How good can databases deal with Netflow data

Bachelorarbeit

Supervisor: bernhard | fabian@net.t-labs.tu-berlin.de

Inteligent Networks Group (INET)

Ernesto Abarca Ortiz

eabarca@net.t-labs.tu-berlin.de

etsinf UNIVERSIDAD POLITECNICA DE VALENCIA

INET berlin

OVERVIEW

INTRODUCTION

METHODOLOGY

RESULTS

- Hardware resource usage
- Data insertion
- Data quering

CONCLUSIONS

FUTURE WORK

QUESTIONS

GROWING TRAFFIC

Networks and Broadband Access are becoming even faster

+

Multimedia streaming, social networks, P2P and more users

+

More applications and protocols are available

=

More data is generated and crossing our networks

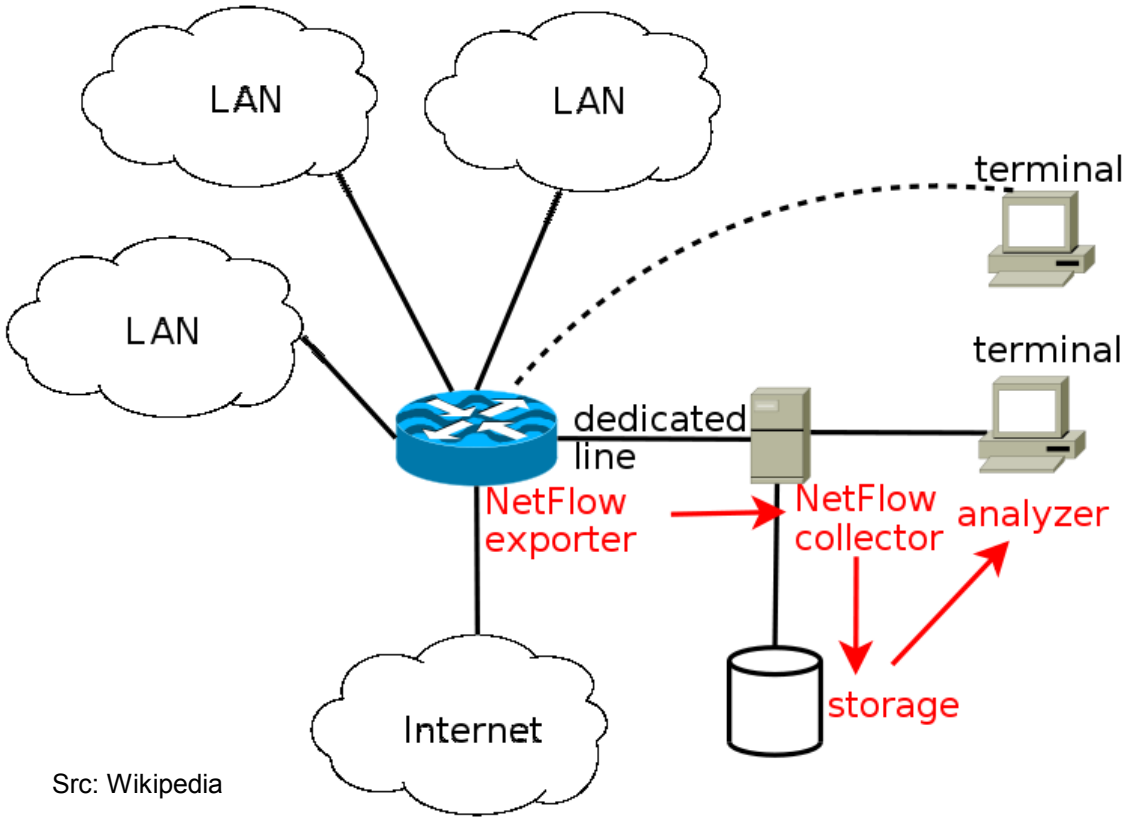
Network monitoring and data analysis is more complex

NETFLOW (I)

FLOW: Network connection's aggregated data

- Time stamp; source/destination for ip, mask, as and port
- Flow length in bytes / packets, other info

NETFLOW: Protocol for router / switches to send flow data



Src: Wikipedia

- How ?
- When ?
- Where ?

NETFLOW (II)

POSSIBLE USES:

- Traffic statistics
- Network Security
- Autonomous Systems (AS) routers behaviour

NETFLOWS ARE STORED IN BINARY FILES

- Vendor specific options/tools
- File-based limitations

WHAT ABOUT A DATABASE TO STORE NETFLOWS ?

- But which one ?
- And how ?

DATABASES

DATABASE: Organized and structured collection of data

DBMS: Complete system to manage databases

- Multi-server, backup, security...

Data can be encoded/stored in different field formats:

- Numeric, text, date/time, raw

Mar 17 2001 05:12:45 → 984802365
192.168.54.345 → 3232249689

- And fields can be indexed

SQL: Common and independent **S**tructured **Q**uery **L**anguage

TRANSACTION: Multiple SQL orders executed together

CHALLENGES

FIELD TYPE ?

- GENERIC: Only integer types
- SPECIFIC: IP Address, Time/Date

FIELD INDEXING ?

- Slow insertions, fast queries
- Is it worth ?

TRANSACTIONS FOR DATA INSERTION ?

HARDWARE REQUIREMENTS ?

BOTTLENECKS ?

METHODOLOGY

Netflow fields selection, database indexes and query design

FLOW-EXPORT tool enhancements:

- Support for postgresql & sqlite
- Field type conversion
- Transaction support
- Generate statistics:
  - > Disk I/O usage
  - > Time spent processing flows

Hardware monitoring for bottlenecks

- MRTG, top, dstat, iostat

Optimize hardware / software only if required

RESOURCE USAGE

STORAGE REQUIREMENTS:

- FLOW-TOOLS binary file:
  - > About 17 Million flows
  - > 1.1 Gbytes

- DATABASE: Integer fields

	NOT INDEXED	INDEXED (Max)
MySQL	0.9 GB	2.3 GB
POSTGRESQL	2.3 GB	5.9 GB
SQLITE	1.4 GB	2.9 GB

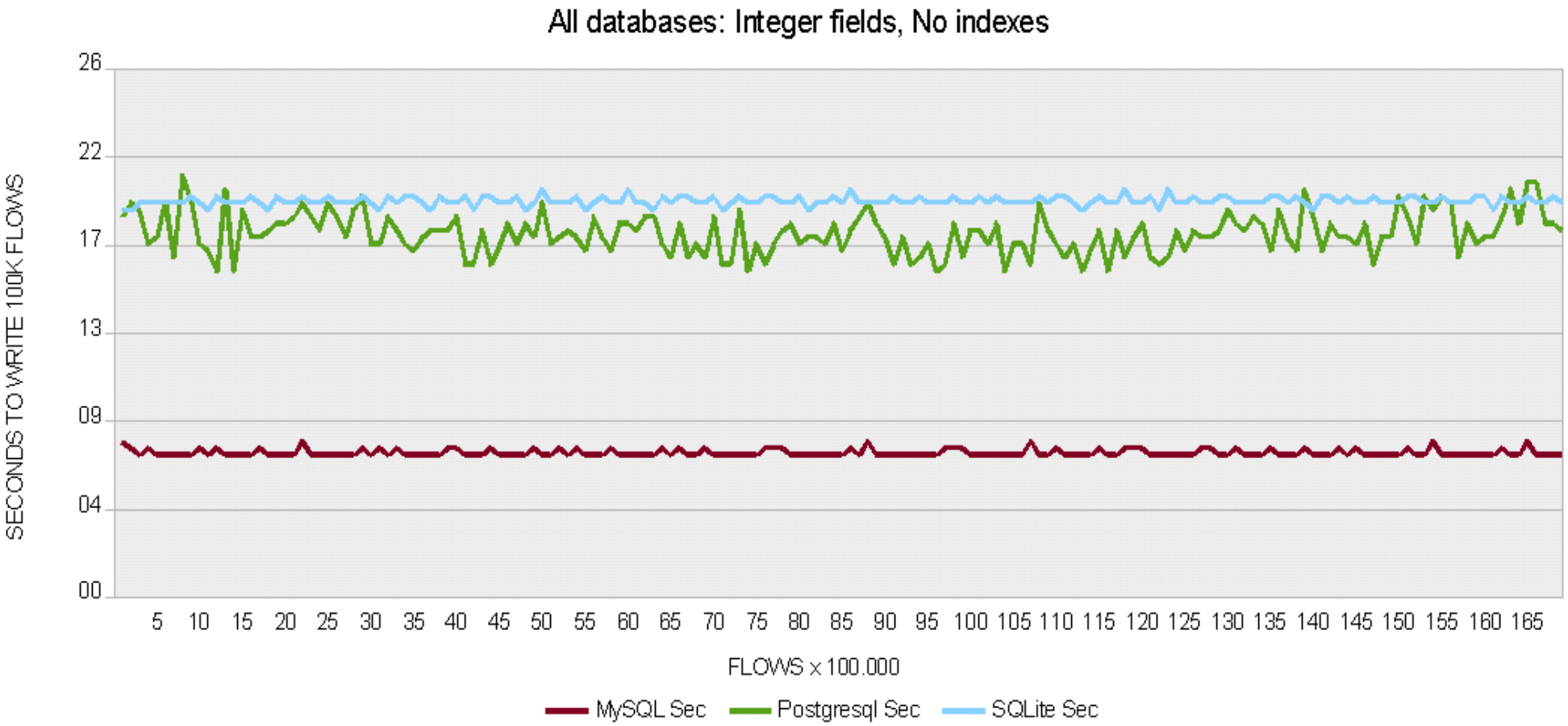
CPU USAGE

- BOTTLENECK: No multiprocessor support
- Mainly used for index creation or complex queries

RAM

- Some GB required for index creation and complex queries

DATA INSERTION (I)

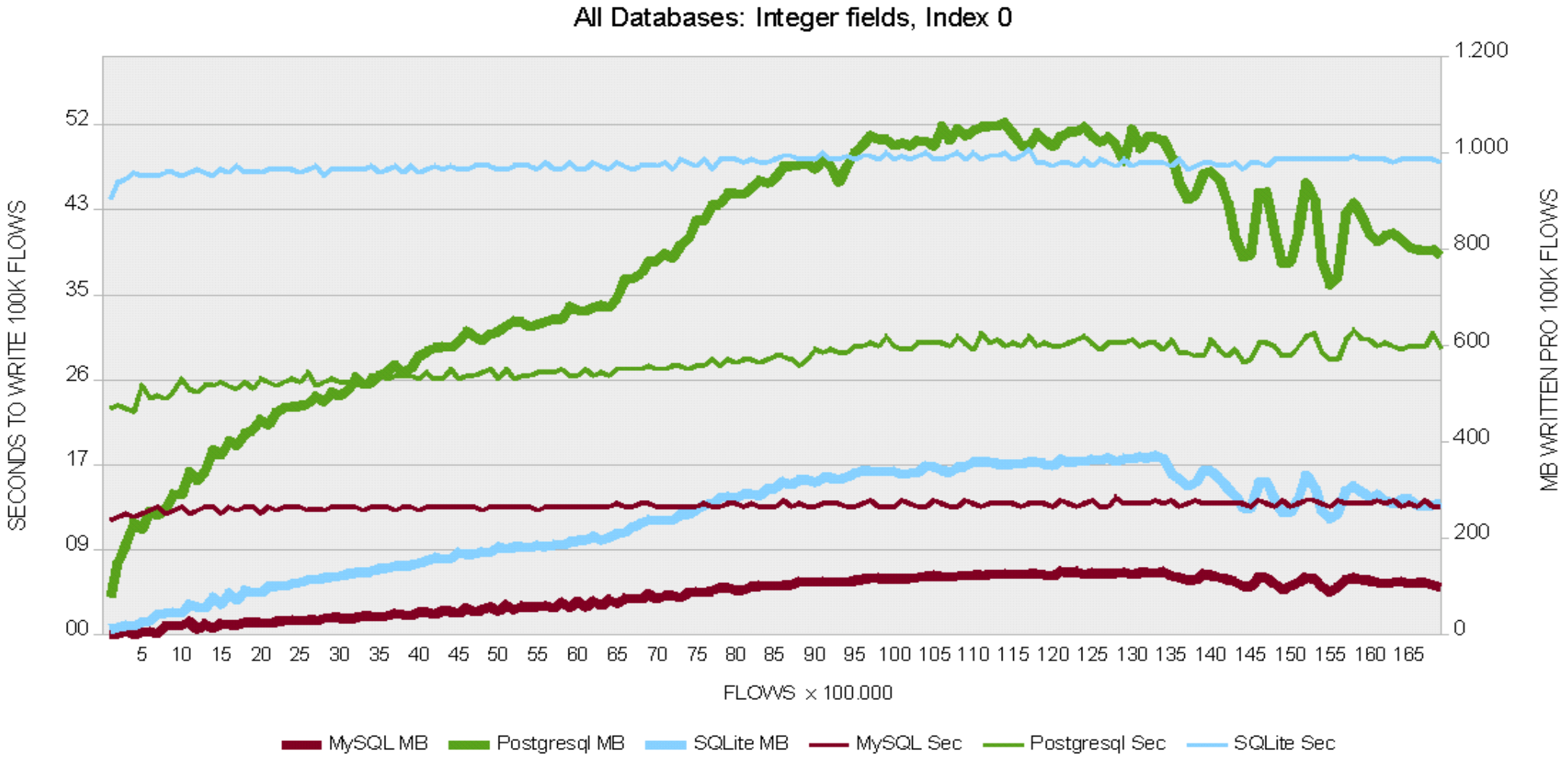


CONSTANT BEHAVIOUR

POSTGRESQL FLUCTUATES AS IT WRITES MORE DATA TO DISK

MYSQL IS TWICE FASTER: 14K FLOWS/SEC

DATA INSERTION (II)

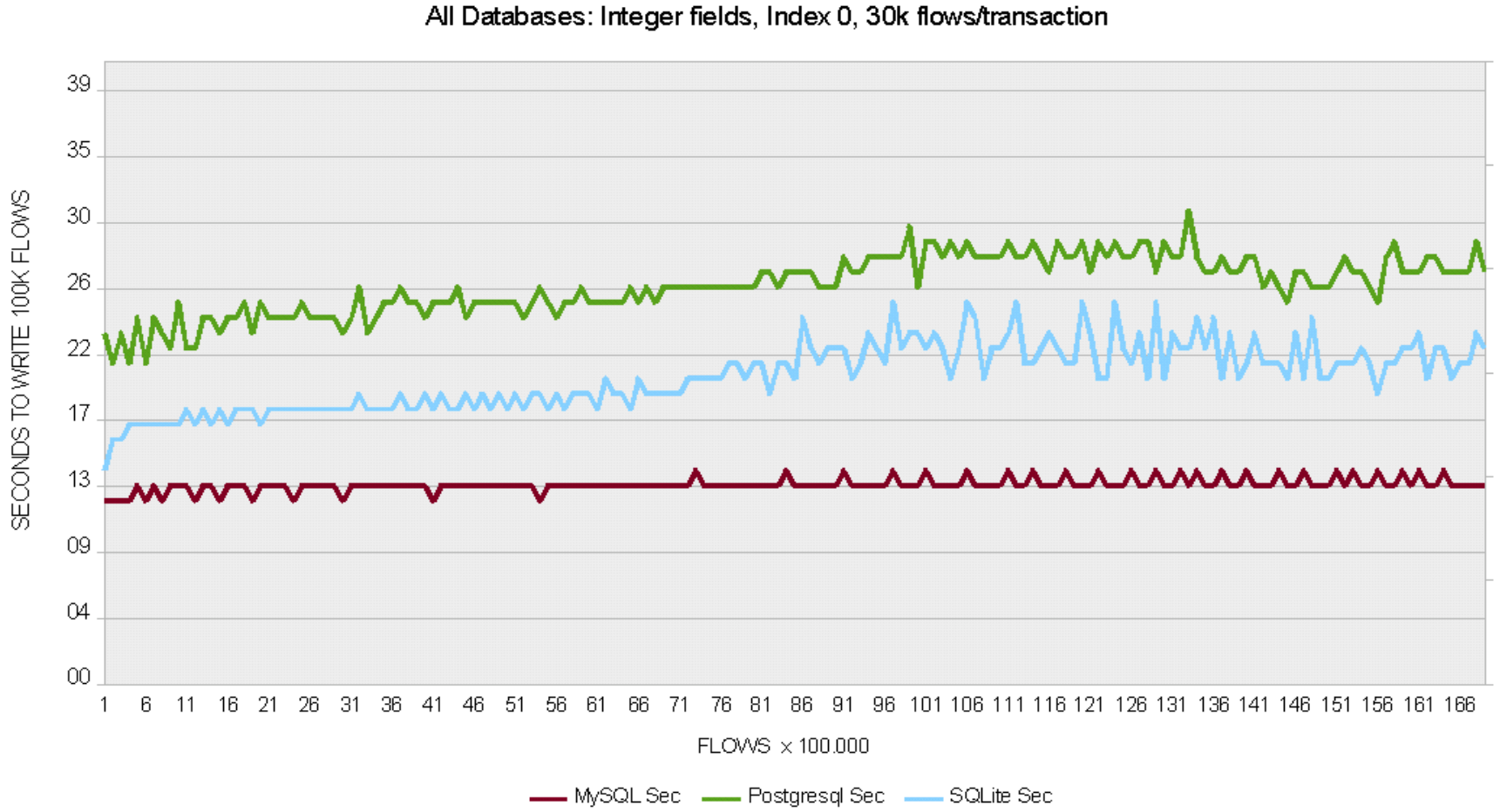


ON INDEXING, TIME DEPENDS ON AMOUNT OF DATA WRITTEN

SQLITE IS NEARLY TWO TIMES SLOWER THAN POSTGRESQL

EXTRANGE BEHAVIOUR AT THE END

DATA INSERTION (III)

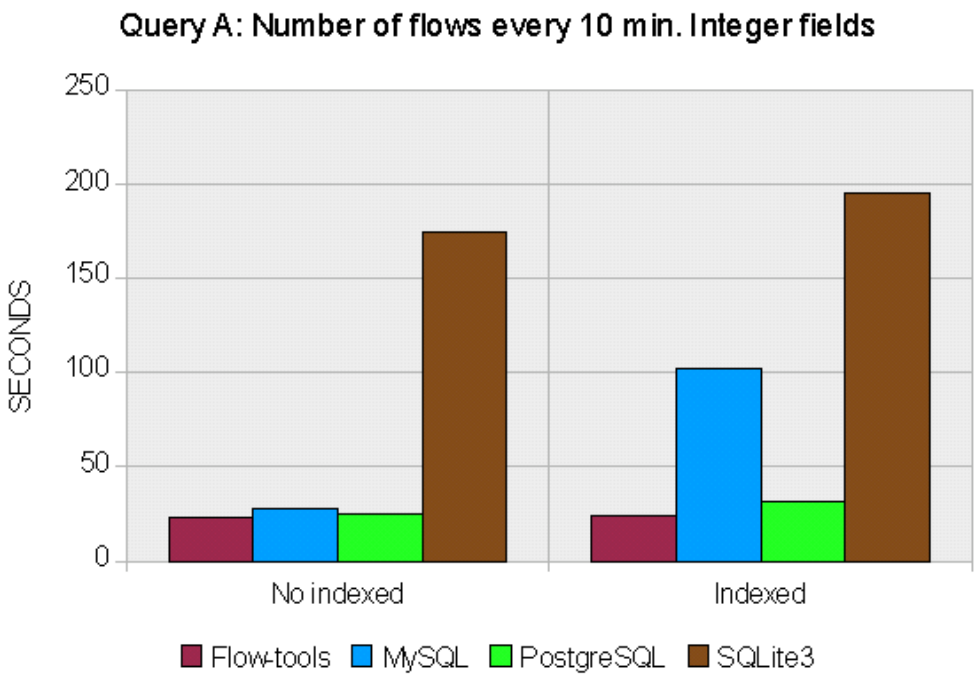


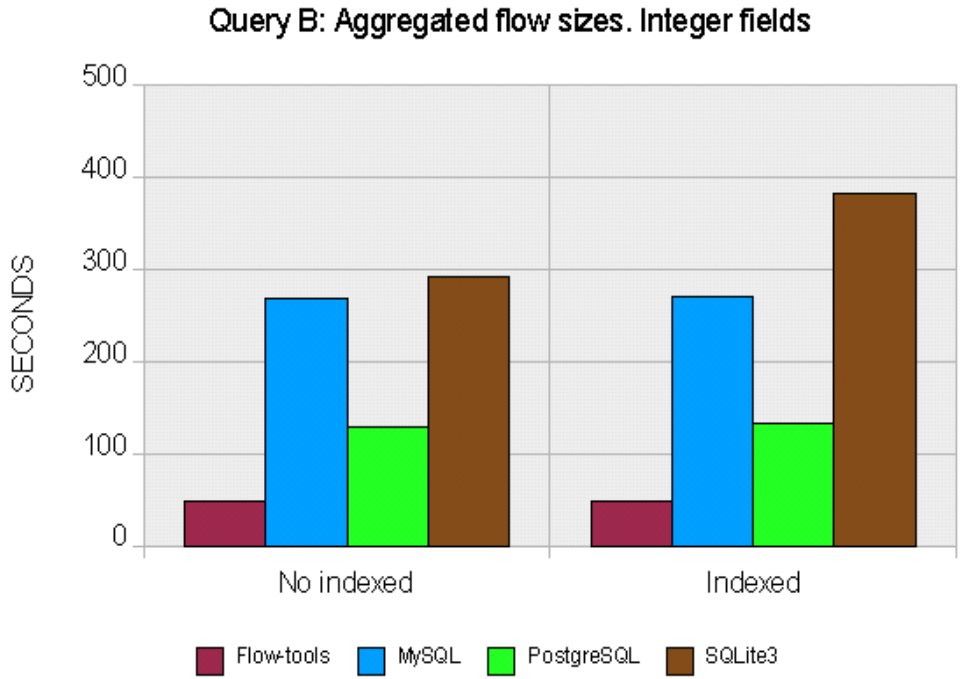
TRANSACTIONS MAKE SQLITE FASTER THAN PGSQL; OTHERS JUST A BIT

INDEXING REQUIRES MUCH MORE RESOURCES

RESULTS

DATA QUERY



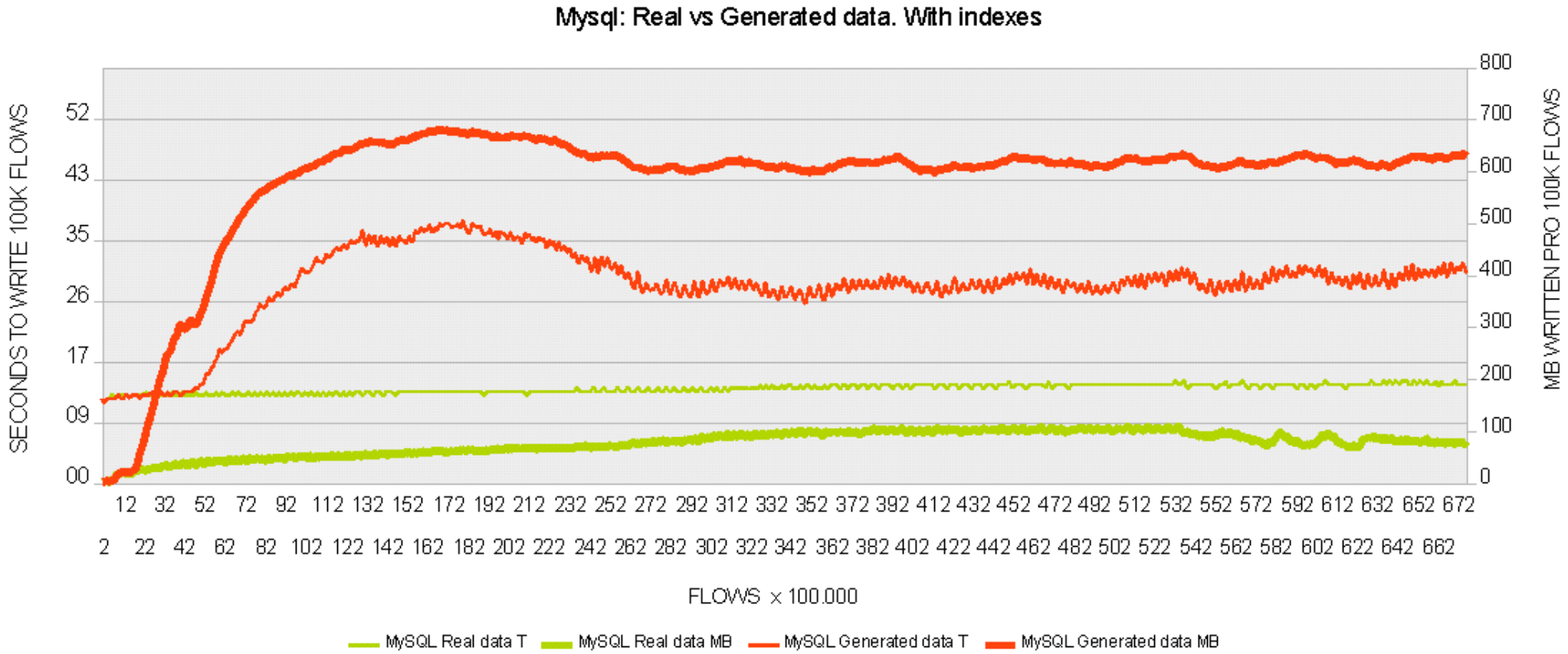


- Flow-tools always require the same execution time
- Databases are faster when specific data is requested

- Flow-tools are much faster aggregating data
- Databases required more than 4 gb of ram

There are no big differences when querying Integer or specific fields

DATA DIVERSITY AND SCALABILITY



DATA DIVERSITY: Important in index creation and aggregation queries

SCALABILITY: Is feasible but can depend on data diversity

CONCLUSIONS

IS IT FEASIBLE TO USE NETFLOW WITH DATABASES ?

- Yes!
- But requires: Twice HDD storage and about 8 GB RAM
- Database multiprocessor support is recommendable
- Data diversity and traffic pattern can be important for indexing

BEST SOLUTION: MySQL

- Integer fields
- No indexes
- Transactions: optional

FUTURE WORK

- DATABASE NORMALIZATION

- DATABASE SERVER OPTIMIZATION

- COMPARE IT WITH NON SQL DATABASE SYSTEMS

Questions ?

Answer: 42