Towards an Optimized Big Data Processing System

The Doctoral Symposium of the IEEE/ACM CCGrid 2013
Delft, The Netherlands

Bogdan Ghiț, Alexandru Iosup, and Dick Epema

Parallel and Distributed Systems Group
Delft University of Technology
Delft, The Netherlands

COMMIT/

TU Delft

Delft University of Technology

PhD at TU Delft

Candidate: **Bogdan Ghiţ**

Group: Parallel and Distributed Systems

Supervisors: Dick Epema, Alexandru Iosup

Research Topic: Resource management in grids and clouds

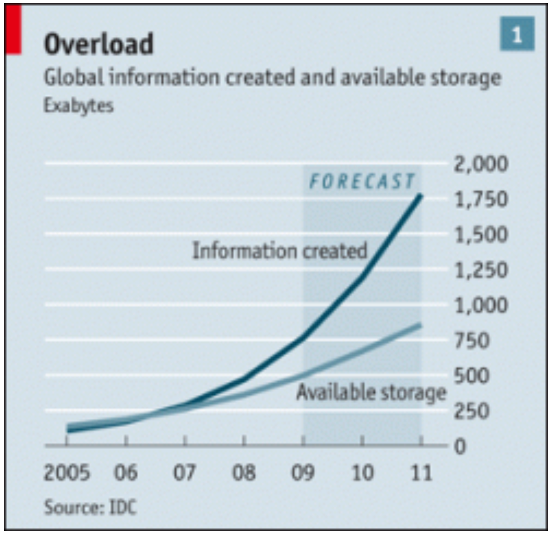
Start date: 24 October 2011

Finish date: 24 October 2015

“The Data Deluge”

“According to one estimate mankind created 150 exabytes (billion gigabytes) of data in 2005. This year, it will create 1,200 exabytes.”

The Data Deluge, The Economist, 25 February 2010



Overload

Global information created and available storage

Exabytes

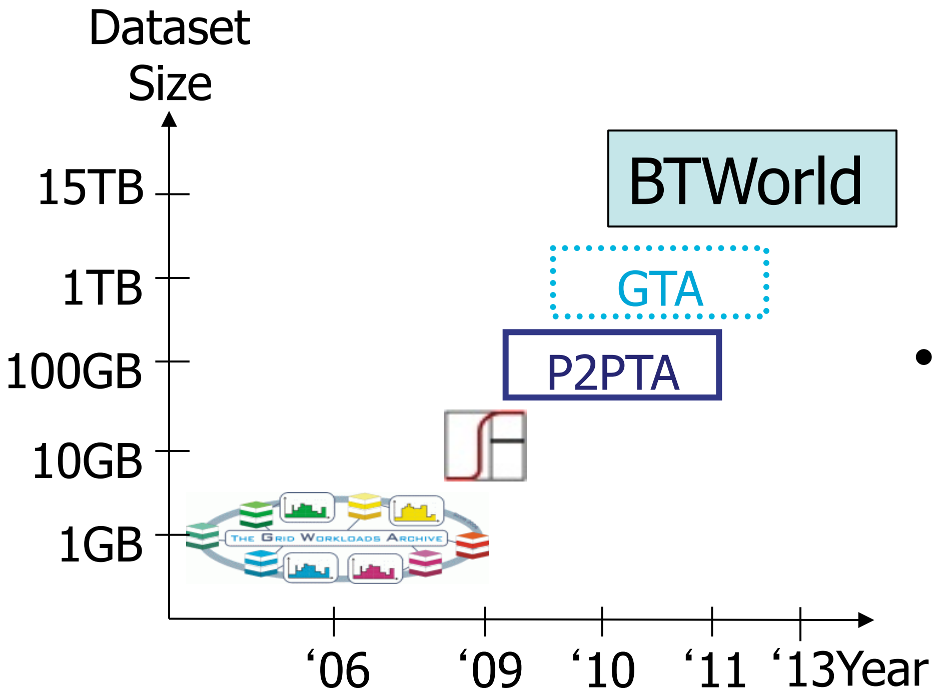
Source: IDC

The Economist
Obama the warrior
Misgoverning Argentina
The economic shift from West to East
Genetically modified crops blossom
The right to eat cats and dogs
The data deluge
AND HOW TO HANDLE IT: A 14-PAGE SPECIAL REPORT

The data is difficult to store, even harder to analyze it

Data Sources

- Computer Science



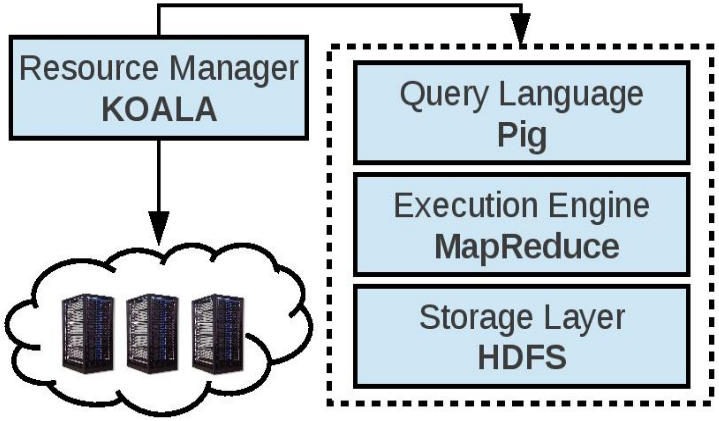
- LinkedIn
  - ➢ Daily batch processing for *"People you may know"* recommendations

The State of LinkedIn
Canada 5,373,475
Denmark 856,635
Sweden 929,829
Netherlands 3,232,249
Norway 563,679
Finland 352,681
Russian Federation 711,991
U.K.. 8,779,696
Germany 1,576,858
Ireland 644,421
Poland 548,975
Belgium 1,080,825
Romania 626,334
Pakistan 1,057,057
China 2,024,628
U.S.A. 60,857,394
France 3,411,161
Turkey 1,079,365
Japan 551,503
Italy 3,026,074
Portugal 824,950
Spain 2,815,541
Egypt 527,746
Hong Kong 445,402
Mexico 2,294,304
Philippines 937,746
Switzerland 643,662
Saudi Arabia 605,837
Malaysia 994,607
Israel 617,266
U.A.E. 901,996
India 13,984,488
Indonesia 1,034,660
Greece 389,690
Brazil 7,344,718
Singapore 735,602
Australia 2,934,712
Argentina 1,829,794
South Africa 1,656,670
New Zeland 556,333
< 9,999,999
3,000,001 9,999,999
1.000,000 3,000,000
> 1,000,000
150,000,000
registered members

- Facebook
  - ➢ 30 PB by March 2011 = 3,000 x Library of Congress

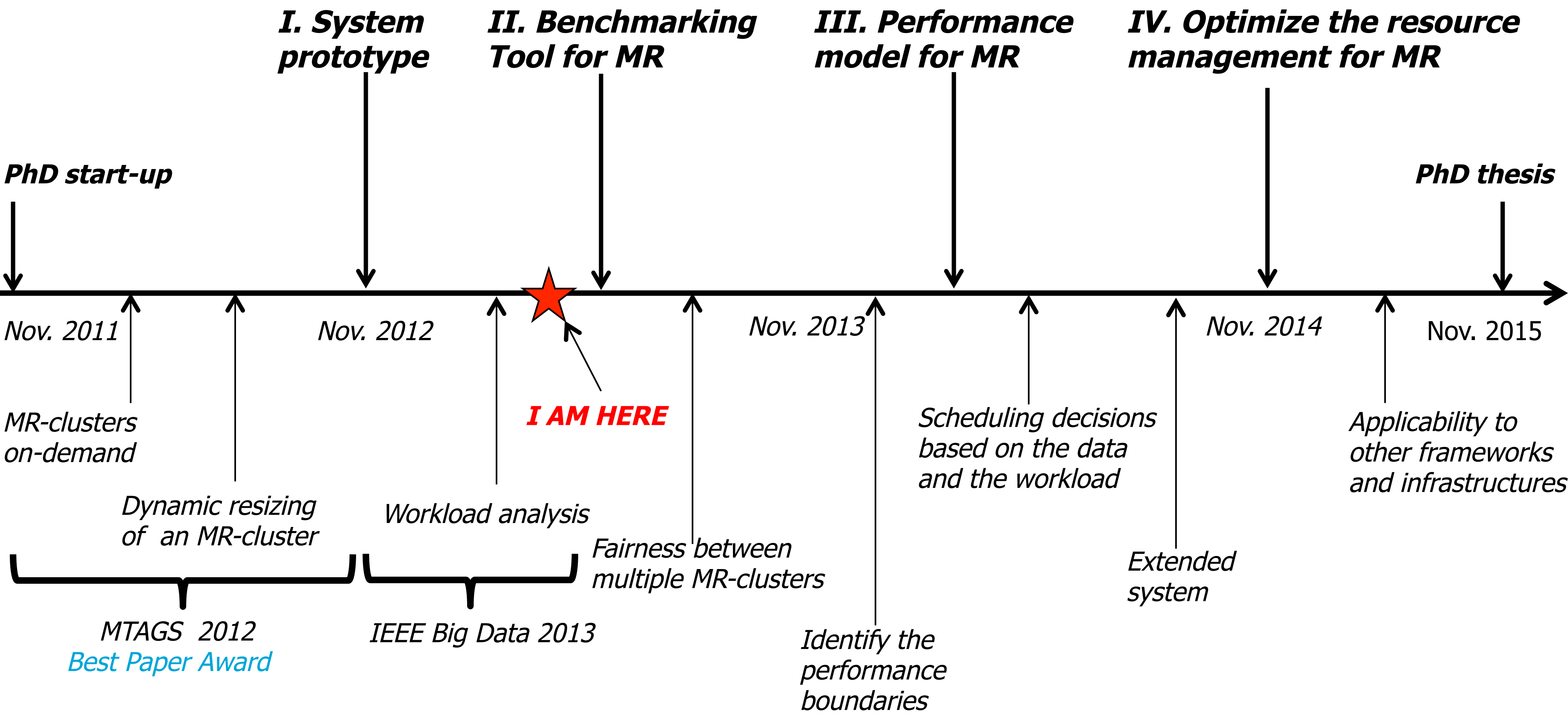
MapReduce and Beyond

- Master-slave model
- MR-cluster
  - Stack of frameworks for large-scale data processing



- *Multiple users vs. Isolation*
  - MR-clusters on-demand
  - Isolation w.r.t. performance, data, failure, and versioning
- *Data volume vs. Limited resources*
  - Use resources from multiple clusters
  - Dynamically change the size
- *Performance vs. Fairness*
  - Capacity-based model
  - Capability-based model

Road Map



PhD start-up

I. System prototype

II. Benchmarking Tool for MR

III. Performance model for MR

IV. Optimize the resource management for MR

PhD thesis

Nov. 2011

Nov. 2012

Nov. 2013

Nov. 2014

Nov. 2015

MR-clusters on-demand

Dynamic resizing of an MR-cluster

MTAGS 2012
Best Paper Award

Workload analysis

IEEE Big Data 2013

I AM HERE

Fairness between multiple MR-clusters

Identify the performance boundaries

Scheduling decisions based on the data and the workload

Extended system

Applicability to other frameworks and infrastructures

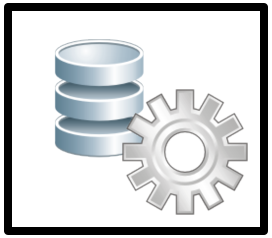
Dynamic MapReduce Clusters

- Complex resource management
  - Single / multiple physical clusters
  - Placement and scheduling policies
  - Change resource allocations at runtime
  - Data management issues

- MR-cluster structure: *data replication vs. data locality*

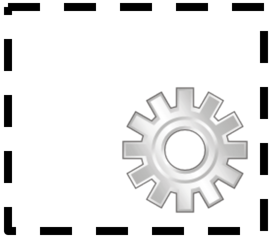
Core nodes

- Execute tasks and store data locally
- Replication required when removed



Transient nodes

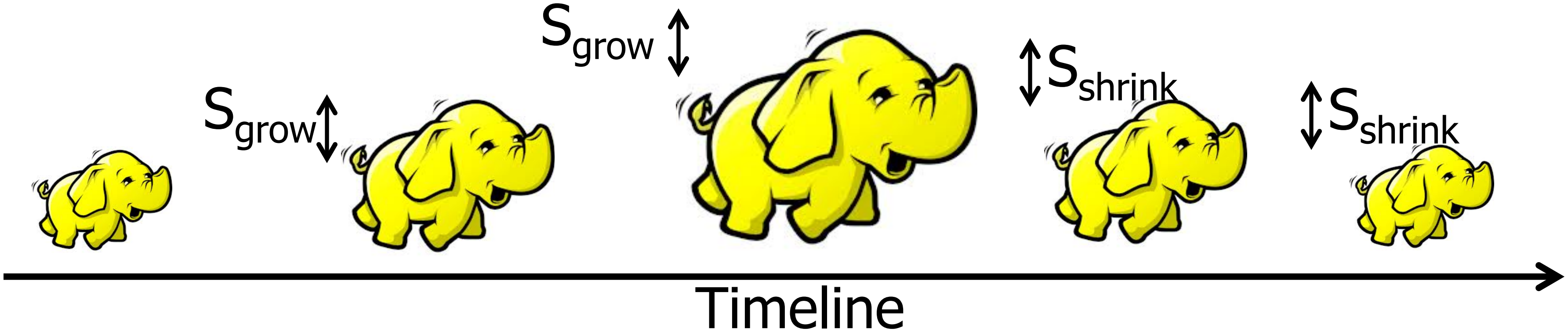
- Execute tasks, do not store data
- Data transfers to read/write data



Resizing Mechanism

Question: Given an MR-cluster, how can you tell if it is overloaded or underloaded?

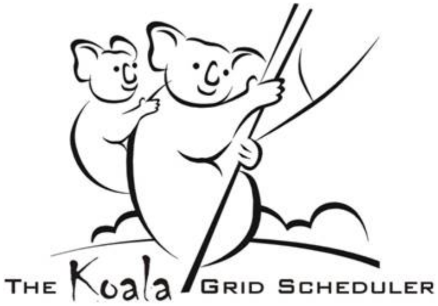
- Monitor the MR cluster utilization: $F_{\min} \le \frac{\#tasks}{\#slots} \le F_{\max}$
- Grow-Shrink Policy (GSP) – with transient nodes
  - Size of grow and shrink steps: $\mathbf{S_{grow}}$ and $\mathbf{S_{shrink}}$
  - Baseline policies: grow with core nodes (GGDP) or grow with transient nodes (GGP)



System Prototype

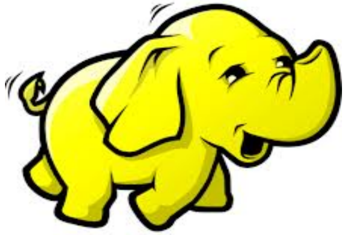
Koala Grid Scheduler

- Enables processor and data co-allocation
- Implements placement and scheduling policies
- Application types: cycle-scavenging, workflows, OpenMPI



Koala and MapReduce

- Developed an MR-Runner module to schedule MR jobs
- Provides isolated MR-clusters on a per-user basis
- Koala mechanism for resizing the MR-clusters
- MR jobs submissions transparent to Koala

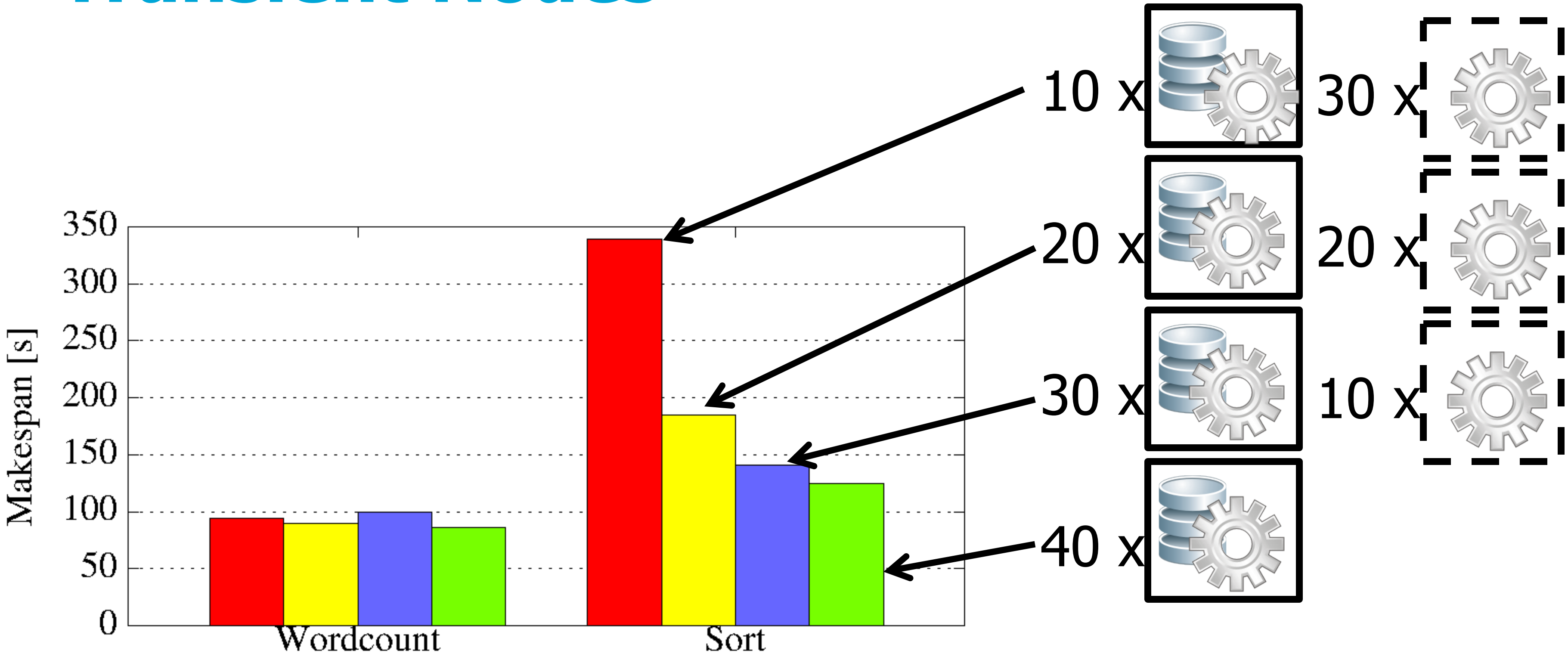


DAS-4 Infrastructure

- Real-world experiments on a multicluster system
- 6 clusters, over 1600 cores, 150 machines, 180 TB , 1-10 Gbit/s

Bogdan Ghit, Nezih Yigitbasi, Dick Epema.
Resource Management for Dynamic MapReduce Clusters in Multicluster Systems (Best Paper Award). *MTAGS 2012*

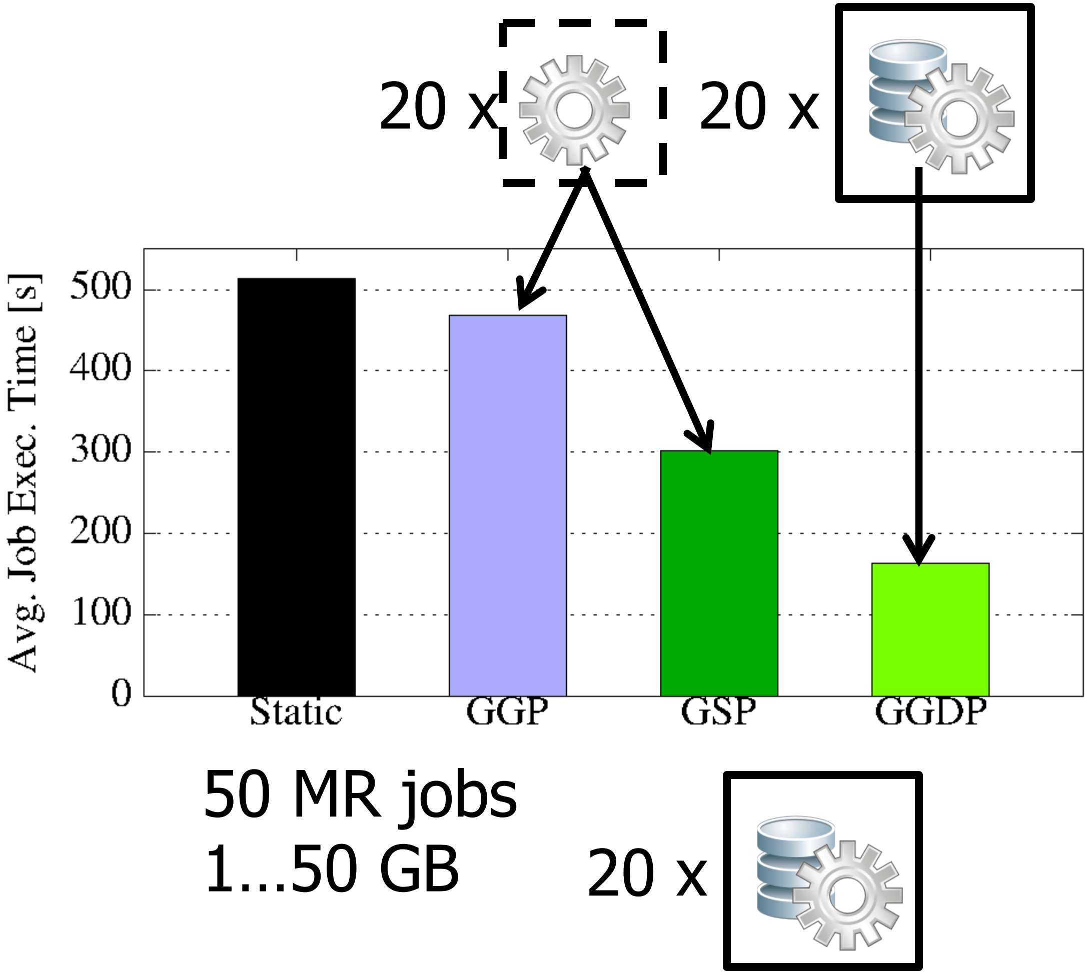
Transient Nodes



50 GB data set

- Wordcount scales better than Sort on transient nodes

Resizing Performance



50 MR jobs
1…50 GB

- Resizing bounds

  $F_{min} = 0.25$

  $F_{max} = 1.25$

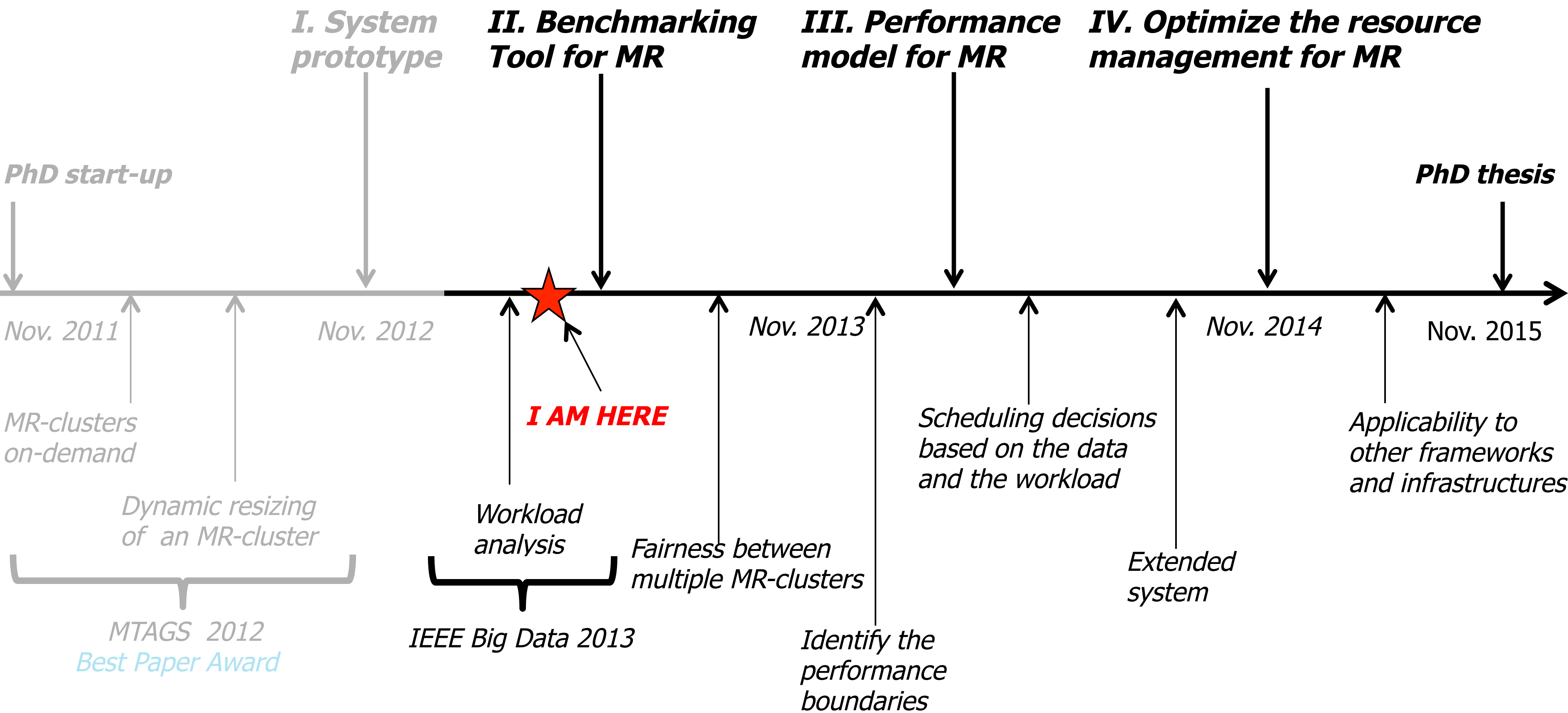
- Resizing steps
  - GSP

    $S_{grow} = 5$

    $S_{shrink} = 2$
  - GG(D)P

    $S_{grow} = 2$

Road Map



I. System prototype

II. Benchmarking Tool for MR

III. Performance model for MR

IV. Optimize the resource management for MR

PhD start-up

PhD thesis

Nov. 2011

Nov. 2012

Nov. 2013

Nov. 2014

Nov. 2015

MR-clusters on-demand

Dynamic resizing of an MR-cluster

MTAGS 2012
Best Paper Award

I AM HERE

Workload analysis

IEEE Big Data 2013

Fairness between multiple MR-clusters

Identify the performance boundaries

Scheduling decisions based on the data and the workload

Extended system

Applicability to other frameworks and infrastructures

Workload Analysis

Question: Which are the major MapReduce use cases?

- Google, Facebook, Yahoo!, Cloudera, Microsoft
  - ➢ Findings from 12 published production traces
  - ➢ Our analysis of other 4 production traces

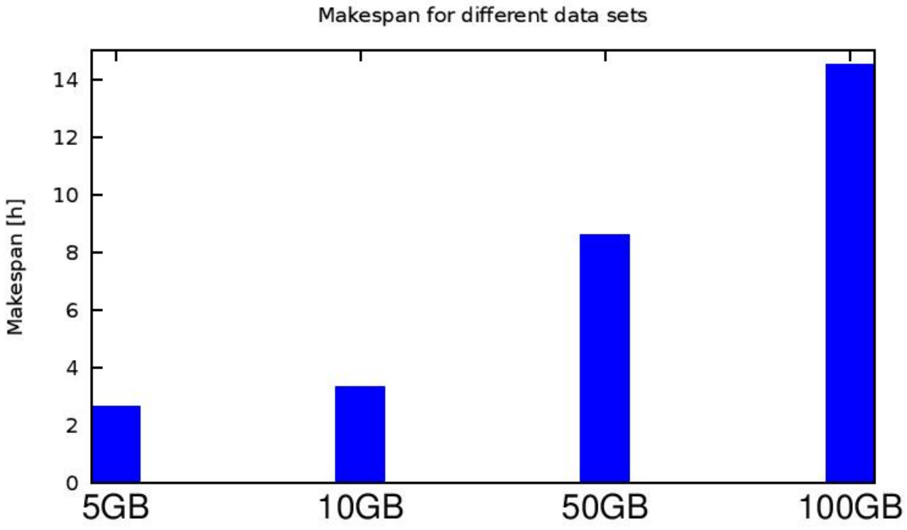
- Complex Workload
  - ➢ Large variations in job submissions rates
  - ➢ 90% of the jobs in all traces process and generate less than 1 GB, and complete in under 1 minute
  - ➢ For large jobs, variations in job sizes vs. job durations
  - ➢ Our PDS group analyzes 15 TB of BitTorrent logs with MapReduce

Benchmarking Tool

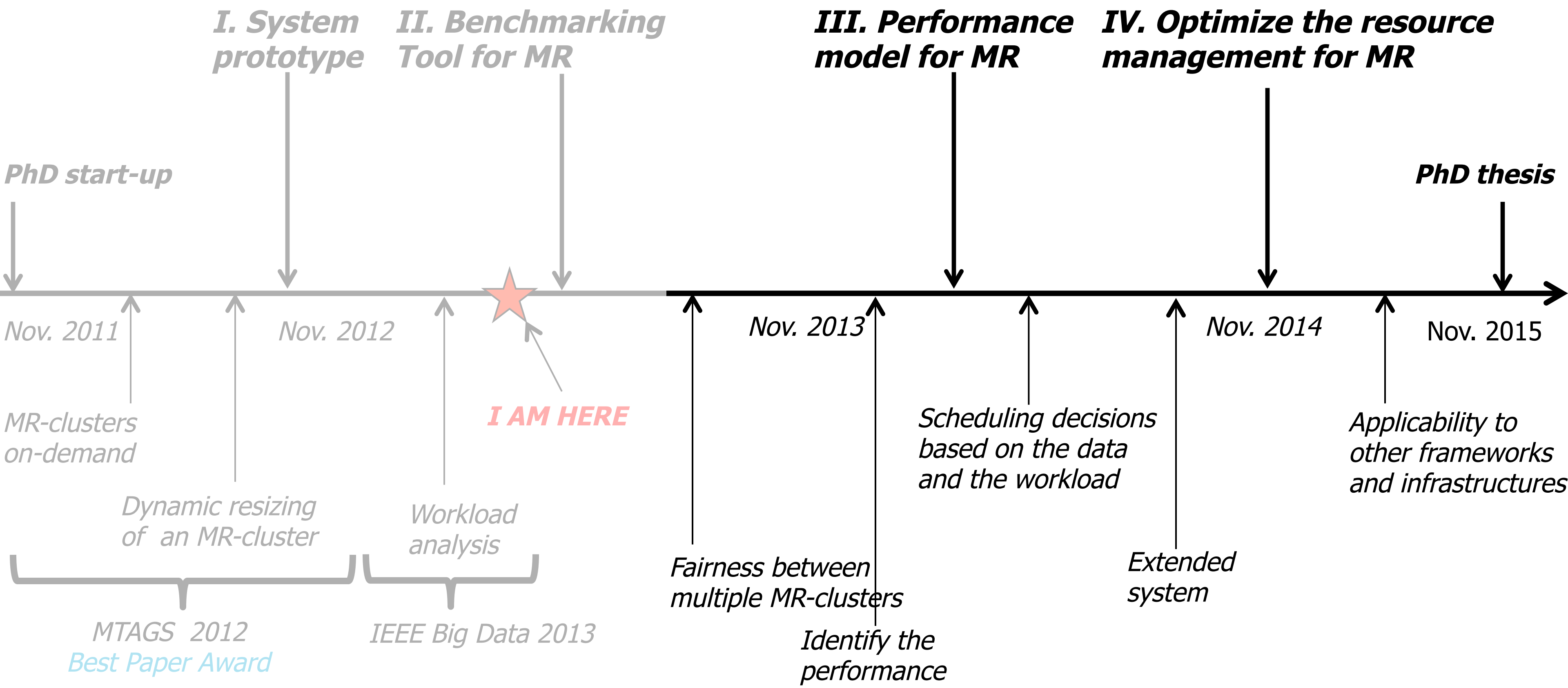
- Real-world applications
  - ➢ Text processing, web searching, machine learning

- Trace-based workloads
  - ➢ Analysis and modeling of traces from production clusters

- BTWorld use case
  - ➢ Complex MR workflow
  - ➢ 14 Pig queries / 33 MR jobs
  - ➢ Aggregations, selections, joins projections



Road Map



I. System prototype

II. Benchmarking Tool for MR

III. Performance model for MR

IV. Optimize the resource management for MR

PhD start-up

PhD thesis

Nov. 2011

Nov. 2012

I AM HERE

Nov. 2013

Nov. 2014

Nov. 2015

MR-clusters on-demand

Dynamic resizing of an MR-cluster

Workload analysis

MTAGS 2012
Best Paper Award

IEEE Big Data 2013

Fairness between multiple MR-clusters

Identify the performance boundaries

Scheduling decisions based on the data and the workload

Extended system

Applicability to other frameworks and infrastructures

Fair-Sharing Across Multiple Users

Question: Given multiple MR-clusters, how can you tell if one is working better than another ?

- Schedule and provision concurrent MR-clusters
- Differentiate users and converge to a division of resources such that they get similar performance

- Weighted proportional allocations:
  - Take snapshots in time of the queue sizes
  - Maintain a history of finished jobs

Provisioning Policies

Question: Can we obtain better performance with the dynamic MR-clusters?

- Data is hard to move
  - Aprox. 3 h to transfer 1 TB between HDFS and the local storage (900 Mbps write speed)
  - Removing a node with 100 GB makes ~ 6 failed jobs (1 Gbit/s Ethernet, avg. map task duration – 24 s, most jobs have less than 150 tasks)

- Explore a large space of policies:
  - Policies for establishing the weights (fair-shares)
  - Policies for growing (core or transient nodes, single or multiple clusters)
  - Policies for shrinking (preemptive or non-preemptive)

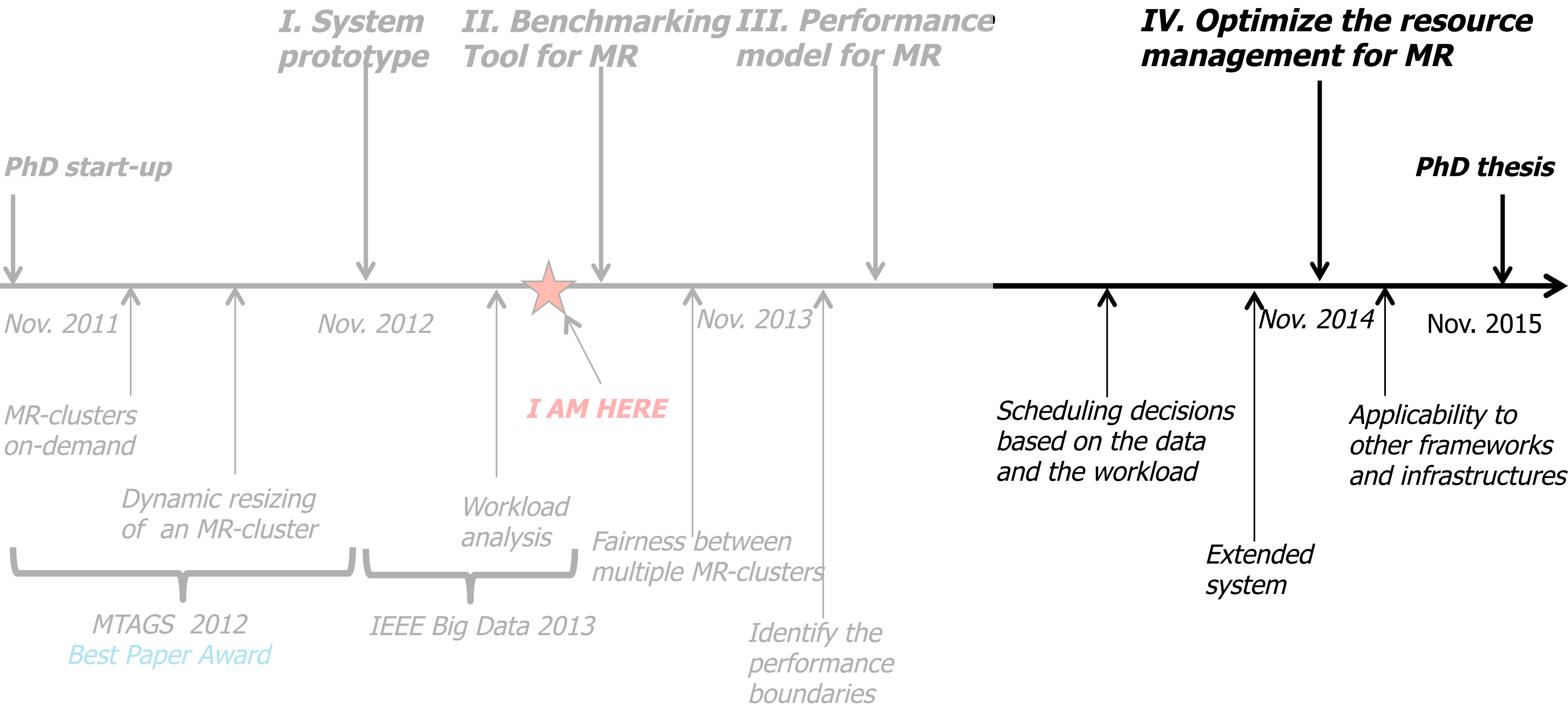
Performance Model

Question: Which are the performance boundaries of the MR processing system?

- Analytical and statistical methods
- Metrics:
  - Fairness – users get similar performance
  - Elasticity – dynamic MR clusters
  - Performace isolation – multiple MR clusters
  - Velocity of data processing
  - Adaptivity to data explosion

http://research.spec.org/

Road Map



PhD start-up

I. System prototype

II. Benchmarking Tool for MR

III. Performance model for MR

IV. Optimize the resource management for MR

PhD thesis

Nov. 2011

Nov. 2012

Nov. 2013

Nov. 2014

Nov. 2015

MR-clusters on-demand

Dynamic resizing of an MR-cluster

MTAGS 2012
Best Paper Award

Workload analysis

IEEE Big Data 2013

I AM HERE

Fairness between multiple MR-clusters

Identify the performance boundaries

Scheduling decisions based on the data and the workload

Extended system

Applicability to other frameworks and infrastructures

Optimize the MapReduce System

Question: Are the results obtained so far relevant for the large domain of data-processing systems?

- Provisioning policies with different optimization targets
- Incorporate knowledge about the workloads in scheduling and provisioning decisions
- Release the extended system with the full functionalities

- Investigate the applicability to other programming models and infrastructures

More Information

- Team: D. Epema, A. Iosup, M. Capotă, T. Hegeman, N. Yigitbasi, L. Fei,…

- PDS publication database
  - ➢ www.pds.ewi.tudelft.nl/research-publications/publications

- Home pages
  - ➢ www.pds.ewi.tudeltf.nl/ghit
  - ➢ www.pds.ewi.tudelft.nl/epema
  - ➢ www.pds.ewi.tudelft.nl/~iosup

- Web sites:

➢ KOALA: www.st.ewi.tudelft.nl/koala