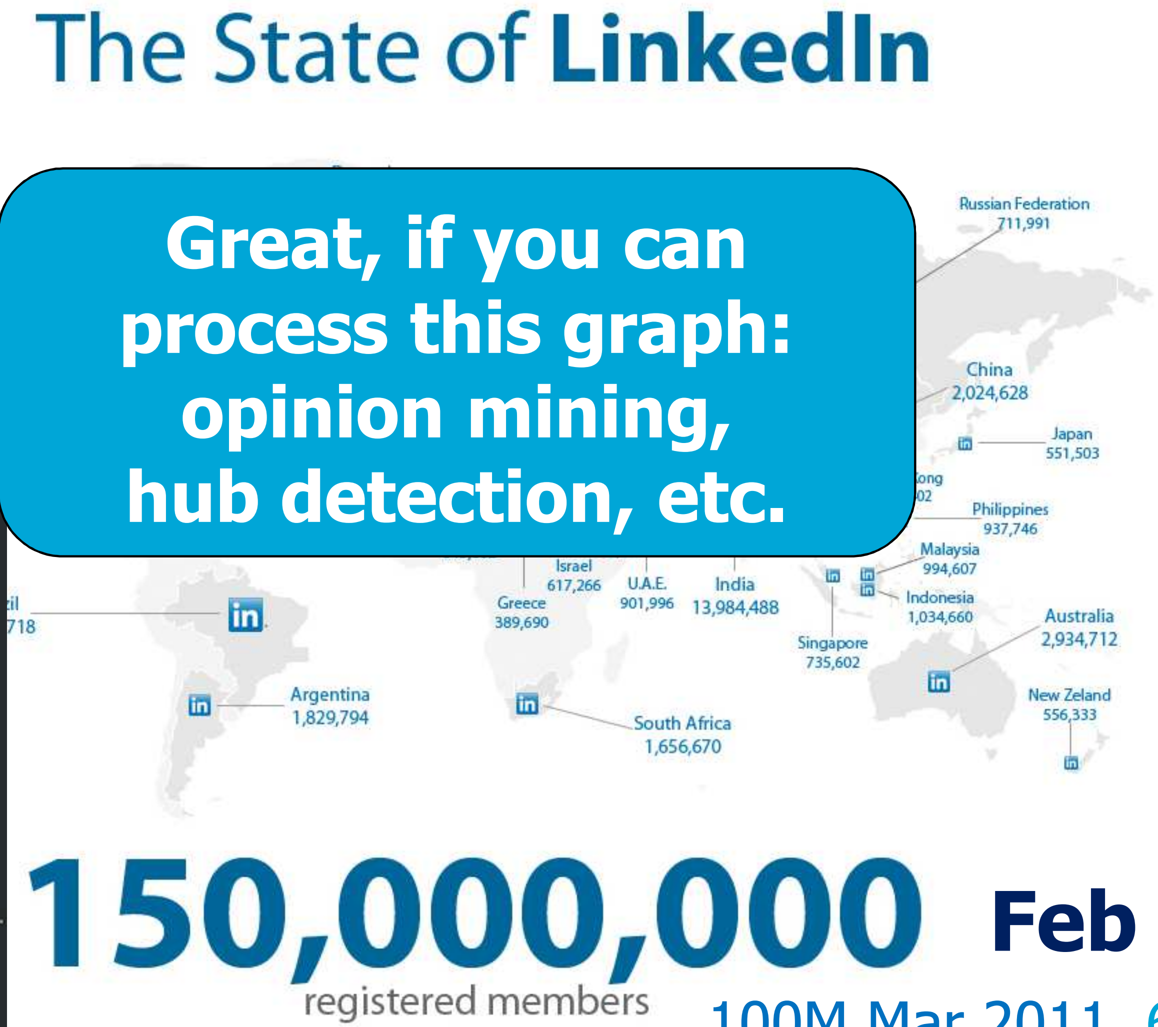
100M Mar 2011, 69M May 2010

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/

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.

Feb 2012

100M Mar 2011, 69M May 2010

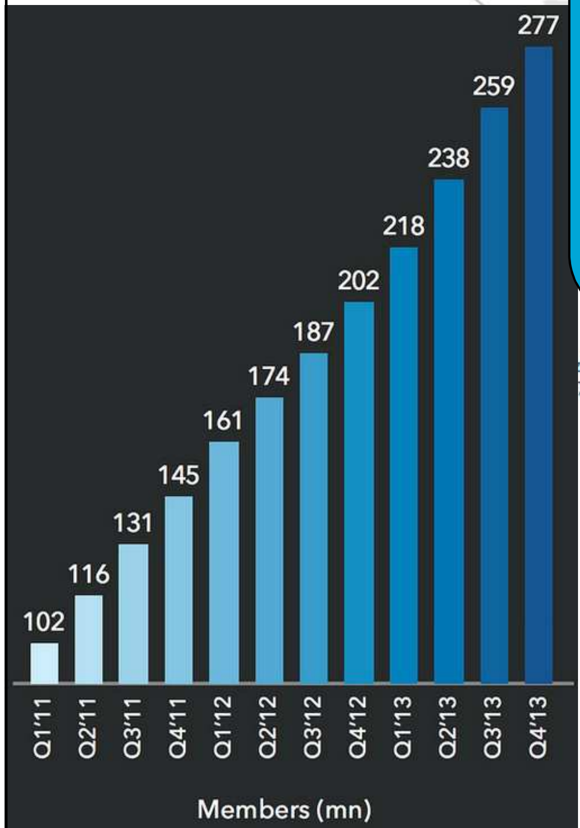
Sources: Vincenzo Cosenza, The State of LinkedIn, http://vincos.it/the-state-of-linkedin/
via Christopher Penn, http://www.shiftcomm.com/2014/02/state-linkedin-social-media-dark-horse/

TU Delft

Data at the Core of Our Society: The LinkedIn Example

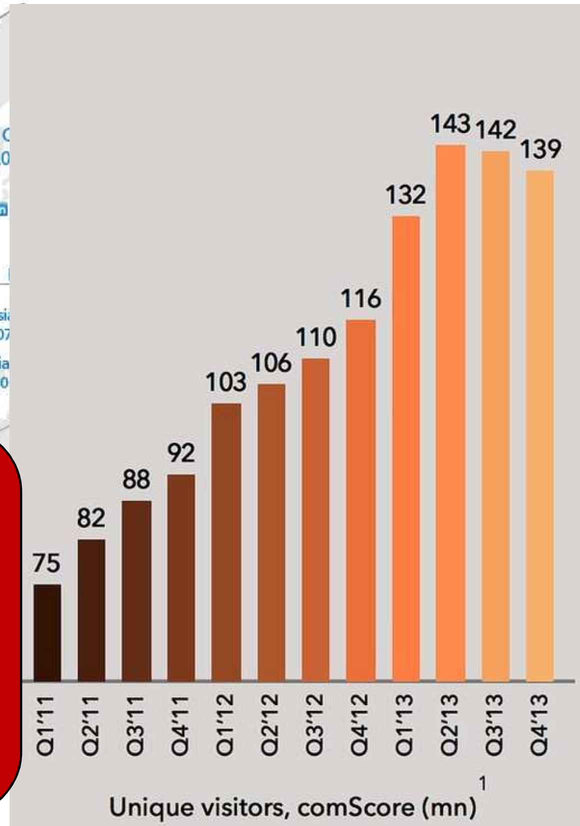
The State of LinkedIn

3-4 new users every second



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

but fewer visitors (and page views)



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

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/

LinkedIn Is Part of the "Data Deluge"

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

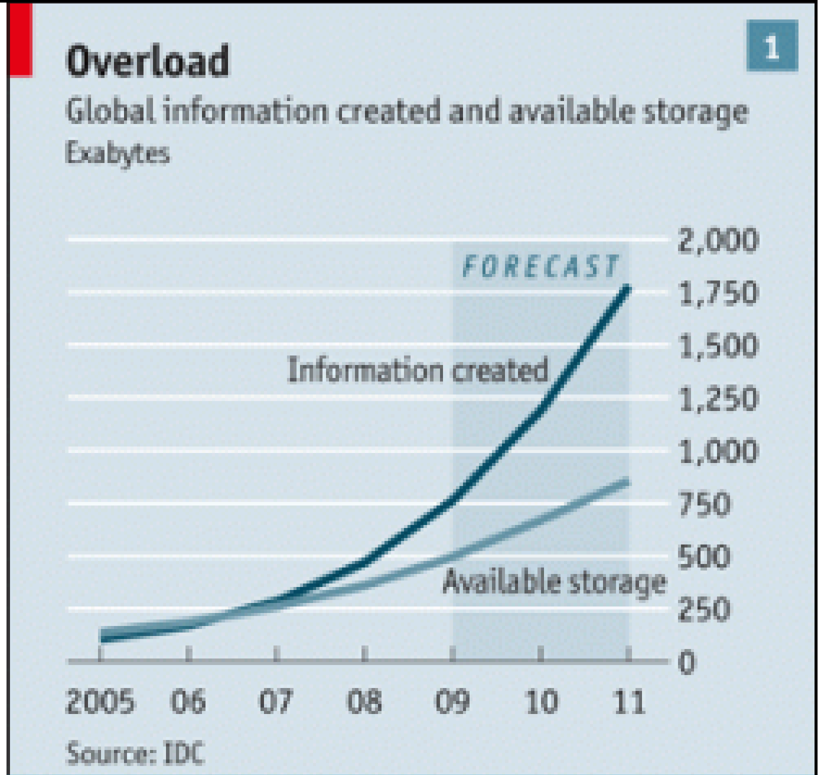
- Interacting
- Understanding
- Deciding
- Creating

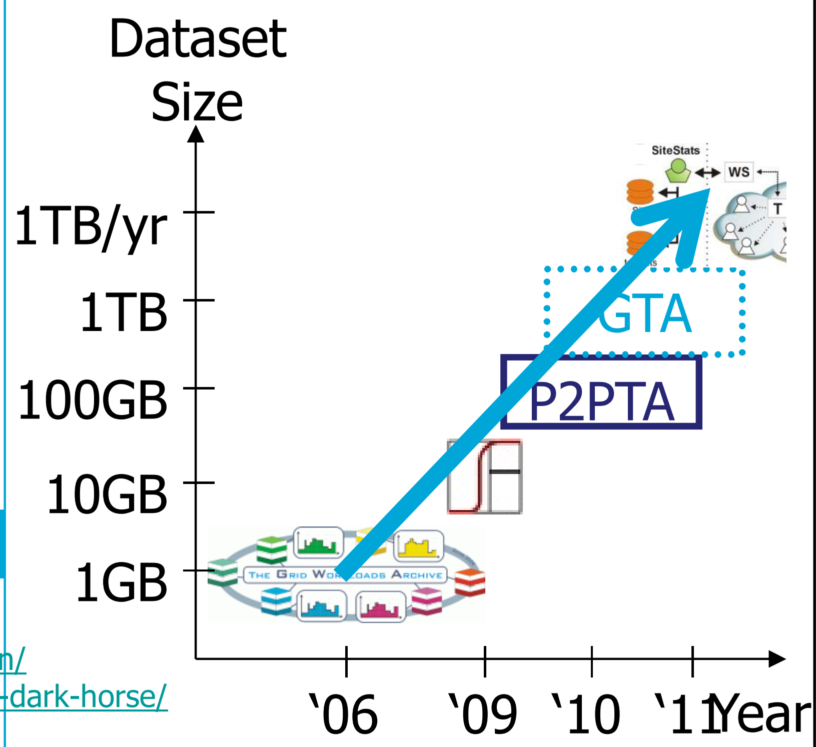
Sources: IDC, EMC.

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 queries later

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

- 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

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 Opportunity, via a Detour (An Anecdotal Example)
The Overwhelming Growth of Knowledge

“When 12 men founded the
Royal Society in 1660, it was
possible for a[illegible]
person to en[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

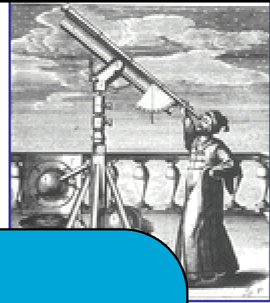
Professionals already know they don’t know [it all]

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.

The Opportunity, via a Detour

From Hypothesis to Data

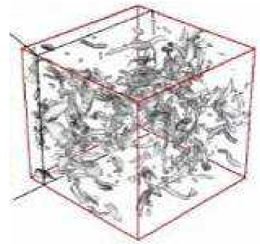


The Fourth Paradigm is suitable for professionals who already know they don't know [enough to formulate good hypotheses], yet need to deliver quickly

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",
http://research.microsoft.com/en-us/collaboration/fourthparadigm/

TU Delft

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

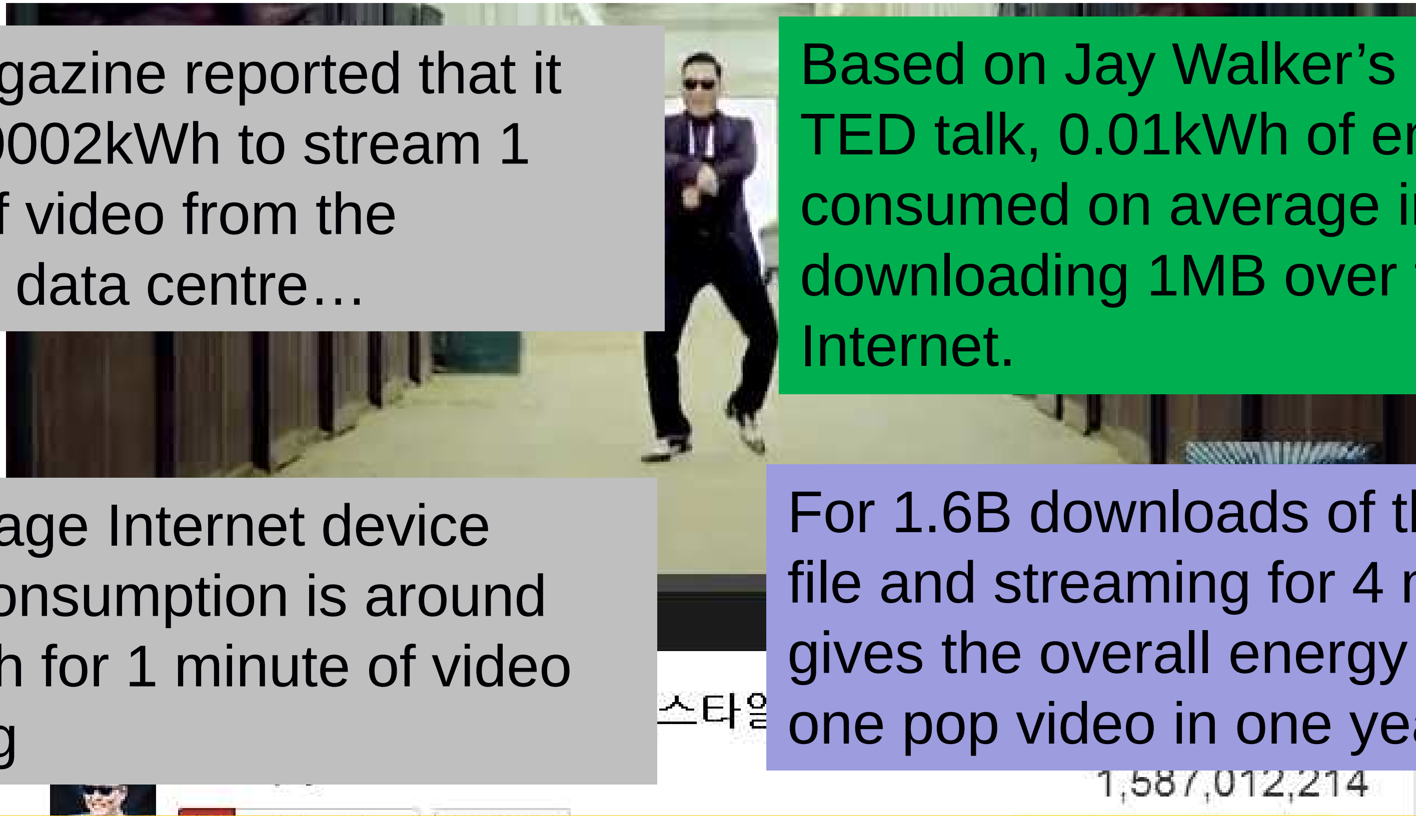
- Data as individual right, enabling private lifestyle and modern **societal services**
- Data as workhorse in creating **services** for SMEs (~60% gross value added, for many years)

EC 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

Can We Afford This Vision, with the Current Technology and 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…

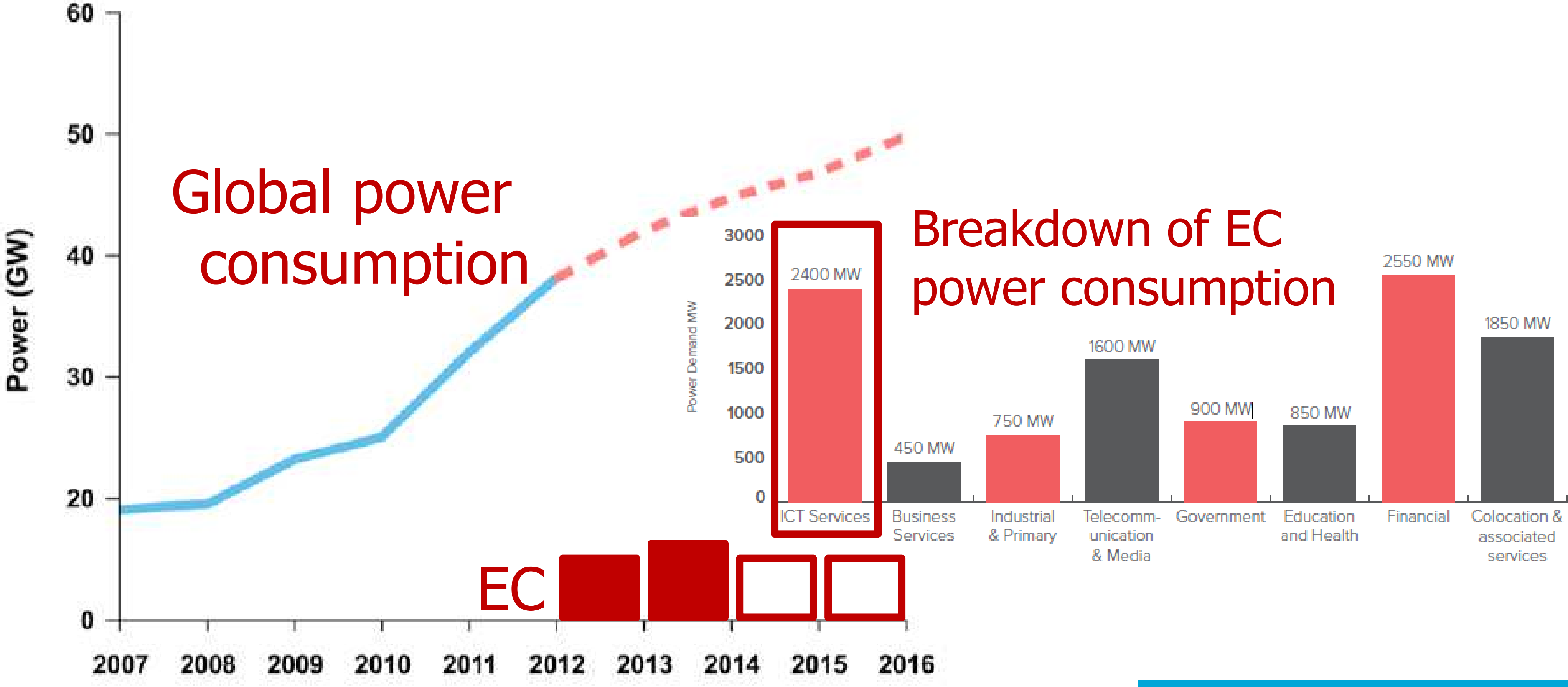
312GWh = more than some countries in a year, 36MW of 24/7/365 diesel, 100M liters of Oil, 80,000 cars running for a year, ...

Source: Ian Bitterlin and Jon Summers, UoL, UK.

TU Delft

Can We Afford This Vision, with the Current Technology and Resources?

- Not with the current technology (in this presentation)
- Not with the current resources (energy, human, …)



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.

Our Big Data Team, PDS Group at TU Delft (http://www.pds.ewi.tudelft.nl/)

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

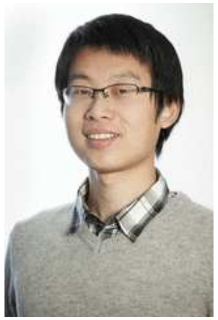
Dick Epema
TU Delft
Big Data & Clouds
Res. management
Systems

Bogdan Ghit
TU Delft
Systems
Workloads

Ana Lucia Varbanescu
U. Amsterdam
Graph processing
Benchmarking

Claudio Martella
VU Amsterdam
Graph processing

Mihai Capota
TU Delft
Big Data apps
Benchmarking

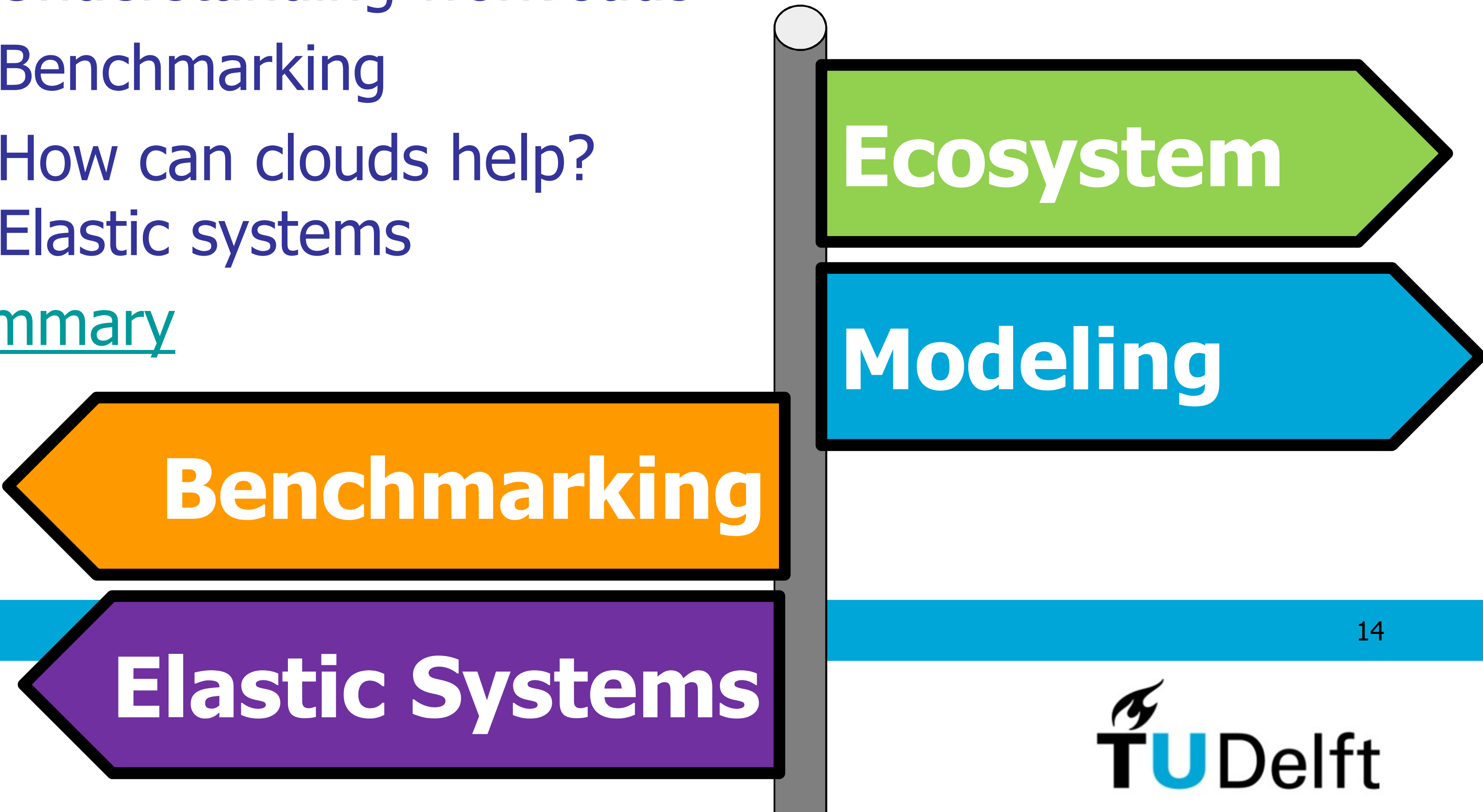


Yong Guo
TU Delft
Graph processing
Benchmarking

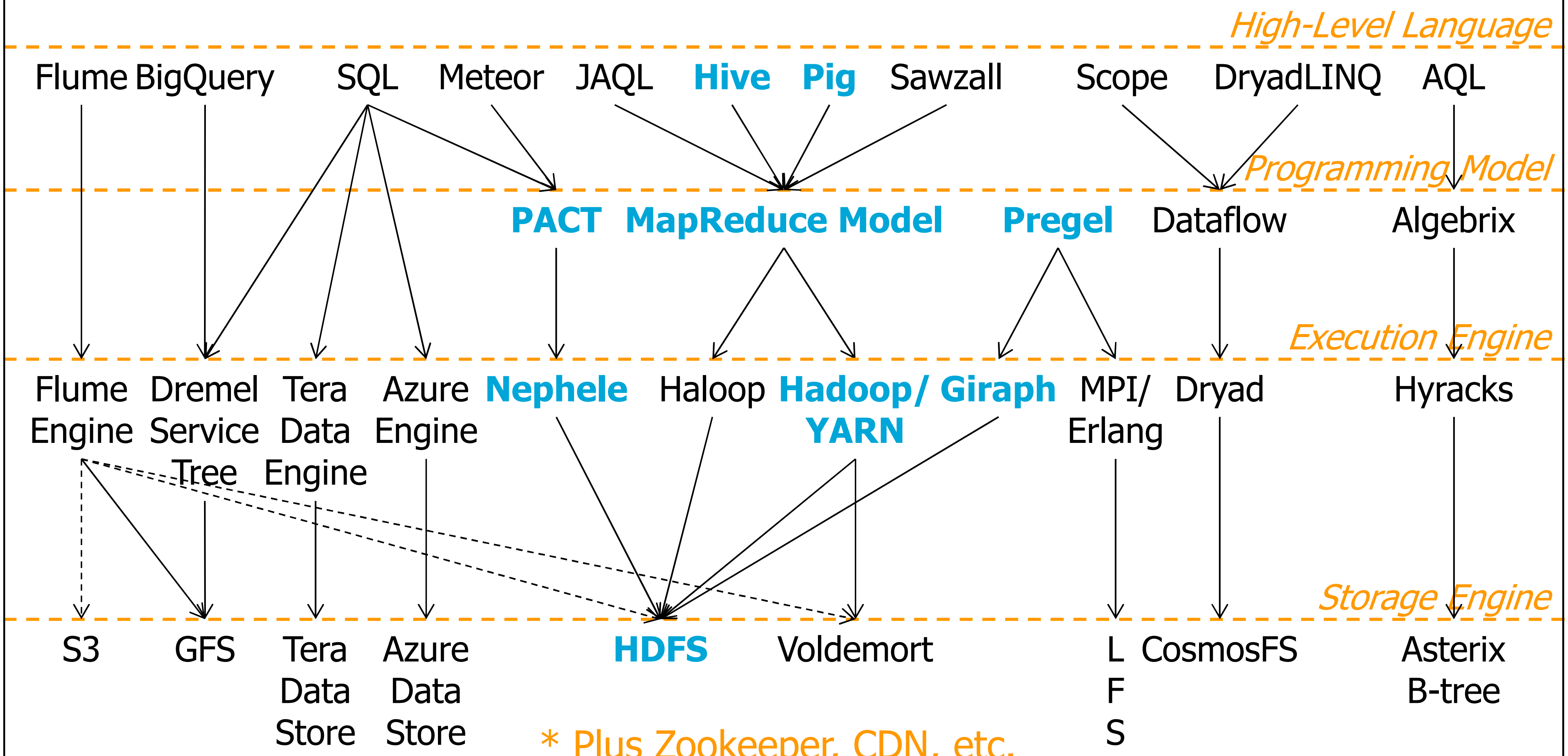
Marcin Biczak
TU Delft
Big Data & Clouds
Performance & Development

Agenda

1. Big Data, Our Vision, Our Team
2. **Big Data on Clouds**
   1. The Big Data ecosystem
   2. Understanding workloads
   3. Benchmarking
   4. How can clouds help? Elastic systems
3. Summary



The Current Technology
Big Data = Systems of Systems



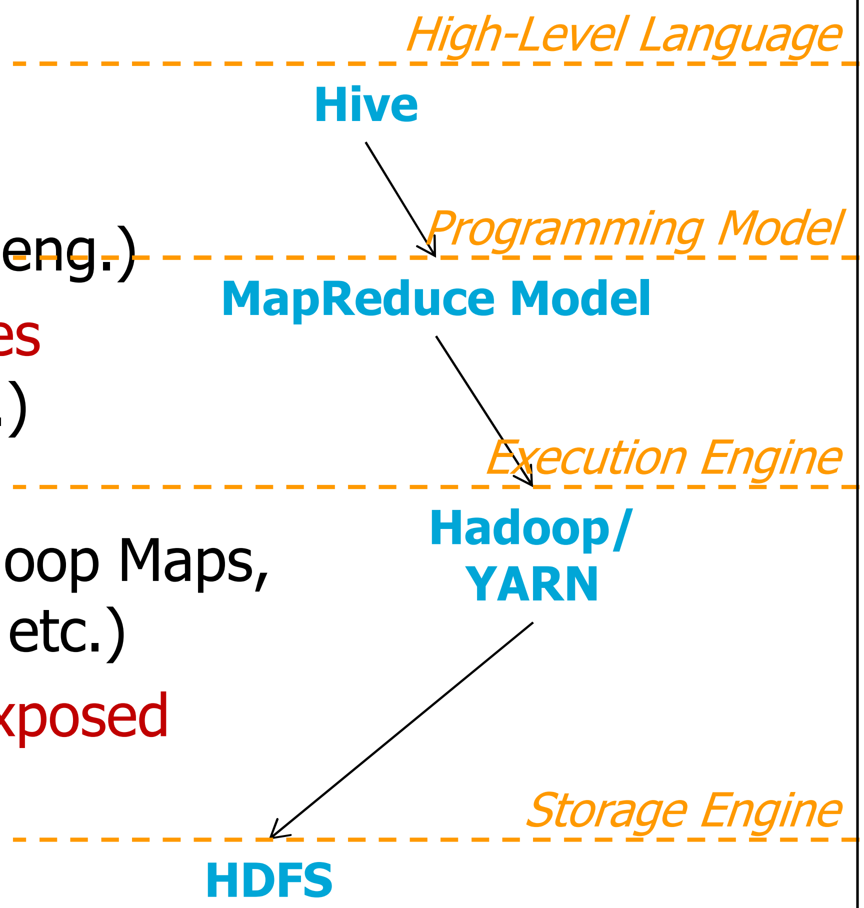
* Plus Zookeeper, CDN, etc.

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

TU Delft

The Problem: 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.)
  - …



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

Instead…

Many-Task Big-Data Processing on Heterogeneous Resources: from GPUs to Clouds

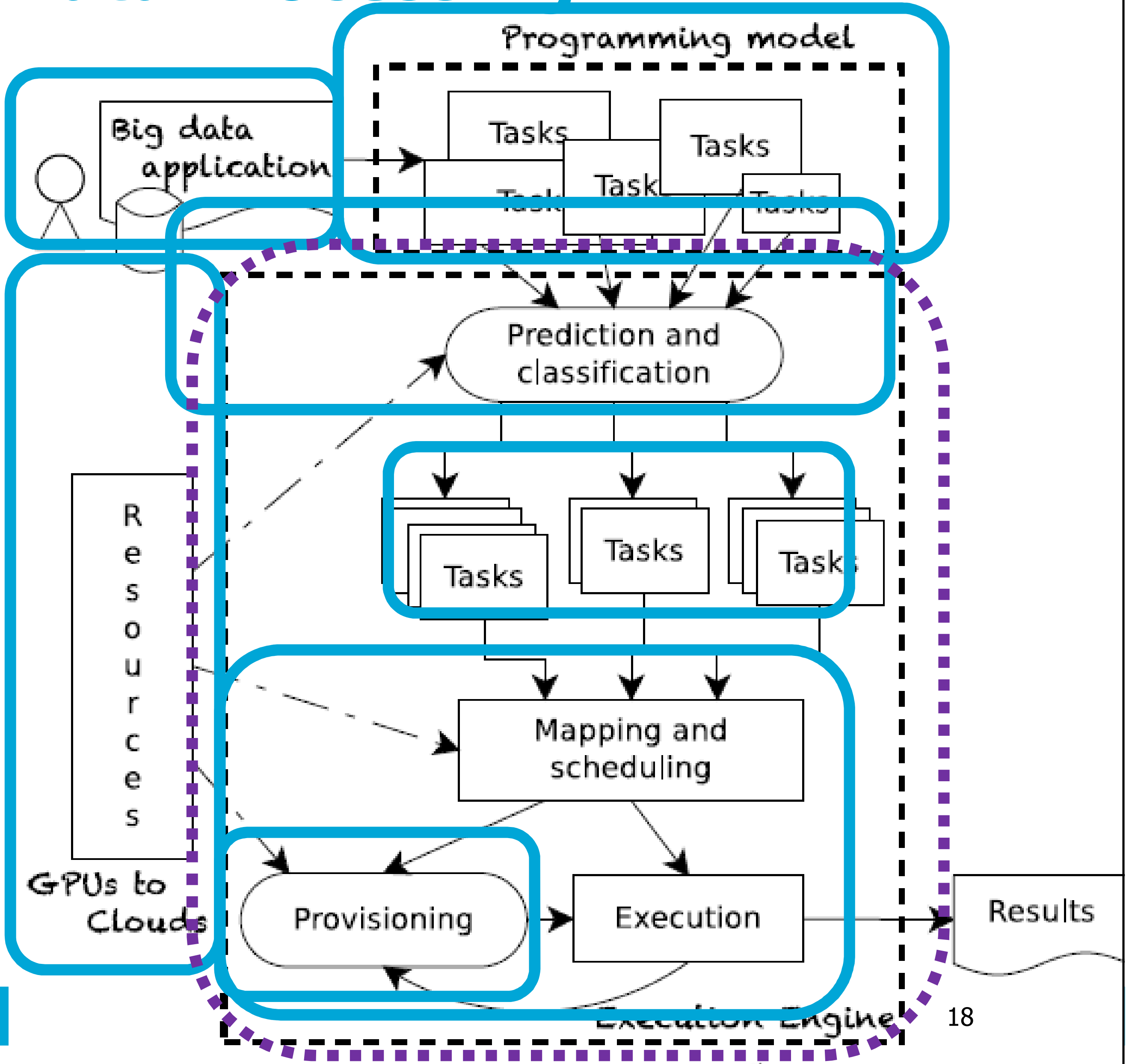
1. Take Big-Data Processing applications
2. Split into Many Tasks
3. Each of the tasks parallelized to match resources
4. Execute each Task on the most efficient resource
5. Exploiting the massive parallelism available now and increasing in the combination multi-core CPUs & GPUs
6. Using the set of resources provided by local clusters
7. And exploiting the efficient elasticity of IaaS clouds

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

A Generic Architecture for Many-Task Big Data Processing

Execute Big Data apps as many tasks using mixed resources:

1. High performance
2. Elasticity
3. Predictability
4. Compatibility



TUDelft

10 Main Challenges in 4 Categories*

* List not exhaustive

High Performance

1. **Parallel architectures and algorithms**—support from start
2. **Heterogeneous platforms**—application and data decomposition
3. Programmability by portability (OpenCL/ACC/...)

Elasticity

1. Performance and cost-awareness under elasticity—**elastic data**
2. **Portfolio scheduling**
3. Social awareness

Predictability

1. **Modeling**
2. **Benchmarking**

Compatibility

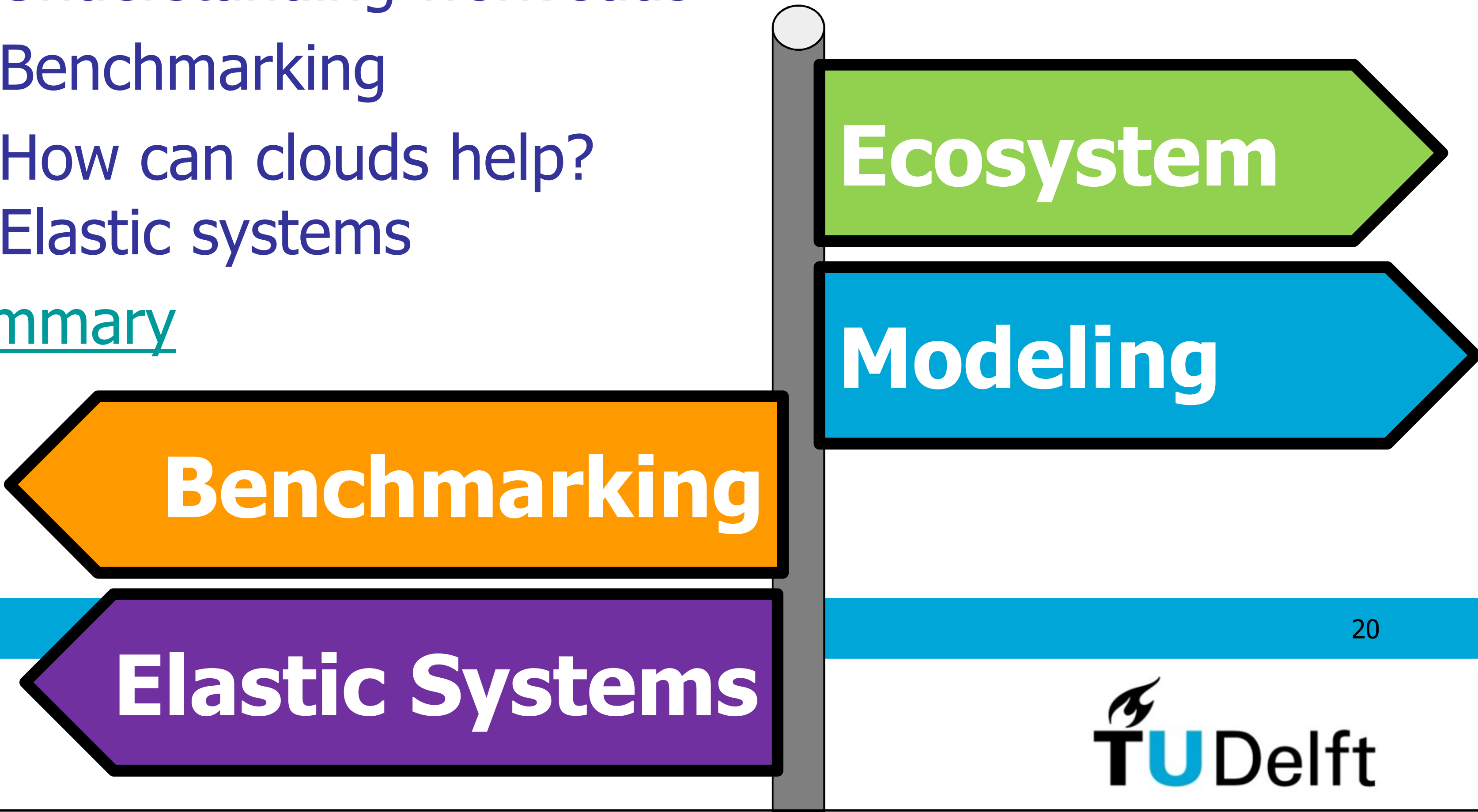
1. Interfacing with the application
2. Storage management

Varbanescu and Iosup, On Many-Task Big Data Processing: from GPUs to Clouds, MTAGS 2013. Proc. of SC13. (invited paper)

Ad: Read our article

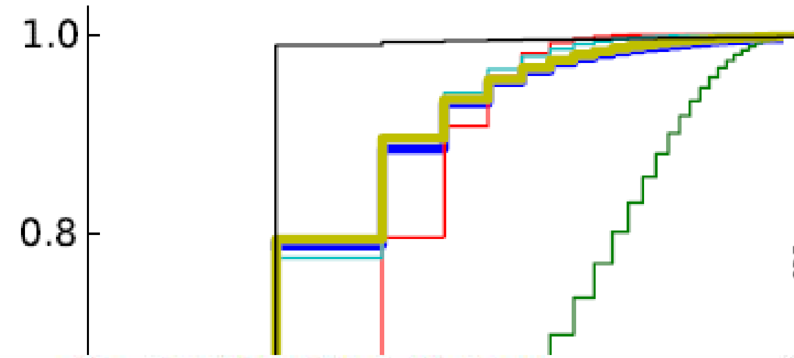
Agenda

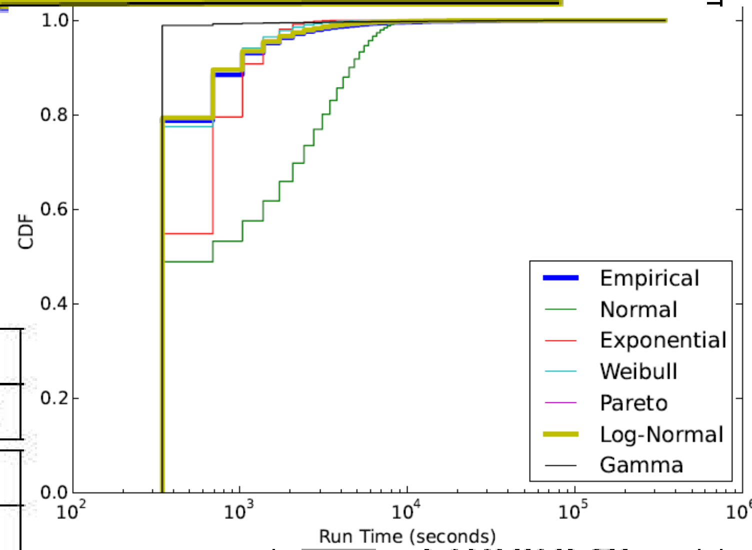
1. Big Data, Our Vision, Our Team
2. **Big Data on Clouds**
    1. The Big Data ecosystem
    2. Understanding workloads
    3. Benchmarking
    4. How can clouds help? Elastic systems
3. Summary



Statistical MapReduce Models From Long-Term Usage Traces

- Real traces
  - Yahoo
  - Google
  - 2 x Social N





Job	Task Count	Best-Fitting Distribution: Task Run Time	Task CPU	Task Memory
1	1	Log-Normal	Weibull	Exponential
2	128	Weibull	Exponential	Weibull
3	128	Log-Normal	Weibull	Weibull
Overall Best Fit		*Log-Normal (129)*	*Weibull (129)*	*Weibull (256)*

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 .

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

- BitTorrent: swarms of people sharing files
  - 100M users
  - At some point 35% of total internet traffic
- Data-driven project: data first, ask questions later
- Over 14TB 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

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

MapReduce-based Workflow for the BTWorld Use Case

Query Diversity

- Queries use different operators, stress different parts of system
- 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

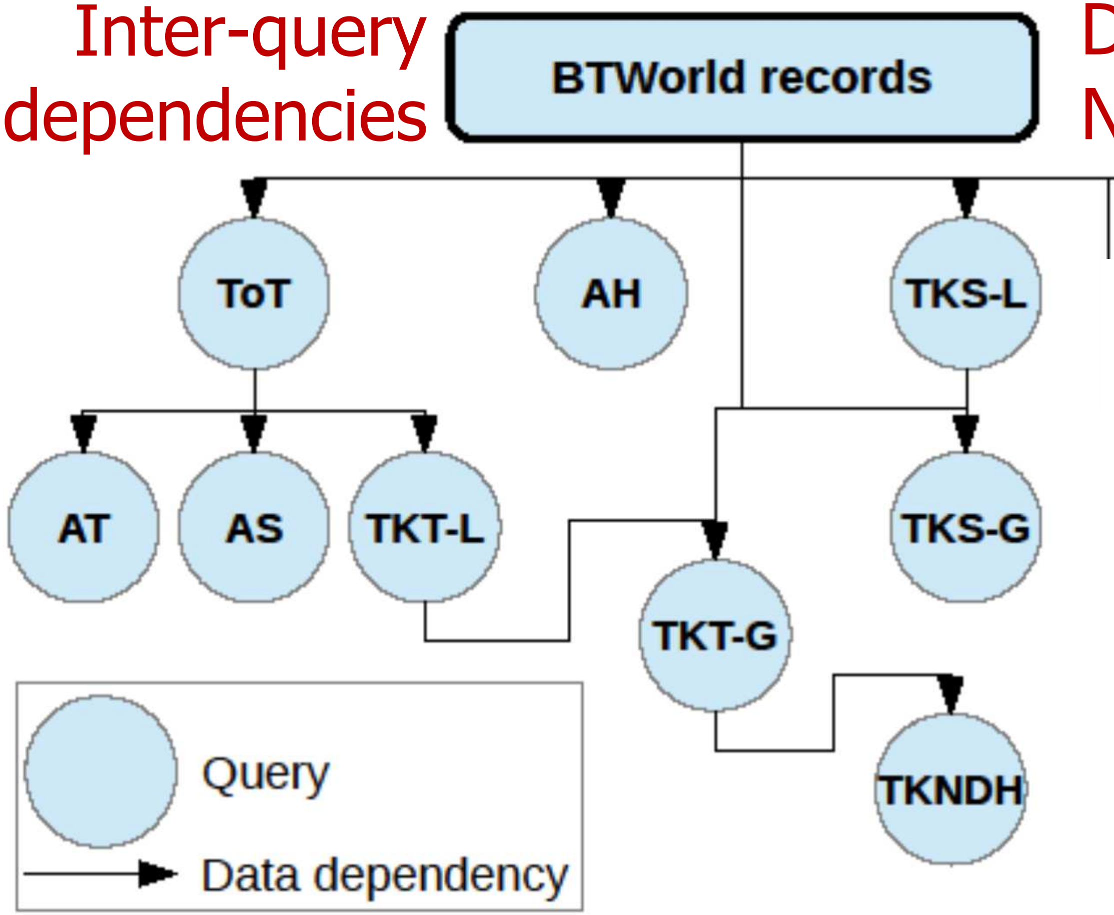
MapReduce Is Now Part of Workflows

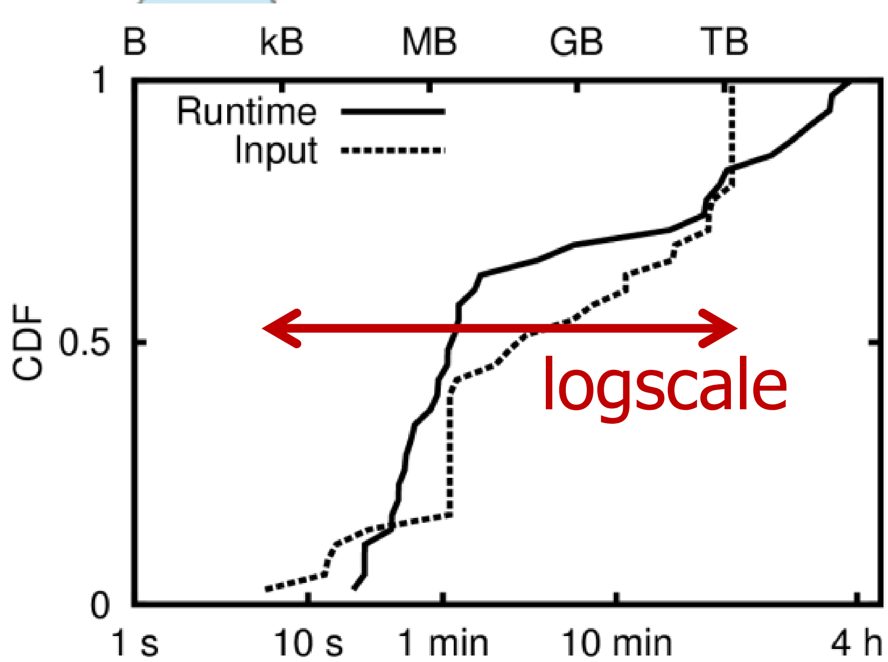
Use Case: Monitoring Large-Scale Distributed Computing System with 160M users

Inter-query dependencies

Diverse queries

New queries during project

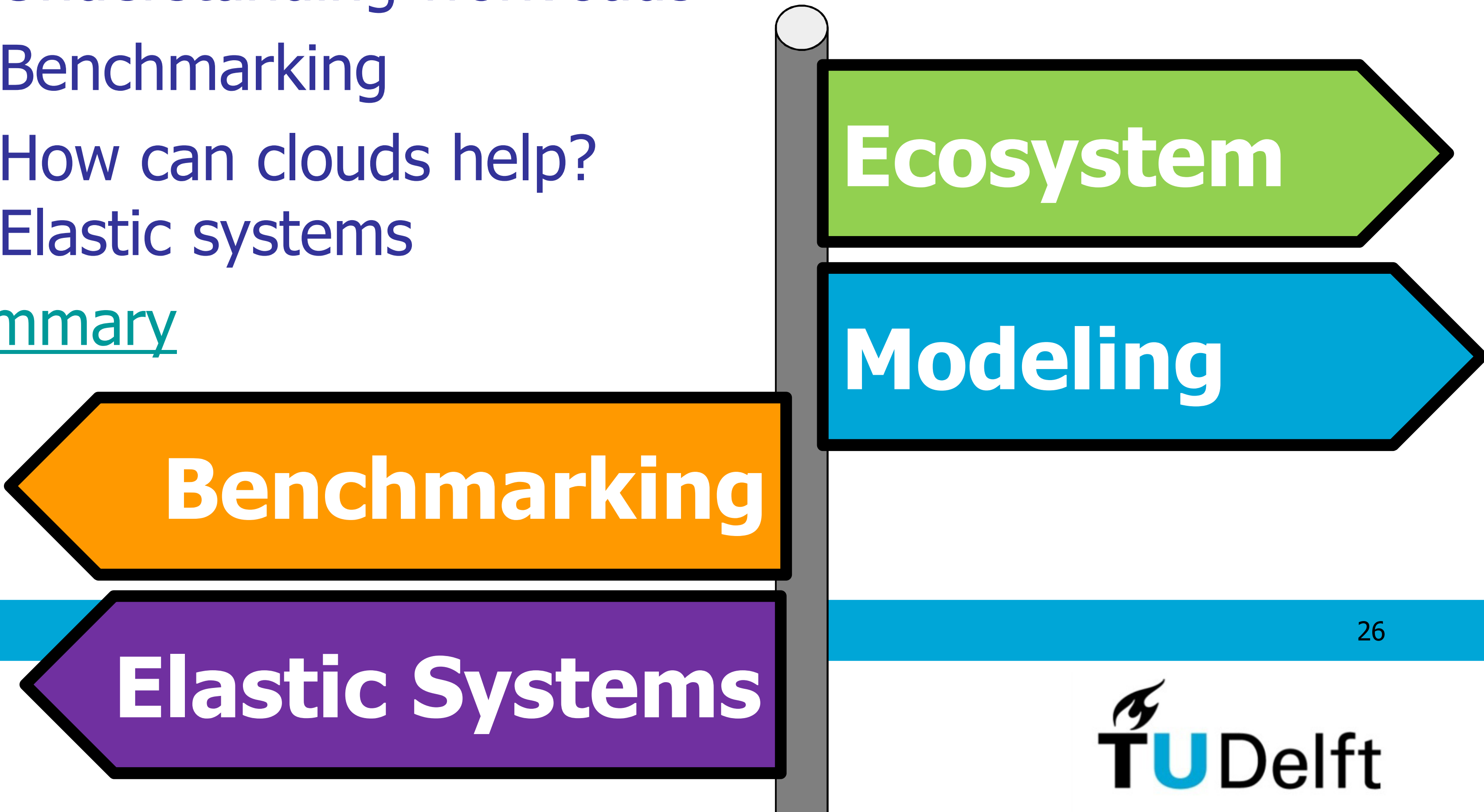




Hegeman, Ghit, Capotã, Hidders, Epema, Iosup. The BTWorld Use Case for Big Data Analytics: Description, MapReduce Logical Workflow, and Empirical Evaluation.IEEE BigData'13

Agenda

1. Big Data, Our Vision, Our Team
2. **Big Data on Clouds**
    1. The Big Data ecosystem
    2. Understanding workloads
    3. Benchmarking
    4. How can clouds help? Elastic systems
3. Summary

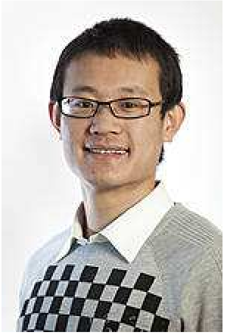


TUDelft

Performance: Our Team Also Includes…

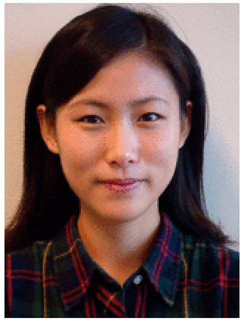
Ana Lucia Varbanescu
U. Amsterdam

Performance modeling
Parallel systems
Multi-core systems



Jianbin Fang
TU Delft

Parallel systems
Multi-core systems
Tianhe/Xeon Phi



Jie Shen
TU Delft

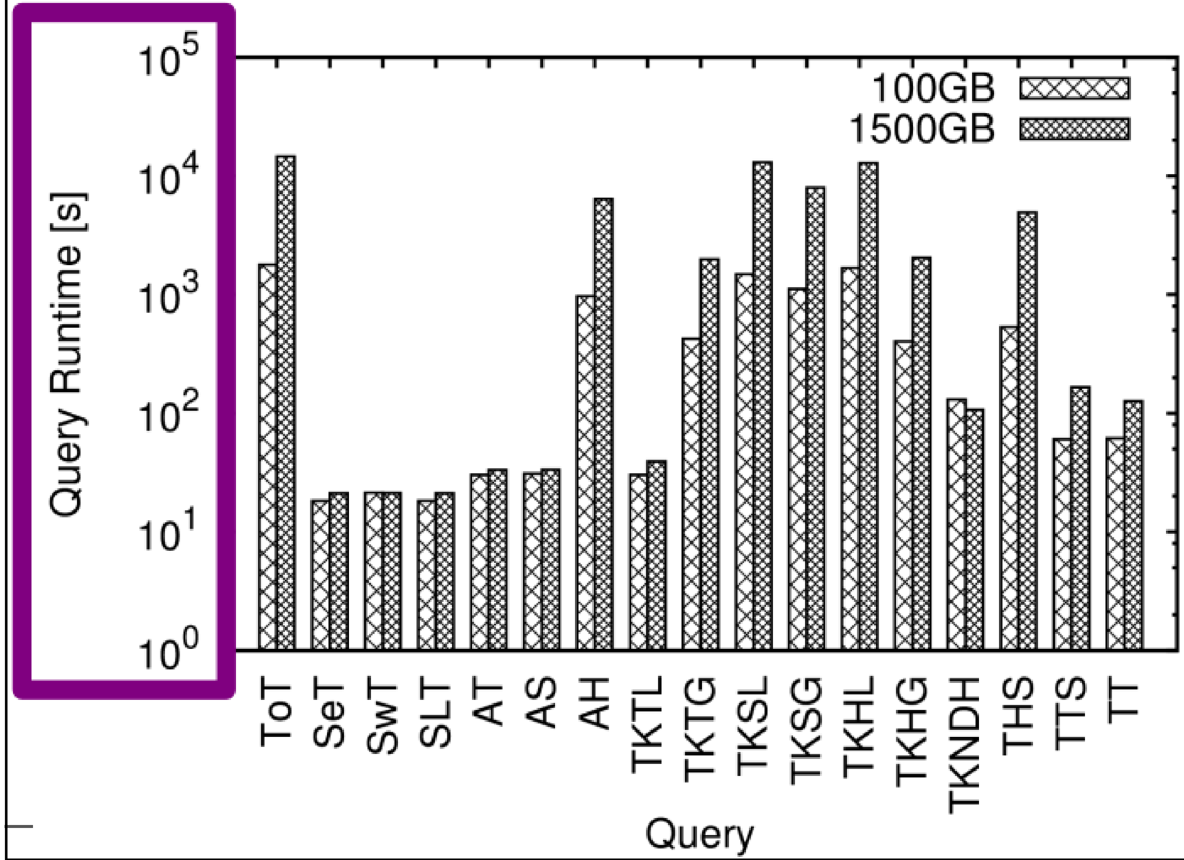
Performance evaluation
Parallel systems
Multi-core systems

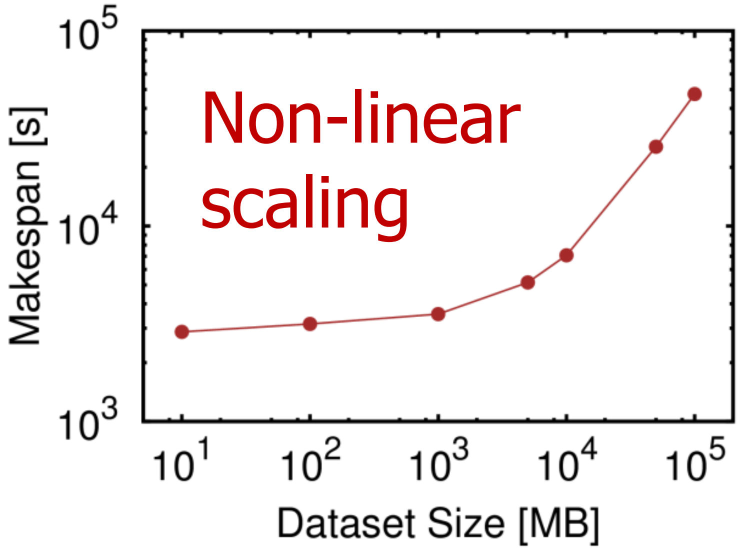
Alexandru Iosup
TU Delft

Performance modeling
Performance evauluation

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

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

More information: http://research.spec.org

TUDelft
Delft University of Technology

Benchmarking

- From single kernel or solitary-kernel suite to …
  Big Data processing workflow
- Derived from modeling …
  Intra-query, intra-job, and inter-job data dependencies
- Can benchmarking be
  - Realistic?
  - Cost- and time-effective?
  - Fair?

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

Survey of graph algorithms

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

Selection of algorithms

A1: General Statistics (STATS: # vertices and edges, LCC)

- Single step, low processing, decision-making

A2: Breadth First Search (BFS)

- Iterative, low processing, building block

A3: Connected Component (CONN)

- Iterative, medium processing, building block

A4: Community Detection (CD)

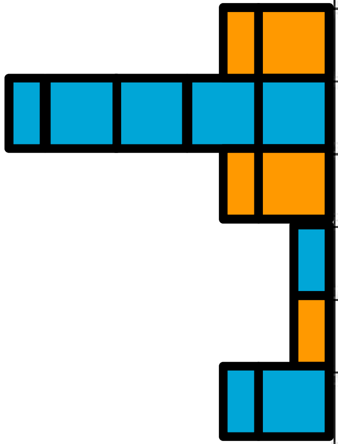
- Iterative, medium or high processing, social network

A5: Graph Evolution (EVO)

- Iterative (multi-level), high processing, prediction

The Science: Which Algorithms?

- (DONE) Our own survey, related to graph-processing
  - Academic publications (CIKM, ICDE, SIGKDD, SIGMOD, VLDB, CCGRID, HPDC, IPDPS, PPoPP, SC)



Class	Typical algorithms
General Statistics	Triangulation [36], Diameter [37], BC [38]
Graph Traversal	BFS, DFS, Shortest Path Search
Connected Components	MIS [39], BiCC [40], Reachability [41]
Community Detection	Clustering, Nearest Neighbor Search
Graph Evolution	Forest Fire Model [1], Preferential Attachment Model [42]
Other	Sampling, Partitioning

Guo, Biczak, Varbanescu, Iosup, Martella, Willke. How Well do Graph-Processing Platforms Perform? An Empirical Performance Evaluation and Analysis, IPDPS14

Selection of graphs

- 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/

The Science: Dataset sizes? Machines in cluster?

- Our own survey, related to graph-processing
  - Academic publications (CIKM, ICDE, SIGKDD, SIGMOD, VLDB, CCGRID, HPDC, IPDPS, PPoPP, SC)

Platforms	Algorithms	Dataset type	Largest dataset	System
Neo4j, MySQL [40]	1 other	synthetic	100 KV	1 C
Neo4j, etc. [4]	3 others	synthetic	1 MV	1 C
Pregel [5]	1 other	synthetic	50 BV	300 C
GPS, Giraph [41]	CONN, 3 others	real	39 MV, 1.5 BE	60 C
			1 BV	16 C
			282 MV	90 C

Dataset size: 100sMB—10s GB

System size: <10—100s nodes

Guo, Biczak, Varbanescu, Iosup, Martella, Willke. How Well do Graph-Processing Platforms Perform? An Empirical Performance Evaluation and Analysis, IPDPS14

TU Delft

Benchmarking suite
Platforms and Process

http://bit.ly/10hYdIU

- Platforms

YARN

Giraph

- Process
  - Evaluate baseline (out of the box) and tuned performance
  - Evaluate performance on fixed-size system
  - Future: evaluate performance on elastic-size system
  - Evaluate scalability

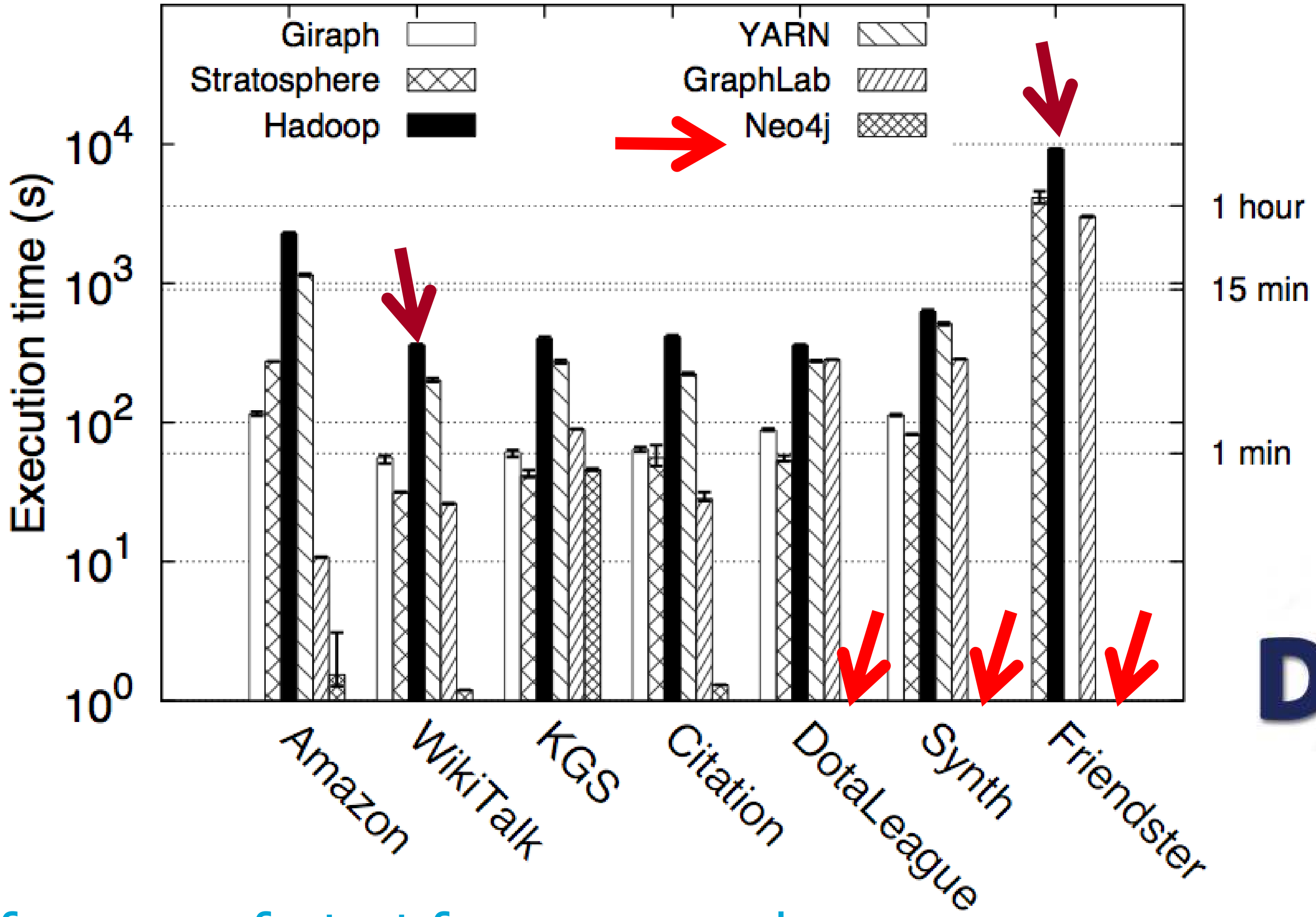
Guo, Biczak, Varbanescu, Iosup, Martella, Willke. Benchmarking Graph-Processing Platforms: A Vision. Proc. of ICPE 2014.

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

Experimental setup

- Size
  - Most experiments take 20 working nodes
  - Up to 50 working nodes
- DAS4: a multi-cluster Dutch grid/cloud
  - Intel Xeon 2.4 GHz CPU (dual quad-core, 12 MB cache)
  - Memory 24 GB
  - 10 Gbit/s Infiniband network and 1 Gbit/s Ethernet network
  - Utilization monitoring: Ganglia
- HDFS used here as distributed file systems

BFS: results for all platforms, all data sets



- No platform runs fastest for every graph
- Not all platforms can process all graphs
- Hadoop is the worst performer

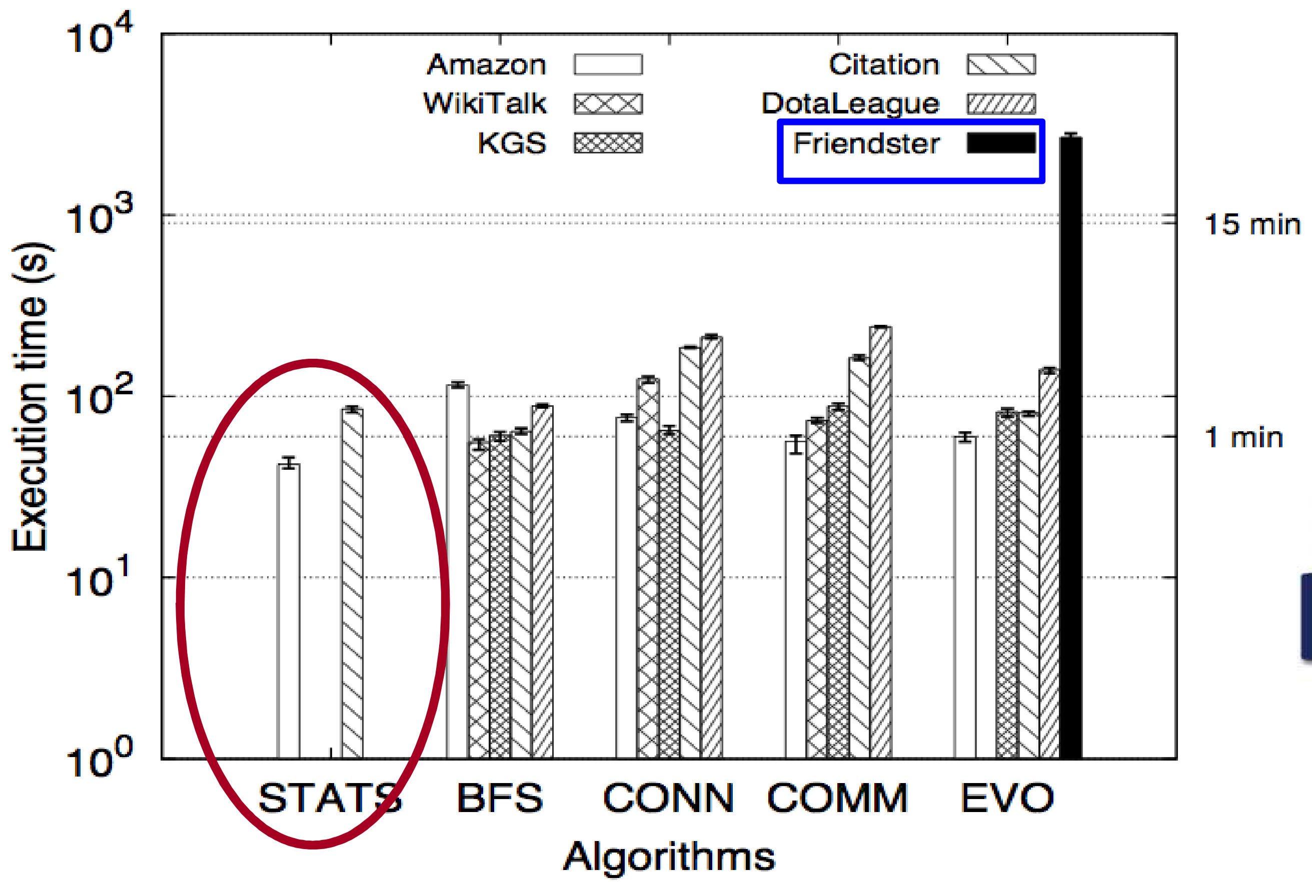
http://bit.ly/10hYdIU

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

Graphitti

Giraph: results for all algorithms, all data sets

http://bit.ly/10hYdIU

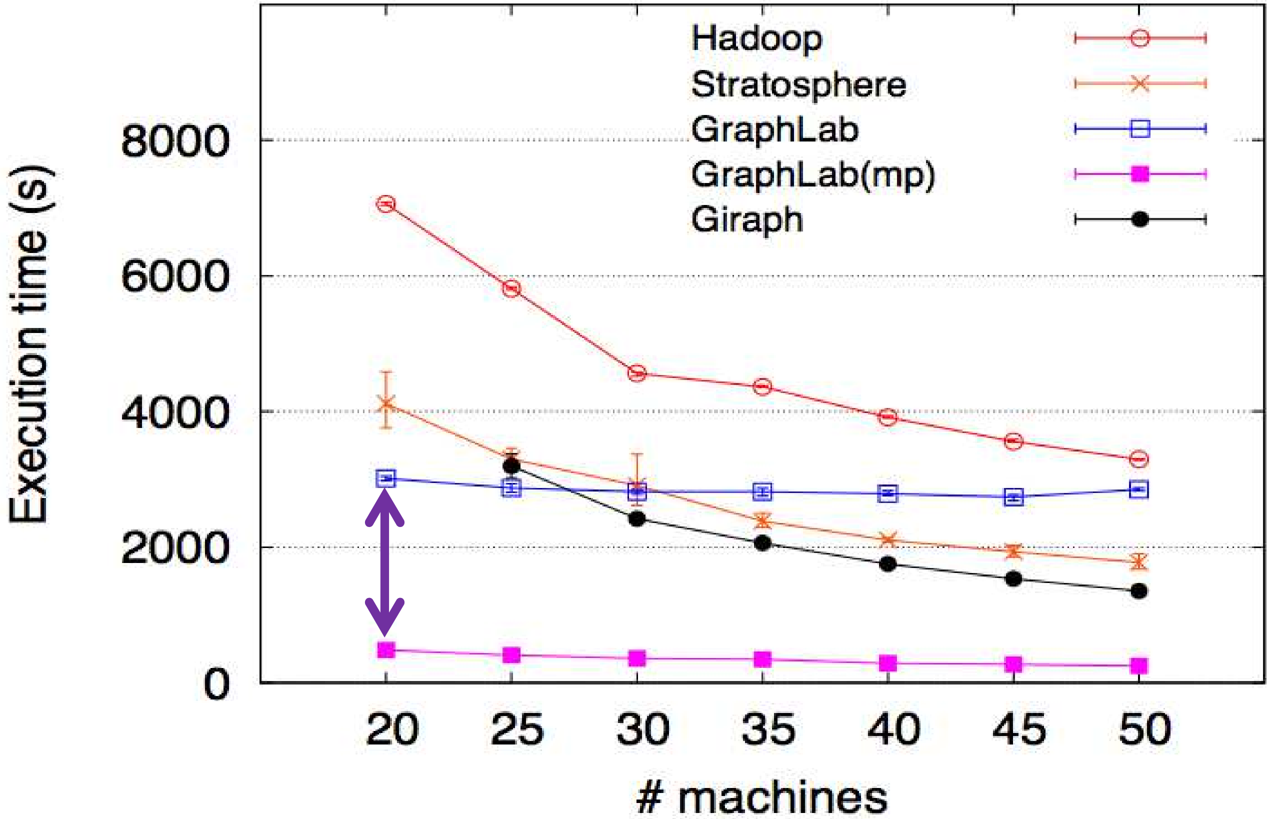


- Storing the whole graph in memory helps Giraph perform well
- Giraph may crash when graphs or number of messages large

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

Horizontal scalability: BFS on Friendster (31 GB)

http://bit.ly/1OhYdIU



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

Guo, Biczak, Varbanescu, Iosup, Martella, Willke. How Well do Graph-Processing Platforms Perform? An Empirical Performance Evaluation and Analysis http://bit.ly/10hYdIU

Additional Overheads
Data ingestion time

- Data ingestion
  - Batch system: one ingestion, multiple processing
  - Transactional system: one ingestion, one processing

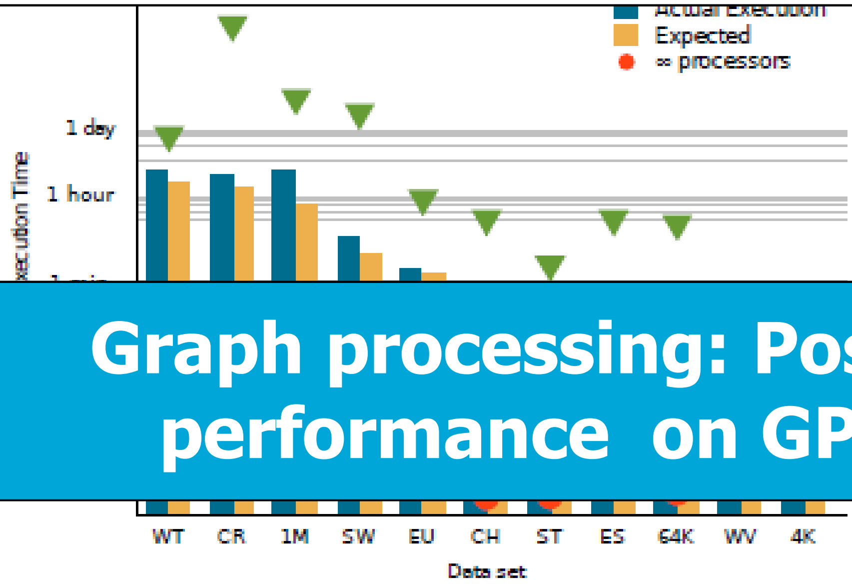
- Data ingestion matters even for batch systems

	Amazon	DotaLeague	Friendster
HDFS	1 second	7 seconds	5 minutes
Neo4J	4 hours	**days**	**n/a**

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

GPUs vs CPUs: All-Pairs Shortest Path

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

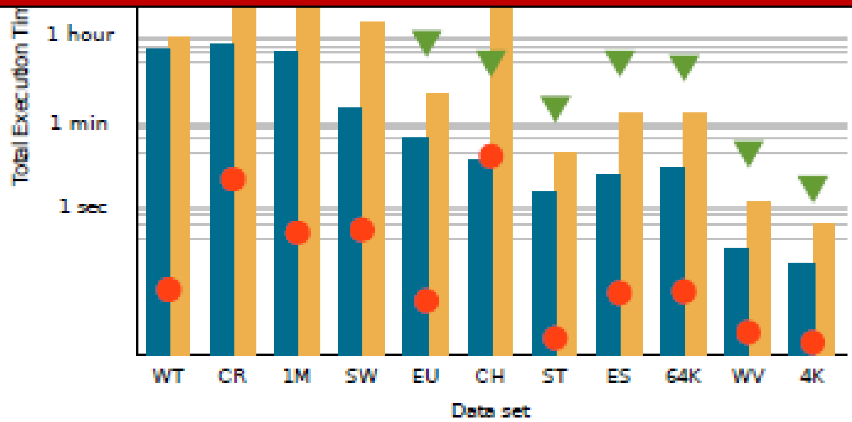


(a) Intel Xeon E5620

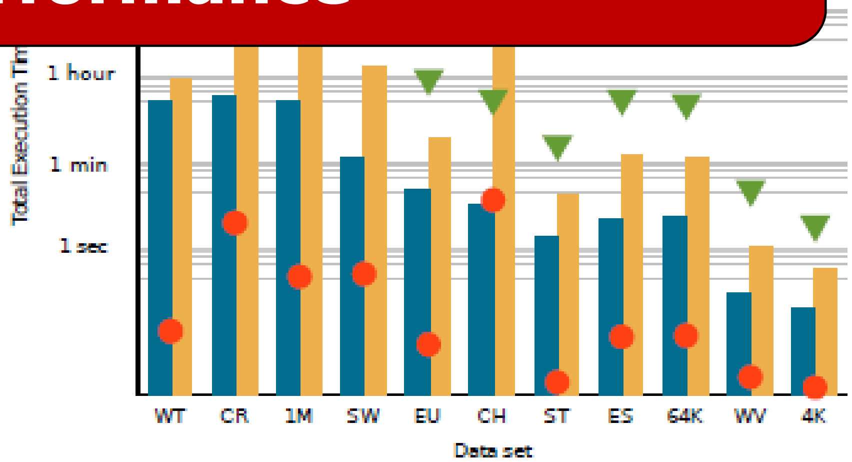
	Dataset
WT	Wikipedia Talk Network
CR	California Road Network
1M	Graph 1M
WV	Wikipedia Vote
4K	Graph 4K

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

However, Algorithm and Dataset also determine performance



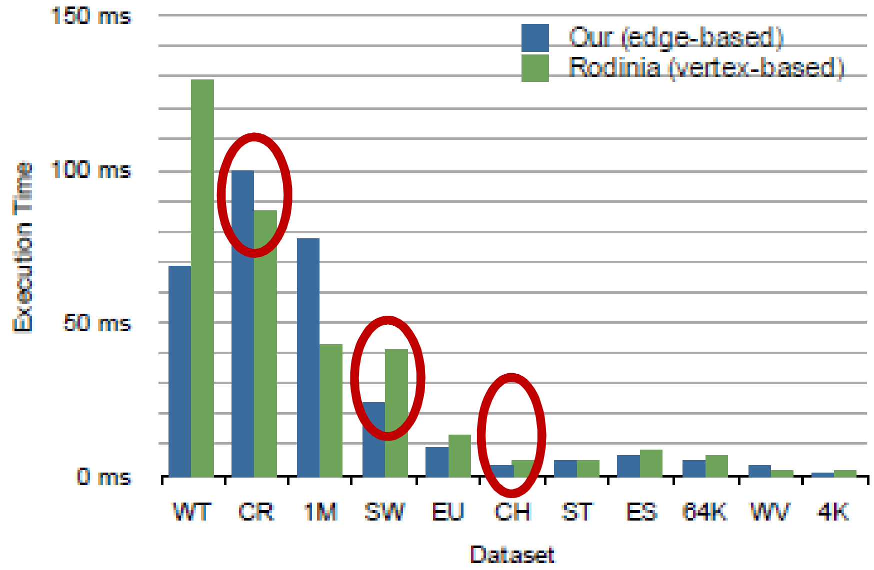
(c) Nvidia Tesla C2050/ C2070



(d) Nvidia GeForce GTX480

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

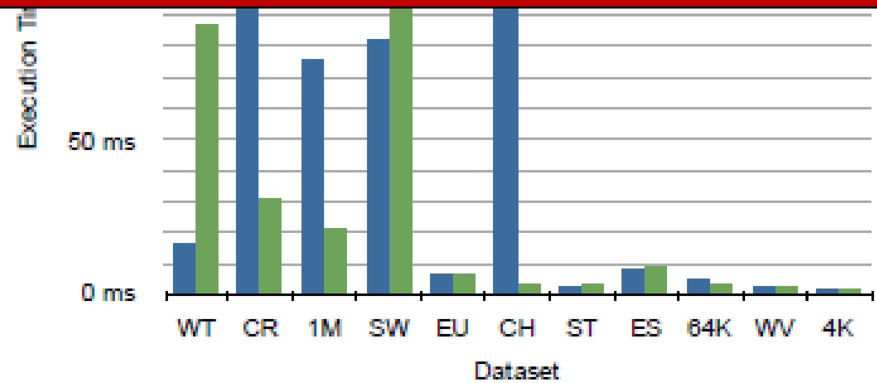
Pender and Varbanescu. MSc thesis at TU Delft. Jun 2012. TU Delft Library, http://library.tudelft.nl .



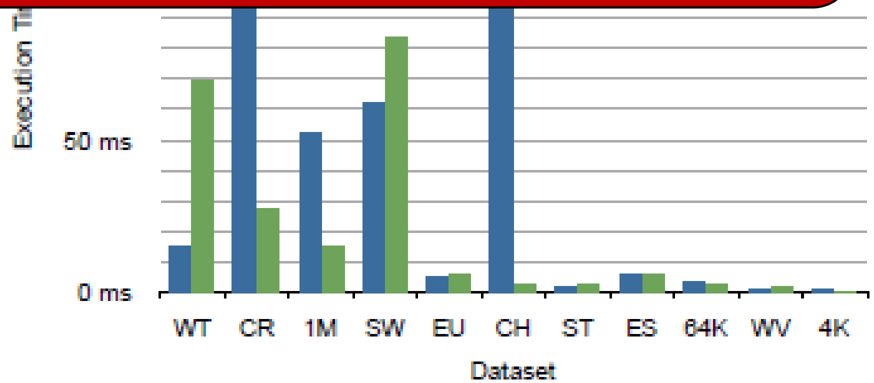
(a) Intel Xeon E5620

	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



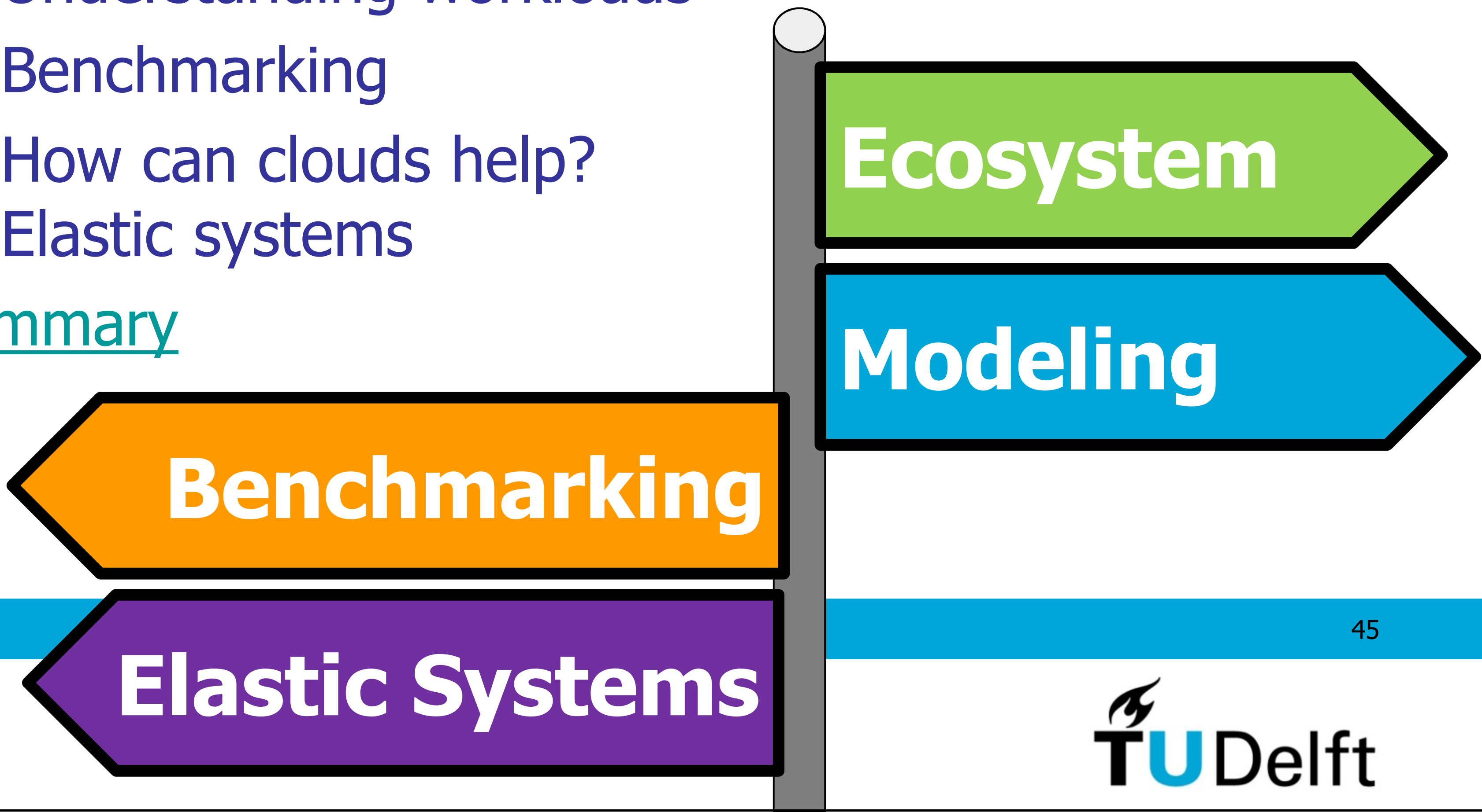
(c) Nvidia Tesla C2050/ C2070



(d) Nvidia GeForce GTX480

Agenda

1. Big Data, Our Vision, Our Team
2. **Big Data on Clouds**
    1. The Big Data ecosystem
    2. Understanding workloads
    3. Benchmarking
    4. How can clouds help? Elastic systems
3. Summary



TUDelft

Elasticity: Our Team Elastically Includes …

Alexandru Iosup
TU Delft

Provisioning
Allocation
Elasticity
Portfolio Scheduling
Isolation
Multi-Tenancy

Athanasios Antoniou
TU Delft

Provisioning
Allocation
Isolation
Utility

David Villegas
FIU/IBM
Elasticity, Utility

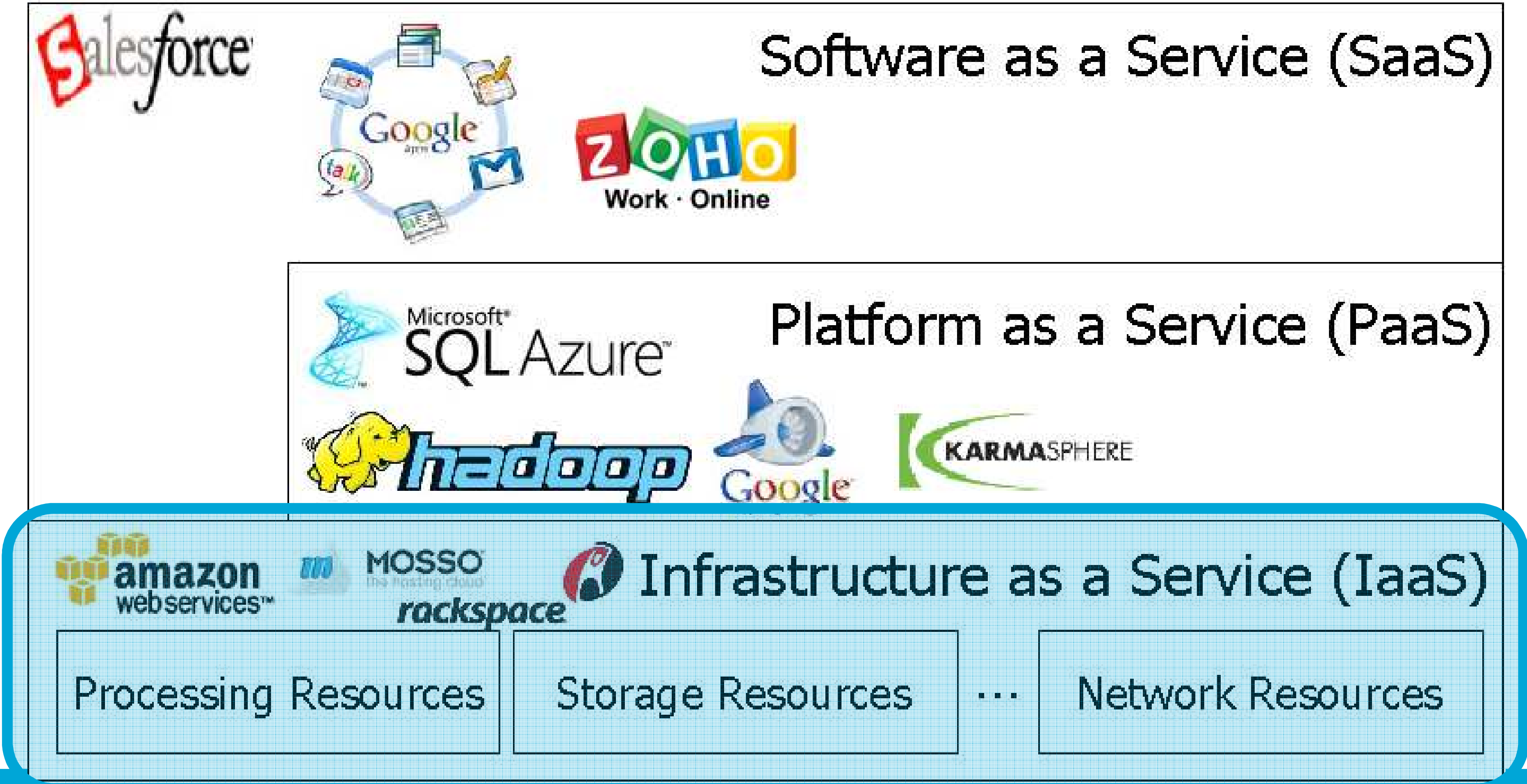
Kefeng Deng
NUDT
Portfolio Scheduling

Orna Agmon-Ben Yehuda
Technion
Elasticity, Utility

TUDelft

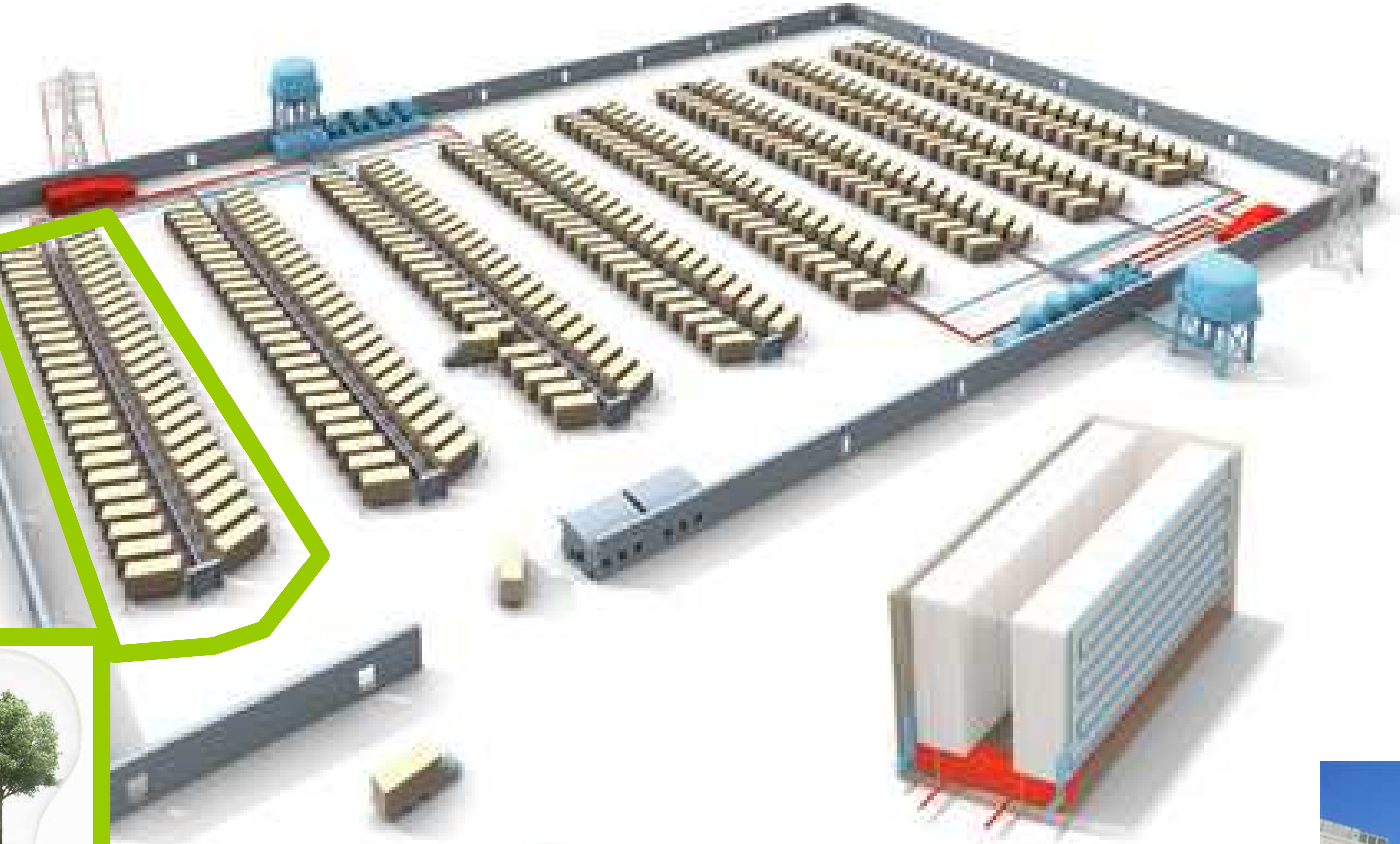
Cloud Computing, the useful IT service

"Use only when you want! Pay only for what you use!"



TUDelft

IaaS Cloud Computing:
Energy-Efficient IT Infrastructure Service



EIT ICT Labs

Uptime Institute
GREEN ENTERPRISE
IT AWARD 2010

amazon webservices™
Windows Azure
rackspace

TU Delft

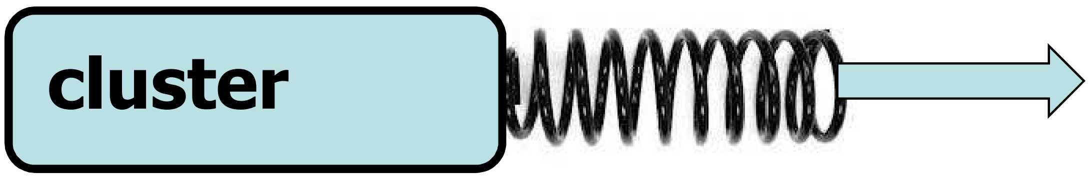
Delft University of Technology

Elasticity, Performance and Cost-Awareness

Why Dynamic Data Processing Clusters?

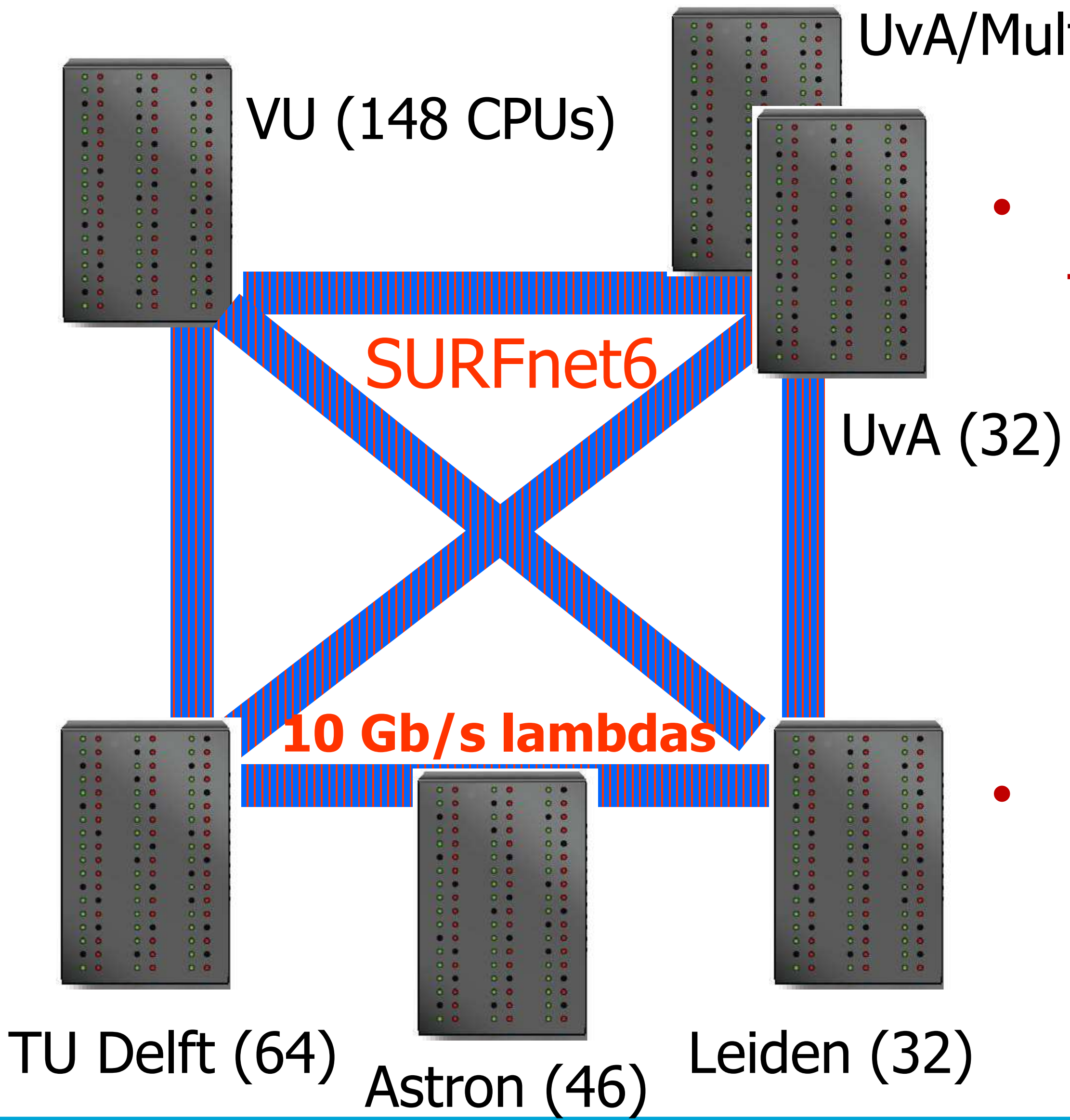
- Improve resource utilization
  - ➢ Grow when the workload is too heavy
  - ➢ Shrink when resources are idle
- Fairness across multiple data processing clusters
  - ➢ Redistribute idle resources
  - ➢ Allocate resources for new MR clusters

Isolation
- Performance
- Failure
- Data
- Version

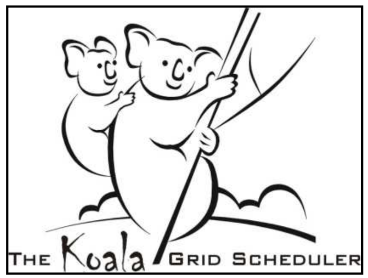


Ghit and Epema. Resource Management for Dynamic MapReduce Clusters in Multicluster Systems. MTAGS 2012. Best Paper Award.

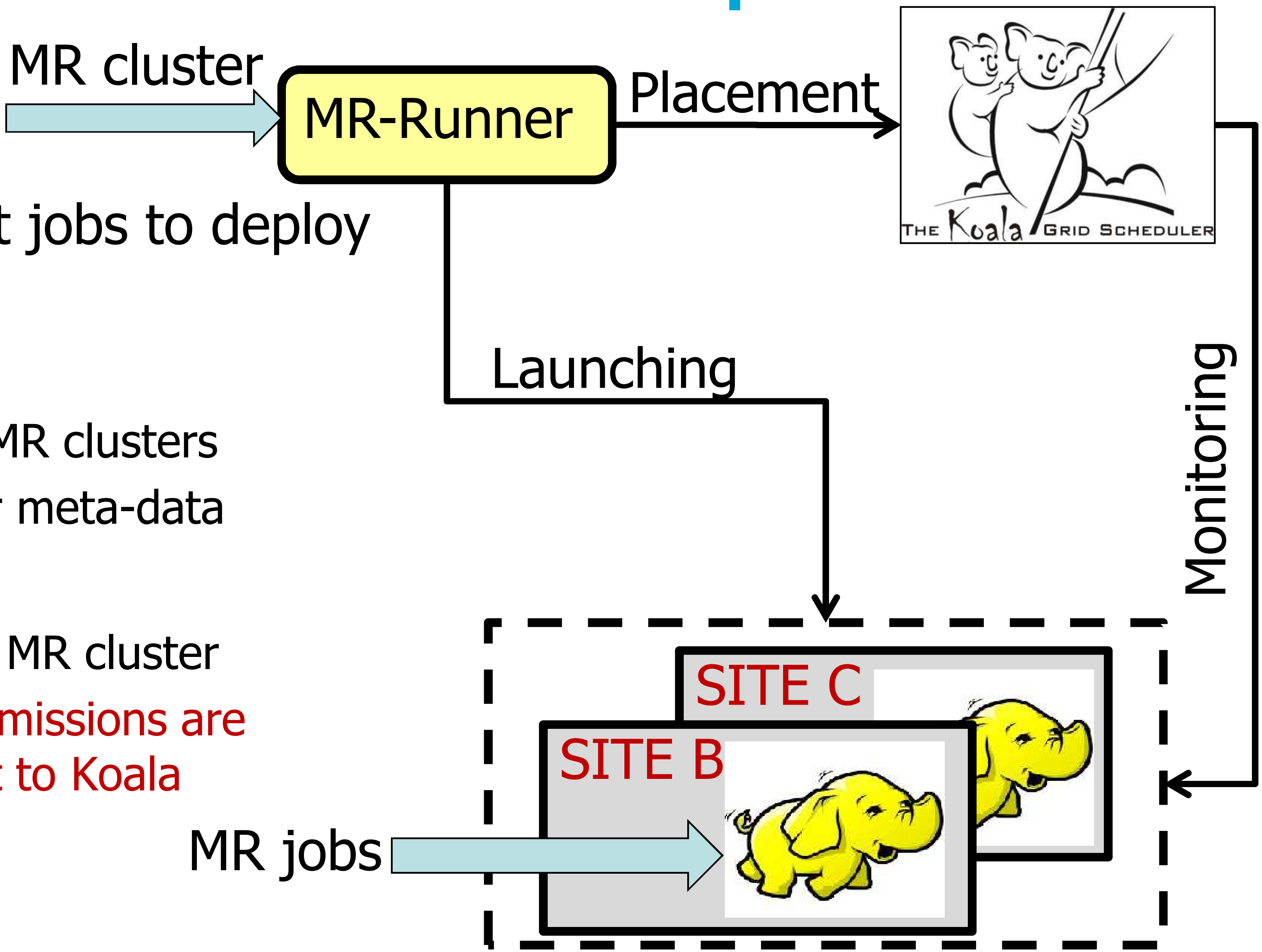
The DAS-4 Infrastructure



- 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



KOALA Grid Scheduler and MapReduce



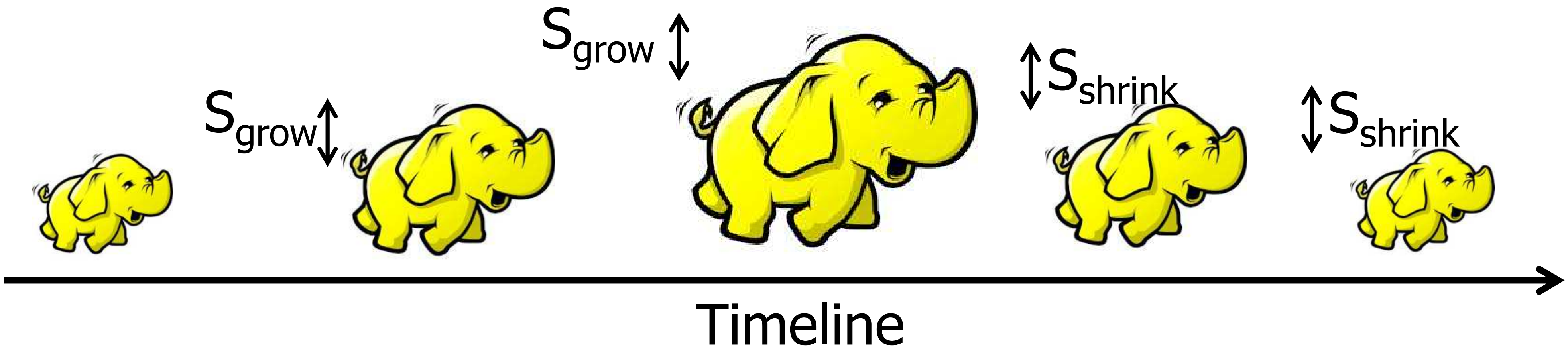
- Users submit jobs to deploy MR clusters
- Koala
  - Schedules MR clusters
  - Stores their meta-data
- MR-Runner
  - Installs the MR cluster
  - MR job submissions are transparent to Koala

Ghit and Epema. Resource Management for Dynamic MapReduce Clusters in Multicluster Systems. MTAGS 2012. Best Paper Award.

Resizing Mechanism

- Two-level provisioning
  - ➢ Koala makes resource offers / reclaims
  - ➢ MR-Runners accept / reject request

- Grow-Shrink Policy (GSP)
  - ➢ MR cluster utilization: $F_{\min} \leq \dfrac{totalTasks}{availSlots} \leq F_{\max}$
  - ➢ Size of grow and shrink steps: $\mathbf{S_{grow}}$ and $\mathbf{S_{shrink}}$

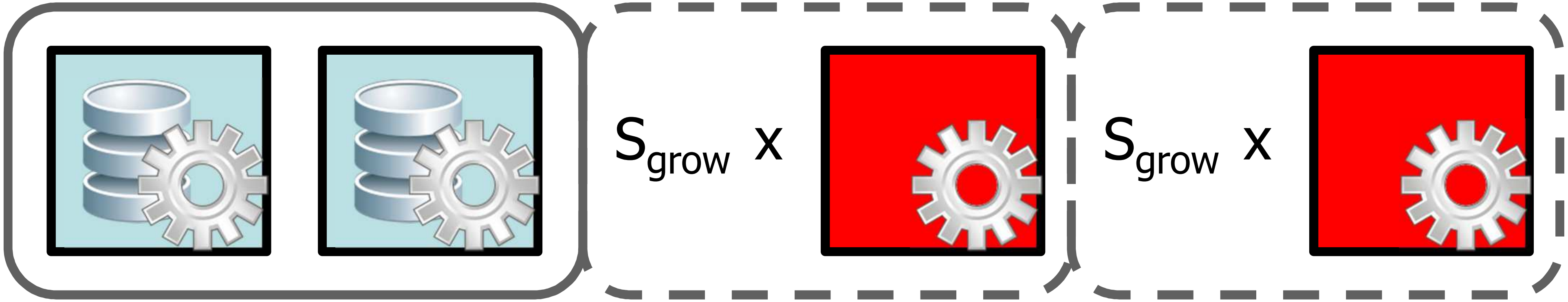


Ghit and Epema. Resource Management for Dynamic MapReduce Clusters in Multicluster Systems. MTAGS 2012. Best Paper Award.

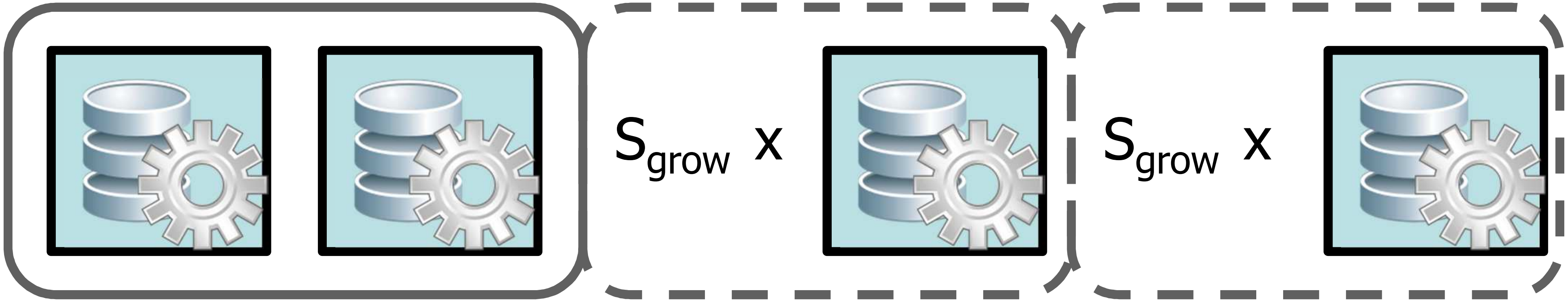
TU Delft

Baseline Policies

- Greedy-Grow Policy (GGP)—only grow with transient nodes:



- Greedy-Grow-with-Data Policy (GGDP)—grow, core nodes:



Ghit and Epema. Resource Management for Dynamic MapReduce Clusters in Multicluster Systems. MTAGS 2012. Best Paper Award.

Setup

- *98% of jobs @ Facebook take less than a minute*
- *Google reported computations with TB of data*

- DAS-4
- Two applications: Wordcount and Sort

Workload 1
- Single job
- 100 GB
- Makespan

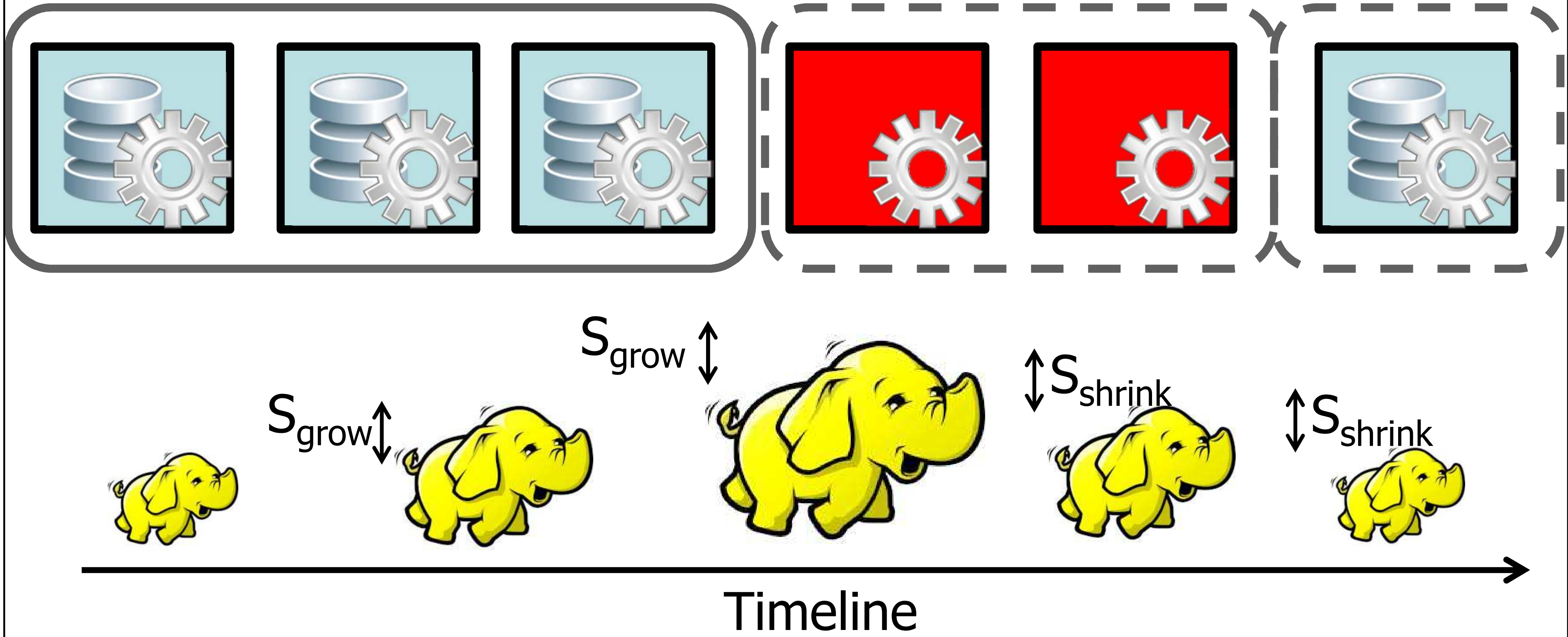
Workload 2
- Single job
- 40 GB, 50 GB
- Makespan

Workload 3
- Stream of 50 jobs
- 1 GB → 50 GB
- Average job execution time

Ghit and Epema. Resource Management for Dynamic MapReduce Clusters in Multicluster Systems. MTAGS 2012. Best Paper Award.

Elastic MapReduce, TUD version

- Two types of nodes
  - Core nodes: compute and data storage (DataNode)
  - Transient nodes: only compute / + data storage



Ghit and Epema. Resource Management for Dynamic MapReduce Clusters in Multicluster Systems. MTAGS 2012. Best Paper Award.

TU Delft

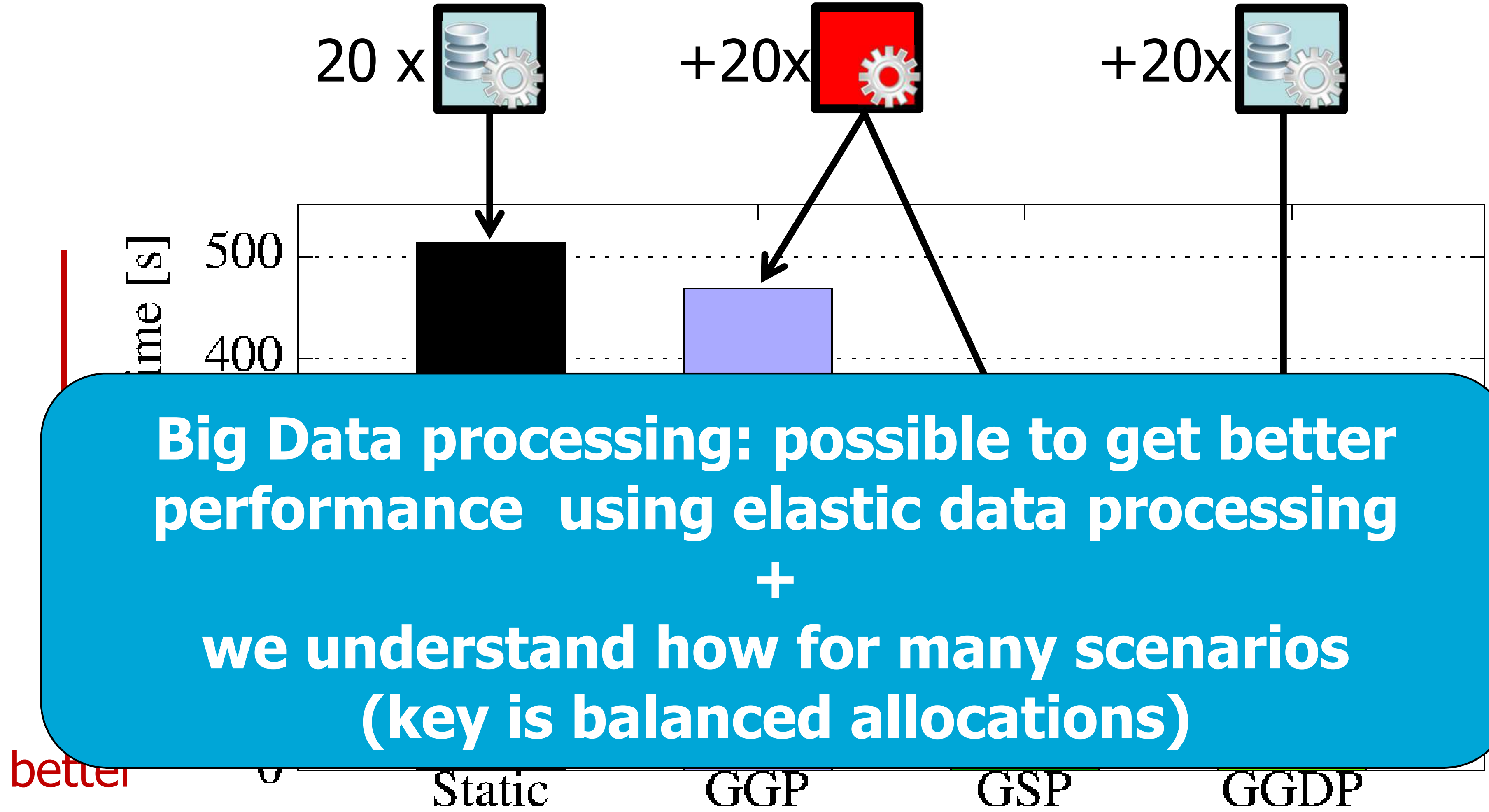
Transient Nodes

Workload 2:
40GB, 50GB

- Wordcount scales better than Sort on transient nodes

Ghit and Epema. Resource Management for Dynamic MapReduce Clusters in Multicluster Systems. MTAGS 2012. Best Paper Award.

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)

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.

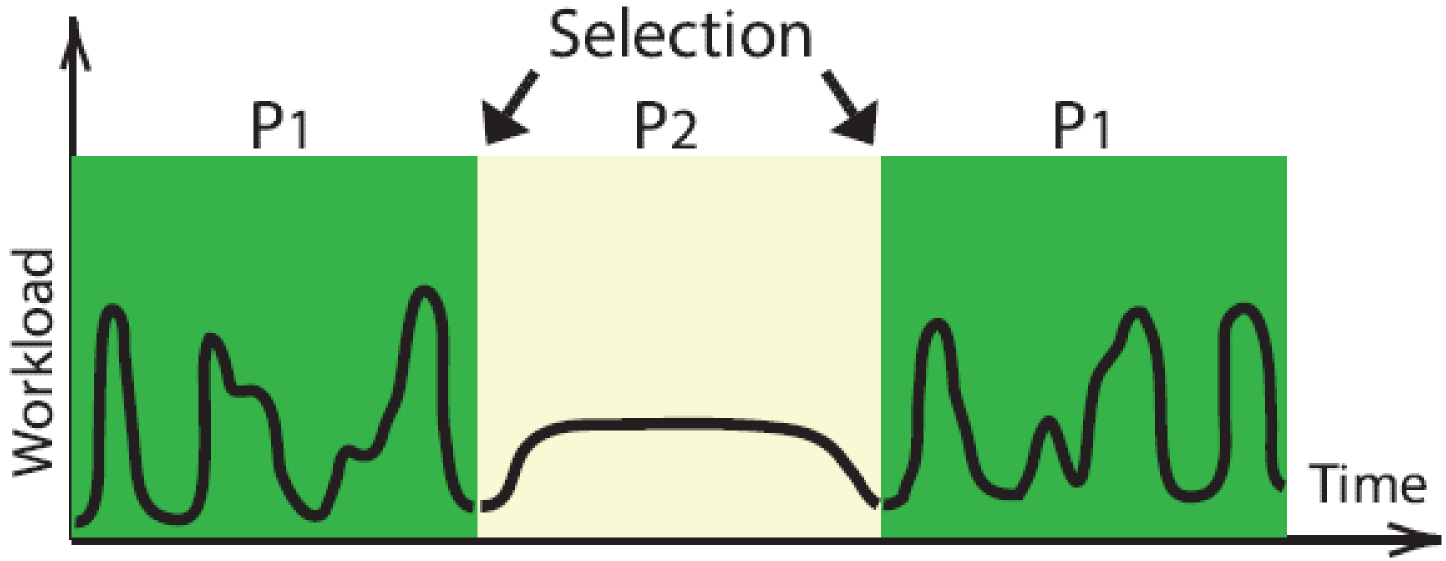
TU Delft

Why Portfolio Scheduling?

- **Old scheduling aspects**
  - Hundreds of approaches, each targeting specific conditions—which to choose? How to configure?
  - No one-size-fits-all policy
- **New scheduling aspects**
  - New workloads, e.g., pretty much all Big Data
  - New data center architectures
  - New cost models, e.g., moving workloads to IaaS clouds

- **Developing a scheduling policy is risky and ephemeral**
- **Selecting a scheduling policy is risky and difficult**

What is Portfolio Scheduling? In a Nutshell, for Elastic Big Data Processing



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

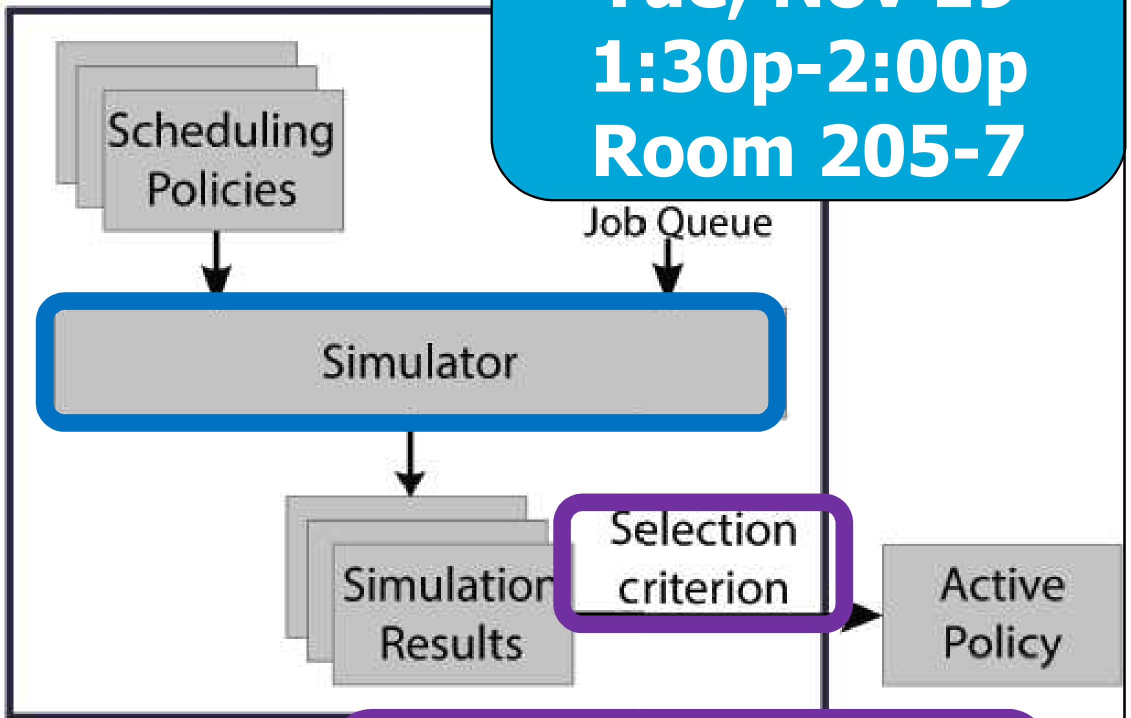
Portfolio Scheduling (Technical)

SC|13
Tue, Nov 19
1:30p-2:00p
Room 205-7

- Periodic execution

- Simulation-based selection
- Utility function

- Alternatives simulator
  - Expert human knowledge
  - WL sample in real env.
  - Mathematical analysis
- Alternatives utility function
  - Well-known and exotic functions



$\alpha=\beta=1$
$K=100$

$$U = \kappa \cdot \left(\frac{R_J}{R_V}\right)^{\alpha} \cdot \left(\frac{1}{S}\right)^{\beta}$$

R_J: Total Runtime of Jobs
R_V: Total Runtime of VMs
S: Slowdown

Deng, Verboon, Iosup. A Periodic Portfolio Scheduler for Scientific Computing in the Data Center. JSSPP'13.

Deng, Song, Ren, Iosup. Exploring portfolio scheduling for long-term execution of scientific workloads in IaaS clouds. SC|13.

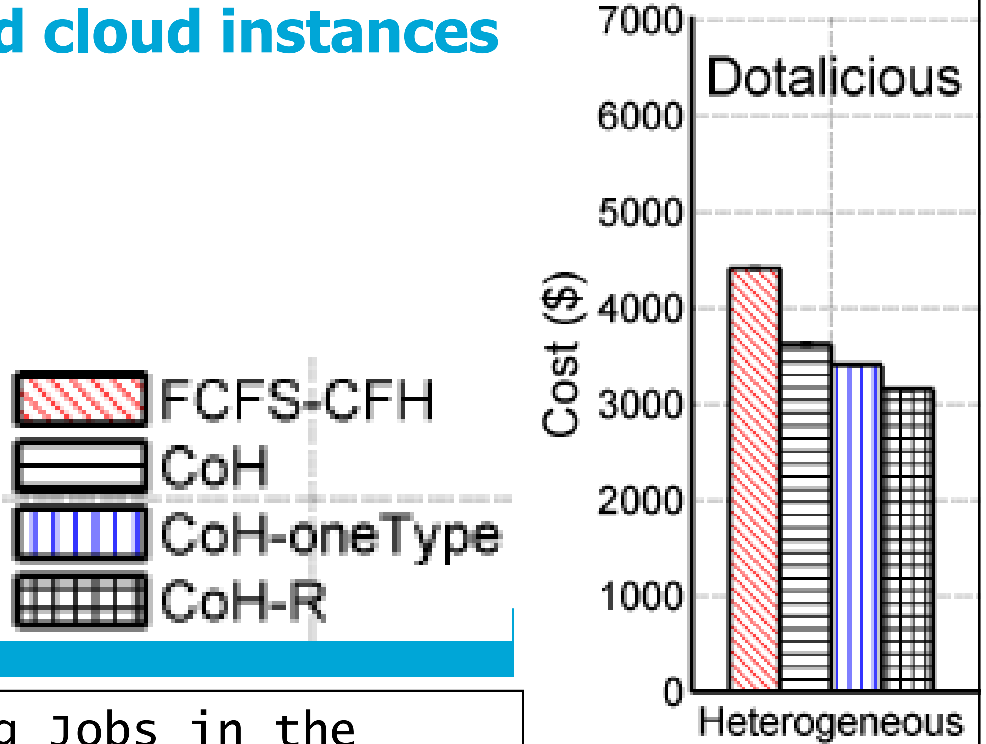
Agmon Ben-Yehuda, Schuster, Sharov, Silberstein, Iosup. ExPERT: pareto-efficient task replication on grids and a cloud. IPDPS'12.

Portfolio Scheduling for Online Gaming
(also for Scientific Workloads)

- **CoH =** Cloud-based, online, Hybrid scheduling
  - Intuition: keep rental cost low by finding good mix of machine configurations and billing options, use **on-demand cloud VMs**
  - Main idea: run *both* solver of an Integer Programming Problem and various heuristics, **pick best schedule periodically (at deadline)**
  - Additional feature: Can use **reserved cloud instances**

Gaming (and scientific) workloads

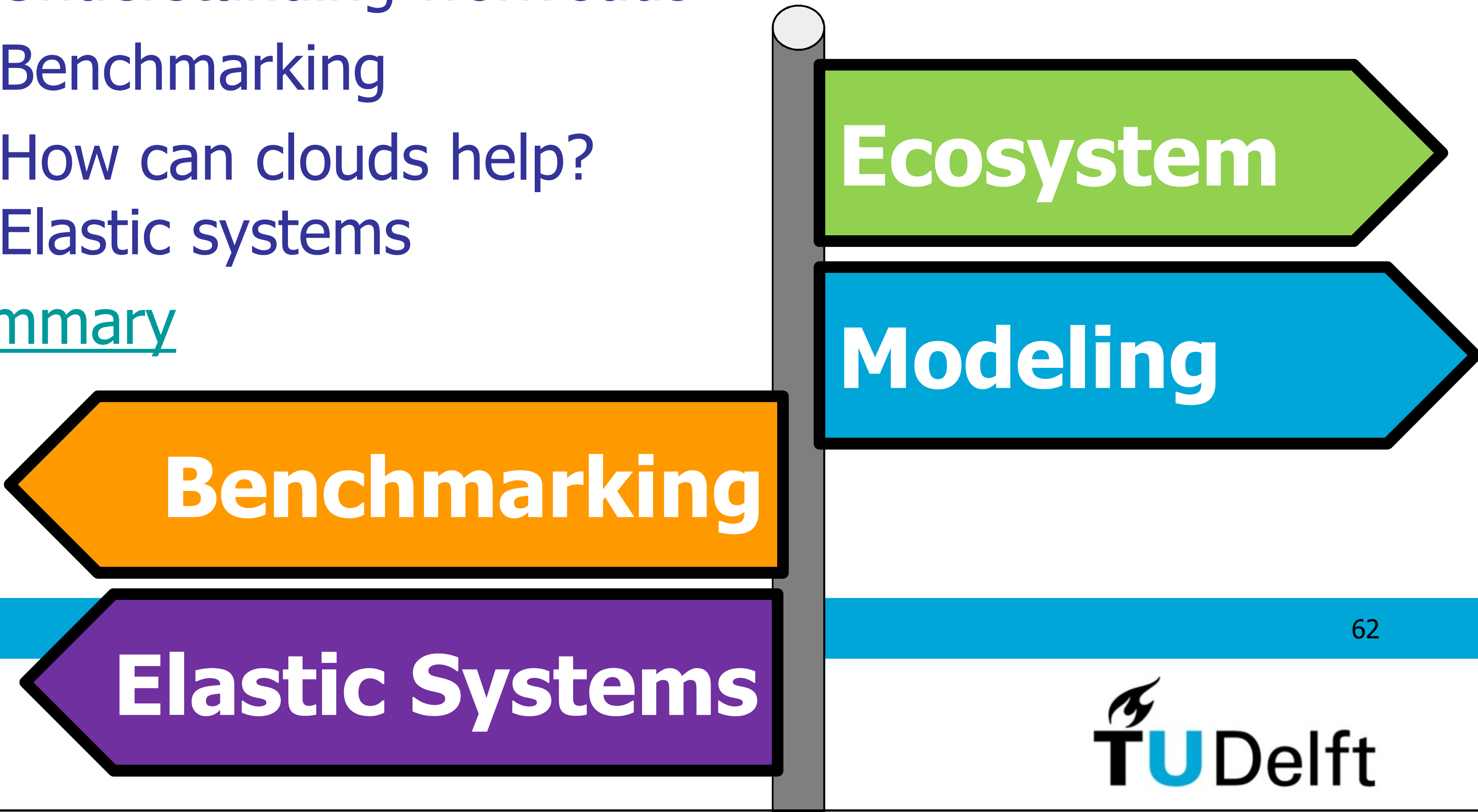
Trace	#jobs	average runtime [s]
Grid5000	200,450	2728
LCG	188,041	8971
DotaLicious	109,251	2231



Shen, Deng, Iosup, and Epema. Scheduling Jobs in the Cloud Using On-demand and Reserved Instances, EuroPar'13.

Agenda

1. Big Data, Our Vision, Our Team
2. **Big Data on Clouds**
    1. The Big Data ecosystem
    2. Understanding workloads
    3. Benchmarking
    4. How can clouds help? Elastic systems
3. Summary



~~Conclusion~~ Take-Home Message

- **Big Data is necessary but grand challenge**
- **Big Data = Systems of Systems**
  - **Big data programming models have ecosystems**
  - **Stuck in stacks!**
  - Many trade-offs, many programming models, many problems

- **Towards a Generic Big-Data Processing System**
  - Looking at the Execution Engine—thrilling moment for this!
  - Predictability challenges: Understanding workload (modeling) and performance (benchmarking)
  - Performance challenges: distrib/parallel from the beginning
  - Elasticity challenges: elastic data processing, portfolio scheduling, etc.
  - etc.

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

More Info:

- http://www.st.ewi.tudelft.nl/~iosup/research.html
- http://www.st.ewi.tudelft.nl/~iosup/research_cloud.html
- http://www.pds.ewi.tudelft.nl/

Do not hesitate to contact me…

Alexandru Iosup

A.Iosup@tudelft.nl
http://www.pds.ewi.tudelft.nl/~iosup/ (or google "iosup")
Parallel and Distributed Systems Group
Delft University of Technology