WILL IT RAIN TODAY?

UNDERSTANDING THE WEATHER OF COMPUTING CLOUDS, BEFORE IT HAPPENS

@Large Research
Massivizing Computer Systems

http://atlarge.science

Sponsored by:

NWO

Many thanks to our collaborators.
Many thanks to our international working groups: spec Research, LDBC The graph & RDF benchmark reference

VU VRIJE UNIVERSITEIT AMSTERDAM

bit.ly/VUCloudPerf20

@AIosup

Prof. dr. ir. Alexandru Iosup

VU AMSTERDAM < SCHIPHOL < THE NETHERLANDS < EUROPE

Amsterdam
founded 10th century
pop: 850,000

VU
founded 1880
pop: 23,500

The Netherlands

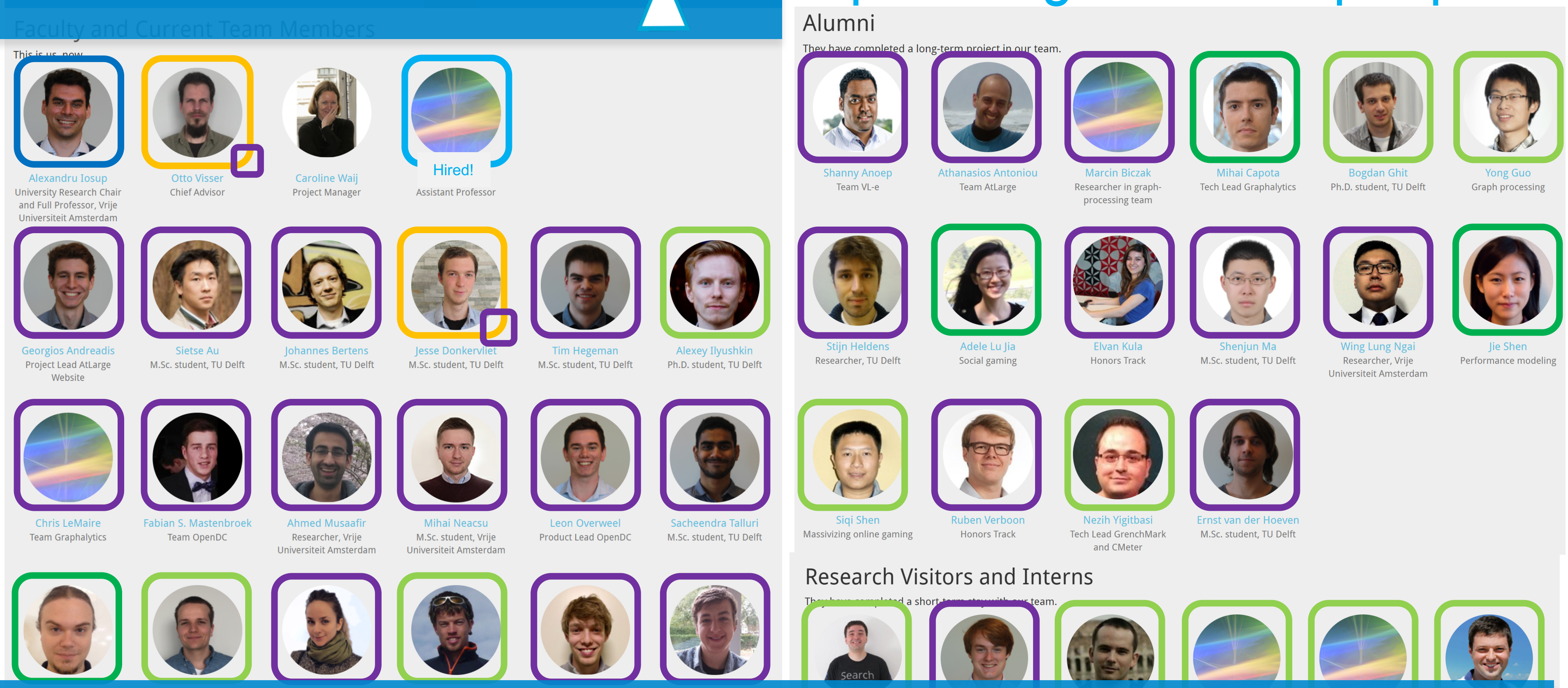
pop: 16.5 M

Europe

ATLARGE RESEARCH, OUR TEAM

http://atlarge.science/people.html

- Professor
- Assistant Prof.
- Teacher
- Post-doc
- Ph.D. student
- Scientist



WE ARE A FRIENDLY, DIVERSE GROUP, OF DIFFERENT RACES AND ETHNICITIES, GENDERS AND SEXUAL PREFERENCES, VIEWS OF CULTURE, POLITICS, AND RELIGION. YOU ARE WELCOME TO JOIN!

WHO AM I?
PROF. DR. IR. ALEXANDRU IOSUP

- Education, my courses:
  > Systems Architecture (BSc)
  > Distributed Systems, Cloud Computing (MSc)
- Research, 15 years in DistribSys:
  > Massivizing Computer Systems

- About me:
  > Worked in 7 countries, NL since 2004
  > I like to help… I train people in need
  > VU University Research Chair + Group Chair
  > NL ICT Researcher of the Year
  > NL Higher-Education Teacher of the Year
  > NL Young Royal Academy of Arts & Sciences

VU

MASSIVIZING COMPUTER SYSTEMS: OUR MISSION

1. Improve the lives of millions through impactful research.

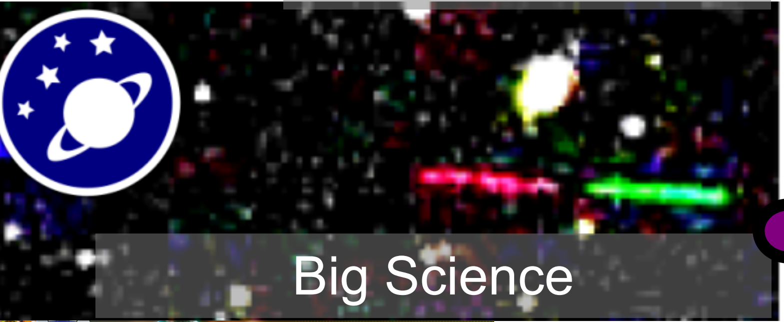
2. Educate the new generation of top-quality, socially responsible professionals.

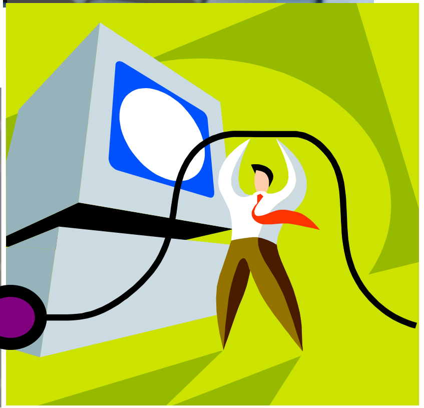
3. Make innovation available to society and industry.

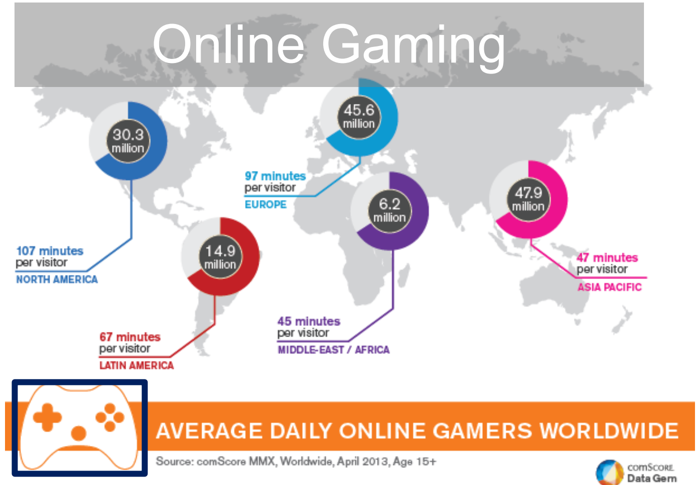
http://atlarge.science/about.html

THIS IS THE GOLDEN AGE OF CLOUD SYSTEMS AND ECOSYSTEMS

ONCE UPON A TIME … THE DAWN OF THE CLOUD







ABN·AMRO

Daily Life

ONCE UPON A TIME … THE DAWN OF THE CLOUD (1960s)

MIT Prof. Martin Greenberger:

" Computing services and establishments will begin to spread throughout every life-sector
[…] medical-information systems,
[…] centralized traffic control,
[…] catalogue shopping from […] home,
[…] integrated management-control systems for companies and factories
[…] "

M. Greenberger (**1964**) The Computers of Tomorrow The Atlantic Monthly. Vol. 213(5), pp. 63-67, May.

ONCE UPON A TIME ... THE DAWN OF THE CLOUD (1960s)

Data Processing ~ SaaS
IBM-Service Bureau Corp.(SBC)

Software/System Dev.~ PaaS
Computer Sciences Corp. (CSC)

Time Sharing ~ IaaS
IBM-SBC, Tymshare, GE Inf.Serv. (GEIS)

Facility management ~ IaaS
Electronic Data Systems (EDS)

Other Services
IBM

Source: J. R. Yost (2017) Making IT Work.

ONCE UPON A TIME … THE DAWN OF THE CLOUD (1970s)

Time Sharing ~ IaaS
GEIS invests in a large network
Tymshare invests in Tymnet
IBM invests in CALL 360

Technology **not ready**	Emergence of **the PC**
Economy: **recession**	Law, anti-trust: **DoJ vs. IBM**

Source: J. R. Yost (2017) Making IT Work.

THIS IS THE GOLDEN AGE OF CLOUD COMPUTING (2010S)

DIVERSE CLOUD SERVICES FOR ALL

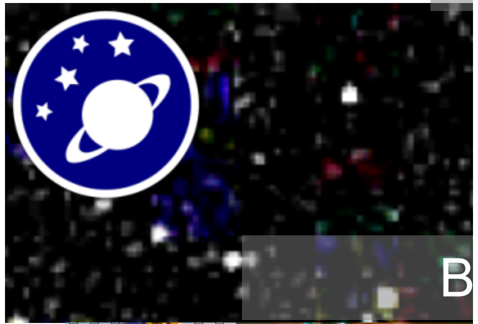
15-30% IT INDUSTRY

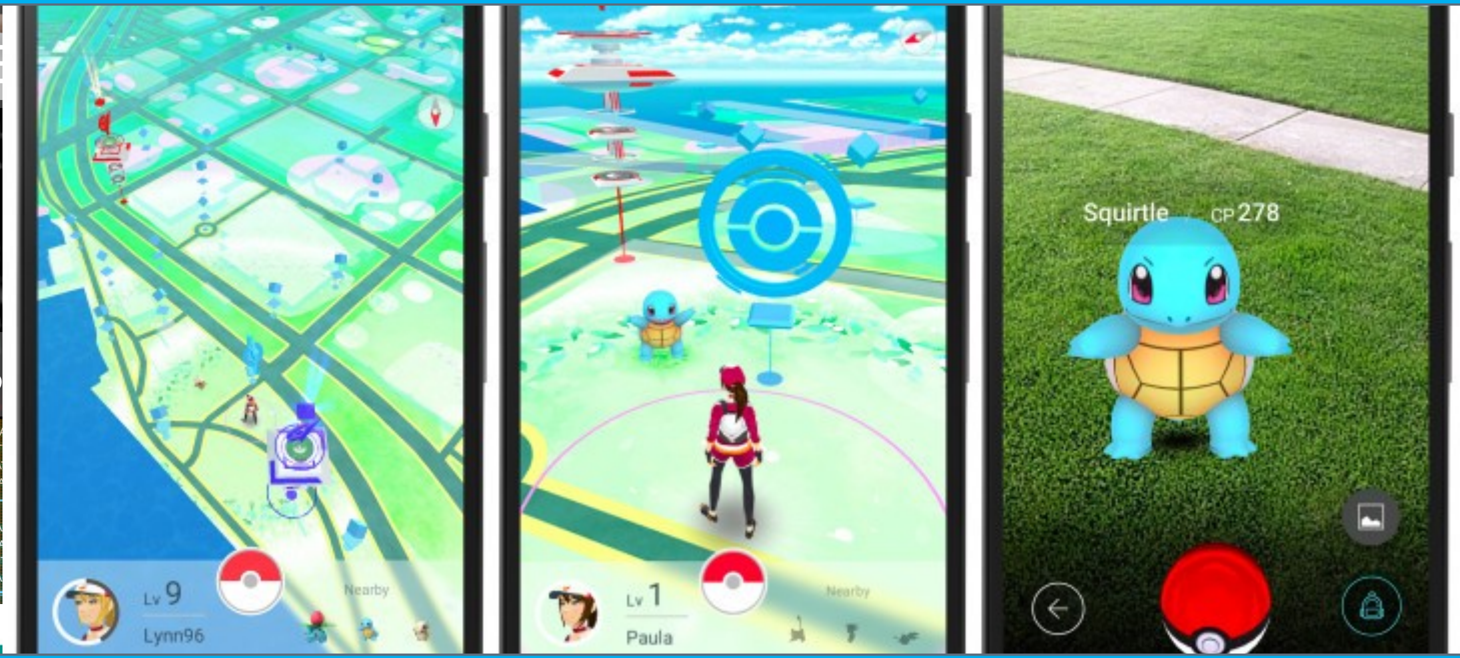
LEARNED OUR LESSON?

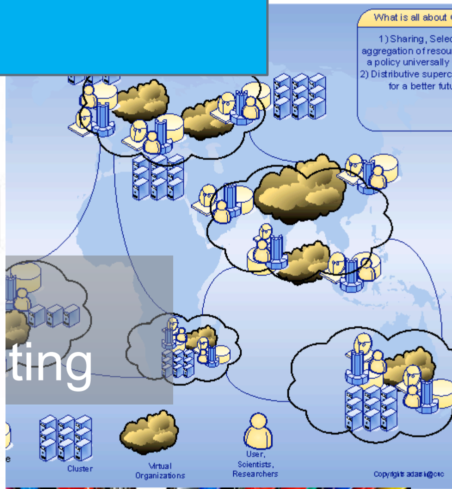
Iosup et al., Massivizing Computer Systems, ICDCS 2018. [Online]

THIS IS THE GOLDEN AGE OF CLOUD COMPUTING (2010S)

Do you recognize this App?



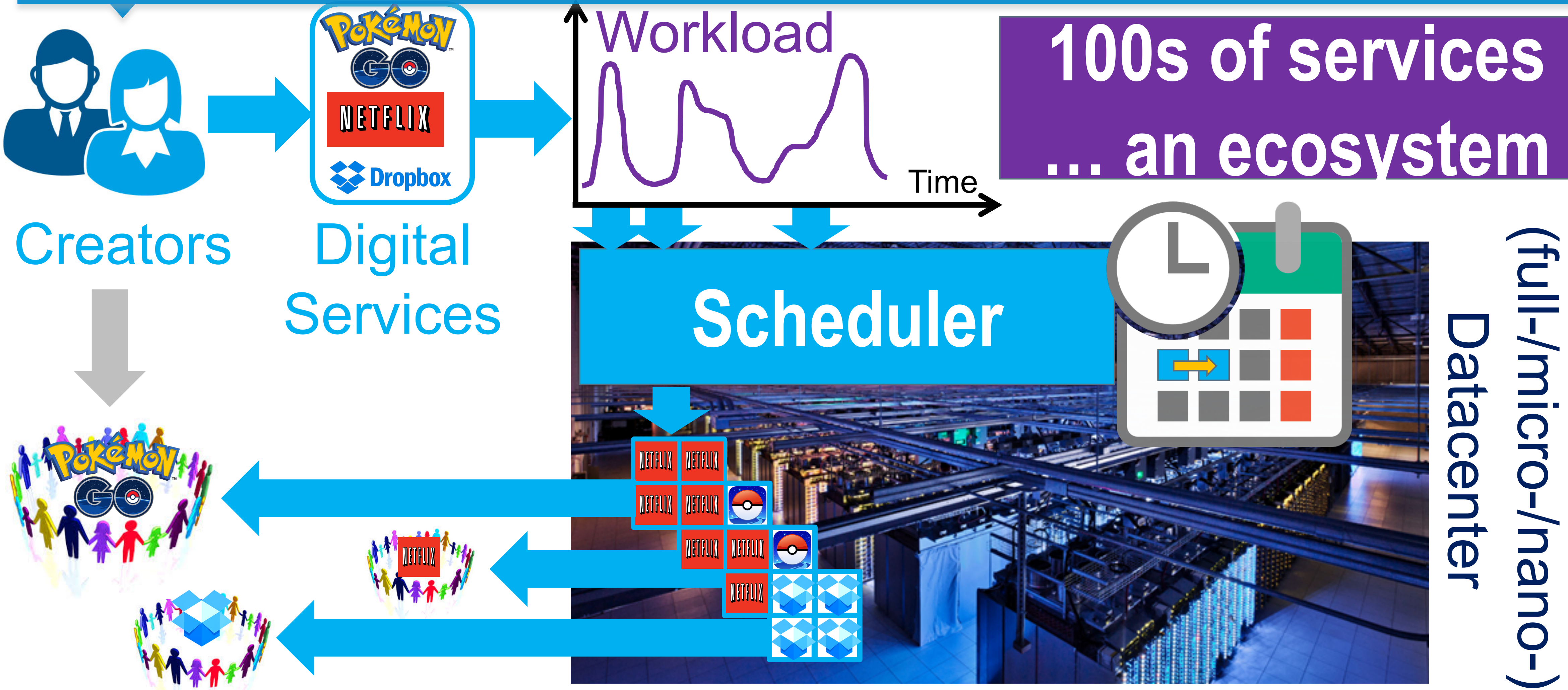




Here is how it operates…

Daily Life

THE CLOUD ECOSYSTEM: SERVICE, DATACENTER, SCHEDULER



Performance, Elasticity, Dependability, Efficiency

DIVERSE CLOUD SERVICES FOR ALL… ARE WE THERE YET?

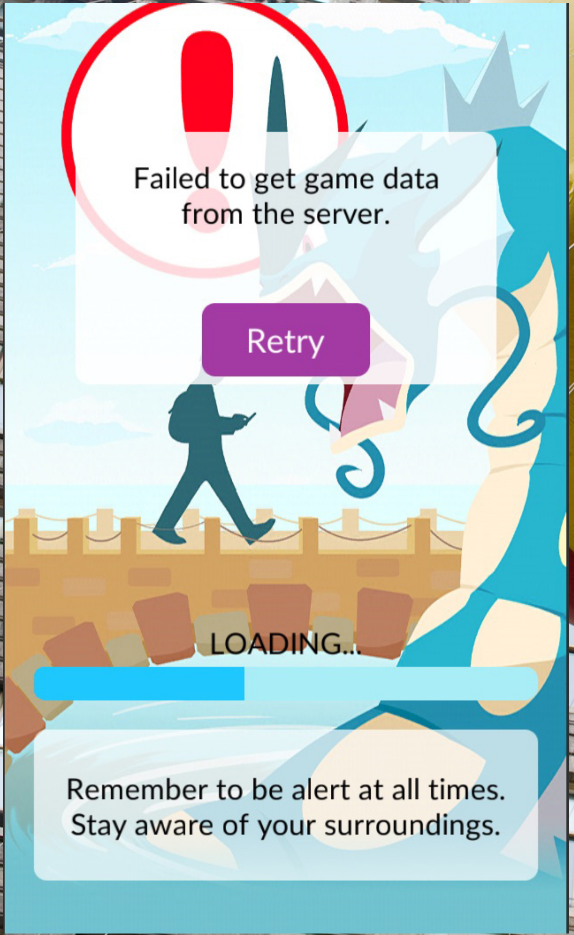
Technology
not ready

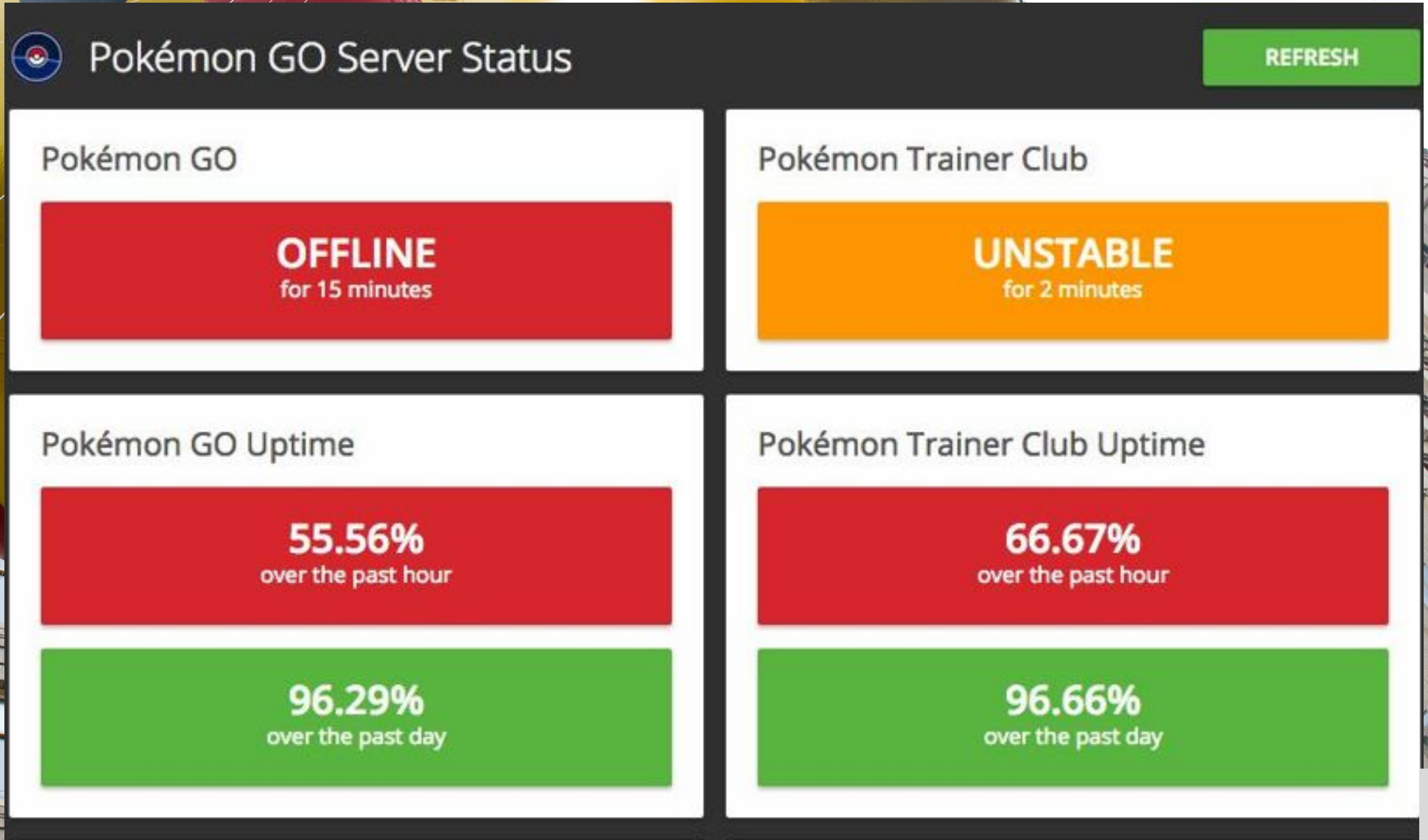
Emergence of
the mobile… 5G

Law, anti-trust:
EU, etc. vs. GAFAM

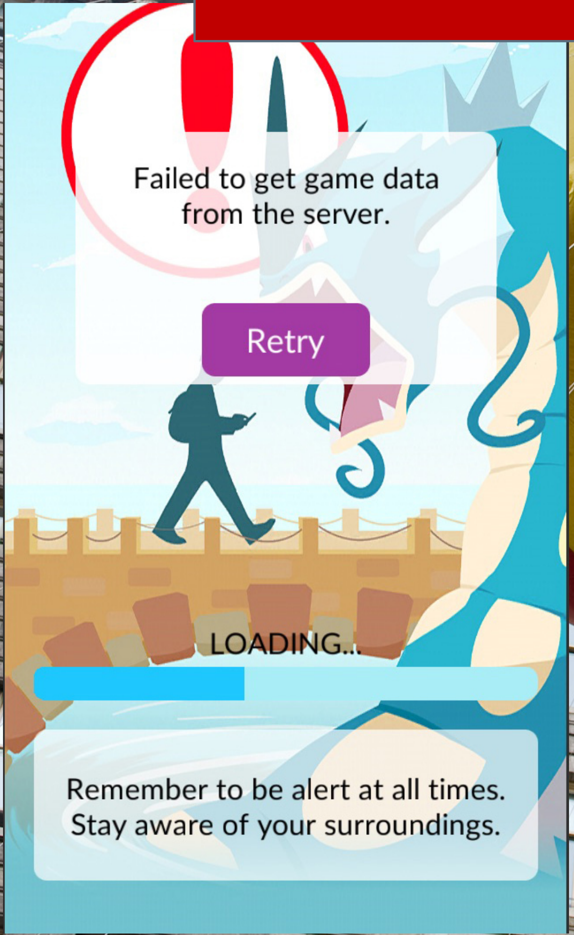
Not covered in this talk

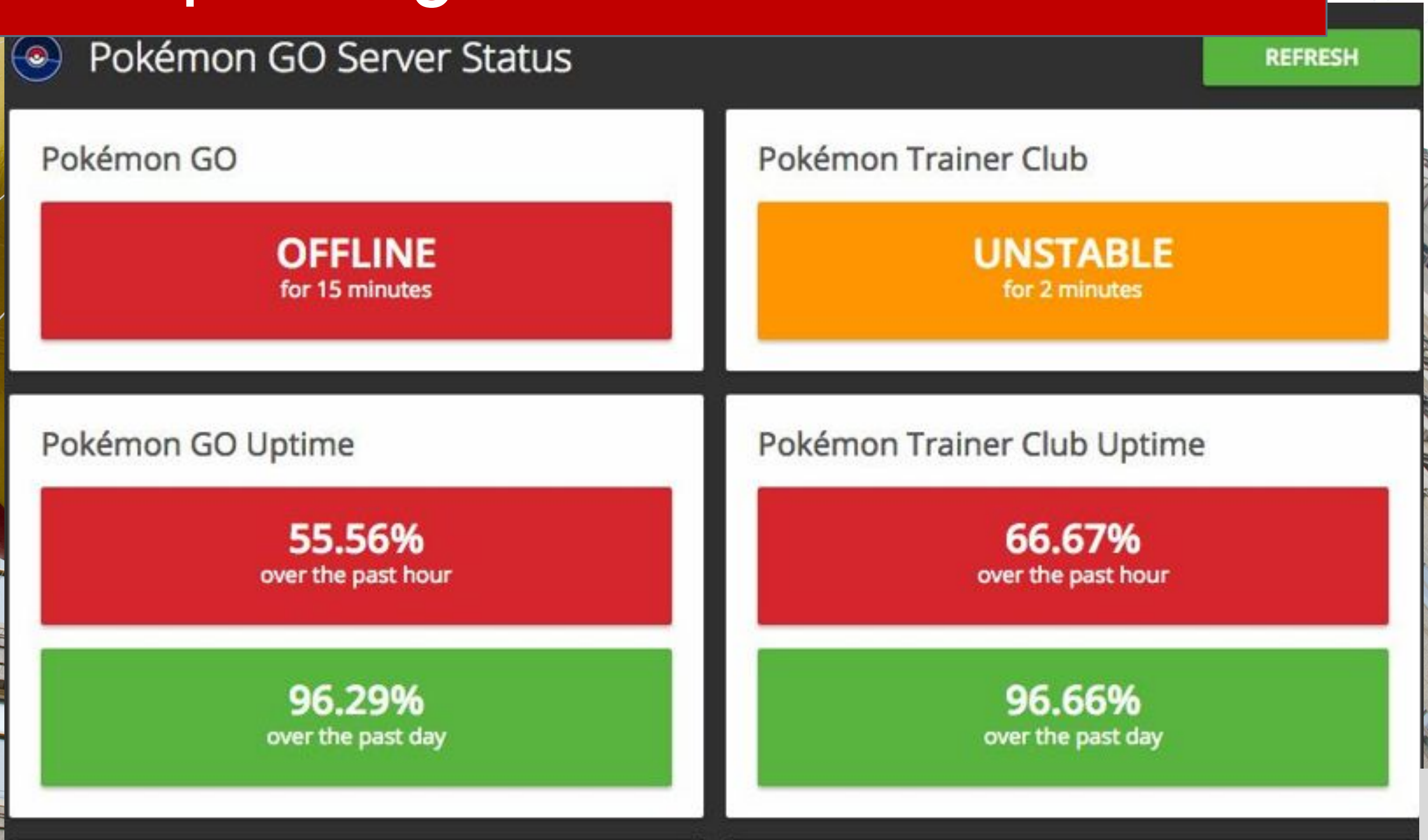
Technology not ready





Is 56% uptime good? 66%? 96%?





My Research: Massivizing Computer Systems

Technology **not ready**

Why does this* happen?

What to do about it*?

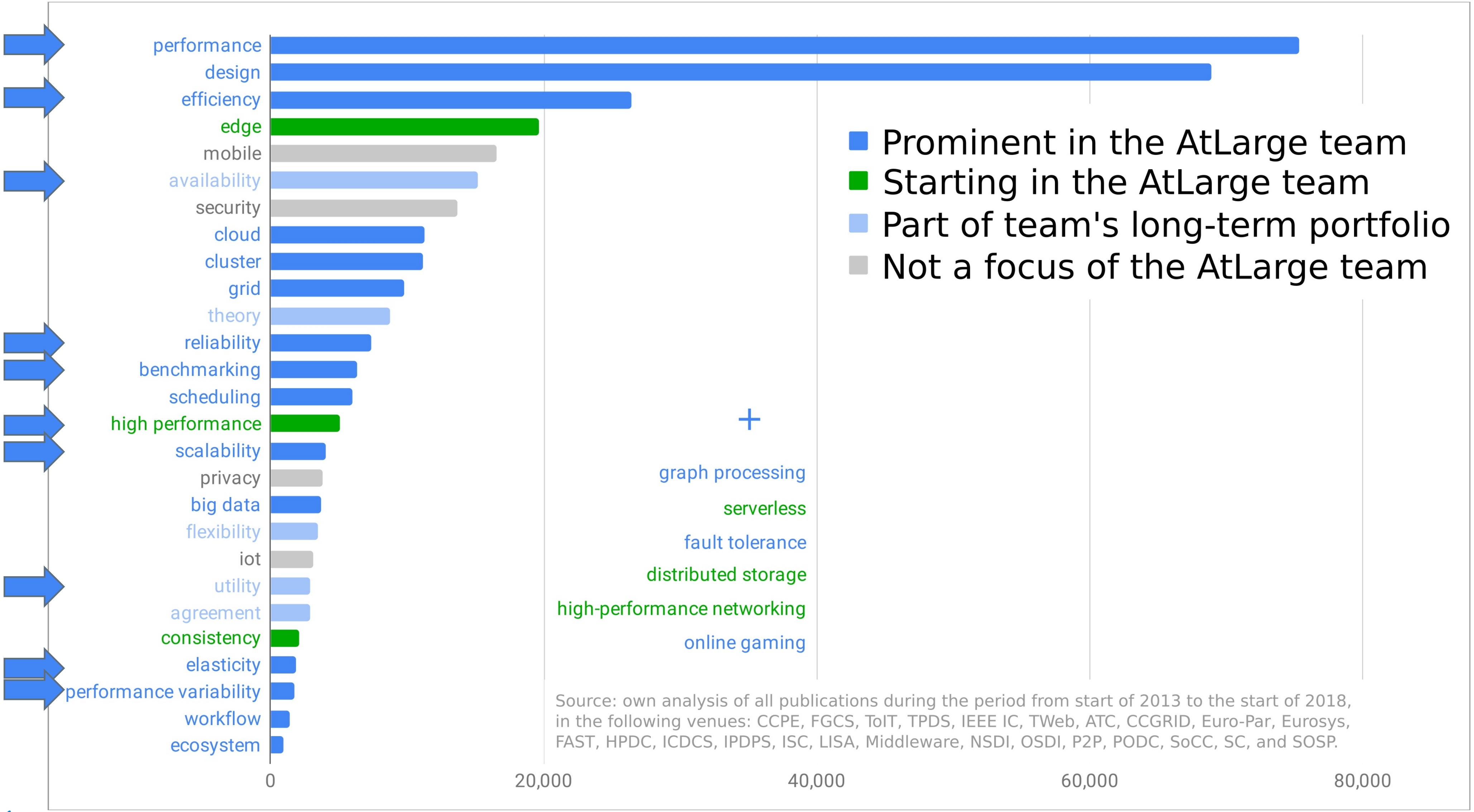
* In modern computer systems, several or all issues may be linked. Thus, looking at any single issue in isolation is no longer sufficient.

We Need Cloud Ecosystems, but We Cannot Take Them for Granted. What Next?

1. Learn from history (science, history)
2. Understand why this happens (science, experimentation)
3. Develop performance-aware solutions (science, design, engineering)

Note: my research group is broader. We build systems!

Science and Practice of Distributed Systems



VU VRIJE UNIVERSITEIT AMSTERDAM

Idea:
Grand experiments that challenge the status quo.
Same for grand observations.

- Ivan inspired me to add this to the talk. Fault is still mine!
- How do scientific theories evolve?
- Many theories, some related to experiments or observations that invalidate current theories.
- … so let's pay some attention to grand experiments or observations, which could invalidate key assumptions

The Theory Prevailing At the Time: Grids to Replace Supercomputers

1

- By fiat,

Mostly large parallel jobs: many CPUs

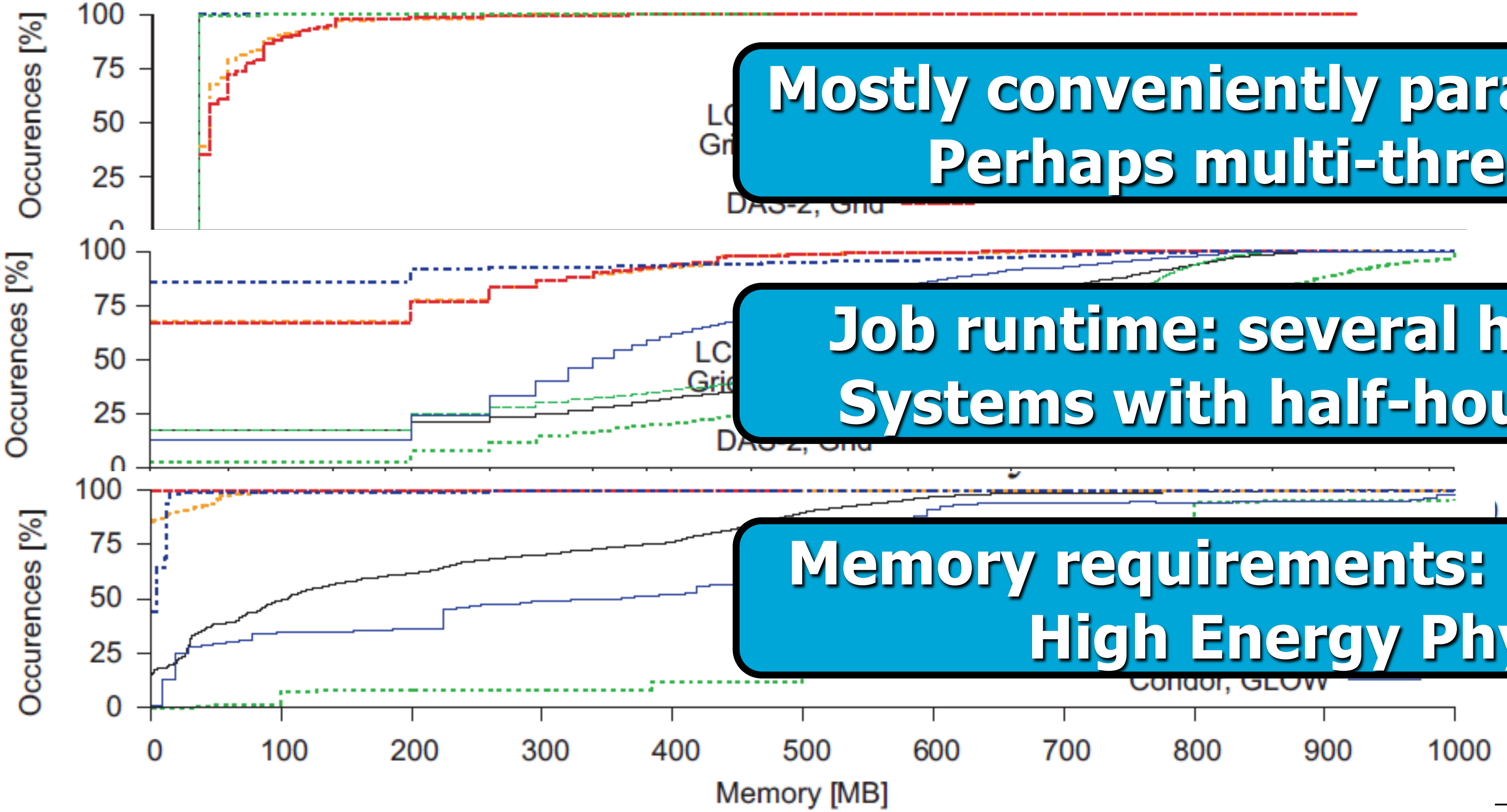
Job runtime: several days on average. Systems with a few hours at least.

Memory requirements: high.

- Experiments only with parallel applications
- Use benchmarks for them: NPB

VU VRIJE UNIVERSITEIT AMSTERDAM

BoTs by Numbers: CPU, Runtime, Mem



Mostly conveniently parallel jobs: 1 CPU
Perhaps multi-threaded apps.

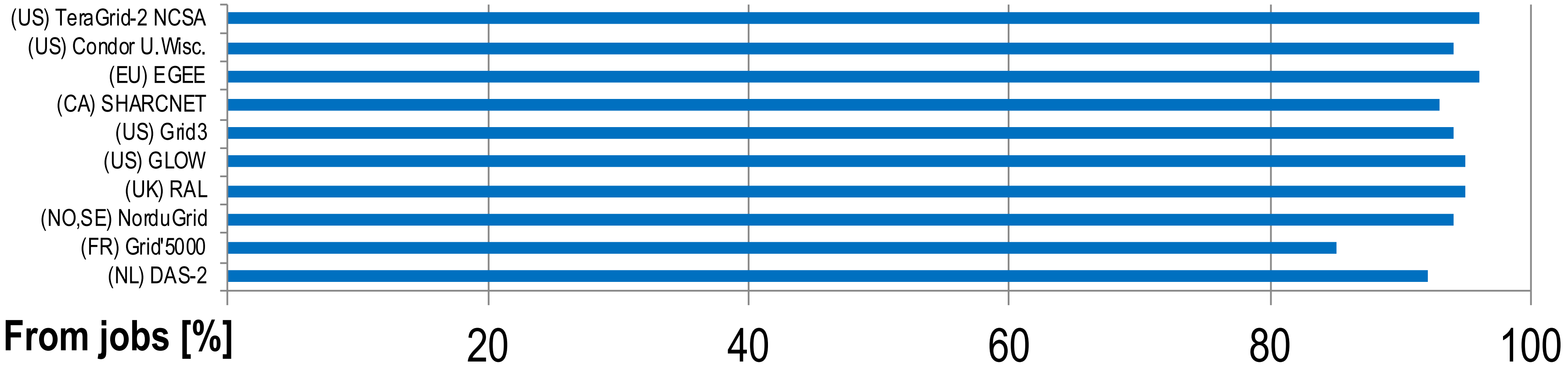
Job runtime: several hours average.
Systems with half-hour average exist.

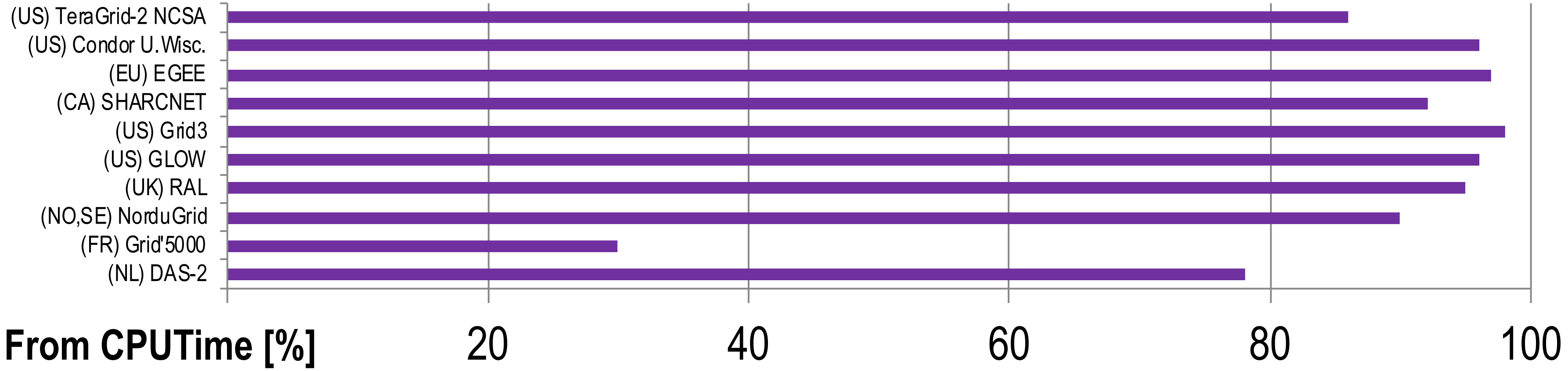
Memory requirements: modest, except
High Energy Physics jobs.

Iosup et al., The Grid Workloads Archive, FGCS, 2008.
Iosup et al., How are Real Grids Used? The Analysis of Four Grid Traces and Its Implications. Grid 2006.

1 BoTs = Dominant Programming Model for Grid Computing

Phenomena





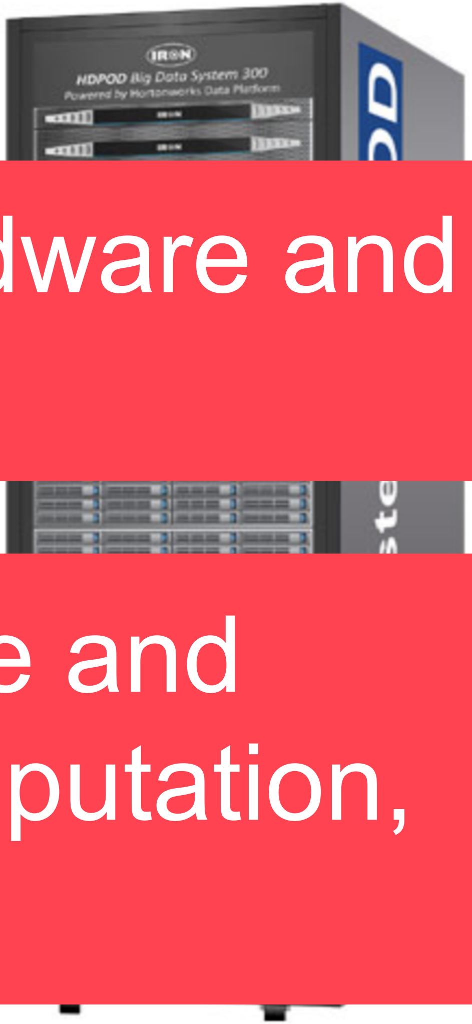
Each bar is for one grid

Each bar is a long-term workload trace

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

VU VRIJE UNIVERSITEIT AMSTERDAM

2 Challenge: HPC and Big Data Infrastructure



Highly divergent in both hardware and software!

Divergence is expensive and unsustainable: energy, computation, human resources!

Uta et al., Exploring HPC and Big Data Convergence: A Graph Processing Study on Intel Knights Landing. IEEE Cluster 2018 [Online]

2

HPC vs. Big Data Software

Most big data stacks are unable to take advantage of (HPC) hardware features.

Uta et al., Exploring HPC and Big Data Convergence: A Graph Processing Study on Intel Knights Landing. IEEE Cluster 2018 [Online]

2 Addressing the HPC and Big Data Convergence

- **Only in software: porting big data to HPC hardware**

Significant effort in porting and tuning!

Can we run big data directly on HPC hardware? What are the trade-offs?

Answer through grand experiment!

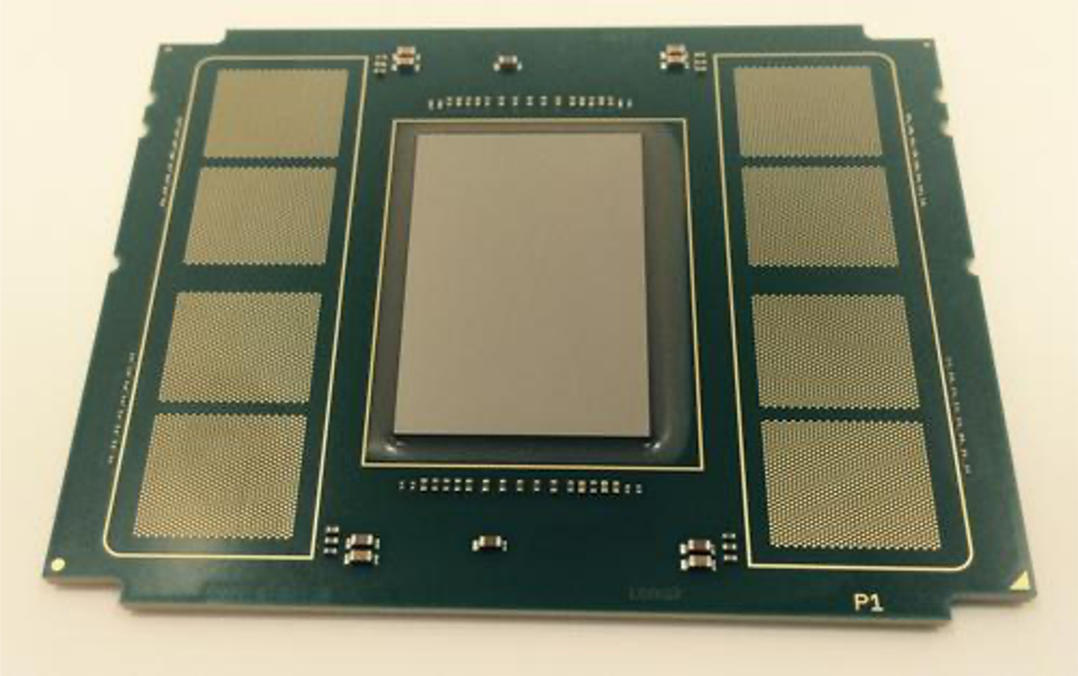
Uta et al., Exploring HPC and Big Data Convergence: A Graph Processing Study on Intel Knights Landing. IEEE Cluster 2018 [Online]

2 Big Data on Intel Knights Landing

- Intel KNL – 2nd generation Xeon Phi, already 10% Top-500 (next up: AMD?)

Can run Big Data:

- Accelerator-like self-booting CPU
- **Full x86_64 compatibility** ← no porting

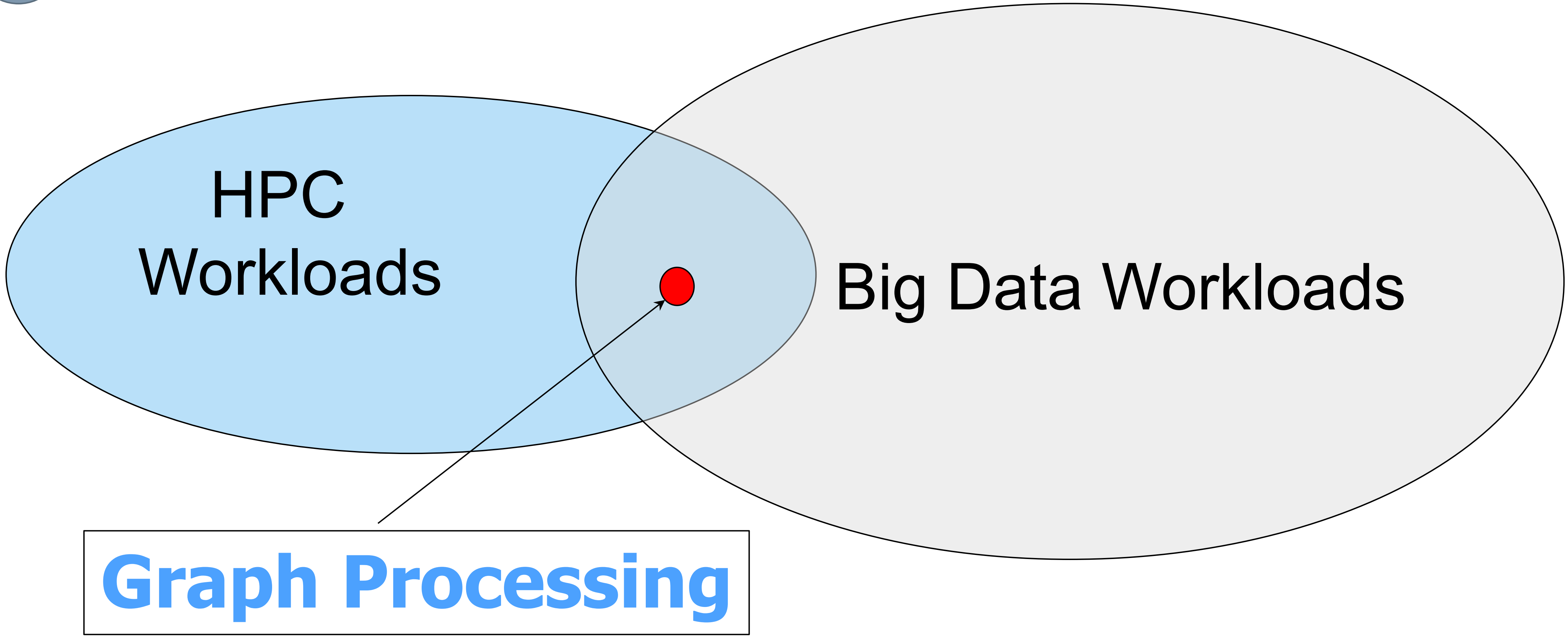


HPC Features:

- (up to) 72 low-power Intel Atom cores
- Wide vector instructions (512B)
- 16GB high-bandwidth on-chip memory
- **3 TFLOPS + 400 GB/s (on-chip) memory bandwidth**

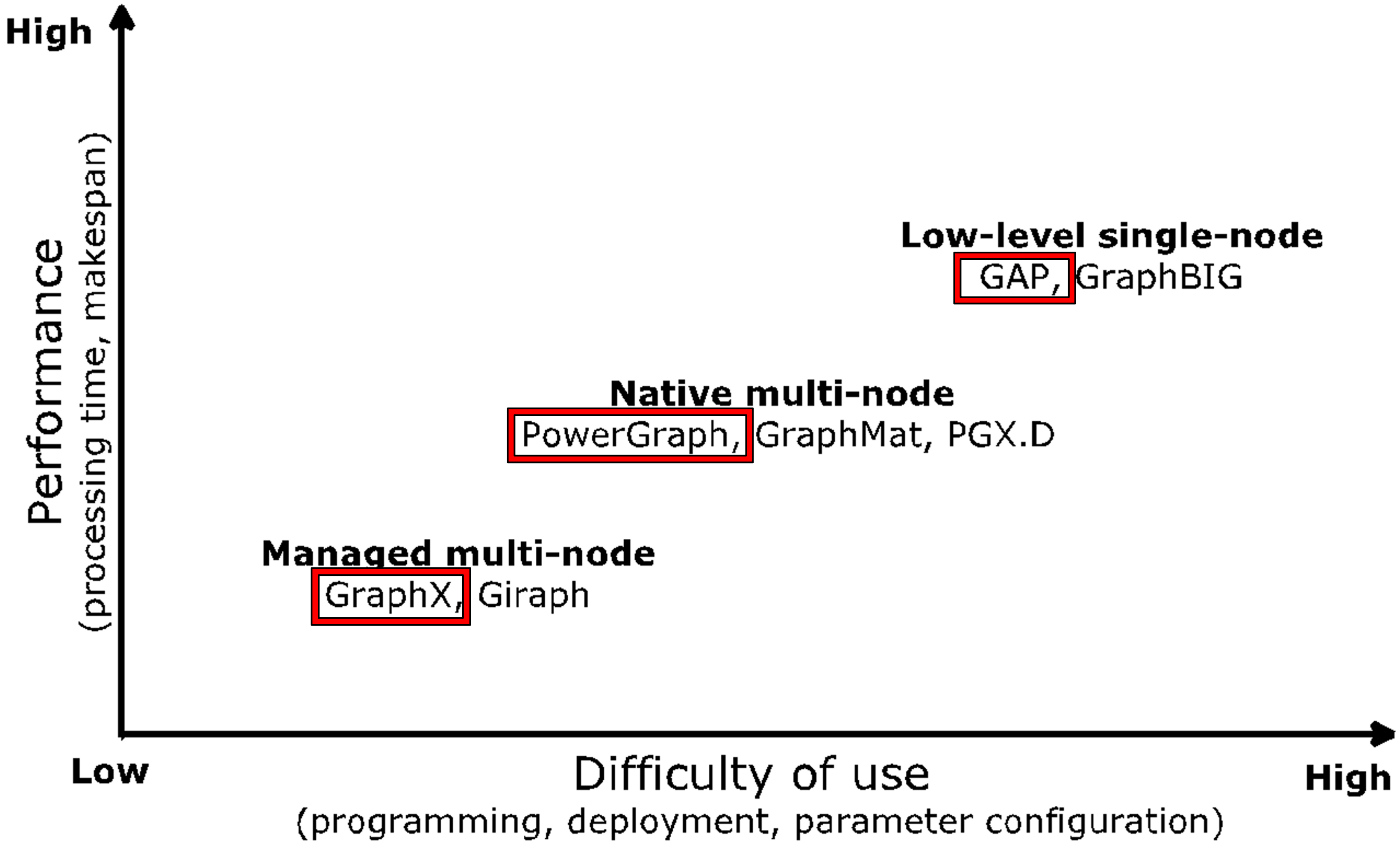
Uta et al., Exploring HPC and Big Data Convergence: A Graph Processing Study on Intel Knights Landing. IEEE Cluster 2018 [Online]

2 Graph Processing – HPC and Big Data



Uta et al., Exploring HPC and Big Data Convergence: A Graph Processing Study on Intel Knights Landing. IEEE Cluster 2018 [Online]

2 Graph Analytics Platforms



Uta et al., Exploring HPC and Big Data Convergence: A Graph Processing Study on Intel Knights Landing. IEEE Cluster 2018 [Online]

2 Quantifying the Convergence

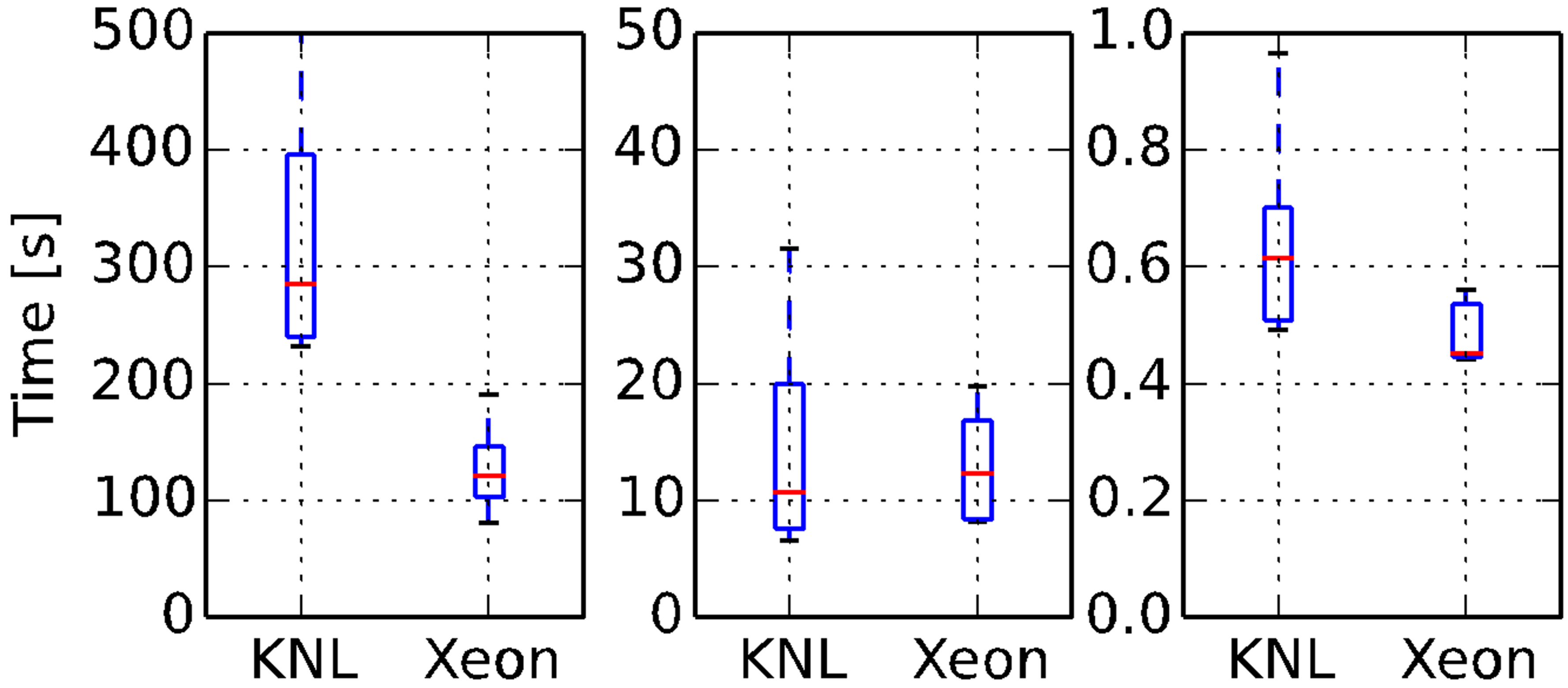
- Large-scale study – over 300,000 compute core-hours [0.3 MCh]

- Experiments run in DAS-5, Cartesius cluster*, Intel Academic cluster*
- **Q1: How does the KNL parameter space influence performance?**
- **Q2: How (difficult it is) to tune the platforms on KNL?**
- **Q3: Is KNL faster than Xeon?**
- **Q4: Does KNL scale?**

	Xeon E5-2630v3	Xeon Phi 7230
Cores	16 (32 hyperthreads)	64 (256 hyperthreads)
Frequency (GHz)	2.4	1.3
Network	56Gbit FDR InfiniBand	56Gbit FDR InfiniBand
Memory	64GB DDR4	96GB DDR4
OS	Linux 3.10.0	Linux 3.10.0

Uta et al., Exploring HPC and Big Data Convergence: A Graph Processing Study on Intel Knights Landing. IEEE Cluster 2018 [Online]

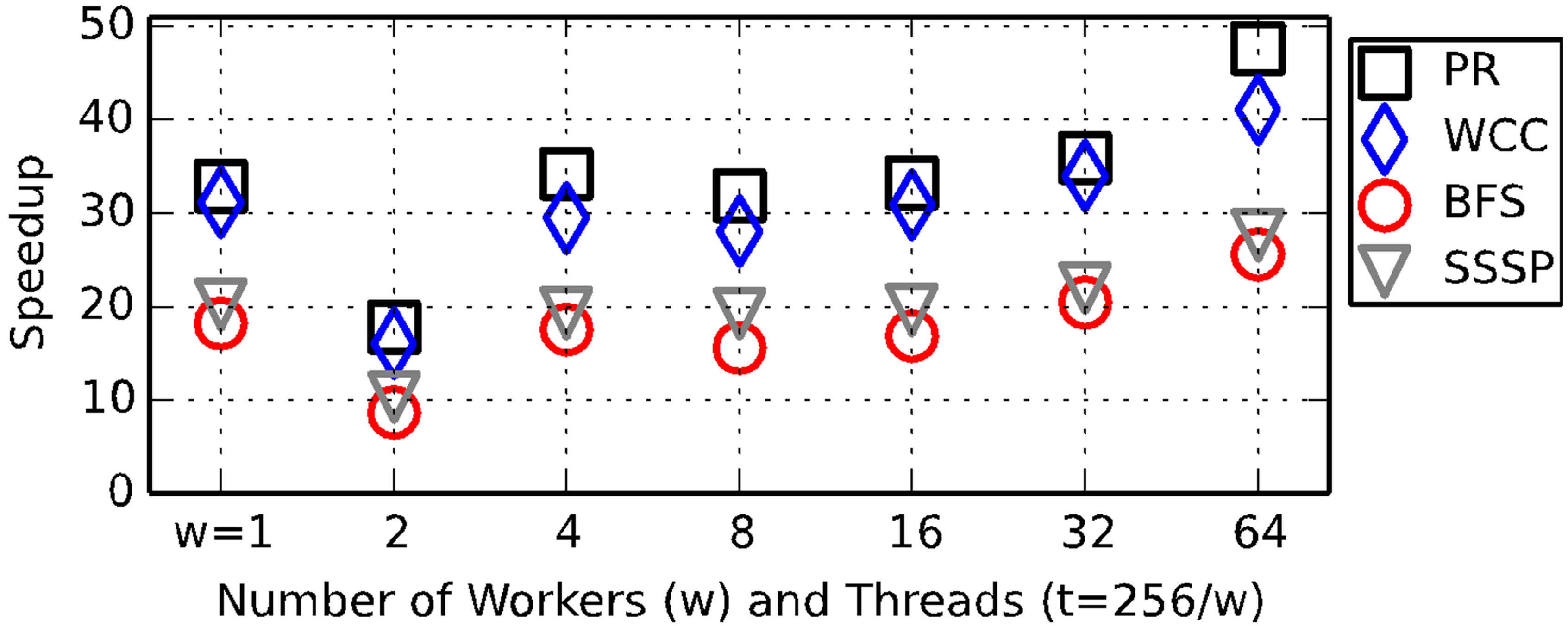
2 Hardware + Software Parameters



MF1: Much larger performance range due to KNL configurability and sw. interactions!

Study on Intel Knights Landing. IEEE Cluster 2018 [Online]

2 KNL Hardware + Platform Interaction and Tuning



MF2: On KNL, tuning (thread pinning) is important!

VU VRIJE UNIVERSITEIT AMSTERDAM

Uta et al., Exploring HPC and Big Data Convergence: A Graph Processing Study on Intel Knights Landing. IEEE Cluster 2018 [Online]

Take-home Message: Main Findings

Alexandru Uta

- **HPC & Big Data can converge at a hardware level! But…**

- MF1: **HPAD** – hardware adds an extra complexity layer
- MF2: **Tuning** – good performance entails significant tuning for KNL
- MF3: **Scaling** – KNL scales well vertically, but cannot scale horizontally
- MF4: **H-P interaction** – platforms closer to hardware perform better on KNL
- MF5: **Convergence** – KNL outperforms Xeon

- Does software need to adapt to KNL? (Or other architectures?)

VU VRIJE UNIVERSITEIT AMSTERDAM

Summary:
Grand experiments that challenge the status quo.

- Many theories, some related to experiments or observations that invalidate current theories.

1. What are the grand observations that could challenge today's assumptions?
2. What are the grand experiments that could challenge today's assumptions?

Idea:
Meaningful discovery requires a mix of experimental science, design, and engineering.

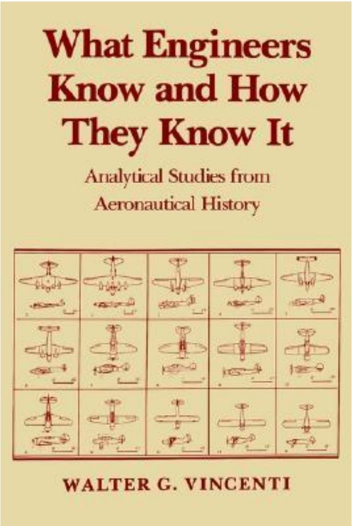
These are very different activities!

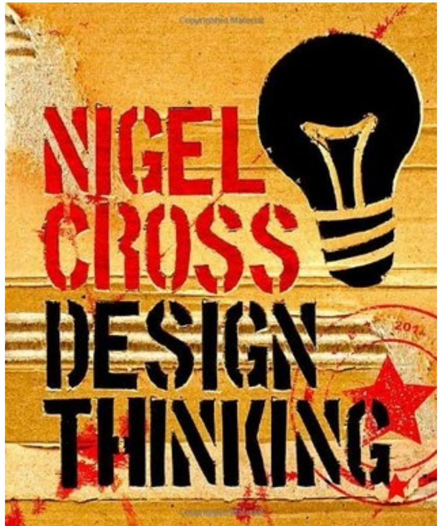
MEANINGFUL DISCOVERY

NO SYSTEMATIC PROCESS FOR COMPUTER SYSTEMS

SO I'LL USE EXAMPLES

science + engineering + design



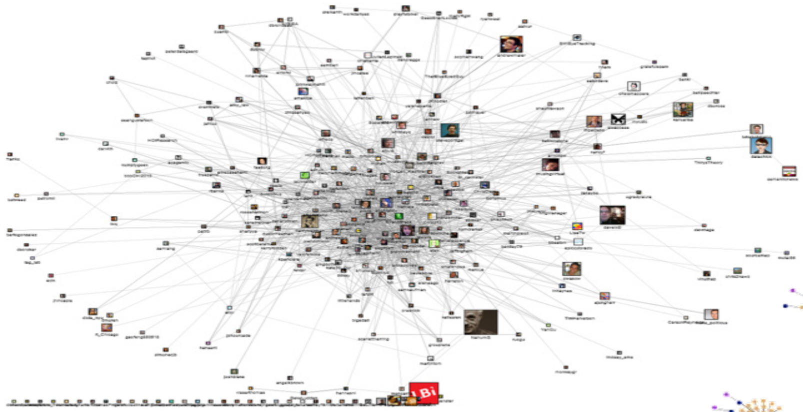


THE COMPUTER SYSTEMS TRIPLET

[Iosup et al. ICDCS'18]
[Iosup et al. ICDCS'19]

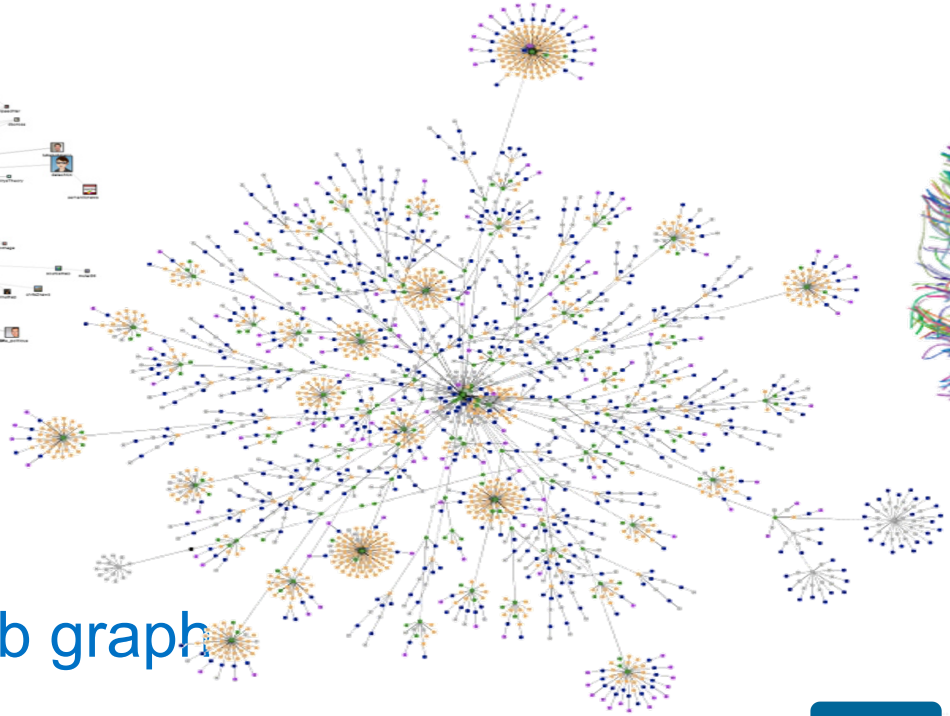
MEANINGFUL DISCOVERY IMPACTING SCIENCE

THE NEED FOR SPEED IN GRAPH PROCESSING



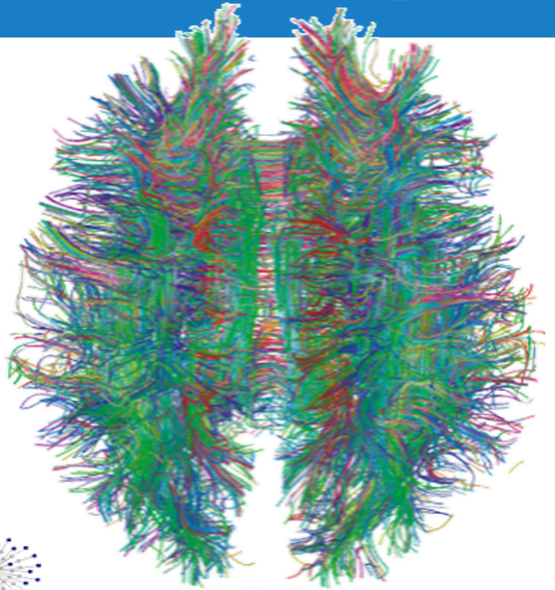
Social network

~1 billion vertices
~100 billion connections



Web graph

~50 billion pages
~1 trillion hyperlinks



Brain network

~100 billion neurons
~100 trillion connections

Sources: Smith, CHI'10; Blog webpage; Gigandet et al., PLoS ONE 3(12)]

MEANINGFUL DISCOVERY IMPACTING SCIENCE

THE NEED FOR SPEED IN GRAPH PROCESSING

ORACLE PGX

Intel Graphmat

hadoop^2

NetSysLab

TOTEM

APACHE GIRAPH

Neo4j the graph database

IBM System G

medusa-gpu

Medusa: Simplified Graph Processing on GPUs

GraphLab

PROJECT PEGASUS

HyperGraphDB

Trinity

StratoSphere Above the Clouds

GraphX

Which platforms perform well?

What to tune?
What to re-design?

AUTOMATED TESTING FOR DISTRIBUTED ECOSYSTEMS?

LDBC GRAPHALYTICS: BENCHMARKING LEADING TO DISCOVERY

LDBC *The graph & RDF benchmark reference*

- Graphalytics:
  - > Benchmark
  - > Many classes of algorithms used in practice
  - > Diverse real and auto-gen datasets
  - > Diverse experiments, representative for practice
  - > Renewal process to keep the workload relevant
  - > Enables comparison of many platforms, community-driven and industrial
  - > Global Competition

- Community endorsed:

graphalytics.org

- Surprising findings:

Performance: orders of magnitude difference due to each of platform, algorithm, dataset, and hardware

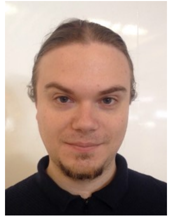
[Iosup et al. PVLDB'16] [Guo et al. CCGRID'15] [Guo et al. IPDPS'14]

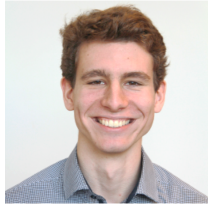
https://graphalytics.org

EXPERIMENTAL METHODS OF DISCOVERY

UNIQUE OPPORTUNITY: WE DRINK OUR OWN CHAMPAGNE (*IN VIVO*)!

Our Prototypes (*in phyisico/in vitro*)

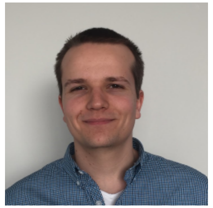

Alex Uta


Georgios Andreadis


Fabian Mastebroek


acheendra


Maria Voinea


Laurens Versluis


Alexey Ilyushkin

\+ amazon web services™

We also use clouds

\+ OpenDC

And simulators (*in silico*)

LOCALIZATION OF BOTTLENECKS → PERF. ISSUES

ENGINEERING LDBC GRAPHALYTICS: MODELING LEADS TO DISCOVERY

LDBC The graph & RDF benchmark reference

- Graphalytics Grade10:
  > Automated bottleneck detection
  > Automated identification of performance issues

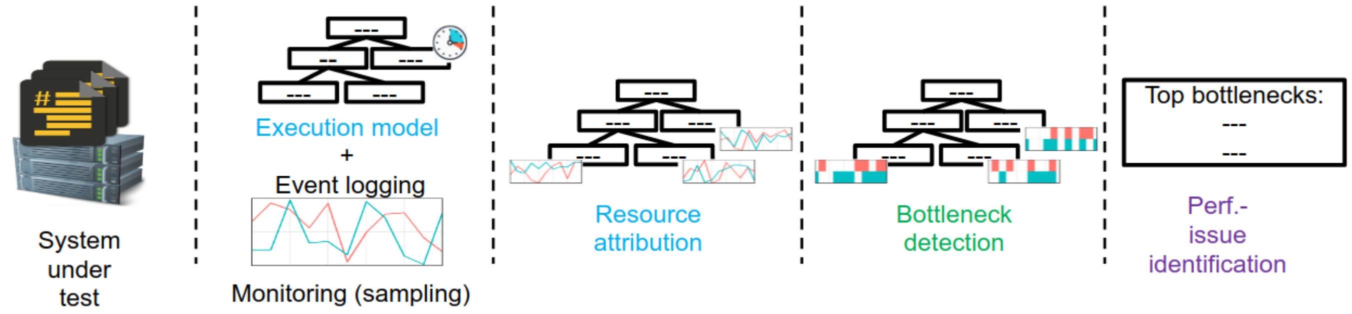
APACHE GIRAPH

- Without Grade10:

No bottleneck at all

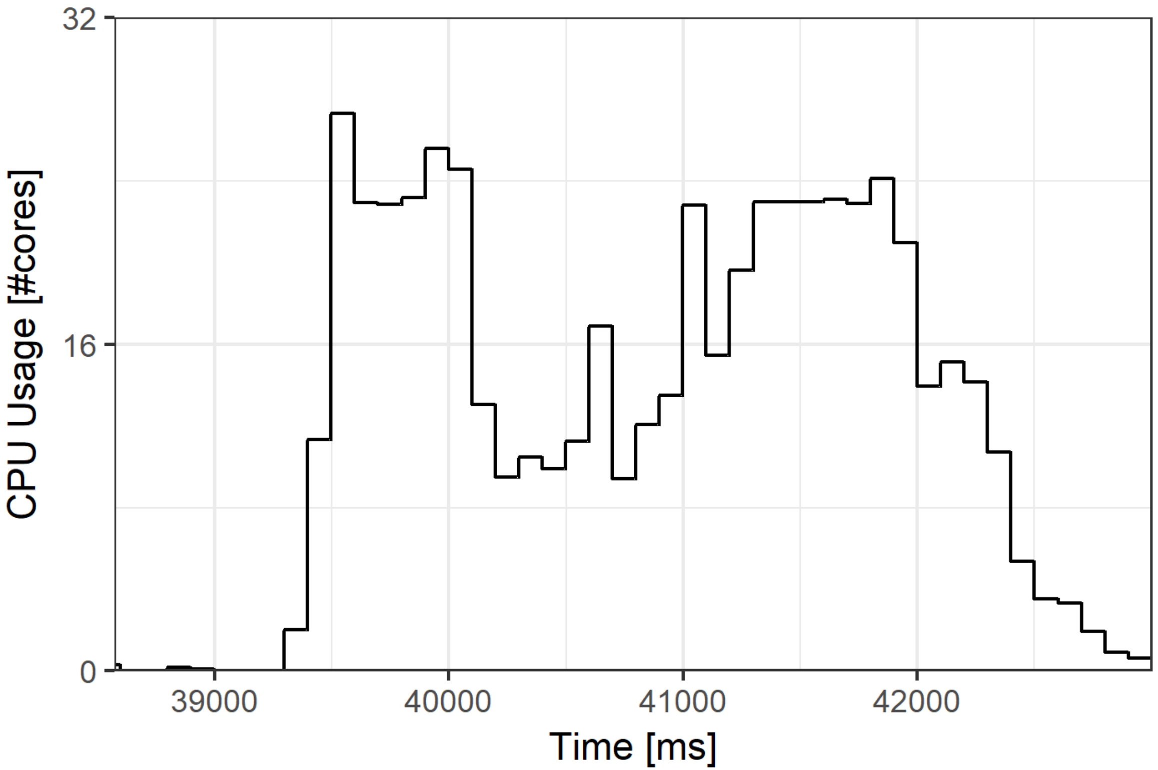
- With Grade10:

Always bottleneck
Can explain causes:
+ Message queue full
+ Garbage collector
+ CPU
+ Others



Multi-stage process, works in ecosystem

Preliminary Result: Analysing a Giraph Job

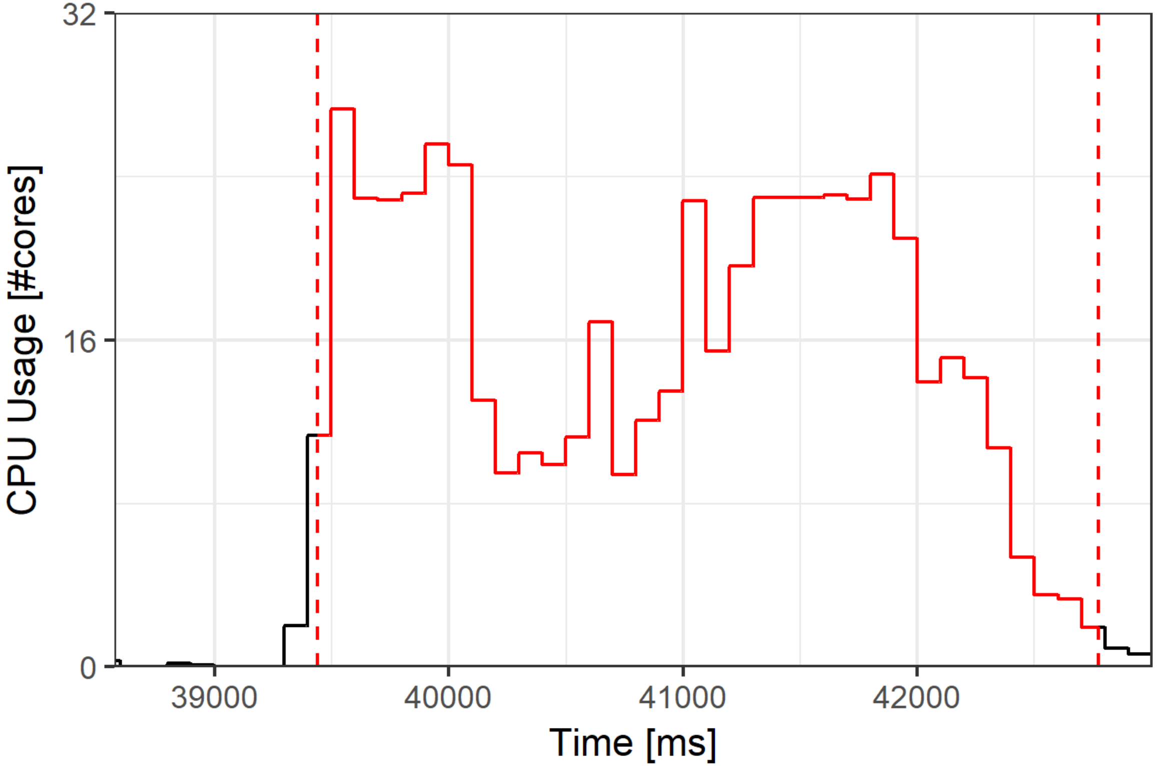


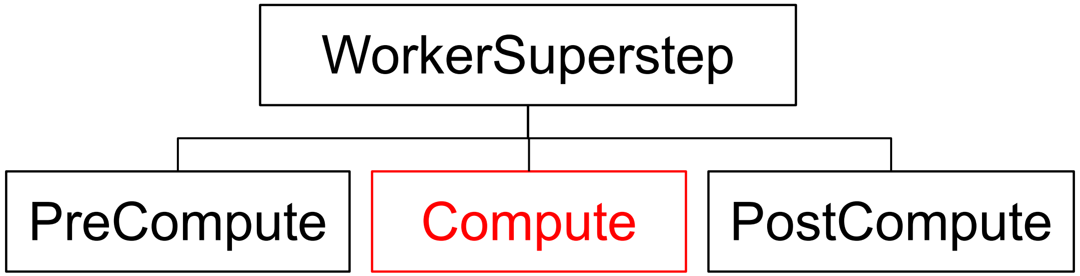
WorkerSuperstep

CPU usage < 32 cores (100%), so no bottleneck

... yet

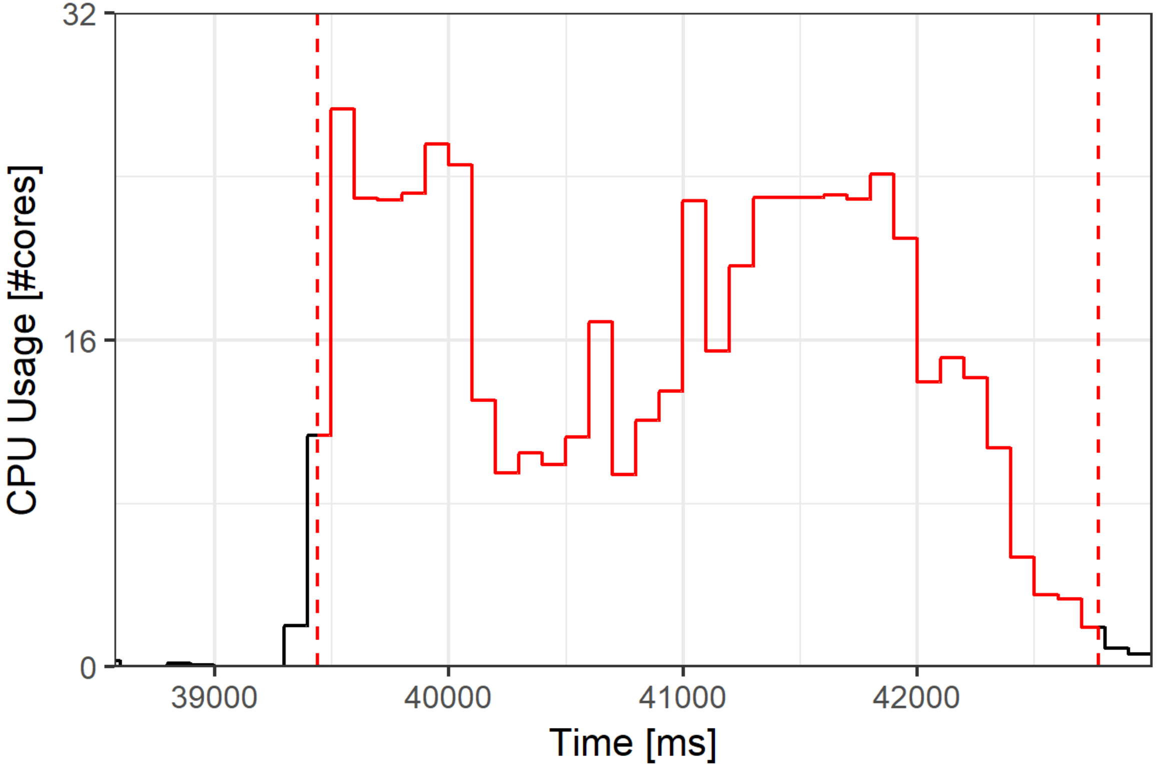
Preliminary Result: Analysing a Giraph Job

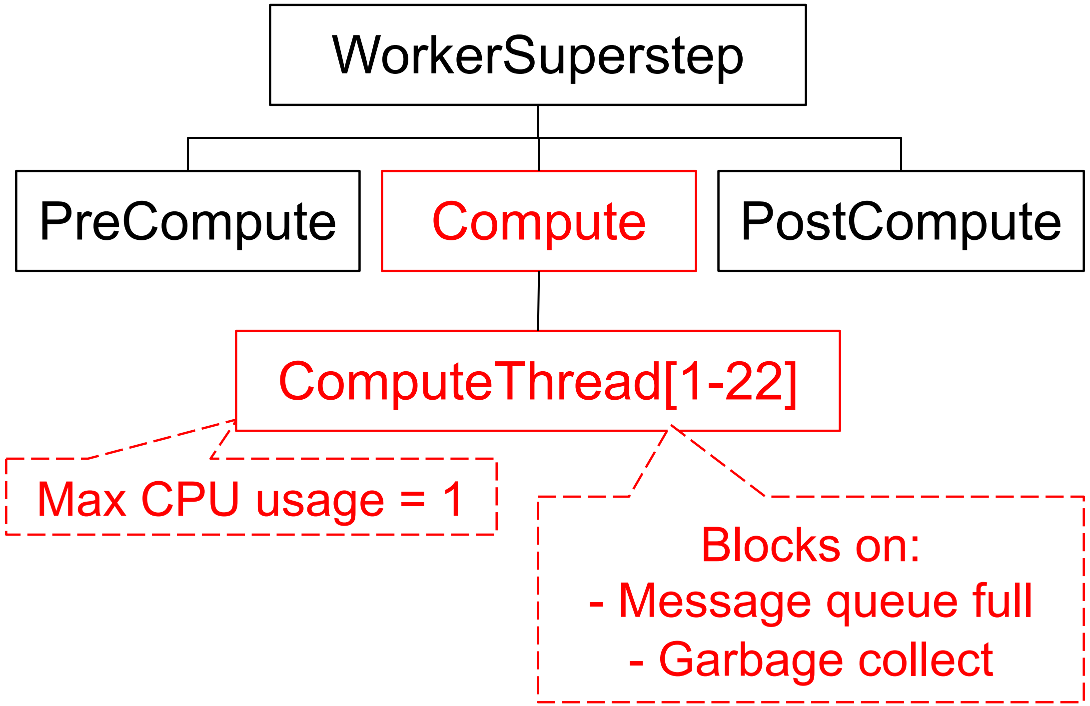




VU VRIJE UNIVERSITEIT AMSTERDAM

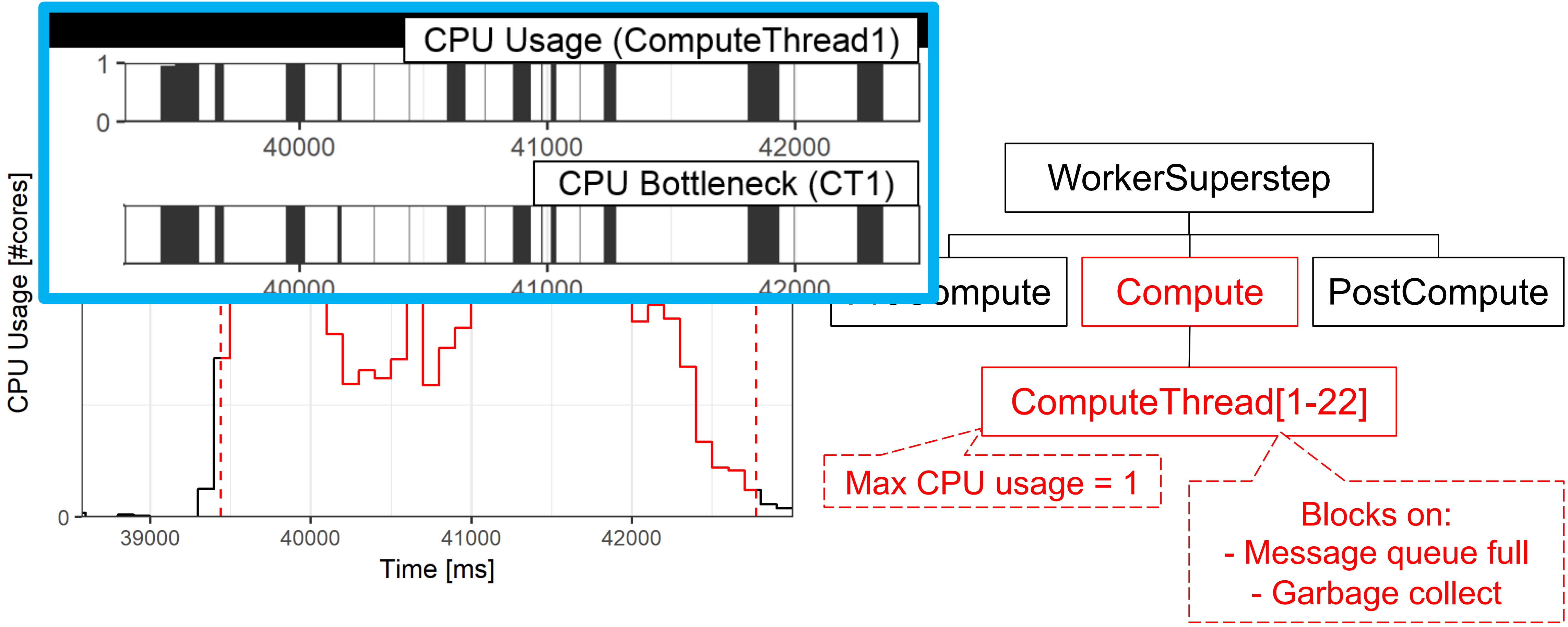
Preliminary Result: Analysing a Giraph Job



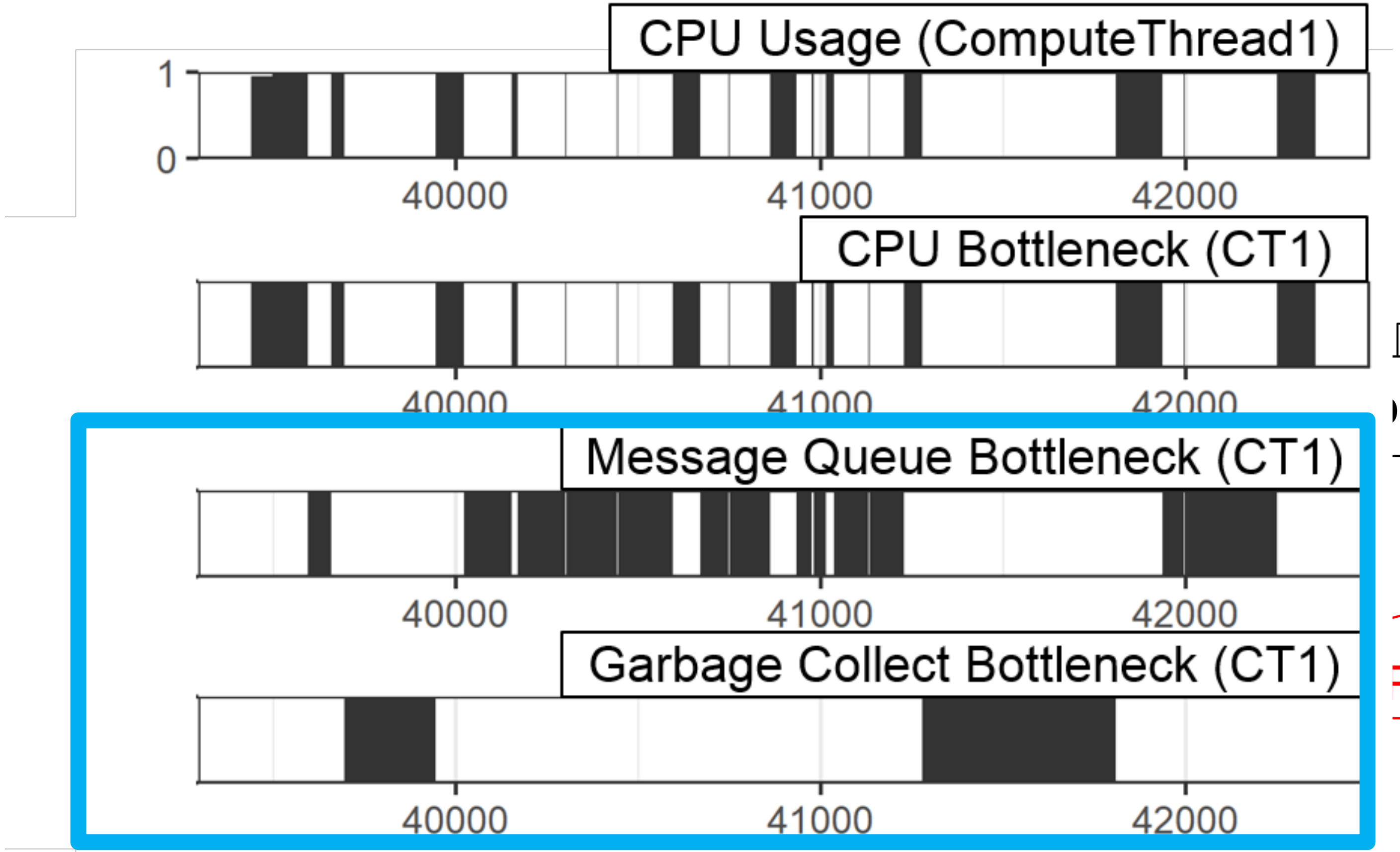


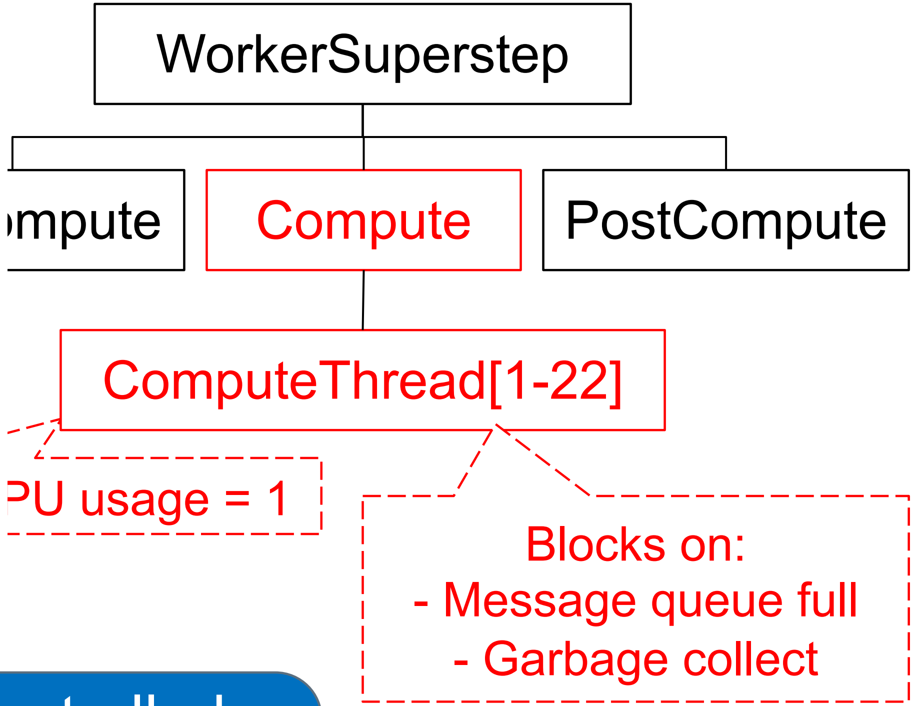
VU VRIJE UNIVERSITEIT AMSTERDAM

Preliminary Result: Analysing a Giraph Job



Preliminary Result: Analysing a Giraph Job





For one class of eco-systems, under tightly controlled experimental conditions, with careful analysis → bottlenecks!

VU

On Shovels vs. Pianos
(hint: it's ok to prefer shovels, but do they exist?)

- Tools ~ Shovels
  - Limited use
  - Easy to use, can just pick up and use
- Instruments ~ Pianos
  - Configurable
  - Must learn how to use, also require the right environment

Source: bol.com

Source: Kawai US

VU VRIJE UNIVERSITEIT AMSTERDAM

Summary:
The interplay of experimental science, design, and engineering leads to important results

1. These are very different activities. Learn them all!
2. Experimentation in vivo, in vitro, in silico
3. Benchmarking reveals graph processing, like many activities, leads to large design space & complex trade-offs
4. Performance ~ f(HW+SW platform, data, algorithm, config)
5. Experimental instruments still drive the field, we need tools (we also need more general findings)

Idea:
Meaningful discovery requires understanding the cloud universe is based on ecosystems.

Ecosystem vs. systems, a primer

- Structure: composites of smaller assemblies, but for ecosystems some constituents are produced elsewhere, with different practices
- Operation: ecosystems exhibit many unknown phenomena, less understood dynamics, and socio-technical issues
- Lifecycle: some ecosystem constituents will perish, or be replaced with others that may not actually fulfill the needs

MEANINGFUL DISCOVERY

BUT … IS THERE A SYSTEMATIC WAY TO APPROACH THESE PHENOMENA?

FUNDING: > 3B USD

- The Human Genome Project:
  - Physical map covering >90% human genome
  - Sequence data made available open-access
- Big Science:
  - Took >10 years to complete
  - Led by US, work by 20 groups in CN, DE, FR, JP, UK, US
- Big impact:
  - Decrease cost of sequencing
  - Facilitate biomedical research

International Human Genome Sequencing Consortium, Initial sequencing and analysis of the human genome, Nature 409, Feb 2011. [Online]

Julie Gould, The Impact of the Human Genome Project, Naturejobs blog, 2015. [Online]

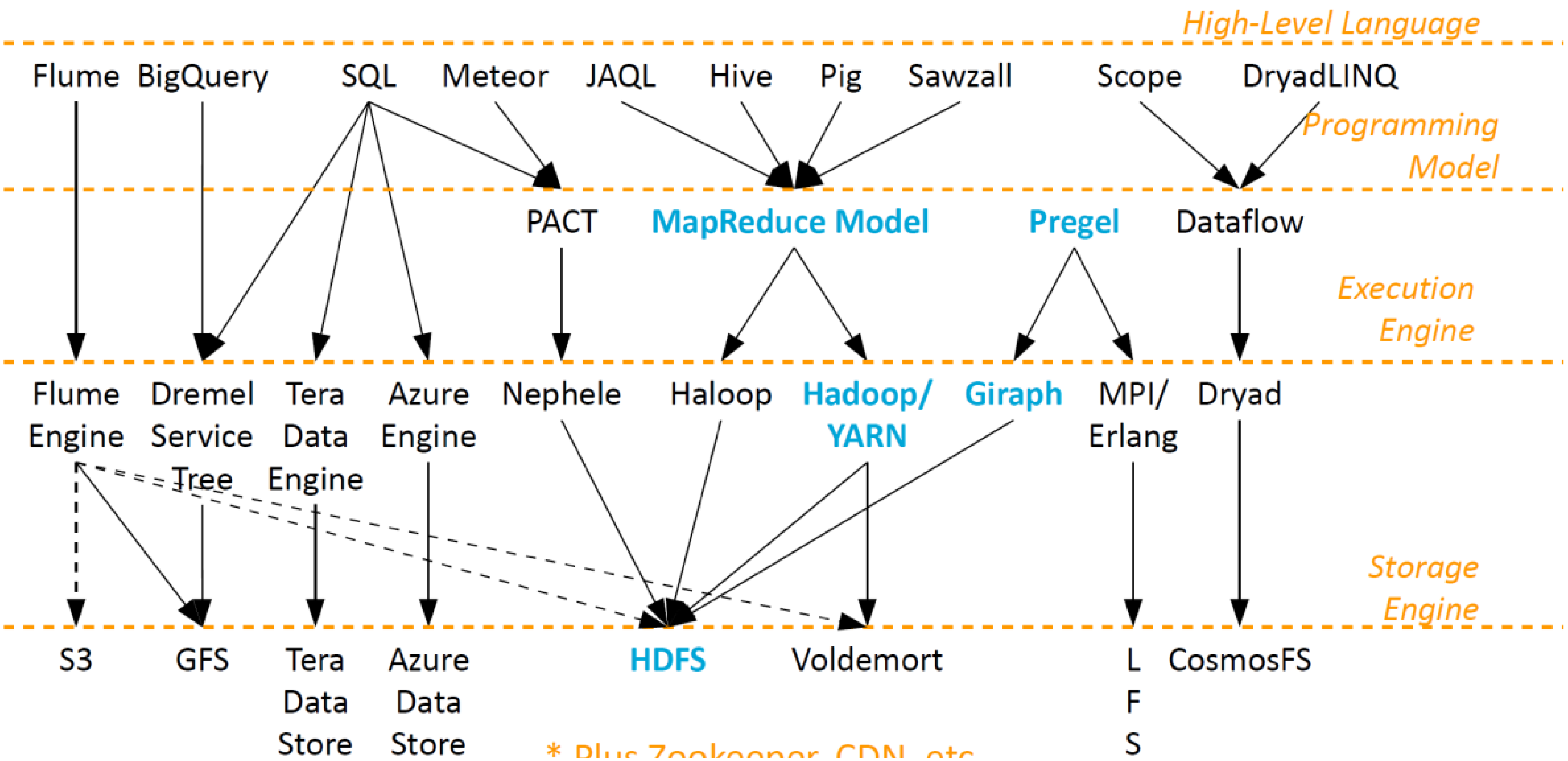
HOW TO MANAGE SYSTEM COMPLEXITY?

THE COMPLEXITY CHALLENGE

DAGSTUHL SEMINAR, 2011

4 Core Layers:

4. High-Level Language
3. Programming Model
2. Execution Engine
1. Storage Engine



[Iosup et al. ICDCS'18]

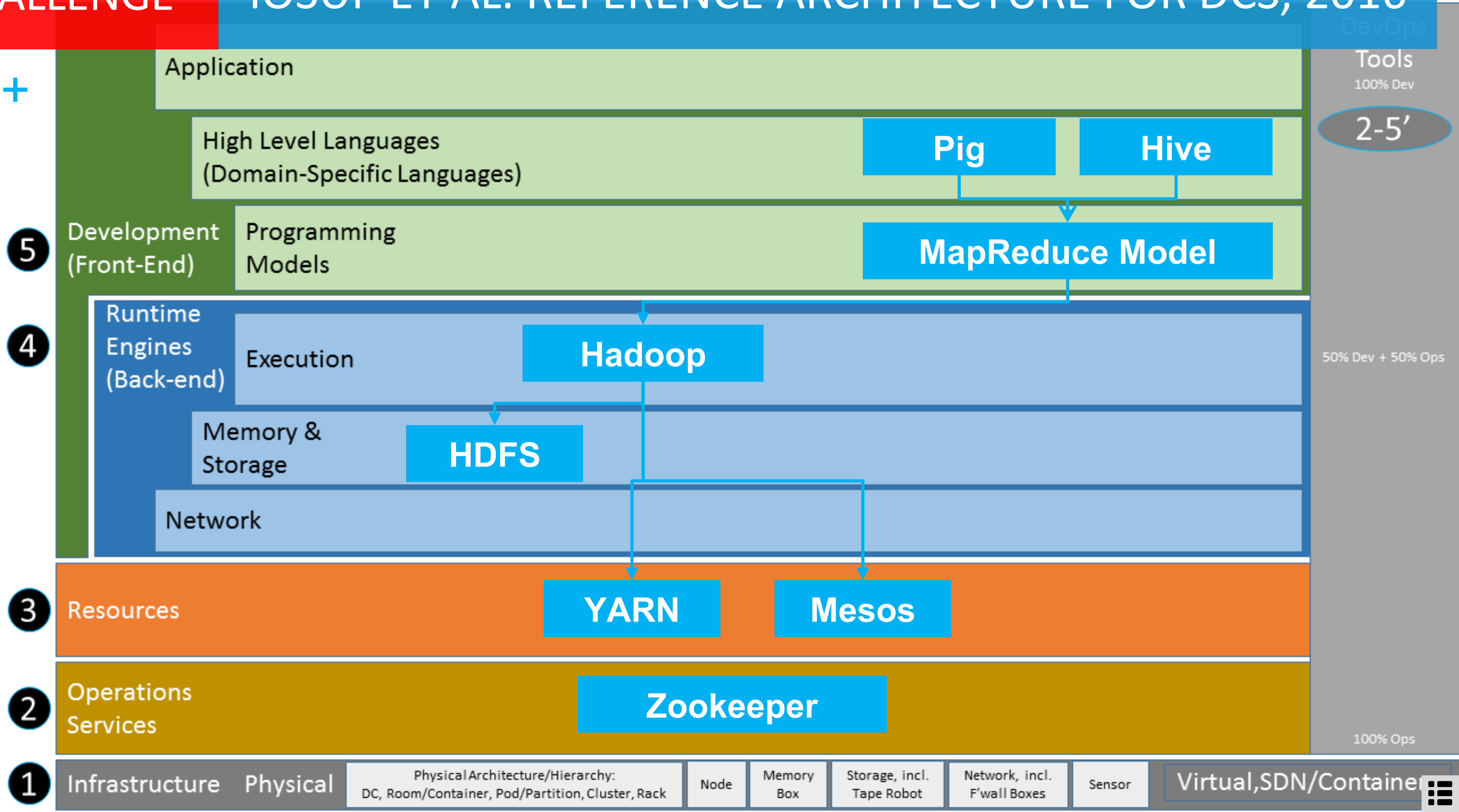
HOW TO MANAGE SYSTEM COMPLEXITY?

THE COMPLEXITY CHALLENGE

IOSUP ET AL. REFERENCE ARCHITECTURE FOR DCS, 2016

Focus on DevOps + Applications, 5 Core Layers:

5. Development (Front-end)
4. Runtime Engines (Back-end)
3. Resources
2. Operations Services
1. Infrastructure



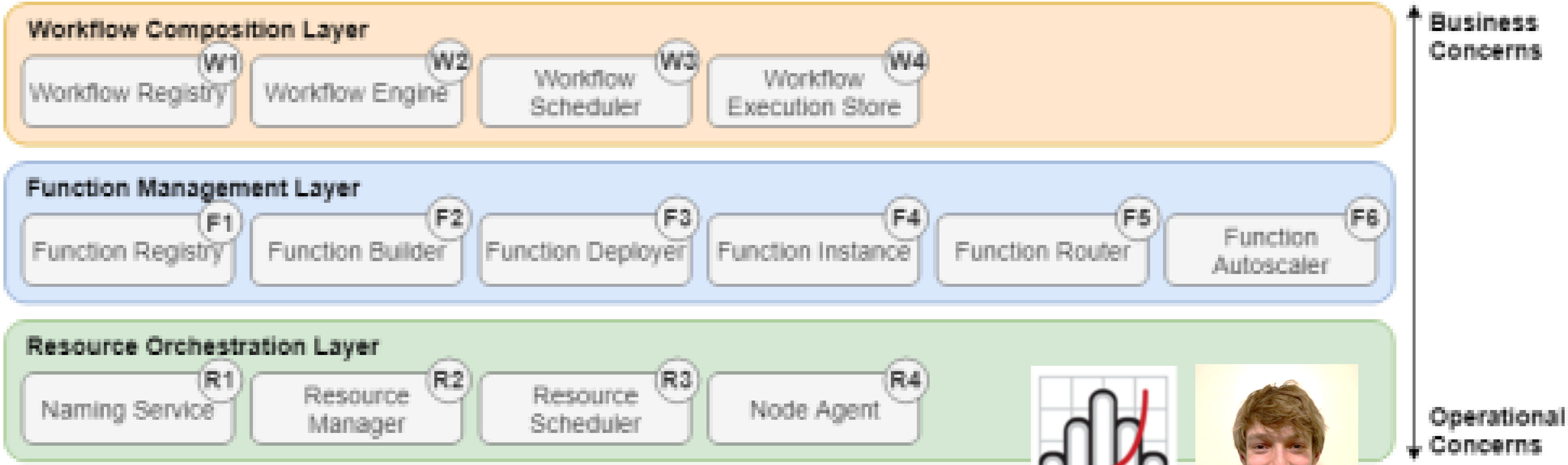
VU

[Iosup et al. ICDCS'18]

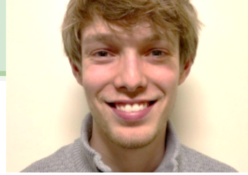
HOW TO MANAGE SYSTEM COMPLEXITY?

THE COMPLEXITY CHALLENGE

REFERENCE ARCHITECTURE OF FAAS PLATFORMS, 2019



[van Eyk et al. (2018) Serverless is More: From PaaS to Present Cloud Computing, IEEE Internet Computing] [Online]



Erwin van Eyk

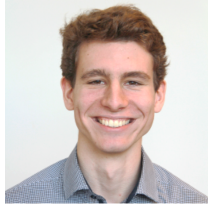
VU

THE SUPER-DISTRIBUTION PRINCIPLE

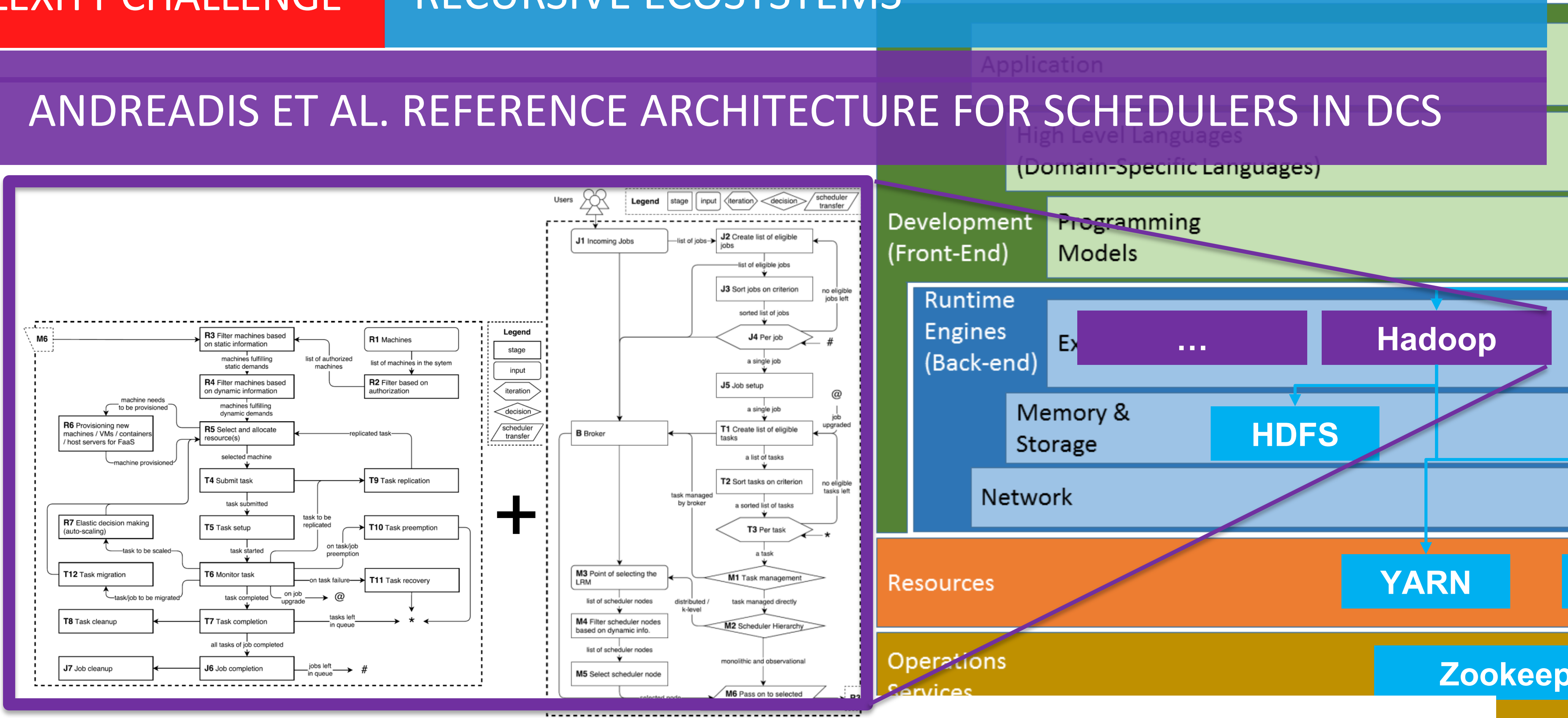
THE COMPLEXITY CHALLENGE

RECURSIVE ECOSYSTEMS

ANDREADIS ET AL. REFERENCE ARCHITECTURE FOR SCHEDULERS IN DCS



Georgios Andreadis



[Andreadis et al. SC'18]

Summary:
Ecosystems are composites. Super-distribution!

1. Ecosystems = composites

2. Conquer complexity with (system-level) reference architectures

3. The super-distribution principle: ecosystems are recursively comprised of systems and even ecosystems

4. Work done in the software engineering community, but less in the computer systems community (after 1960s)

Idea:
Meaningful discovery requires understanding the cloud universe is based on ecosystems.

Ecosystem vs. systems, a primer

- Structure: composites of smaller assemblies, but for ecosystems some constituents are produced elsewhere, with different practices
- Operation: ecosystems exhibit many unknown phenomena, less understood dynamics, and socio-technical issues
- Lifecycle: some ecosystem constituents will perish, or be replaced with others that may not actually fulfill the needs

MEANINGFUL DISCOVERY

UNCOVERING THE MYSTERIES OF OUR UNIVERSE

GALILEO GALILEI, 1608-9, 3-8X TELESCOPE

MERELY AN INSTRUMENT?

FUNDAMENTAL SCIENCE?

Garney. The Inquisition's Semicolon: Punctuation, Translation, and Science in the 1616 Condemnation of the Copernican System, ArXiv document 1402.6168. [Online]

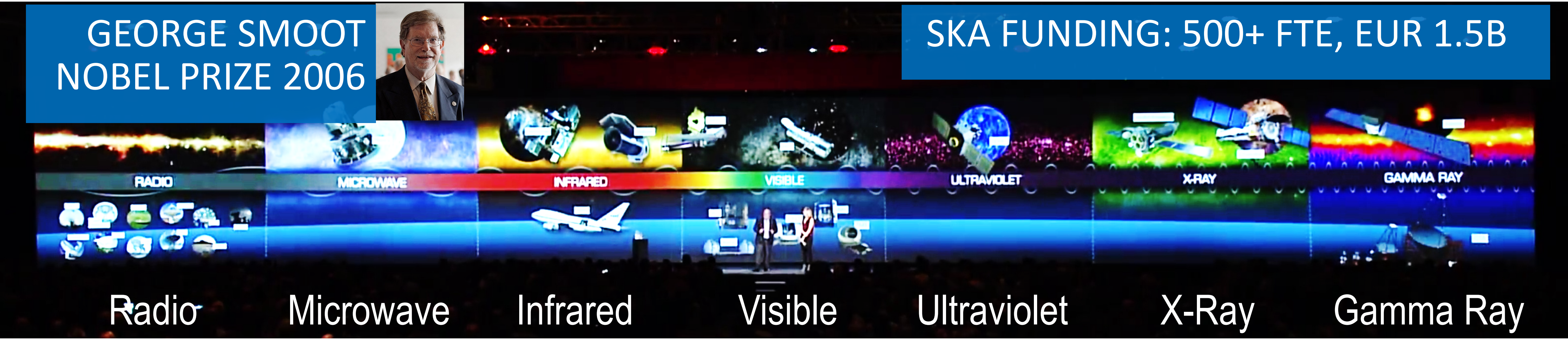
Phil Diamond and Rosie Bolton, Life, the Universe & Computing: The story of the SKA Telescope, SC17 Keynote. [Online]

DISCOVERY = LARGE-SCALE, LONG-TERM STUDY

UNCOVERING THE MYSTERIES OF OUR PHYSICAL UNIVERSE

GEORGE SMOOT
NOBEL PRIZE 2006

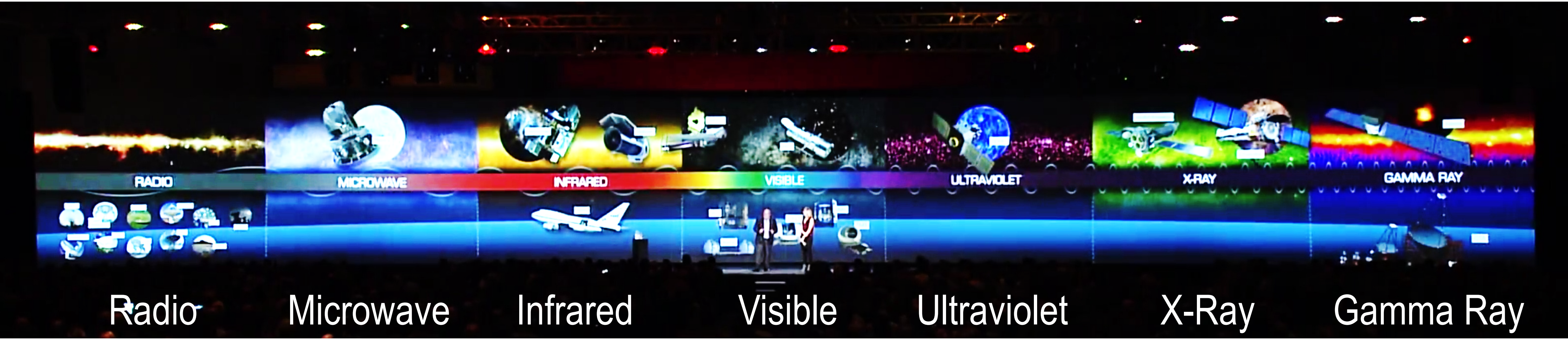
SKA FUNDING: 500+ FTE, EUR 1.5B



James Cordes, The Square Kilometer Array, Project Description, 2009 [Online]
The Square Kilometer Array Factsheet, How much will it cost?, 2012 [Online]
Phil Diamond and Rosie Bolton, Life, the Universe & Computing: The story of the SKA Telescope, SC17 Keynote. [Online]

DISCOVERY = LARGE-SCALE, LONG-TERM STUDY

UNCOVERING THE MYSTERIES OF OUR UNIVERSE, PHYSICAL AND DIGITAL



Radio	Microwave	Infrared	Visible	Ultraviolet	X-Ray	Gamma Ray
Cloud, Grid, Edge, Fog, etc.	One aspect: BigData, P2P	Sci.&Eng. Apps+Sys.	Consumer Apps+Sys.	Enterprise Sys.	Systems, Ecosystems	Performance, Availability, etc.
	P2P T.A.	BoTs, Groups, Workflows	Game Trace Archive	Business -Critical		
[Iosup et al. FGCS'08]	[Zhang et al. CoNext'10]	[Iosup et al. IEEE IC'11]	[Guo et al. NETGAMES'12]	[Shen et al. CCGRID'15]	[Ghiț et al. CCGRID'14]	[Iosup et al. CCGRID'10]

THE DISTRIBUTED SYSTEMS MEMEX

Bush (1945) As we may think. The Atlantic, Jul 1945.

UNCOVERING THE MYSTERIES OF OUR UNIVERSE, PHYSICAL AND DIGITAL

Find and eradicate performance issues

Quantitative evidence

Enable new designs and automation

Cultural and ethical concerns

Understand how entire ecosystems behave and evolve

Cloud, Grid, Edge, Fog, etc.

[Iosup et al. FGCS'08]

One aspect: BigData, P2P

P2P T.A.

[Zhang et al. CoNext'10]

Sci.&Eng. Apps+Sys.

Groups, BoTs, Workflows

[Versluis et al.'19]

[Iosup et al. IEEE IC'11]

Consumer Apps+Sys.

Game Trace Archive

[Guo et al. NETGAMES'12]

Enterprise Sys.

Business-Critical

[Shen et al. CCGRID'15]

Systems, Ecosystems

[Ghiț et al. CCGRID'14]

Performance, Availability, etc.

[Iosup et al. CCGRID'10]

THE WORKFLOW TRACE ARCHIVE

96 traces

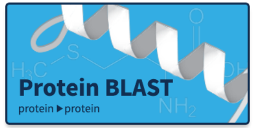
Laurens Versluis

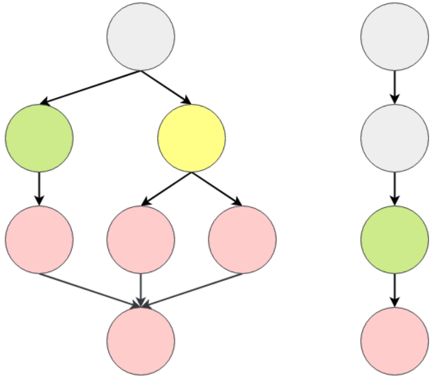
METADATA AND TRACES FOR YOUR WORKFLOW SYSTEMS

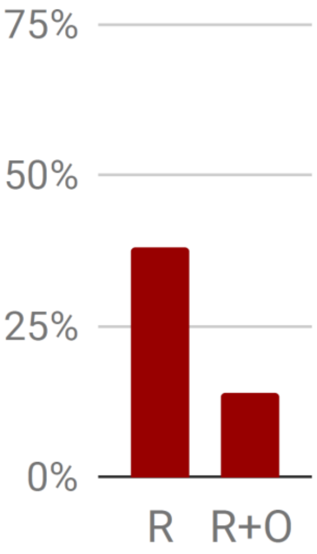
WORKFLOWS ARE COMMON IN MANY DOMAINS

EXCEPT IN SCI., DESIGN, & ENG.

THE WORKFLOW TRACE ARCHIVE CORRECTS THIS





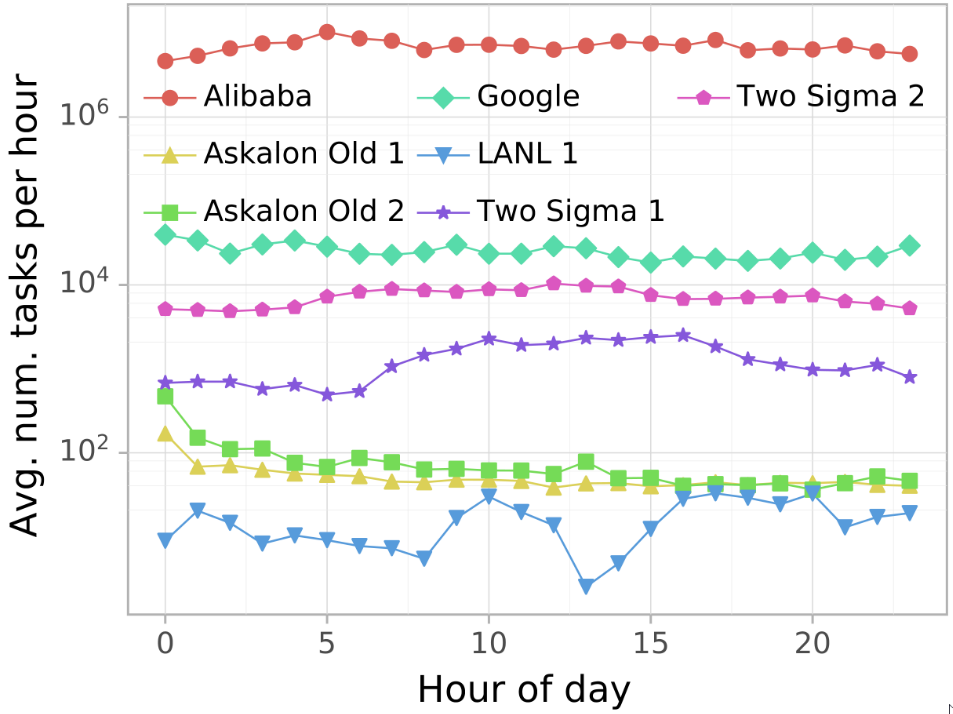


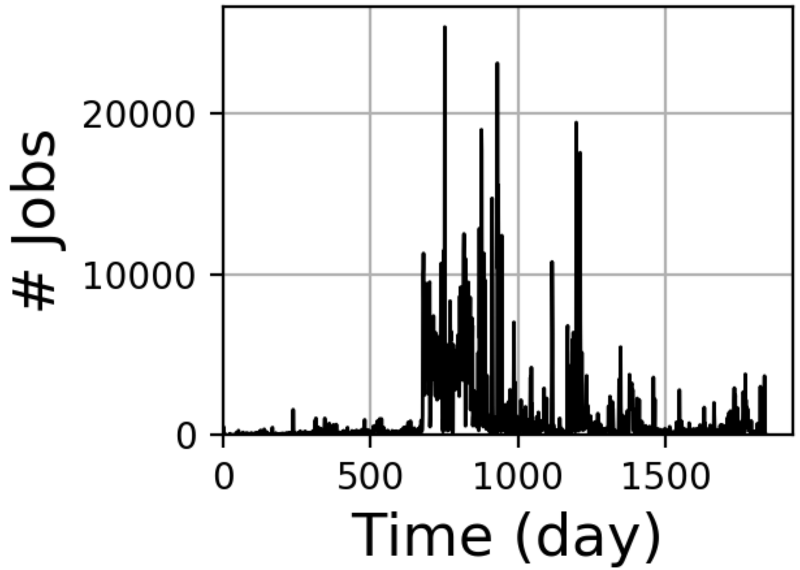
[Versluis et al. The Workflow Trace Archive] Tech.rep.: http://arxiv.org/pdf/1906.07471

http:// wta.atlarge.science

Feature 3: Detailed Statistics

- Generic Statistics: #Workflows, #Tasks, Memory & CPU info, etc.
- Job-level Statistics: Job Arrival, CDF of Job Critical Path, etc.
- Task-level Statistics: Task Arrival graphs, CDFs of properties, etc.





http:// wta.atlarge.science

@Large Research
Massivizing Computer Systems

VU VRIJE UNIVERSITEIT AMSTERDAM

Summary: Ecosystems exhibit short- and long-term dynamics

1. Many facets of workloads (input), processing (system), and performance (output) → many are less understood than they should
2. The Distributed System Memex tries to capture these facets, one archive at a time
3. Already existing archives*:
   - The Grid Workloads Archive (established 2006)
   - The Failure Trace Archive (est. 2009)
   - The P2P Trace Archive (est. 2010)
   - The Game Trace Archive (est. 2012)
   - The Workflow Trace Archive (est. 2019)

* Also several other archives established by various USENIX members: networking, failures, etc. Plus more archives. No Memex yet, though.

Idea:
Meaningful discovery requires understanding the cloud universe is based on ecosystems.

Ecosystem vs. systems, a primer

- Structure: composites of smaller assemblies, but for ecosystems some constituents are produced elsewhere, with different practices
- Operation: ecosystems exhibit many unknown phenomena, less understood dynamics, and socio-technical issues
- Lifecycle: some ecosystem constituents will perish, or be replaced with others that may not actually fulfill the needs

MEANINGFUL DISCOVERY IN DISTRIBUTED ECOSYSTEMS

UNCOVERING THE MYSTERIES OF OUR UNIVERSE, PHYSICAL AND DIGITAL

BUT … WHY WOULD YOU NEED TO UNCOVER AN ARTIFICIAL UNIVERSE?! YOU BUILT IT!

Cloud, Grid, Edge, Fog, etc.	One aspect: BigData, P2P	Sci.&Eng. Apps+Sys.	Consumer Apps+Sys.	Enterprise Sys.	Systems, Ecosystems	Performance, Availability, etc.
	P2P T.A.	BoTs, Groups, Workflows	Game Trace Archive	Business -Critical		
[Iosup et al. FGCS'08]	[Zhang et al. CoNext'10]	[Iosup et al. IEEE IC'11]	[Guo et al. NETGAMES'12]	[Shen et al. CCGRID'15]	[Ghiț et al. CCGRID'14]	[Iosup et al. CCGRID'10]

UNKNOWN PHENOMENA: INTER-, ADAPT-, EXAPTATION

UNCOVERING THE MYSTERIES OF OUR UNIVERSE, PHYSICAL AND DIGITAL

SOME OF OUR DISCOVERIES

BOTS, NOT PARALLEL JOBS	GROUPS NOT RARE, DOMINANT	COMMUNITY FORMATION	SYSTEMIC VARIABILITY	CORRELATED, NOT IID FAILURES
…	…	…		…

Cloud, Grid, Edge, Fog, etc. | One aspect: BigData, P2P | Sci.&Eng. Apps+Sys. | Consumer Apps+Sys. | Enterprise Sys. | Systems, Ecosystems | Performance, Availability, etc.

[Iosup et al. FGCS'08]

P2P T.A.

[Zhang et al. CoNext'10]

BoTs, Groups, Workflows

[Iosup et al. IEEE IC'11]

Game Trace Archive

[Guo et al. NETGAMES'12]

Business -Critical

[Shen et al. CCGRID'15]

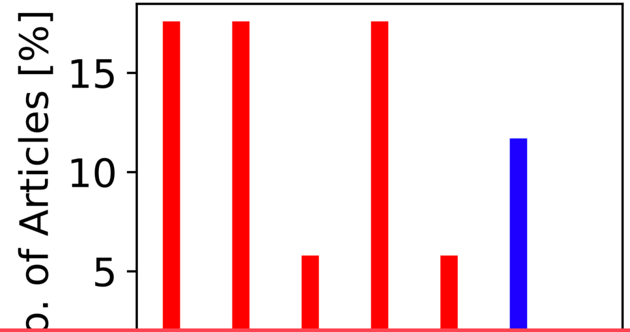
[Ghiț et al. CCGRID'14]

[Iosup et al. CCGRID'10]

Variability is disconsidered in real-world systems

Main findings re. 44 articles*, SC/NSDI/OSDI/SOSP 2010-2018:

Most articles report 3-10 repetitions, few report > 10



Can we trust these evaluations?

Big reproducibility problem in general!

median

What did those articles report?

Cited articles > 11,000 citations !!!

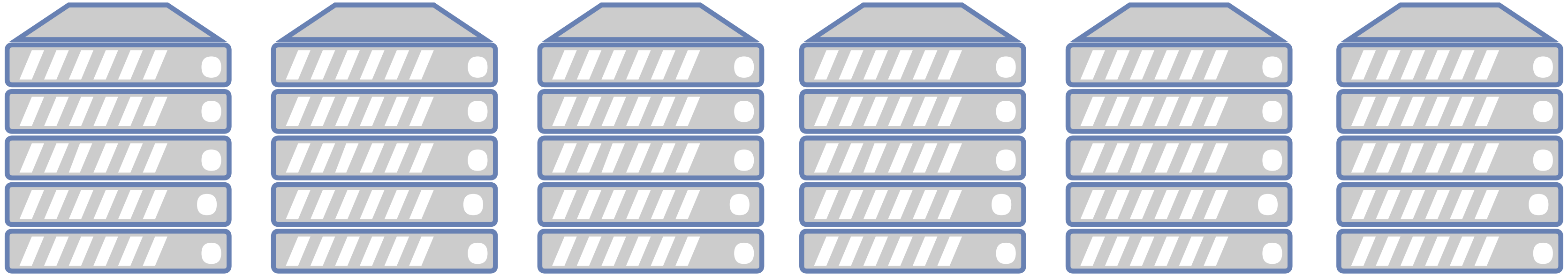
Uta et al. Is Big Data Performance Reproducible In Modern Cloud Networks? USENIX NSDI'20. Tech.rep.: https://arxiv.org/pdf/1912.09256

Q: How to Check? A: Through Benchmarking!

Batch Processing	Stream Processing	SQL-like Analytics	Graph Processing	Machine Learning

Big Data Processing Platform (Spark)

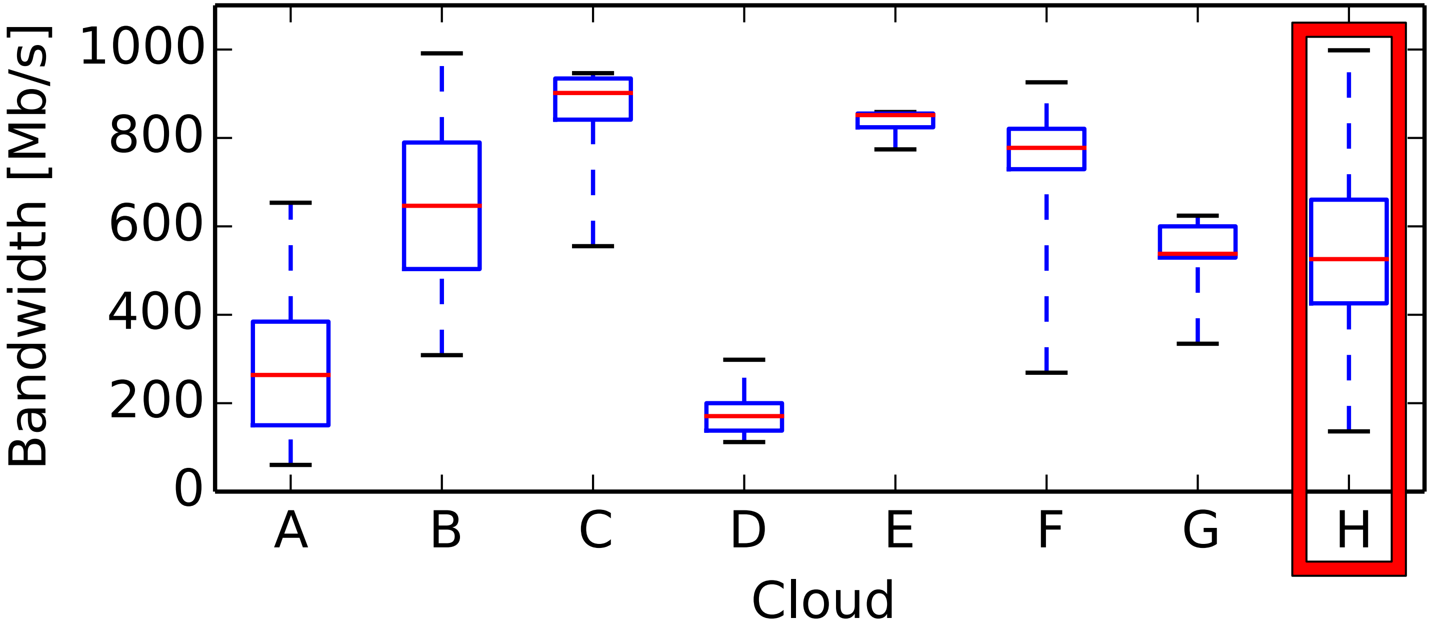
Resource Variability Emulation Layer (Network, CPU, Disk)

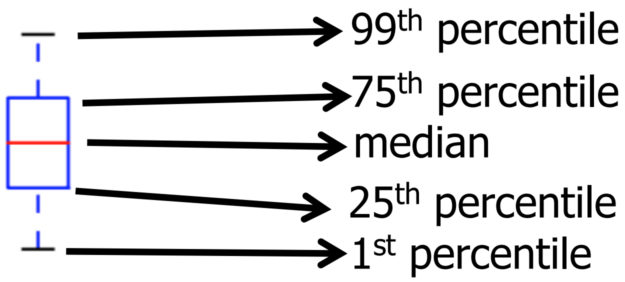


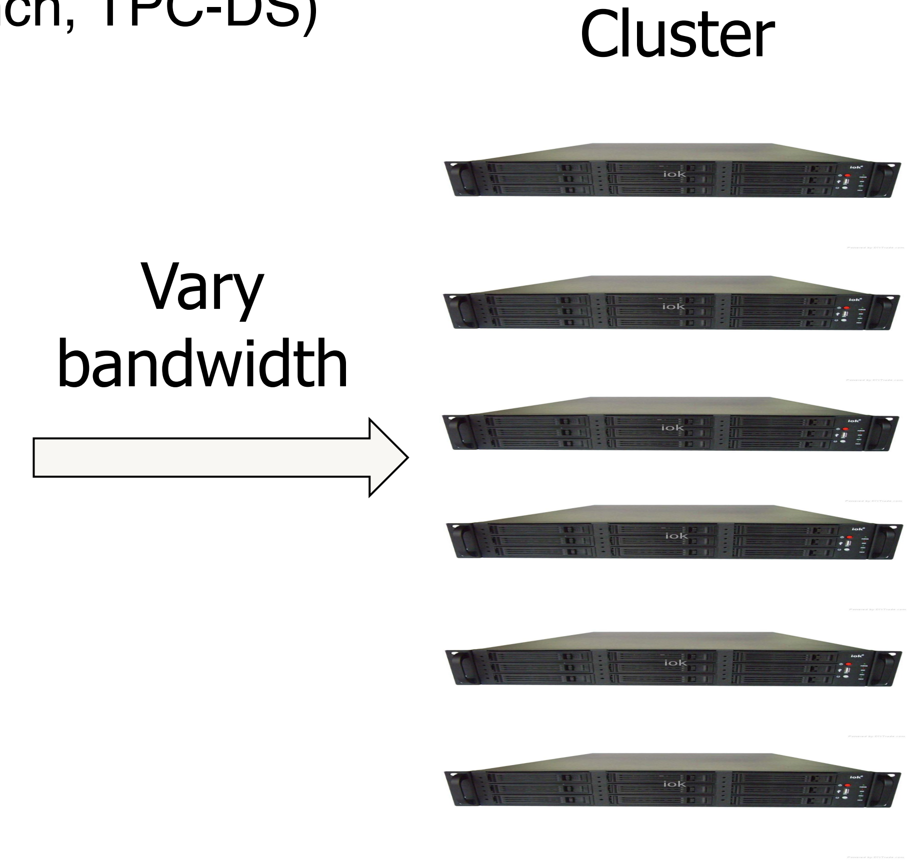
Uta et al. Is Big Data Performance Reproducible In Modern Cloud Networks? USENIX NSDI’20. Tech.rep.: https://arxiv.org/pdf/1912.09256

Q: How to check? A: Traces Emulation.

Systematic study using A-H cloud bandwidth distributions. For each:
Run a series of big data applications/benchmarks (HiBench, TPC-DS)
The distributions are:

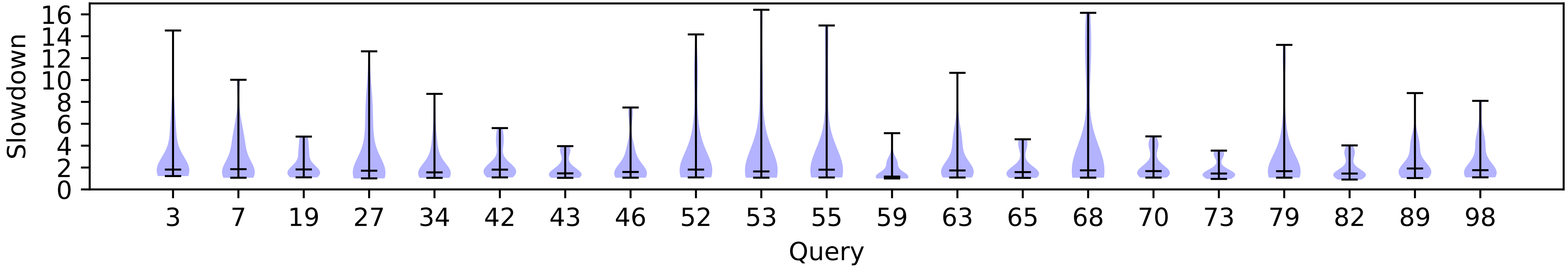






Uta et al. Is Big Data Performance Reproducible In Modern Cloud Networks? USENIX NSDI'20. Tech.rep.: https://arxiv.org/pdf/1912.09256

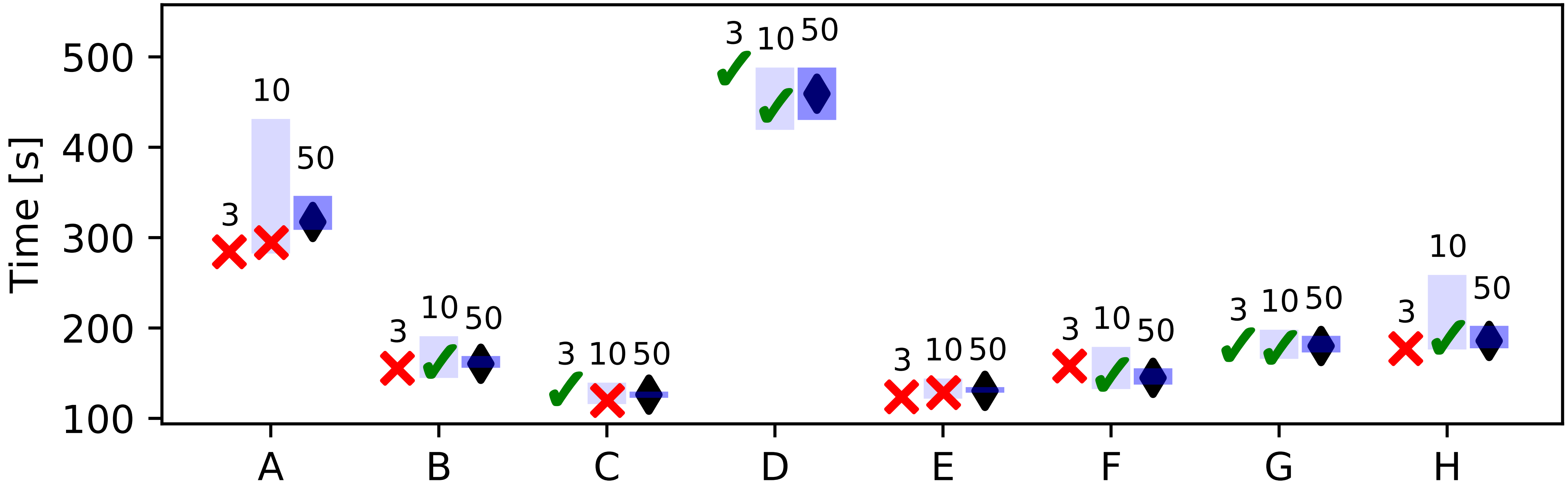
Large Variable Slowdowns – TPC-DS

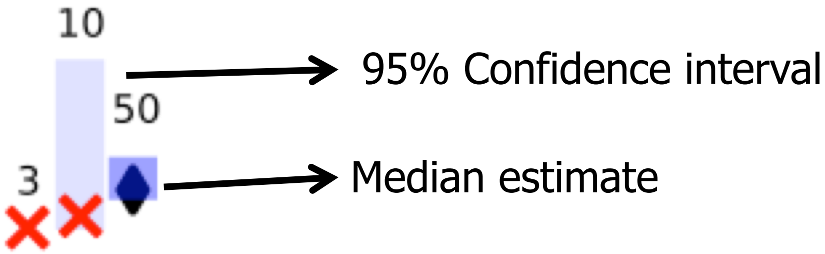


Large amounts of slowdown!
3-10 repetitions anywhere in this range!

Uta et al. Is Big Data Performance Reproducible In Modern Cloud Networks? USENIX NSDI'20. Tech.rep.: https://arxiv.org/pdf/1912.09256

Number of trials – Estimating Median Performance

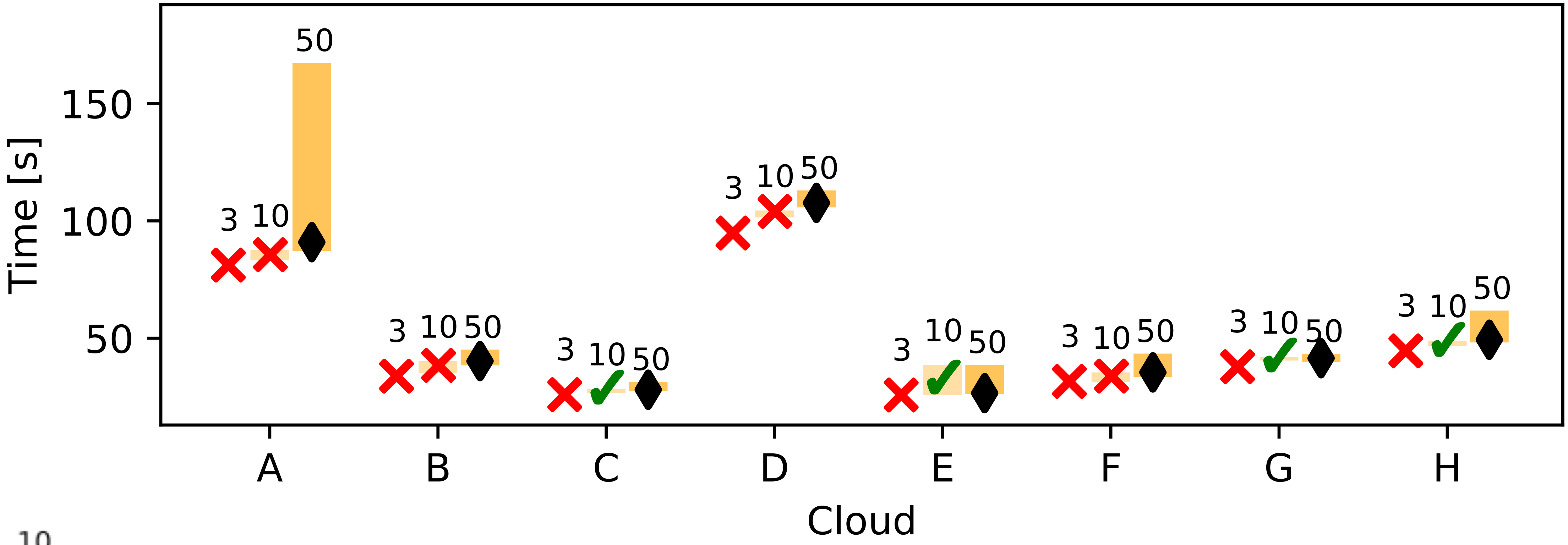




- 50 trials, computed CI and median
- For 3 and 10 runs:
- X = median estimate is **not** within CI for 50 trials
- V = median estimate is within CI for 50 trials

Uta et al. Is Big Data Performance Reproducible In Modern Cloud Networks? USENIX NSDI'20. Tech.rep.: https://arxiv.org/pdf/1912.09256

Number of trials – Estimating Tail Performance



3-10 repetitions not good enough!

Uta et al. Is Big Data Performance Reproducible In Modern Cloud Networks? USENIX NSDI'20. Tech.rep.: https://arxiv.org/pdf/1912.09256

Summary: Ecosystems exhibit many unknown phenomena

1. (Network) performance variability is a **widespread** cloud phenomenon

   Iosup et al. On the Performance Variability of Production Cloud Services. CCGRID 2011. Tech.rep.: http://www.st.ewi.tudelft.nl/iosup/tech_rep/cloud-perf-var10tr.pdf

   Uta et al. Is Big Data Performance Reproducible In Modern Cloud Networks? USENIX NSDI'20. Tech.rep.: https://arxiv.org/pdf/1912.09256

2. Systems community **neglects** performance variability
3. Network performance variability due to: resource sharing, provider QoS
4. High impact in result reporting, experiment design, replication
5. An ongoing reproducibility problem in computer systems

Idea:
Get meaningful, reproducible results.
Avoid the reproducibility wars.

Note: many types of reproducibility. Won’t go into the technical details of the semantics of reproducibility.

REPRODUCIBILITY AND VALIDATION OF DISCOVERY

A PERENNIALLY TOUGH PROBLEM, IN COMPUTING BUT ALSO IN ALL OTHER SCIENCES

METHODOLOGY

OPEN SCIENCE

REPORTING & DISSEMINATION

REPRODUCIBILITY

* Conferences do not accept such material… except when they do…

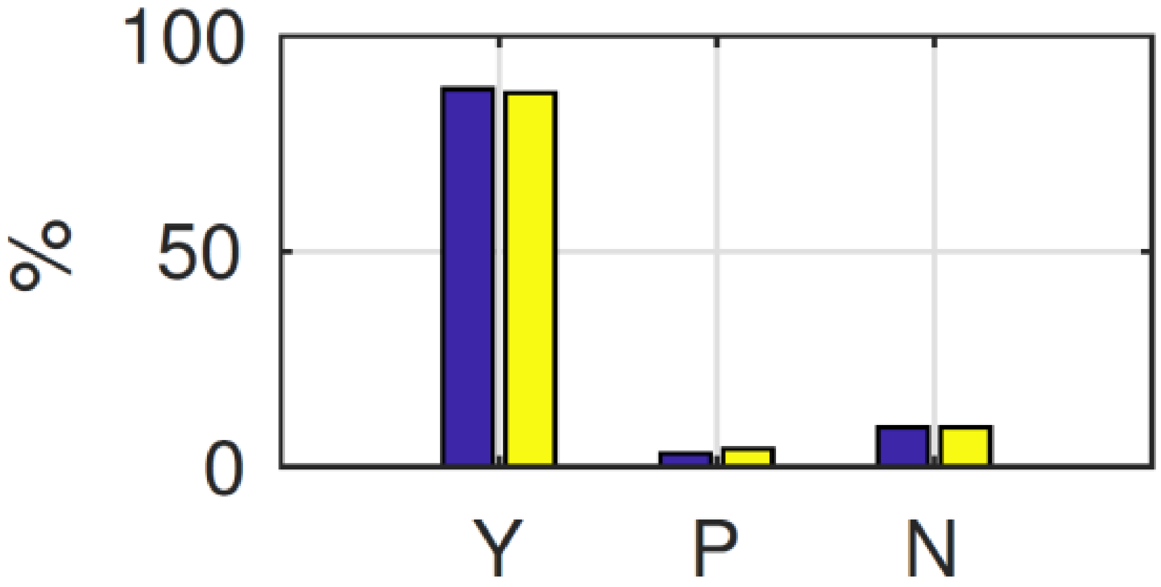
Munafò et al., A manifesto for reproducible science, Nature Human Behaviour, Jan 2017. [Online]

REPRODUCIBILITY AND VALIDATION OF DISCOVERY

A PERENNIALLY TOUGH PROBLEM, IN COMPUTING BUT ALSO IN ALL OTHER SCIENCES

- METHODOLOGY
- OPEN SCIENCE
- REPORTING & DISSEMINATION
- REPRODUCIBILITY

P7: Measurement units. For all the reported quantities, report the corresponding unit of measurement.



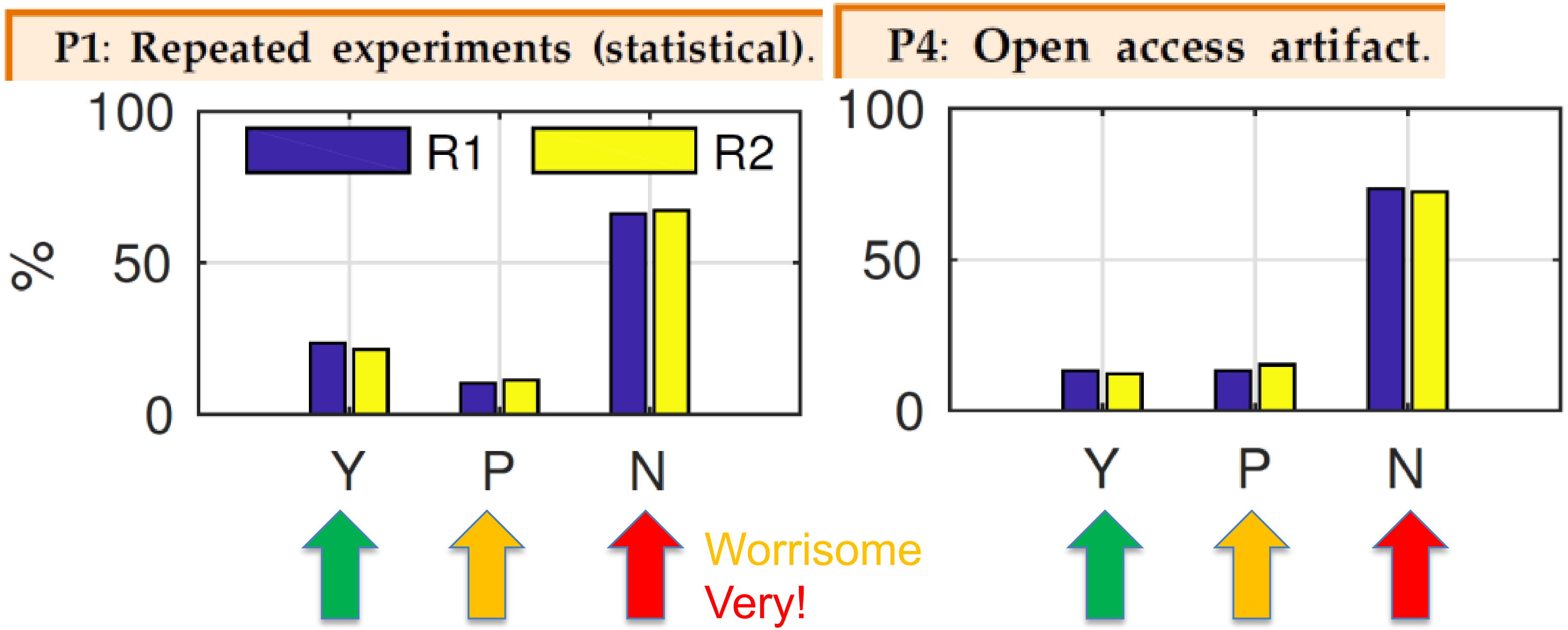
* Conferences do not accept such material… except when they do…

Papadopoulos et al., Methodological Principles for Reproducible Performance Evaluation in Cloud Computing. IEEE TSE 2019/ICSE Journal-First 2020 [Online]

REPRODUCIBILITY AND VALIDATION OF DISCOVERY

A PERENNIALLY TOUGH PROBLEM, IN COMPUTING BUT ALSO IN ALL OTHER SCIENCES

- METHODOLOGY
- OPEN SCIENCE
- REPORTING & DISSEMINATION
- REPRODUCIBILITY



* Conferences do not accept such material… except when they do…

Papadopoulos et al., Methodological Principles for Reproducible Performance Evaluation in Cloud Computing. IEEE TSE 2019/ICSE Journal-First 2020 [Online]

REPRODUCIBILITY AND VALIDATION OF DISCOVERY

AVOID REPRODUCIBILITY WARS. WE ALREADY SETTLED SW. TESTING VS. VERIFICATION

- METHODOLOGY
- OPEN SCIENCE
- REPORTING & DISSEMINATION
- REPRODUCIBILITY

Remember the 1960s’ debate on software testing vs. verification?
→
Apply basic guidelines, not strict rules

* Conferences do not accept such material… except when they do…

Papadopoulos et al., Methodological Principles for Reproducible Performance Evaluation in Cloud Computing. IEEE TSE 2019/ICSE Journal-First 2020 [Online]

CLOUD PERFORMANCE ~ ECOSYSTEMS

PART OF THE LARGER VISION OF MASSIVIZING COMPUTER SYSTEMS

- Golden Age of Cloud Ecosystems ... Yet many challenges
  1. Experimental Science, Design, and Engineering
  2. Reproducibility through guidelines, not strict
  3. Reference architectures can conquer complexity
  4. Distributed Systems Memex (Archives)
  5. Phenomena: performance variability, etc.
  6. Grand observations and experiments
  7. [Extra] In clouds, nobody can hear you scale

Many thanks to 200+ collaborators

VU

@Large Research
Massivizing Computer Systems

http://atlarge.science

MASSIVIZING COMPUTER SYSTEMS

FURTHER READING

https://atlarge-research.com/publications.html

1. Iosup et al. The AtLarge Vision on the Design of Distributed Systems and Ecosystems. ICDCS 2019 ← Start here
2. Uta et al. Is big data performance reproducible in modern cloud networks? NSDI 2020
3. Van Eyk et al. The SPEC-RG Reference Architecture for FaaS: From Microservices and Containers to Serverless Platforms, IEEE IC 2019
4. Papdopoulos et al. Methodological Principles for Reproducible Performance Evaluation in Cloud Computing. TSE 2019 and (journal-first) ICSE 2020
5. van Beek et al. Portfolio Scheduling for Managing Operational and Disaster-Recovery Risks in Virtualized Datacenters Hosting Business-Critical Workloads. ISPDC 2019
6. van Beek et al. A CPU Contention Predictor for Business-Critical Workloads in Cloud Datacenters. HotCloudPerf19

\+ Iyushkin et al. Performance-Feedback Autoscaling with Budget Constraints for Cloud-based Workloads of Workflows. Under submission

Etc.

MASSIVIZING COMPUTER SYSTEMS

FURTHER READING

https://atlarge-research.com/publications.html

1. Iosup et al. Massivizing Computer Systems. ICDCS 2018 ← start here
2. Andreadis et al. A Reference Architecture for Datacenter Scheduling, SC18
3. Van Eyk et al. Serverless is More: From PaaS to Present Cloud Computing, IEEE IC Sep/Oct 2018
4. Uta et al. Exploring HPC and Big Data Convergence: A Graph Processing Study on Intel Knights Landing, IEEE Cluster 2018

5. Talluri et al. Big Data Storage Workload in the Cloud. ACM/SPEC ICPE 2019.
6. Toader et al. Graphless. IEEE ISPDC'19.
7. Jiang et al. Mirror. CCPE 2018.
8. Ilyushkin et al. Autoscalers. TOMPECS 2018.
9. Versluis et al. Autoscaling Workflows. CCGRID'18.
10. Uta et al. Elasticity in Graph Analytics? IEEE Cluster 2018.

11. Herbst et al. Ready for rain? TOMPECS 2018.
12. Guo et al. Streaming Graph-partitioning. JPDC'18.
13. Iosup et al. The OpenDC Vision. ISPDC'17.
14. Iosup et al. Self-Aware Computing Systems book.
15. Iosup et al. LDBC Graphalytics. PVLDB 2016.

Etc.

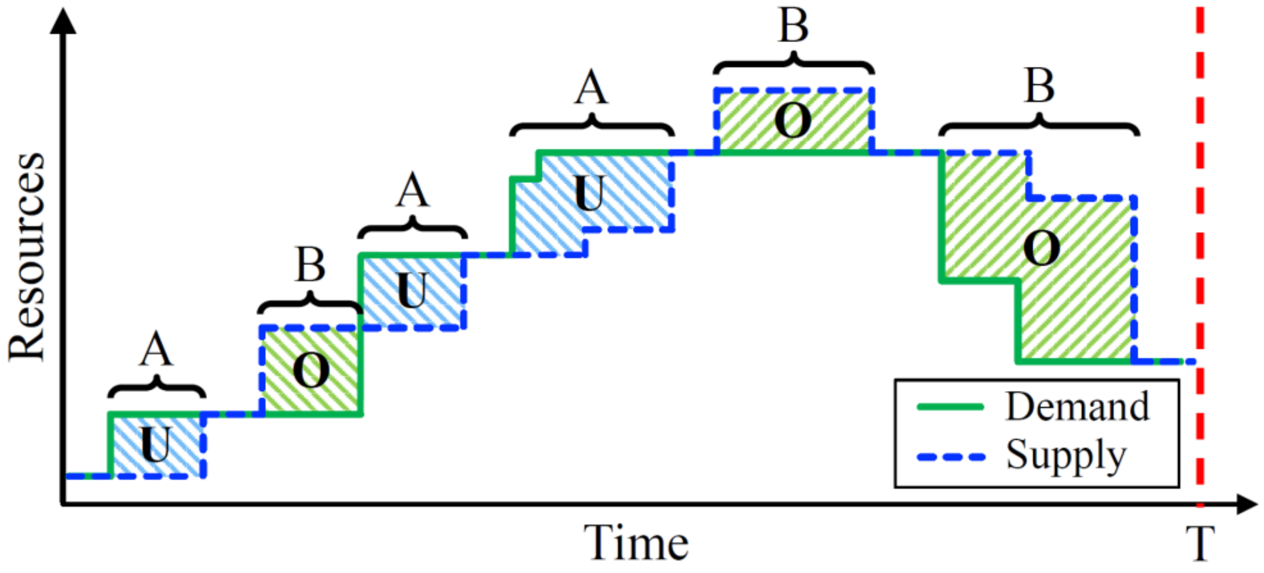
[EXTRA] Idea:
In clouds, nobody can hear you scale.
Unless you're scaling elastically.

- Scalability ~ system ability to do more with more resources
  - Strong and weak scalability for important (HPC) parallel applications
  - Also good for vanity projects
- Elasticity ~ … more with proportionally more resources
  - Sharing of resources for multi-tenancy
  - Partitioning distributed resources according to application phases
  - Ethical science and engineering

1

Elasticity: New Metrics for the Cloud World

= Degree to which a system adapts to workload changes by provisioning and de-provisioning resources autonomically, s.t. the supply (the provisioned resources) matches the demand



Zone type A:
D > S ~ Underprovisioning

Zone type B:
D < S ~ Overprovisioning

A. Ilyushkin, A. Ali-Eldin, N. Herbst, A. Bauer, A. V. Papadopoulos, D. H. J. Epema, A. Iosup (2018)
An Experimental Performance Evaluation of Autoscalers for Complex Workflows. TOMPECS 3(2).

>

1

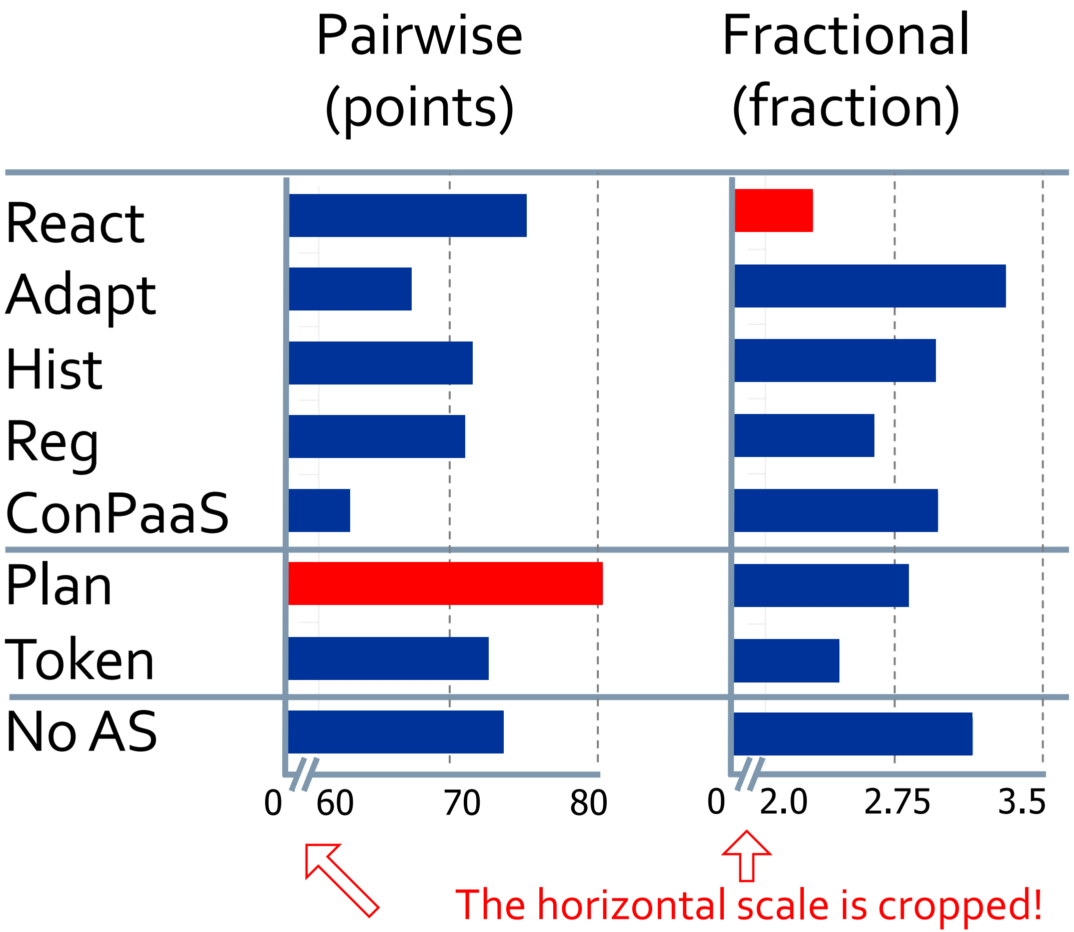
More Cloud Metrics: Elasticity, Isolation, Risk, …

= many new metrics trying to capture cloud operations

Quality Attribute	Metric		Value Range	Unit	How to Measure	Show Case
Elasticity	Accuracy	θ_U	[0; ∞), opt: 0	%	ex post, calibration required	✓ Sec. 3.8 [7, 37]
		θ_O	[0; ∞), opt: 0	%		
		θ'_U	[0; 100], opt: 0	%		
		θ'_O	[0; 100], opt: 0	%		
	Timeshare	τ_U	[0; ∞), opt: 0	%		
		τ_O	[0; ∞), opt: 0	%		
	Instability	υ	[0; 100], opt: 0	%		
	Deviation	σ	[0; ∞), opt: 0	%	ex post	
	Speedup	ϵ_k	[0; ∞), opt: ∞	None	ex post	
Perf. Isolation	QoS	I_{QoS}	[0; ∞], opt: 0	None	ex post	✓Sec. 4.2 [45]
Perf. Variability	Deviation	$PVDC$	[0; 100], opt: 0	%	ex post	✓Sec. 4.3 [21]
Availability	Adherence	S_a	[0; 100], opt: 100	%	ex post	✓Sec. 5.3
	Strictness	S_s	[0; ∞], opt: ∞	%	der. from SLO def.	appl. on pub. clouds[5]
Operational Risk	Provision	r_p	[−1; 1], opt: 0	None	ex post	✓ Sec. 6.5
	Contention	r_c	[0; 1], opt: 0	None		
	Service	r_e	[0; 1], opt: 0	None		
	System	r_s	[0; 1], opt: 0	None		

N. Herbst, A. Bauer, S. Kounev, G. Oikonomou, E. Van Eyk, G. Kousiouris, A. Evangelinou, R. Krebs, T. Brecht, C. L. Abad, A. Iosup: Quantifying Cloud Performance and Dependability: Taxonomy, Metric Design, and Emerging Challenges. TOMPECS 3(4): 19:1-19:36 (2018)

The State of the Art in Empirical Studies Example: (Autoscaling, Scientific WLs, Cluster)

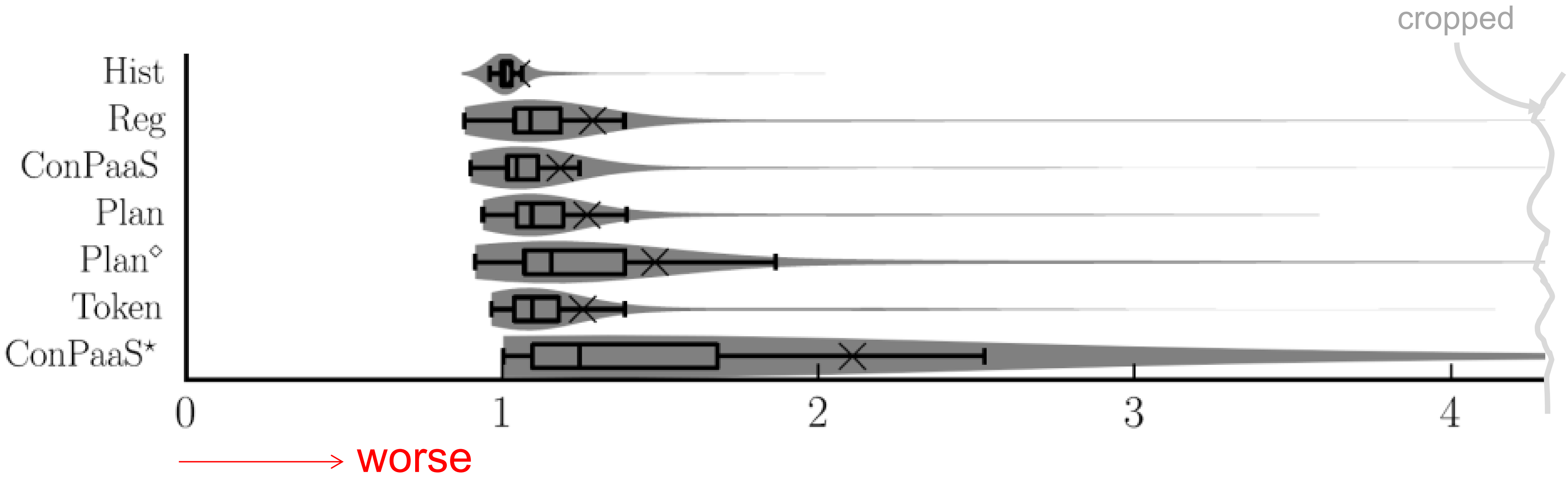


A. Ilyushkin, A. Ali-Eldin, N. Herbst, A. Bauer, A. V. Papadopoulos, D. H. J. Epema, A. Iosup (2018)
An Experimental Performance Evaluation of Autoscalers for Complex Workflows. TOMPECS 3(2).

2

… But Each Approach Can Have Drawbacks…

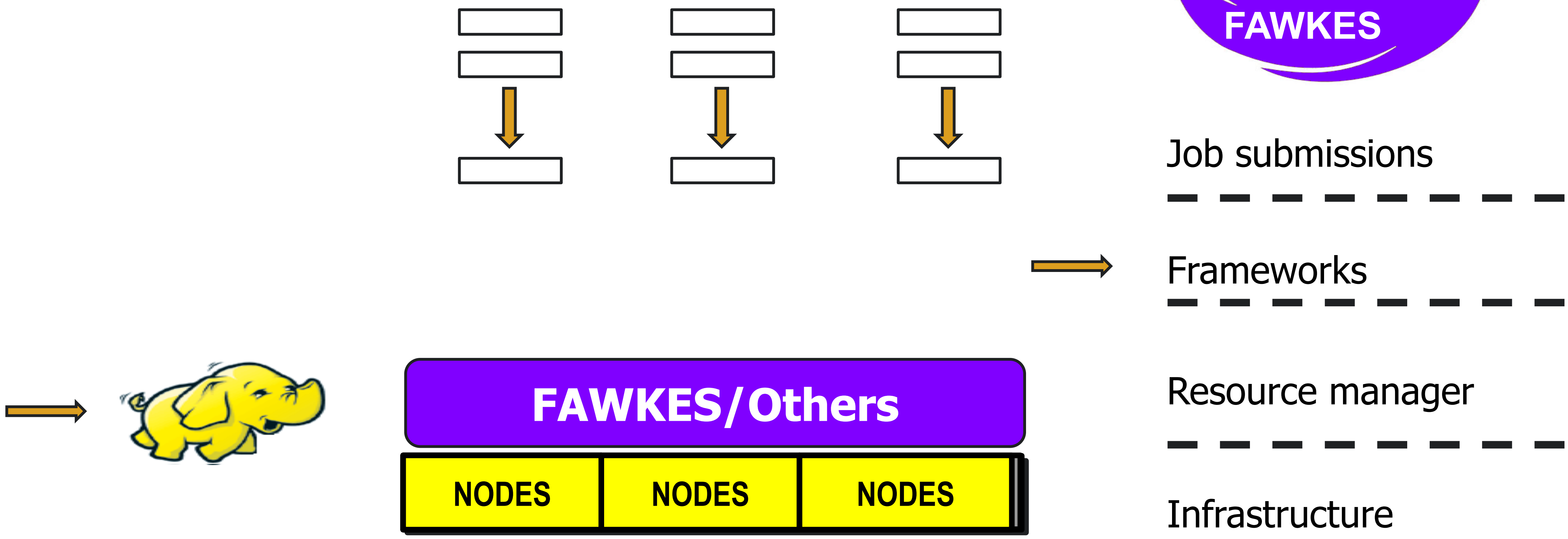
- High tail latency ~ large differences between slowdowns



A. Ilyushkin, A. Ali-Eldin, N. Herbst, A. Bauer, A. V. Papadopoulos, D. H. J. Epema, A. Iosup (2018) An Experimental Performance Evaluation of Autoscalers for Complex Workflows. TOMPECS 3(2).

Dynamic Big Data Processing

Fawkes = Elastic MapReduce



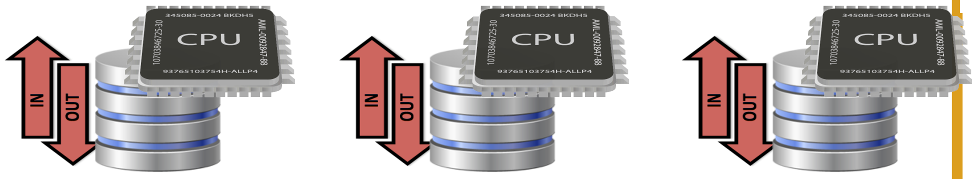
B. Ghit et al. Balanced Resource Allocations Across Multiple Dynamic MapReduce Clusters. SIGMETRICS 2014

End Example→

3

Mechanisms for Elasticity in MapReduce Frameworks

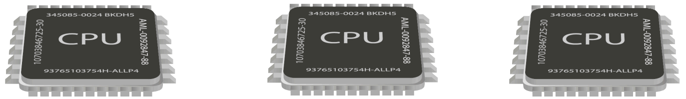
Core nodes



INPUT/OUTPUT DATA

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

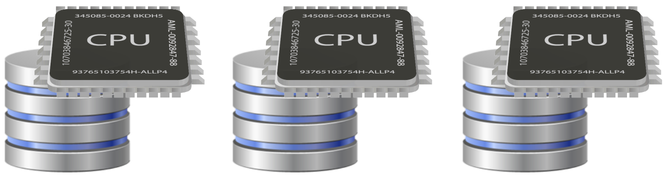
Transient nodes (TR)



NO DATA

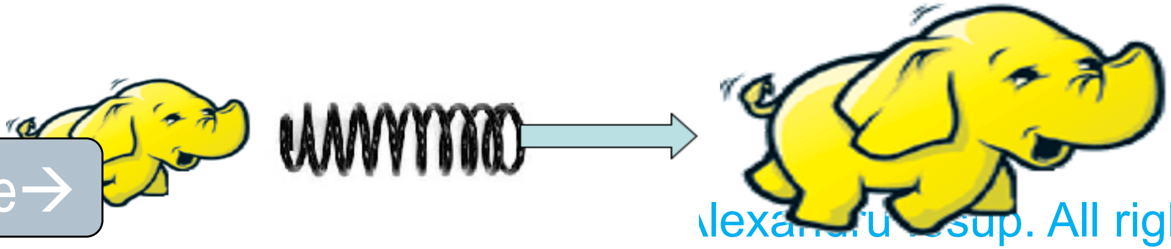
- No local storage
- R/W from/to core nodes
- **Instant removal**

Trans-core nodes (TC)



OUTPUT DATA

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



End Example→

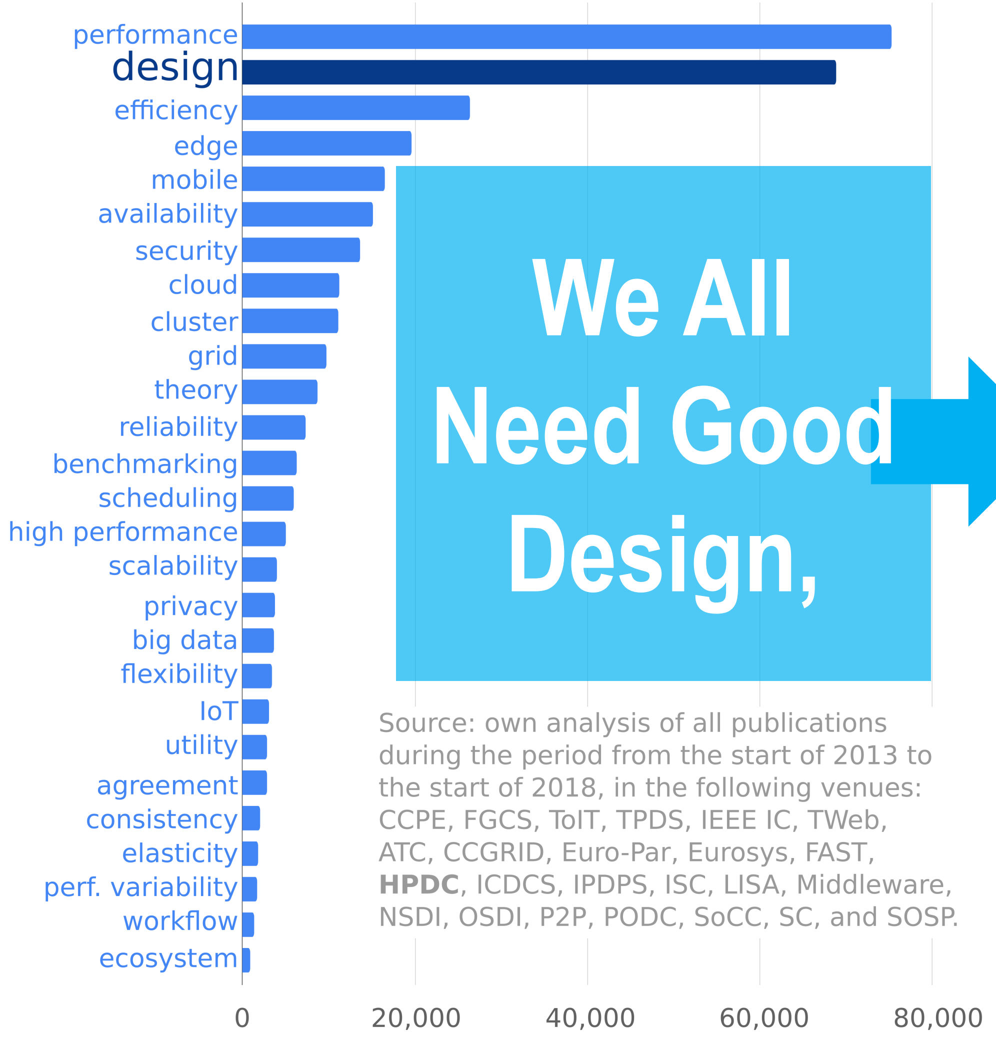
>

[EXTRA] Summary:
Focus on elasticity, not (merely) on scalability

1. Metrics for assessing elasticity
2. When looking at multiple metrics,
   use performance *tournaments* to compare
3. Elasticity raises interesting system-level challenges
4. Lots to discuss – elasticity for graph processing,
   heterogeneous hardware, etc.

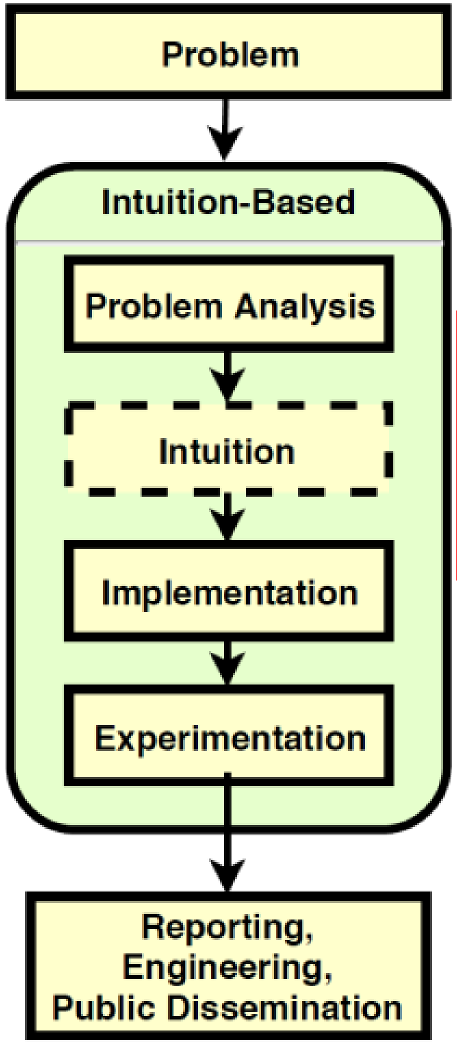
Extra: Have you thought about design?

THE DESIGN OF DISTRIBUTED SYSTEMS AND ECOSYSTEMS



We All Need Good Design,

Source: own analysis of all publications during the period from the start of 2013 to the start of 2018, in the following venues: CCPE, FGCS, ToIT, TPDS, IEEE IC, TWeb, ATC, CCGRID, Euro-Par, Eurosys, FAST, **HPDC**, ICDCS, IPDPS, ISC, LISA, Middleware, NSDI, OSDI, P2P, PODC, SoCC, SC, and SOSP.

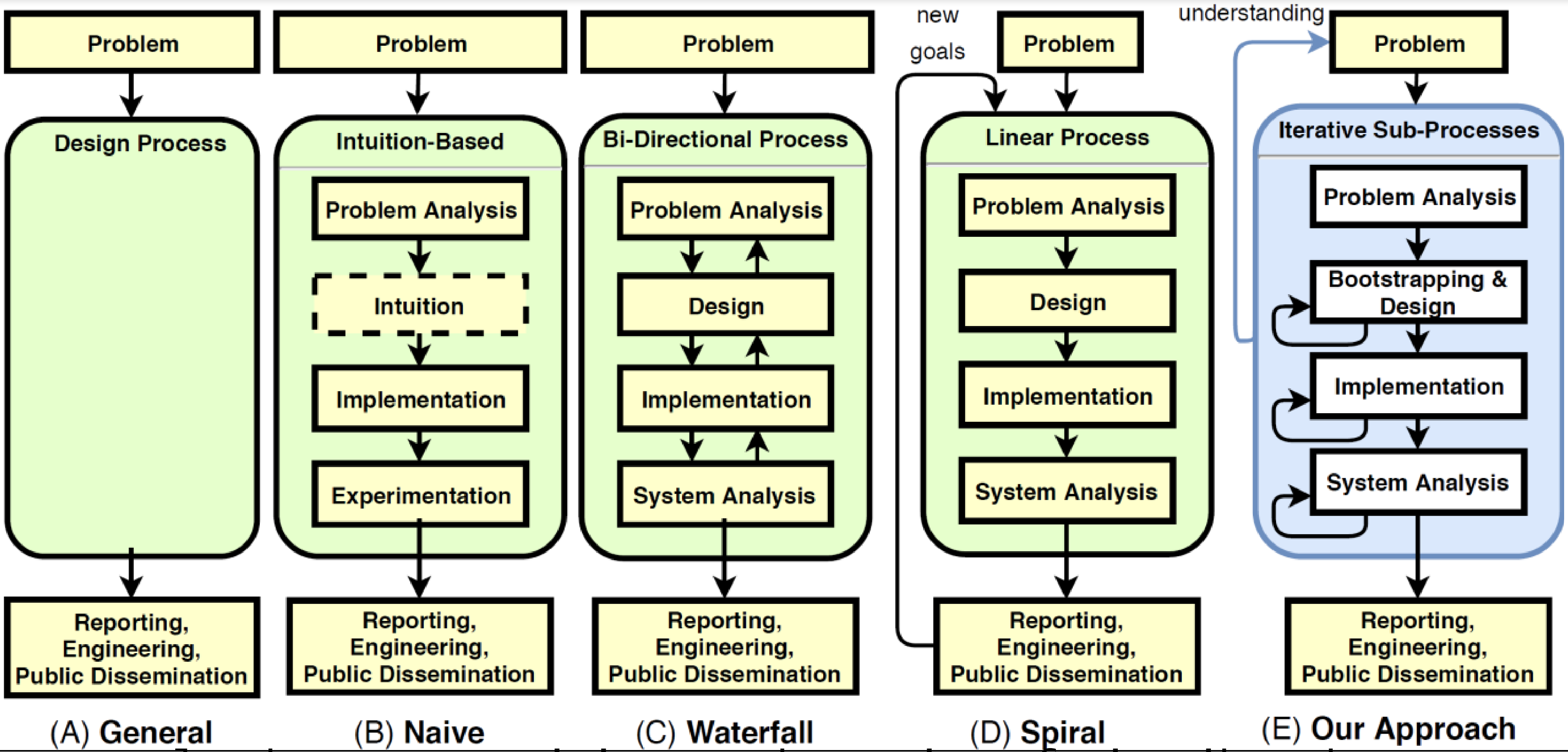


But We Are Not All Steve Jobs

Photo by Matthew Yohe, 2008.

THE ATLARGE DESIGN PROCESS FOR DISTRIBUTED SYSTEMS AND ECOSYSTEMS

bit.ly/AtLargeDesign1Talk



[Iosup et al. The AtLarge Vision on the Design of Distributed Systems and Ecosystems. ICDCS'19] Online version: http://arxiv.org/pdf/1902.05416

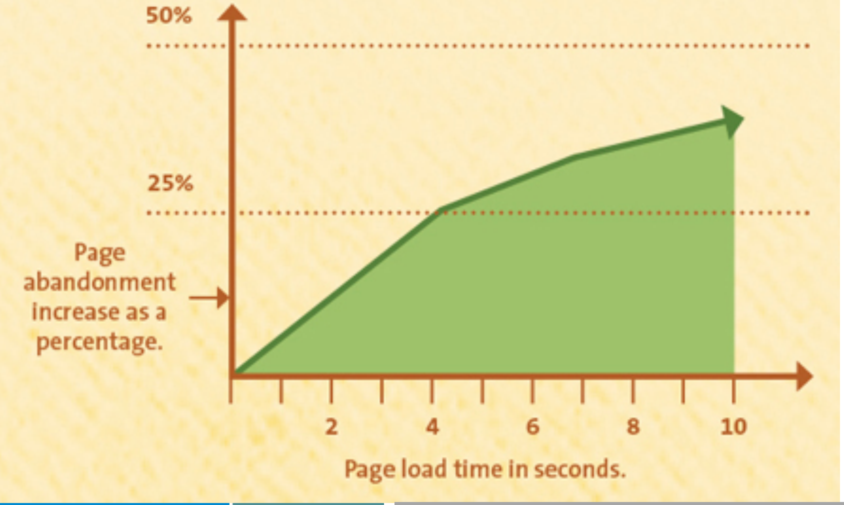
Extra: Some of the challenges, in bright colors

VU VRIJE UNIVERSITEIT AMSTERDAM

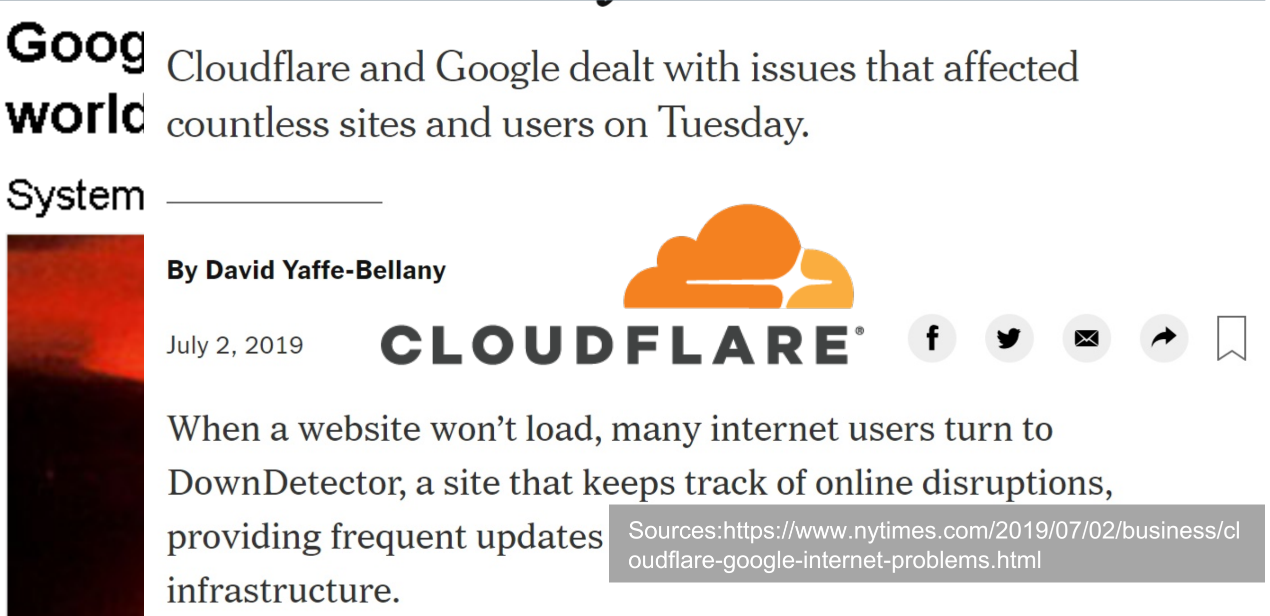
CHALLENGE: MEET SERVICE LEVEL AGREEMENTS

PERFORMANCE, DEPENDABILITY, AND OTHER NON-FUNCTIONAL CHALLENGES

We Cannot Maintain the Ecosystems we Have Built (and Thought We've Tested, and Validated)



Sources: https://www.fastcompany.com/1825005/how-one-second-could-cost-amazon-16-billion-sales


Goog
world
System
Cloudflare and Google dealt with issues that affected countless sites and users on Tuesday.

By David Yaffe-Bellany

July 2, 2019

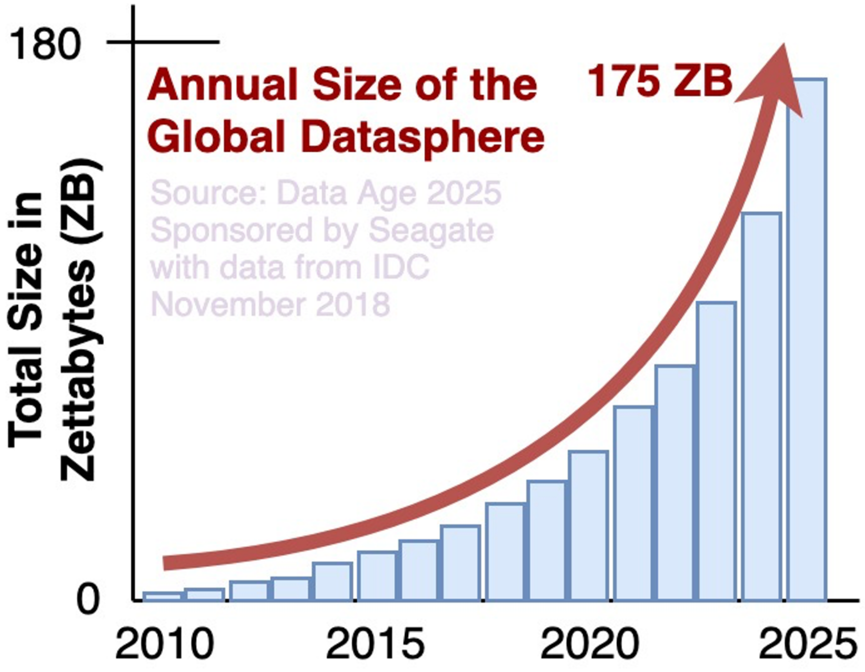
When a website won't load, many internet users turn to DownDetector, a site that keeps track of online disruptions, providing frequent updates ... infrastructure.

Sources:https://www.nytimes.com/2019/07/02/business/cloudflare-google-internet-problems.html

CHALLENGE: SYSTEMATIC DESIGN & DESIGN-SPACE EXPLORATION

THE COMPLEXITY CHALLENGE

We Build and Test
Isolated Computer Systems, Yet
Everything Works in Stacks and Ecosystems
... Need to Reason About Them!



<<1% OF BIG DATA BY MATT TURK (2017)
“SW. IS EATING THE WORLD”

CHALLENGE: EFFICIENCY, SUSTAINABILITY, RESPONSIBILITY!

THE RESOURCE MANAGEMENT CHALLENGE

Based on Jay Walker's recent TED talk.

Need To Be Much More Efficient,

Need to Also Be Ethical, and to Educate Our Clients

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 >3.5 billion views (last update, May 2018).

THIS IS THE GOLDEN AGE OF DISTRIBUTED COMPUTER SYSTEMS

YET WE ARE IN A CRISIS – 5 CORE CHALLENGES

1. Ecosystem ≠ 1 System/Stack
But the Laws and Theories are made for Isolated Computer Systems (or Silos, or Narrow Stacks)

TRADITIONAL DISTRIBUTED SYSTEMS COURSES TEACH YOU ALL ABOUT THIS

2. Need to Understand **How to Maintain Ecosystems**

3. Need to Understand **How to Make Ecosystems Automated, Efficient (Smarter)**

4. **Beyond Tech**: How to Be Ethical, Socially Useful?

5. Need to **Address the Peopleware Problems**

Extra: References for the Distributed Systems Memex

The Distributed Systems Memex: References

Bush (1945) As we may think. The Atlantic, Jul 1945.

First idea recorded publicly: Iosup (2012) Towards logging and preserving the entire history of distributed systems. Is the Future of Preservation Cloudy? (Dagstuhl Seminar 12472), pp. 126–127.

Prototypes of the idea:

[Guo et al. NETGAMES'12] Yong Guo, Alexandru Iosup: The Game Trace Archive. NetGames 2012: 1-6

[Iosup et al. FGCS'08] Alexandru Iosup, Hui Li, Mathieu Jan, Shanny Anoep, Catalin Dumitrescu, Lex Wolters, Dick H. J. Epema: The Grid Workloads Archive. Future Generation Comp. Syst. 24(7): 672-686 (2008). Highly cited.

[Iosup et al. CCGRID'10] Derrick Kondo, Bahman Javadi, Alexandru Iosup, Dick H. J. Epema: The Failure Trace Archive: Enabling Comparative Analysis of Failures in Diverse Distributed Systems. CCGRID 2010. Best Paper Award. Highly cited.

[Iosup et al. IEEE IC'11] Alexandru Iosup, Dick H. J. Epema: Grid Computing Workloads. IEEE Internet Computing 15(2) Highly cited.

[Shen et al. CCGRID'15] Siqi Shen, Vincent van Beek, Alexandru Iosup: Statistical Characterization of Business-Critical Workloads Hosted in Cloud Datacenters. CCGRID 2015. Highly cited.

[Versluis et al.'19] Laurens Versluis, Roland Mathá, Sacheendra Talluri, Tim Hegeman, Radu Prodan, Ewa Deelman, Alexandru Iosup: The Workflow Trace Archive: Open-Access Data from Public and Private Computing Infrastructures - Technical Report. CoRR abs/1906.07471 (2019)

[Zhang et al. CoNext'10] Boxun Zhang, Alexandru Iosup, Johan Pouwelse, Dick H. J. Epema: The peer-to-peer trace archive: design and comparative trace analysis. CoNEXT 2010

VRIJE UNIVERSITEIT AMSTERDAM