An architecture for distributed Network Intrusion Detection based on the Map-Reduce Framework

Marcelo D. Holtz, **Bernardo M. David**,
Laerte Peotta, Rafael T. de Sousa Jr.,

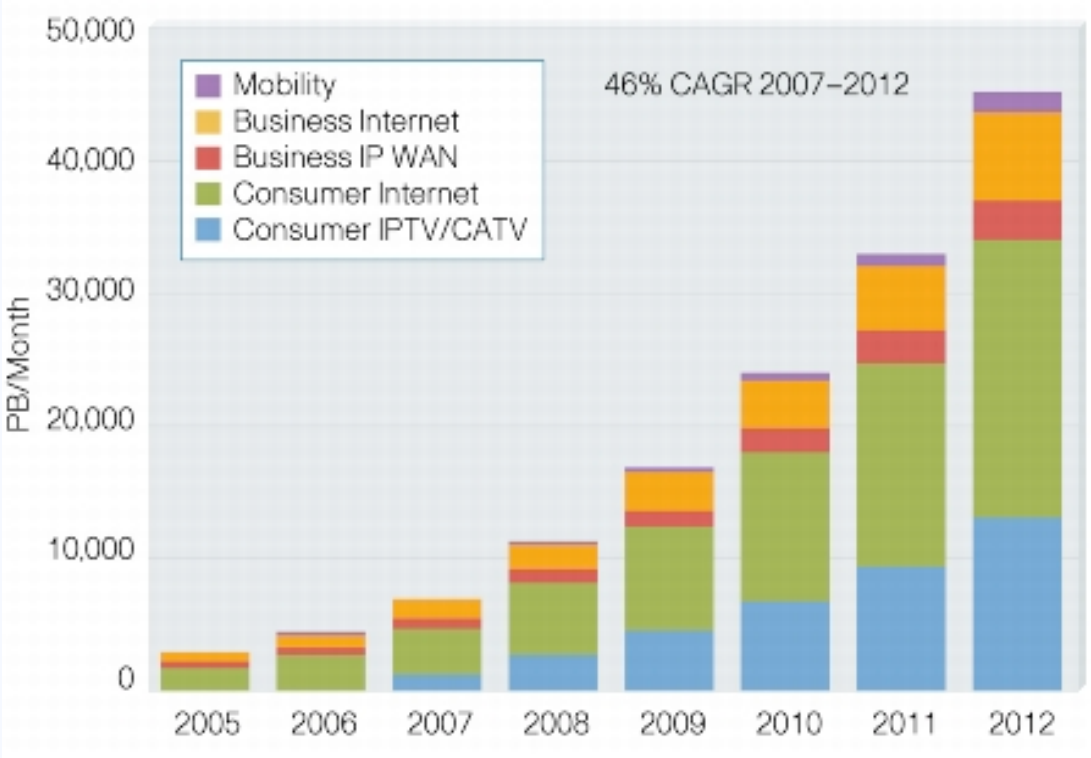
crypto & information theory Group

University of Brasilia

Introduction

- Intrusion Detection Systems (IDS) collect and analyse network traffic data detecting attack signatures [DEBAR1999]
- Regular IDS solutions are based on centralized traffic captures (e.g. at gateways).
- As network traffic grows and distributed data is collected, regular IDS systems become overwhelmed
- Previous works show that moving data storage and analysis to cloud environment is a viable solution for honeypot log analysis [AMARAL2011][LEE2010]
- In this talk we describe efficient and highly scalable cloud based IDS architectures

The Problem: Huge Datasets



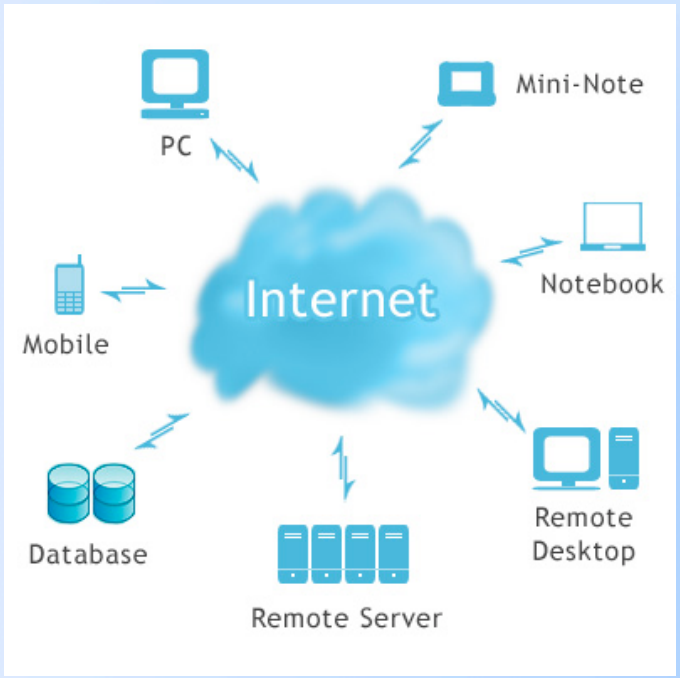
IP Traffic forecast

Source: Approaching the Zettabyte Era, Cisco 2008

- Network traffic is continuously growing
- Larger traffic volumes mean larger analysis datasets for IDS systems.

Possible Solution: The Cloud

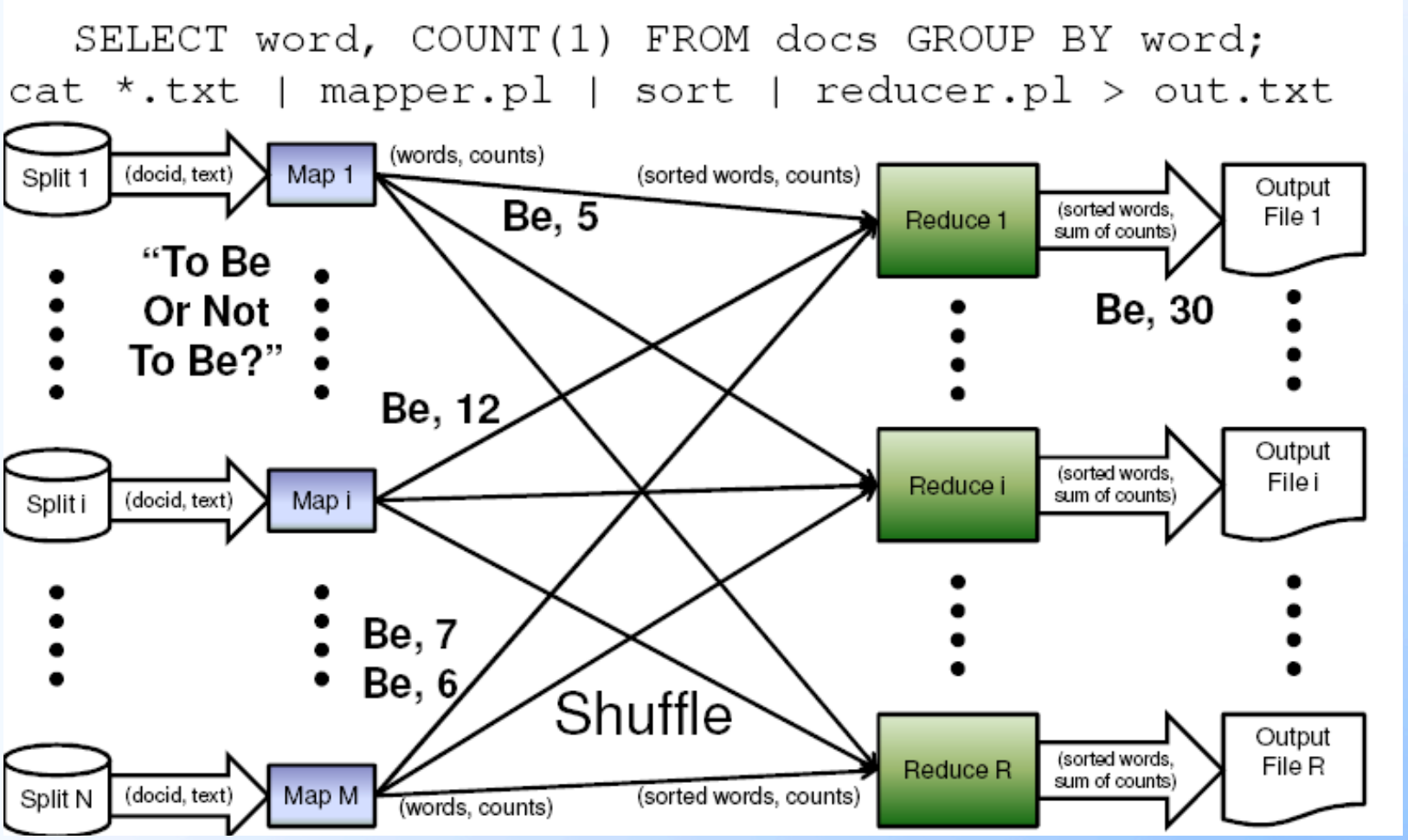
- Distributed file systems
- Distributed data processing
- High Availability through data replication
- Self healing fault tolerant systems



Hadoop: MapReduce and Beyond

- The Hadoop project includes efficient and scalable implementations of distributed filesystems, databases and processing techniques [DEAN2008]
- It is mainly composed by :
  - **HDFS:** Self-healing, high-bandwidth distributed filesystem
  - **MapReduce:** A framework for dsitributed data procesing
  - **Hive:** A distributed database
  - **Pig:** A data processing language for developing MapReduce programs
- Hadoop achieves high scalability supporting large scale clusters/ clouds:
  - Yahoo.com: 4000 nodes
  - Facebook.com: 2000 nodes

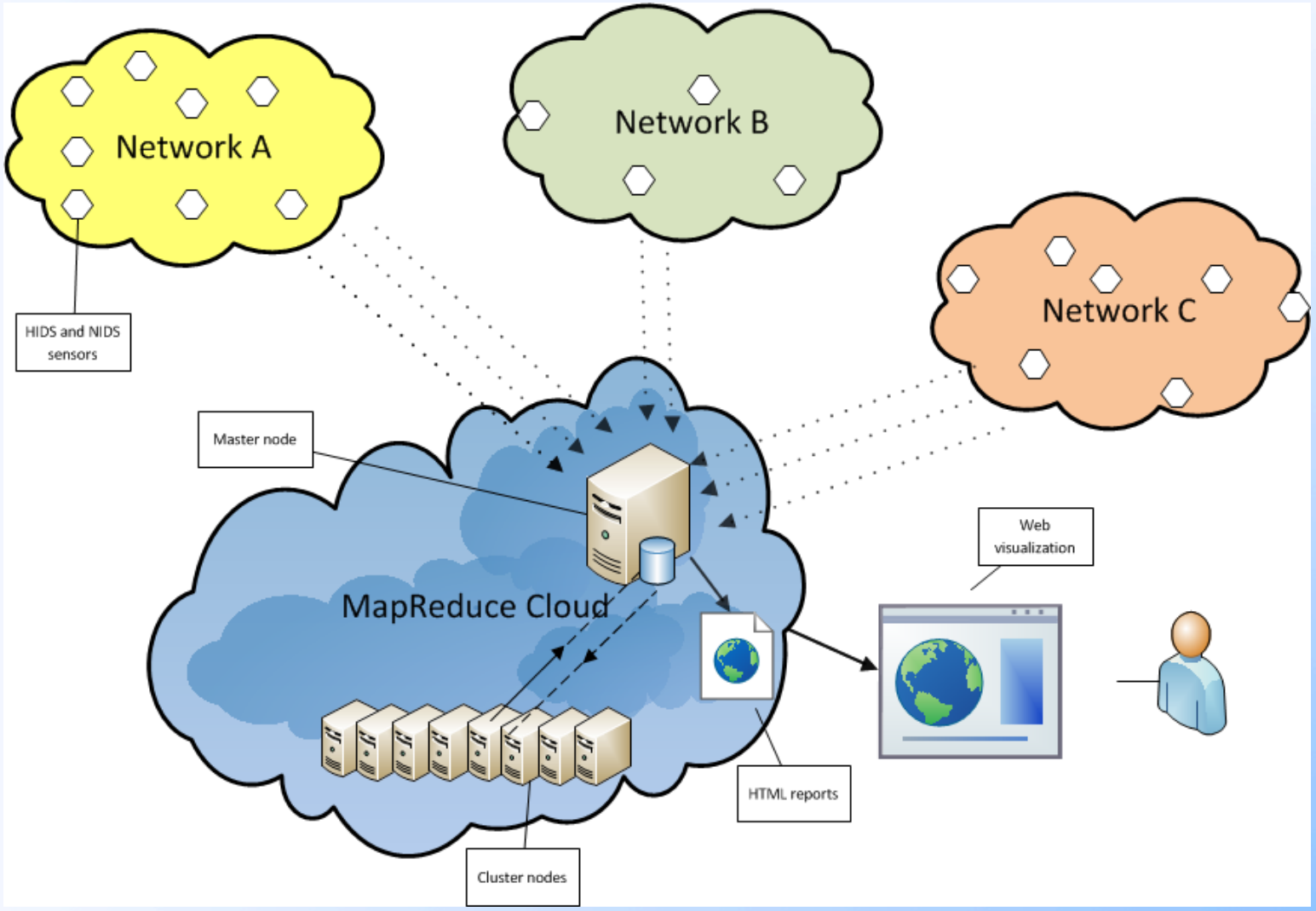
The MapReduce Framework



Distributed Intrusion Detection Systems

- DIDS collects data from different sources in different areas of the networks [HEADY1990][SNAPP1998].
- This data is combined and correlated to detect complex attack signatures [BASS2000][TIAN2010].
- We show that such a system can be efficiently implemented in current networks using cloud infrastructure for data storage and processing

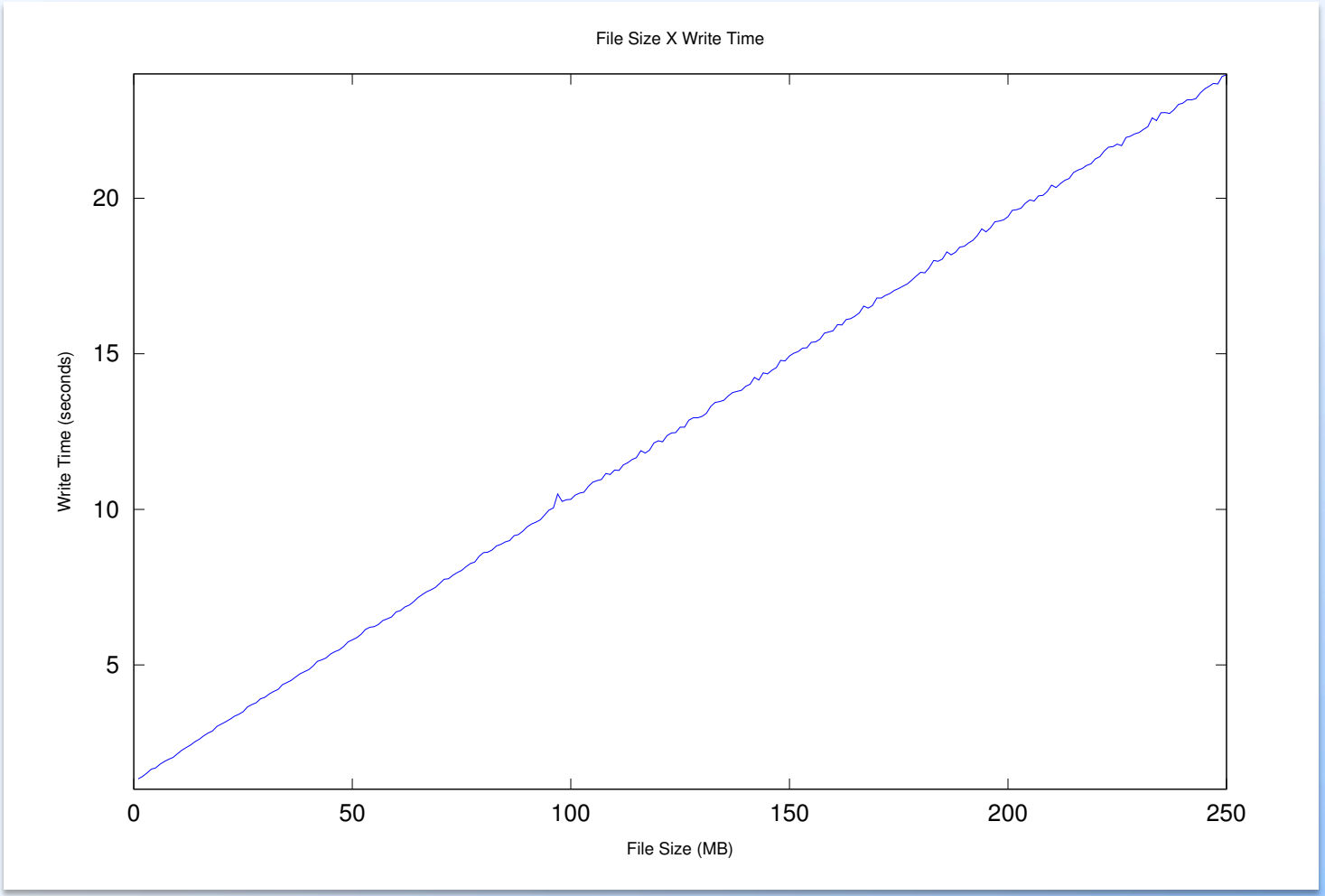
The DIDS architecture



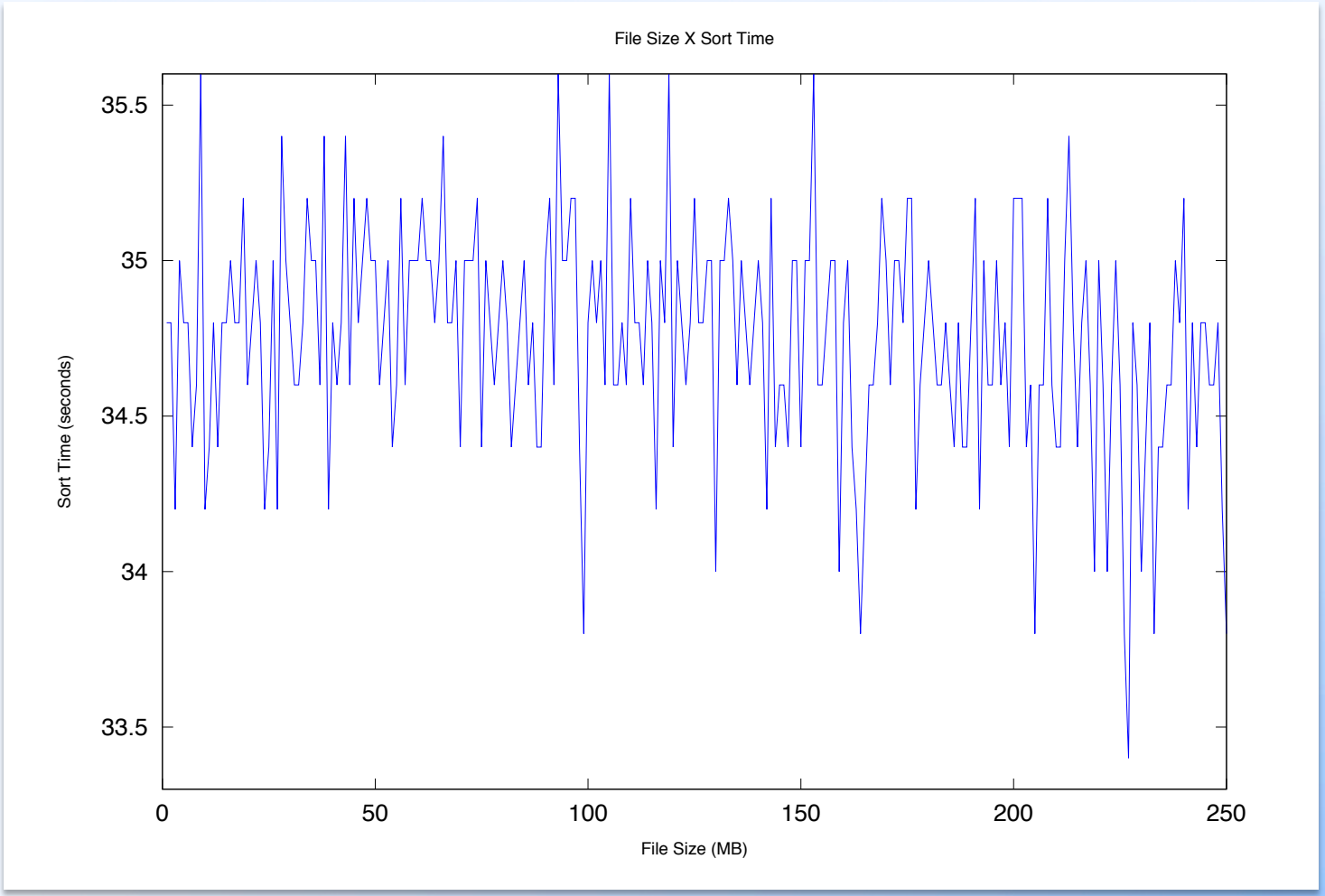
Case Study

- Cloud infrastructure:
  - 1 master node and 5 slave nodes
  - Intel(R) Core 2 Duo, CPU E7500 2.93GHz, 4 GB DDR667 RAM, 300 GB hard disk and 10/100 Mpbs Ethernet network interface.
  - 1.42 TB HDFS volume.
- Dataset of randomly generated logs of up to 250 MB
- We simulated the following operations: distributed data storage and sorting using the MapReduce framework.
- The results show it is feasible to implement such an architecture, since it is capable of processing large quantities of data in short periods.

Filesystem Write Performance



Data Sort Performance



Conclusion

- File system write time increases linearly in file size.
- The data sort performance is fairly stable, varying with network performance for all file sizes.
- Clearly, the synchronization overhead is more significant than actual data process for the file sizes analysed.
- Scalable distributed intrusion detection systems may be implemented based on cloud computing infrastructure.
- Cloud based solutions can be applied to obtain highly efficient and scalable network security mechanisms.
- Future works: leverage cloud infrastructure and standard log collecting mechanisms to perform log correlations for automatic attack detection.

Thanks!
Are there any questions?

Contact Info:

holtz@redes.unb.br
bernardo.david@redes.unb.br

http://www.redes.unb.br

crypto & information theory Group

University of Brasilia