Self-* Datacenter Management for Business Critical Workloads

Alexandru Iosup
Delft University of Technology
The Netherlands

Team: **Undergrad** Tim Hegeman, … **Grad** Vincent van Beek, Lipu Fei, Yong Guo, Mihai Capota, Bogdan Ghit **Researchers** Marcin Biczak, Otto Visser **Staff** Henk Sips, Dick Epema **Collaborators*** Ana Lucia Varbanescu (UvA, Ams), Claudio Martella (VU, Giraph), Intel Research Labs, IBM TJ Watson, Oracle Research Labs SF, …

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

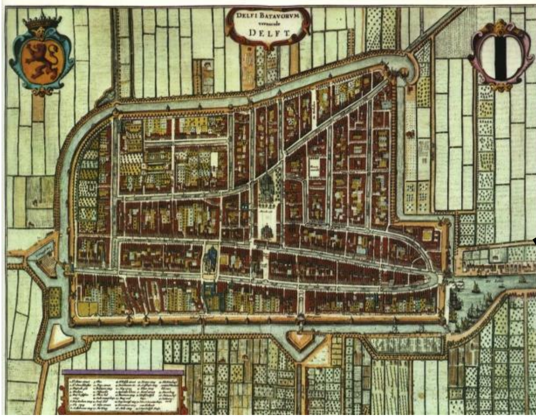
January 19, 2015

Dagstuhl seminar Model-driven Algorithms and Architectures for Self-Aware Computing Systems, Aug 2014

spec Research

TU Delft
Delft University of Technology

(TU) Delft – the Netherlands – Europe



founded 13th century
pop: 100,000

founded 1842
pop: 13,000

Delft

pop: 16.5 M

TU Delft

The Parallel and Distributed Systems Group at TU Delft

VENI

Alexandru Iosup

Grids/Clouds
P2P systems
Big Data
Online gaming

Dick Epema

Grids/Clouds
P2P systems
Video-on-demand
e-Science

VENI

Ana Lucia Varbanescu (now UvA)

HPC systems
Multi-cores
Big Data
e-Science

Henk Sips

HPC systems
Multi-cores
P2P systems

VENI

Johan Pouwelse

P2P systems
File-sharing
Video-on-demand

Home page

- www.pds.ewi.tudelft.nl

Publications

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

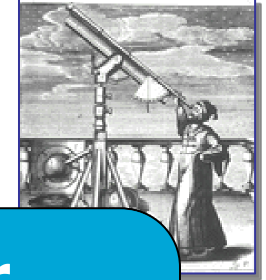
August 31, 2011

Winners IEEE TCSC Scale Challenge 2014

TU Delft

Lessons From Grids

From Hypothesis to Data

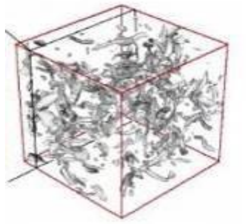


The Fourth Paradigm is suitable for professionals who already know they don't know [enough to formulate good hypotheses], yet need to deliver quickly

ena

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

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

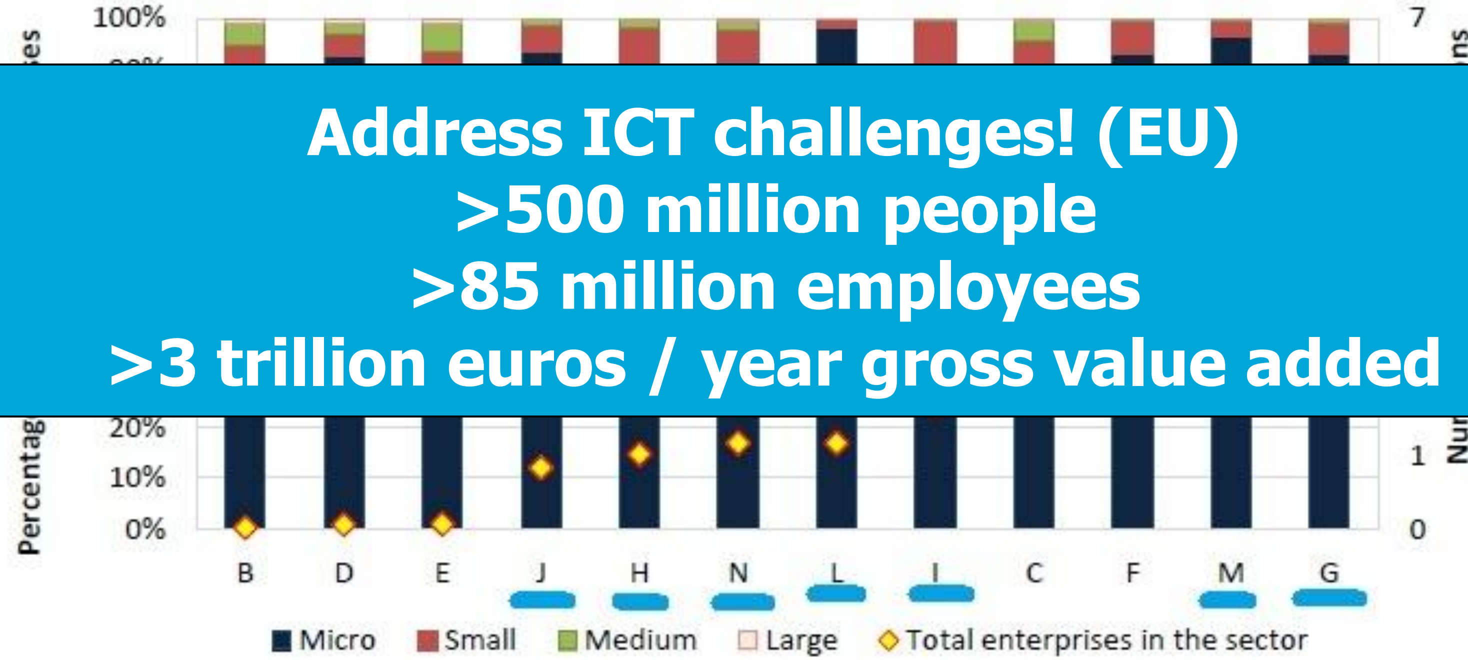


- Today (**the Fourth Paradigm**):
  data exploration
  unify theory, experiment, and simulation
  - Data captured by instruments or generated by simulator
  - Processed by software
  - Information/Knowledge stored in computer
  - Scientist analyzes results using data management and statistics

Source: Jim Gray and "The Fourth Paradigm" (Jan 2007 and, posthumously, 2011), http://research.microsoft.com/en-us/collaboration/fourthparadigm/

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

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



Address ICT challenges! (EU)
>500 million people
>85 million employees
>3 trillion euros / year gross value added

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

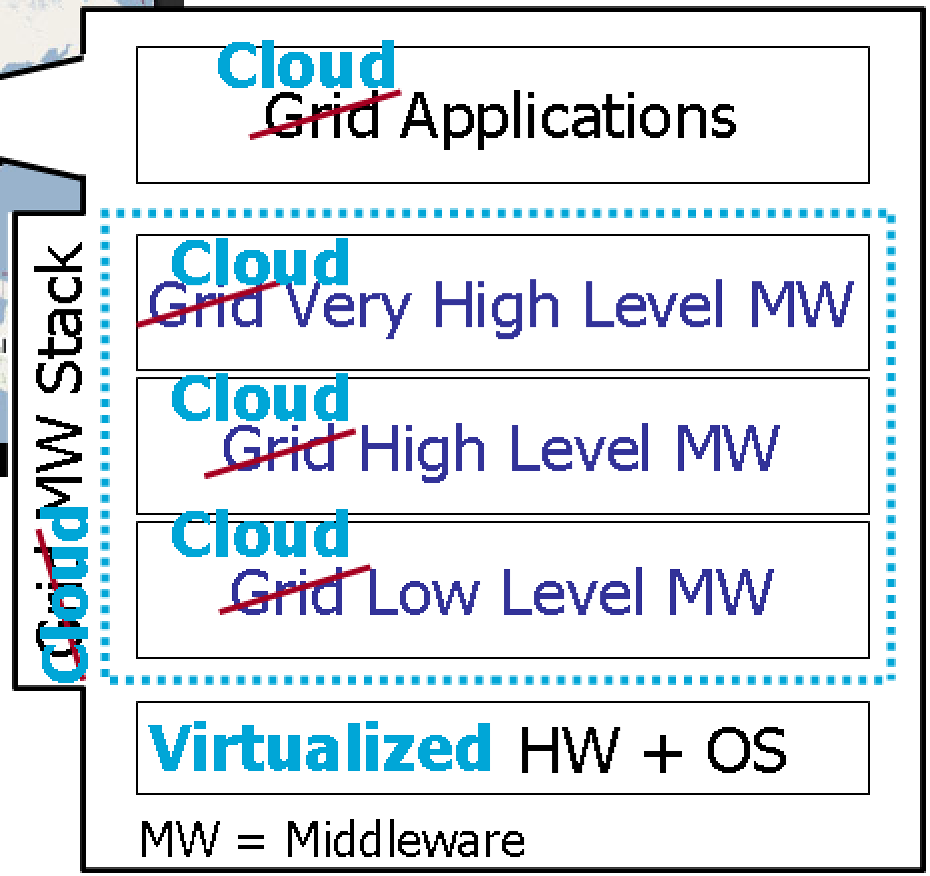
TU Delft

What is Cloud Computing? A Descendant* of the Grid Idea

* Subset.

Source: http://royal.pingdom.com/2008/04/11/map-of-all-google-data-center-locations/

"A computational grid is a hardware and software infrastructure that provides dependable, consistent, pervasive, and inexpensive access to high-end computational capabilities [+ for] nontrivial QoS." I. Foster, 1998 + 1999



TU Delft

The Energy Ceiling (Can We Afford this Vision?)

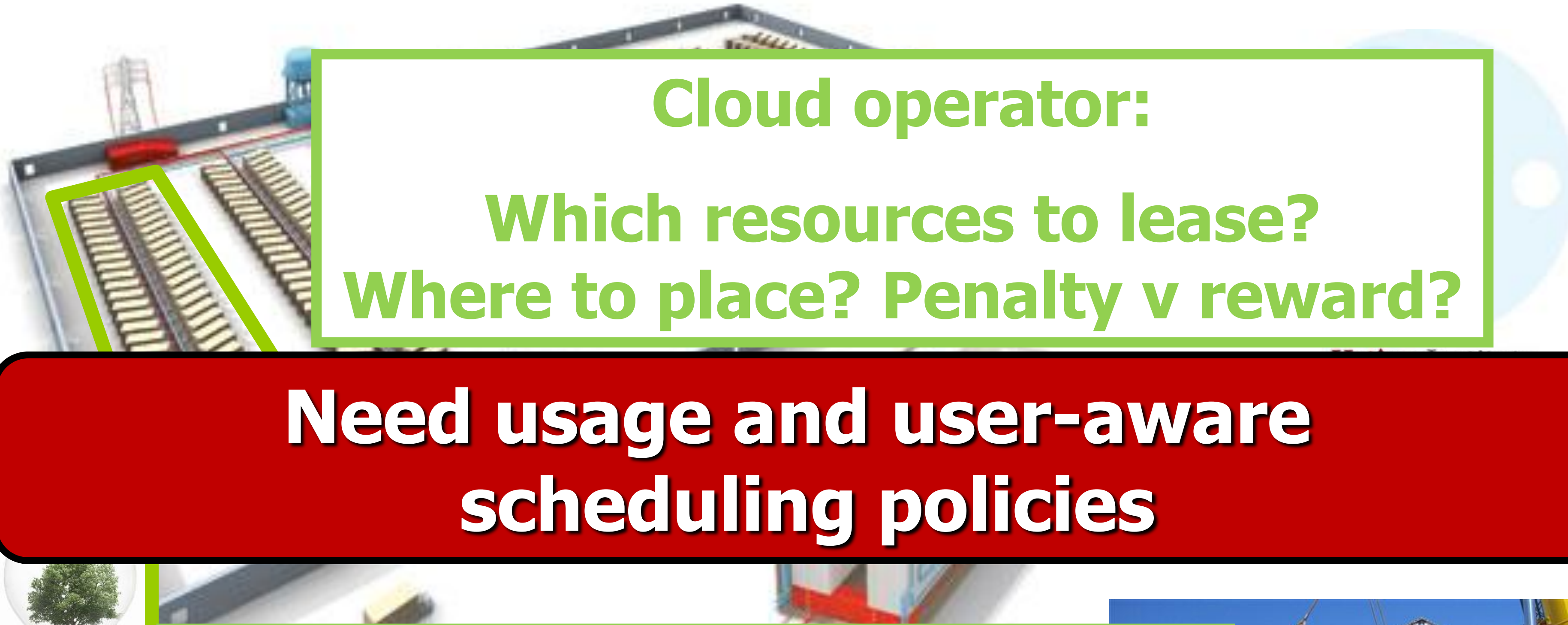
Over 500 YouTube videos have at least 100,000,000 viewers each.

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

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

Source: Ian Bitterlin and Jon Summers, UoL, UK, Jul 2013.
Adapted (Sep 2014).

Scheduling in IaaS Clouds
An Overview



Cloud operator:

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

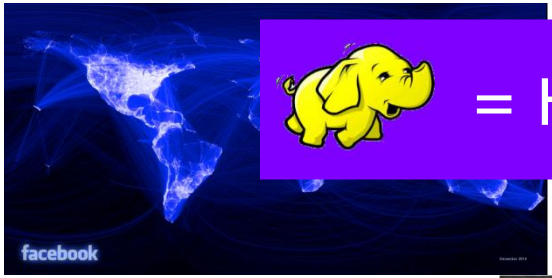
Need usage and user-aware scheduling policies

Cloud customer:

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

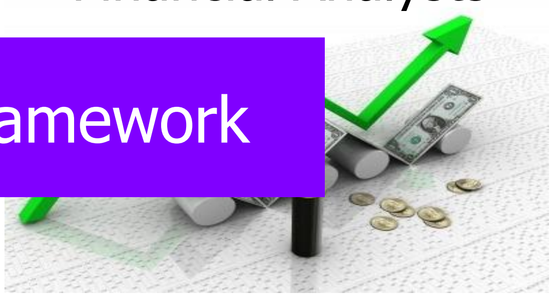
The "Big Data cake" in the Data Center

Online Social Networks



= Hadoop / MapReduce framework

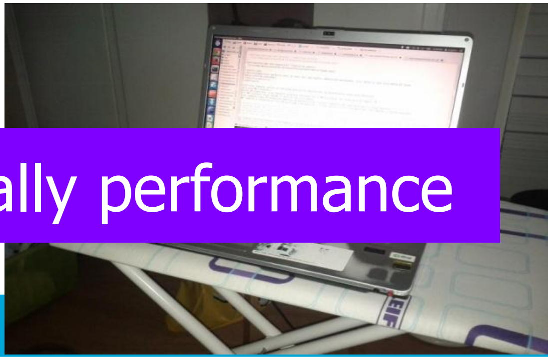
Financial Analysts



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

Universe Explorers

Big Data Enthusiast

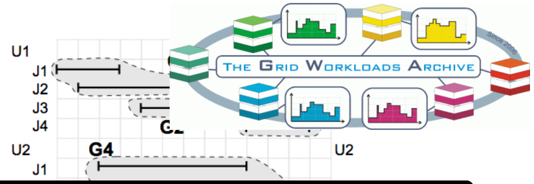


Multiple frameworks = Isolation, especially performance

Agenda

Everyone is a Scientist

Can we afford it?



Workloads

Scheduling

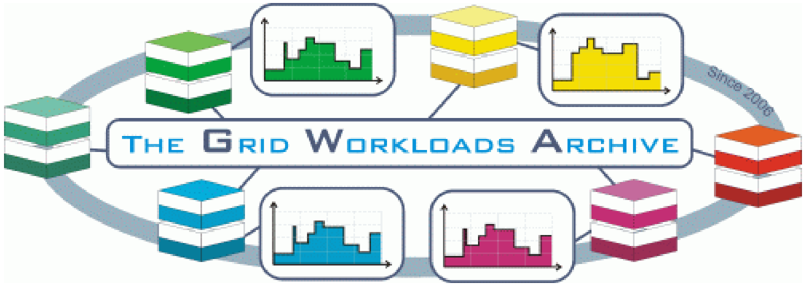
Scheduling

Scheduling

Conclusion

TU Delft

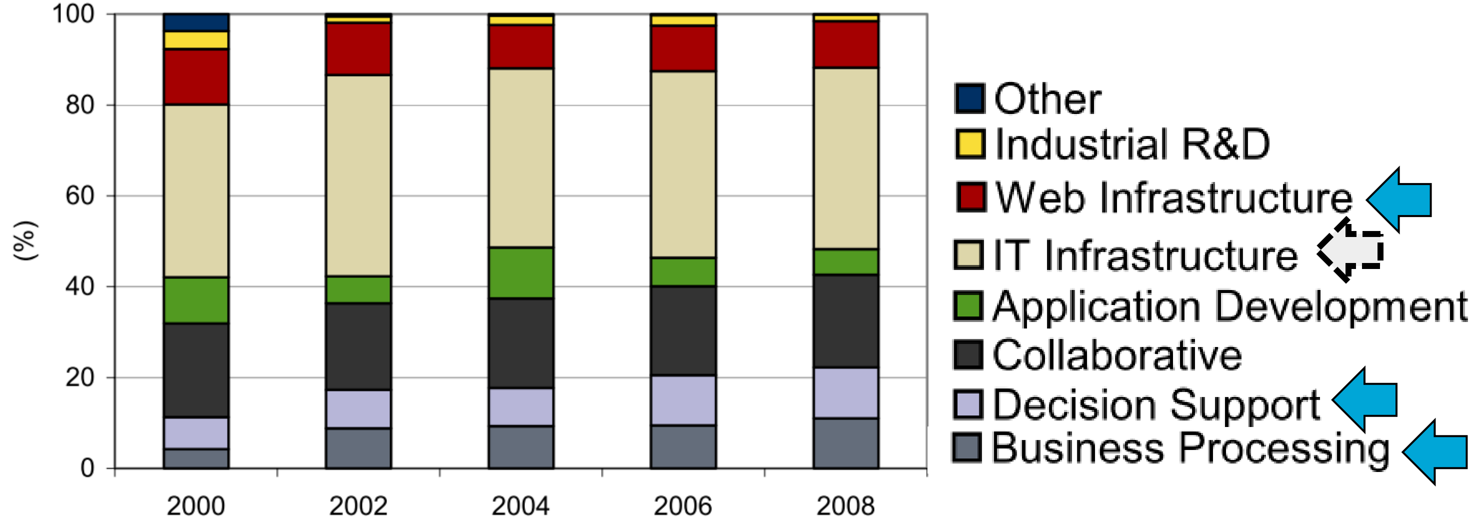
Workloads



Business-Critical Workloads

Business-Critical Workloads

- Growing user base (IDC'10)



Worldwide Windows Workloads Shipment Share, 2000–2008

- Applications
  - Business intelligence and decision
  - Office back-end
  - Other Business support

TUDelft

Workload Characterization

Requested resources				Used resources			
CPU	Mem	Disk	Net	CPU	Mem	Disk	Net
—	—	—	—	yes	—	yes	—
yes	yes	—	—	yes	yes	—	—
—	—	—	—	yes	—	yes	—
—	—	—	—	yes	yes	—	—
—	—	—	—	yes	yes	yes	yes
—	yes	—	—	yes	yes	—	—
yes	yes	yes	yes	yes	yes	yes	yes

- Basic statistics
- Correlation analysis
- Time pattern analysis

Van Beek, Shen, Iosup: Statistical Characterization of Business-Critical Workloads Hosted in Cloud Datacenters. CCGRID 2015

Workload Characterization Results

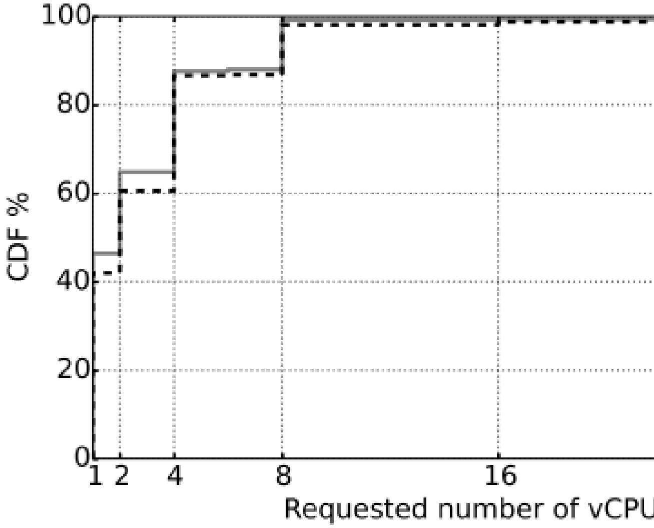
1. More than 60% of the VMs use less than 4 cores and 8GB of memory.

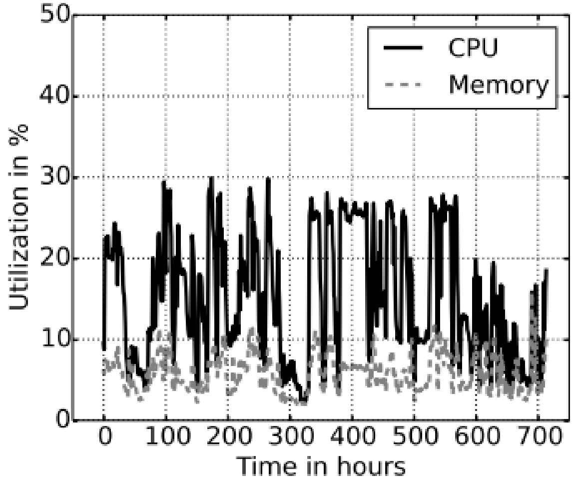
2. There is a strong positive correlation between requested CPU and memory.

3. Resource usage is low, under 10% of the requested resources, and the correlation between requested and used resources is also low.

4. Peak workloads can be 10–10,000 times higher than mean workloads, depending on resource type.

5. The CPU and memory resource usage is often predictable over the short-term. Disk and network I/O follow daily patterns.





Van Beek, Shen, Iosup: Statistical Characterization of Business-Critical Workloads Hosted in Cloud Datacenters. CCGRID 2015

Scheduling

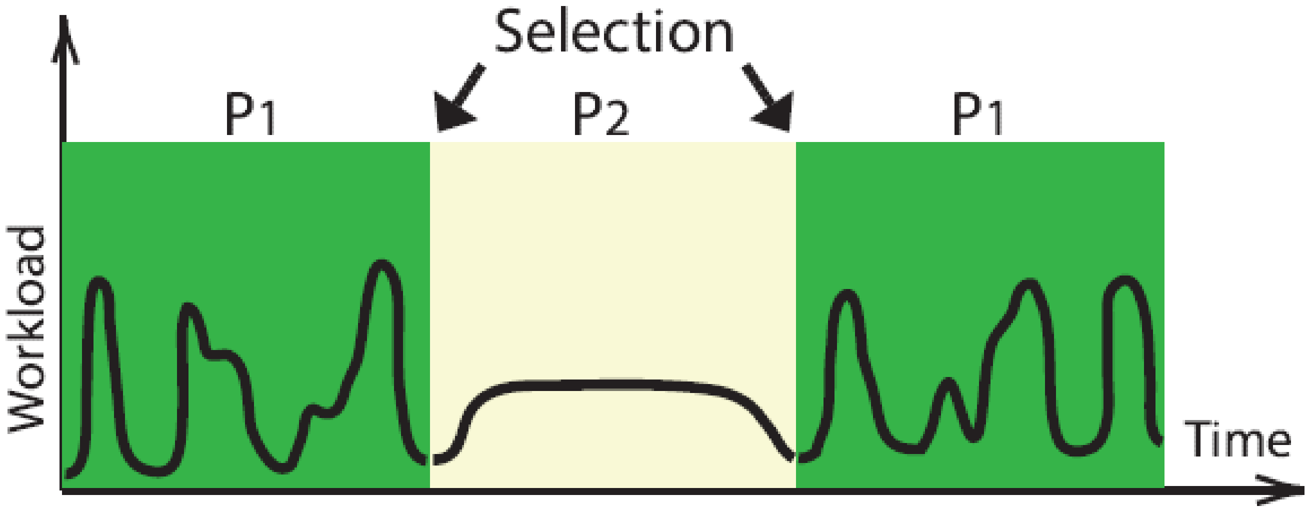
Portfolio Scheduling

Why Portfolio Scheduling?

- **Data centers increasingly popular**
  - Constant deployment since mid-1990s
  - Users moving their computation to IaaS clouds
  - Consolidation efforts in mid- and large-scale companies
- **Old scheduling aspects**
  - Hundreds of approaches, each targeting specific conditions—which?
  - No one-size-fits-all policy
- **New scheduling aspects**
  - New workloads
  - New data center architectures
  - New cost models

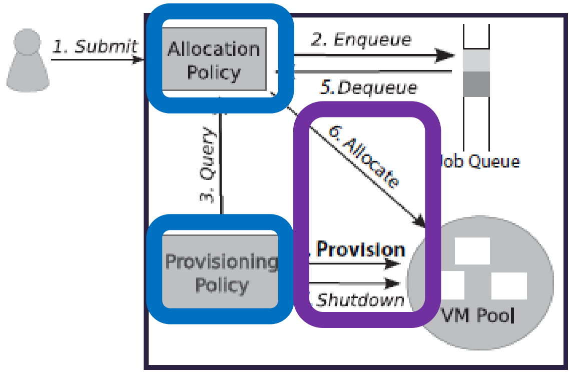
- **Developing a scheduling policy is risky and ephemeral**
- **Selecting a scheduling policy for your data center is difficult**

What is Portfolio Scheduling? In a Nutshell, for Data Centers



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

Portfolio Scheduling
The Process



Which policies to include?

Creation

Which policy to activate?

Selection

Application

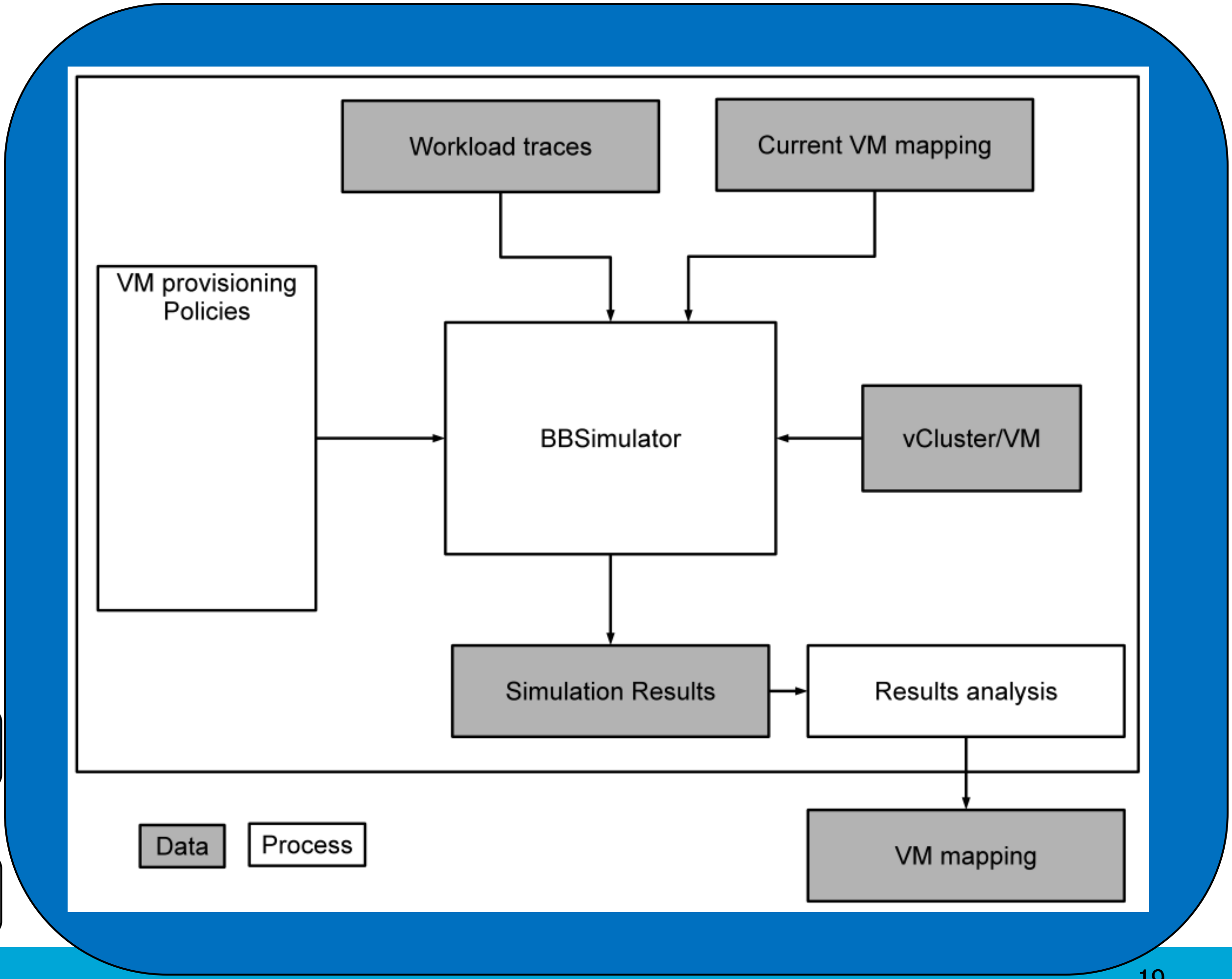
Which resources? What to log?

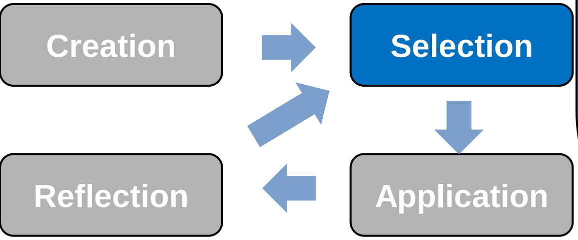
Reflection

Which changes to the portfolio?

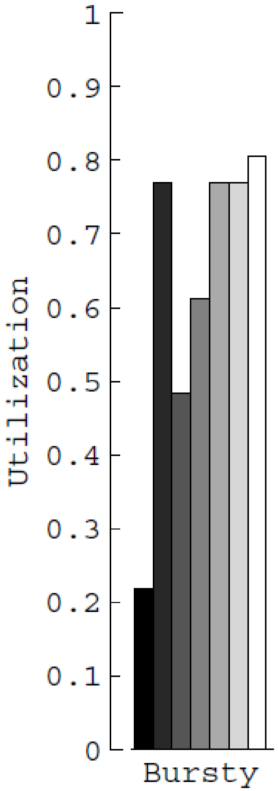
Bonus: The portfolio scheduler can explain each selection, like your engineers would.

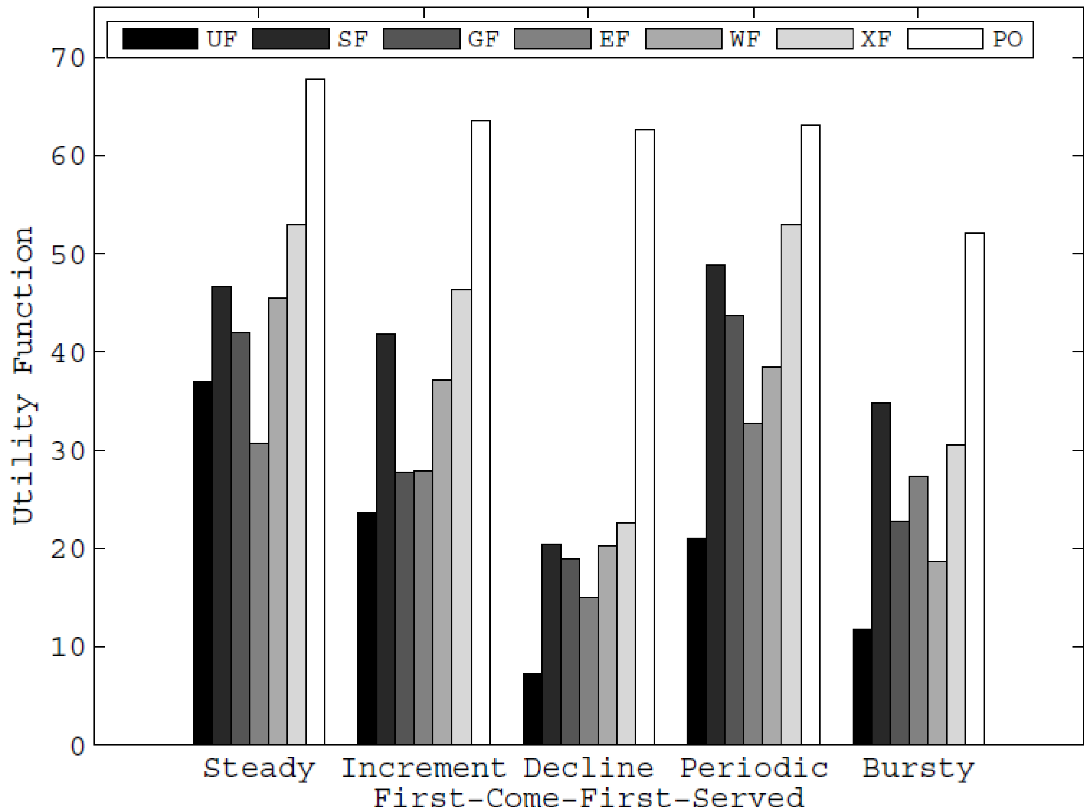
Portfolio Scheduling
An Implementation of the Selection Step





Experimental Results, Synthetic Workloads
Resource Utilization + Workload Utility





- POrtfolio leads to high utilization
- Start-Up leads to poor utilization

- POrtfolio leads to better utility
- Start-Up leads to poor utility

Deng, Song, Ren, Iosup: Exploring portfolio scheduling for long-term execution of scientific workloads in IaaS clouds. SC 2013: 55

Experimental Results: Real-World Business-Critical Workloads at Bitbrains

- Additional requirements: affinity and anti-affinity, worst-fit/best-fit
- Risk score = Risk(CPU/mem/netw/IO)
  - Considers oversubscription, latency
  - Combines severity and probability
  - Lower is better

What the portfolio optimizes for (same set of policies)

Configurations:	**replay**	**MinScore**	**MaxScore**	**MinMem**	**MaxMem**	**MinCPU**	**MaxCPU**
Memory	1.68	0.31	2.18	6.16	1.56	0.26	15.09
CPU	0.0	0.0	0.0	0.0	0.0	0.0	0.0
IO write	2.6	1.53	2.99	1.49	1.69	1.35	1.93
IO read	2.31	1.25	2.15	1.2	1.23	2.05	1.41
Network send	1.86	1.18	2.51	1.13	1.32	0.9	1.86
Network receive	1.17	1.13	2.19	1.09	1.1	1.11	1.49
Total:	**9.62**	**5.4**	**12.01**	**11.07**	**6.9**	**5.66**	**21.78**

- Portfolio scheduler can lead to lower risk in the datacenter
- Policy selection very important, otherwise portfolio can perform badly

Van Beek, et al.: Mnemos: Self-Expressive Management of Business-Critical Workloads in Virtualized Datacenters, IEEE Computer SI on Self-Aware and Self-Expressive Computing Systems(submitted Dec 2014)

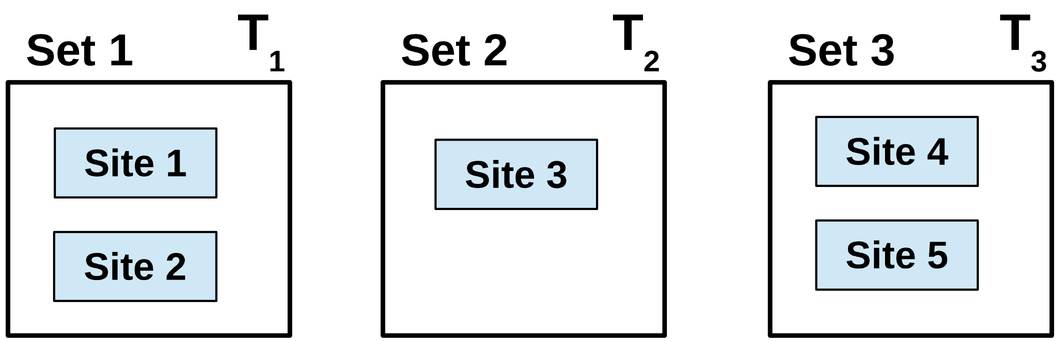
Scheduling

Multi-Cluster, Multi-DC TAGS

TAGS-based Policy Design

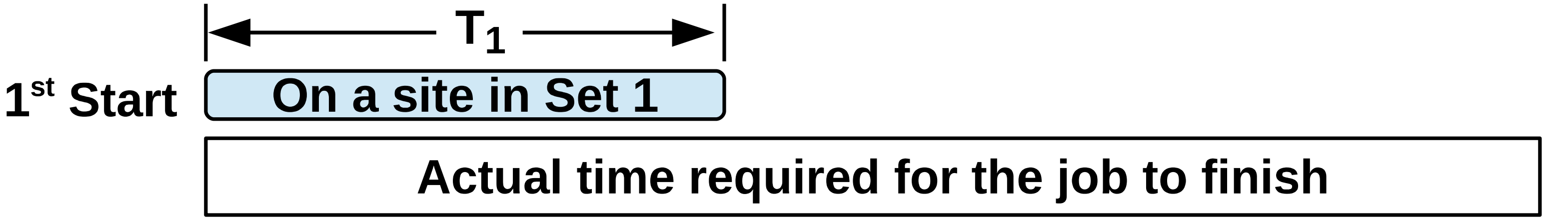
- Goal: to achieve low slowdown **without prediction**
- Method: to partition the sites into sets to serve jobs of different runtime ranges

- A number of sets of sites
- Set i allows jobs to run for T_i amount of time ($T_i < T_{i+1}$)
- The last set has a T of ∞ (all jobs will finish without being killed)



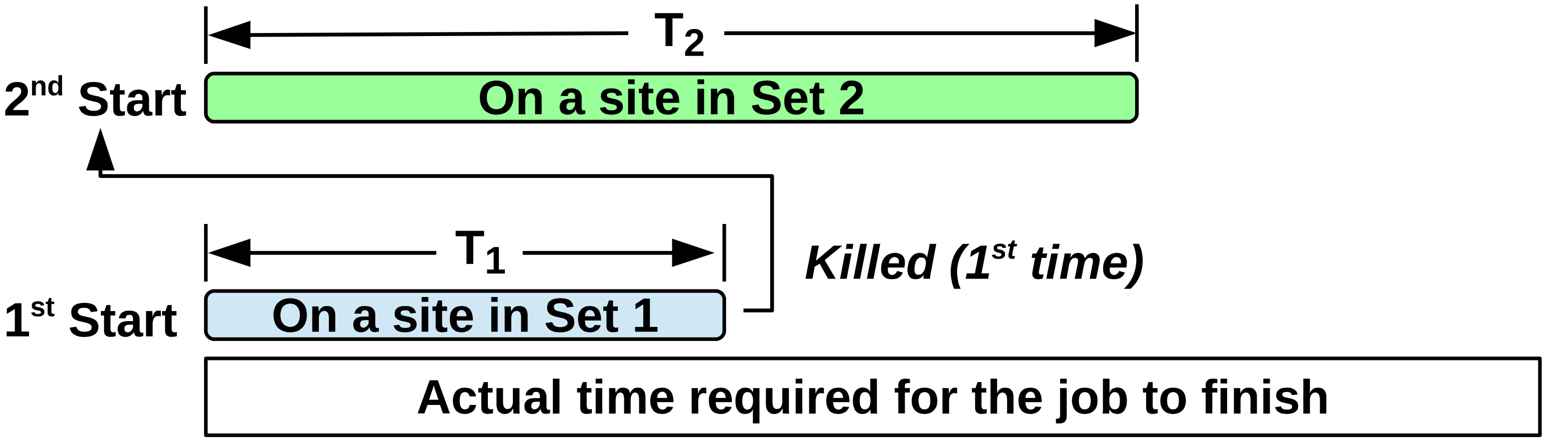
$T_1 < T_2 < T_3$
$T_3 = \infty$

TAGS-based Policy Design



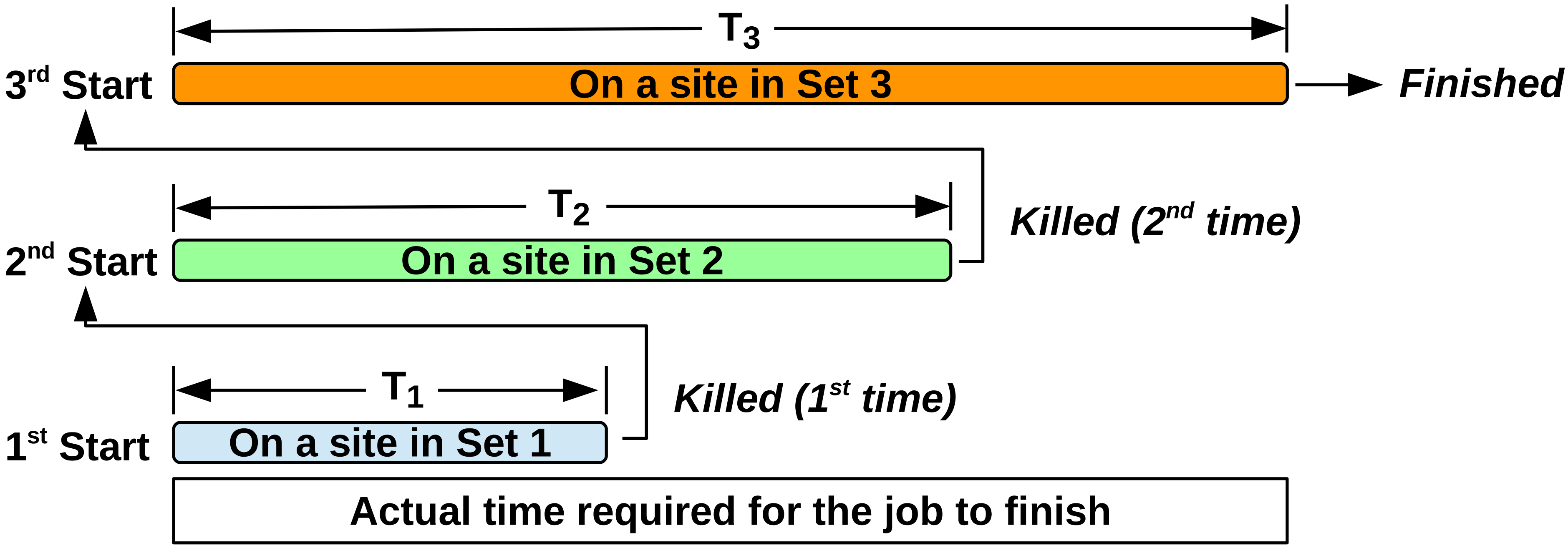
Policy Design

TAGS-based Policies in General



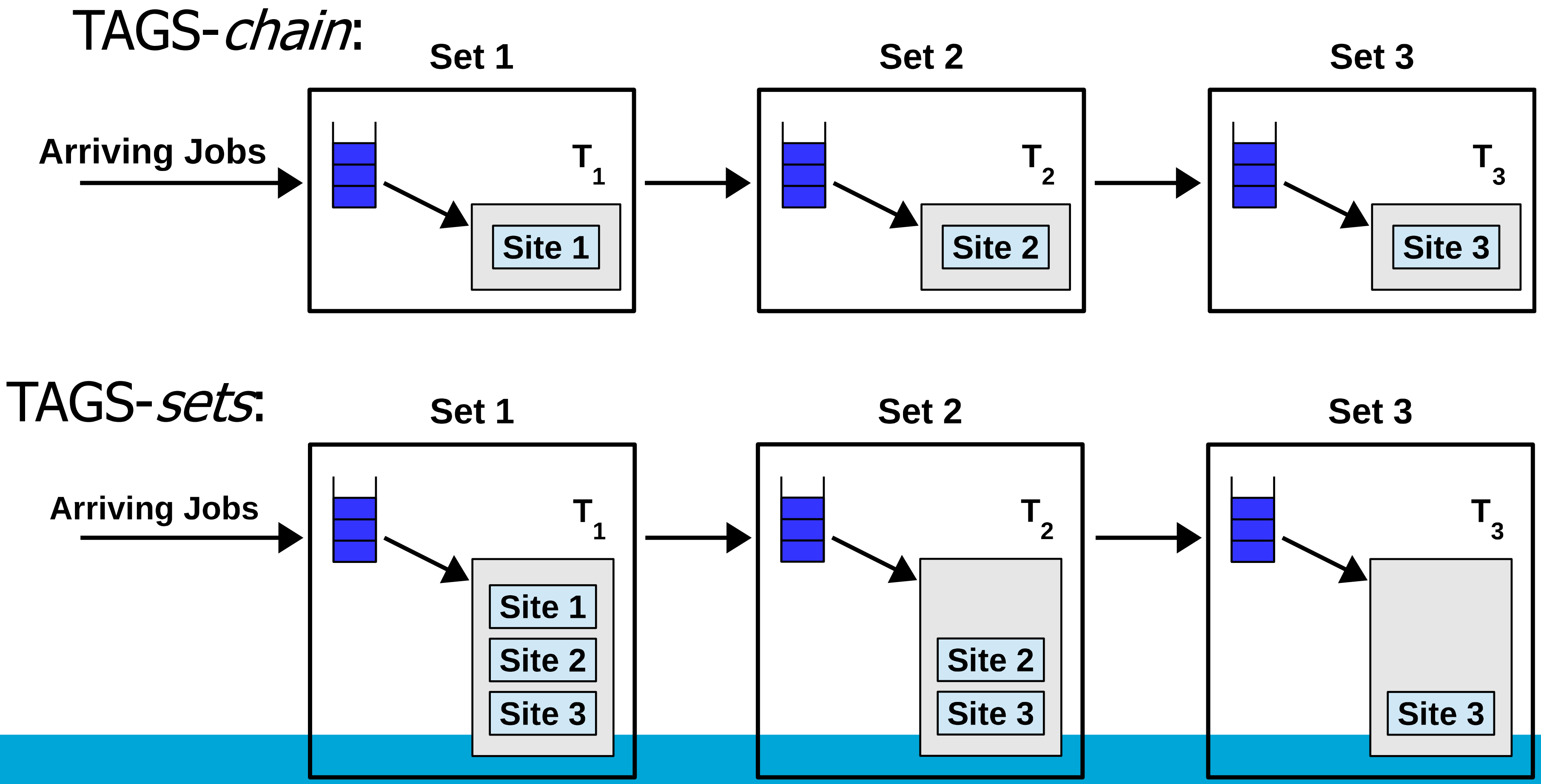
Policy Design

TAGS-based Policies in General



Our Policy Design for Multi-Cluster Multi-DCs

TAGS-*chain* and TAGS-*sets*



Fei, Ghit, Iosup, Epema: KOALA-C: A task allocator for integrated multicluster and multicloud environments. CLUSTER 2014

Policy Design Space

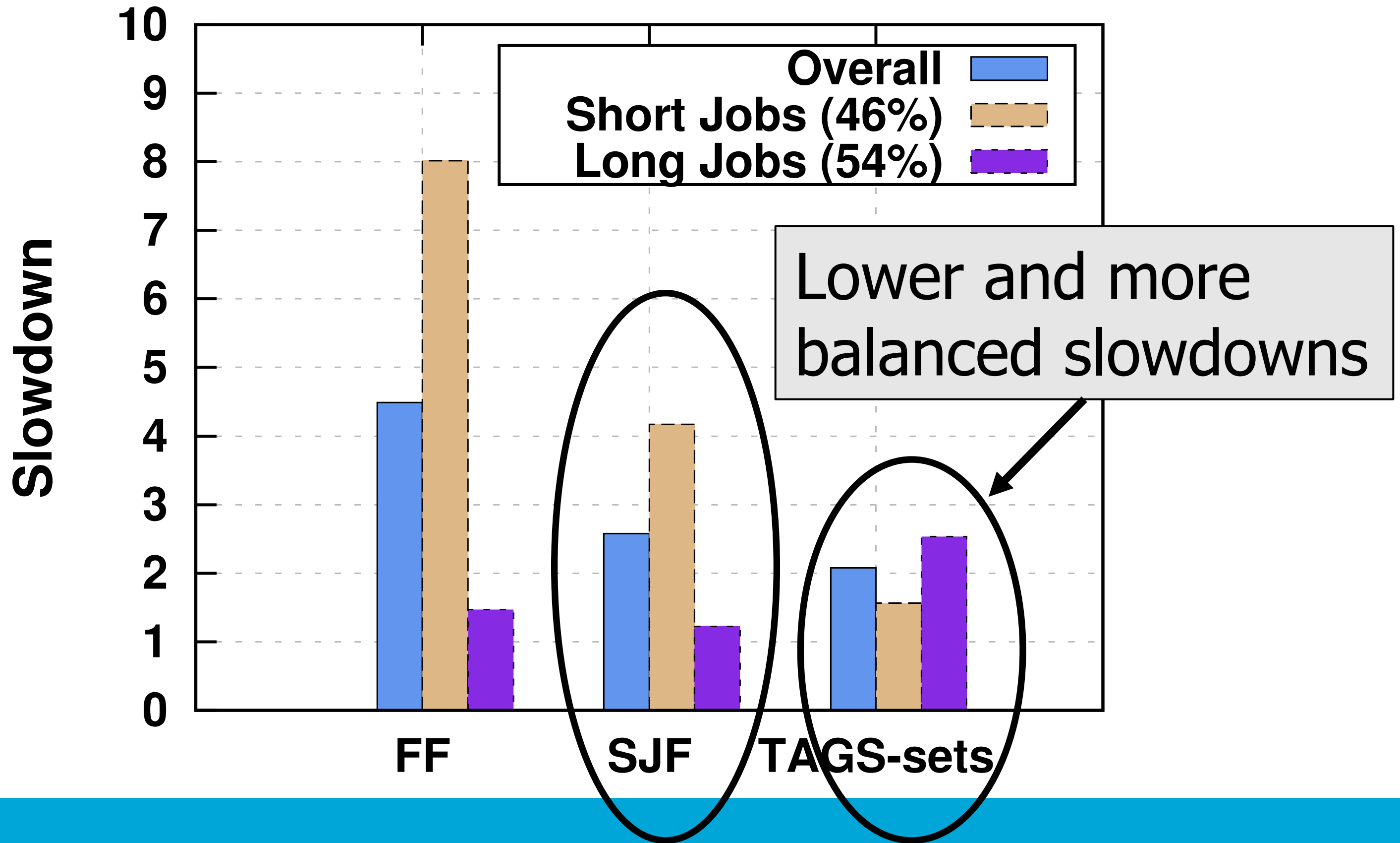
	Prediction	Slowdown	Preemption	MC+MC support
Uninformed (FF, RR)	−	↑	✗	✗
Informed (SJF, SJF-*ideal*, HSDF)	+	↓	✗	✗
Traditional TAGS	−	↓	✓	✗
TAGS extensions (TAGS-*chain*, TAGS-*sets*) new	−	↓	✓	✓

TAGS-chain and TAGS-sets have nice properties both in simulation and in real-world experimentation.

Fei, Ghit, Iosup, Epema: KOALA-C: A task allocator for integrated multicluster and multicloud environments. CLUSTER 2014

TU Delft

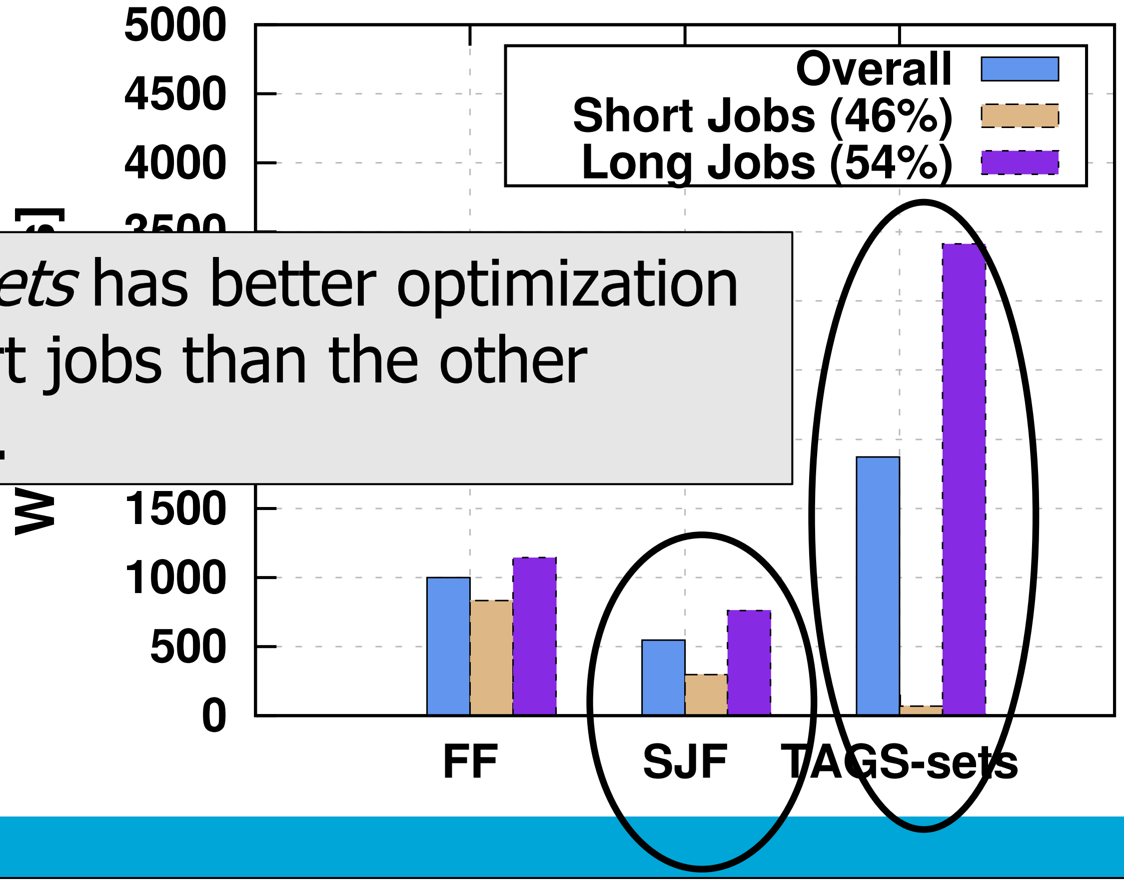
Real-World Experimental Results



Fei, Ghit, Iosup, Epema: KOALA-C: A task allocator for integrated multicluster and multicloud environments. CLUSTER 2014

TU Delft

Real-world Experimental Results



TAGS-*sets* has better optimization for short jobs than the other policies.

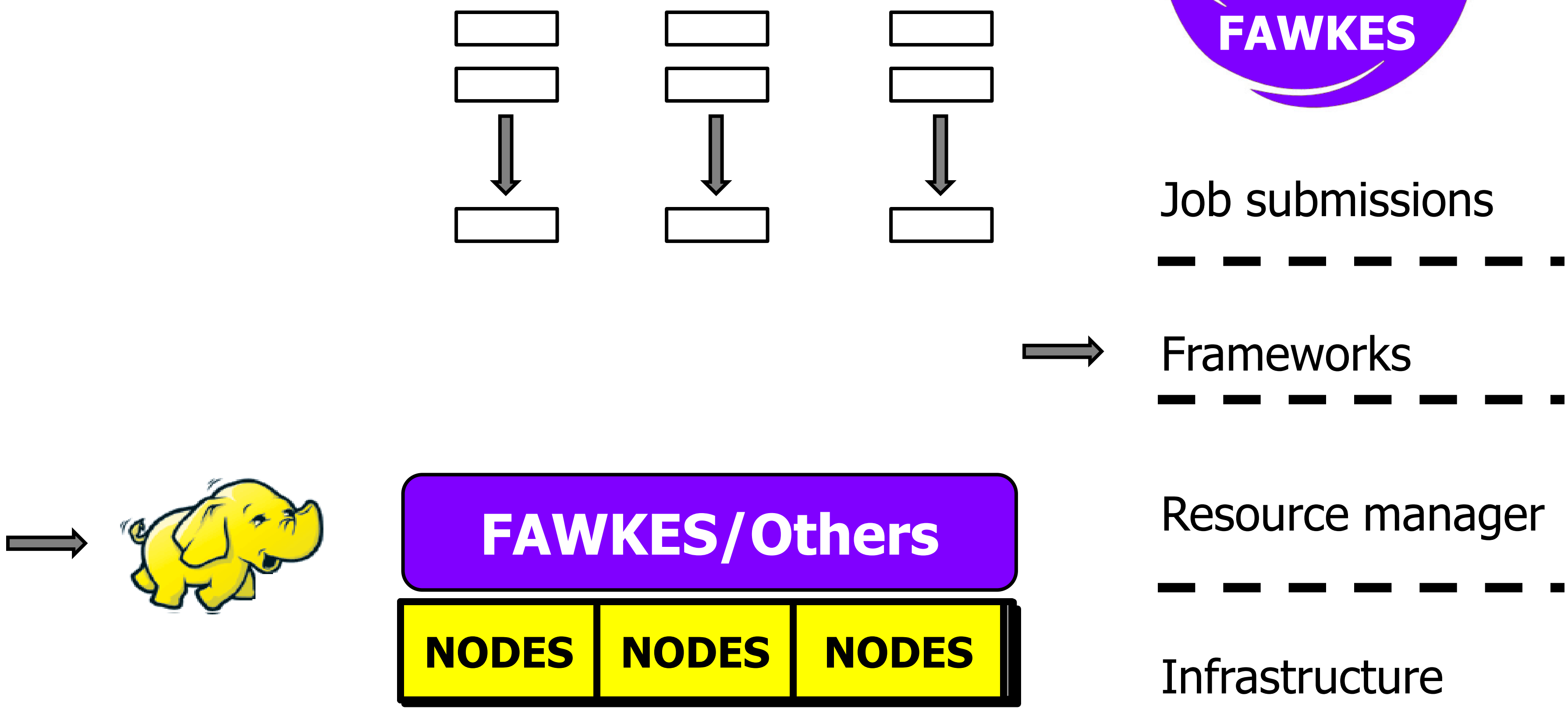
Fei, Ghit, Iosup, Epema: KOALA-C: A task allocator for integrated multicluster and multicloud environments. CLUSTER 2014

Scheduling

Elastic MapReduce

Dynamic Big Data Processing

Fawkes = Elastic MapReduce via Two-level scheduling architecture

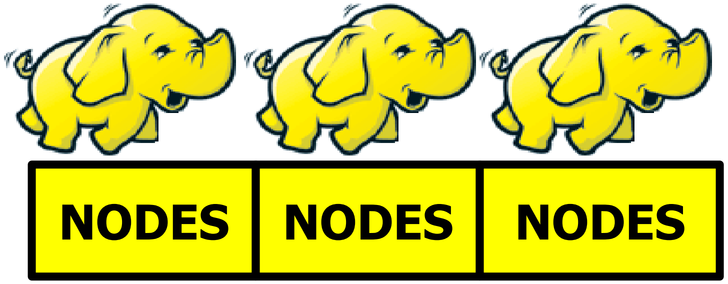


Ghit, Yigitbasi, Iosup, Epema, Iosup. Balanced Resource Allocations Across Multiple Dynamic MapReduce Clusters. ACM SIGMETRICS 2014.

Elastic MapReduce

MapReduce framework

- Distributed file system
- Execution engine
- Data locality constraints

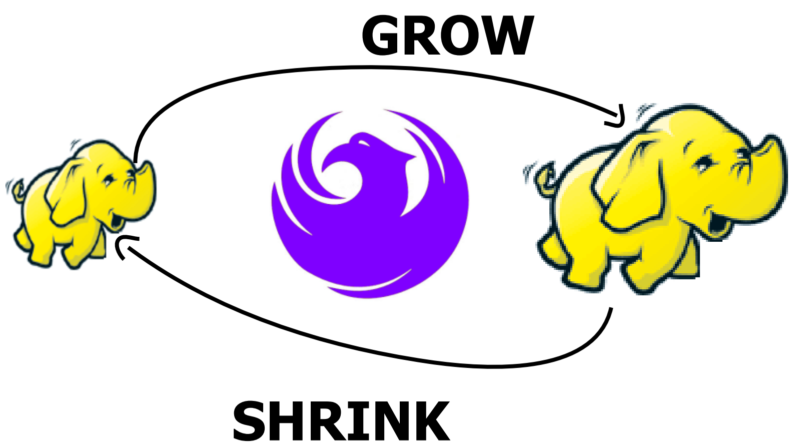


Because workloads may be time-varying:

- Poor resource utilization
- Imbalanced service levels

Grow and shrink MapReduce

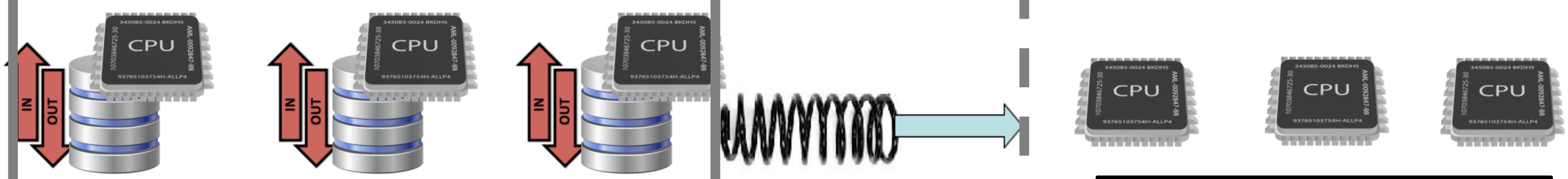
- High resource utilization
- Reconfiguration for balanced service levels
- Break data locality



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

No data locality

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**

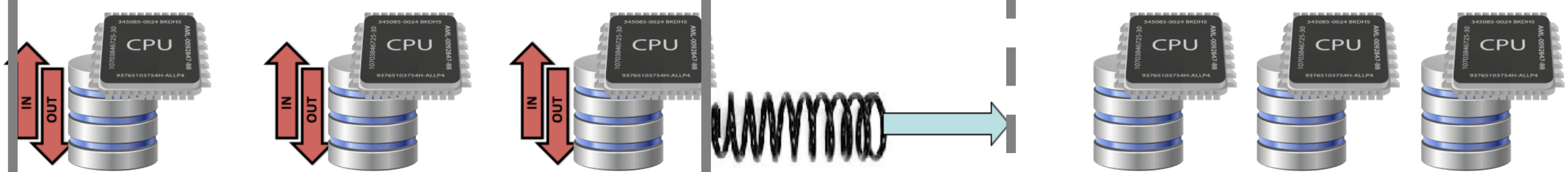
Performance?

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

Relaxed data locality

Core nodes

Trans-core nodes (TC)



INPUT/OUTPUT DATA

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

OUTPUT DATA

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

Better performance?

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

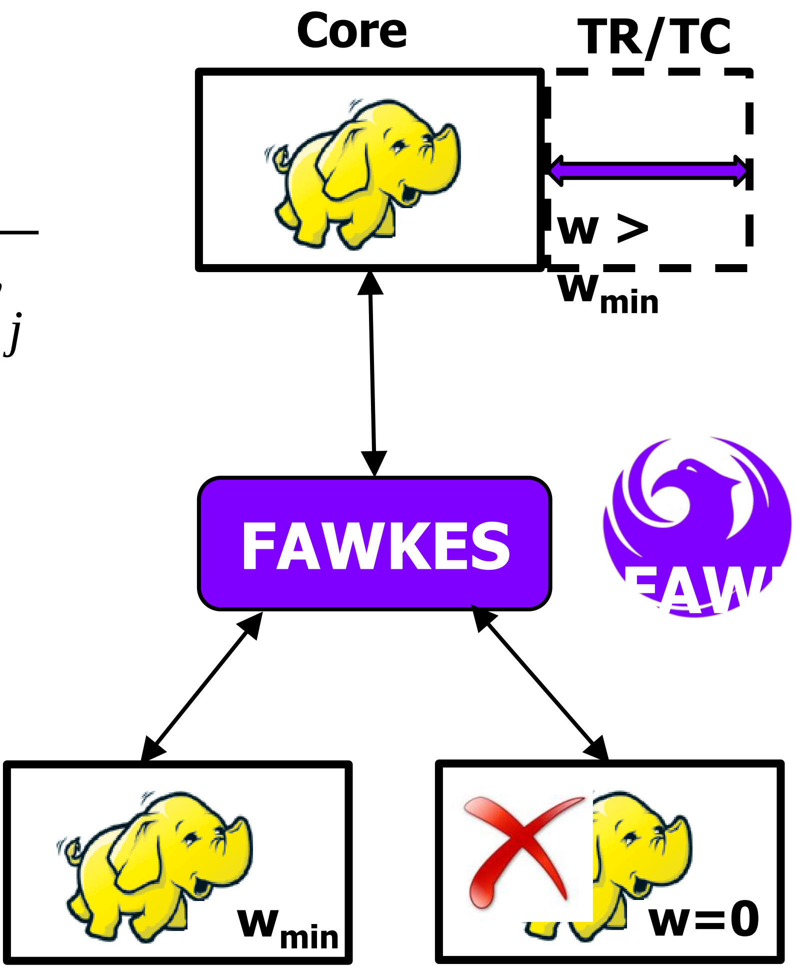
FAWKES in a nutshell

1. Size of MapReduce cluster
   - Changes dynamically
   - Balanced by weight
   - Weight by demand, usage, actual service

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

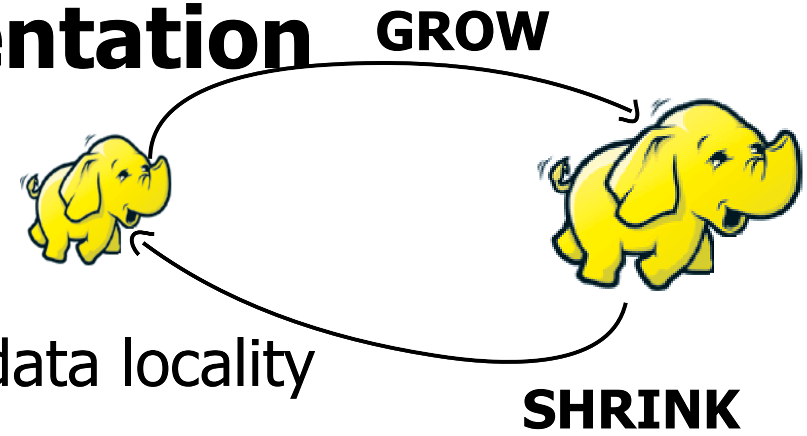
2. Updates dynamic weights when
   - New frameworks arrive
   - Framework states change

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



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

Real-World Experimentation



1. Dynamic MapReduce relaxes data locality

2. FAWKES policies can reduce imbalance between frameworks

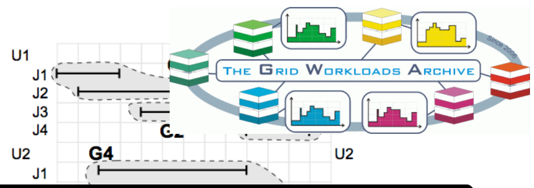
FAWKES

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

Agenda

Everyone is a Scientist

Can we afford it?



Workloads

Scheduling

Scheduling

Scheduling

Conclusion

TU Delft

~~Conclusion~~ Take-Home Message

- **"Everyone is a Scientist!"**
  Our vision of a growing, leading Europe
- **Grand challenge**
  - Managing the datacentre
  - Helping demanding users
- **In this talk**
  - Business-critical workloads
  - Scheduling with portfolio scheduling
  - Scheduling by guessing size
  - Scheduling by load-balancing across elastic-MapReduce frameworks

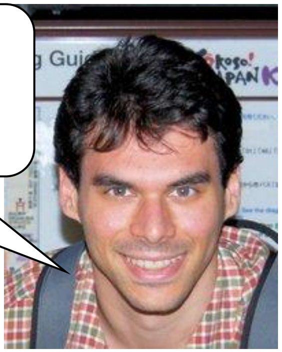
TU Delft

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

More Info:

- http://www.st.ewi.tudelft.nl/~iosup/
- http://www.pds.ewi.tudelft.nl/
- http://research.spec.org

Do not hesitate to contact me…



Alexandru Iosup

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