Scalable High Performance Systems

@Alosup

dr. ir. Alexandru Iosup

Parallel and Distributed Systems Group

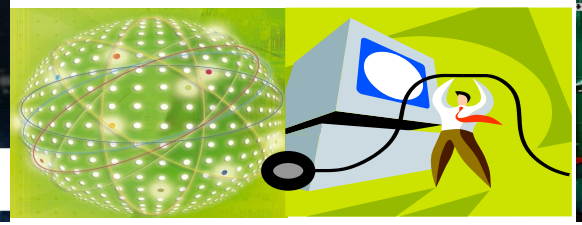
TU Delft

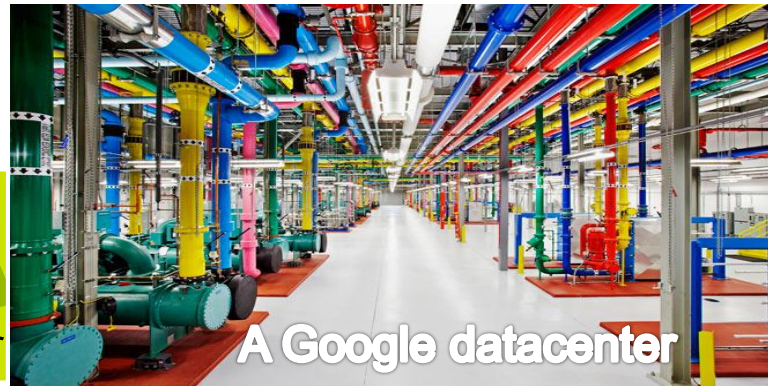
KIVI Engineering Society

Won IEEE Scale Challenge 2014!

DELFT DATA SCIENCE

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

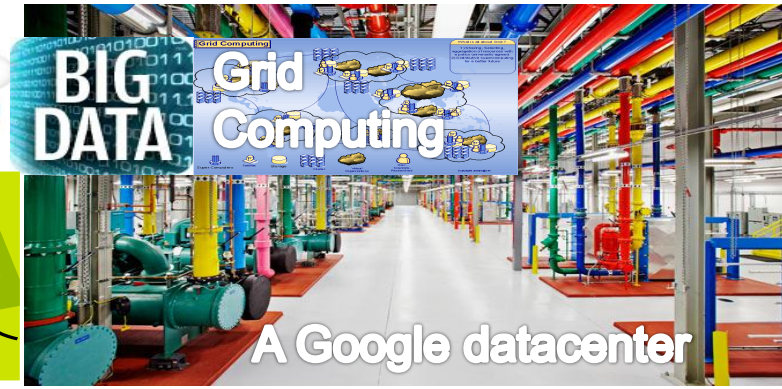
"my other computer is a datacenter"

TU Delft

KIVI Engineering Society

DELFT DATA SCIENCE

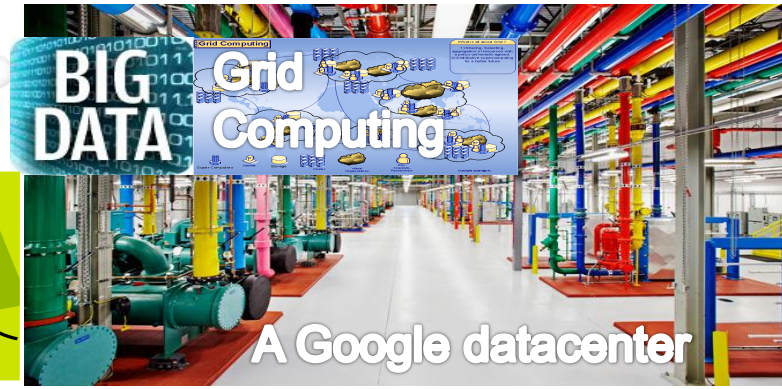
Scientific Challenges

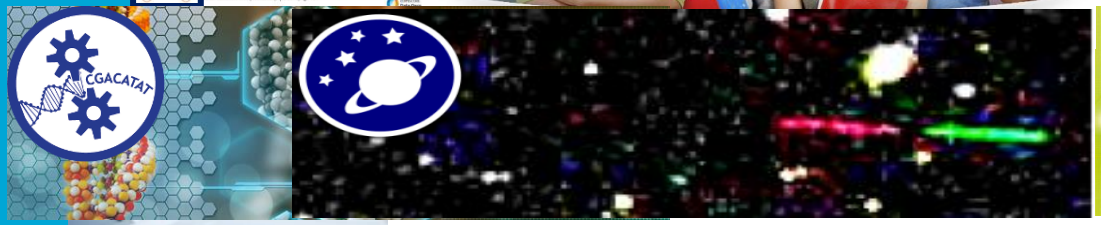


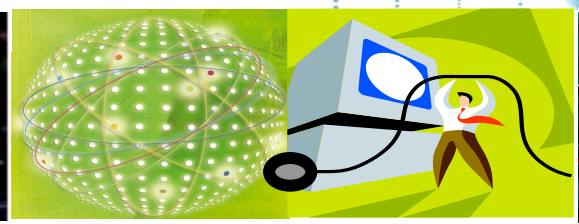
How to massivize datacenters?

- 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

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)
- “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: Netherlands as a world-wide ICT hub

TU Delft

KIVI Engineering Society

DELFT DATA SCIENCE

Scalable High Performance Systems

Interactive Masterclass

5’ — Pitch on Scalable High Performance Systems

5’ — The Golden Age of Datacenters

20’ — A Delft Data Science View on Datacenters

- The core idea of datacenter computing
- The main enabling technologies for datacenter computing
- The main challenges and techniques

35’ — Delft Data Science Makes Datacenters Tick

- Addressing the Scheduling challenge
- Addressing the Ecosystem Navigation challenge
- Addressing the Big Cake challenge
- Addressing Jevon’s Effect in Data Science

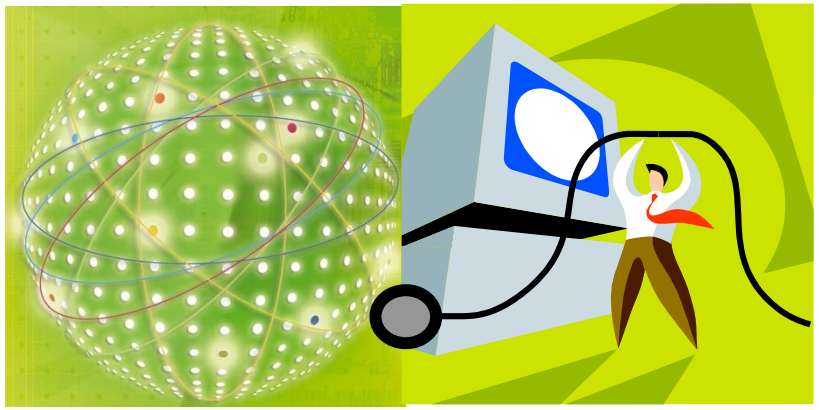
10’ — Towards a KIVI Taskforce on Data Science as a Service

TU Delft

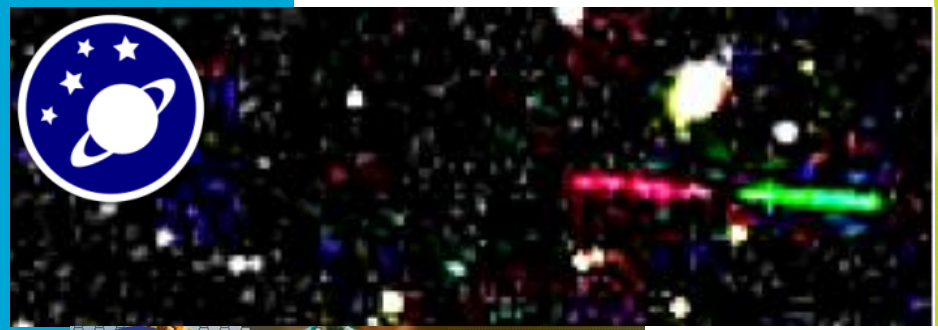
KIVI Engineering Society

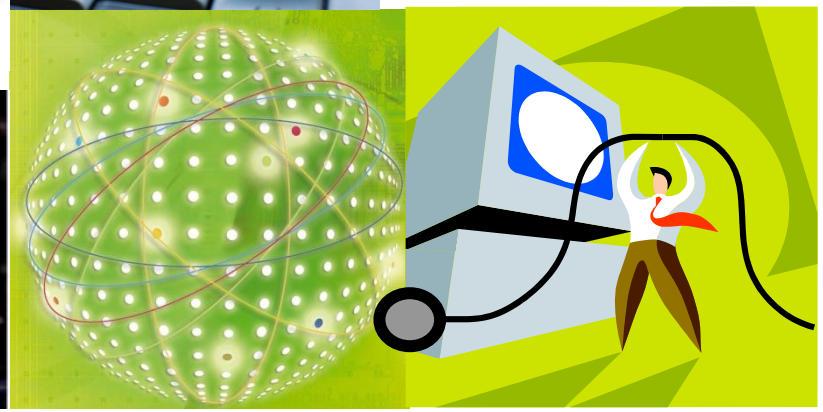
DELFT DATA SCIENCE

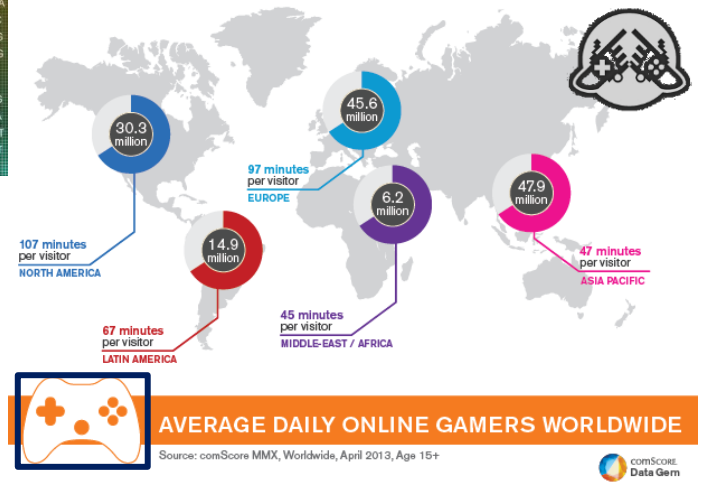
This Is the Golden Age of Datacenters



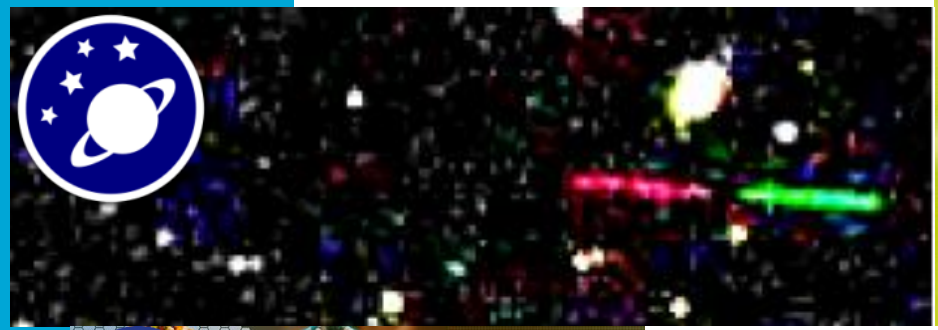
This Is the Golden Age of Datacenters

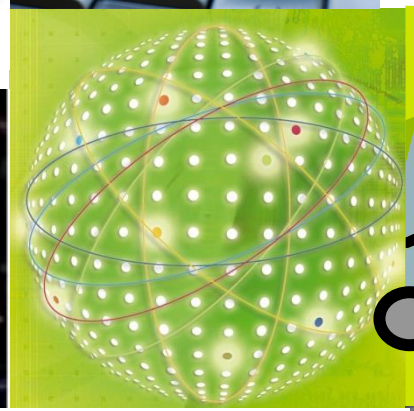


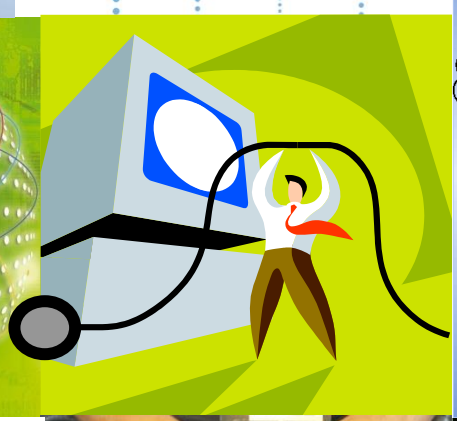


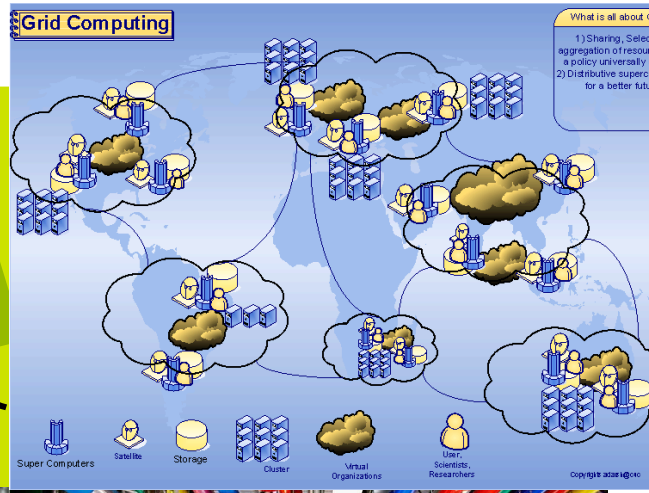


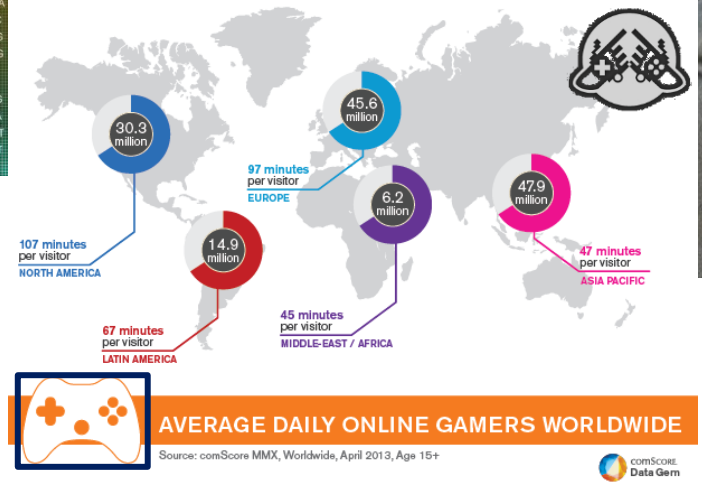
This Is the Golden Age of Datacenters

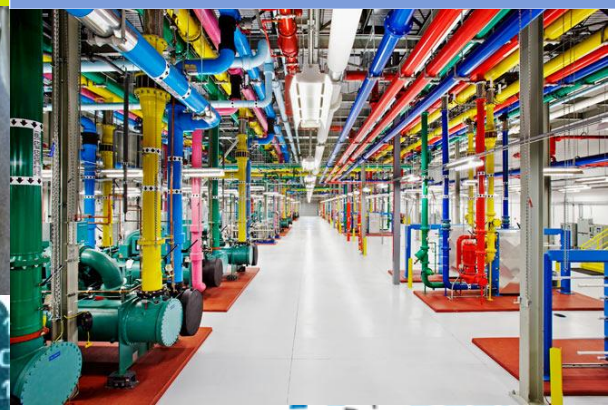












DATA SCIENCE

TU Delft

Scalable High Performance Systems

Interactive Masterclass

5’ — Pitch on Scalable High Performance Systems

5’ — The Golden Age of Datacenters

20’ — A Delft Data Science View on Datacenters

- The core idea of datacenter computing
- The main enabling technologies for datacenter computing
- The main challenges and techniques

35’ — Delft Data Science Makes Datacenters Tick

- Addressing the Scheduling challenge
- Addressing the Ecosystem Navigation challenge
- Addressing the Big Cake challenge
- Addressing Jevon’s Effect in Data Science

10’ — Towards a KIVI Taskforce on Data Science as a Service

TU Delft

KIVI Engineering Society

DELFT DATA SCIENCE

Scalable High Performance Systems

Interactive Masterclass

5’ — Pitch on Scalable High Performance Systems

5’ — The Golden Age of Datacenters

20’ — A Delft Data Science View on Datacenters

- The core idea of datacenter computing
- The main enabling technologies for datacenter computing
- The main challenges and techniques

35’ — Delft Data Science Makes Datacenters Tick

- Addressing the Scheduling challenge
- Addressing the Ecosystem Navigation challenge
- Addressing the Big Cake challenge
- Addressing Jevon’s Effect in Data Science

10’ — Towards a KIVI Taskforce on Data Science as a Service

TU Delft

KIVI Engineering Society

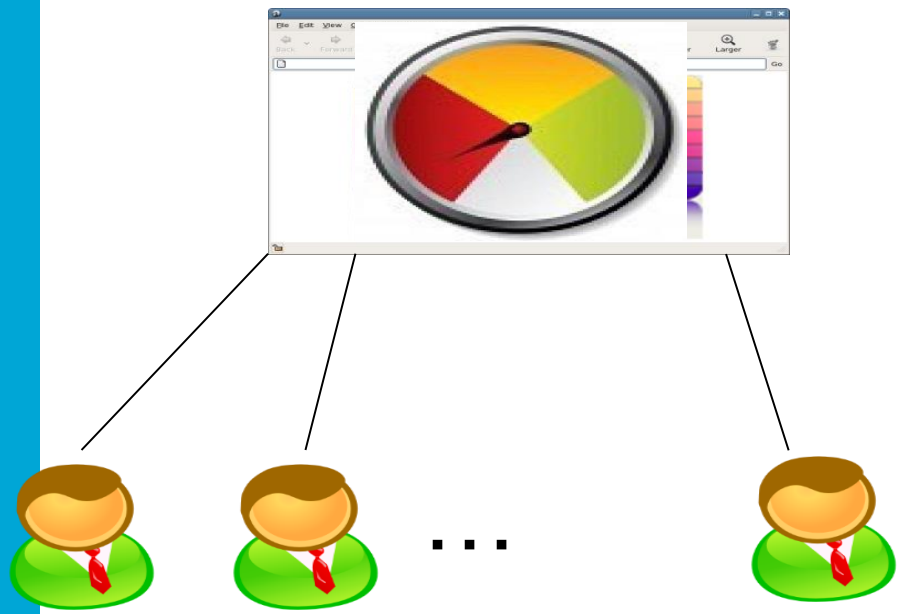
DELFT DATA SCIENCE

Joe Has an Idea ($$$)

Solution #1

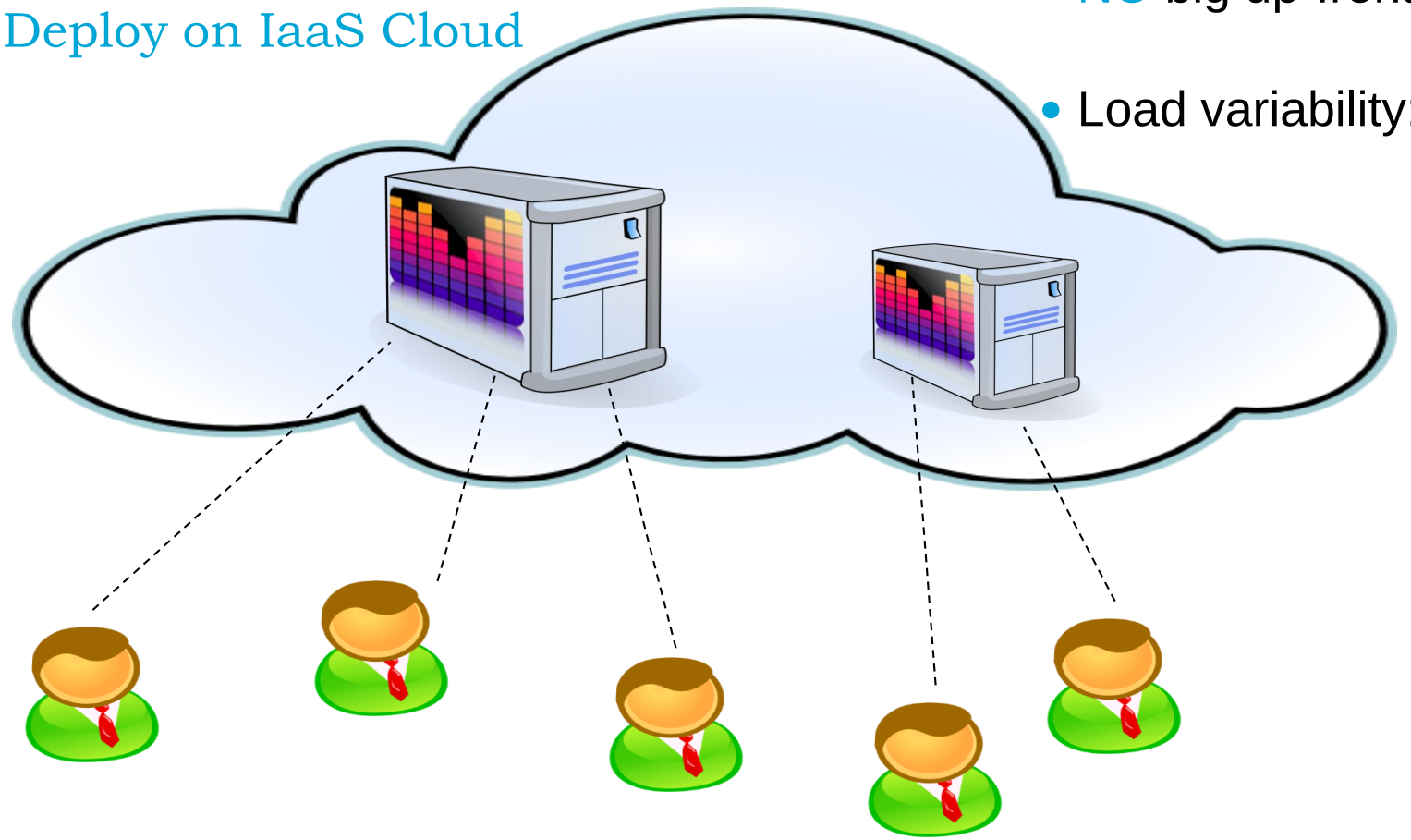
Buy then Maintain

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



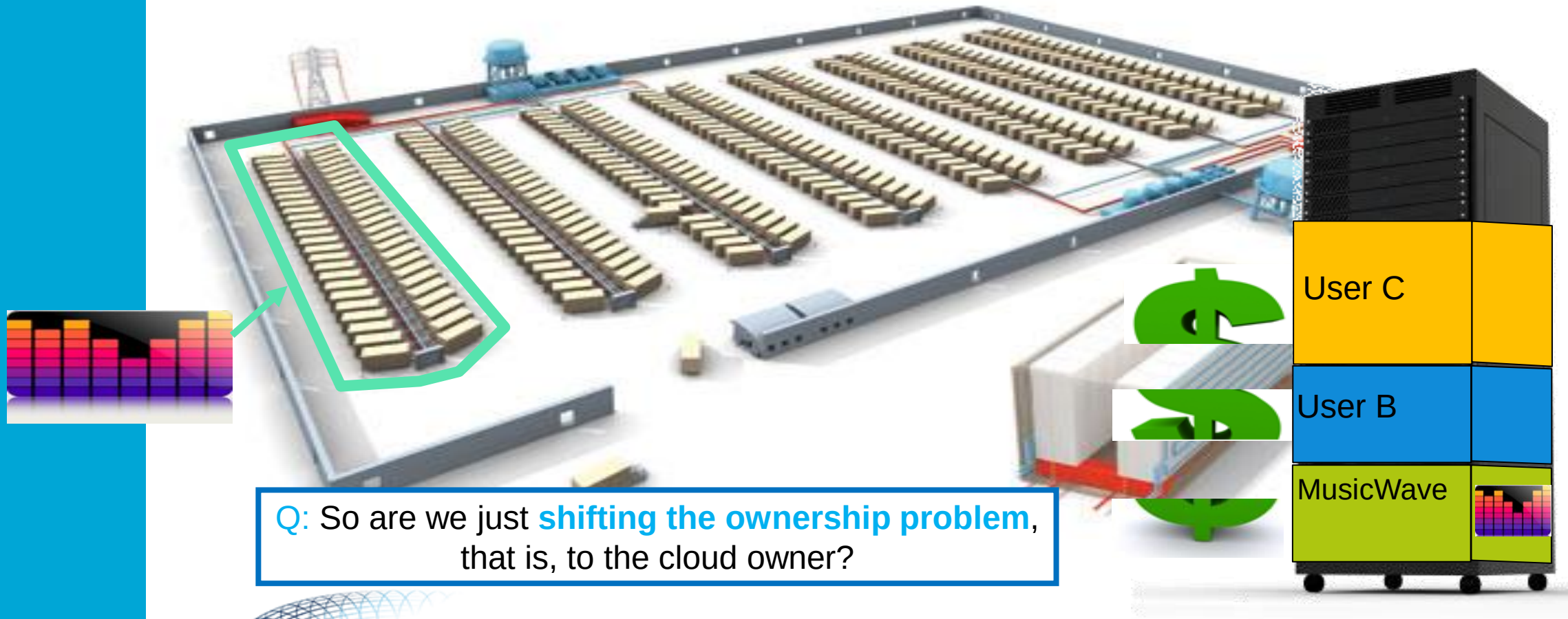
Solution #2

Deploy on IaaS Cloud



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

Inside a Cloud Datacenter: Infrastructure as a Service



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

Scalable High Performance Systems

Interactive Masterclass

5’ — Pitch on Scalable High Performance Systems

5’ — The Golden Age of Datacenters

20’ — A Delft Data Science View on Datacenters

- The core idea of datacenter computing
- The main enabling technologies for datacenter computing
- The main challenges and techniques

35’ — Delft Data Science Makes Datacenters Tick

- Addressing the Scheduling challenge
- Addressing the Ecosystem Navigation challenge
- Addressing the Big Cake challenge
- Addressing Jevon’s Effect in Data Science

10’ — Towards a KIVI Taskforce on Data Science as a Service

TU Delft

KIVI Engineering Society

DELFT DATA SCIENCE

The Pizza-Box Stack

- The 1U server

The Pizza-Box Stack

- The 1U server

The Pizza-Box Stack

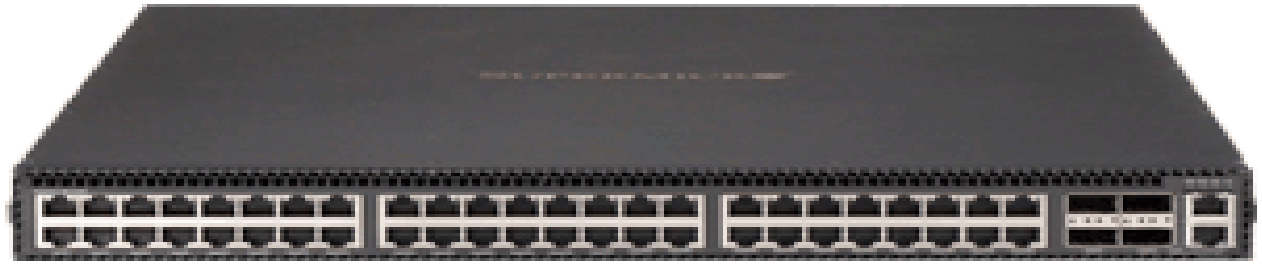
- The 1U server

The Pizza-Box Stack

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

The Data Center Network

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



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

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

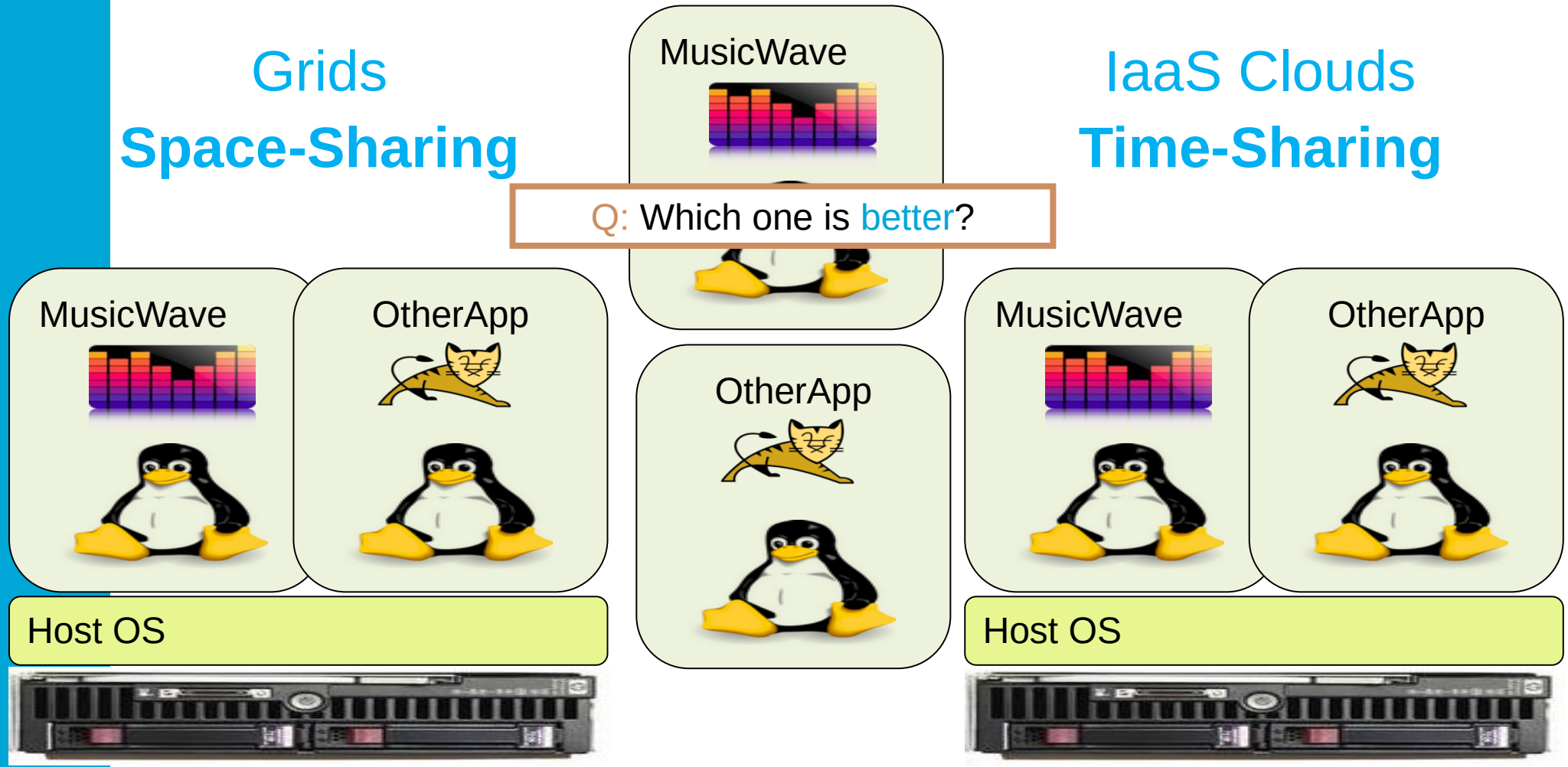

Servers + Server Racks +
Intra-Rack Network + Inter-Rack Network
An Entire Floor in a
Google Datacenter
Google | google.com/datacenters

Resource Sharing Models

Grids
Space-Sharing

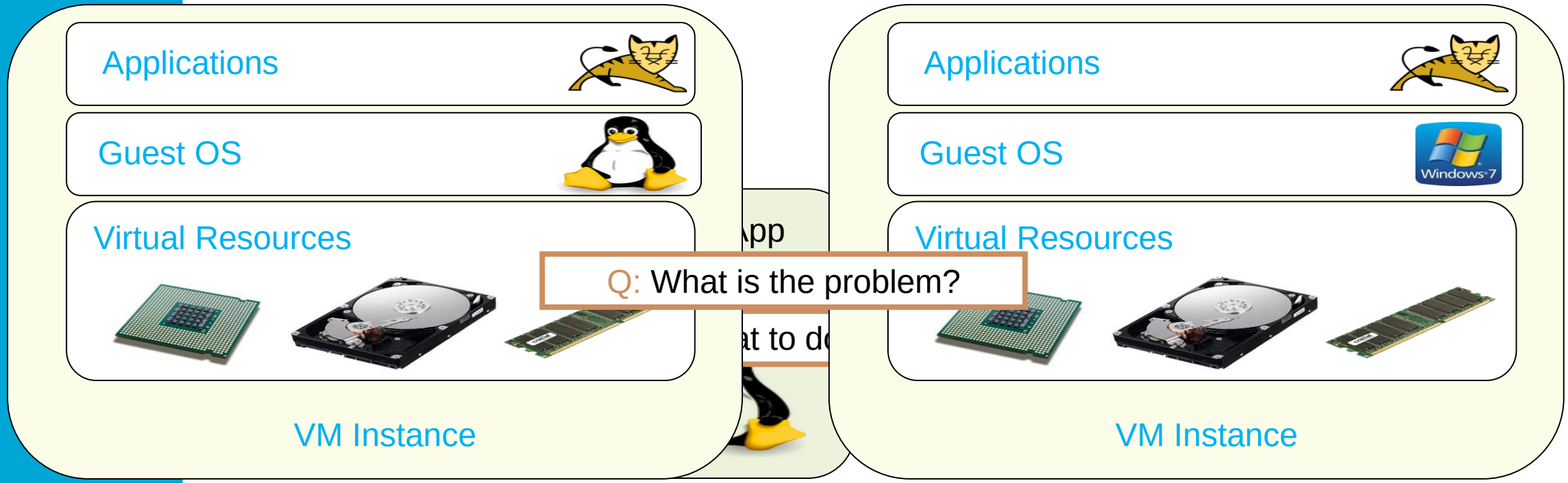
IaaS Clouds
Time-Sharing

Q: Which one is better?



(Source: A. Iosup, Course in 2012, 2014)

Virtualization



Virtualization

Host OS

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

Scalable High Performance Systems

Interactive Masterclass

5’ — Pitch on Scalable High Performance Systems

5’ — The Golden Age of Datacenters

20’ — A Delft Data Science View on Datacenters

- The core idea of datacenter computing
- The main enabling technologies for datacenter computing
- The main challenges and techniques

35’ — Delft Data Science Makes Datacenters Tick

- Addressing the Scheduling challenge
- Addressing the Ecosystem Navigation challenge
- Addressing the Big Cake challenge
- Addressing Jevon’s Effect in Data Science

10’ — Towards a KIVI Taskforce on Data Science as a Service

TU Delft

KIVI Engineering Society

DELFT DATA SCIENCE

The Scheduling Challenge

amazon web services™ Windows Azure rackspace

Cloud operator:

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

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

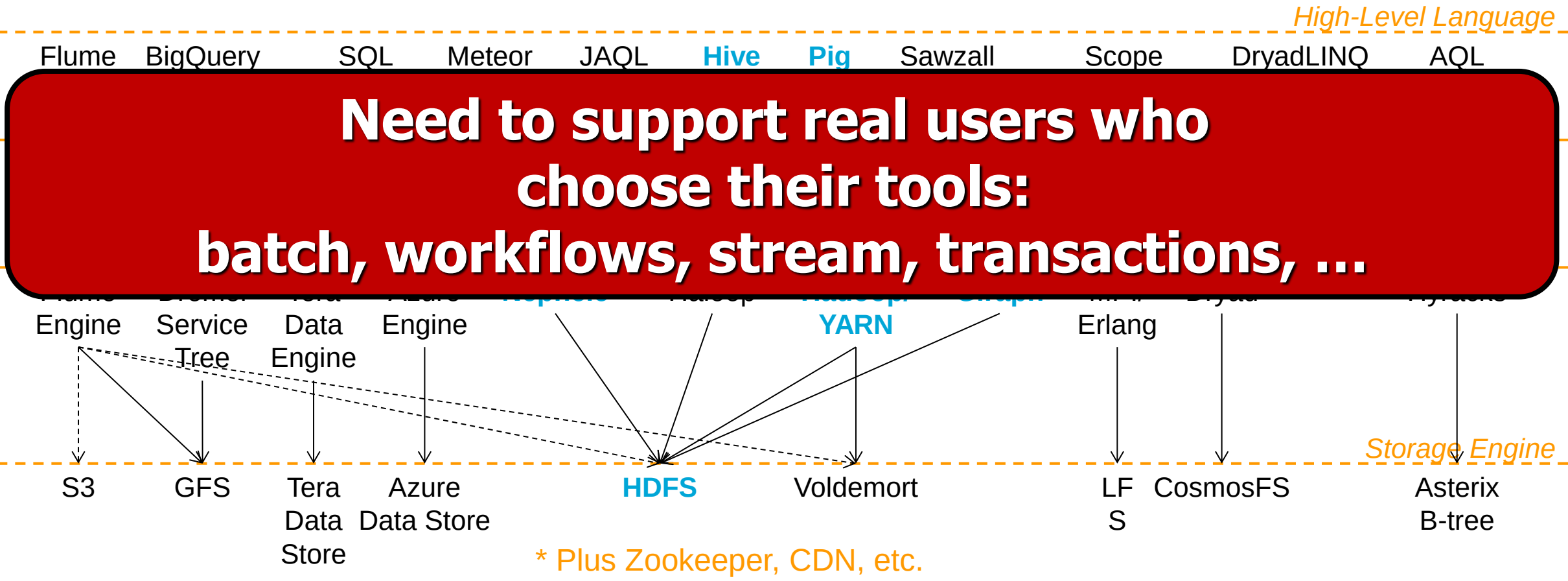
Cloud customer:

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

TU Delft

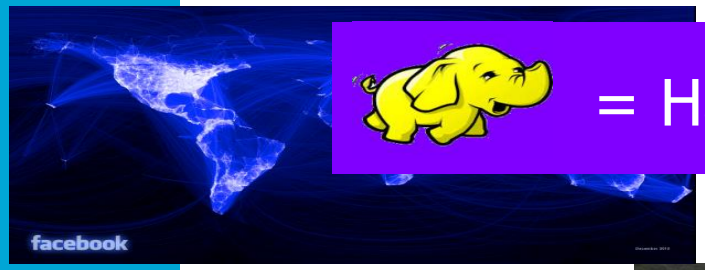
DELFT DATA SCIENCE

The Ecosystem Navigation Challenge

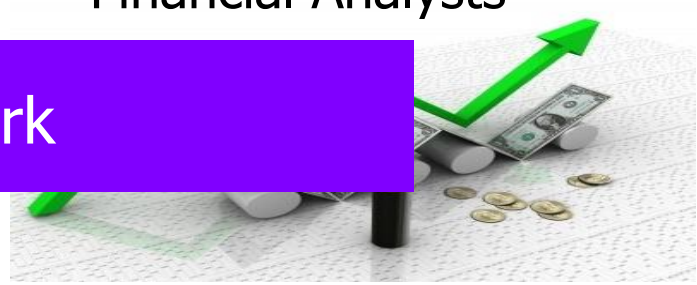


The “Big Cake” Challenge In the Datacenter

Online Social Networks



Financial Analysts

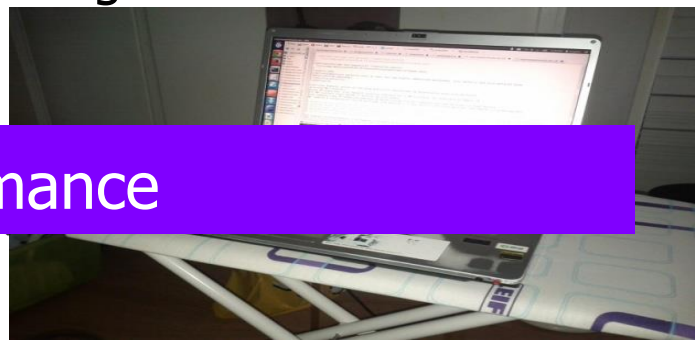


= Hadoop / MapReduce framework

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

Universe Explorers

Big Data Enthusiast



Multiple frameworks = Isolation, especially performance

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

Jevon's Effect: More Efficient, Less Capable

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

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

PSY Gangnam consumed ~500GWh

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

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

The New “Jevon’s Effect”: The “Data Deluge” vs Capability

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

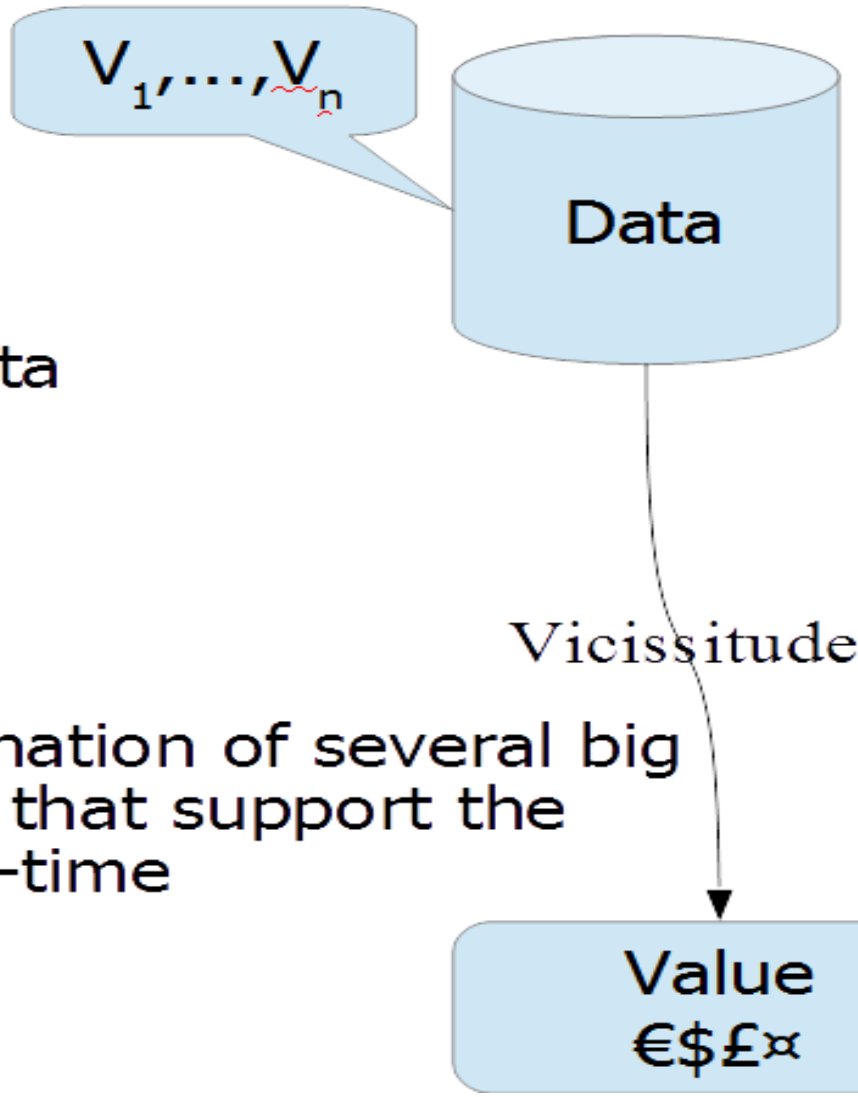
- Interacting
- Understanding
- Deciding
- Creating

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

* Other Vs possible: ours is “vicissitude”

Sources: IDC, EMC.

Vs of big data



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

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

Scalable High Performance Systems

Interactive Masterclass

5’ — Pitch on Scalable High Performance Systems

5’ — The Golden Age of Datacenters

20’ — A Delft Data Science View on Datacenters

- The core idea of datacenter computing
- The main enabling technologies for datacenter computing
- The main challenges and techniques

35’ — Delft Data Science Makes Datacenters Tick

- Addressing the Scheduling challenge
- Addressing the Ecosystem Navigation challenge
- Addressing the Big Cake challenge
- Addressing Jevon’s Effect in Data Science

10’ — Towards a KIVI Taskforce on Data Science as a Service

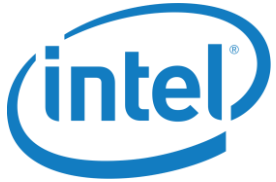
TU Delft

KIVI Engineering Society

DELFT DATA SCIENCE

Our Industry Collaborators

AZAVISTA



Scalable High Performance Systems

Interactive Masterclass

5’ — Pitch on Scalable High Performance Systems

5’ — The Golden Age of Datacenters

20’ — A Delft Data Science View on Datacenters

- The core idea of datacenter computing
- The main enabling technologies for datacenter computing
- The main challenges and techniques

35’ — Delft Data Science Makes Datacenters Tick

- Addressing the Scheduling challenge
- Addressing the Ecosystem Navigation challenge
- Addressing the Big Cake challenge
- Addressing Jevon’s Effect in Data Science

10’ — Towards a KIVI Taskforce on Data Science as a Service

TU Delft

KIVI Engineering Society

DELFT DATA SCIENCE

The Scheduling Challenge

amazon web services™ Windows Azure rackspace

Cloud operator:

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

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

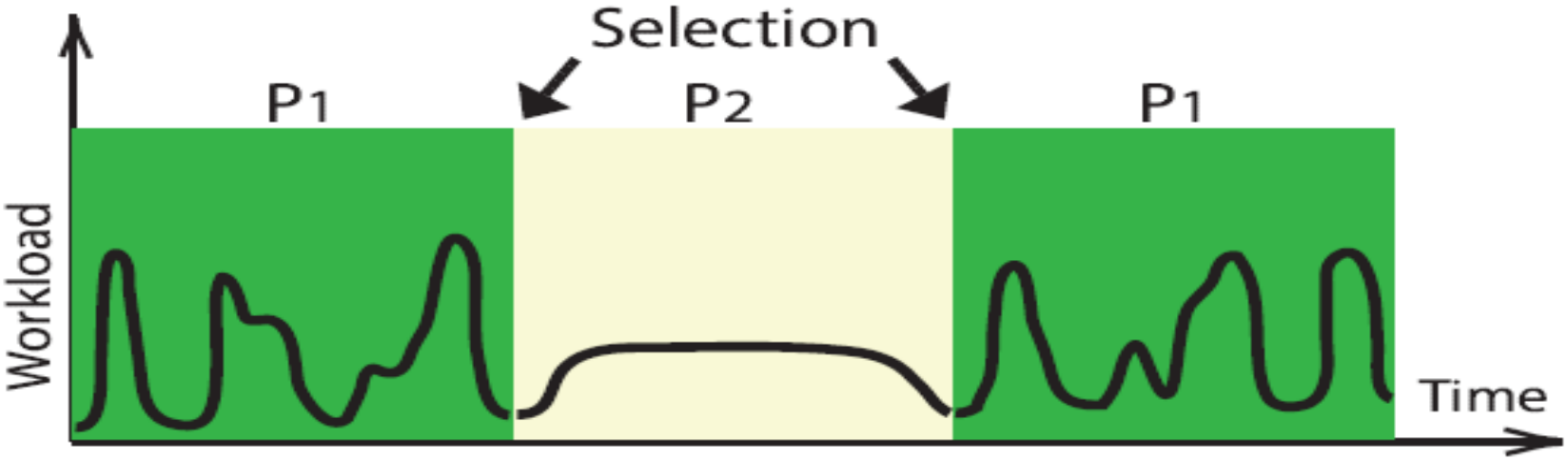
Cloud customer:

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

TU Delft

DELFT DATA SCIENCE

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?

Creation

Which policy to activate?

Selection

Application

Which resources? What to log?

Reflection

Which changes to the portfolio?

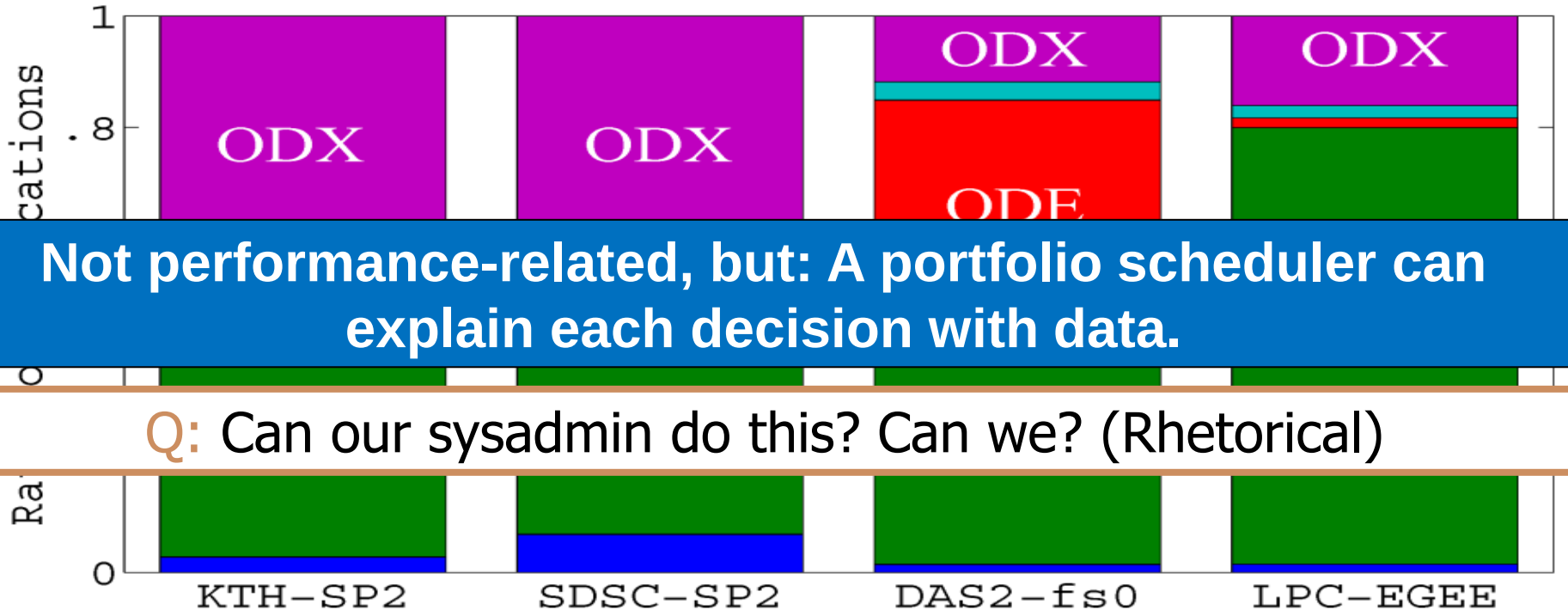
TU Delft

KIVI Engineering Society

bitbrains IT SERVICES

DELFT DATA SCIENCE

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

Scalable High Performance Systems

Interactive Masterclass

5’ — Pitch on Scalable High Performance Systems

5’ — The Golden Age of Datacenters

20’ — A Delft Data Science View on Datacenters

- The core idea of datacenter computing
- The main enabling technologies for datacenter computing
- The main challenges and techniques

35’ — Delft Data Science Makes Datacenters Tick

- Addressing the Scheduling challenge
- Addressing the Ecosystem Navigation challenge
- Addressing the Big Cake challenge
- Addressing Jevon’s Effect in Data Science

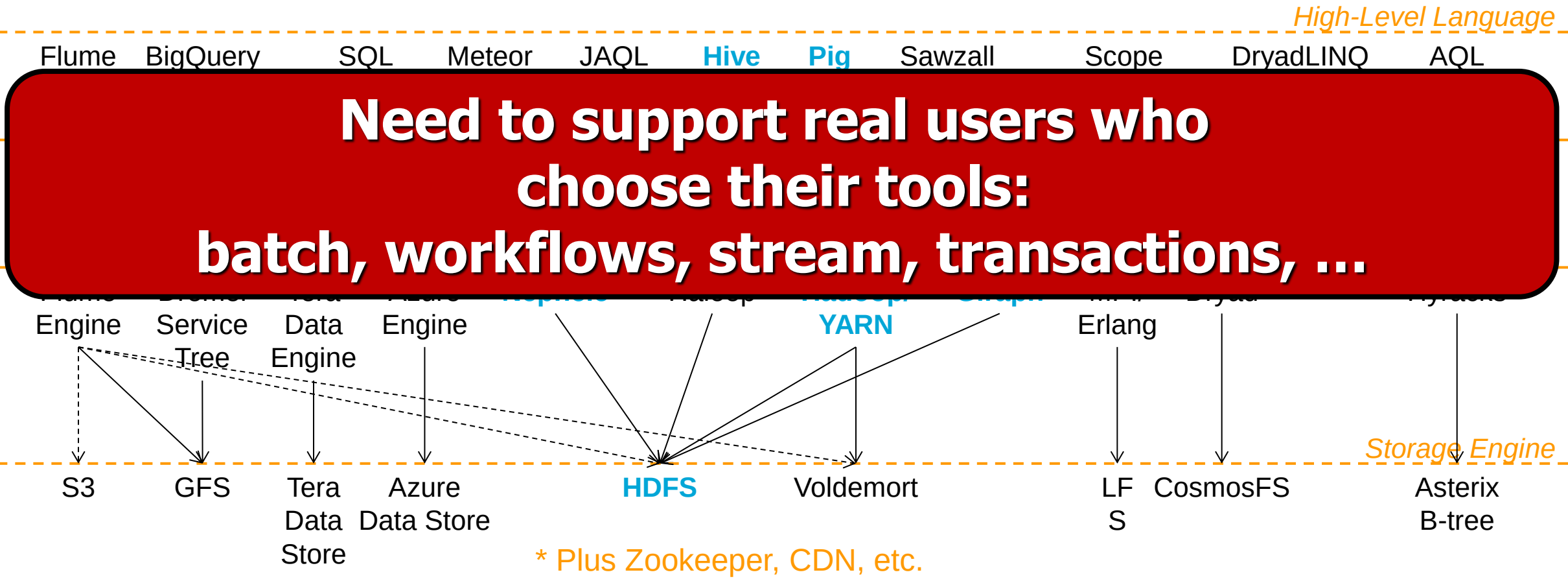
10’ — Towards a KIVI Taskforce on Data Science as a Service

TU Delft

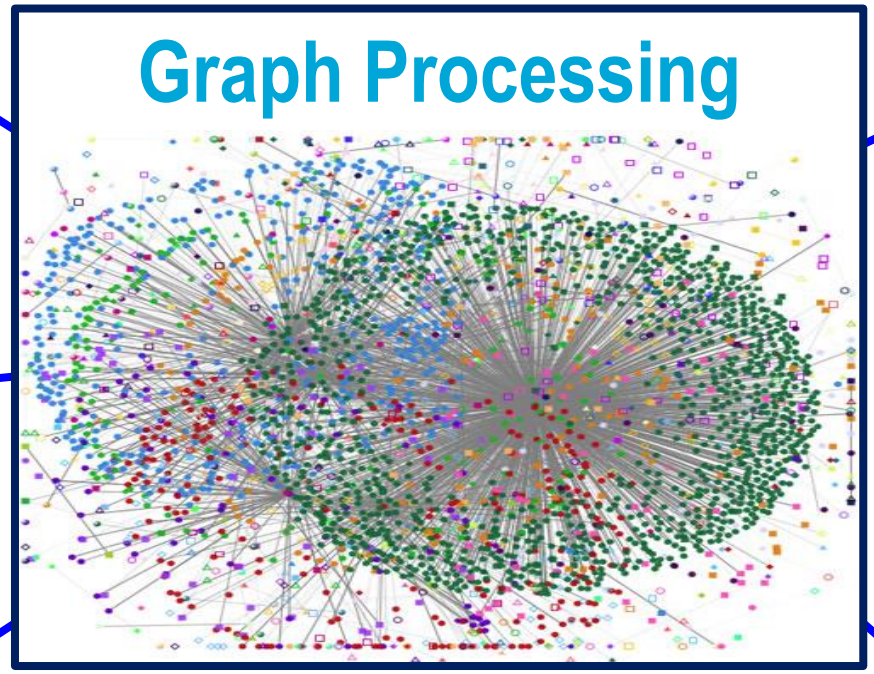
KIVI Engineering Society

DELFT DATA SCIENCE

The Ecosystem Navigation Challenge



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



LinkedIn

amazon.com

Spotify

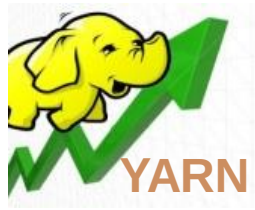
KIVI Engineering Society

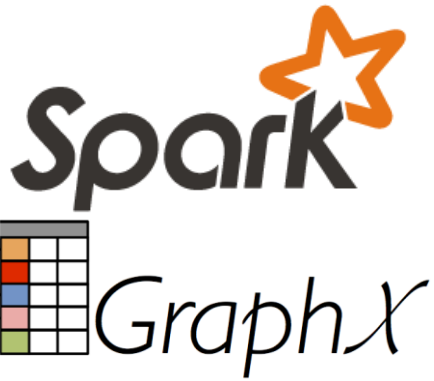
STEAM

DELFT DATA SCIENCE

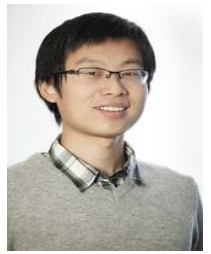
Platform Diversity

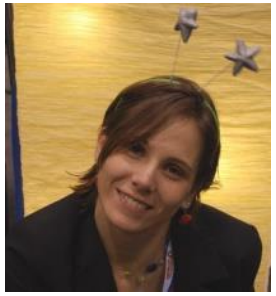
Oracle Labs
PGX

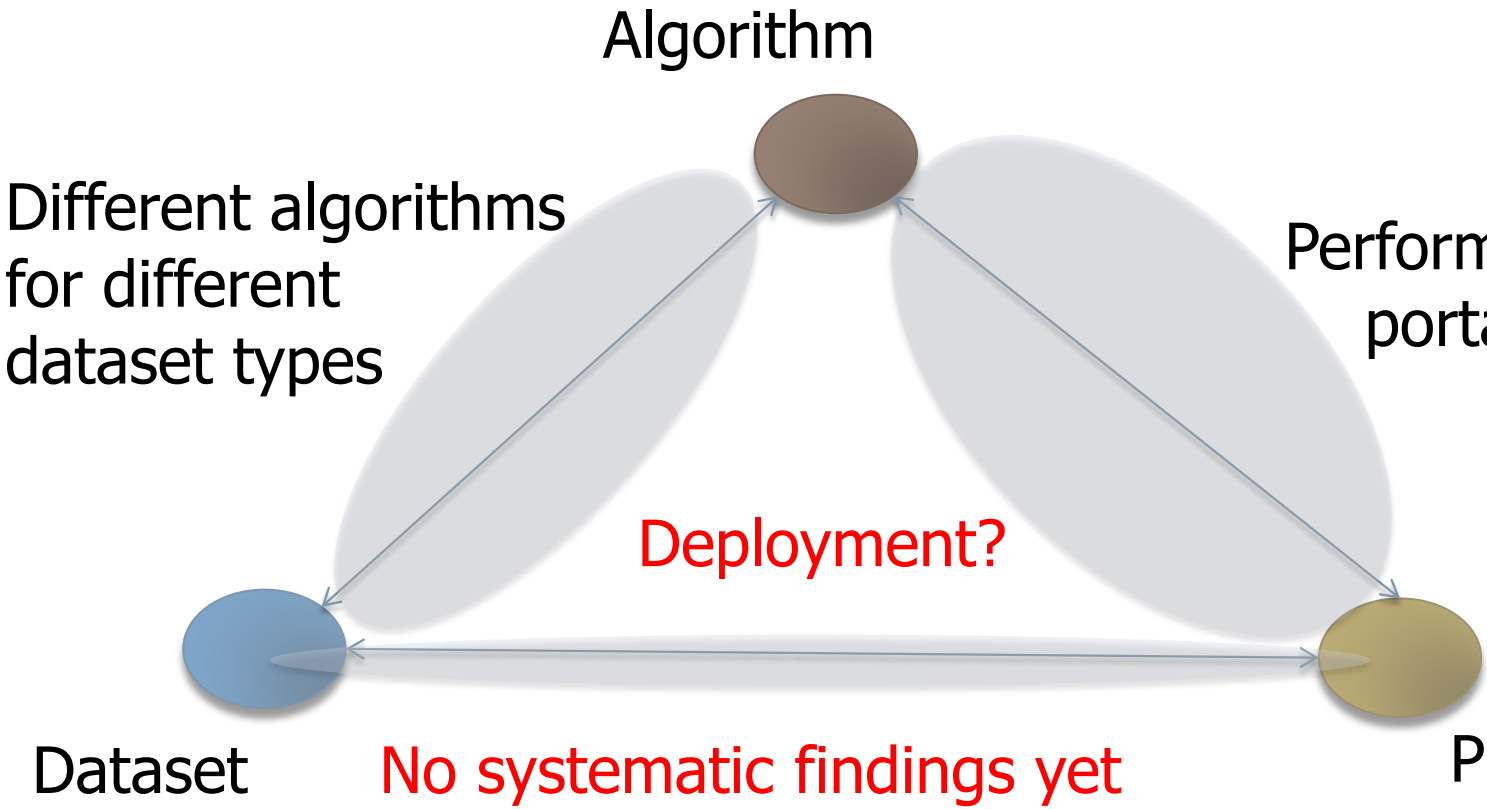




Ecosystem Navigation = Understanding the PAD Triangle







TU Delft

KIVI Engineering Society

DELFT DATA SCIENCE

Graphalytics: The first comprehensive benchmark for big data graph processing

A PAD triangle explorer for Graph Processing

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

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

TU Delft

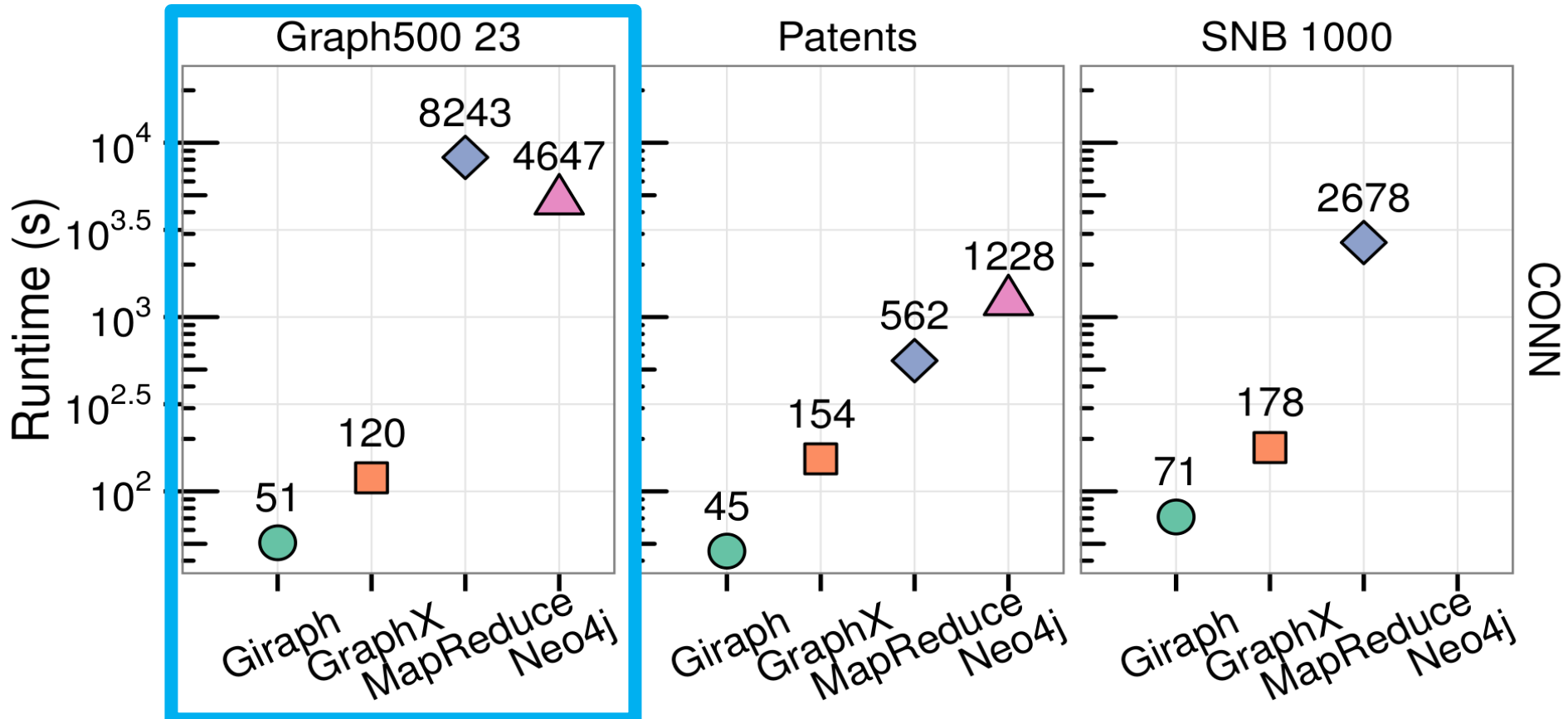
KIVI Engineering Society

ORACLE

LDBC The graph & RDF benchmark reference

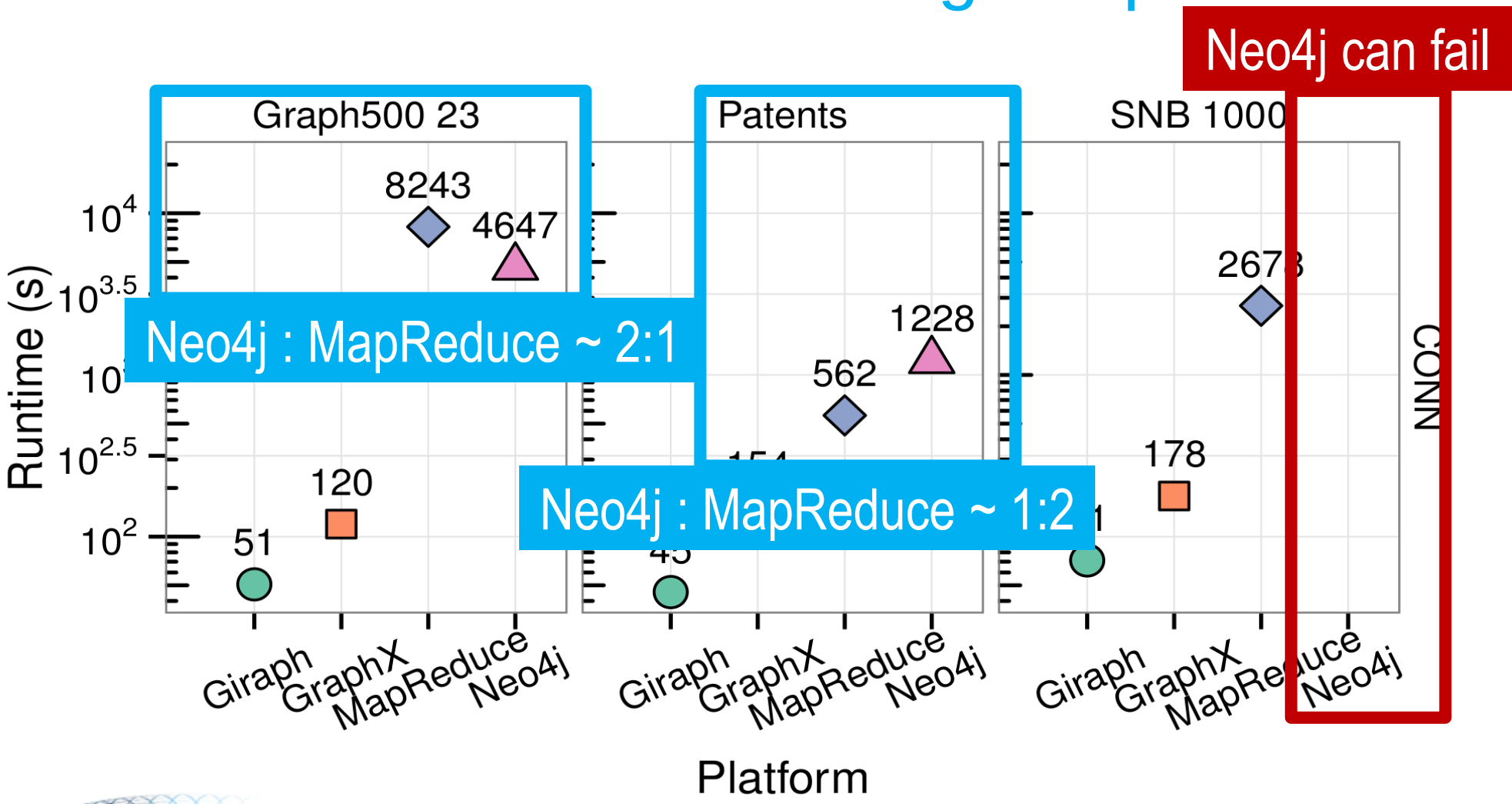
DELFT DATA SCIENCE

Runtime: the Platform has large impact



2 orders of magnitude difference due to platform

Runtime: The Dataset has large impact

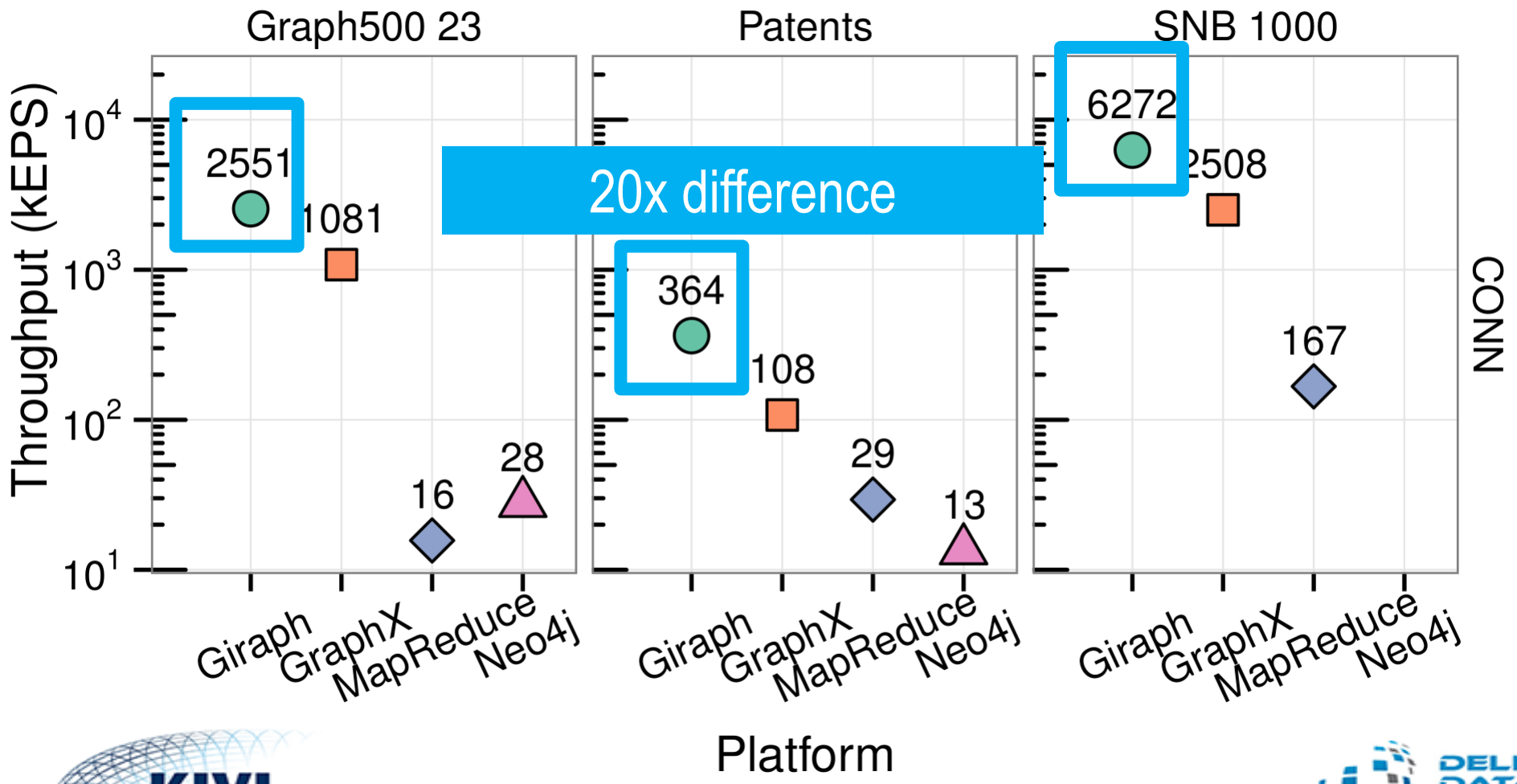


TU Delft

KIVI Engineering Society

DELFT DATA SCIENCE

Throughput: The Dataset structure matters!



Scalable High Performance Systems

Interactive Masterclass

5’ — Pitch on Scalable High Performance Systems

5’ — The Golden Age of Datacenters

20’ — A Delft Data Science View on Datacenters

- The core idea of datacenter computing
- The main enabling technologies for datacenter computing
- The main challenges and techniques

35’ — Delft Data Science Makes Datacenters Tick

- Addressing the Scheduling challenge
- Addressing the Ecosystem Navigation challenge
- Addressing the Big Cake challenge
- Addressing Jevon’s Effect in Data Science

10’ — Towards a KIVI Taskforce on Data Science as a Service

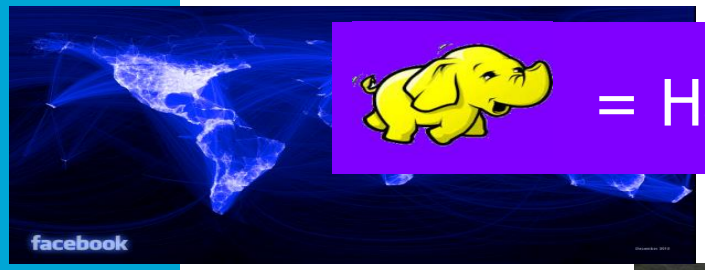
TU Delft

KIVI Engineering Society

DELFT DATA SCIENCE

The “Big Cake” Challenge In the Datacenter

Online Social Networks



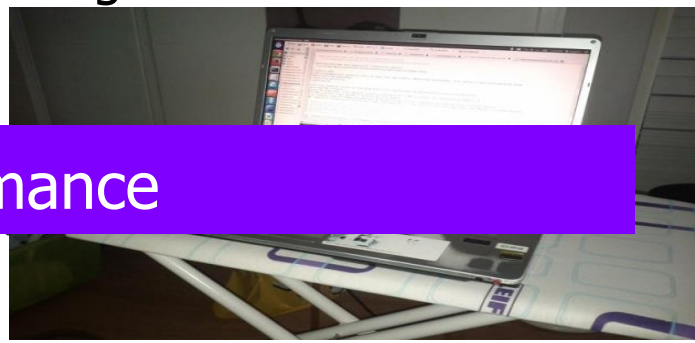
Financial Analysts

= Hadoop / MapReduce framework

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

Universe Explorers

Big Data Enthusiast



Multiple frameworks = Isolation, especially performance

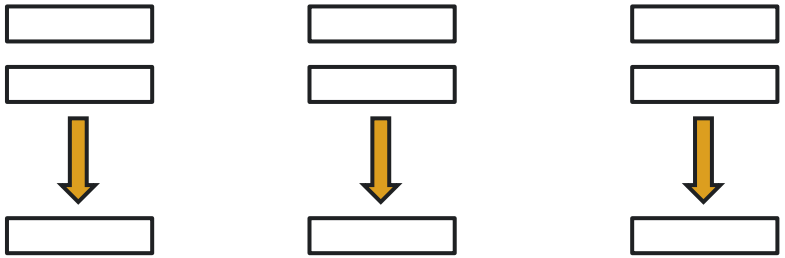
TU Delft

KIVI Engineering Society

DELFT DATA SCIENCE

Dynamic Big Data Processing

Fawkes = Elastic MapReduce

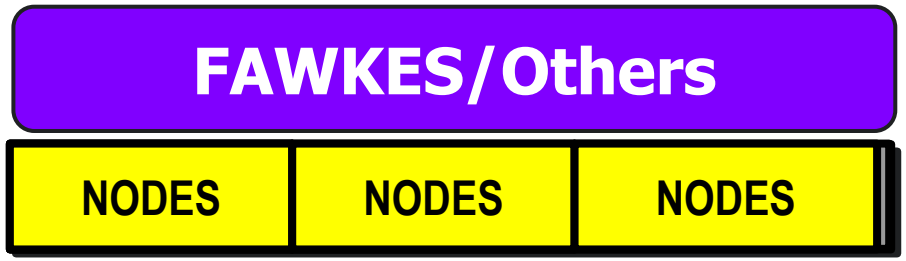


Job submissions

Frameworks

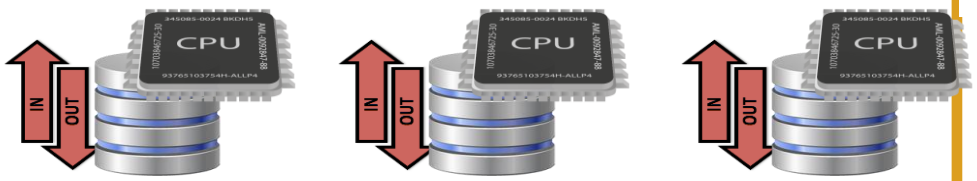
Resource manager

Infrastructure



Elasticity for MapReduce Frameworrks

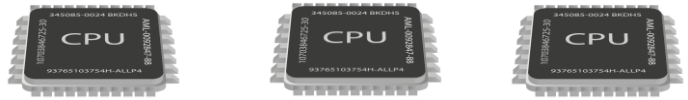
Core nodes



INPUT/OUTPUT DATA

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

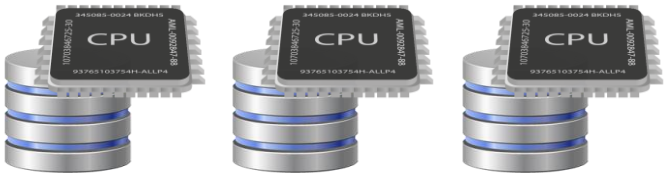
Transient nodes (TR)



NO DATA

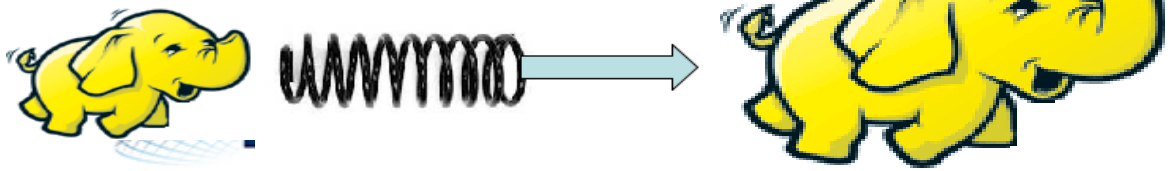
- No local storage
- R/W from/to core nodes
- **Instant removal**

Trans-core nodes (TC)



OUTPUT DATA

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



TU Delft

Fawkes in a Nutshell [1/2]

Because workloads may be time-varying:

- Poor resource utilization
- Imbalanced service levels

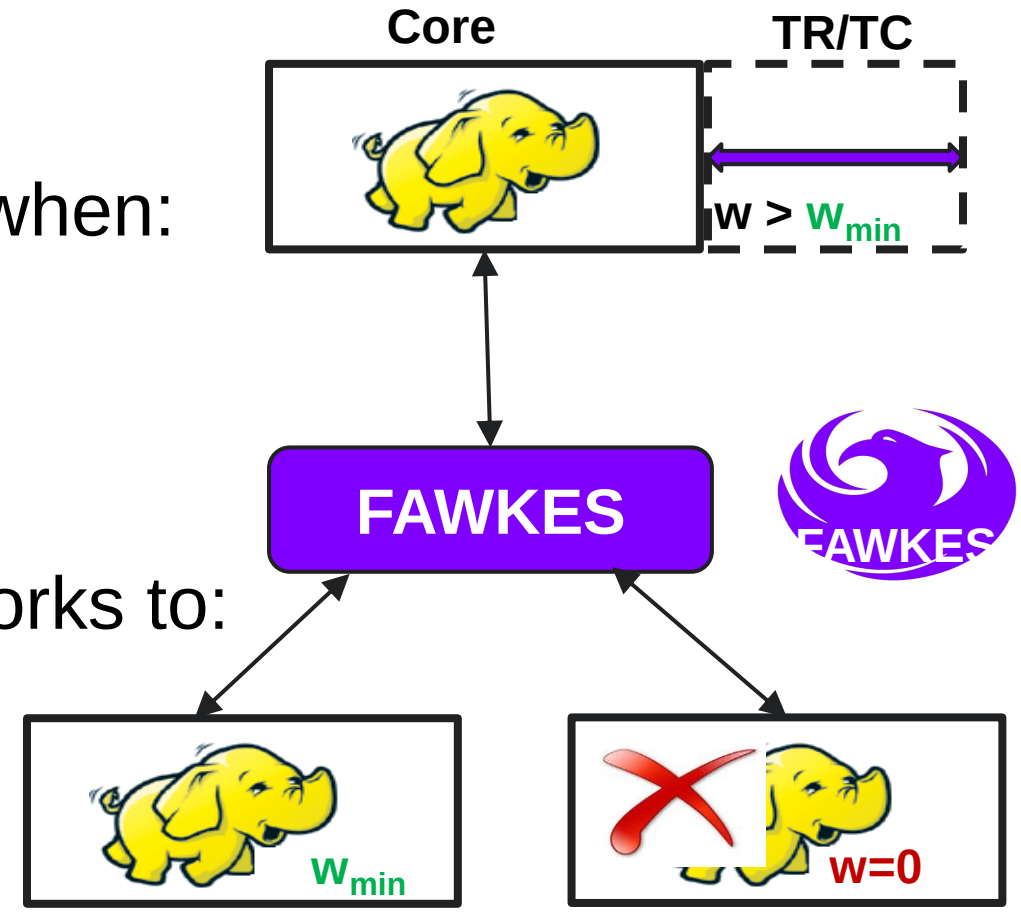
1. Fair framework size:

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

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

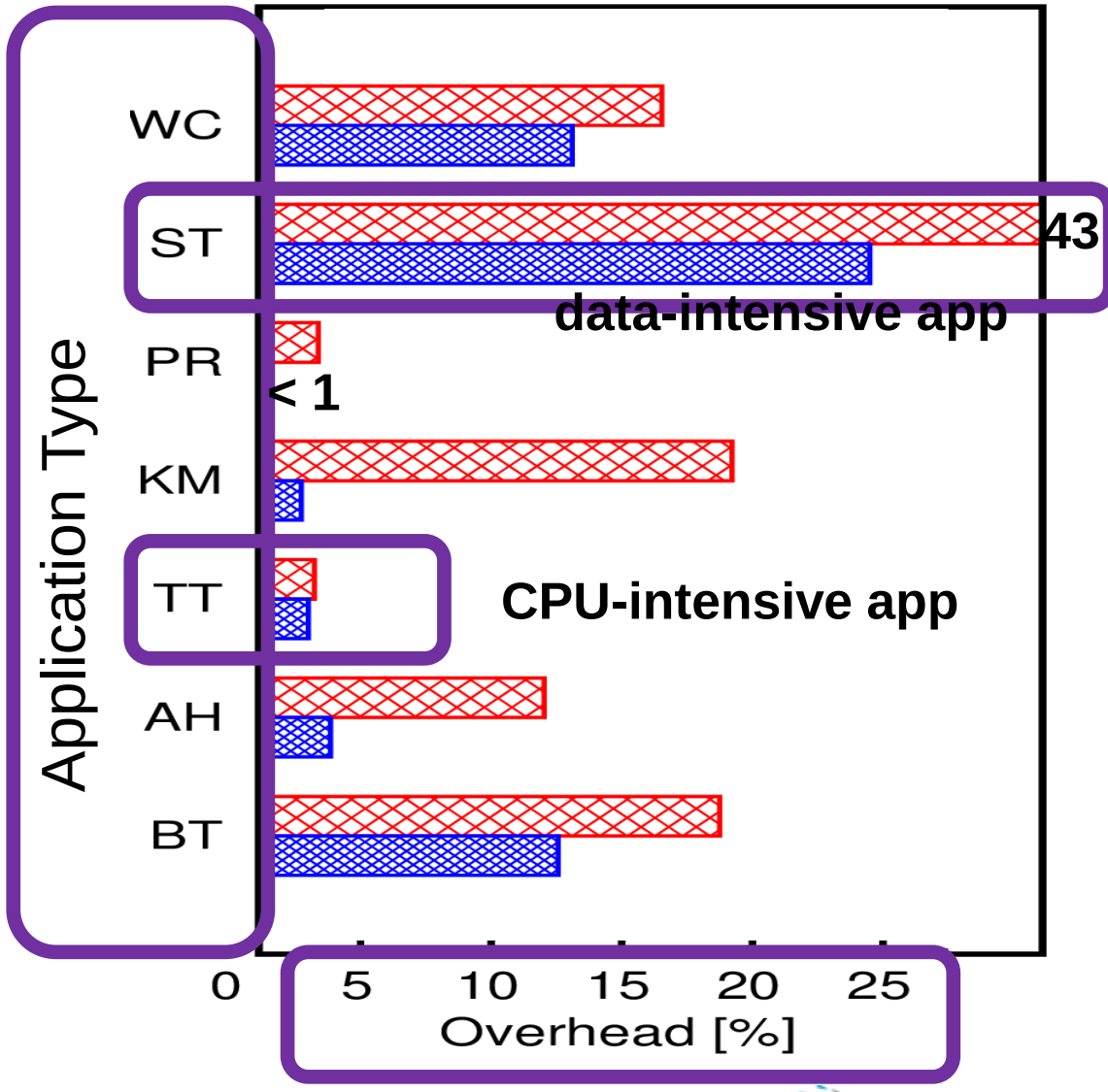
10 core +10xTR
10 core +10xTC
vs.
20 core nodes (baseline)

TR - good for compute-intensive workloads.

TC - needed for disk-intensive workloads.

Dynamic MapReduce:
< 25% overhead

Fawkes also reduces imbalance



TU Delft

KIVI Engineering Society

DELFT DATA SCIENCE

Scalable High Performance Systems

Interactive Masterclass

5’ — Pitch on Scalable High Performance Systems

5’ — The Golden Age of Datacenters

20’ — A Delft Data Science View on Datacenters

- The core idea of datacenter computing
- The main enabling technologies for datacenter computing
- The main challenges and techniques

35’ — Delft Data Science Makes Datacenters Tick

- Addressing the Scheduling challenge
- Addressing the Ecosystem Navigation challenge
- Addressing the Big Cake challenge
- Addressing Jevon’s Effect in Data Science

10’ — Towards a KIVI Taskforce on Data Science as a Service

TU Delft

KIVI Engineering Society

DELFT DATA SCIENCE

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

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

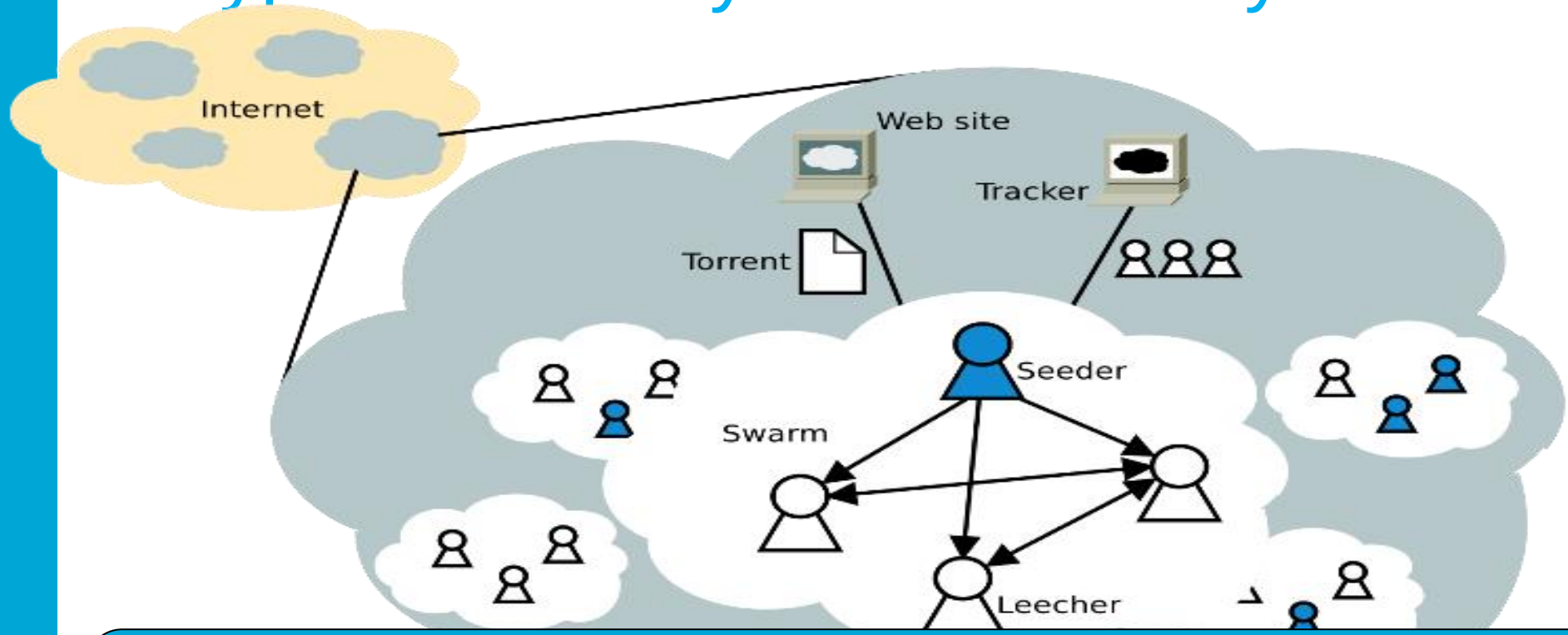
Vicissitude of Big Data = dynamic mix of big data issues (Vs) that lead in big data systems to different bottlenecks over time

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

- Interacting
- Understanding
- Deciding
- Creating

Sources: IDC, EMC.

Observing BitTorrent: Managing A Typical Globally Distributed System

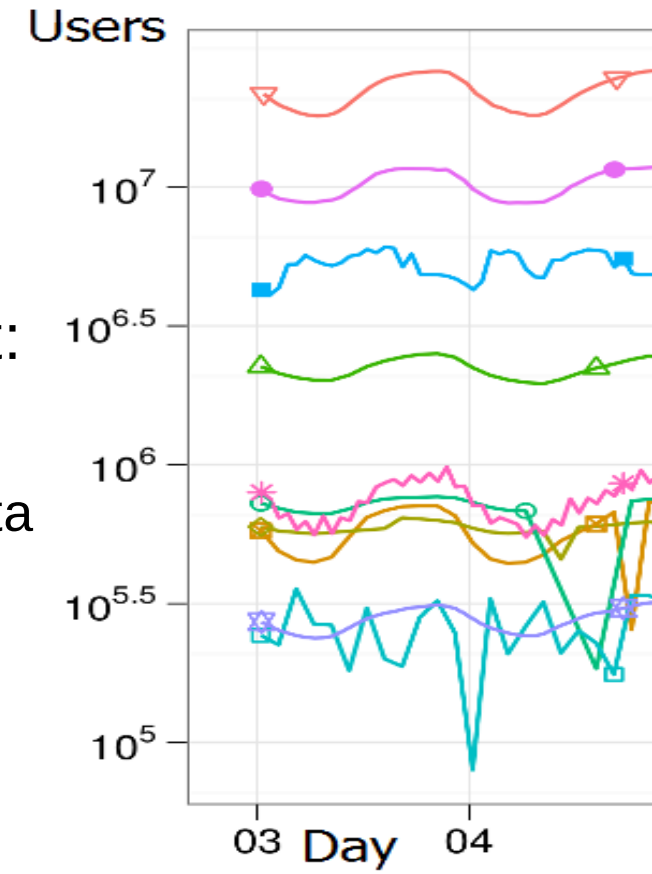


Most used protocol on Internet, by upload volume [1]
One third (US) to half (EU) of residential upload
Over 100 million users [2]

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

BTWorld: a Typical Big Data Project (and Our Use Case)

- 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?

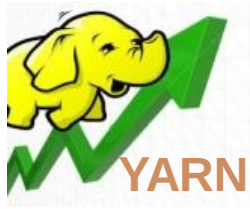


TU Delft

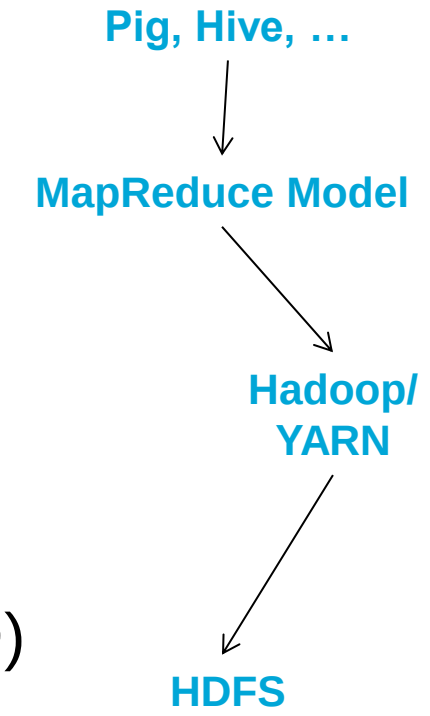
KIVI Engineering Society

DELFT DATA SCIENCE

The MapReduce ecosystem (a big problem in big data)

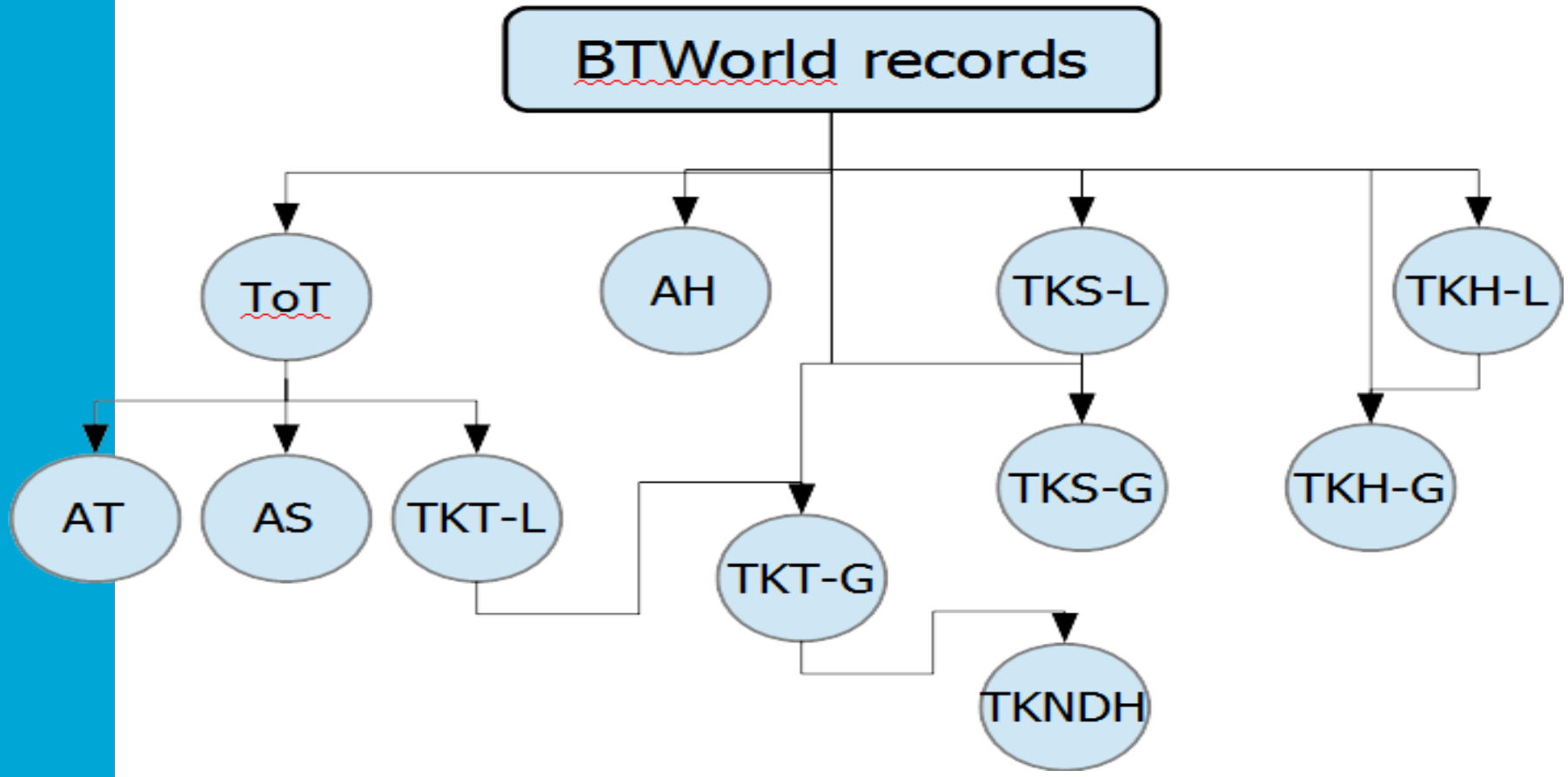


- Widely used in industry and academia
  - Similar to other big data stacks
- Complex software to tune
  - 100s of parameters
  - Non-linear effects common
- Lots of issues cause crashes [1]
- Focus on Small and Medium Enterprises (60% GPD)
  - No resources or even competence to fix issues
  - Difficult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



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

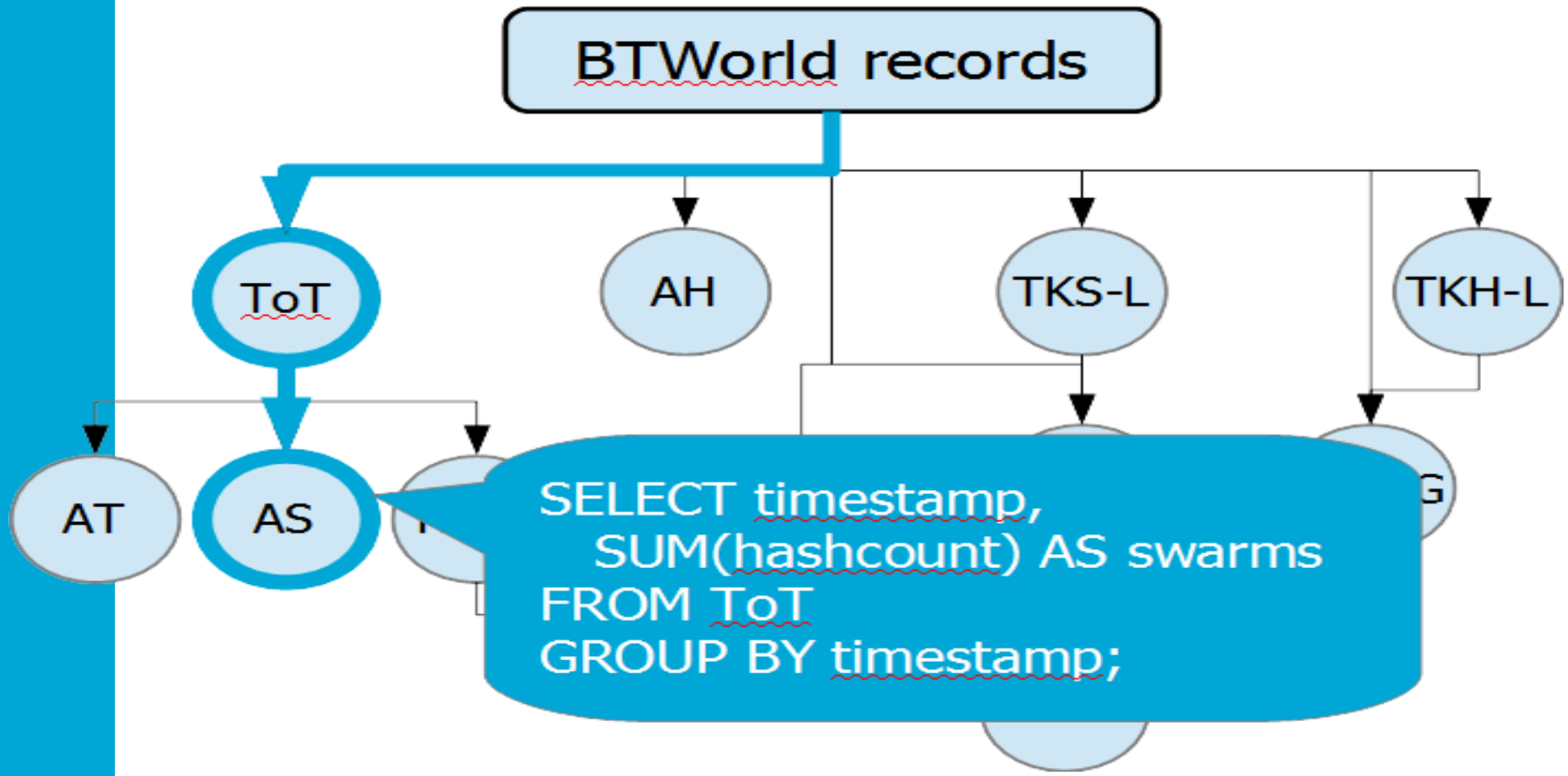
The Abstract BTWorld Workflow



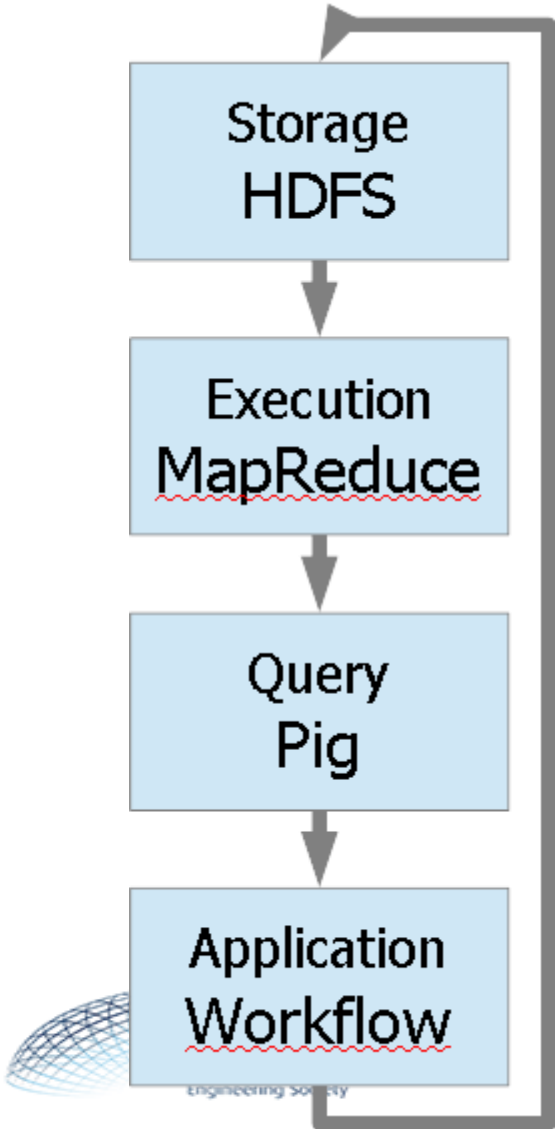
TKH-L Query

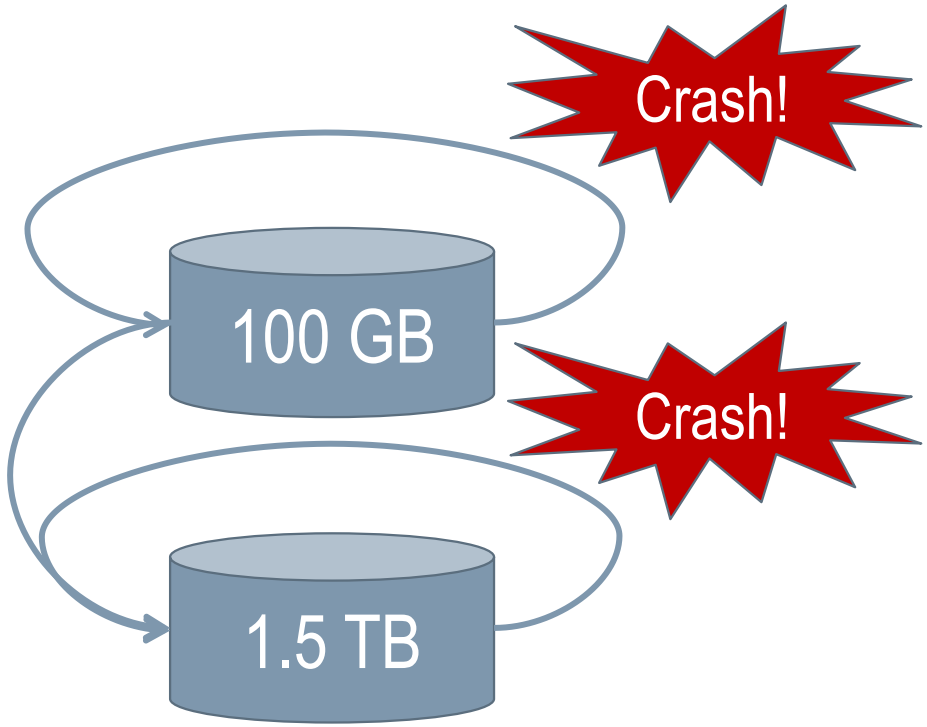
Data path

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

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!

spec Research

TU Delft

DELFT DATA SCIENCE

Scalable High Performance Systems

Interactive Masterclass

5’ — Pitch on Scalable High Performance Systems

5’ — The Golden Age of Datacenters

20’ — A Delft Data Science View on Datacenters

- The core idea of datacenter computing
- The main enabling technologies for datacenter computing
- The main challenges and techniques

35’ — Delft Data Science Makes Datacenters Tick

- Addressing the Scheduling challenge
- Addressing the Ecosystem Navigation challenge
- Addressing the Big Cake challenge
- Addressing Jevon’s Effect in Data Science

10’ — Towards a KIVI Taskforce on Data Science as a Service

TU Delft

KIVI Engineering Society

DELFT DATA SCIENCE

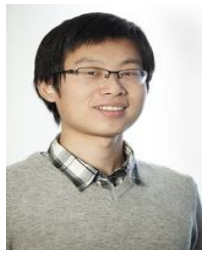
Take-Home Message

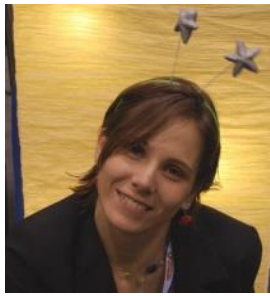
The Golden Age of datacenters

Cloud computing + Big Data

Important New Challenges

1. The scheduling challenge
2. The ecosystem navigation challenge
3. The big cake challenge
4. Jevon's Effect for Big Data





TU Delft

KIVI Engineering Society

DELFT DATA SCIENCE

Research Agenda for Datacenter-based Data Science

1. “Data Science as a Service” as functional goal.
2. Compute- & data-intensive models can coexist in the datacenter.
3. Non-functional targets: high performance and availability, elasticity, etc.
4. Fundamental models of data science platforms.
5. Fundamental knowledge on Platform-Algorithm-Data interaction.
6. New generation of resource management techniques, including scheduling.
7. Benchmarking data science services.

TU Delft

spec Research

DELFT DATA SCIENCE

Next? A New KIVI Taskforce on Data Science as a Service

Identify industry needs in the Netherlands

- Stakeholders: datacenter operators, ICT designers, ICT analysts, ICT researchers, governance, ICT media

Establish a joint research agenda, between fundamental and applied research

- Groundbreaking ideas for important challenges
- Prototypes and Proof-of-Concepts, not only ideas

Build a solid, pragmatic collaboration

- Relevant recommendations for relevant problems
- Embedding of human resources, joint networking, etc.

TU Delft

KIVI Engineering Society

DELFT DATA SCIENCE

More?

to out-compute is to out-compete—Collaborate with Delft Data Science on datacenter infrastructure

- Work together on complex engineering problems
- Two-way transfer of knowledge and expertise
- With impact on society, industry, academia, and governance

Attend ICT with Industry

- “direct and rapid interaction between the ICT researchers and Industrial partners”
- Dutch doctoral schools in ICT
- Co-organized with NWO and STW
- 7—11 December 2015

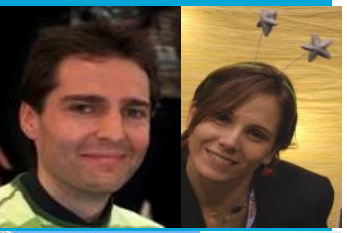
TU Delft

KIVI Engineering Society

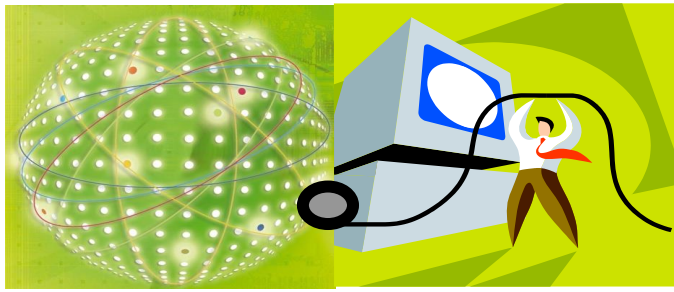
DELFT DATA SCIENCE

Contact Me or Our Team

Staff members



Collaboration or discussion about:



- Leveraging open-source cloud computing and big data systems in your organization
- Introducing MapReduce and graph-processing, and distributed computing systems in your organization
- Optimizing your high performance and high throughput clusters

A.Iosup@tudelft.nl
+31-15-2784433
@Alosup
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

TU Delft

KIVI Engineering Society

DELFT DATA SCIENCE

Recommended Reading

Elastic Big Data and Computing

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

Time-Based Analytics

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

Graph Processing / Benchmarking

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

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!