www.bsc.es

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

ALOJA-HDI: A characterization of cost-effectiveness of PaaS Hadoop on the Azure Cloud

Aaron Call, Nicolas Poggi, David Carrera

Sixth Workshop on Big Data Benchmarking
Toronto, 2015

BSC~Microsoft Research Centre

ALOJA project

- The ALOJA project intends to characterize cost-effectiveness of Hadoop deployments
  - Currently moving towards predictive analytics

- With a testbed of more than 40k executions on the cloud and on our on-premise cluster

- In this paper we start the work of characterization of Hadoop PaaS

Advantages of Hadoop as PaaS

- Automatically installed during cluster provisioning
  - Ready to run
- Fine-tuned according to the underlying hardware
  - This is the most time-consuming part when setting up a Hadoop environment
- Maintenance usually provided
  - Either hardware or software failures
- As consequence, you save money
  - No need of having data engineers to install, tune Hadoop or fix issues, which is required on IaaS deployments

Hadoop PaaS solutions

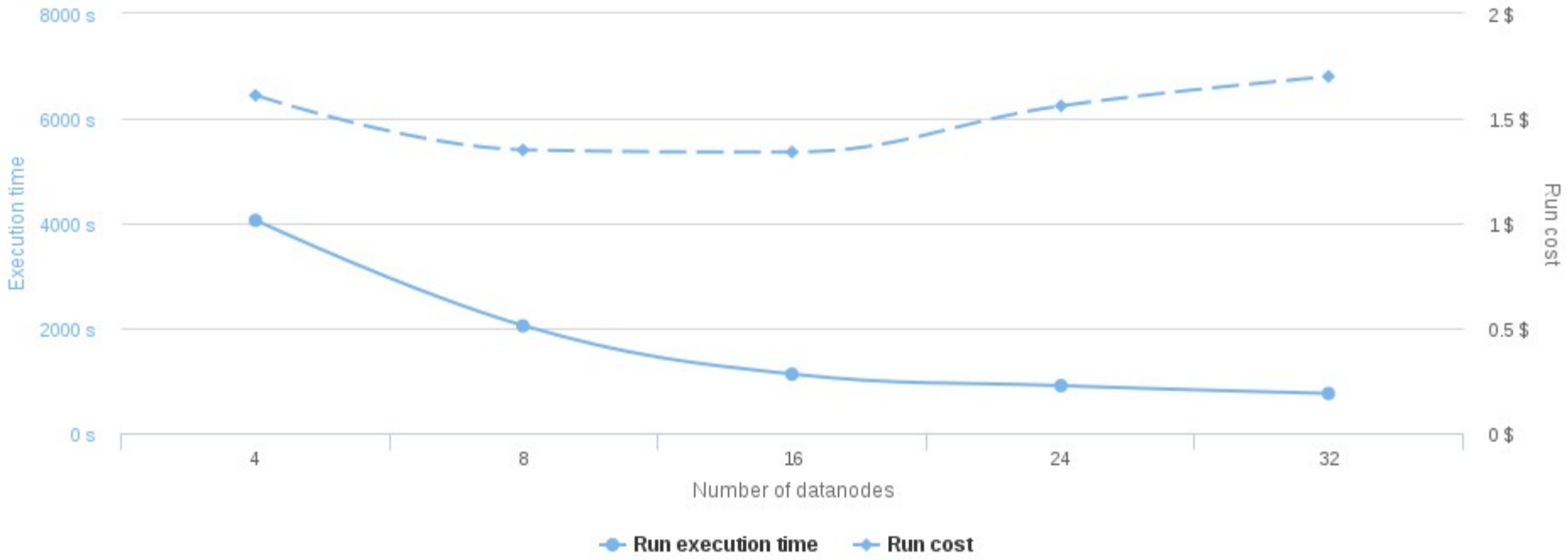
- Amazon Elastic MapReduce
- Rackspace Cloud BigData service
- Microsoft Azure HDInsight
  - Not limited to Hadoop, but also offers Storm, Hbase, Hive, and others.
  - In this work we evaluate its cost-effectiveness

Evaluation environment

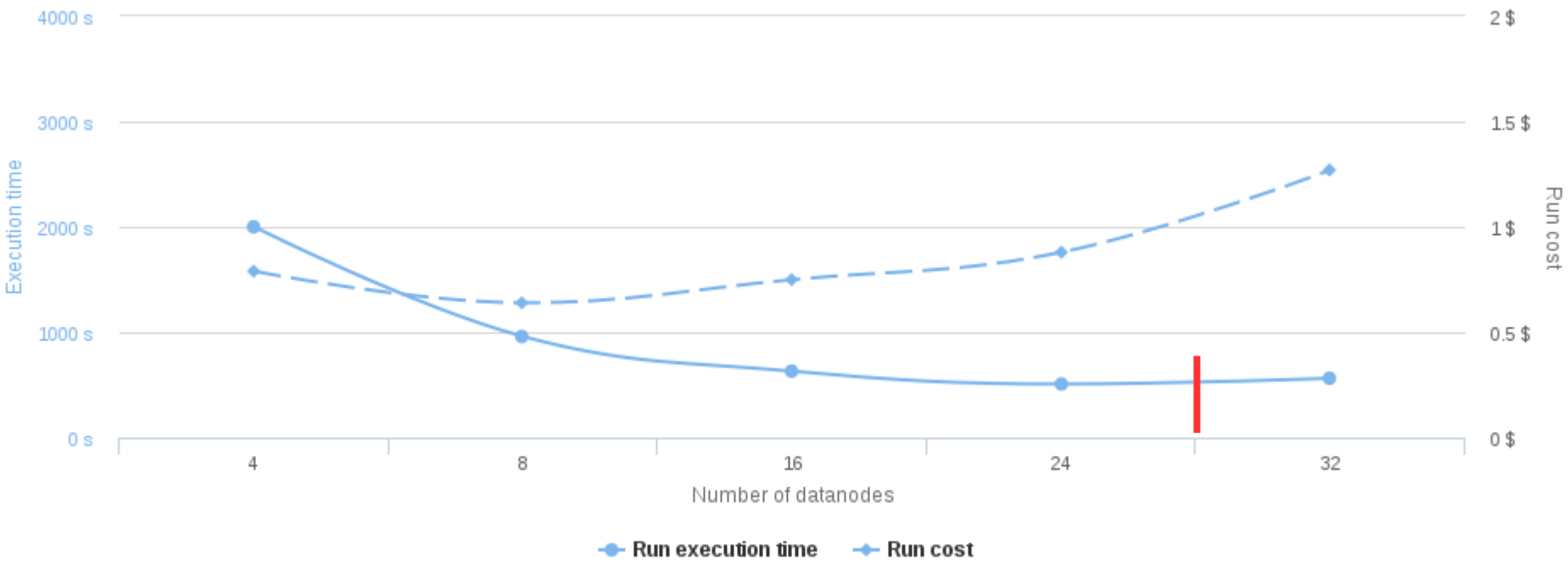
- On our initial results we used the default Windows-based clusters, with 4,8,16,24 and 32 datanodes, and later we added the recent Linux-based clusters

- On both OS we used A3 machines:
  - 4 cores, 7GB RAM, 0.2384$/hr
  - Using the default Microsoft implementation of HDFS on Azure Storage (data on remote volumes, local disks of nodes not being used)
- Tested workloads: wordcount (128GB generated data) and terasort (100GB data)

Scalability of wordcount

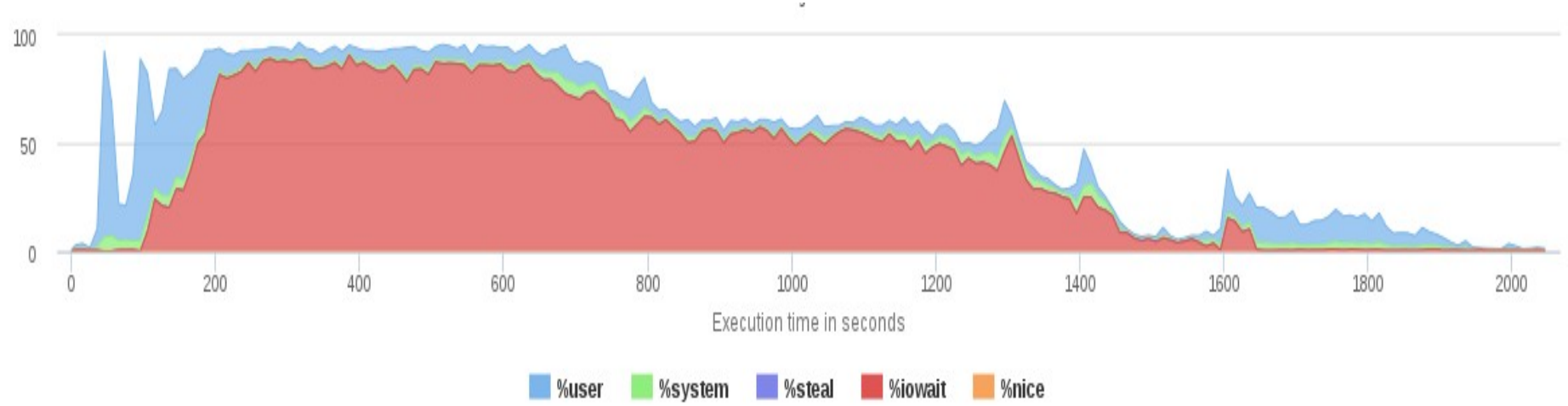


Scalability of terasort

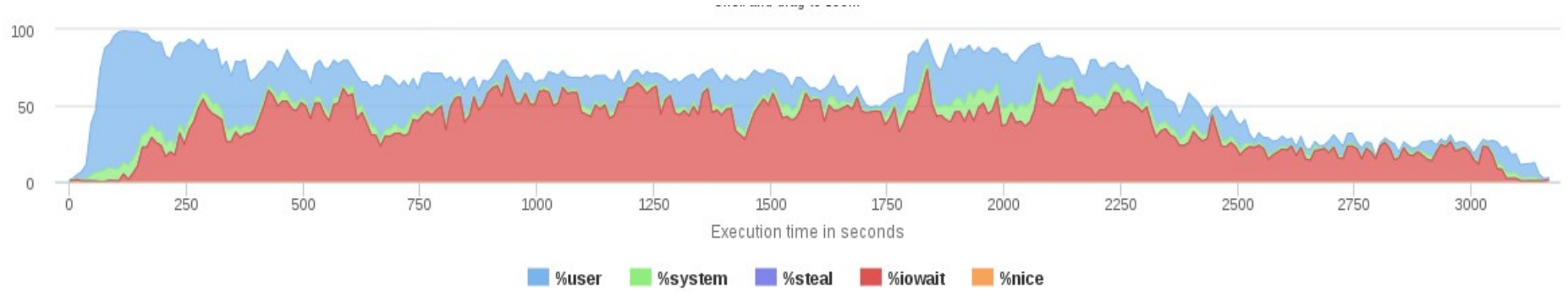


Bottleneck: IO

Terasort 32 nodes CPU usage

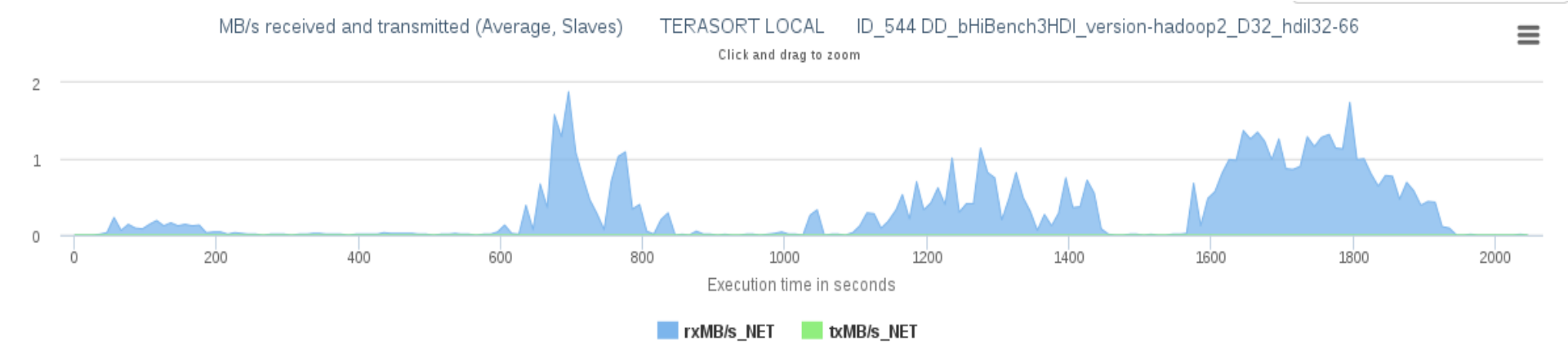


Terasort 8 nodes CPU usage

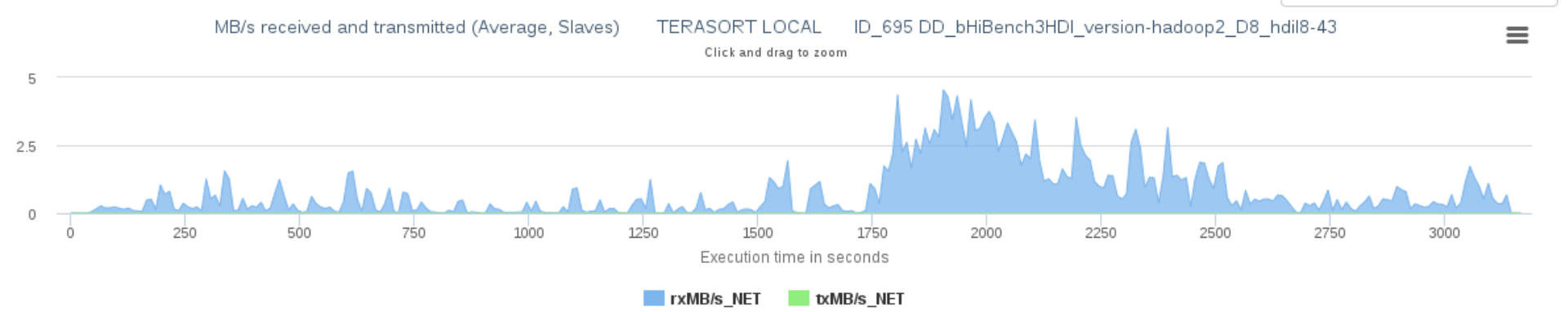


IO problem: throttling

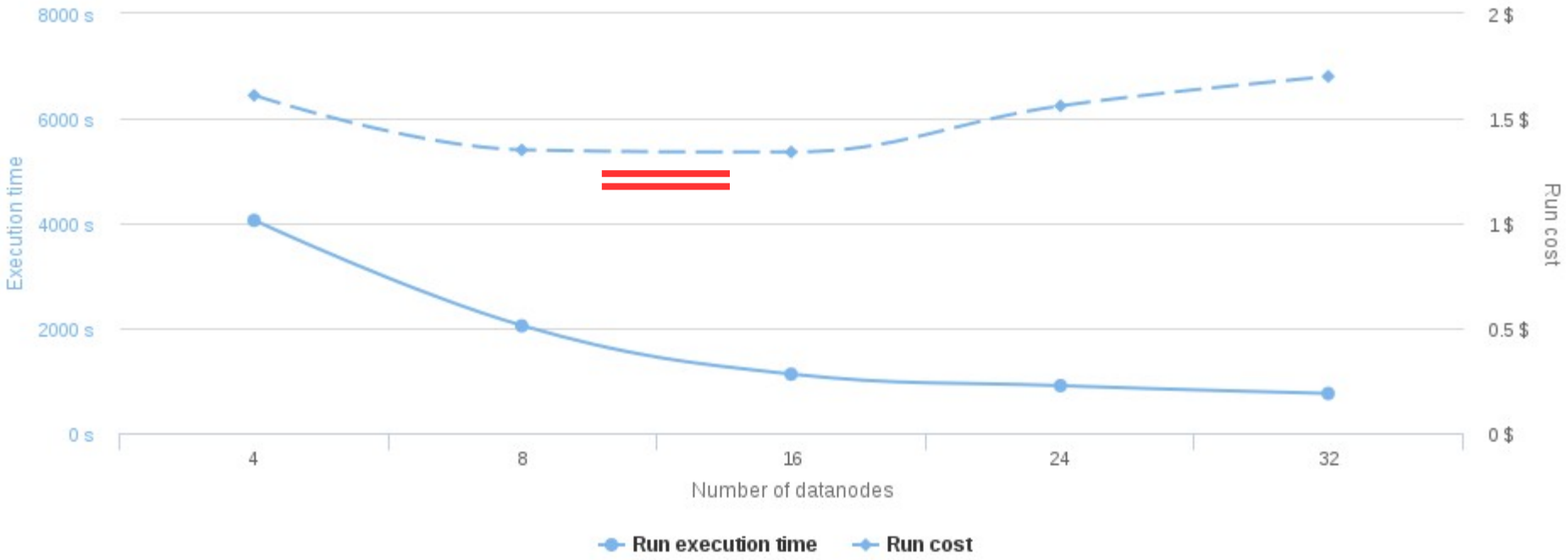
Terasort 32 node network usage



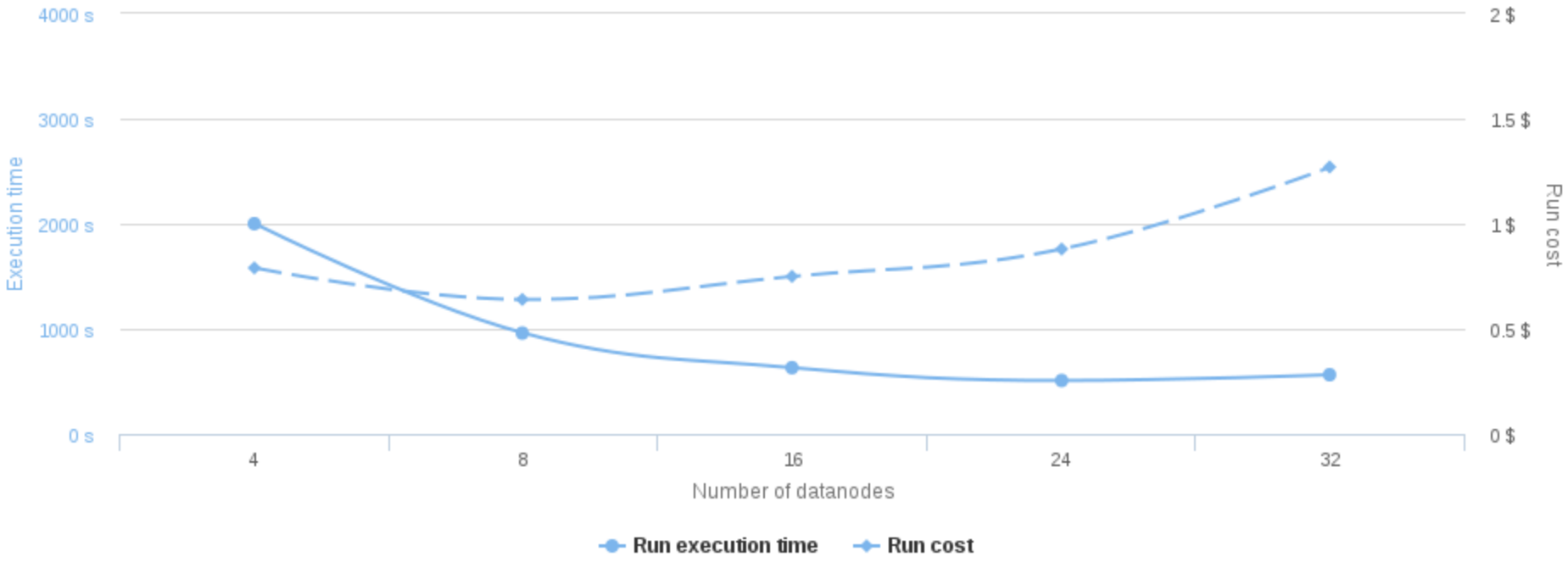
Terasort 8 nodes network usage



Cost performance of wordcount



Cost performance of terasort



PaaS Benefits

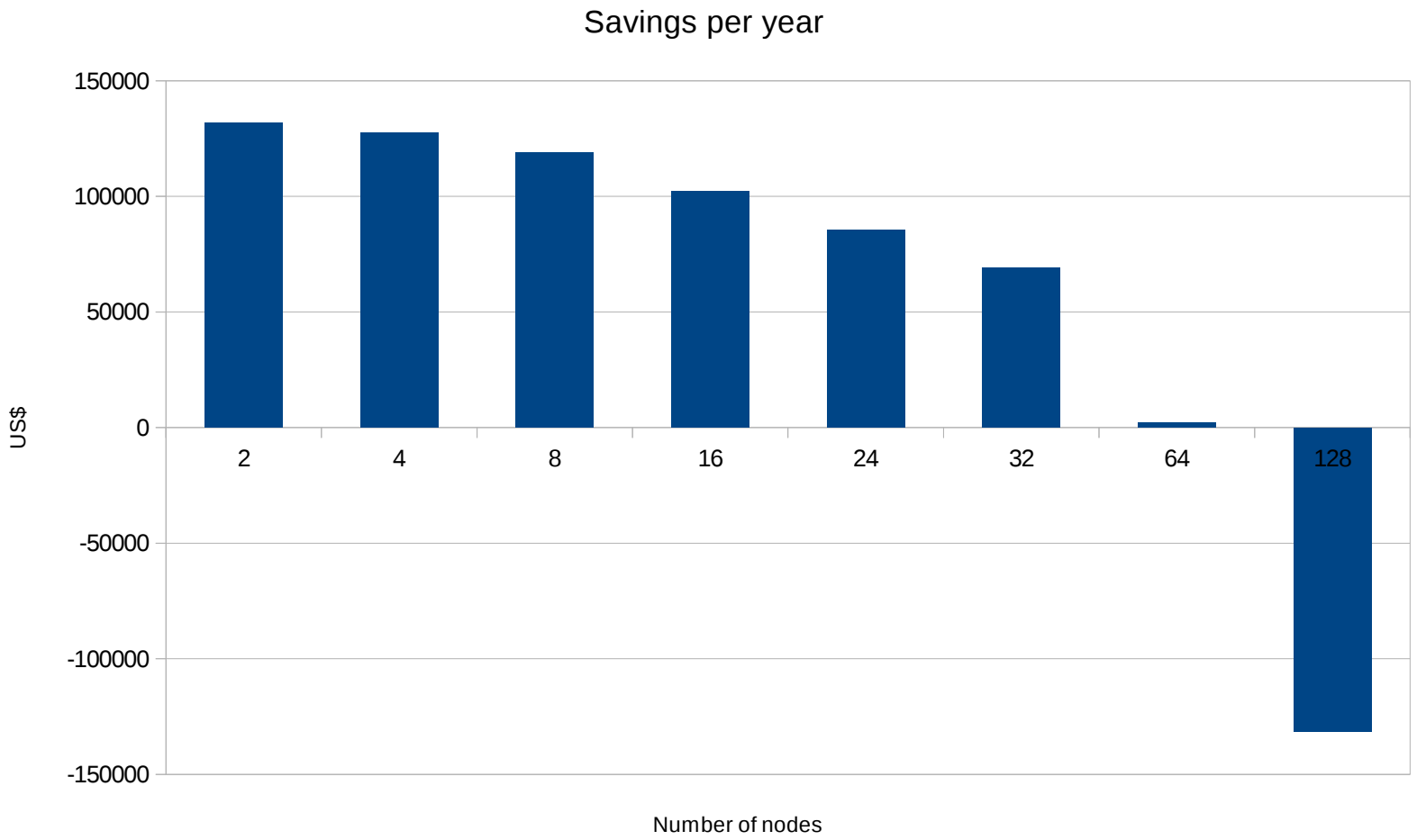
- On IaaS deployments a data engineer is always required to install, tune and fix issues on our clusters
- On PaaS, however, this is provided by the provider, saving us of a hiring
- O'Reily made a survey on data science salaries and estimated an average salary of 140.000 US$ for a data engineer
- Calculating the most cost-effective cluster cost (16 datanodes) and its cost during a year, we get:

  (16 datanodes + 2 headnodes) * 0.2384/hr = 4.2912 US$/hr

  4.2912*24*365 = 37,590.912 US$/year

- Therefore we can potentially save 102,409.088 US$ per year

PaaS benefits (2)



Conclusions

- To have cost-effective large clusters we need to improve significantly enough the execution time, otherwise it will be more expensive even though we get better times
- Azure storage experiments throttling when running with many datanodes, limiting HDInsight performance
- PaaS might save lots of money if you use it for a limited period of time, or if you only run small clusters

Future work

- Testing over HDI-Linux
- Use different VM SKUs
  - Use of SSDs, InfiniBand, faster cores, larger nodes...
- New workloads
  - Run full HiBench and BigBench
  - Run TPC-H (over Hive)
- Test other features such as using local datanodes as HDFS, use them only for temporary data. Use azure storage premium (more IOPS)
- Test other providers
  - Rackspace CBD, Amazon Elastic MapReduce
- Compare with IaaS deployments

www.bsc.es

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Questions?

aaron.call@bsc.es

BSC~Microsoft Research Centre