www.bsc.es

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

ALOJA: Cost-effective Big Data deployments

Nicolas Poggi, Senior Researcher

EXCELENCIA SEVERO OCHOA

BSC~Microsoft Research Centre

February 2015

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

INSTITUTIONAL
ABOUT THE PROJECT

Barcelona Supercomputing Center (BSC)

- 22 year history in Computer Architecture research
  - European Center for Parallelism of Barcelona (CEPBA)
    - Based at the Technical University of Catalonia (UPC) in 1991
  - Long track record with IBM in chip Architecture & Parallelism

- Led by Mateo Valero
  - ACM fellow, Eckert-Mauchly award 2007, Goode award 2009
  - Active research staff with 1000+ publications
  - Large ongoing life science computational projects
    - Computational Genomics
    - Molecular modeling & Bioinformatics
    - Protein Interactions & Docking
  - In place computational capabilities
    - Mare Nostrum Super Computer

Marenostrum Supercomputer

- Prominent body of research activity around Hadoop since 2008
  - Previous to **ALOJA**
    - SLA-driven scheduling (Adaptive Scheduler), in memory caching, etc.
  - Research group page: http://www.bsc.es/autonomic

BSC-MSRS Centre and ALOJA

BSC Microsoft Research Centre

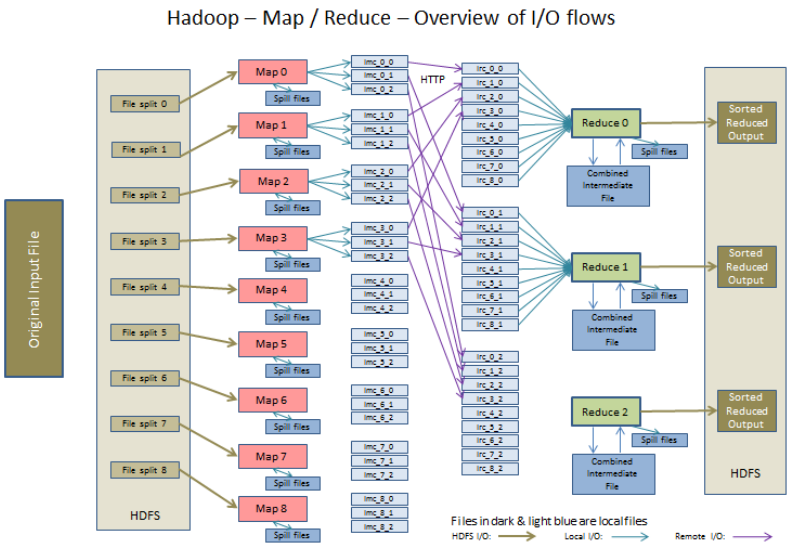
- Long-term relationship between
  - BSC and Microsoft Research and Microsoft product teams
- Previous research at the intersection of computer architecture, language implementation, and systems software, and performance profiling
- Open model:
  - **No patents, public IP, publications and open source main focus**
  - 87 publications, 4 Best paper awards

- **ALOJA** is the latest phase of the engagement
- With intent to explore:
  - upcoming hardware architectures
  - and building automated mechanism
- for deploying cost-effective Hadoop clusters.

BSC-MSRS Centre

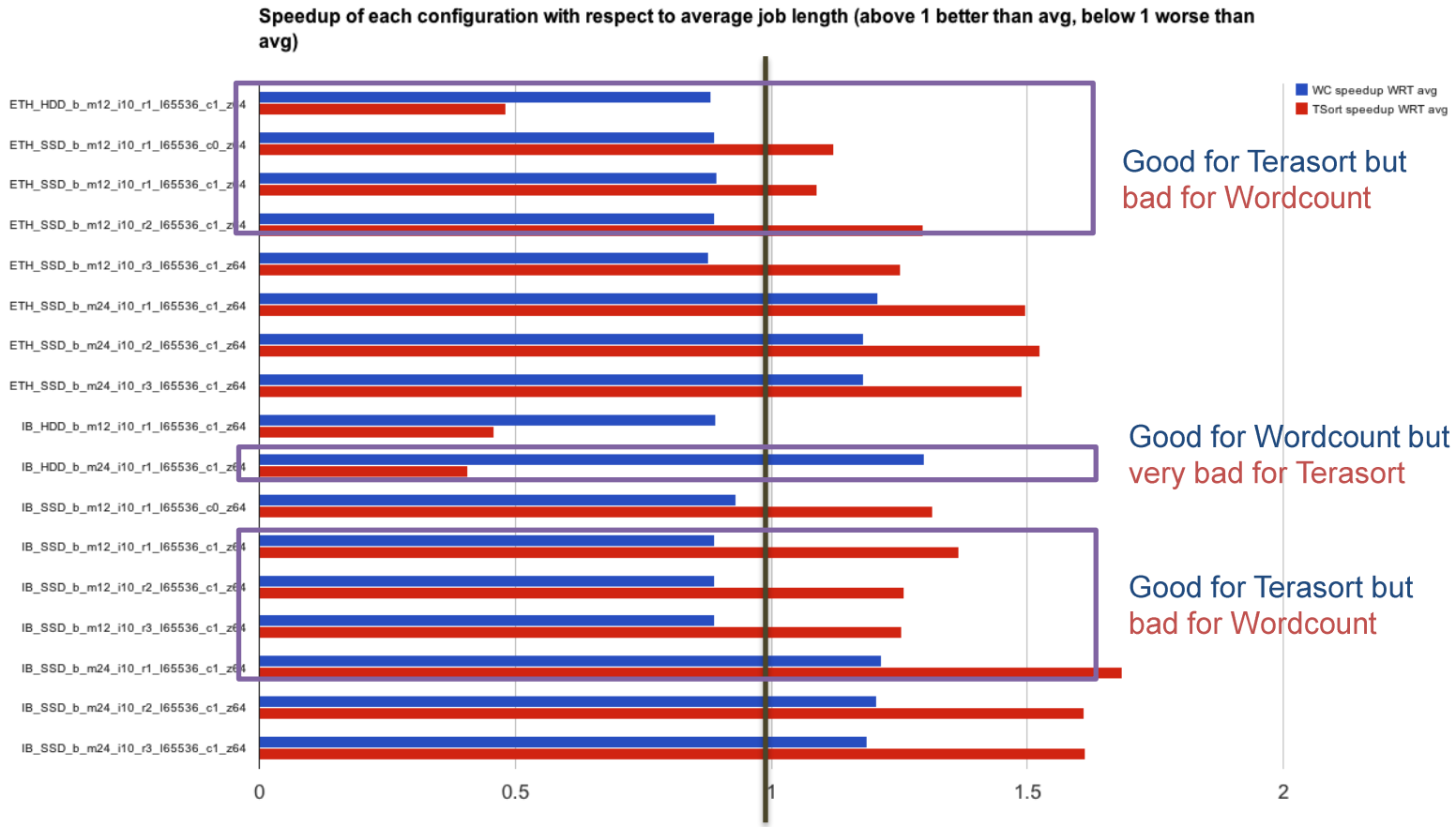
Motivation

- The Hadoop framework implements a complex distributed execution model
  - Over **100** interrelated config parameters
  - Requires manual iterative benchmarking and tuning
- Early results show that Hadoop's price/performance
  - are affected by relatively simple **SW >3x**
  - and **HW** configuration choices **> 3x**
- Commodity HW no longer low-end
  - new affordable hardware from original design (ie., *SSDs*)
  - Hadoop performs poorly on scale-up
  - or low power HW
- **New Cloud services** for Hadoop
  - **IaaS** and **PaaS**
  - Direct vs. remote attached volumes
- Spread Hadoop ecosystem
  - Dominated by vendors
  - Lack of verifiable benchmarks



Early exploration: Job resource configs

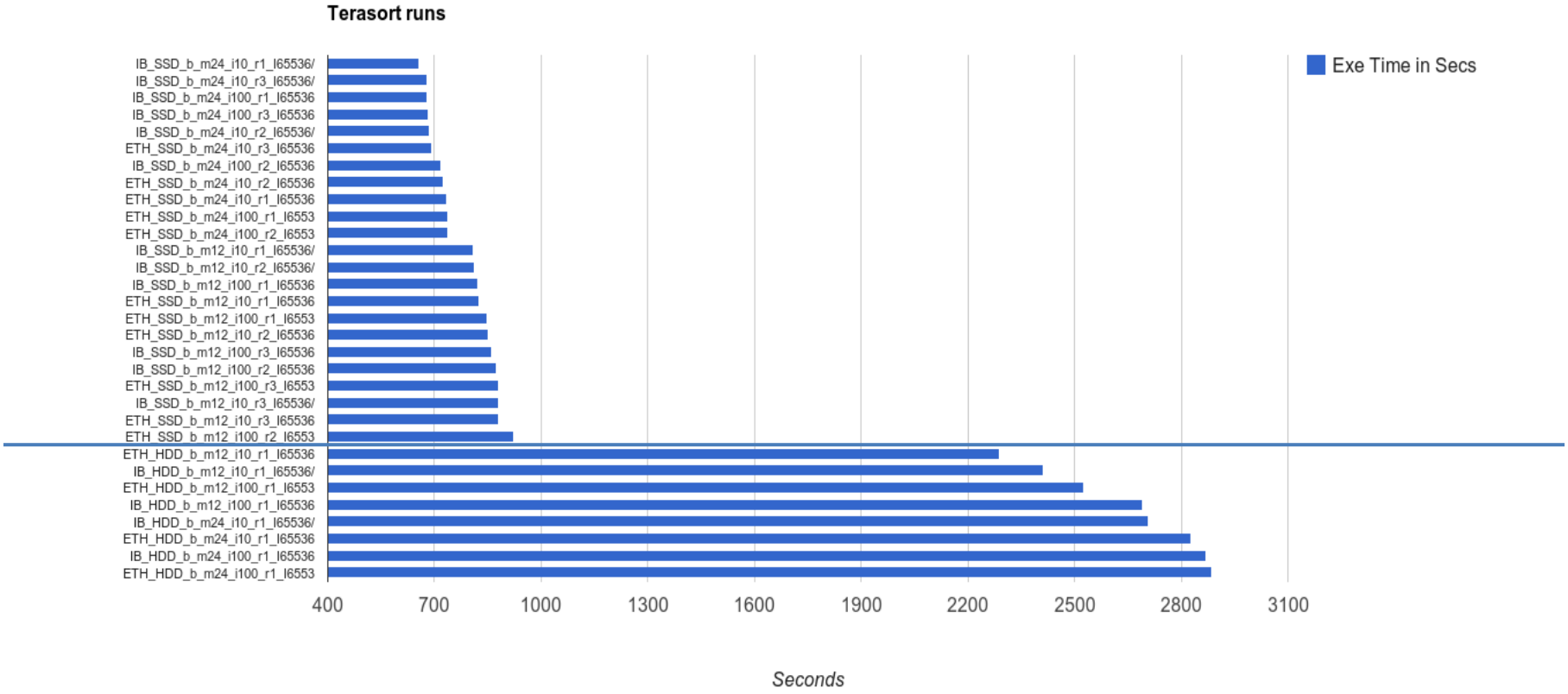
Is there one software configuration iteration that fits everybody?



Vertical line: Average performance for this workload across configurations
Values to the right: above average
Values to the left: below average

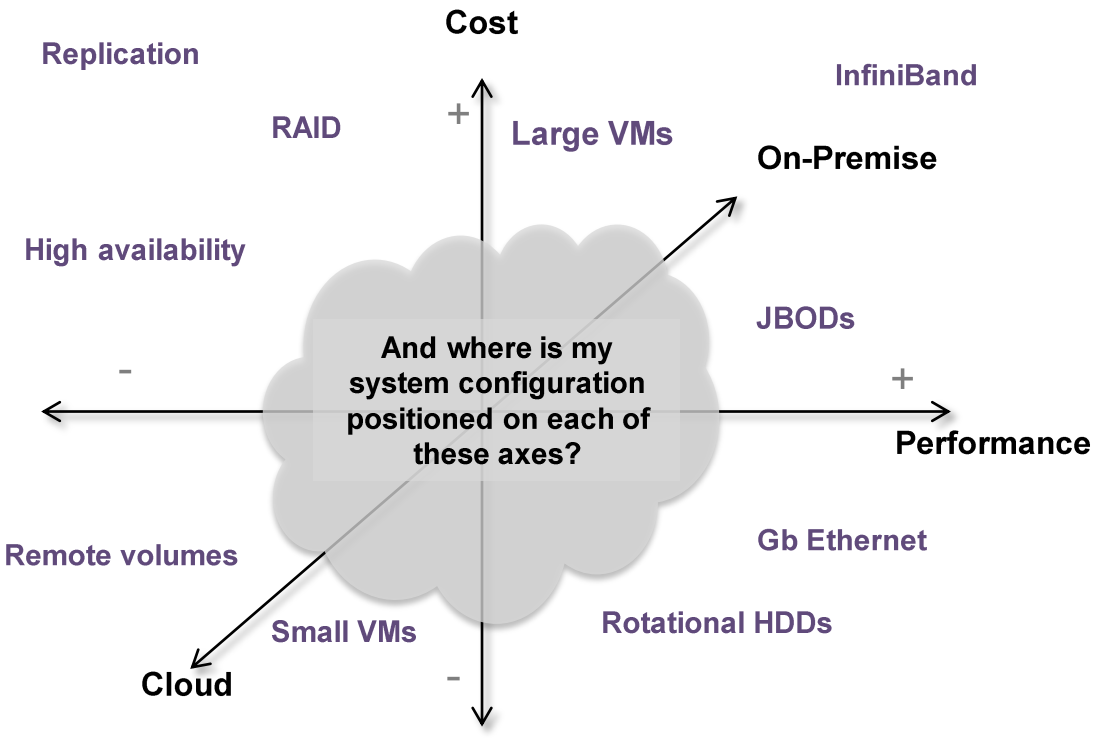
Early exploration: HW technology impact

- Impact of SSDs to running time of Terasort

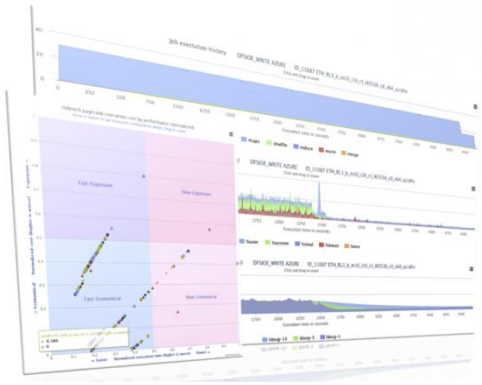


Current scenario and problematic

- What is the most cost-effective configuration for my needs?
  - Multidimensional problem



Project ALOJA



- Joint initiative to produce mechanisms for an
  - automated characterization of cost-effectiveness
  - of Big Data deployments
- Results from of a growing need of the community to understand job execution details
- Explore different configuration deployment options and their tradeoffs
  - Both software and hardware
  - Cloud services and on-premise
- Seeks to provide knowledge, tools, and an online service
  - to with which users make better informed decisions
  - reduce the TCO for their Big Data infrastructures
  - Guide the future development and deployment of Big Data clusters and applications

ALOJA Project phases

1. Systematic study of Hadoop runtime executions
   - across a range of hardware components
   - software parameters, job types,
   - and deployment patterns
2. Analytical models of Hadoop cost-effectiveness
   - Price vs. performance evaluation
   - Expand exploration in Cloud: IaaS (HDI) vs. PaaS, storage
     - VM flavor and cluster characterization
3. Automation and Prediction
   - Modeling and Prediction of executions
     - Minimize # of executions
     - Job similarity characterization
   - Online learning of configuration and recommendation
   - …

We are here


BSC
Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación
PLATFORM

ALOJA Platform: Evolution and status

- Benchmarking, Repository, and Analytics tools for Big Data

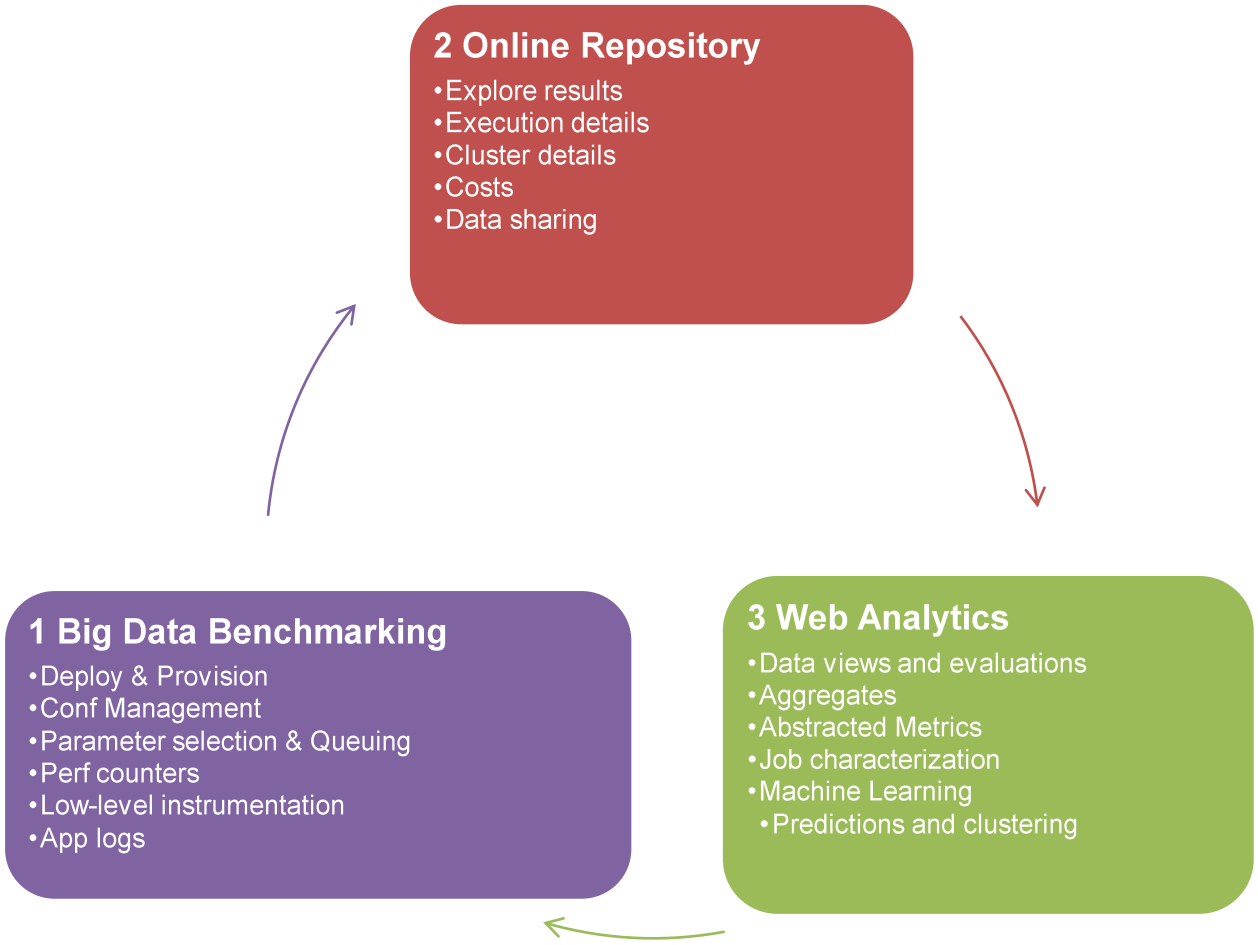
- Composed of open-source
  - Benchmarking, provisioning and orchestration tools,
  - high-level system performance metric collection,
  - low-level Hadoop instrumentation based on **BSC Tools**
  - and Web based data analytics tools
    - And recommendations

- Online Big Data Benchmark repository of:
  - **8000+** runs (from HiBench)
  - Sharable, comparable, repeatable, verifiable executions

- Abstracting and leveraging tools for BD benchmarking
  - Not reinventing the wheel but,
  - most current BD tools designed for production, **not for benchmarking**
  - leverages current compatible tools and projects
    - When possible

- Dev VM toolset and sandbox
  - via Vagrant

ALOJA Platform main components



1.) BIG DATA BENCHMARKING TOOLS

1.) Big Data Benchmarking

- **ALOJA-DEPLOY** Composed of scripts to:
  - Automatically create, stop, delete clusters in the cloud
    - From a simple and abstracted node and cluster definition files
    - Both for Linux and Windows
    - IaaS and PaaS (HDInsight)
    - Abstracted to support multiple providers
  - Provision and configuration of base software to servers
    - Both for cloud based as on premise
    - Composed of portable configuration management scripts
    - Designed for benchmarking needs
  - Orchestrate benchmark executions
    - Prioritized job queues
    - Results gathering and packaging

- **ALOJA-BENCH**
  - Multi-benchmark support
  - Flexible performance counter options
  - Dynamic SW and HW configurations

Node and Cluster definitions

- The configuration files can be used with the scripts to automate the definition of clusters (at provisioning stage)
- Example: **cluster_al-14**.conf

```
defaultProvider="azure”
clusterID='14' #from 03 0 99
clusterName="al-${clusterID}"
numberOfNodes="8" #starts at 0 (max 99) 0 is assigned to master
vmSize='extralarge' #extralarge are A4s
attachedVolumes="3"
diskSize="512”
queueJobs="true" #enable on cluster config to queue benchmarks after deploy
vmCores="8"
vmRAM="14GB”
clusterCostHour="2.664"
clusterType="IaaS"
```

Sources: https://github.com/Aloja/aloja/tree/master/shell/conf

Provisioning scripts

https://github.com/Aloja/aloja/tree/master/aloja-deploy

cache	Adding cache dir	2 months ago
include	Changed variable name	a month ago
providers	Added a comment to quickly edit host line	23 hours ago
README.md	Changes for multi cloud provider2	3 months ago
connect_cluster.sh	Change in providers and default values	3 months ago
connect_node.sh	Changes for multi cloud provider2	3 months ago
delete_cluster.sh	Calculating total time	3 months ago
delete_node.sh	Changes for multi cloud provider2	3 months ago
deploy_cluster.sh	Renaming of global vars	3 months ago
deploy_node.sh	Download and install ARM JDK	3 months ago
start_cluster.sh	Calculating total time	3 months ago
start_node.sh	Improvements in cluster deployments	3 months ago
stop_cluster.sh	Calculating total time	3 months ago
stop_node.sh	Multi provider cleanup	3 months ago
sync_node.sh	Added a new command to sync code changes without deploy	a month ago

Running benchmarks in ALOJA

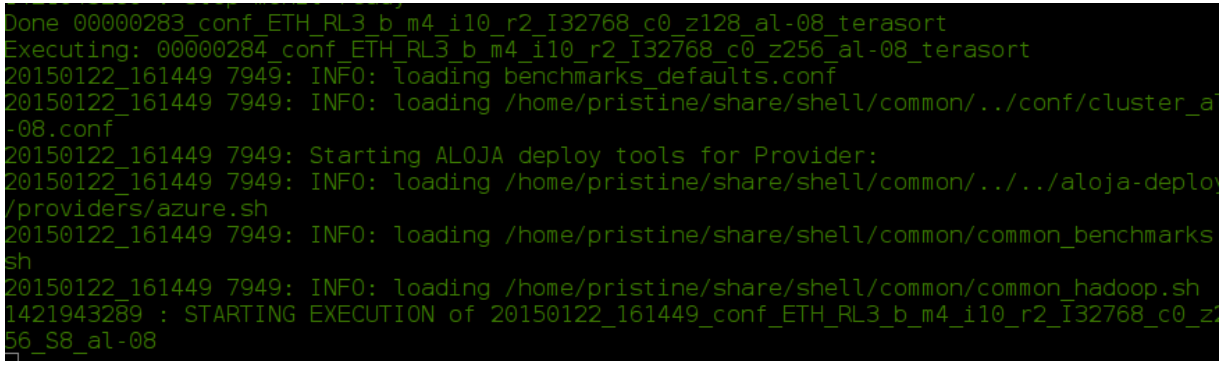
- https://github.com/Aloja/aloja/tree/master/shell

- Example of submitting a job to run:
  - https://github.com/Aloja/aloja/blob/master/shell/run_benchmarks.sh

```
run_bench marks.sh -C al-04 -n IB -d HDD -r 1 -m 12 -i 10 -p 3 -b -min -I 4096 -l wordcount -c 1
```

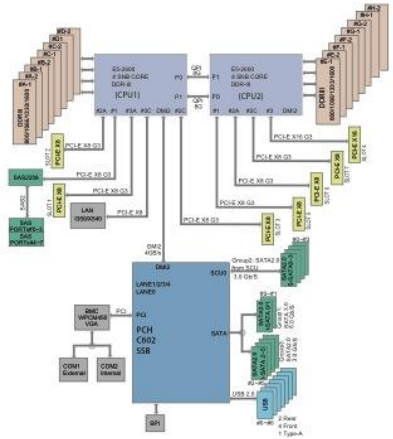
*al-04 cluster must be previously provisioned with the provisioning scripts

- Controls de Jobs in execution:
  - https://github.com/Aloja/aloja/blob/master/shell/exeq.sh



Initial testing infrastructure

- High-End Cluster:
  - 4 nodes, 12 real cores, 128GB RAM, 6x SSDs , 2x 56Gb InfiniBand, 4Gb GbE (bonding)
- Mid-end Cluster:
  - 18 nodes, 12 real cores, 64GB RAM, 1x SSD, 6x HDDs, 1Gb GbE
    - Evaluating different number of datanodes performance
- Cloud **IaaS** (Azure)
  - 1 head node 8 datanodes of A3, A4, A6, A7 VMs
- Cloud **PaaS** (HDInsight)
  - 4, 8, 16, 32 datanodes
- Low-powered cluster:
  - 10-node ARM based cluster*



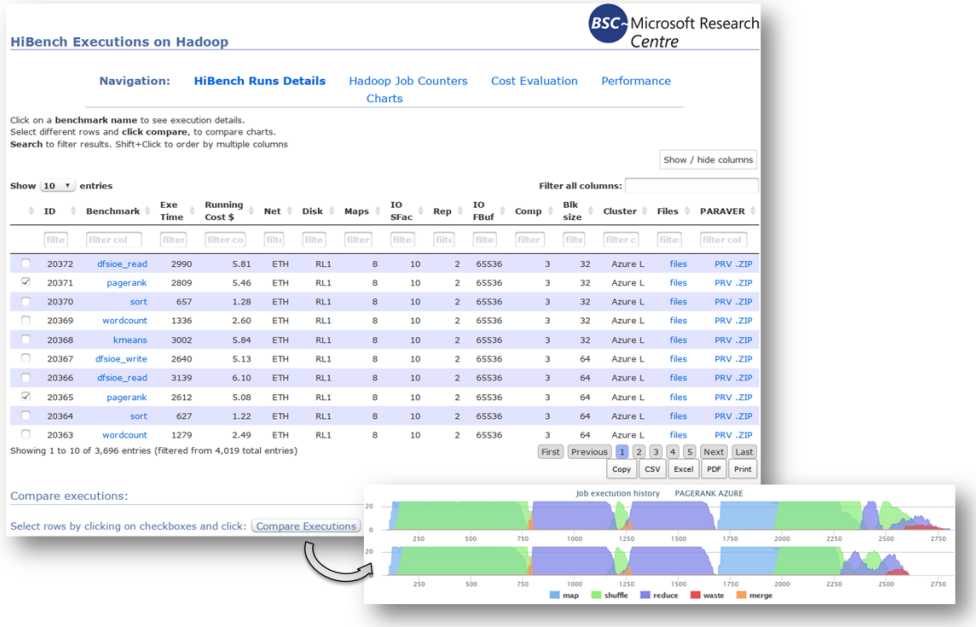
* Result numbers not online yet

ONLINE BENCHMARK REPOSITORY

2.) ALOJA-WEB Online Repository

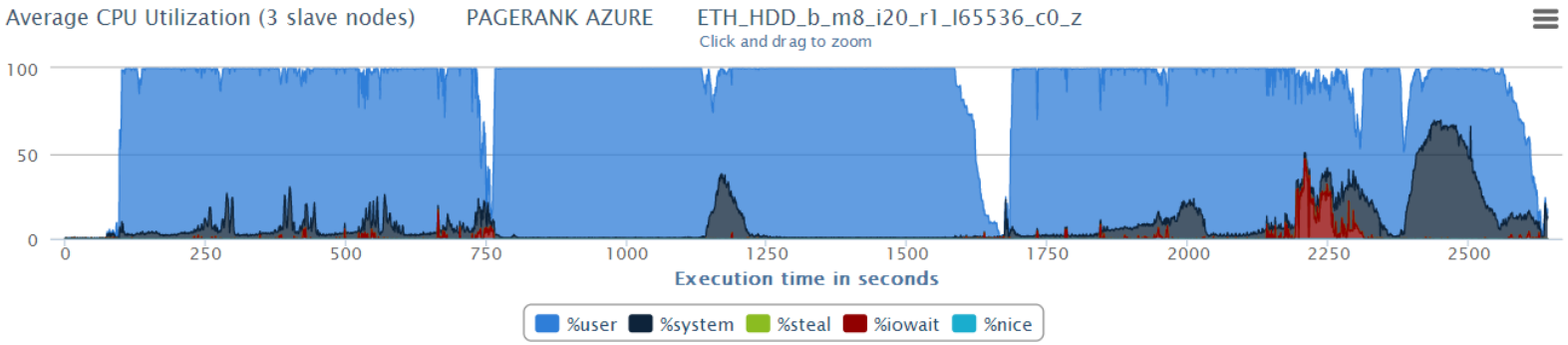
- Entry point for explore the results collected from the executions
  - Index of executions
    - Quick glance of executions
    - Searchable, Sortable
  - Execution details
    - Performance charts and histograms
    - Hadoop counters
    - Jobs and task details
- Data management of benchmark executions
  - Data importing from different clusters
  - Execution validation
  - Data management and backup
- Cluster definitions
  - Cluster capabilities (resources)
  - Cluster costs
- Sharing results
  - Download executions
  - Add external executions
- Documentation and References
  - Papers, links, and feature documentation

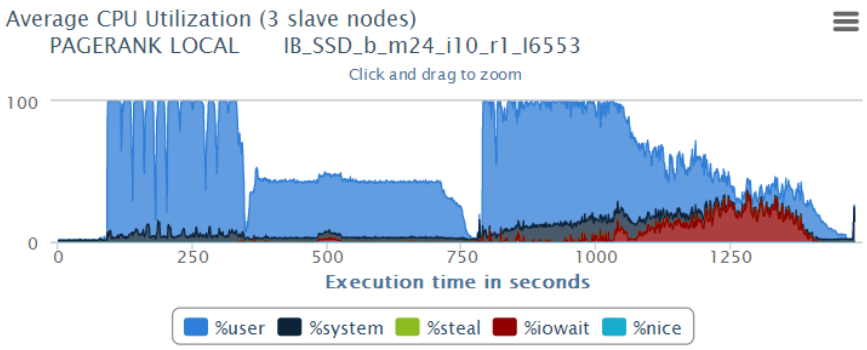
Available at: http://hadoop.bsc.es



Benchmarks Execution comparisons

- You can compare, side by side, all execution parameters:
  - CPU, Memory, Network, Disk, Hadoop parameters….





Example: 24 maps in parallel, SSD vs HDD vs ETH vs IB

Terasort

ETH+HDD

IB+HDD

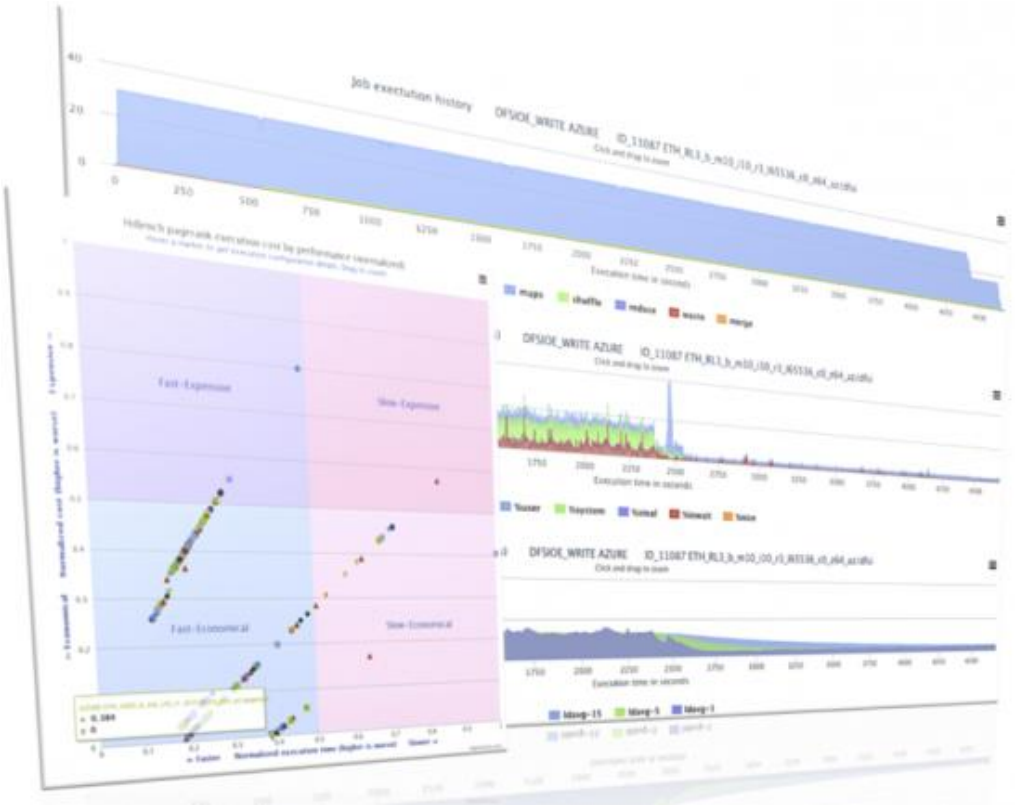
ETH+SSD

IB+SSD

ALOJA-WEB

- Entry point for explore the results collected from the executions,
  - Provides insights on the obtained results through continuously evolving data views.

- Online **DEMO** at: http://hadoop.bsc.es

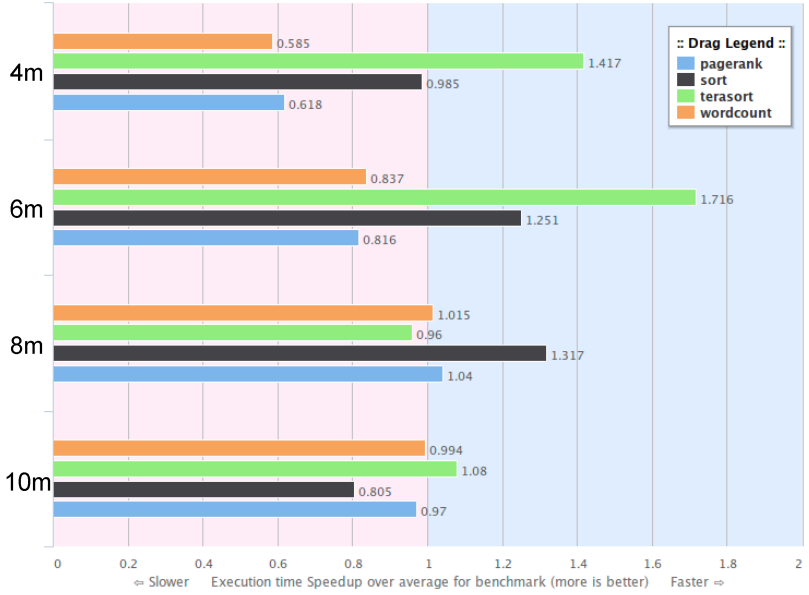


Barcelona Supercomputing Center
Centro Nacional de Supercomputación

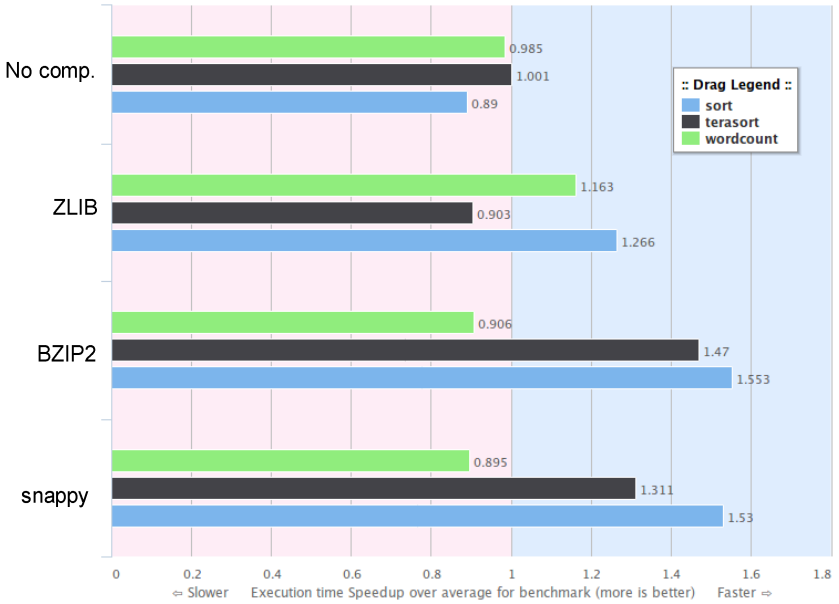
EARLY FINDINGS IN SW AND HW CONFIGURATIONS

Impact of SW configurations in Speedup

Number of mappers



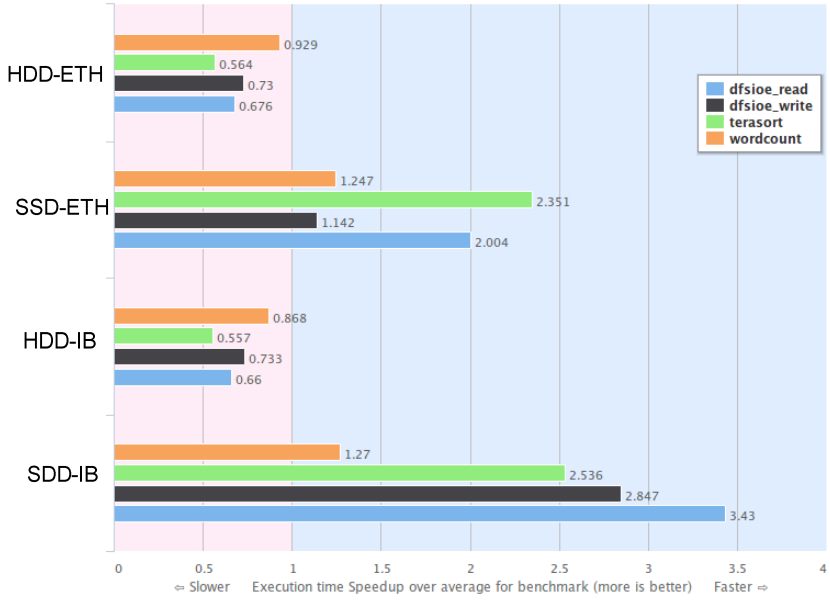
Compression algorithm



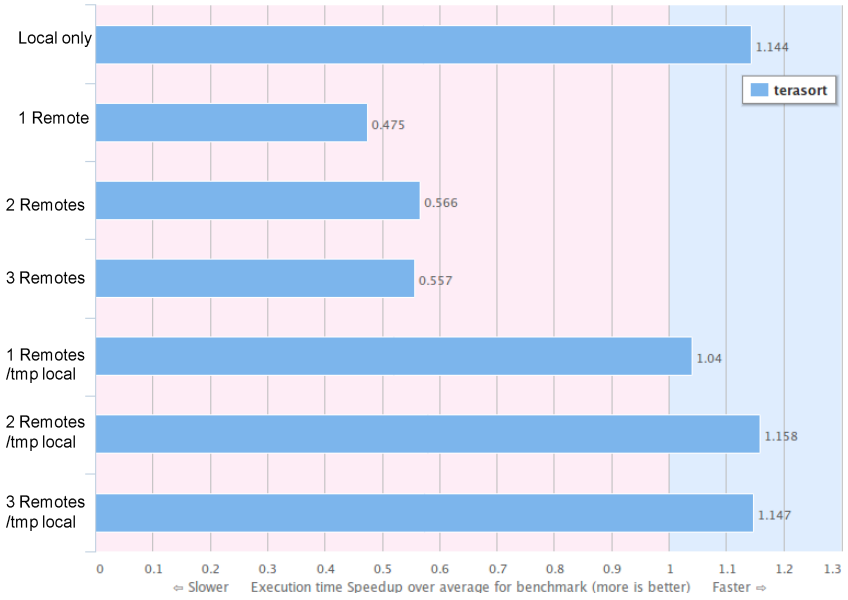
Speedup (higher is better)

Impact of HW configurations in Speedup

Disks and Network

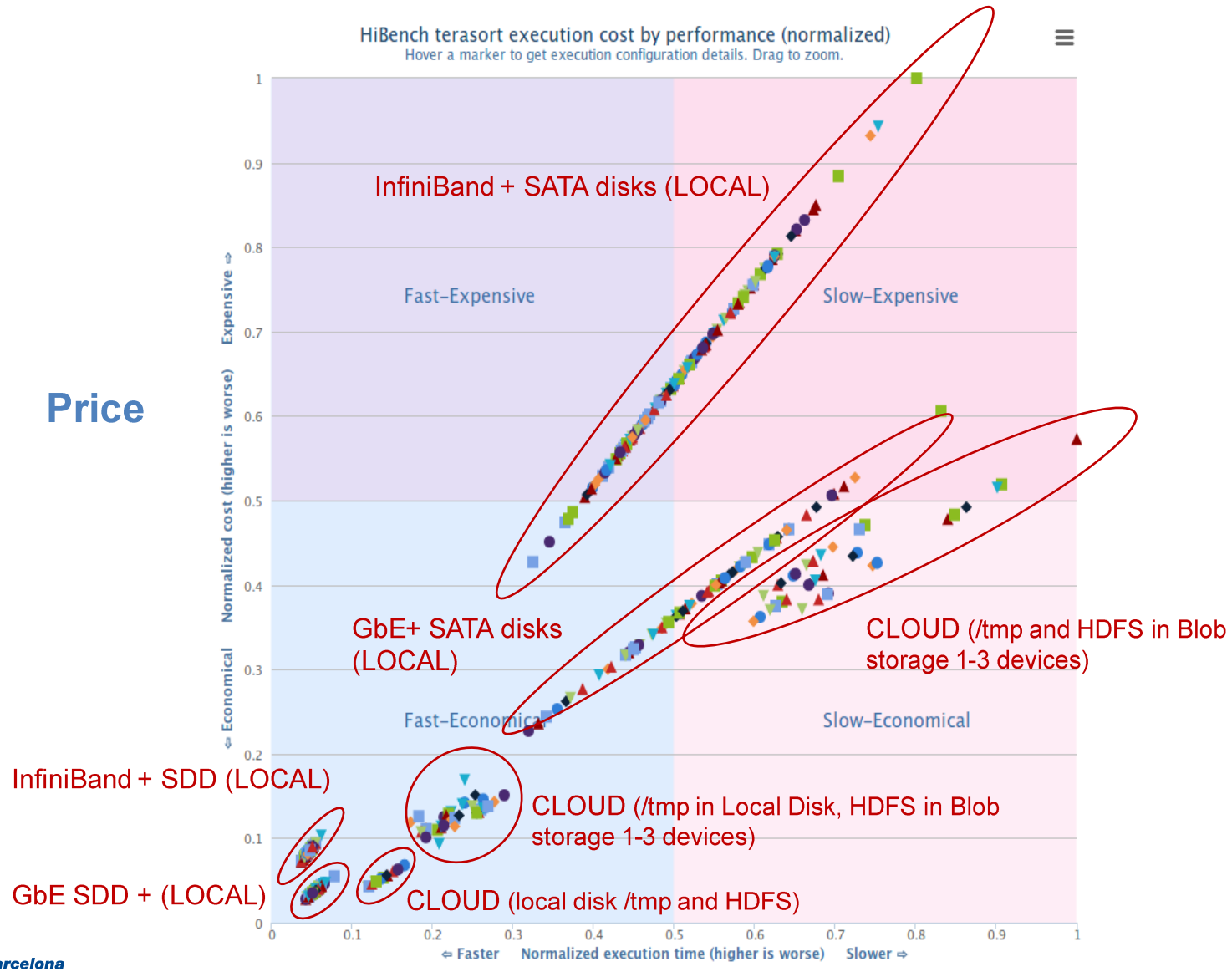


Cloud remote volumes



Speedup (higher is better)

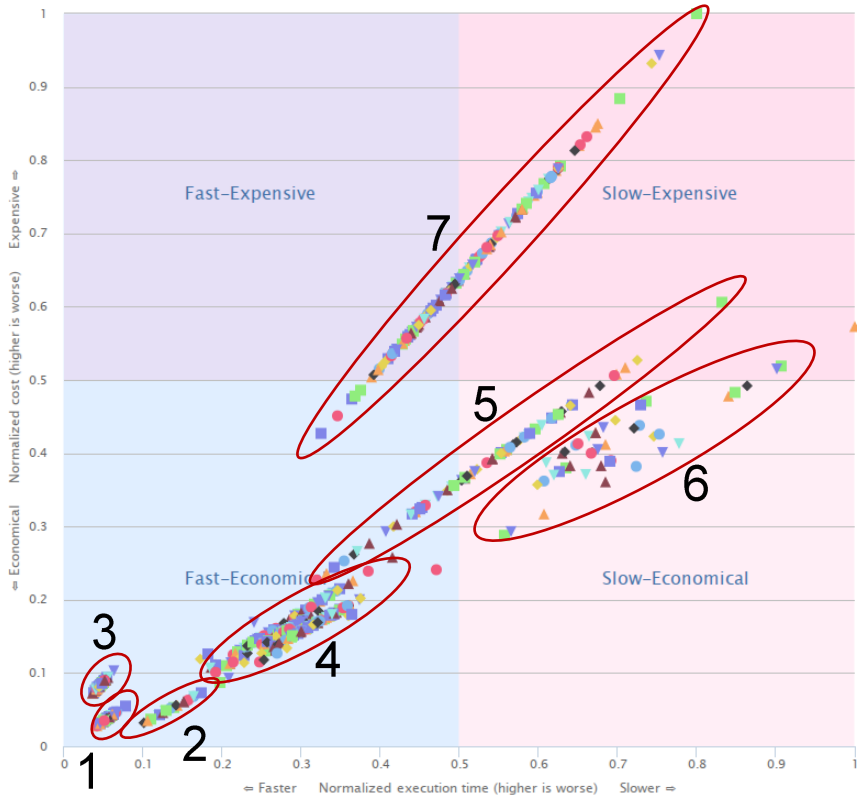
Cost-effectiveness of SW and HW (On-premise vs. Cloud)



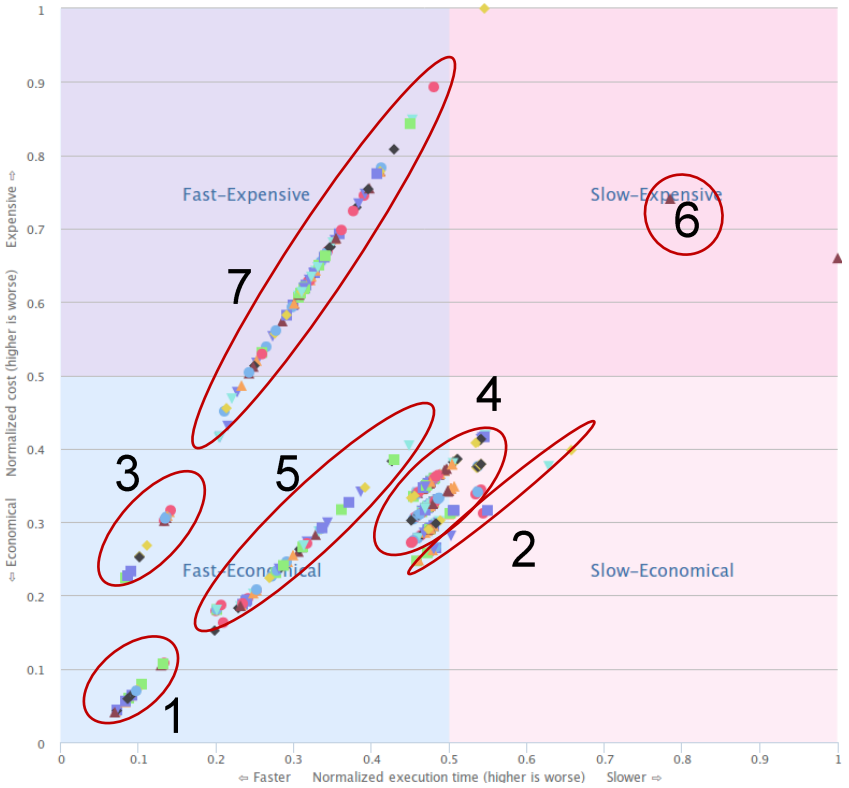
Cost-effectiveness of SW and HW

Point (0,0) represents most cost-effective execution

Terasort



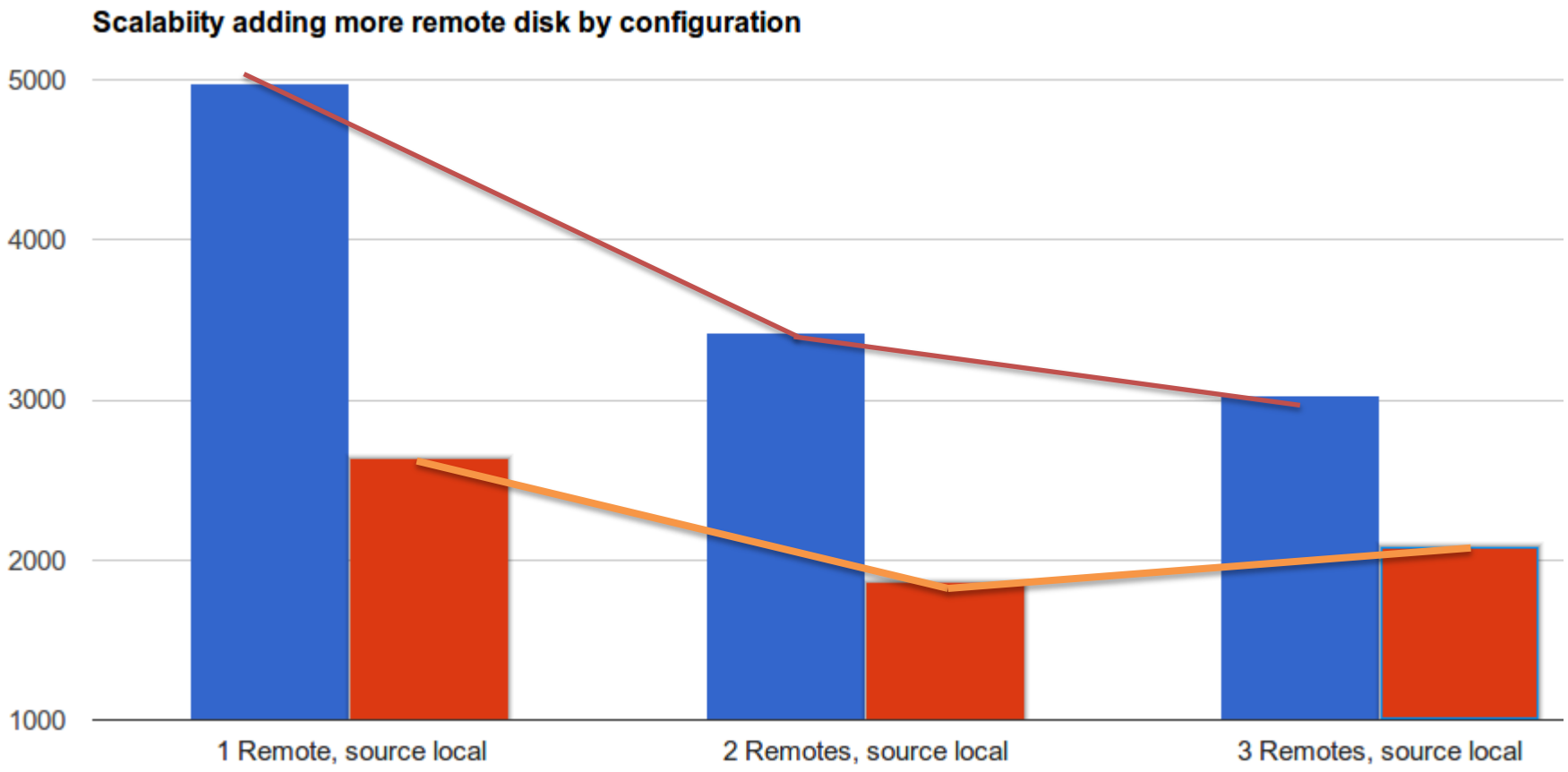
Wordcont



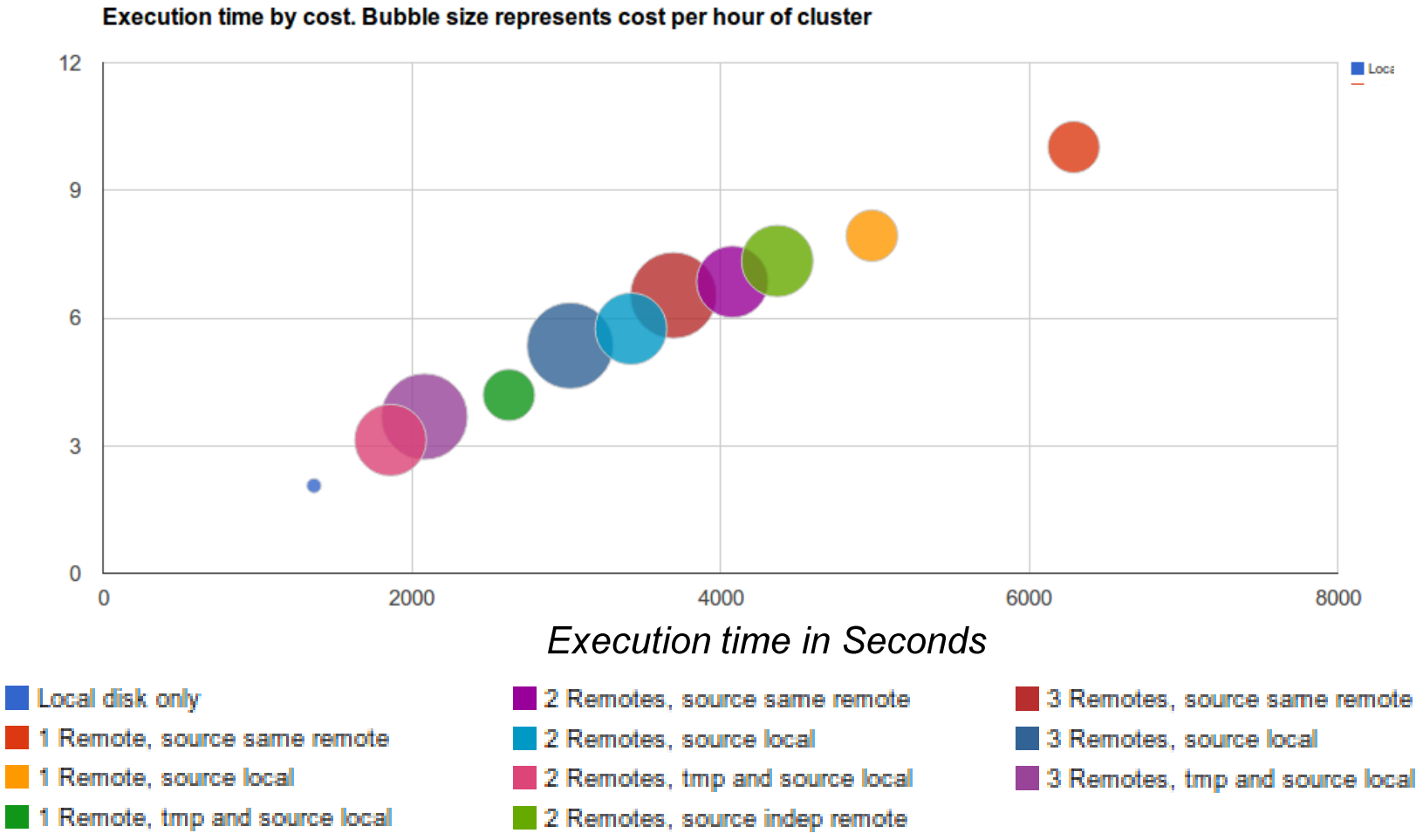
1) On-premise cluster: SSD disks + GbE.
2) Azure IaaS: Only local disk, virtualized SSD and GbE (baseline).
3) On-premise cluster: SSD disks + InfiniBand.

4) Azure IaaS: 1-3 remote vol. and Hadoop /tmp to local disk (SSD) and GbE
5) On-premise cluster: 1 SATA disk + GbE.
6) Azure IaaS: 1-3 remote volumes (Blob storage).
7) On-premise cluster: 1 SATA disk + InfiniBand.

Cloud IaaS impact of utilizing local vs. remote storage



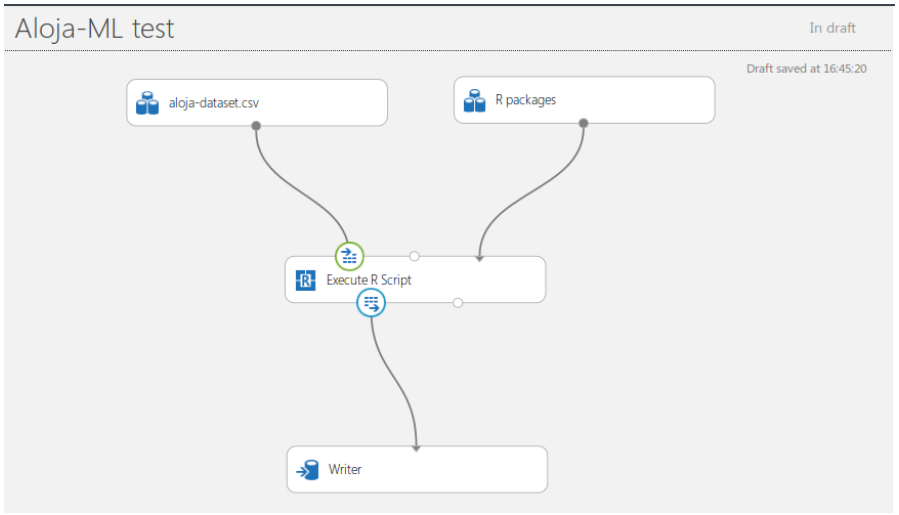
Cost efficiency of different Cloud deployment options

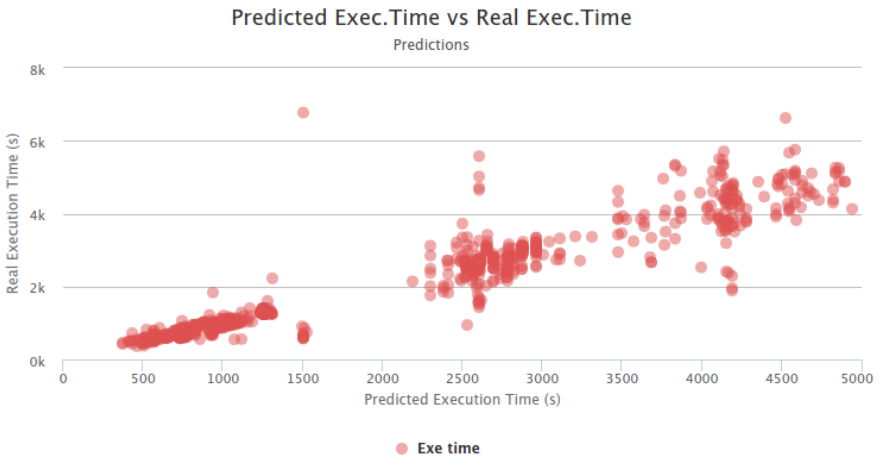


3.) ADVANCED WEB ANALYTICS WITH ALOJA-ML

3.) ALOJA-ML: Advanced Web Analytics

- Data views and evaluations
  - Best configuration recommendation
  - Configuration improvement
  - Parameter evaluation
- Cost / Performance analysis
- Aggregation and data filters
  - Aggregated performance metrics
  - Data filters
  - Abstracted Metrics
- Job characterization
  - Job resource consumption and bottlenecks
  - Job execution characteristics
- Machine Learning
  - Execution time prediction
    - By SW and HW configuration
  - Estimation of missing values
  - Clustering
    - DBScan and K-means for different views



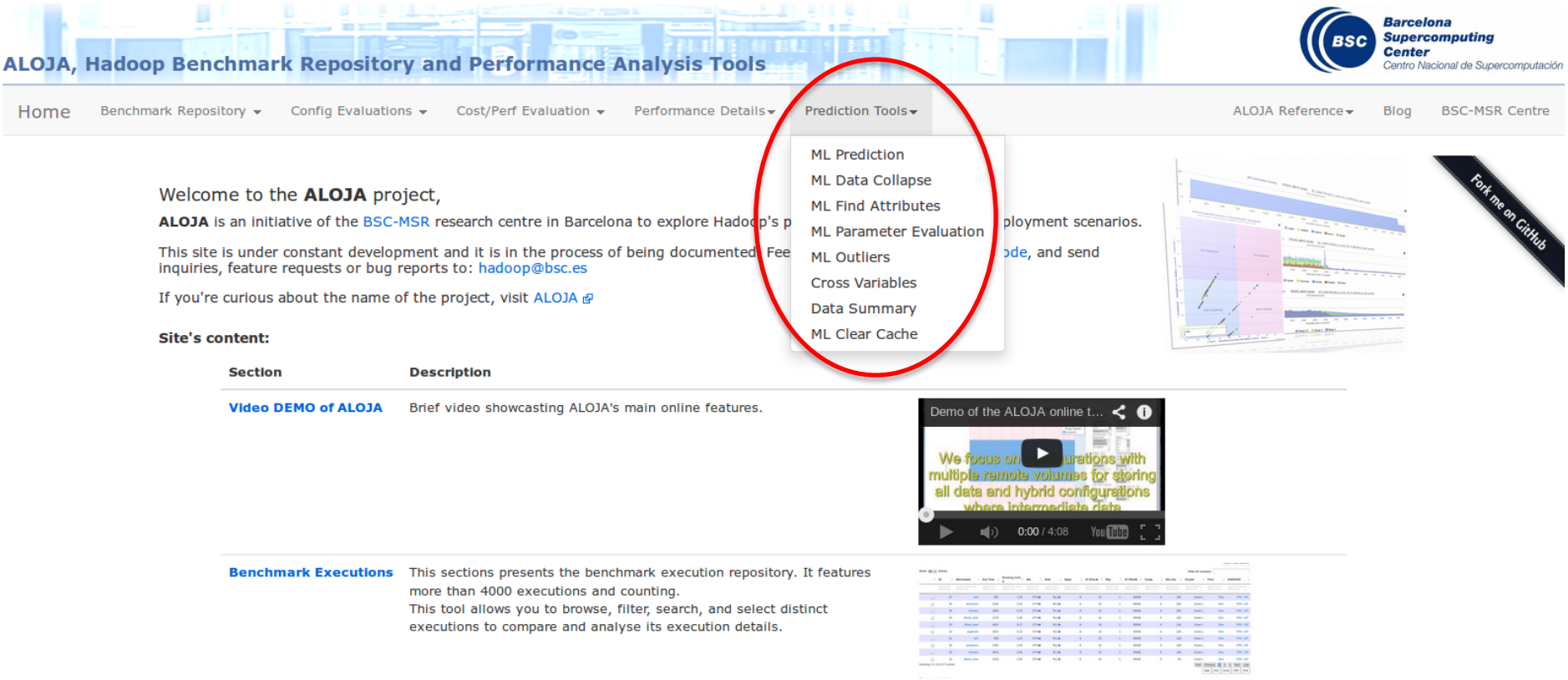


New feature: DBSCAN

- **DBSCAN** is a data clustering algorithm
  - It finds a number of clusters starting from the estimated density distribution of corresponding nodes.
- General overview of all executions of a benchmark
  - Select two metrics, Optionally filter desired parameters
- **Automatic clustering of similar executions**

Upcoming component : ALOJA-ML

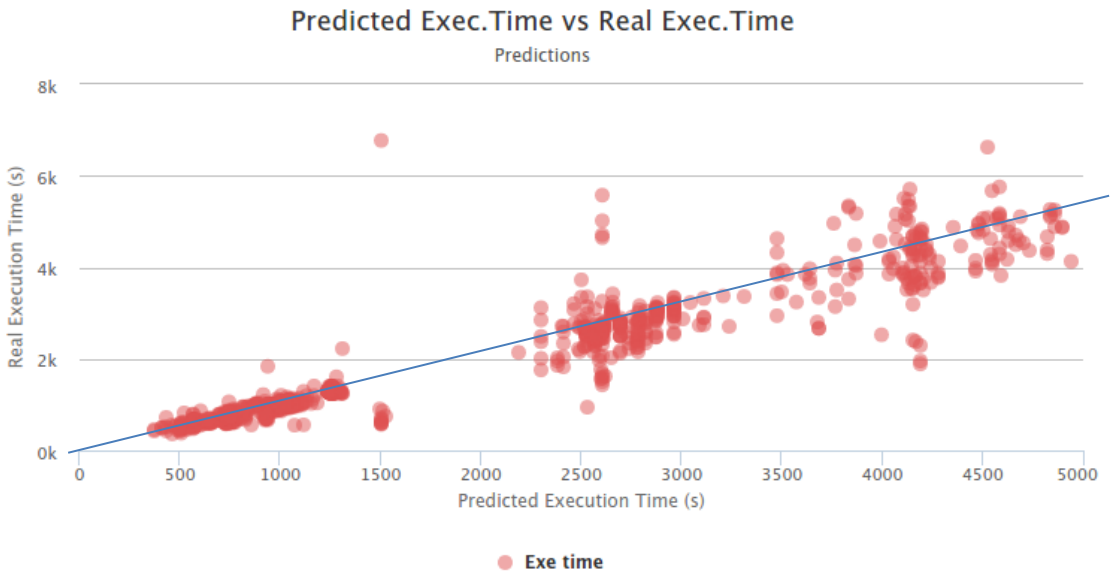
- ALOJA-ML
  - Provides automatic means to characterize job executions and clusters



The ALOJA-ML tool-set

1. Modeling and Prediction
   - From ALOJA dataset → Find a model for ‹Workld,Conf ~ Exe.Time›
2. Configuration recommendation
   - Rank (un)seen confs. for a benchmark from their expected Exe.Time
3. Anomaly detection (outliers)
   - Statistic + Model-based detection of anomalous executions
4. Behavior observation and statistic information
   - Aggregate variables around the ones we want to observe
   - Show frequency, percentiles and other useful information from ALOJA datasets

Modeling and Prediction



- Real vs. Predicted execution times

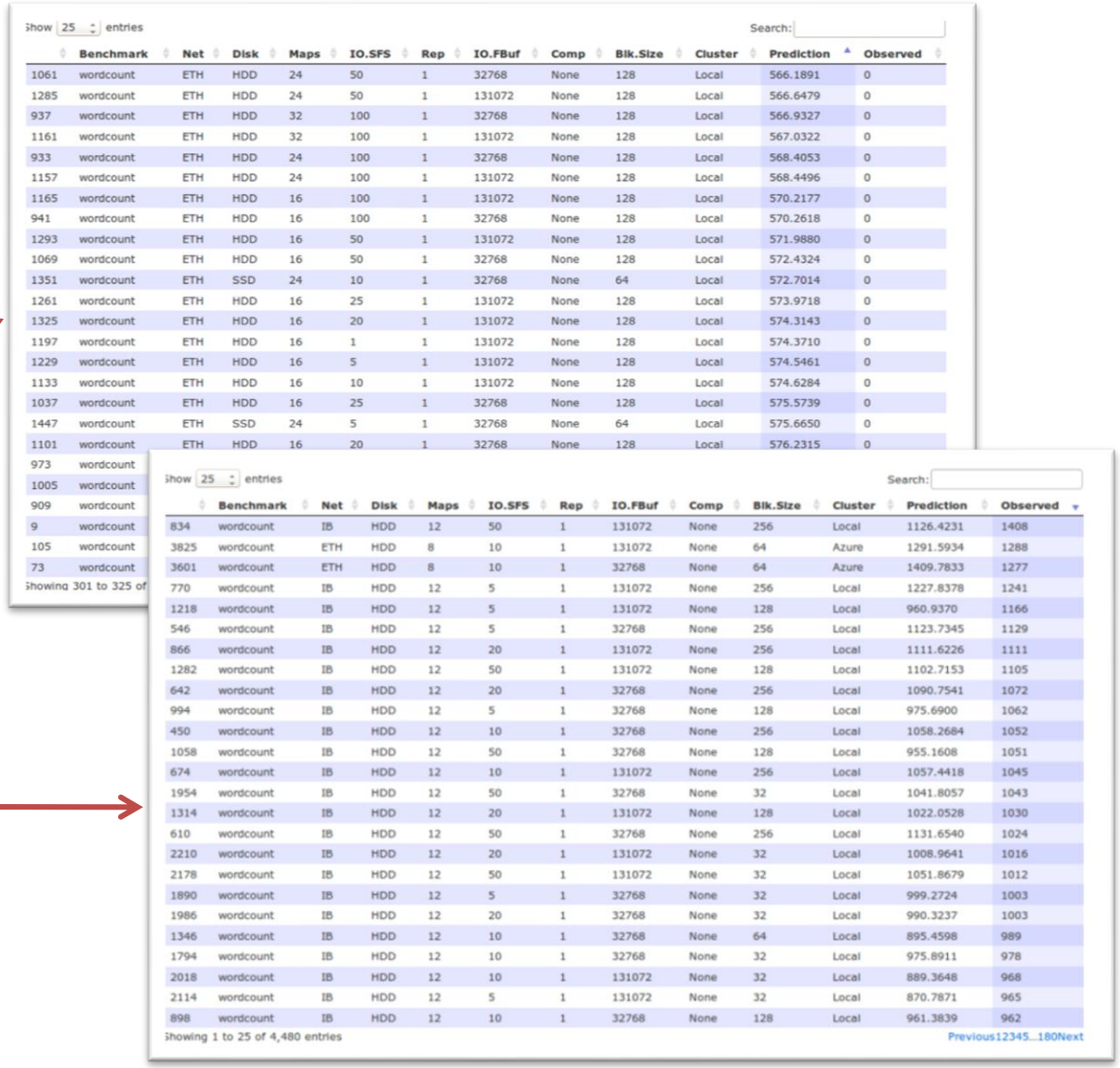
- Current available methods:
  - *Regression Trees*
  - *Nearest Neighbors*
  - *FFA Neural Networks*
  - *Multinomial Regression*

Prediction capabilities

- Different techniques used
  - *Regression Trees, Nearest Neighbors, FFANNs, Multinomial* Regs…
  - Mean Absolute Errors around **200s** [ranges from 100 to 6000]
  - Relative Absolute Errors of 0.12 to 0.15 (that's actually good!)
  - Without going deeper, we can learn from < **1000** different random observations [Current tested dataset: 4400 instances]
- A model can be used to:
  - Predict expected execution times for unobserved configurations
  - Determine if an observation is an **outlier** (anomaly)
  - Determine which configuration properties influences more a run

Rank and Recommend Configurations

- Predict a range of configurations, previously seen or unseen
- Order them by predicted execution time and rank configurations
- Compare also predicted execution times vs. observed execution times, if there are any.



Show 25 entries — Search:

	Benchmark	Net	Disk	Maps	IO.SFS	Rep	IO.FBuf	Comp	Blk.Size	Cluster	Prediction	Observed
1061	wordcount	ETH	HDD	24	50	1	32768	None	128	Local	566.1891	0
1285	wordcount	ETH	HDD	24	50	1	131072	None	128	Local	566.6479	0
937	wordcount	ETH	HDD	32	100	1	32768	None	128	Local	566.9327	0
1161	wordcount	ETH	HDD	32	100	1	131072	None	128	Local	567.0322	0
933	wordcount	ETH	HDD	24	100	1	32768	None	128	Local	568.4053	0
1157	wordcount	ETH	HDD	24	100	1	131072	None	128	Local	568.4496	0
1165	wordcount	ETH	HDD	16	100	1	131072	None	128	Local	570.2177	0
941	wordcount	ETH	HDD	16	100	1	32768	None	128	Local	570.2618	0
1293	wordcount	ETH	HDD	16	50	1	131072	None	128	Local	571.9880	0
1069	wordcount	ETH	HDD	16	50	1	32768	None	128	Local	572.4324	0
1351	wordcount	ETH	SSD	24	10	1	32768	None	64	Local	572.7014	0
1261	wordcount	ETH	HDD	16	25	1	131072	None	128	Local	573.9718	0
1325	wordcount	ETH	HDD	16	20	1	131072	None	128	Local	574.3143	0
1197	wordcount	ETH	HDD	16	1	1	131072	None	128	Local	574.3710	0
1229	wordcount	ETH	HDD	16	5	1	131072	None	128	Local	574.5461	0
1133	wordcount	ETH	HDD	16	10	1	131072	None	128	Local	574.6284	0
1037	wordcount	ETH	HDD	16	25	1	32768	None	128	Local	575.5739	0
1447	wordcount	ETH	SSD	24	5	1	32768	None	64	Local	575.6650	0
1101	wordcount	ETH	HDD	16	20	1	32768	None	128	Local	576.2315	0
973	wordcount											
1005	wordcount											
909	wordcount											
9	wordcount											
105	wordcount											
73	wordcount											

Showing 301 to 325 of

Show 25 entries — Search:

	Benchmark	Net	Disk	Maps	IO.SFS	Rep	IO.FBuf	Comp	Blk.Size	Cluster	Prediction	Observed
834	wordcount	IB	HDD	12	50	1	131072	None	256	Local	1126.4231	1408
3825	wordcount	ETH	HDD	8	10	1	131072	None	64	Azure	1291.5934	1288
3601	wordcount	ETH	HDD	8	10	1	32768	None	64	Azure	1409.7833	1277
770	wordcount	IB	HDD	12	5	1	131072	None	256	Local	1227.8378	1241
1218	wordcount	IB	HDD	12	5	1	131072	None	128	Local	960.9370	1166
546	wordcount	IB	HDD	12	5	1	32768	None	256	Local	1123.7345	1129
866	wordcount	IB	HDD	12	20	1	131072	None	256	Local	1111.6226	1111
1282	wordcount	IB	HDD	12	50	1	131072	None	128	Local	1102.7153	1105
642	wordcount	IB	HDD	12	20	1	32768	None	256	Local	1090.7541	1072
994	wordcount	IB	HDD	12	5	1	32768	None	128	Local	975.6900	1062
450	wordcount	IB	HDD	12	10	1	32768	None	256	Local	1058.2684	1052
1058	wordcount	IB	HDD	12	50	1	32768	None	128	Local	955.1608	1051
674	wordcount	IB	HDD	12	10	1	131072	None	256	Local	1057.4418	1045
1954	wordcount	IB	HDD	12	50	1	32768	None	32	Local	1041.8057	1043
1314	wordcount	IB	HDD	12	20	1	131072	None	128	Local	1022.0528	1030
610	wordcount	IB	HDD	12	50	1	32768	None	256	Local	1131.6540	1024
2210	wordcount	IB	HDD	12	20	1	131072	None	32	Local	1008.9641	1016
2178	wordcount	IB	HDD	12	50	1	131072	None	32	Local	1051.8679	1012
1890	wordcount	IB	HDD	12	5	1	32768	None	32	Local	999.2724	1003
1986	wordcount	IB	HDD	12	20	1	32768	None	32	Local	990.3237	1003
1346	wordcount	IB	HDD	12	10	1	32768	None	64	Local	895.4598	989
1794	wordcount	IB	HDD	12	10	1	32768	None	32	Local	975.8911	978
2018	wordcount	IB	HDD	12	10	1	131072	None	32	Local	889.3648	968
2114	wordcount	IB	HDD	12	5	1	131072	None	32	Local	870.7871	965
898	wordcount	IB	HDD	12	10	1	32768	None	128	Local	961.3839	962

Showing 1 to 25 of 4,480 entries — Previous 1 2 3 4 5 ... 180 Next

Anomaly Detection

Anomaly and Outlier Detection

- Use of statistic and model-based outlier detections
- Highlight executions with high probability of anomaly
- Mark down executions with high probability of being errors

Data aggregation and Statistics

- Tools for data aggregation (also predicting their aggregates)
  - Find relevance or discard parameters

Exe.Time: Observed / Estimated

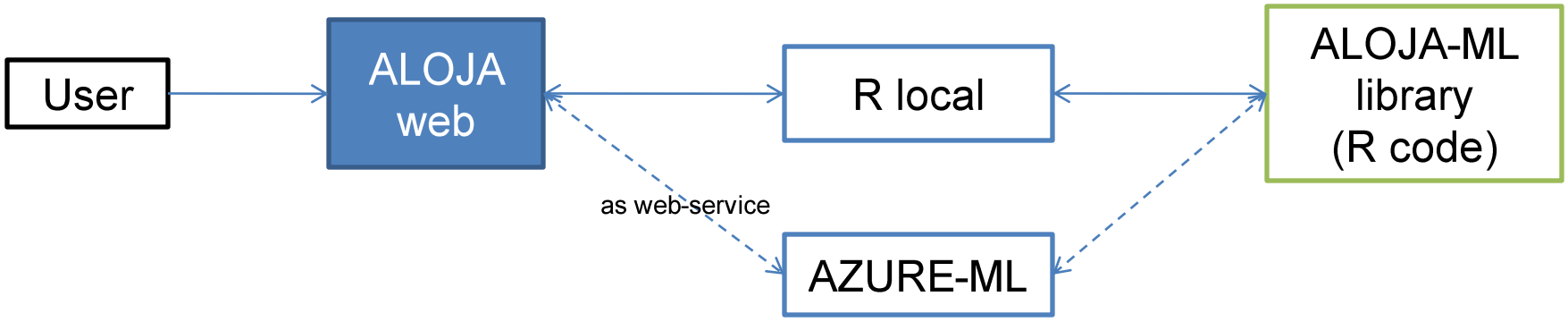
Show 10 entries Search:

	sort	terasort	wordcount
HDD:Cmp0:1:131072	747	1622	1288
HDD:Cmp0:1:32768	2810	3300	992
SSD:Cmp0:1:131072	457	1022	718
SSD:Cmp0:1:32768	454	967	674

Showing 1 to 4 of 4 entries Previous1Next

ALOJA-ML engine as a Cloud Service

- ALOJA-ML works with **R**
  - AZURE-ML has incorporated recently R to its workbench
  - We can run the ML engine locally, also use AZURE-ML as an option

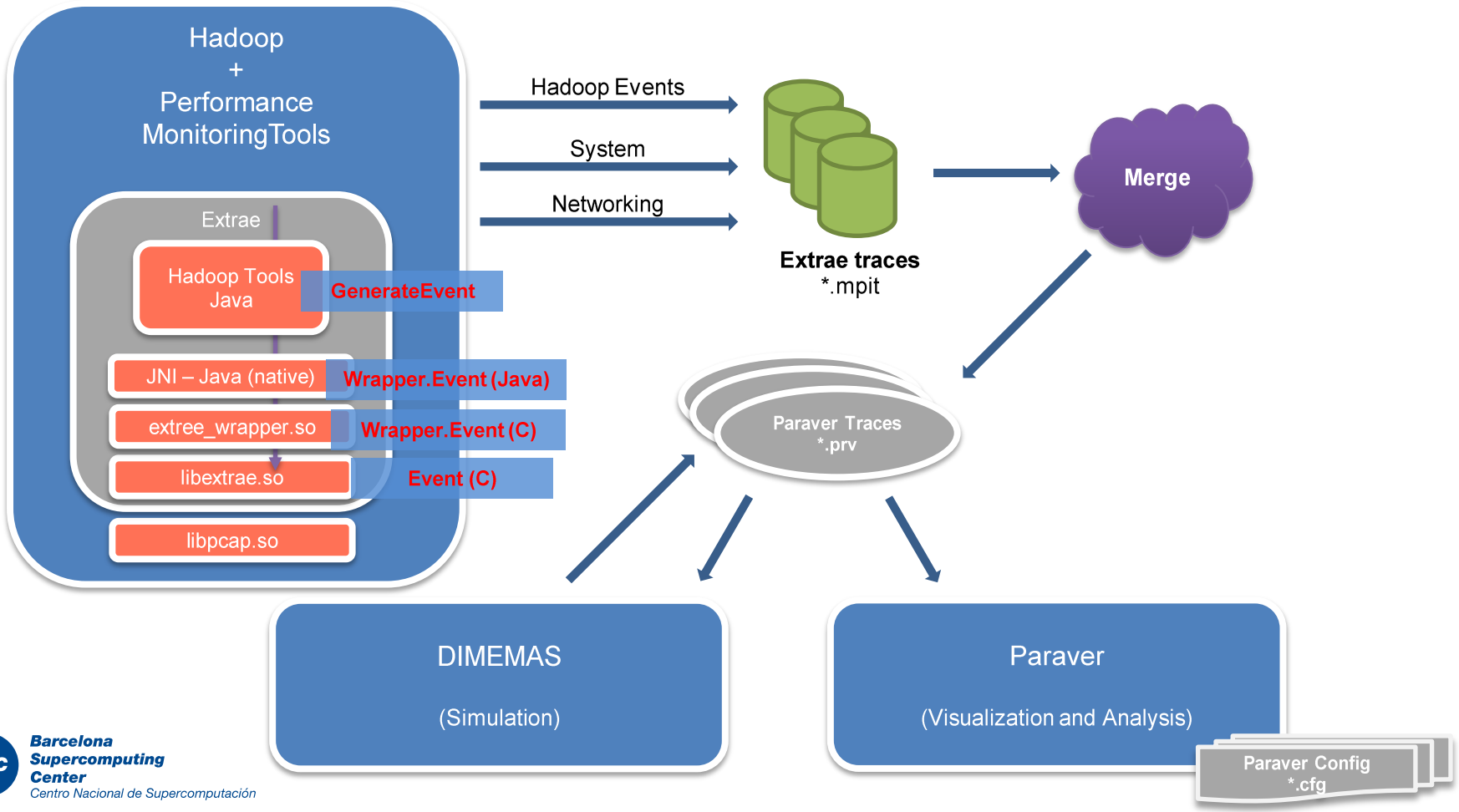


- Most of the R code of ALOJA-ML library can run on AZML+R
- The ML process can be delegated to a AZML web-service

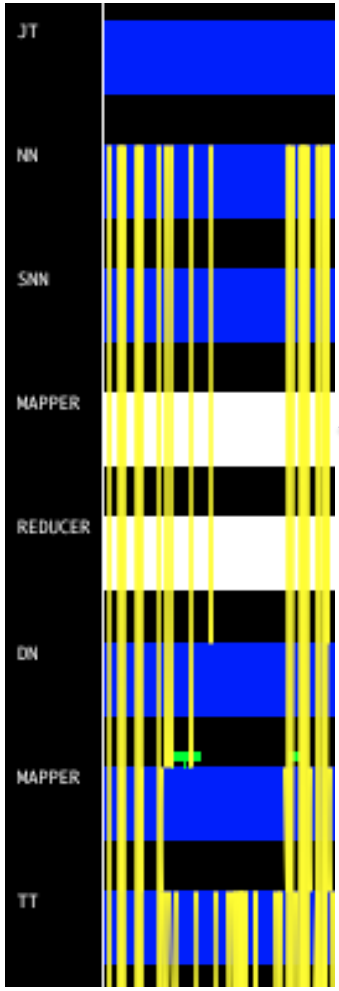
DEEP INSTRUMENTATION WITH BSC HPC TOOLS

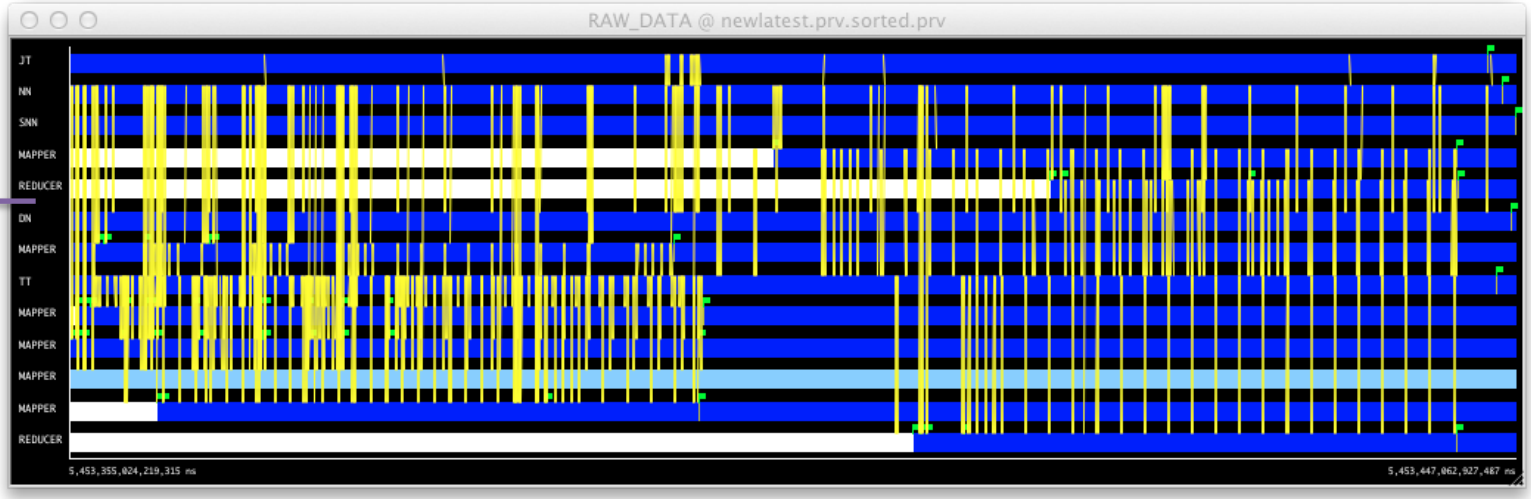
Overview

Hadoop Analysis Toolkit and BSC tools

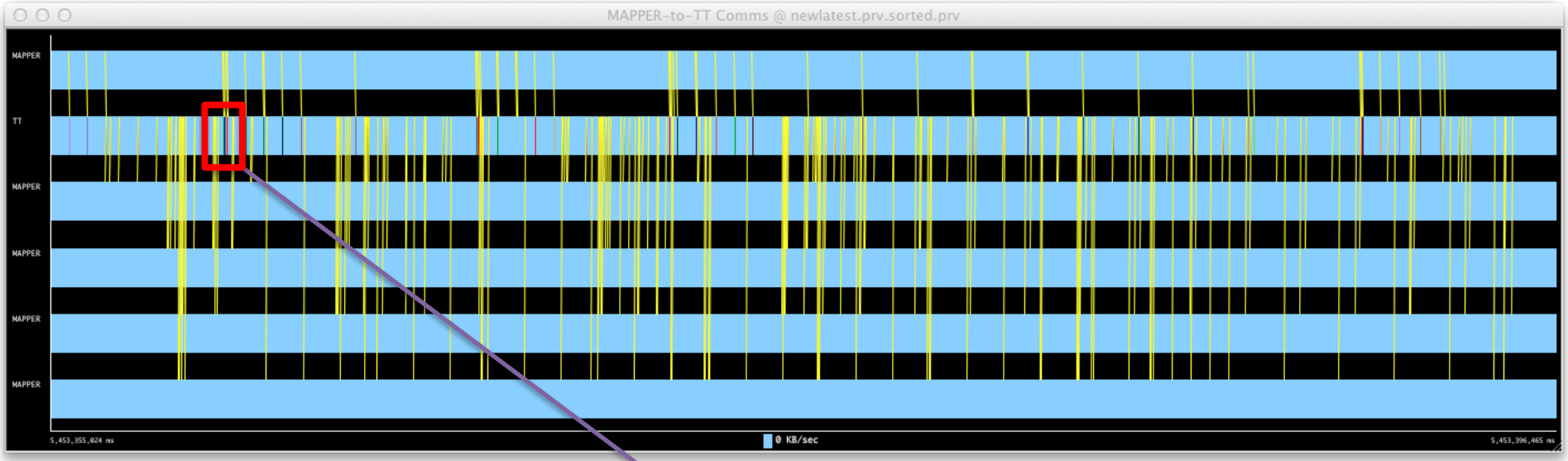


Deep instrumentation: threads, states, events, communication



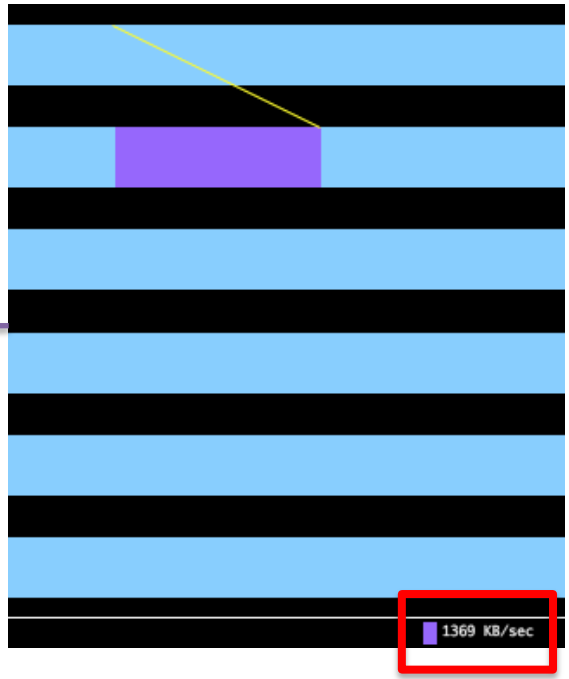


Example: Packet level communications

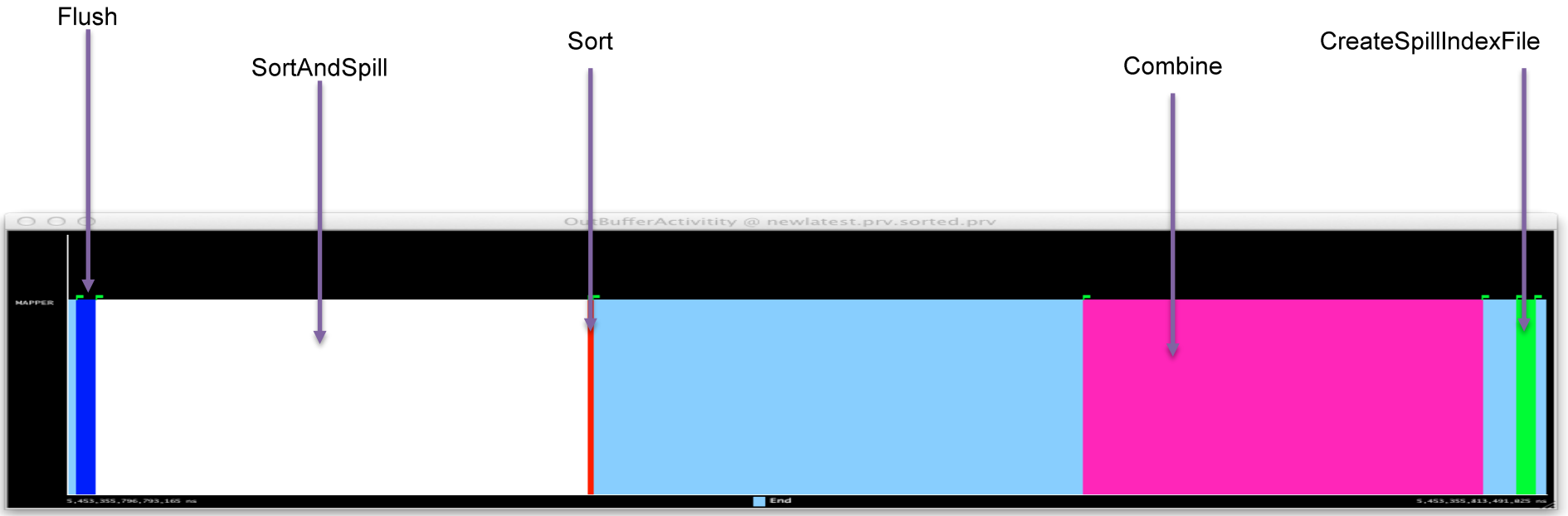


Data Sent to TT by Mappers @ newlatest.prv.sorted.prv

	MAPPER	MAPPER	MAPPER	MAPPER	MAPPER	MAPPER
JT	-	-	-	-	-	-
NN	-	-	-	-	-	-
SNN	-	-	-	-	-	-
MAPPER	-	-	-	-	-	-
REDUCER	-	-	-	-	-	-
DN	-	-	-	-	-	-
MAPPER	-	-	-	-	-	-
TT	792,866.50	453,508.20	450,403.59	528,267.08	457,878.90	484,062.35
MAPPER	-	-	-	-	-	-
MAPPER	-	-	-	-	-	-
MAPPER	-	-	-	-	-	-
MAPPER	-	-	-	-	-	-
REDUCER	-	-	-	-	-	-
Total	792,866.50	453,508.20	450,403.59	528,267.08	457,878.90	484,062.35

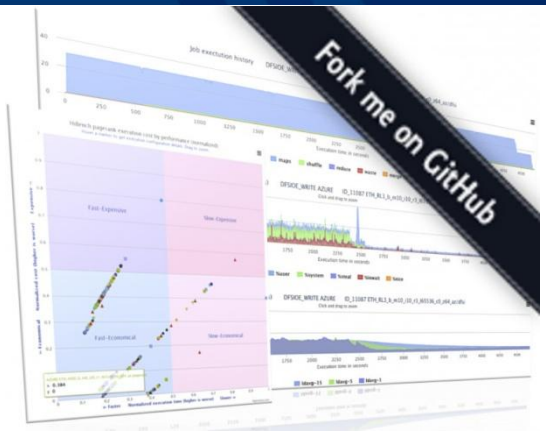


Example: Low-level Hadoop events



PROJECT DISSEMINATION, REFERENCE, AND CONCLUDING REMARKS

Extending and collaborating in ALOJA



1. Install prerequisites
   - vagrant

2. `git clone` https://github.com/Aloja/aloja.git

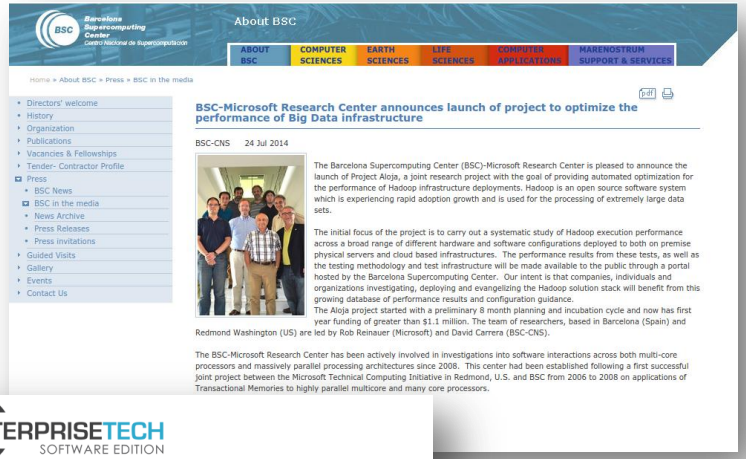
3. `cd aloja/vagrant`

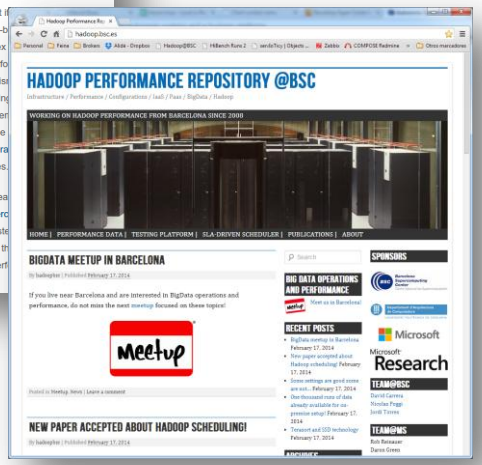
4. `vagrant up`

5. Open your browser at: http://localhost:8080

Community engagement and PR

- Recurrent
  - Blogging on Big Data
  - Presentations at local IT Big Data meetup events
- Upcoming
  - Cloudscape DEMO
  - Strata EU, Hadoop Summit (TBC)
- Research Center
  - Barcelona Big Data Center of Excellence
  - … seeking EU funding
- In talks/tests with Cloud providers to expand search
  - Rackspace, Amazon, Google…





Additional reference and publications

- Online repository and tools available at:
  - http://hadoop.bsc.es

- Publications: http://hadoop.bsc.es/publications

  - Project description on:
    - "ALOJA: a Systematic Study of Hadoop Deployment Variables to Enable Automated Characterization of Cost-Effectiveness“
  - Upcoming:
    - ALOJA-ML for KDD15’
      1. ALOJA-ML: A first dive into Hadoop behavior using Machine Learning
    - Working on:
      2. The Economics of Hadoop in the Cloud
         - » An evaluation of cost-effectiveness of Hadoop in the Cloud
      3. Cluster performance Characterization for Big Data
         - » A performance modeling comparison of Hadoop in different cluster sizes and OS configurations

Concluding remarks

- The early findings of the project already show significant value in understanding Hadoop's runtime
  - for optimizing executions times
  - understanding the cost-effectiveness of different configuration and deployment options

- Our intent is that researchers and organizations evaluating or deploying the Hadoop stack will benefit
  - from this growing database of performance results and configuration guidance

BSCMSCR Team

BSC Microsoft Research Centre

- **ALOJA Team members:**

David Carrera	Senior Researcher, Barcelona Super Computing Center (BSC) Associate Professor, Universitat Politecnica de Catalunya (UPC)
Nicolas Poggi	Post Doctorate Researcher, BSC
Aaron Call	Research support engineer, BSC
Josep Lluis Berral	Post Doctorate Researcher, BSC/UPC
Josep Cugat	Research support engineer, BSC
Fabrizio Gagliardi	Senior Strategy Consultant, BSC Distinguished Research Director, UPC Chairman, ACM Europe Council
Jordi Torres	Research Manager, BSC Professor, UPC
Rob Reinauer	Partner Systems Architect, Microsoft SQL Server
Jose Blakeley	Partner Software Architect, Microsoft SQL Server
Nikola Vujic	Software Development Engineer, Microsoft HDInsight
Daron Green	Sr. Director, Regional Research, Microsoft Research
J. Eduardo Campos	Director, Business Strategy, Microsoft Emerging Markets

www.bsc.es

Barcelona Supercomputing Center

Centro Nacional de Supercomputación

Thanks!

Q&A