Dynamic MapReduce Clusters on Demand

5th Workshop on Many-Task Computing on Grids and Supercomputers
Salt Lake City, Utah

Bogdan Ghit, Nezih Yigitbasi, and Dick Epema

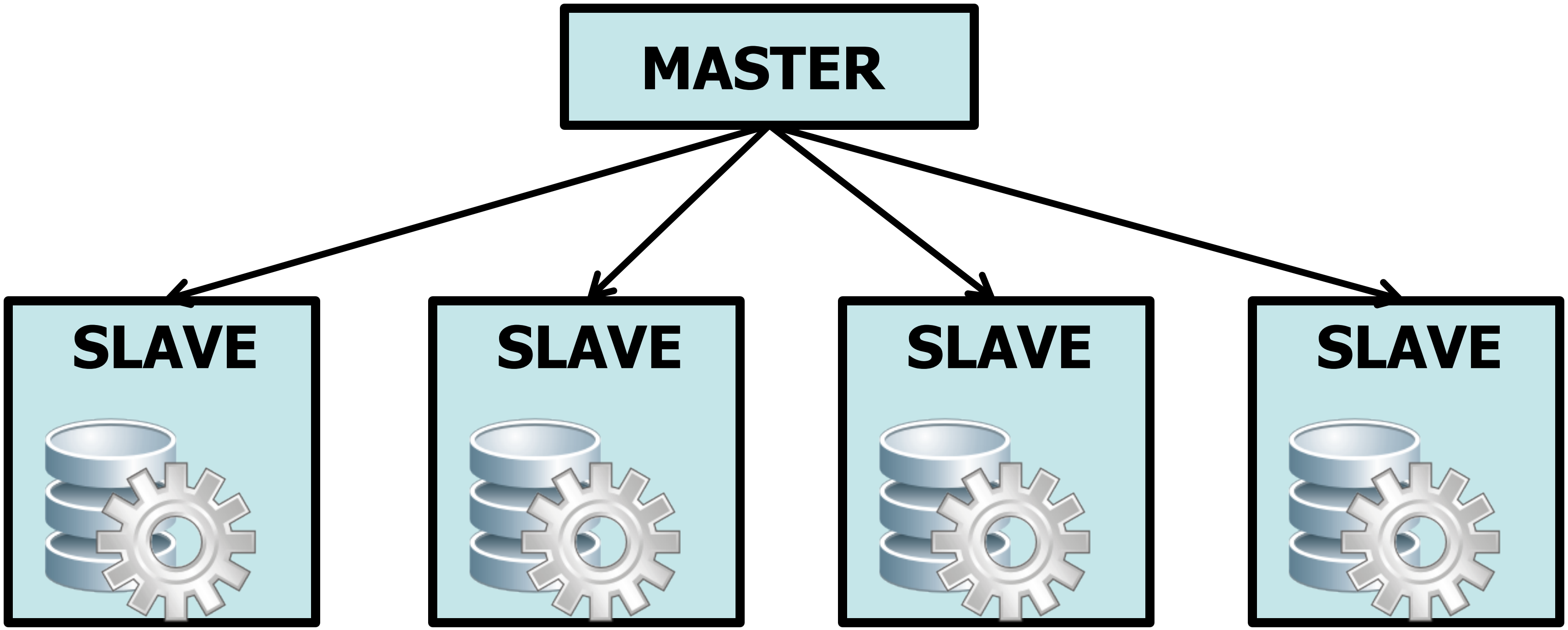
Parallel and Distributed Systems Group
Delft University of Technology
Delft, The Netherlands

COMMIT/

TU Delft
Delft University of Technology

MapReduce Overview

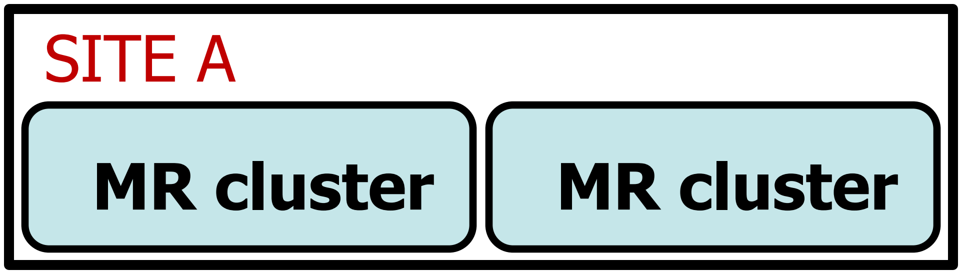
- MR cluster
  - Large-scale data processing
  - Master-slave paradigm
- Components
  - Distributed file system (storage)
  - MapReduce framework (processing)



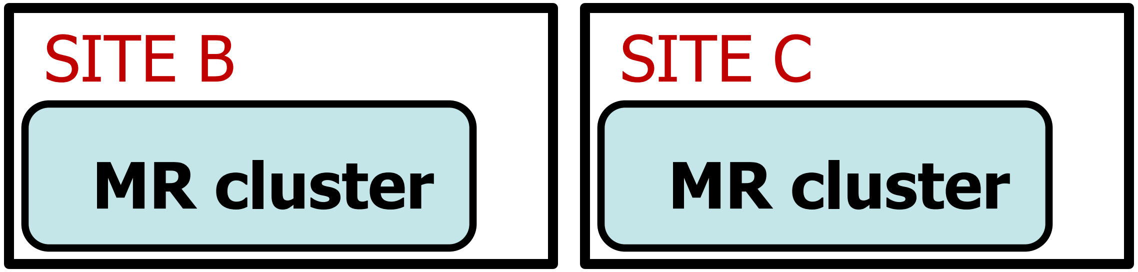
TUDelft

Why Multiple MapReduce Clusters?

- Intra-cluster Isolation

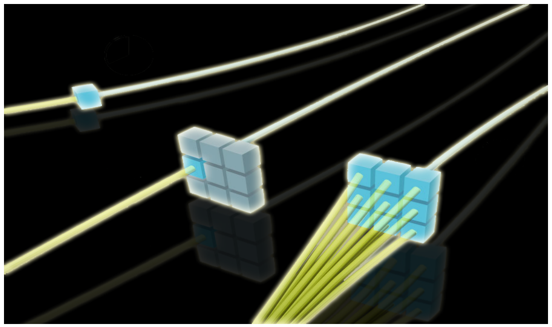


- Inter-cluster Isolation



Types of Isolation

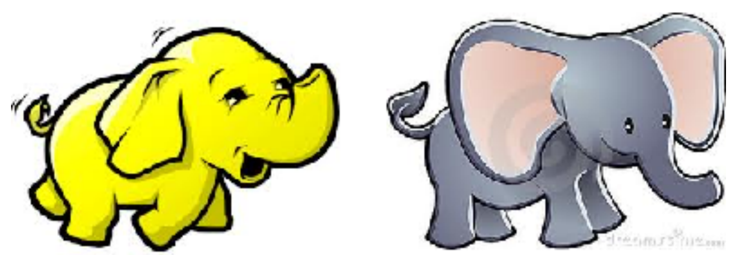
- Performance Isolation



- Failure Isolation

- Data Isolation

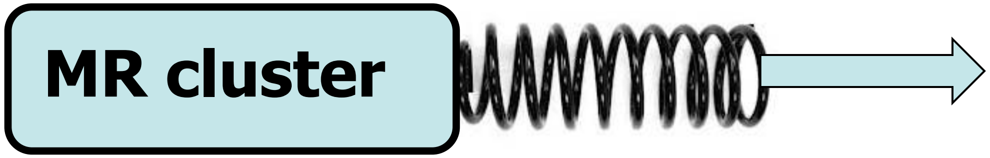
- Version Isolation



Why Dynamic MapReduce Clusters?

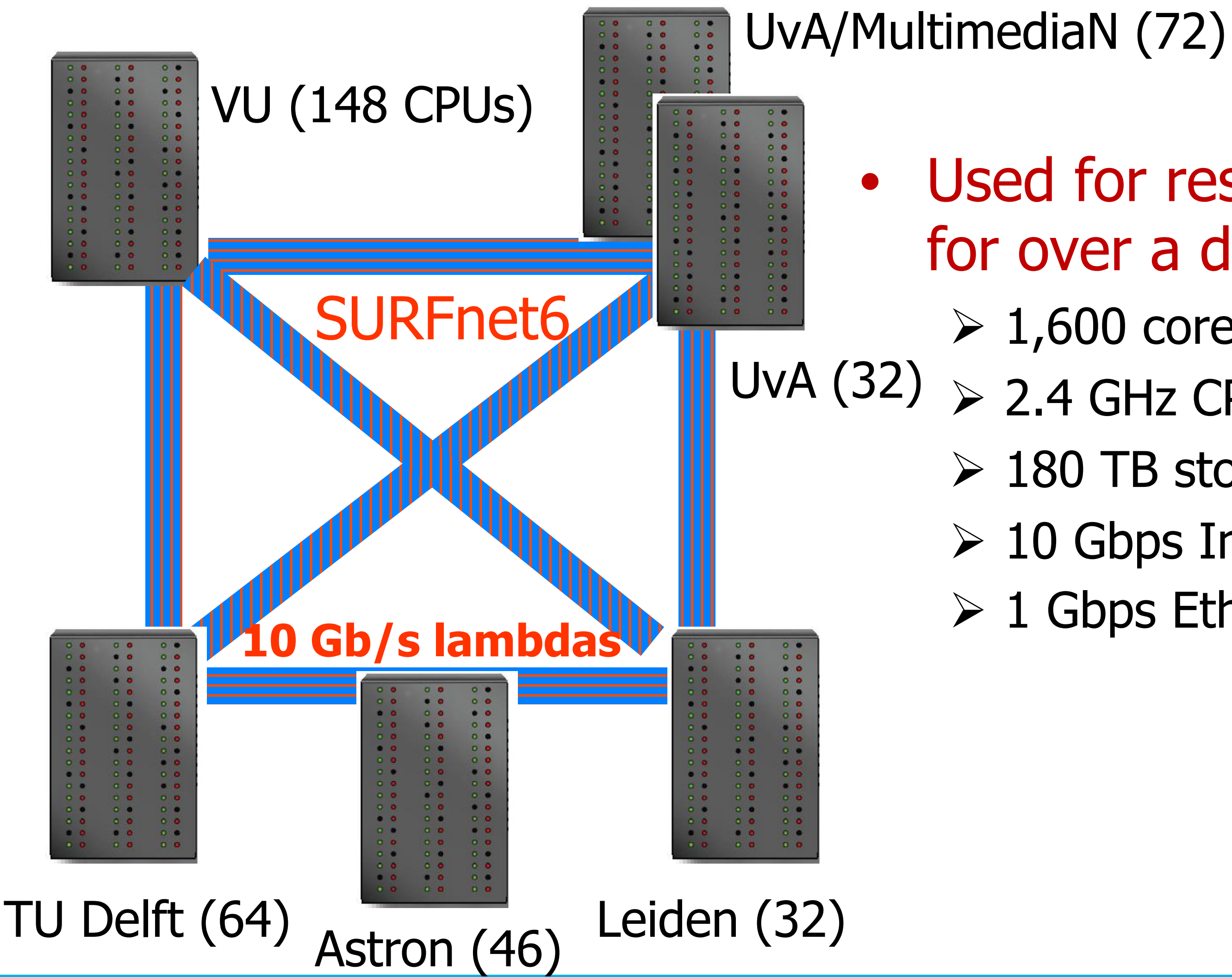
- Improve resource utilization
  - Grow when the workload is too heavy
  - Shrink when resources are idle

- Fairness across multiple MR clusters
  - Redistribute idle resources
  - Allocate resources for new MR clusters



TU Delft

The DAS-4 Infrastructure



- Used for research in systems for over a decade
  - 1,600 cores (quad cores)
  - 2.4 GHz CPUs, GPUs
  - 180 TB storage
  - 10 Gbps Infiniband
  - 1 Gbps Ethernet

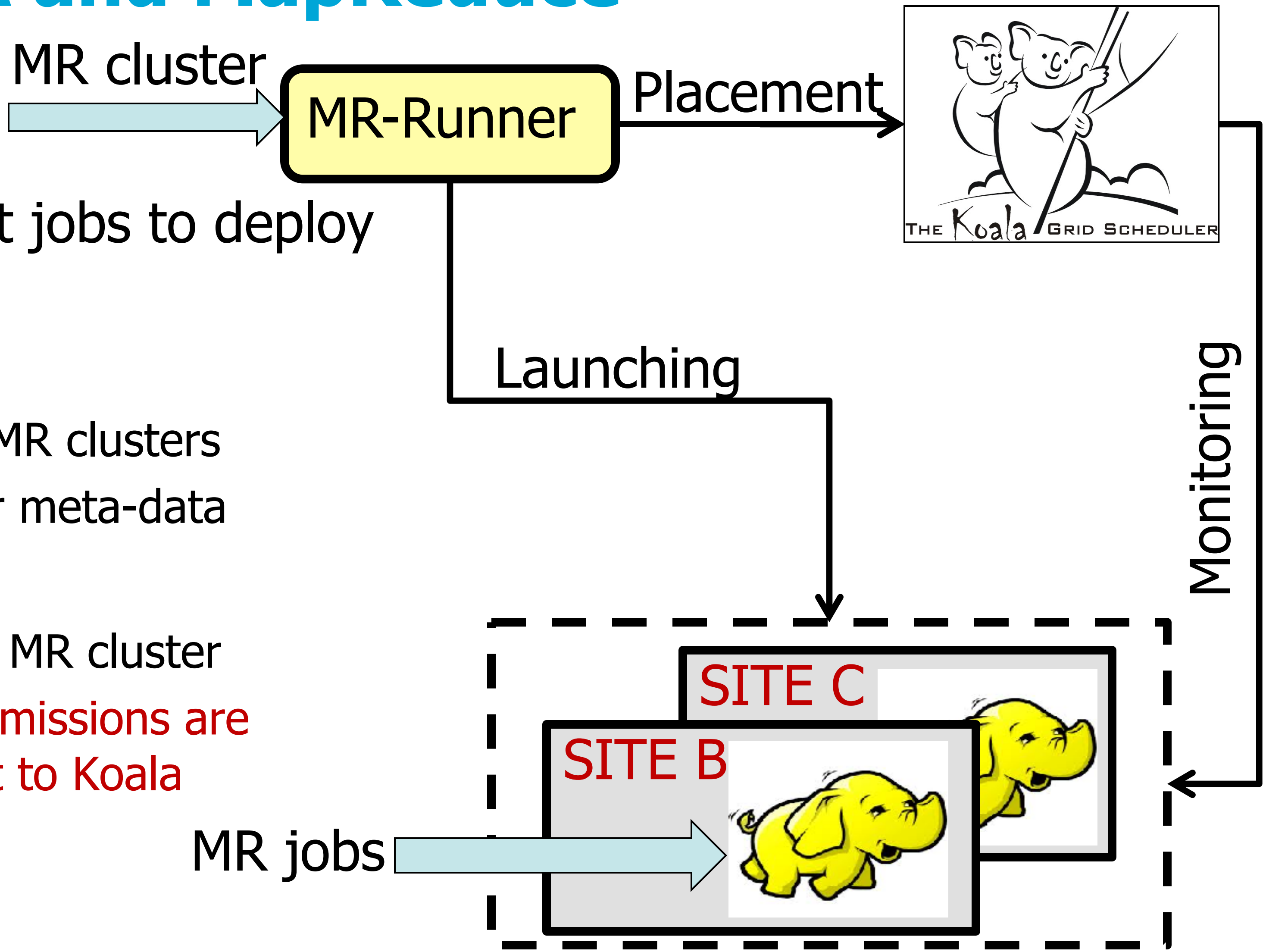
Koala Grid Scheduler

- Deployed on DAS-4
- Meta-scheduler, transparent for local schedulers
- Research vehicle in grid and cloud computing

- Features:
  - Resource co-allocation
  - Scheduling policies
  - Various application types

- Current runners:
  - CSRunner: cycle scavenging apps.
  - OMRunner: co-allocated OpenMPI apps.
  - Wrunner: co-allocated workflows
  - MR-Runner: MapReduce clusters

KOALA and MapReduce

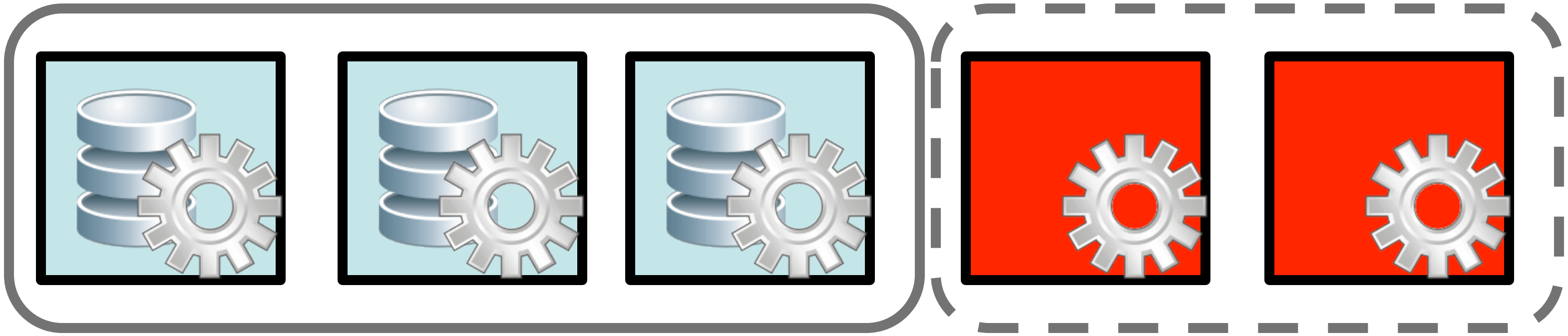


- Users submit jobs to deploy MR clusters
- Koala
  - ➢ Schedules MR clusters
  - ➢ Stores their meta-data
- MR-Runner
  - ➢ Installs the MR cluster
  - ➢ MR job submissions are transparent to Koala

TU Delft

System Model

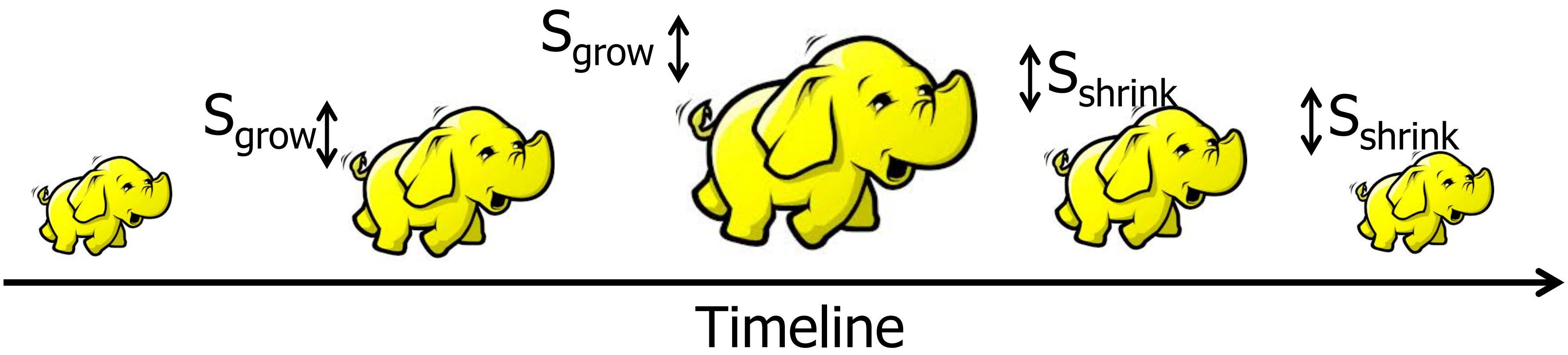
- Two types of nodes
  - Core nodes: TaskTracker and DataNode
  - Transient nodes: only TaskTracker



Resizing Mechanism

- Two-level provisioning
  - ➢ Koala makes resource offers / reclaims
  - ➢ MR-Runners accept / reject request

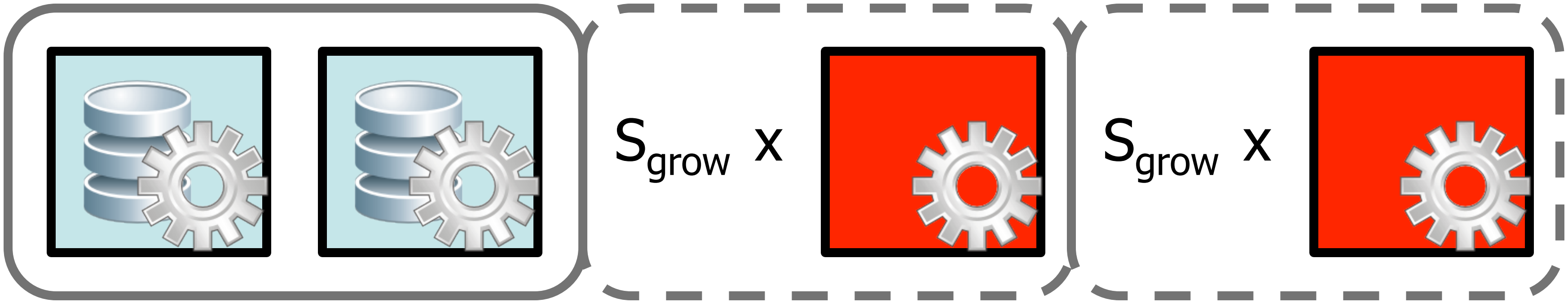
- Grow-Shrink Policy (GSP)
  - ➢ MR cluster utilization: $F_{\min} \leq \dfrac{totalTasks}{availSlots} \leq F_{\max}$
  - ➢ Size of grow and shrink steps: $\mathbf{S_{grow}}$ and $\mathbf{S_{shrink}}$



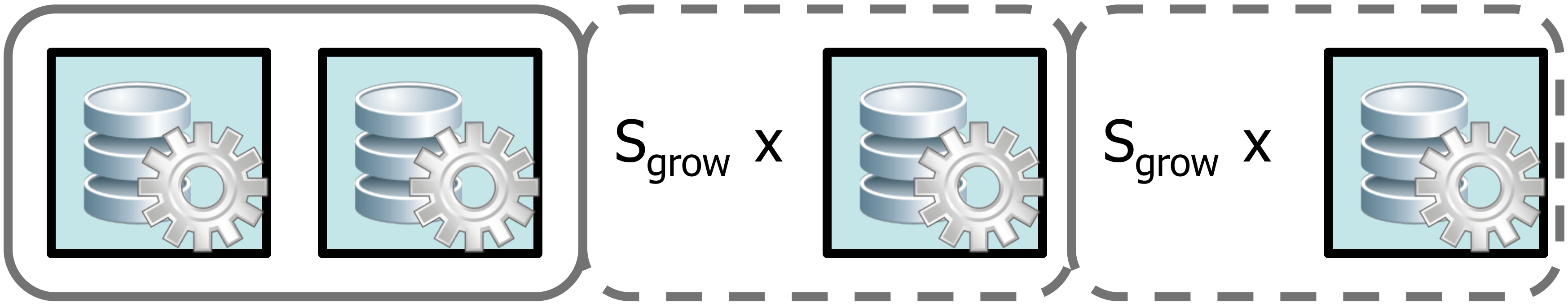
TU Delft

Baseline Policies

- Greedy-Grow Policy (GGP):



- Greedy-Grow-with-Data Policy (GGDP):



TU Delft

Setup

- *98% of jobs @ Facebook take less than a minute*
- *Google reported computations with TB of data*

- Two applications: Wordcount and Sort

<u>Workload 1</u>

- Single job
- 100 GB
- Makespan

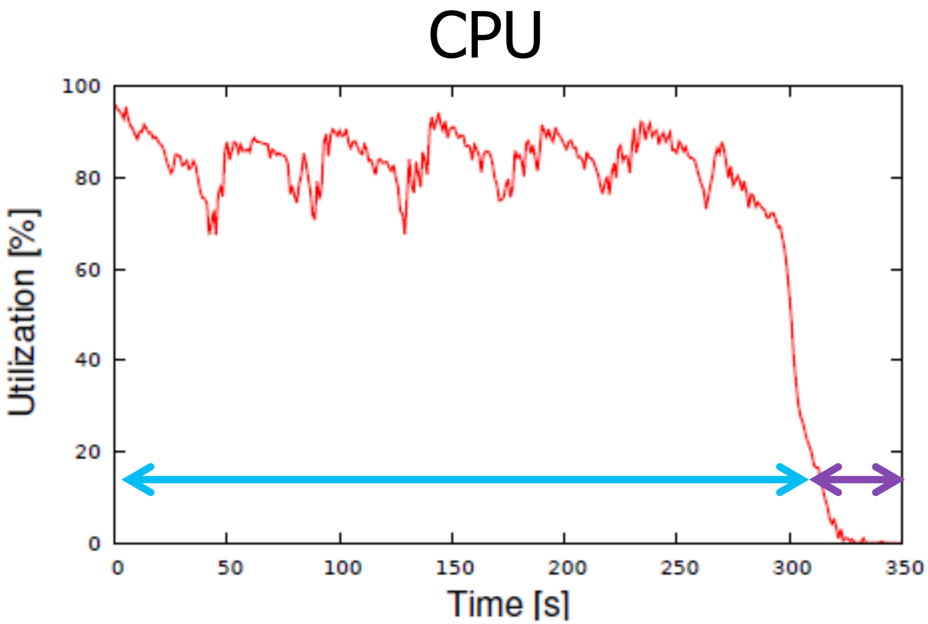
<u>Workload 2</u>

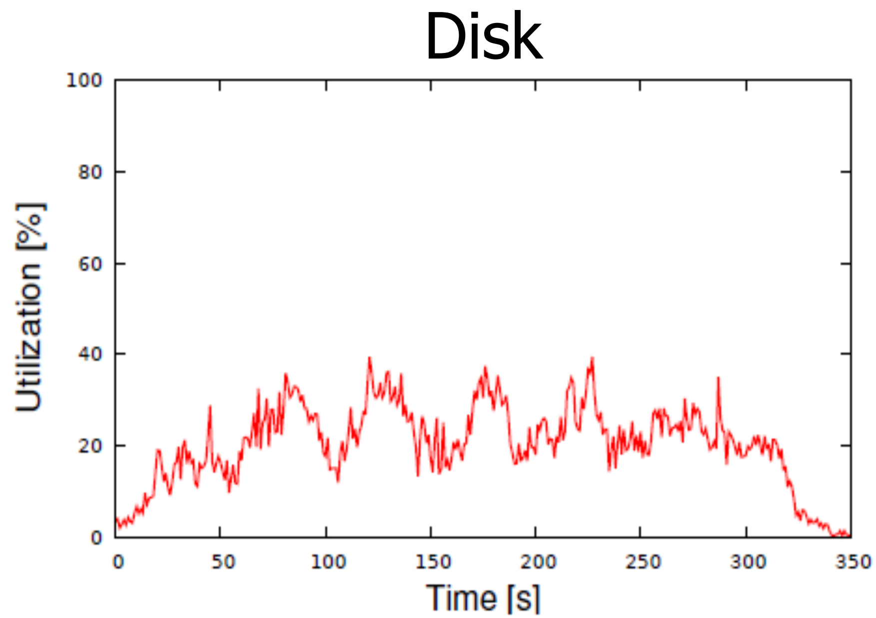
- Single job
- 40 GB, 50 GB
- Makespan

<u>Workload 3</u>

- Stream of 50 jobs
- 1 GB → 50 GB
- Average job execution time

Wordcount



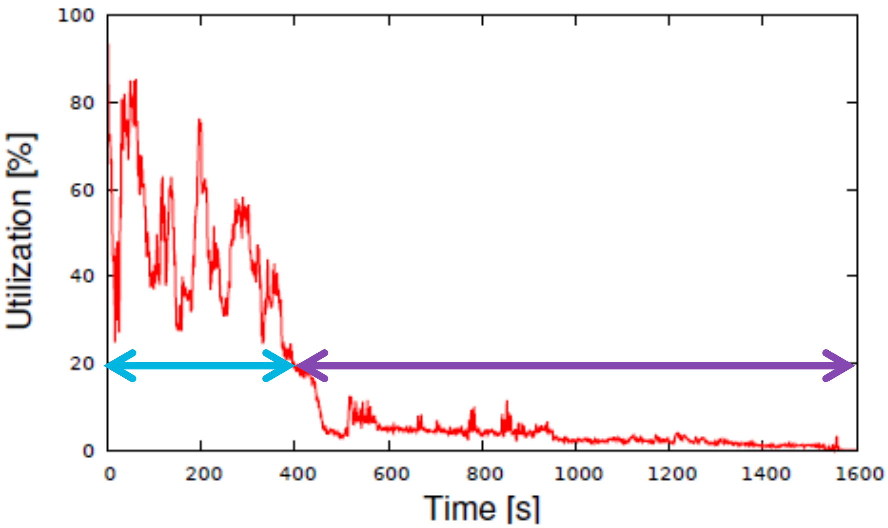


Workload 1 => 10 x 

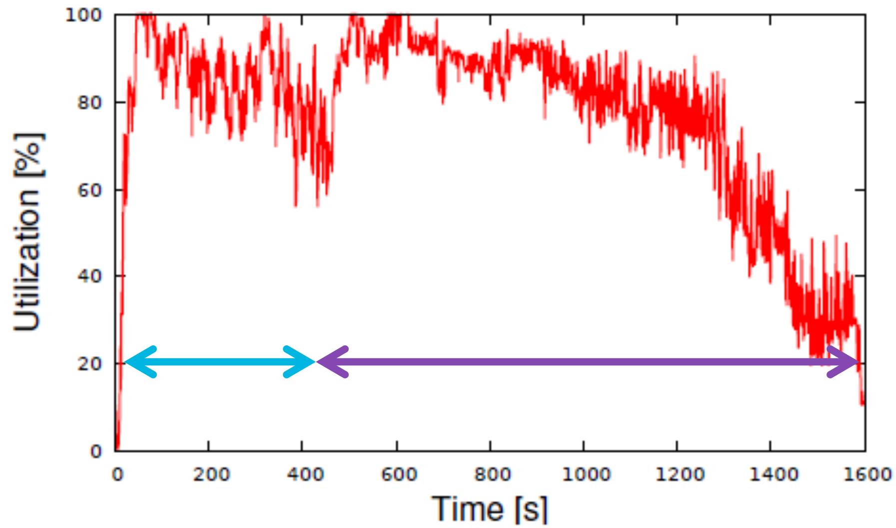
- Wordcount is CPU-bound in the map phase
- Short reduce phase with low CPU utilization

Sort

CPU



Disk

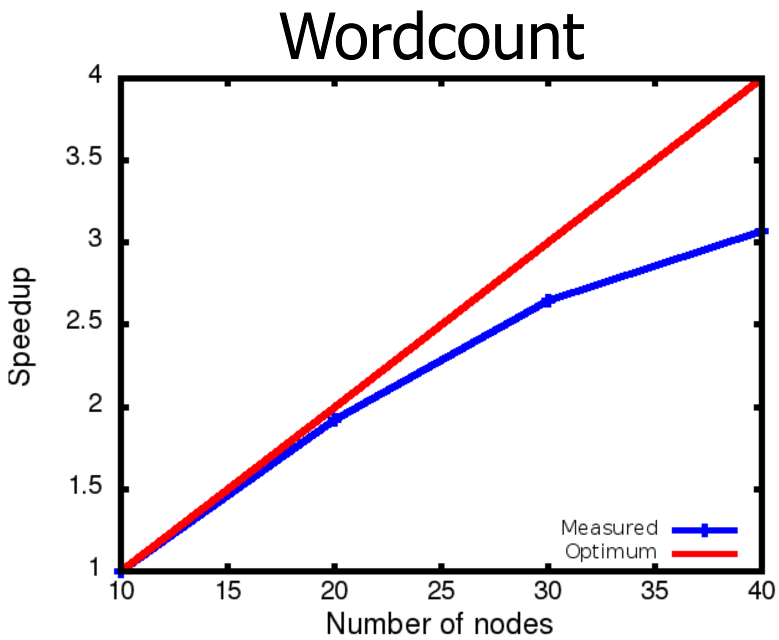


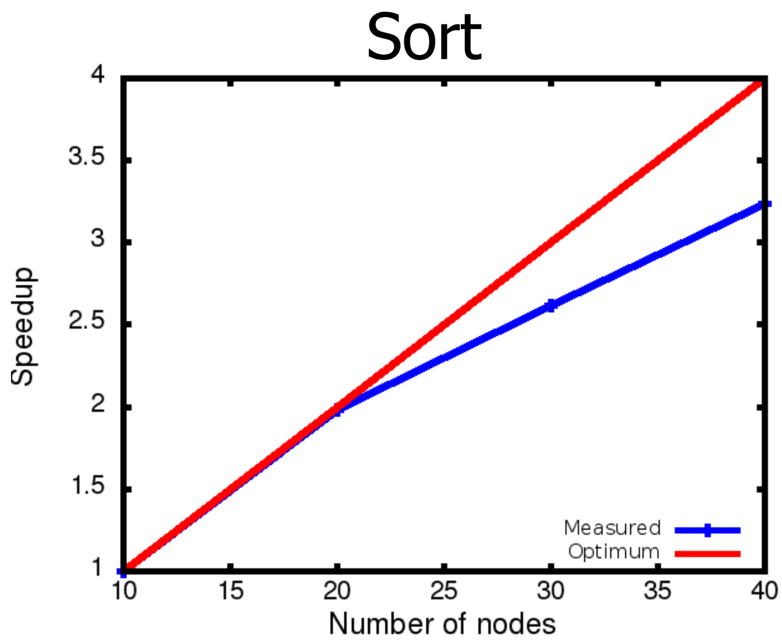
Workload 1 => 10 x 

- Short map phase with 40%-60% CPU utilization
- Long reduce phase which is highly disk intensive

TU Delft

Speedup



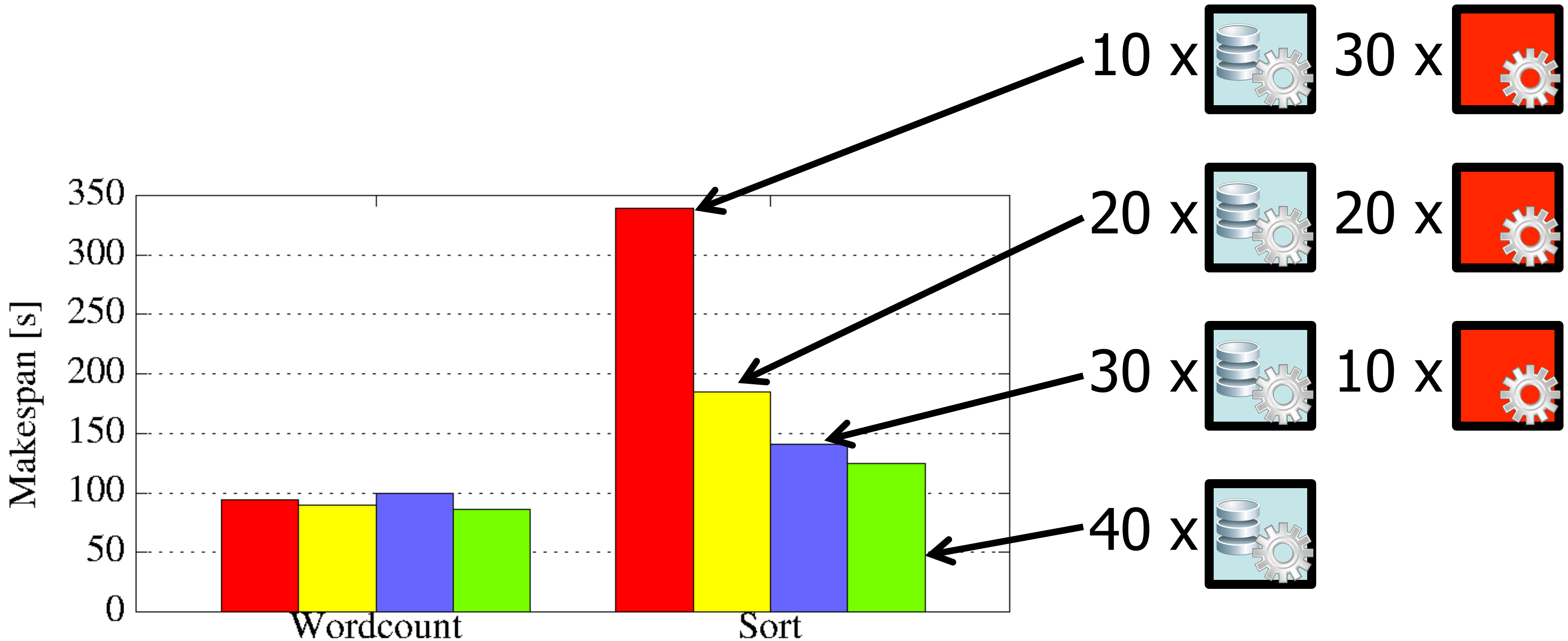


Workload 1 => {10,20,30,40} x 

- Speedup relative to an MR cluster with 10 core nodes
- Close to linear speedup on core nodes

TU Delft

Transient Nodes

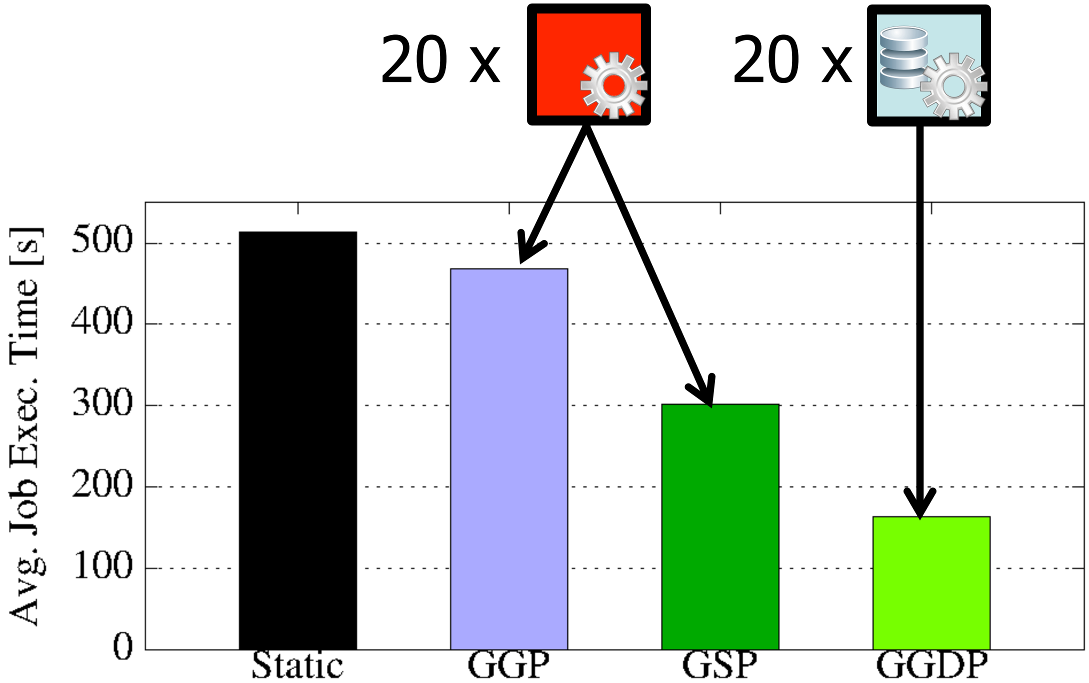


Workload 2

- Wordcount scales better than Sort on transient nodes

TUDelft

Resizing Performance



Workload 3 => 20 x

- Resizing bounds

  $F_{min} = 0.25$

  $F_{max} = 1.25$

- Resizing steps
  - GSP

    $S_{grow} = 5$

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

    $S_{grow} = 2$

Conclusions

- MR clusters on demand
  - System deployed on DAS-4
  - Resizing mechanism
- Performance evaluation
  - Single jobs workloads
  - Stream of jobs workload

- Distinct applications behave differently with transient nodes
- GSP reduces the job average execution time

- Future Work
  - More policies, more thorough parameter analysis

TU Delft

More Information

- Team: D. Epema, A. Iosup, N. Yigitbasi, S. Shen, Y. Guo, …

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

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

- Web sites:
  - KOALA: www.st.ewi.tudelft.nl/koala

THE Koala GRID SCHEDULER