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

@AIosup

dr. ir. Alexandru Iosup
Parallel and Distributed Systems Group

(TU) Delft – the Netherlands – Europe

founded 13th century
pop: 100,000

founded 1842
pop: 15,000

Delft

pop: 16.5 M

Barcelona

TU Delft

COMMIT/

DELFT DATA SCIENCE

THE PARALLEL AND DISTRIBUTED SYSTEMS GROUP AT TU DELFT

Winners IEEE TCSC Scale Challenge 2014

VENI

Alexandru Iosup

Grids/Clouds
P2P systems
Big Data/graphs
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/graphs

Henk Sips

HPC systems
Multi-cores
P2P systems

VENI

Johan Pouwelse

P2P systems
File-sharing
Video-on-demand

Home page

- www.pds.ewi.tudelft.nl

Publications

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

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

TU Delft · COMMIT/ · Delft Data Science

Our Industry Collaborators

Thank you for your invitation!

Scalable High Performance Systems

Interaction Encouraged!

2' — Where and What Is TU Delft/the PDS group?

5' — The Golden Age of Datacenters

5' — A Delft View on Datacenter Technology

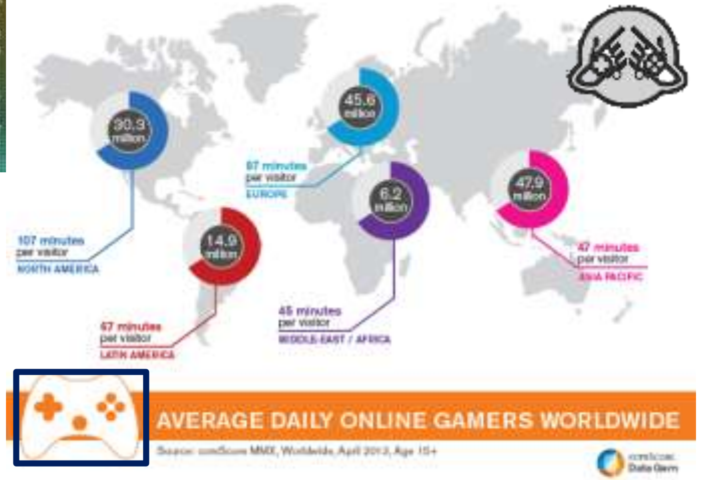
- The main challenges

35' — The Delft Approach to Making Datacenters Tick

- Addressing the New World Challenge
- Addressing the Scheduling challenge
- Addressing the Ecosystem Navigation challenge
- Addressing the Big Cake challenge
- Addressing Jevons Effect in Datacenters

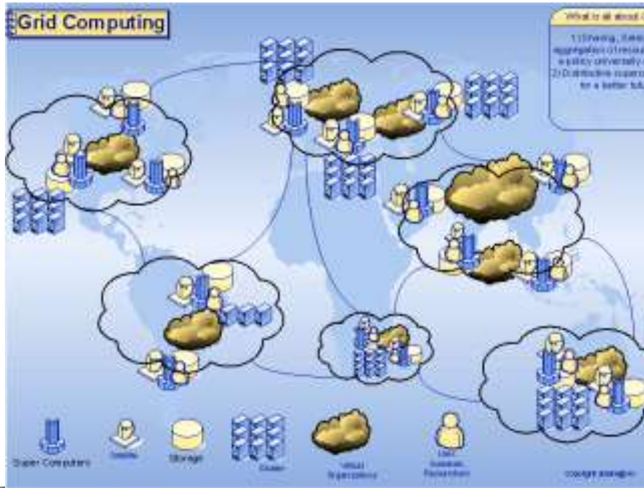
10' — Towards a Collaboration on Datacenter Technology

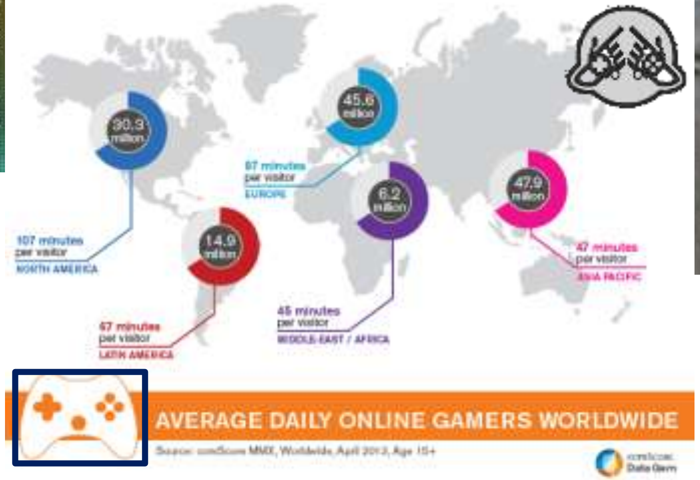
This Is the Golden Age of Datacenters



DELFT DATA SCIENCE

This Is the Golden Age of Datacenters

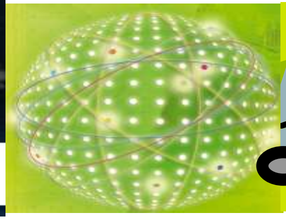
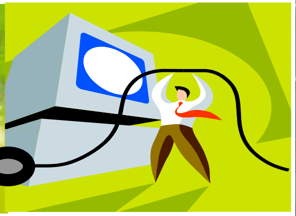
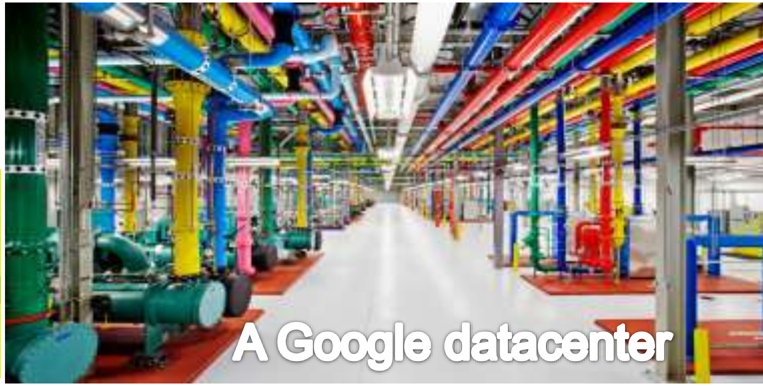




DATA SCIENCE

This Is the Golden Age of Datacenters

This Is the Golden Age of Datacenters

Datacenters = commodity high performance systems

- Large-scale infrastructure
- High-tech automated software to manage
- Inter-connected computer clusters
- High-end computation, storage, network
- Large memory capacity

"my other computer is a datacenter"

KIVI Engineering Society

TUDelft

DELFT DATA SCIENCE

Societal Challenges

The quadruple helix: prosperous society & blooming economy & inventive academia & wise governance depend on datacenters

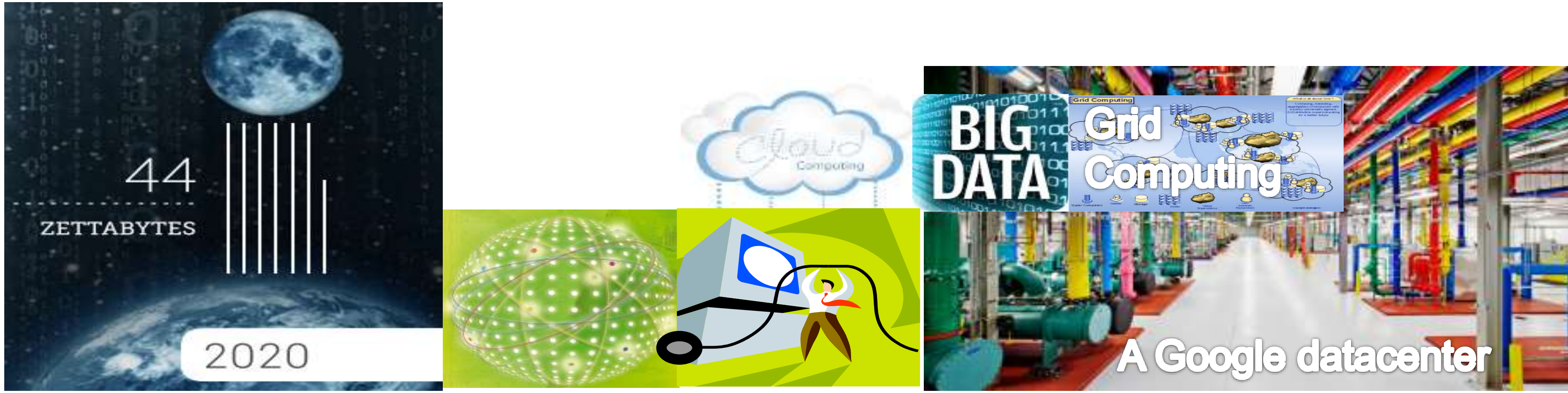
- Enable data access & processing as a fundamental right in Europe
- Enable big science and engineering (2020: €100 bn., 1 mil. jobs in Europe)
- “To out-compute is to out-compete”, but with energy footprint <5%
- Keep Internet-services affordable yet high quality in Europe
- The Schiphol of computation: building world-wide ICT hubs

KIVI Engineering Society

DELFT DATA SCIENCE

TU Delft

Scientific and Technical Challenges



How to massivize datacenters?

- Super-scalable, super-flexible, yet efficient ICT infrastructure
- End-to-end automation of large-scale processes
- Dynamic, compute- and data-intensive workloads
- Evolving, heterogeneous hardware and software
- Strict performance, cost, energy, reliability, and fairness requirements

TUDelft

KIVI Engineering Society

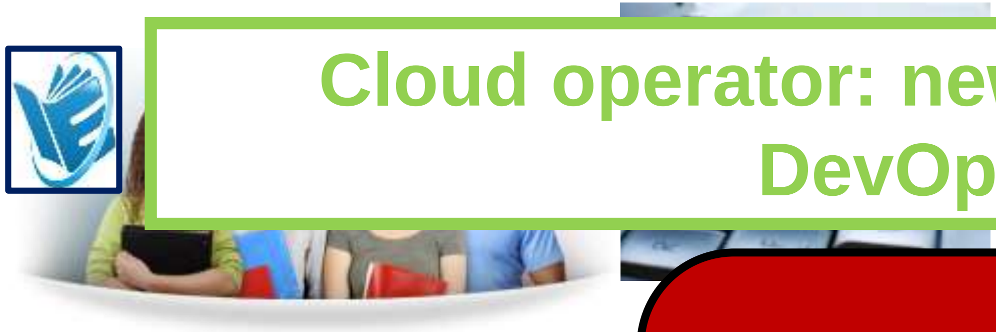
DELFT DATA SCIENCE

Scalable High Performance Systems

Interaction Encouraged!

- 2' — Where and What Is TU Delft/the PDS group?
- 5' — The Golden Age of Datacenters
- 5' — A Delft View on Datacenter Technology
  - The main challenges
- 35' — Delft Data Science Makes Datacenters Tick
  - Addressing the New World Challenge
  - Addressing the Scheduling challenge
  - Addressing the Ecosystem Navigation challenge
  - Addressing the Big Cake challenge
  - Addressing Jevons Effect in Datacenters
- 10' — Towards a Collaboration on Datacenter Technology

The New World Challenge



Cloud operator: new value-adding services, DevOps workloads

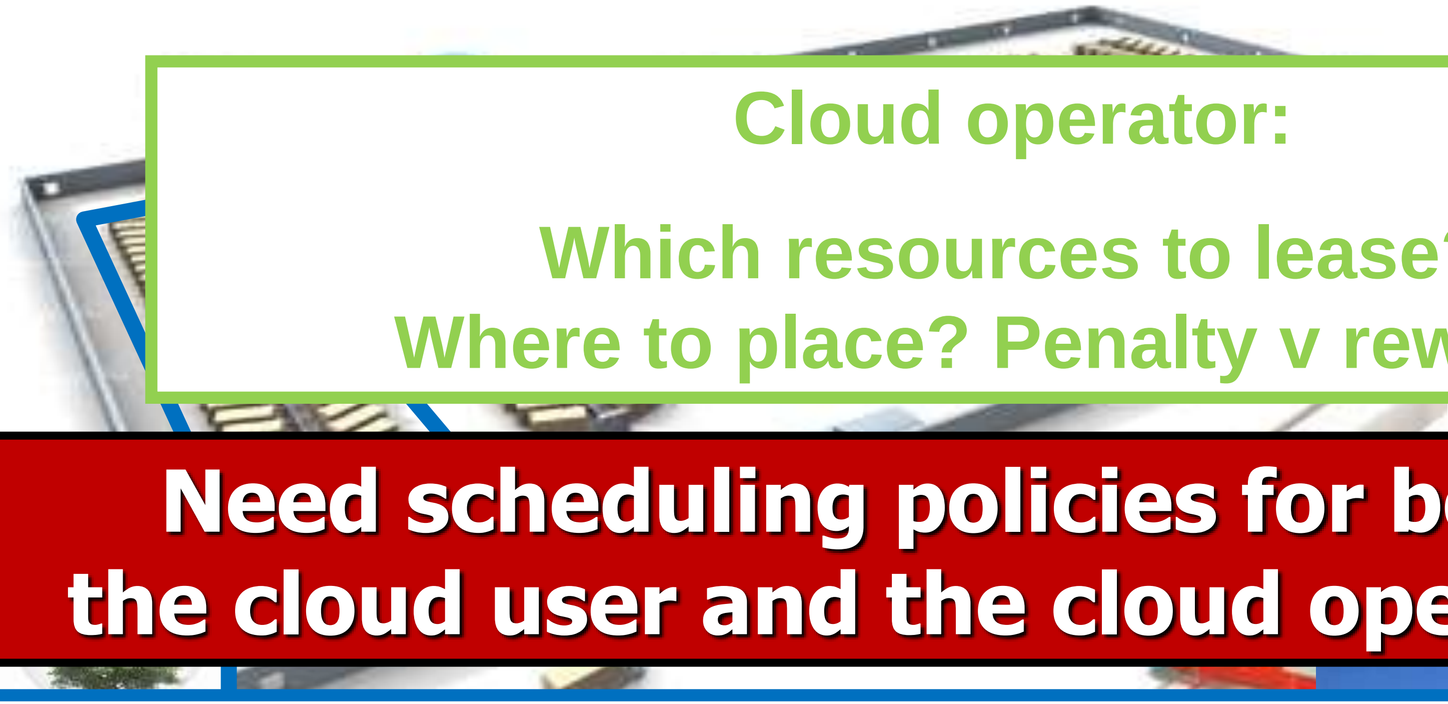
Need to understand and model workloads in datacenters, both new customer apps and datacenter Dev Ops workloads

Cloud customer: new apps, new services, customers can become operators (value-chain)

TU Delft

DELFT DATA SCIENCE

The Scheduling Challenge



Cloud operator:

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

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

Cloud customer:

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

The Ecosystem Navigation Challenge

Cloud operator: how to prove capabilities? How to tune the tool? In which technology to invest? Which tech to DevOp in-house?

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

Cloud customer: how to choose the right tool? (Stonebraker: no one size fits all!)

Programming Model

Tree

Engine

S3

GFS

Data Store

Data Store

S

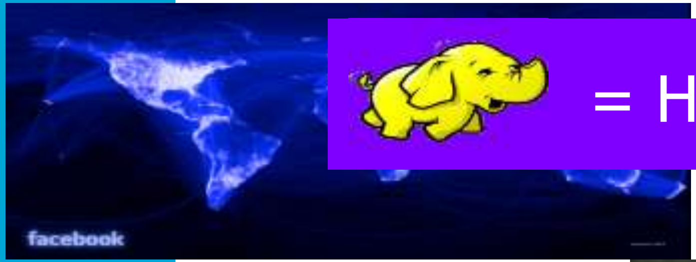
Asterix B-tree

Storage Engine

* Plus Zookeeper, CDN, etc.

The “Big Cake” Challenge In the Datacenter

Online Social Networks



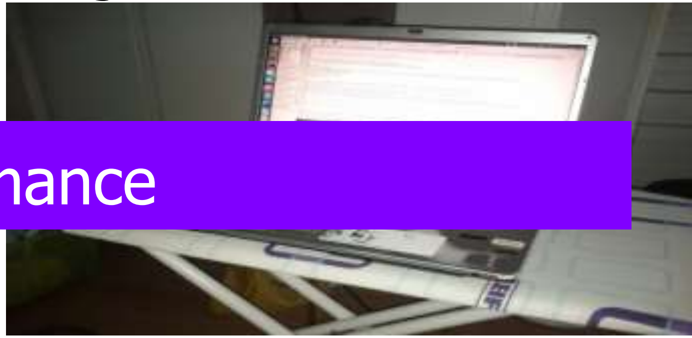
Financial Analysts

= Hadoop / MapReduce framework

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

Universe Explorers

Big Data Enthusiast



Multiple frameworks = Isolation, especially performance

TUDelft COMMIT/

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

DELFT DATA SCIENCE

Jevons Effect: More Efficient, Less Capable

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

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

PSY Gangnam consumed ~500GWh

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

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

The New “Jevons Effect”: The “Data Deluge” vs Capability

Data Deluge =

data generated by humans and devices (IoT)

- Interacting
- Understanding
- Deciding
- Creating

To be capable of processing Big Data, need to address Volume, Velocity, Variety of Big Data*

* Other Vs possible: ours is “vicissitude”

Sources: IDC, EMC.

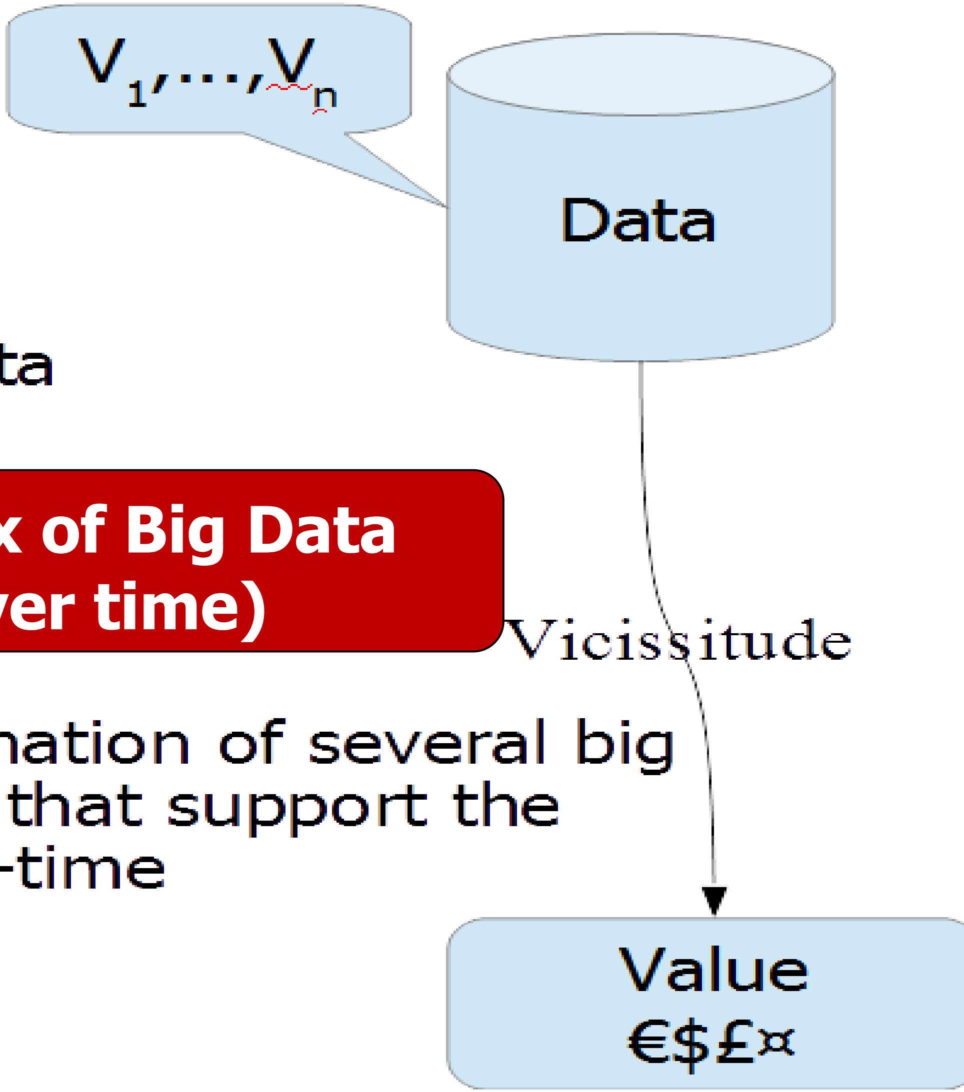
TUDelft COMMIT/ DELFT DATA SCIENCE

Vs of big data

- Volume – large scale of data
- Variety – different forms of data

Need to address viccisitude (mix of Big Data Volume, Velocity, Variety over time)

- **Vicissitude** – dynamic combination of several big data Vs in processing systems that support the addition of new queries at run-time



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

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

A Delft View on Datacenter Technology

- The New World Challenge: knowing operator + customer workload
- The Scheduling challenge: using resources efficiently
- The Ecosystem Navigation challenge: benchmarking efficiently
- The Big Cake challenge: sharing resources efficiently
- Jevons Effect in Datacenters: addressing vicissitude efficiently

Addressing these challenges =

Massivizing Datacenter through Technology

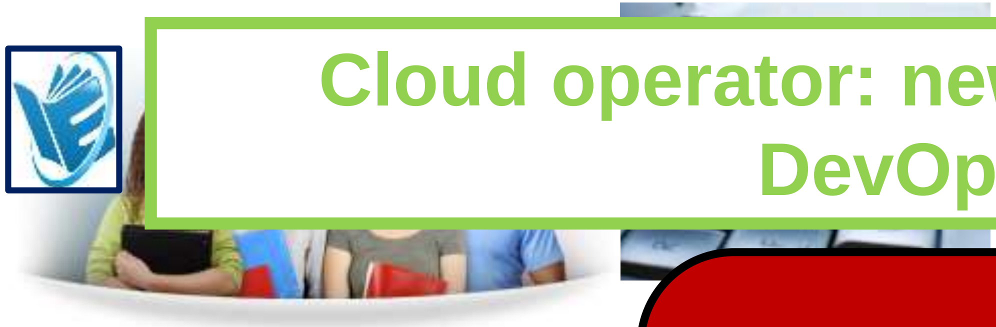
TUDelft COMMIT/ DELFT DATA SCIENCE

Scalable High Performance Systems

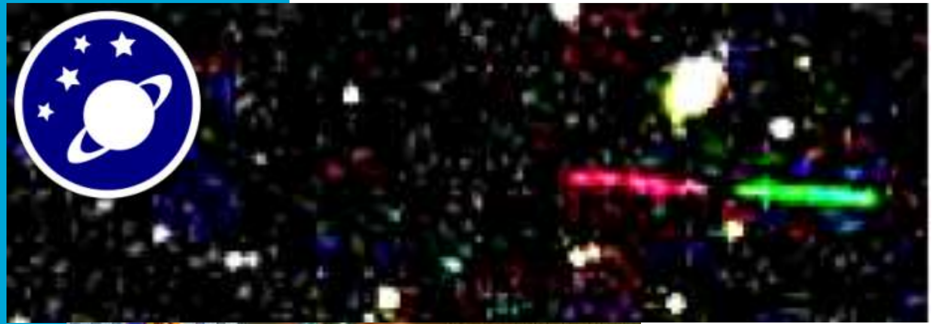
Interaction Encouraged!

- 2' — Where and What Is TU Delft/the PDS group?
- 5' — The Golden Age of Datacenters
- 5' — A Delft View on Datacenter Technology
  - The main challenges
- 35' — The Delft Approach to Making Datacenters Tick
  - Addressing the New World Challenge
  - Addressing the Scheduling challenge
  - Addressing the Ecosystem Navigation challenge
  - Addressing the Big Cake challenge
  - Addressing Jevons Effect in Datacenters
- 10' — Towards a Collaboration on Datacenter Technology

The New World Challenge



Cloud operator: new value-adding services, DevOps workloads



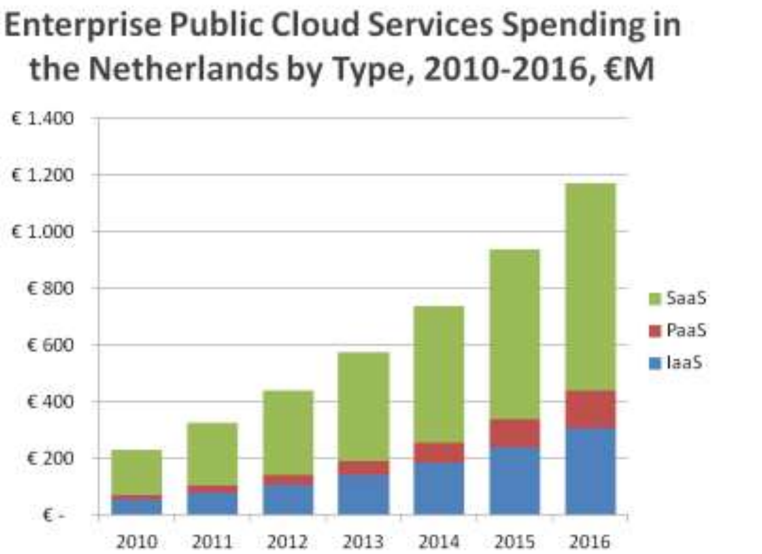
Need to understand and model workloads in datacenters, both new customer apps and datacenter Dev Ops workloads

Cloud customer: new apps, new services, customers can become operators (value-chain)

TU Delft

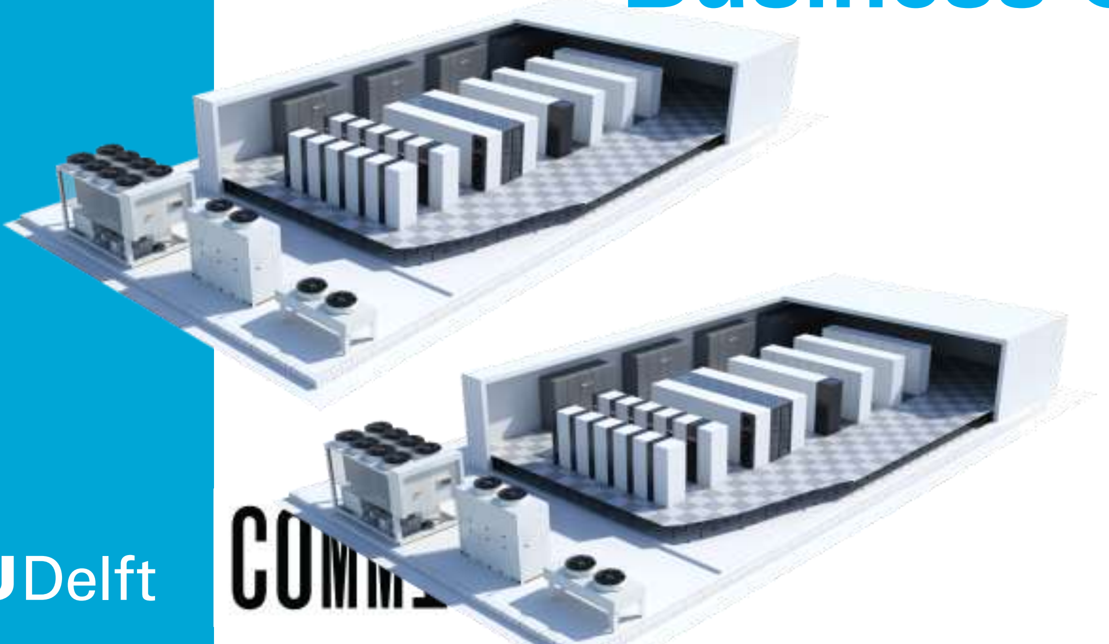
DELFT DATA SCIENCE

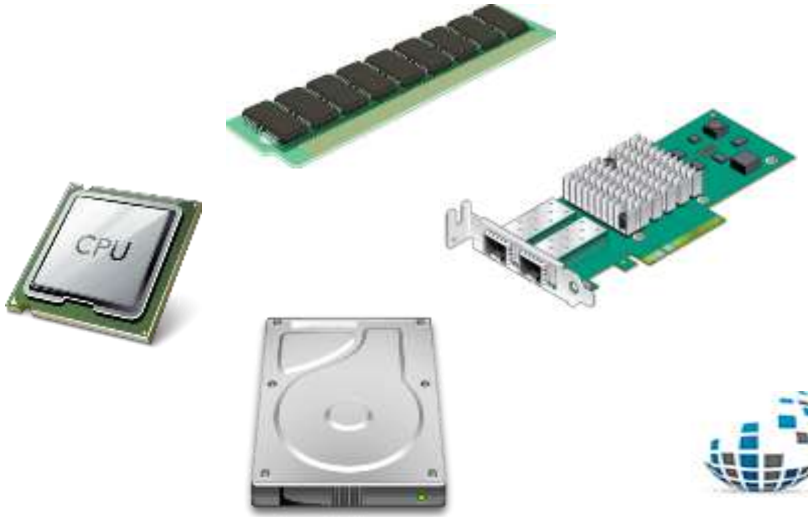
Monte Carlo simulation



Source: http://www.themetisfiles.com

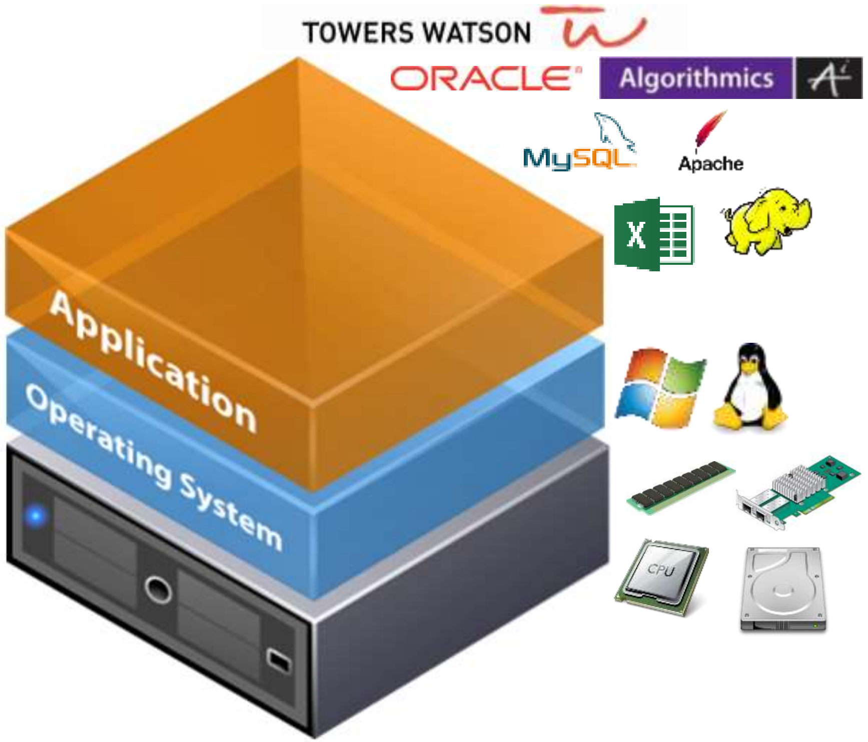
Business Critical Workloads



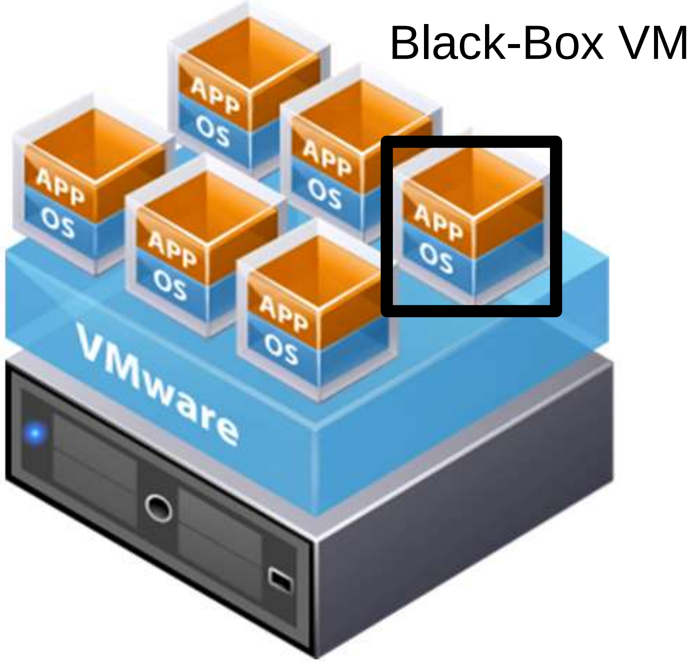


TU Delft

DELFT DATA SCIENCE



Traditional Architecture



Virtual Architecture

TU Delft

COMMIT/

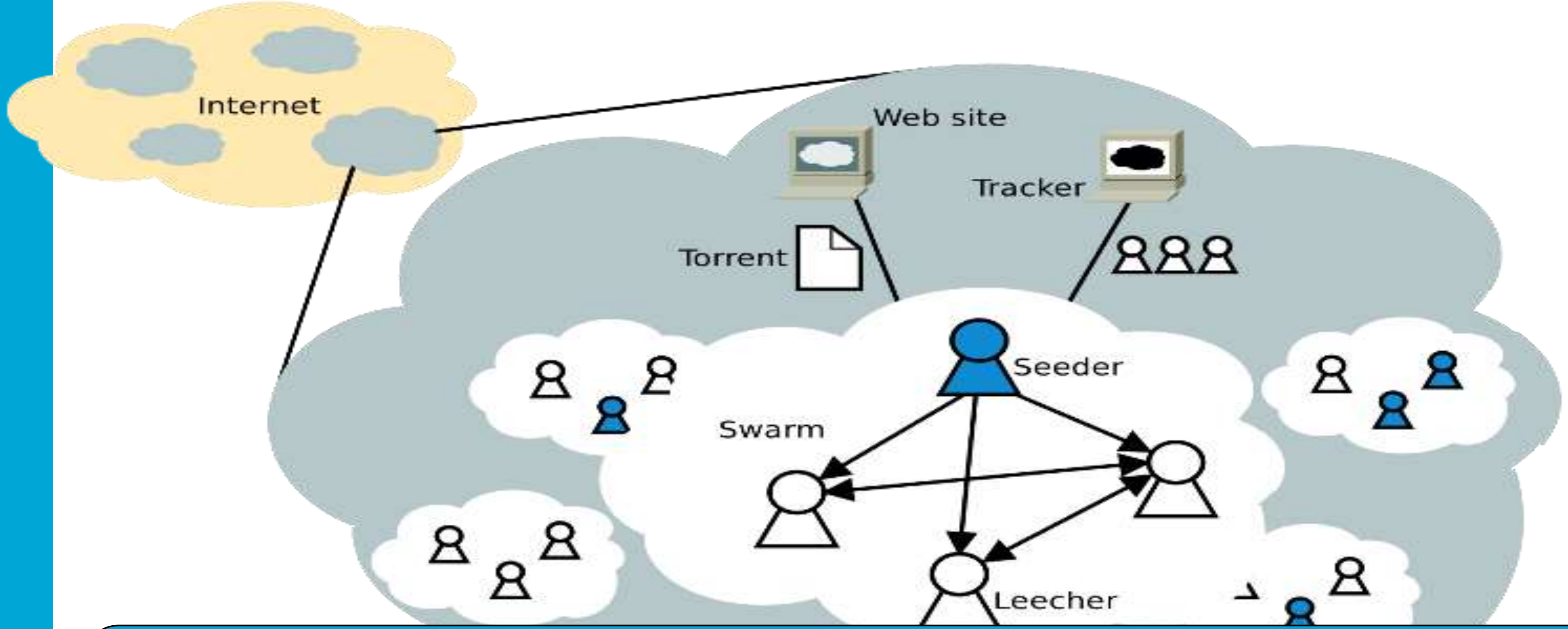
DELFT DATA SCIENCE

Our findings:
Business-Critical vs Known workloads

- Long running VMs vs short running jobs
- Compared to parallel workloads, small in size (cpu and memory)
  - Many opportunities for scheduling efficiency (e.g., used<<requested, pow-2, periodicity)
- Much more diverse in nature, compared to
  data analysis workloads from Facebook, Google, and Tabao
  - Monte Carlo Simulation (e.g., finance)
  - Data analysis of business data (e.g., finance)
  - Office automation (e.g., web, mail)
  - High available web-services for complex applications (e.g., retail, CC systems)
  - DC value-adding services, e.g., backup

S. Shen et al. Statistical Characterization of Business-Critical Workloads Hosted in Cloud Datacenters. CCGRID 2015: 465-474

Monitoring A Typical Global System: BitTorrent

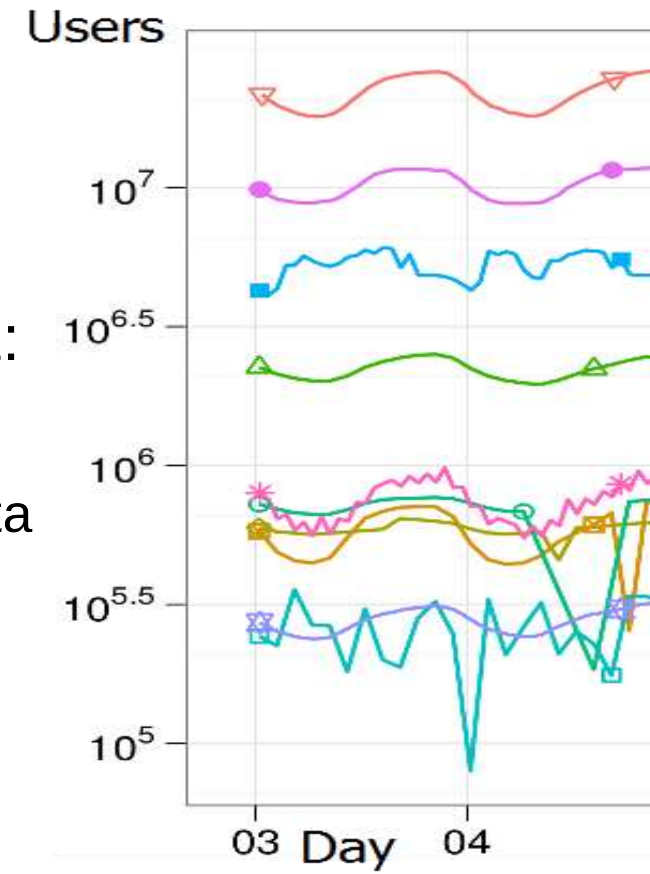


Most used protocol on Internet, by upload volume [1]
One third (US) to half (EU) of residential upload
Over 100 million users [2]

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

BTWorld: a Typical Big Data Project

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



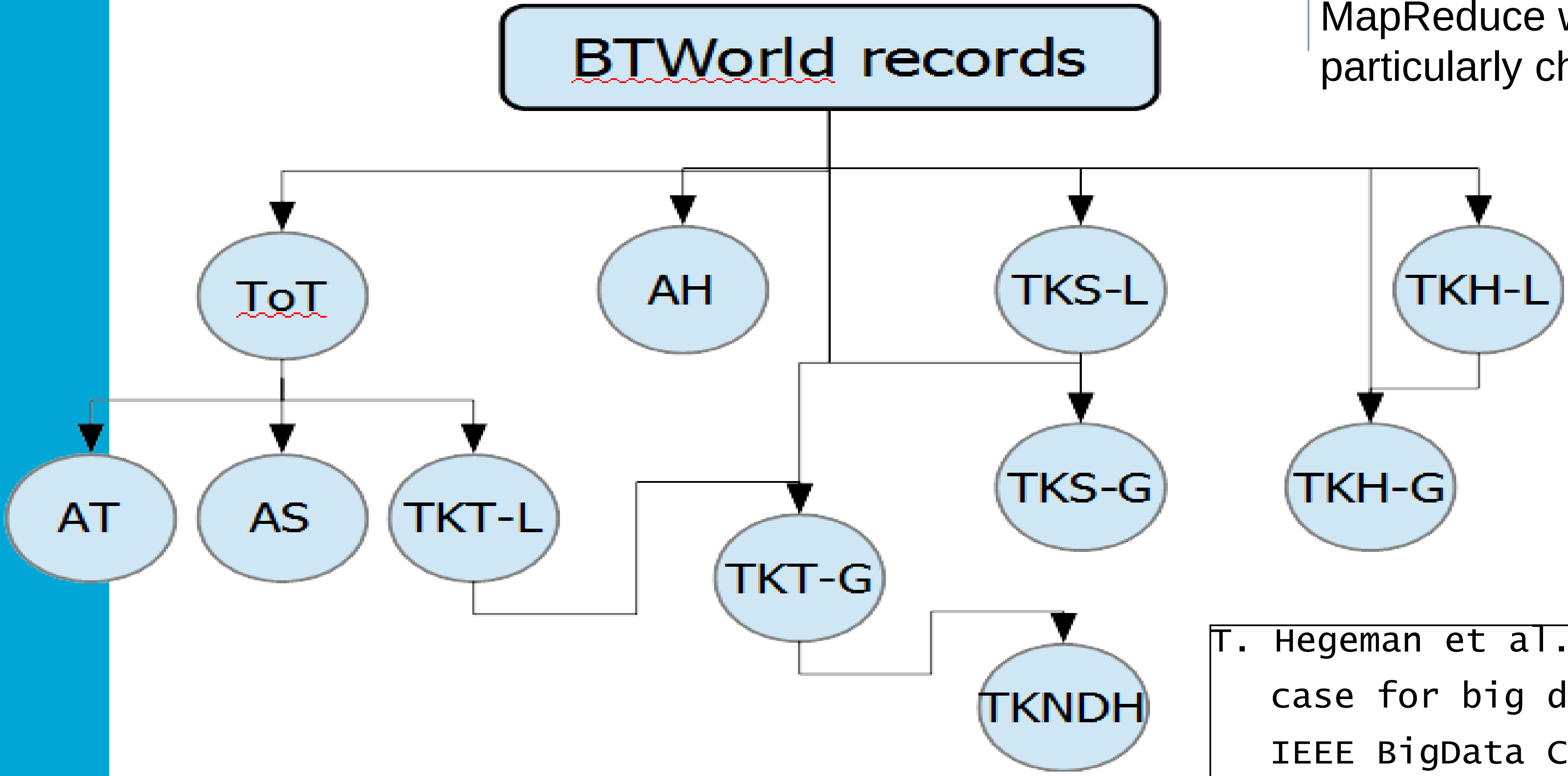
TU Delft

COMMIT/

DELFT DATA SCIENCE

The Abstract BTWorld Workflow

Workflows pose significant scheduling challenges, and MapReduce workflows can be particularly challenging



T. Hegeman et al. The BTWorld use case for big data analytics. IEEE BigData Conference 2013

Scalable High Performance Systems

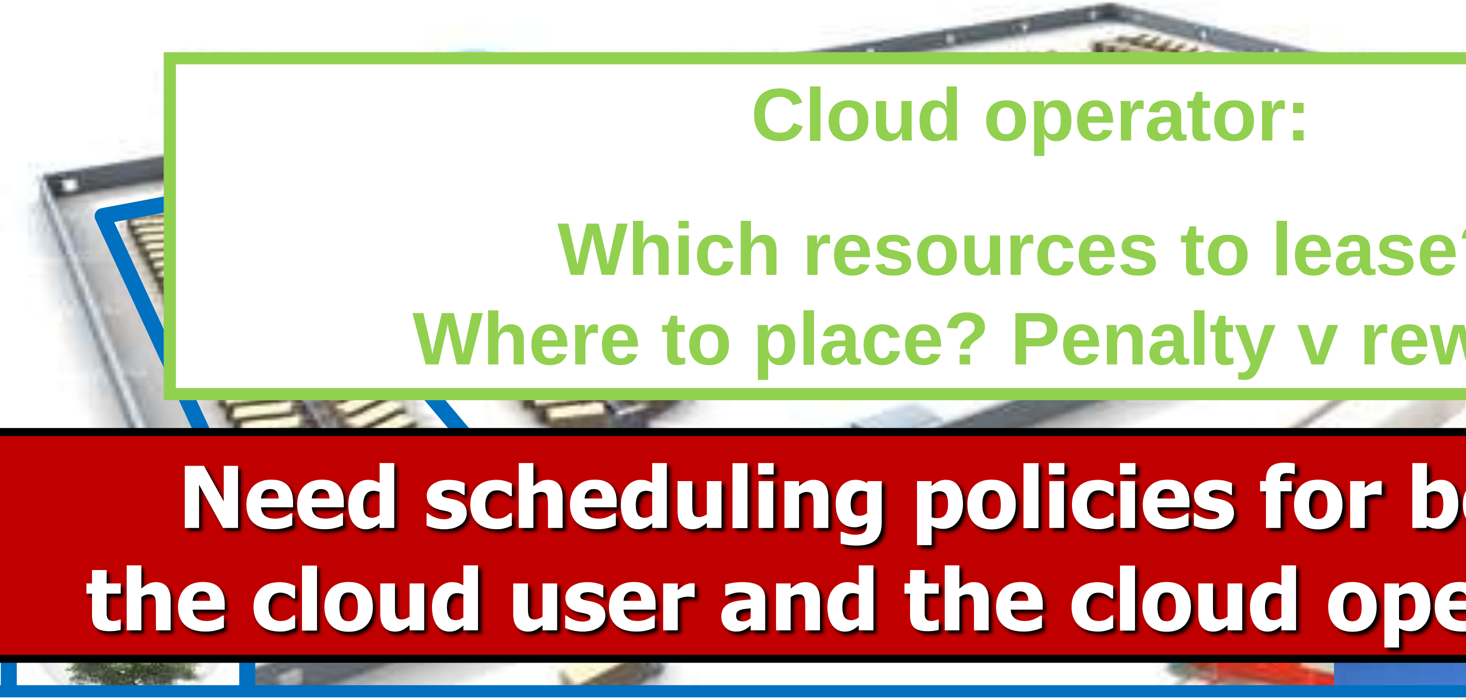
Interaction Encouraged!

- 2' — Where and What Is TU Delft/the PDS group?
- 5' — The Golden Age of Datacenters
- 5' — A Delft View on Datacenter Technology
  - The main challenges
- 35' — The Delft Approach to Making Datacenters Tick
  - Addressing the New World Challenge
  - Addressing the Scheduling challenge
  - Addressing the Ecosystem Navigation challenge
  - Addressing the Big Cake challenge
  - Addressing Jevons Effect in Datacenters
- 10' — Towards a Collaboration on Datacenter Technology

TU Delft

DELFT DATA SCIENCE

The Scheduling Challenge



Cloud operator:

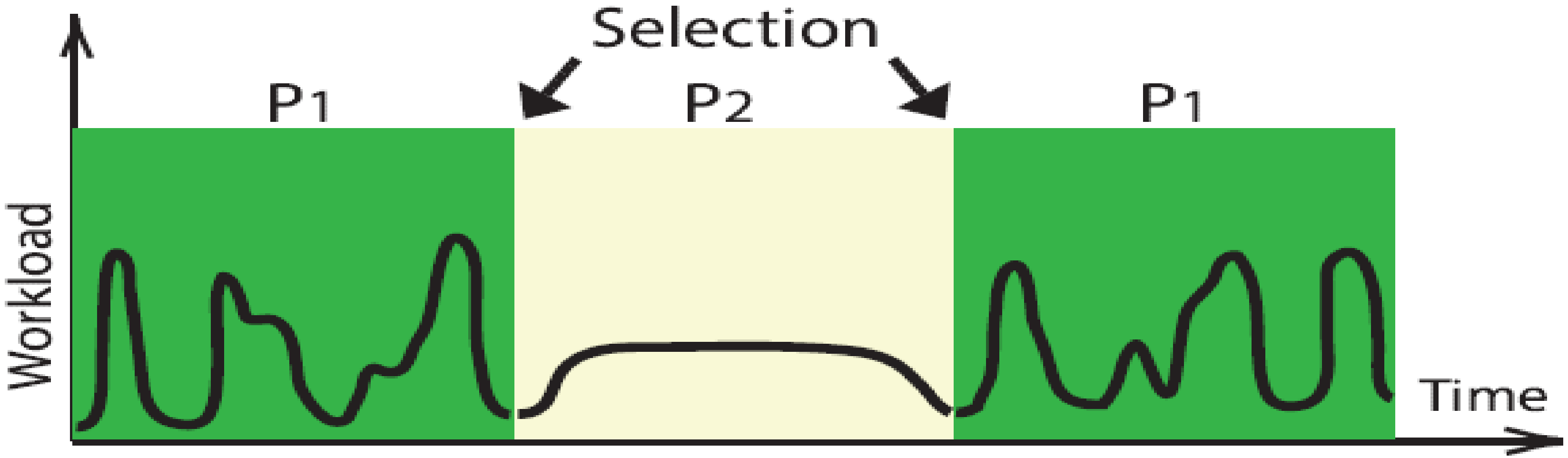
Which resources to lease?
Where to place? Penalty v reward?

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

Cloud customer:

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

Portfolio Scheduling, In A Nutshell

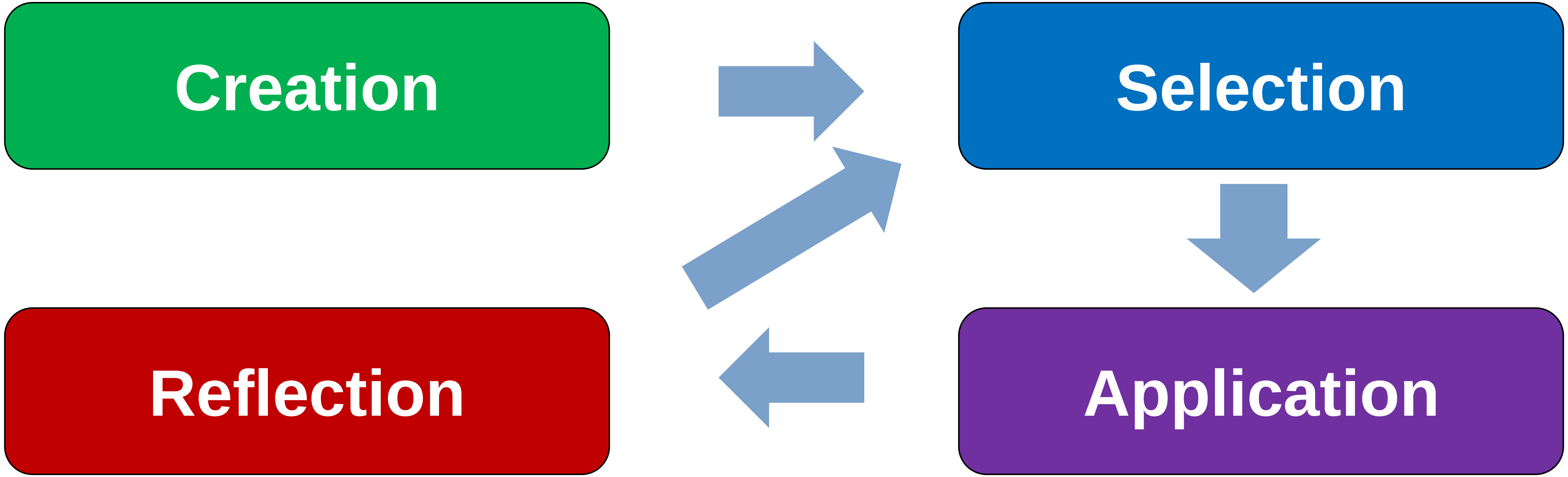


- Create a set of scheduling policies
  - Resource provisioning and allocation policies for datacenters
- Online selection of the active policy, at important moments

Portfolio Scheduling: Process

Which policies to include?

Which policy to activate?



Which changes to the portfolio?

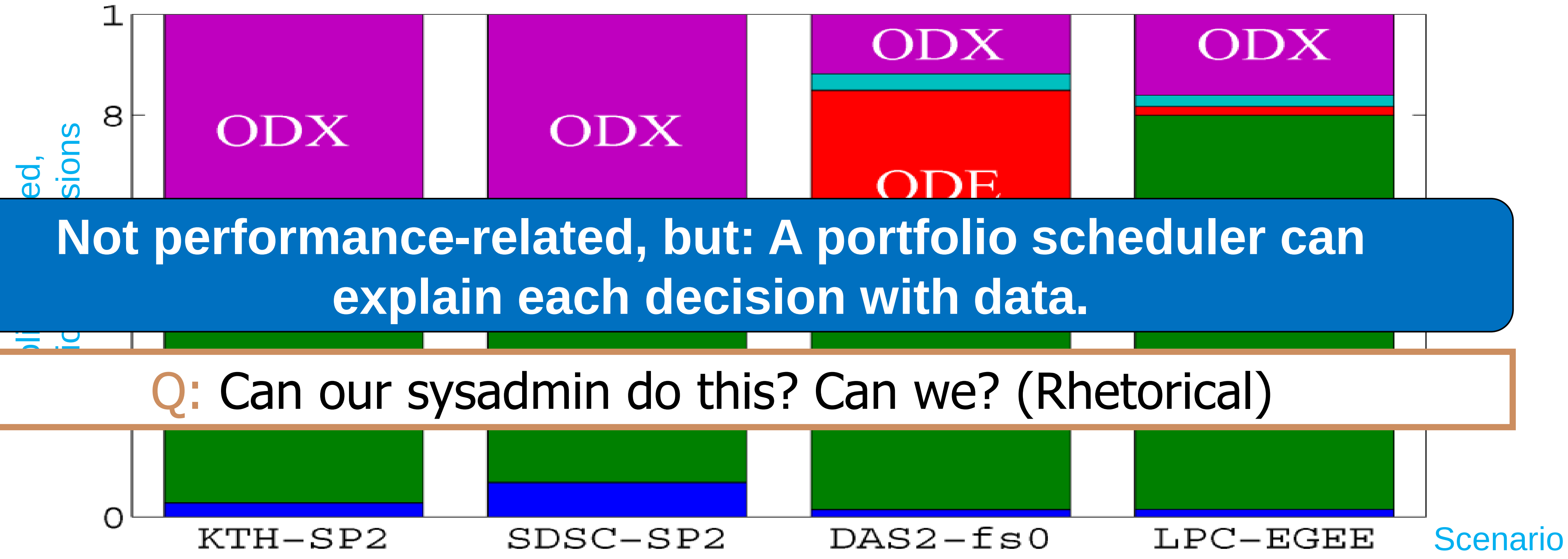
Which resources? What to log?

TUDelft

COMMIT/

V. van Beek et al. Mnemos: Self-Expressive Management of Business-Critical Workloads in Virtualized Datacenters. IEEE Computer 2015

Promising Results for Scientific Computing, Business-Critical, and Online Gaming



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

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

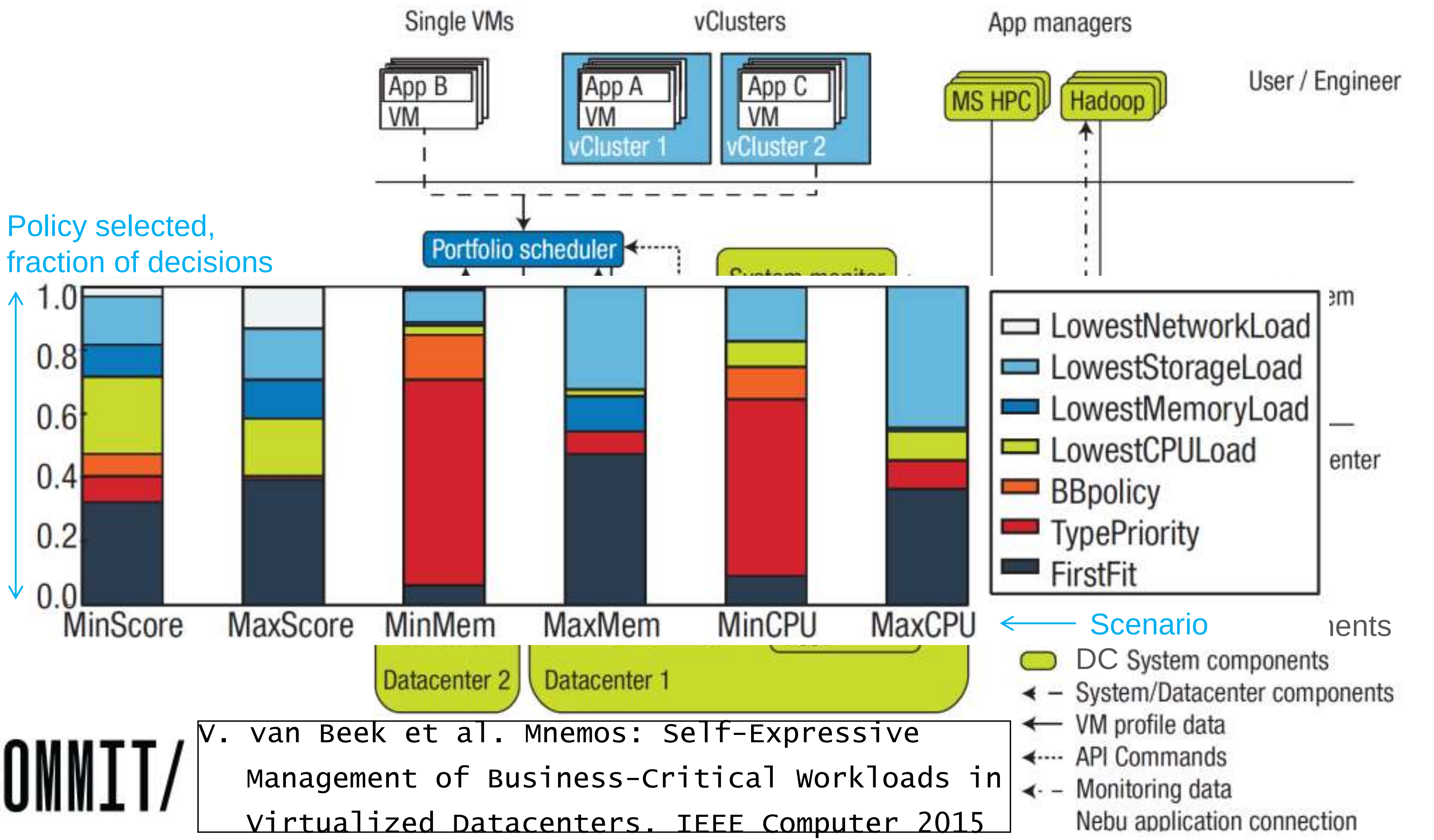
- No single dominant policy, even for complex policies

COMMIT/

K. Deng et al. Exploring Portfolio Scheduling for Long-Term Execution of Scientific Workloads in IaaS Clouds. SC|13

Portfolio Scheduling in Practice

Policy selected,
fraction of decisions



Scenario

V. van Beek et al. Mnemos: Self-Expressive Management of Business-Critical Workloads in Virtualized Datacenters. IEEE Computer 2015

COMMIT/

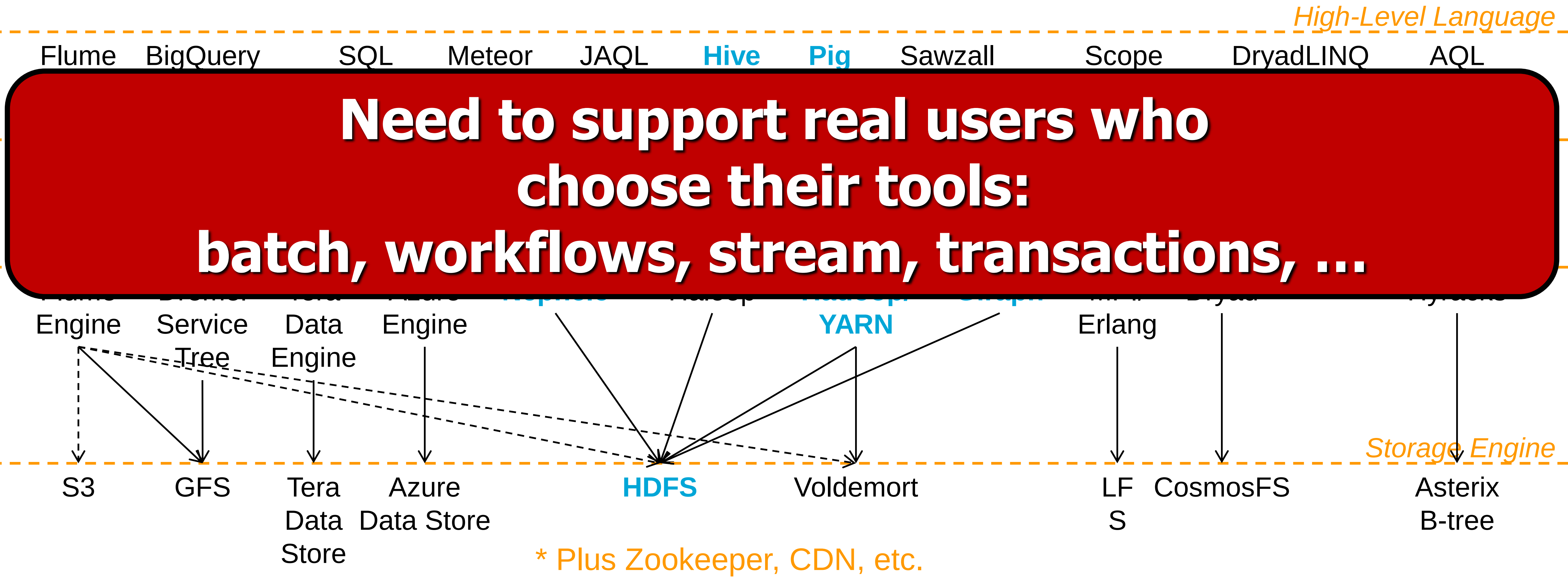
TU Delft

Scalable High Performance Systems

Interaction Encouraged!

- 2' —— Where and What Is TU Delft/the PDS group?
- 5' —— The Golden Age of Datacenters
- 5' —— A Delft View on Datacenter Technology
  - The main challenges
- 35' —— The Delft Approach to Making Datacenters Tick
  - Addressing the New World Challenge
  - Addressing the Scheduling challenge
  - Addressing the Ecosystem Navigation challenge
  - Addressing the Big Cake challenge
  - Addressing Jevons Effect in Datacenters
- 10' —— Towards a Collaboration on Datacenter Technology

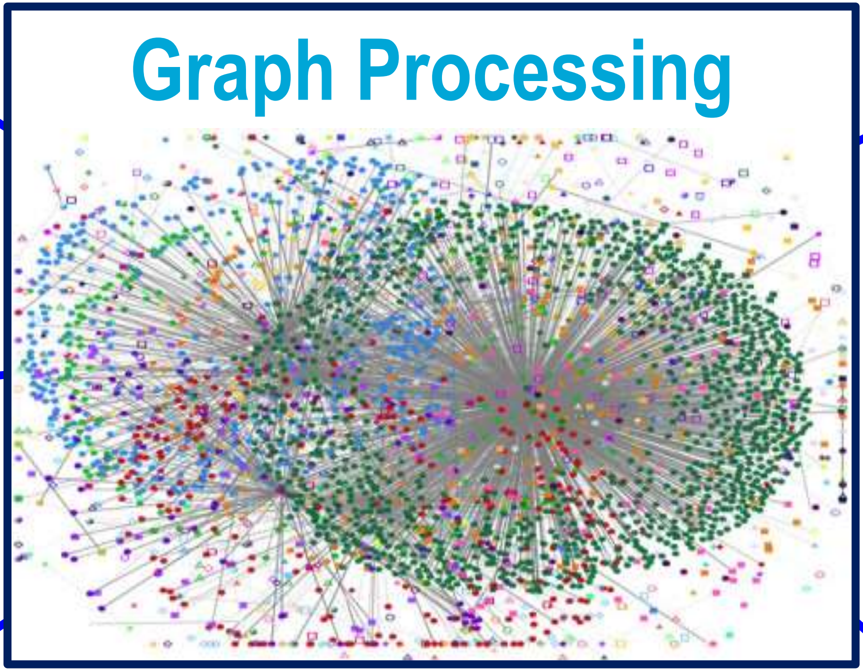
The Ecosystem Navigation Challenge



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

* Plus Zookeeper, CDN, etc.

The data deluge: large-scale graphs tens of Billions of Edges



Platform Diversity

Neo4j

Apache Giraph

Oracle Labs PGX

YARN

hadoop

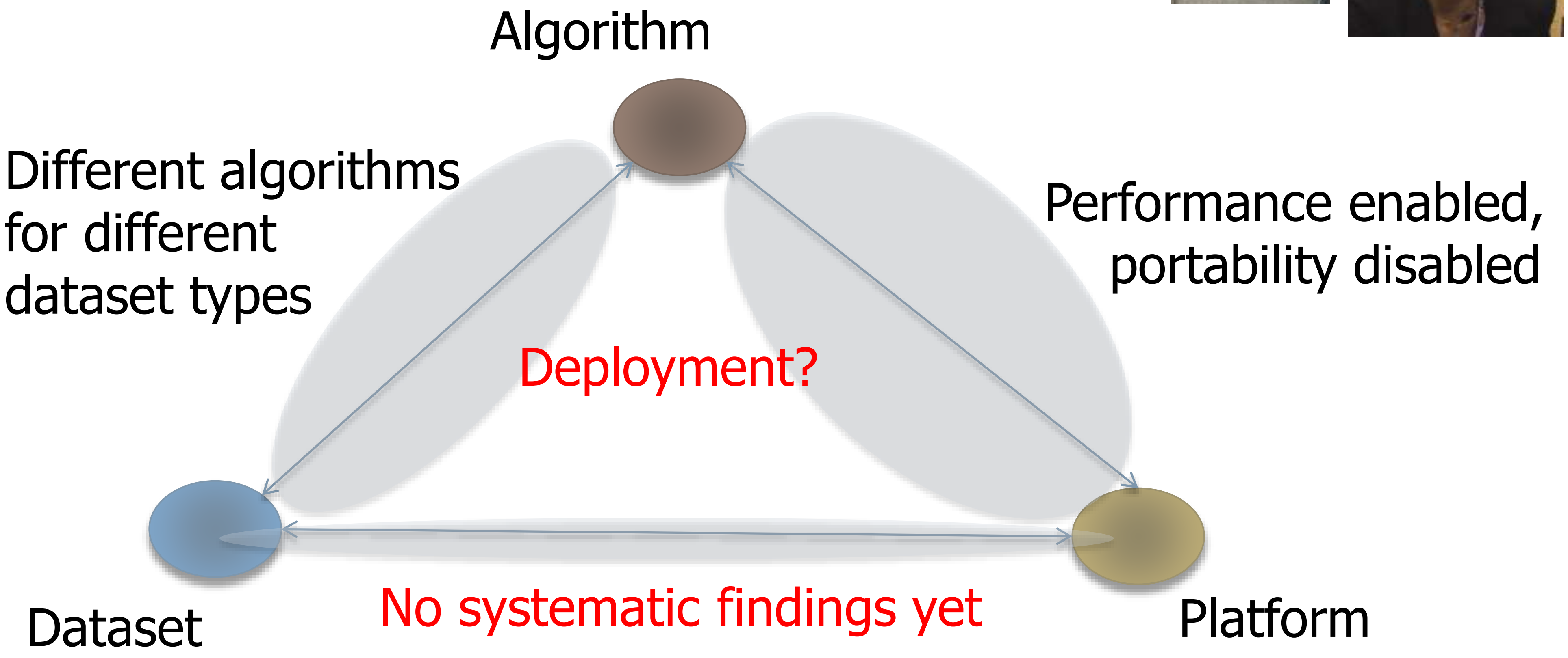
Spark GraphX

GraphLab

StratoSphere Above the Clouds

Y. Guo et al. How Well Do Graph-Processing Platforms Perform? An Empirical Performance Evaluation and Analysis. IPDPS 2014: 395-404

Ecosystem Navigation = Understanding the PAD Triangle



A. L. Varbanescu et al. Can Portability Improve Performance? An Empirical Study of Parallel Graph Analytics. ICPE 2015: 277-287

A. Iosup et al. Towards Benchmarking IaaS and PaaS Clouds for Graph Analytics. WBDB 2014: 109-131

TU Delft

COMMIT/

DELFT DATA SCIENCE

Graphalytics: The first comprehensive benchmark for big data graph processing

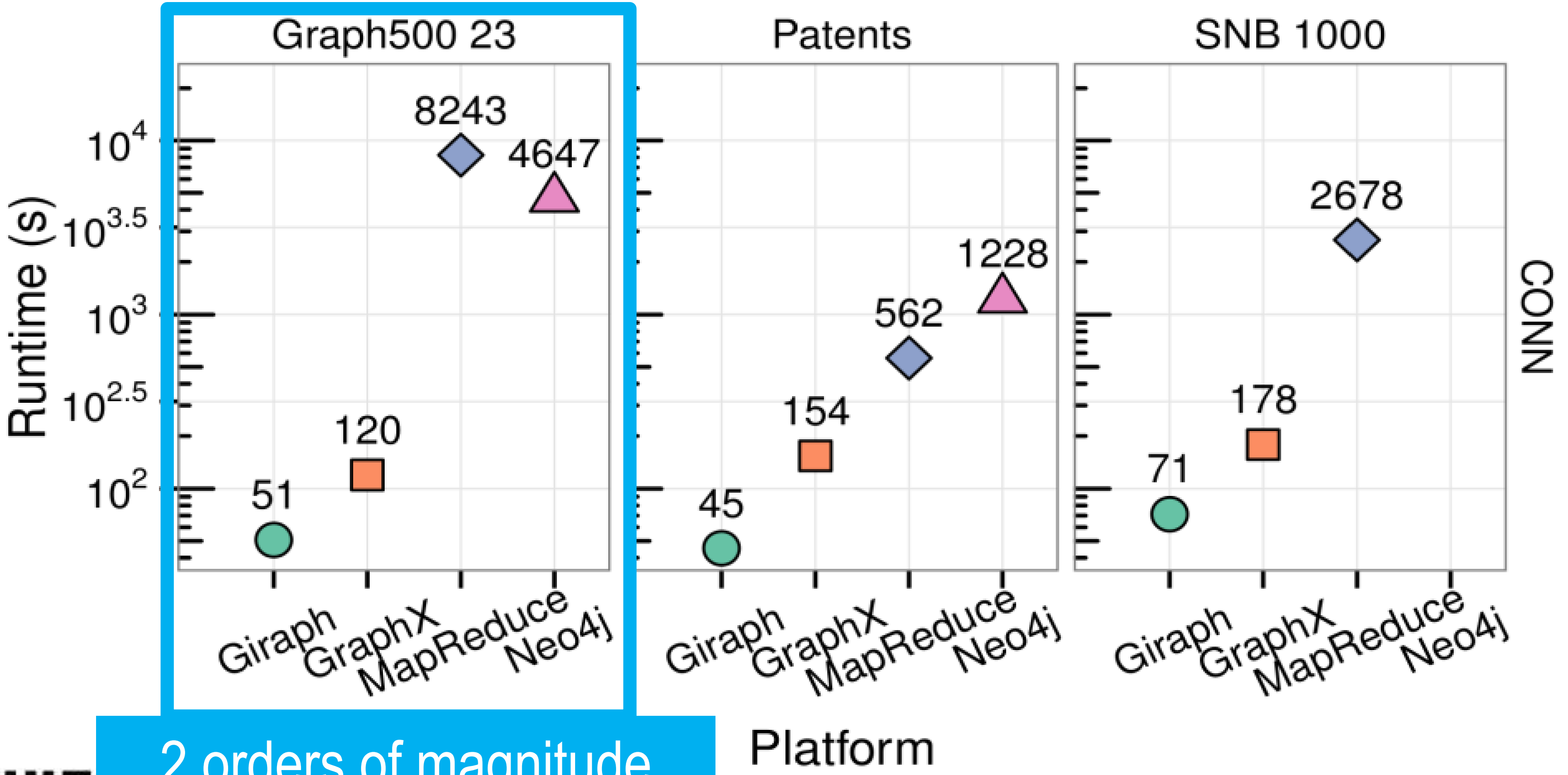
https://github.com/tudelft-atlarge/graphalytics/

A PAD triangle explorer for Graph Processing

- Advanced benchmarking harness
- Choke-point analysis
- Real data + Realistic graph generator, many algos

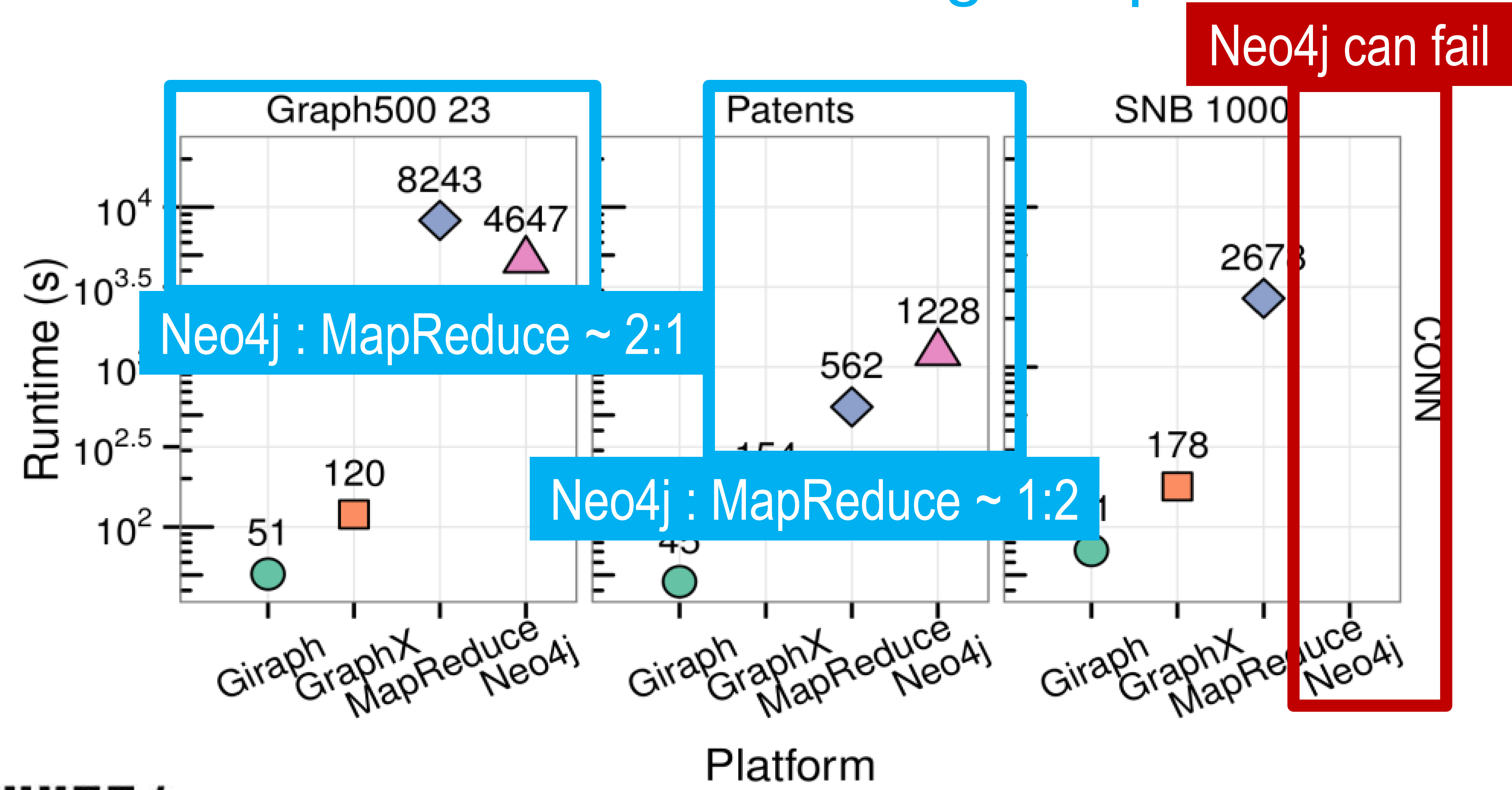
- Co-sponsored by Oracle Labs, Intel Labs
- Supported by LDBC, partially developed through SPEC RG

Runtime: the Platform has large impact

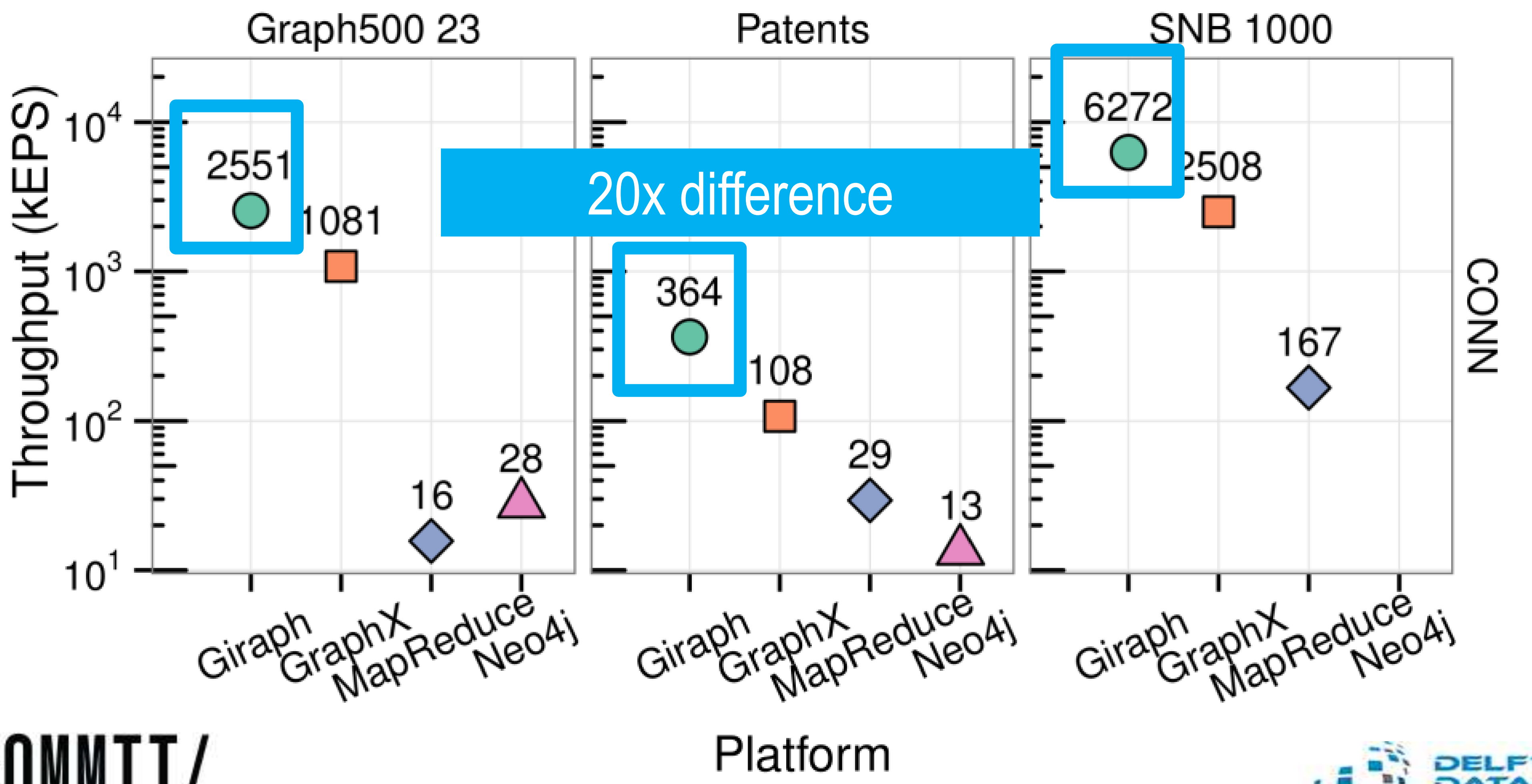


2 orders of magnitude difference due to platform

Runtime: The Dataset has large impact



Throughput: The Dataset structure matters!



COMMIT/

TU Delft

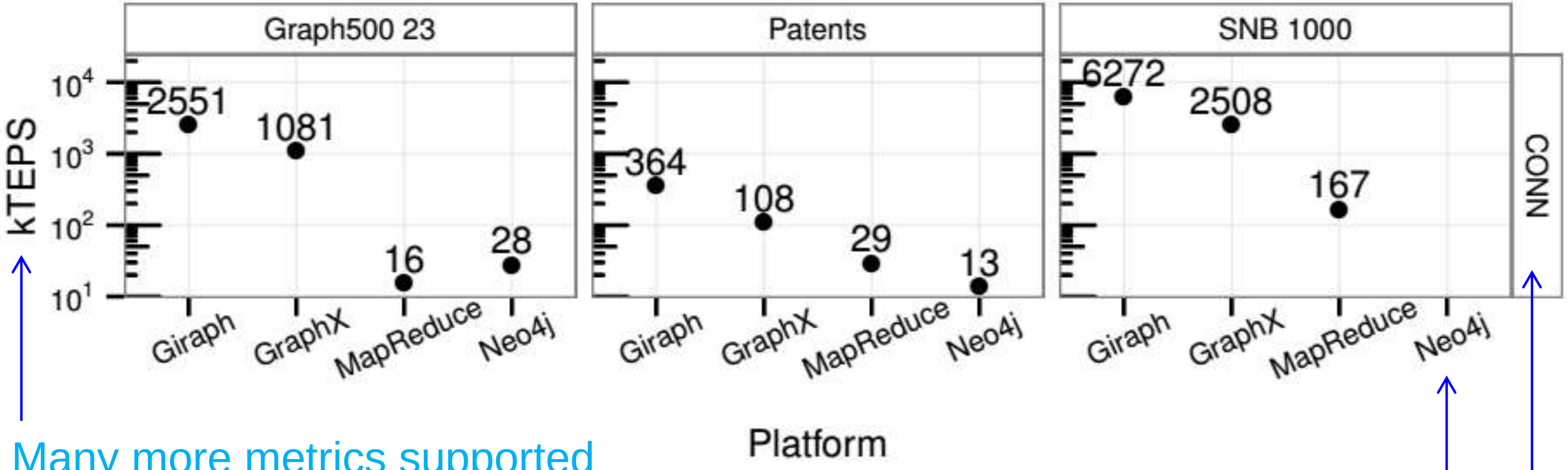
DELFT DATA SCIENCE

Graphalytics in Practice

https://github.com/tudelft-atlarge/graphalytics/

6 real-world datasets +
2 synthetic generators

Data ingestion not included in this graph.



Many more metrics supported

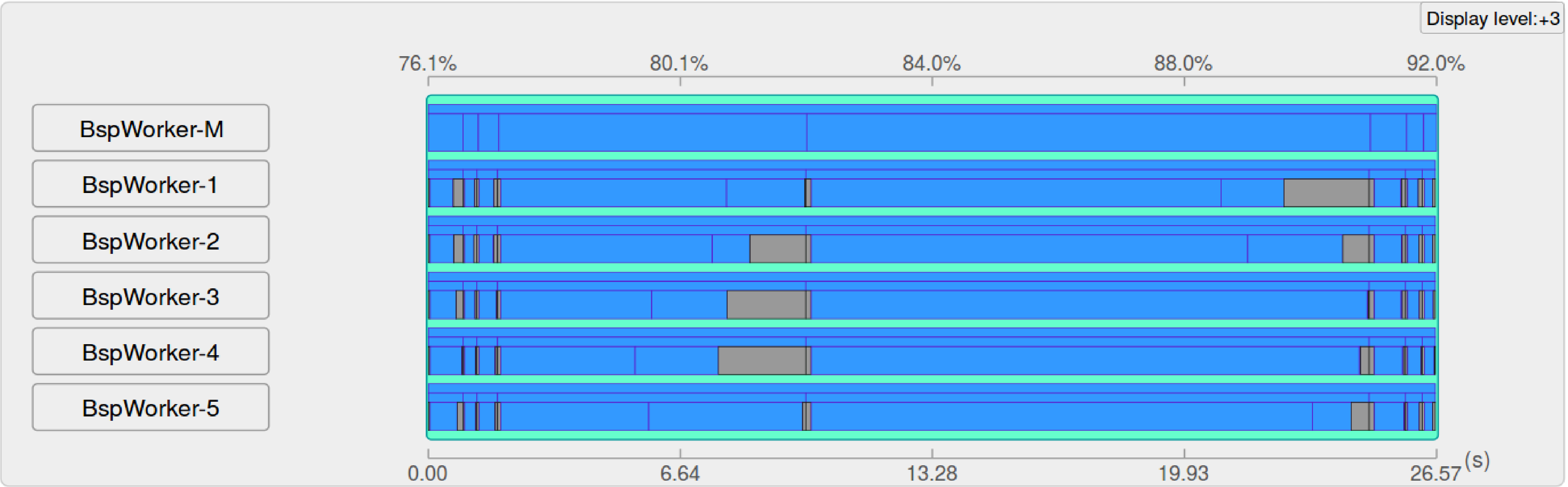
10 platforms tested w prototype implementation

5 classes of algorithms

M. Capota et al., Graphalytics: A Big Data Benchmark for Graph-Processing Platforms. SIGMOD GRADES 2015

Graphalytics in Practice

https://github.com/tudelft-atlarge/graphalytics/



Join us for the SC2015 tutorial, Nov 15 (tut149)

TUDelft COMMIT/ DELFT DATA SCIENCE

Scalable High Performance Systems

Interaction Encouraged!

2' — Where and What Is TU Delft/the PDS group?

5' — The Golden Age of Datacenters

5' — A Delft View on Datacenter Technology

- The main challenges

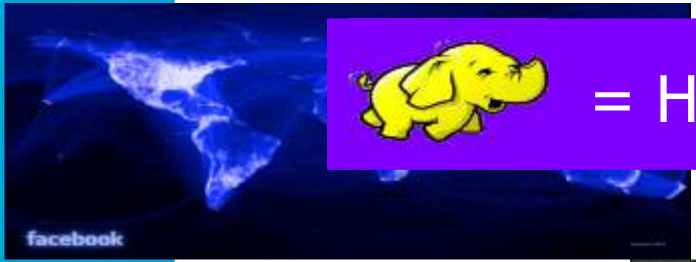
35' — The Delft Approach to Making Datacenters Tick

- Addressing the New World Challenge
- Addressing the Scheduling challenge
- Addressing the Ecosystem Navigation challenge
- Addressing the Big Cake challenge
- Addressing Jevons Effect in Datacenters

10' — Towards a Collaboration on Datacenter Technology

The "Big Cake" Challenge In the Datacenter

Online Social Networks



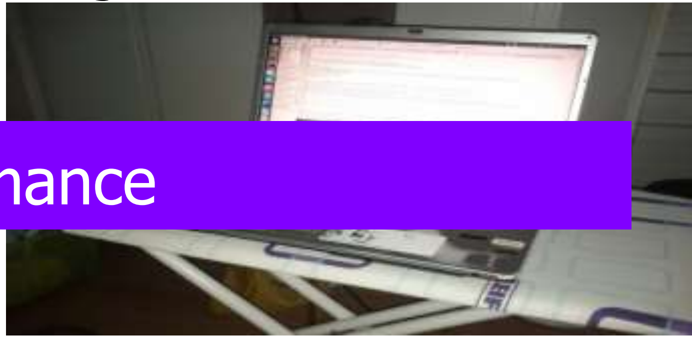
Financial Analysts

= Hadoop / MapReduce framework

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

Universe Explorers

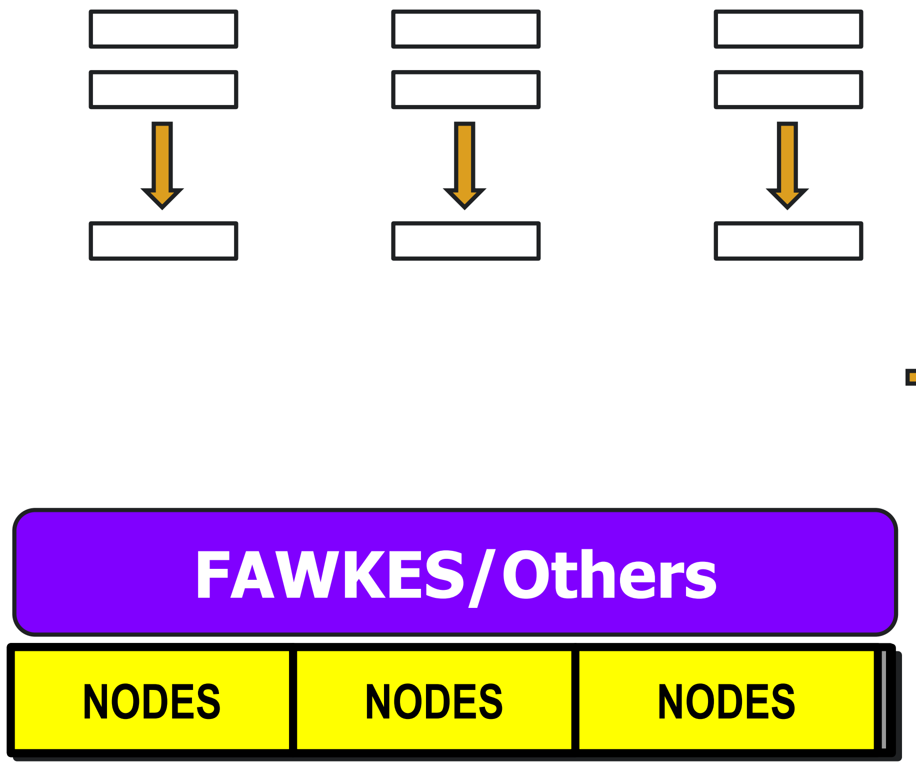
Big Data Enthusiast



Multiple frameworks = Isolation, especially performance

Dynamic Big Data Processing

Fawkes = Elastic MapReduce



Job submissions

Frameworks

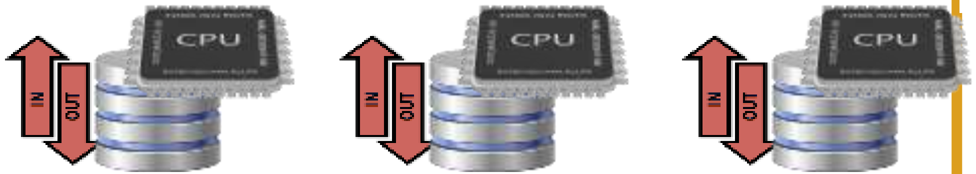
Resource manager

Infrastructure

B. Ghit et al. Balanced Resource Allocations Across Multiple Dynamic MapReduce Clusters. SIGMETRICS 2014

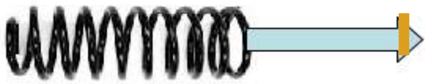
Elasticity for MapReduce Frameworrks

Core nodes

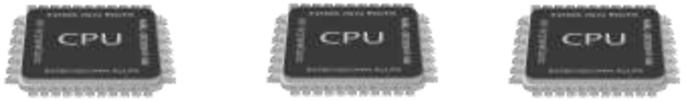


INPUT/OUTPUT DATA

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

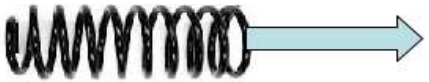


Transient nodes (TR)



NO DATA

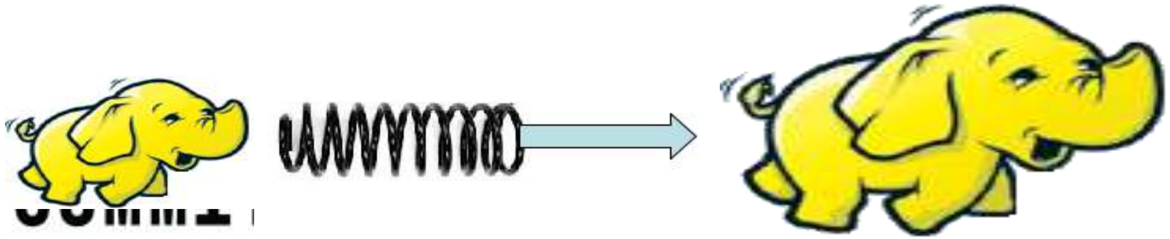
- No local storage
- R/W from/to core nodes
- **Instant removal**



Trans-core nodes (TC)

OUTPUT DATA

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



Fawkes in a Nutshell [1/2]

Because workloads may be time-varying:

- Poor resource utilization
- Imbalanced service levels

$$w_1 < w_2 < w_3$$

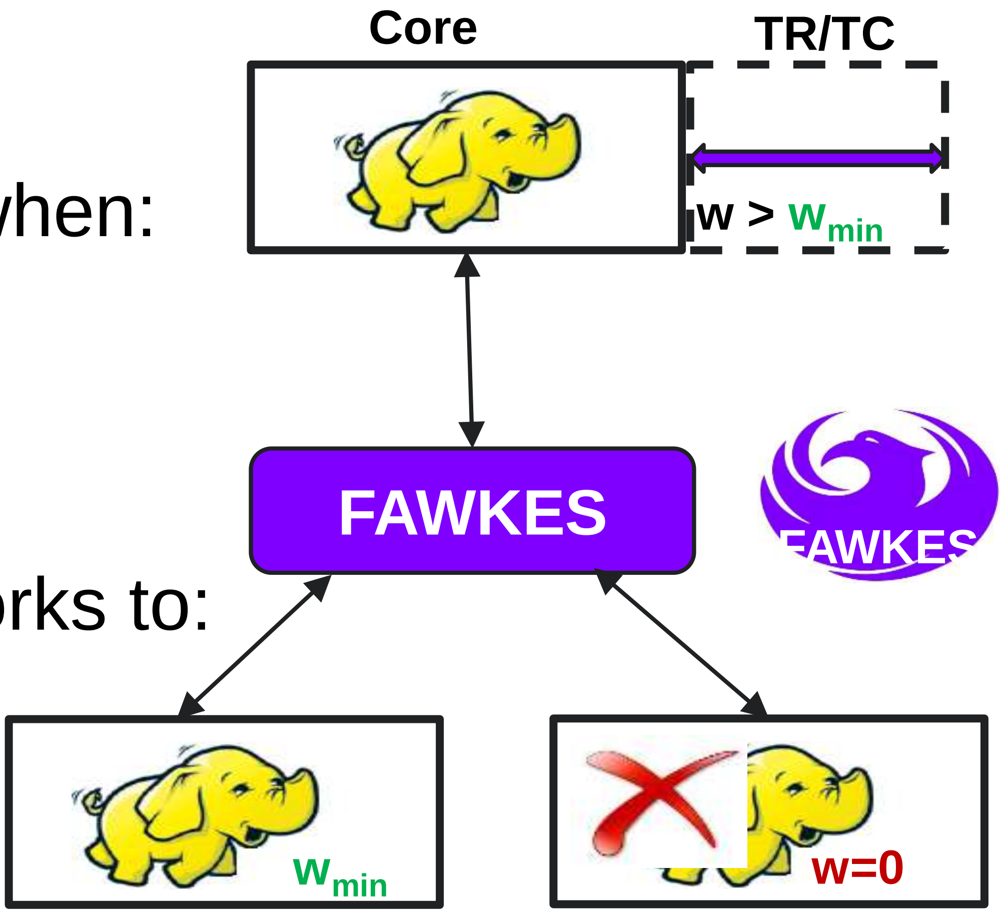
1. Fair framework size:

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

Fawkes in a Nutshell [2/2]

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

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



Performance of dynamic MapReduce

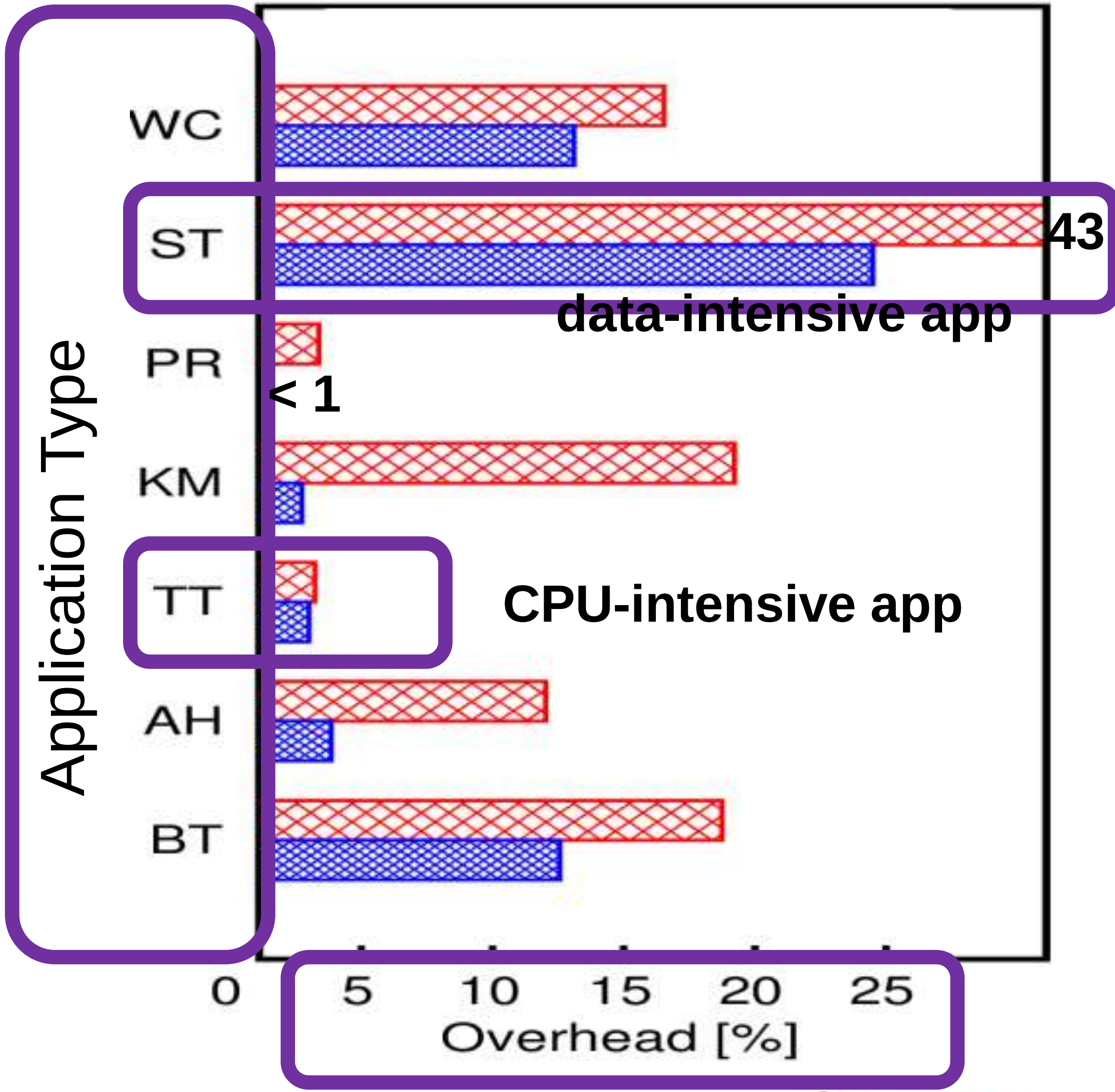
10 core + 10xTR
10 core + 10xTC
vs.
20 core nodes (baseline)

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

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

Dynamic MapReduce:
< 25% overhead

Fawkes also reduces imbalance



Scalable High Performance Systems

Interaction Encouraged!

- 2' — Where and What Is TU Delft/the PDS group?
- 5' — The Golden Age of Datacenters
- 5' — A Delft View on Datacenter Technology
  - The main challenges
- 35' — The Delft Approach to Making Datacenters Tick
  - Addressing the New World Challenge
  - Addressing the Scheduling challenge
  - Addressing the Ecosystem Navigation challenge
  - Addressing the Big Cake challenge
  - Addressing Jevons Effect in Datacenters
- 10' — Towards a Collaboration on Datacenter Technology

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

Data Deluge =

data generated by humans and devices (IoT)

- Interacting
- Understanding
- Deciding
- Creating

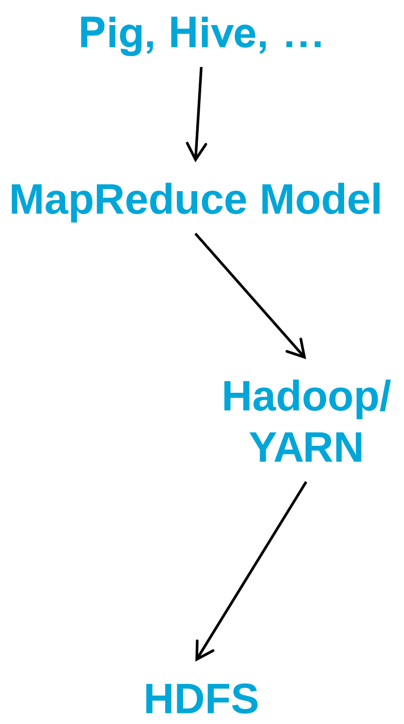
Need to address
Volume, Velocity, Variety of Big Data*

Vicissitude of Big Data = dynamic mix of big data issues (Vs) that lead in big data systems to different bottlenecks over time

Sources: IDC, EMC.

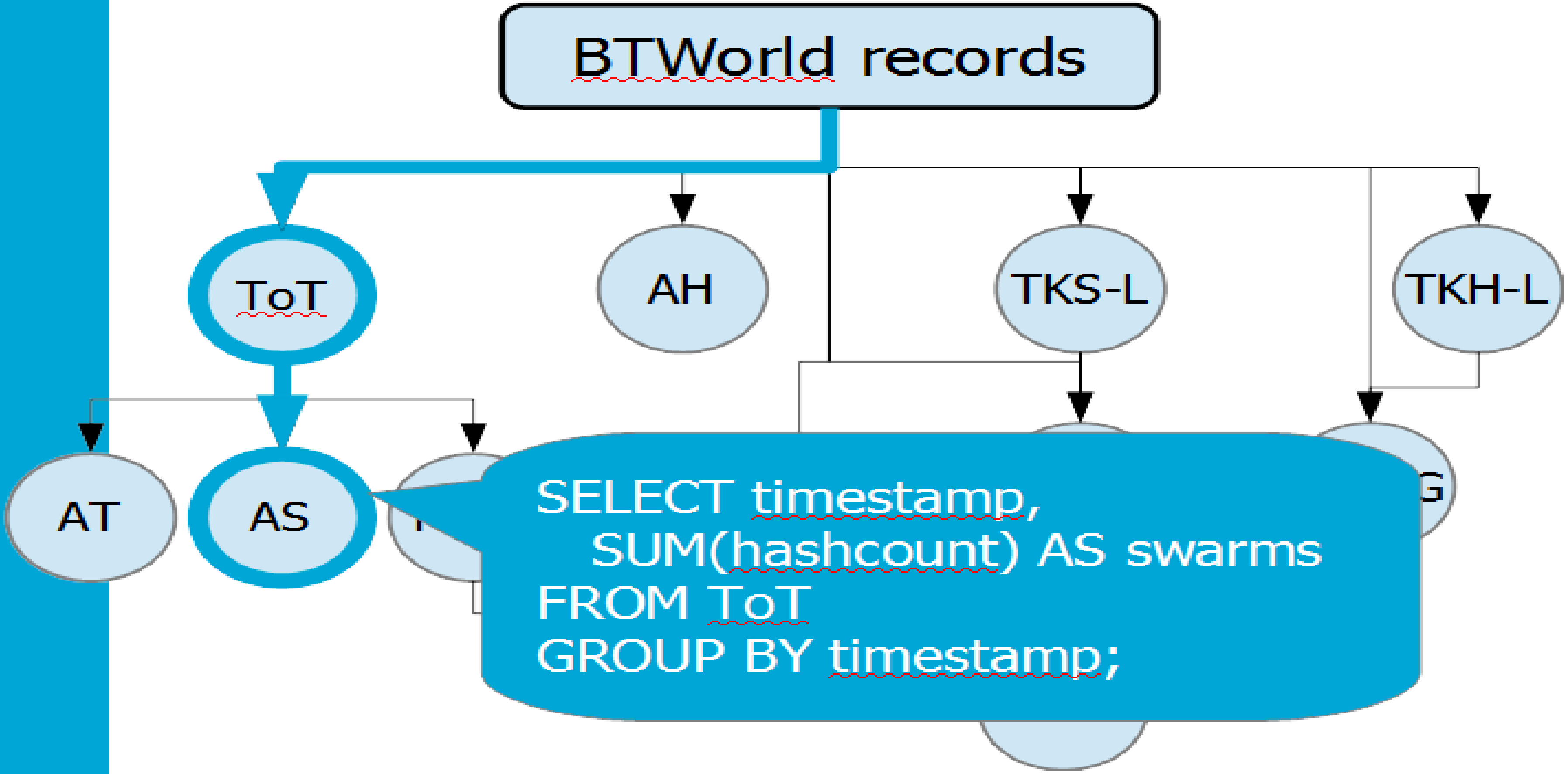
The MapReduce ecosystem (a big problem in big data)

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

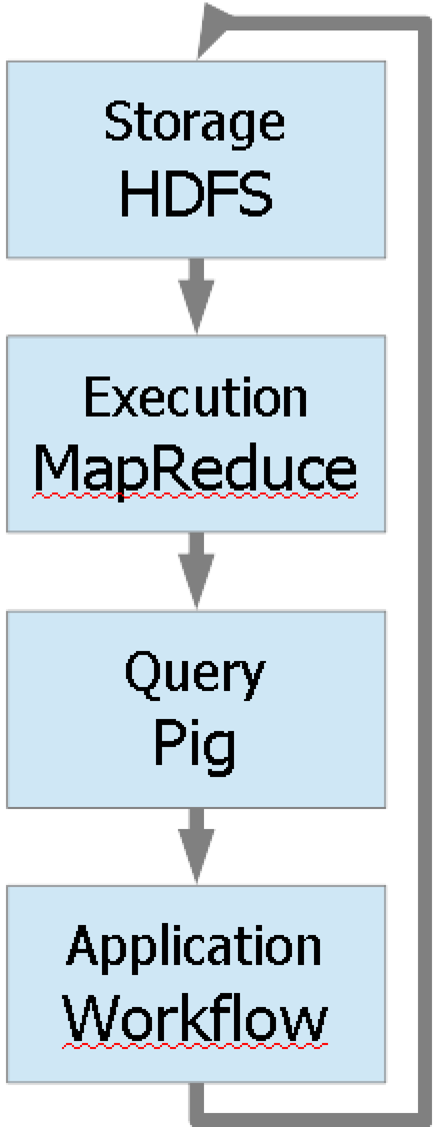


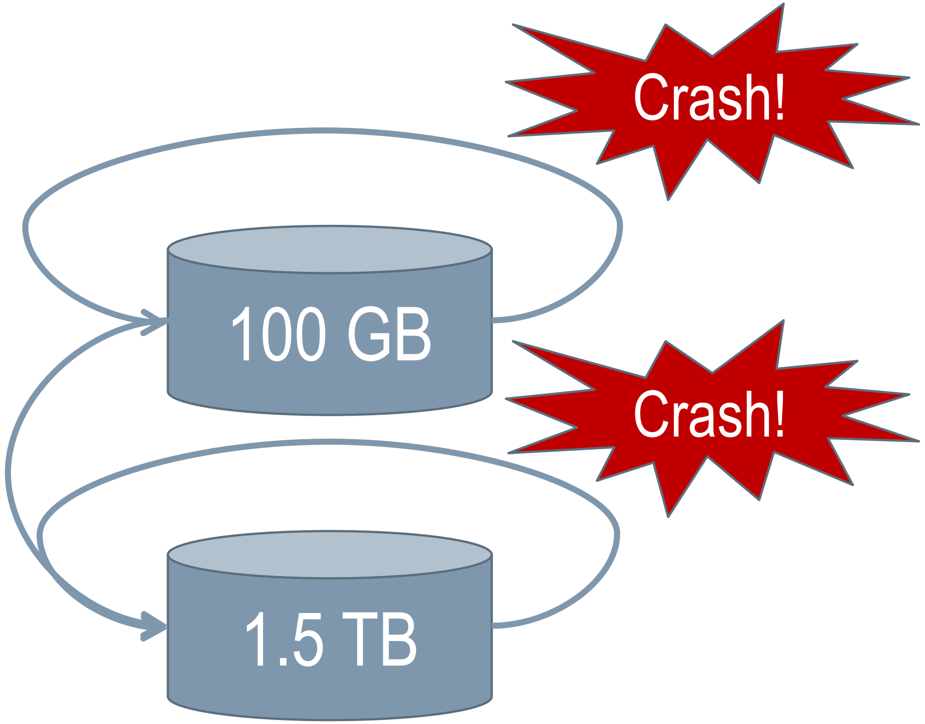
[1] Ewen et al., "Spinning Fast Iterative Data Flows", PVLDB 2012

The BTWorld Workload



Optimization Cycle





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

B. Ghit et al. V for Vicissitude: The Challenge of Scaling Complex Big Data Workflows. CCGRID 2014

COMMIT/

TUDelft

General Problem

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

Won IEEE Scale Challenge 2014!

Scalable High Performance Systems

Interaction Encouraged!

2' — Where and What Is TU Delft/the PDS group?

5' — The Golden Age of Datacenters

5' — A Delft View on Datacenter Technology

- The main challenges

35' — The Delft Approach to Making Datacenters Tick

- Addressing the New World Challenge
- Addressing the Scheduling challenge
- Addressing the Ecosystem Navigation challenge
- Addressing the Big Cake challenge
- Addressing Jevons Effect in Datacenters

10' — Towards a Collaboration on Datacenter Technology



Take-Home Message

The Golden Age of datacenters

Cloud computing + Big Data

Important New Challenges

1. The New World challenge
2. The scheduling challenge
3. The ecosystem navigation challenge
4. The big cake challenge
5. Jevons Effect for Big Data






TU Delft

DELFT DATA SCIENCE

Research Agenda for Datacenter-related Research

1. Characteristics and models of datacenter workloads.
2. Compute- & data-intensive models can coexist in the datacenter.
3. Non-functional targets: high performance and availability, elasticity, etc.
4. Fundamental models of datacenter operation.
5. Fundamental knowledge on Datacenter-Framework-App-Data interaction.
6. New generation of resource management techniques, including scheduling.
7. Benchmarking datacenter services.

TUDelft

DELFT DATA SCIENCE

Contact Us!

Staff members

A.Iosup@tudelft.nl
+31-15-2784433
@AIosup
http://pds.twi.tudelft.nl/~iosup/
https://www.linkedin.com/in/aiosup
PDS Group, Faculty EEMCS, TU Delft
Room HB07.050, Mekelweg 4, 2628CD Delft

TUDelft

DELFT DATA SCIENCE

Recommended Reading

Elastic Big Data and Computing

- V. van Beek (Solvinity/Bitbrains), J. Donkervliet, T. Hegeman, S. Hugtenburg, A. Iosup: Self-Expressive Management of Business-Critical Workloads in Virtualized Datacenters. IEEE Computer 48(7): 46-54 (2015)
- B. Ghit, N. Yigitbasi (Intel Research Labs, Portland), A. Iosup, and D. Epema. Balanced Resource Allocations Across Multiple Dynamic MapReduce Clusters. SIGMETRICS 2014
- L. Fei, B. Ghit, A. Iosup, D. H. J. Epema: KOALA-C: A task allocator for integrated multicluster and multicloud environments. CLUSTER 2014: 57-65

Time-Based Analytics

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

Graph Processing / Benchmarking

- A. Iosup, M. Capota, T. Hegeman, Y. Guo, W. L. Ngai, A. L. Varbanescu, M. Verstraaten: Towards Benchmarking IaaS and PaaS Clouds for Graph Analytics. WBDB 2014: 109-131
- Y. Guo, M. Biczak, A. L. Varbanescu, A. Iosup, C. Martella (Apache Giraph), T. L. Willke: How Well Do Graph-Processing Platforms Perform? An Empirical Performance Evaluation and Analysis. IPDPS 2014: 395-404
- A. L. Varbanescu, M. Verstraaten, C. de Laat, A. Penders, A. Iosup, H. J. Sips: Can Portability Improve Performance?: An Empirical Study of Parallel Graph Analytics. ICPE 2015: 277-287

Workloads

- S. Shen, V. van Beek (Solvinity/BitBrains), A. Iosup: Statistical Characterization of Business-Critical Workloads Hosted in Cloud Datacenters. CCGRID 2015: 465-474
- T. Hegeman, B. Ghit, M. Capota, J. Hidders, D. H. J. Epema, A. Iosup: The BTWorld use case for big data analytics: Description, MapReduce logical workflow, and empirical evaluation. IEEE BigData Conference 2013: 622-630.

Disclaimer: images used in this presentation obtained via Google Images.

- Images used in this lecture courtesy to many anonymous contributors to Google Images, and to Google Image Search.
- Many thanks!