VITESSE DATA

DeepGreen DB: 性能优化、开发方向

CK Tan

Vitesse Data, Inc.

Postgres Conference China 2016 中国用户大会

PostgreSQL

创始人

- CK Tan
- 田丰
  - 美国威斯康星大学硕士 / 博士、数据库系
  - 原 Greenplum Database 团队 2006-2009
    - gpfdist, external table, column store, executor opt, storage opt, gpmon, hashagg opt, etc.

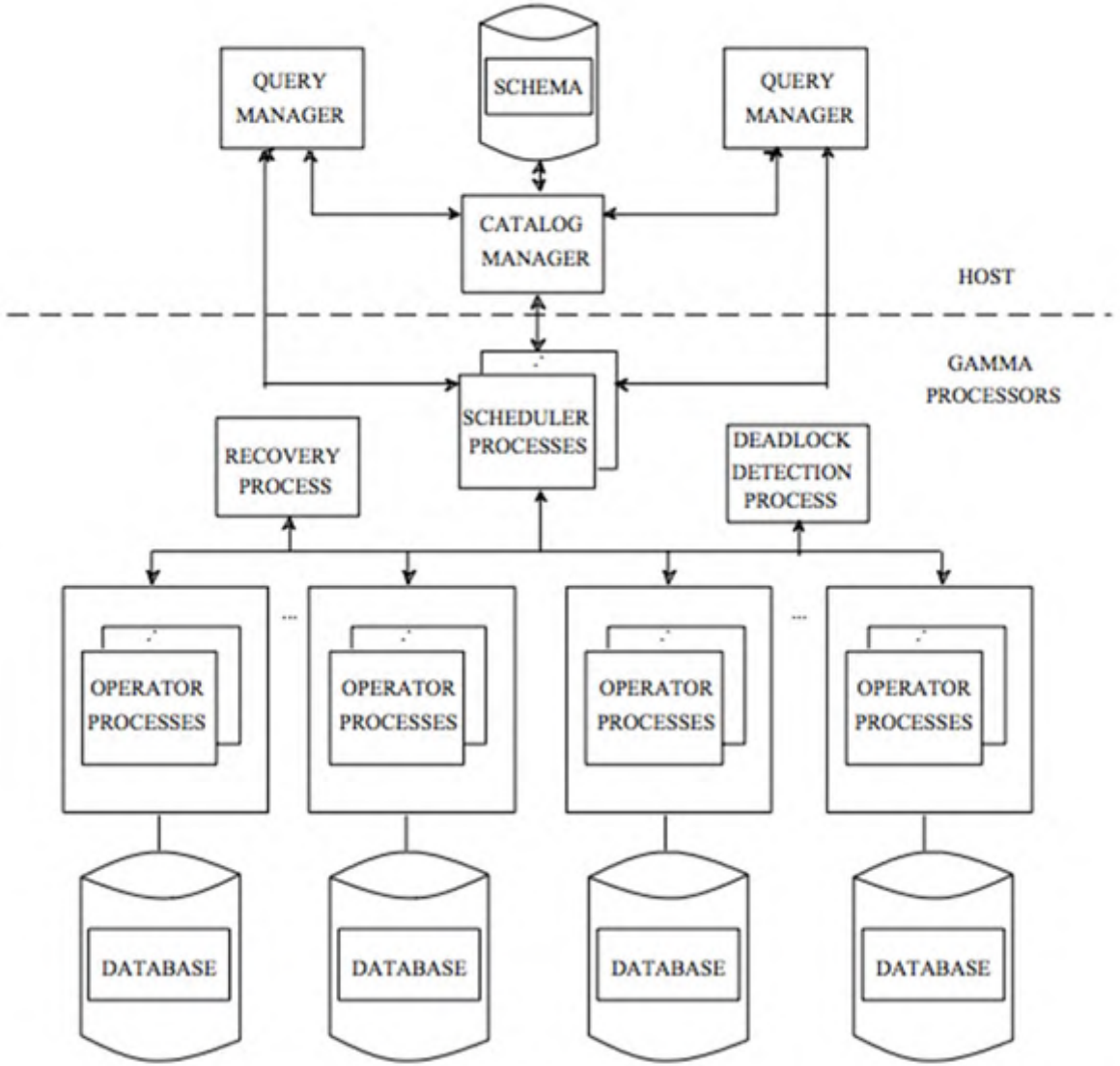

UNIVERSITY OF WISCONSIN - MADISON
NUMEN LUMEN

MPP 起源：威斯康星大学

- Gamma Database Machine, 1985-1990
- 20 VAX 11/750
- 32 intel iPSC/2

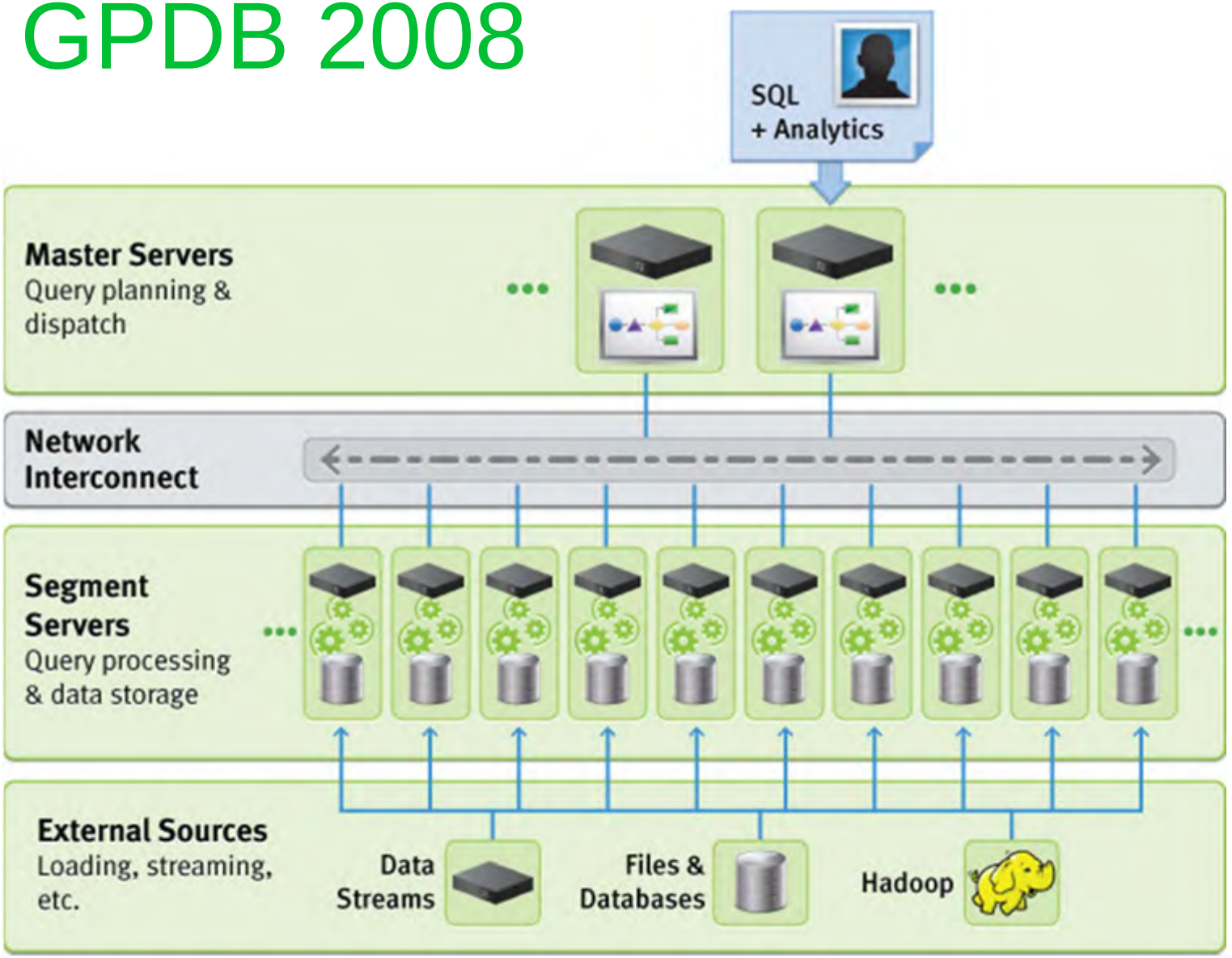

intel

Gamma 1990

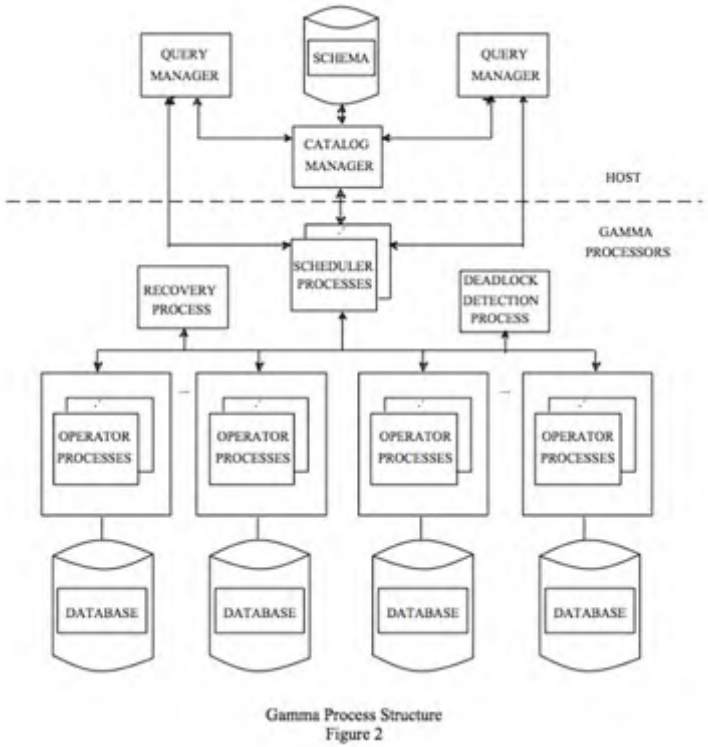


Gamma Process Structure
Figure 2

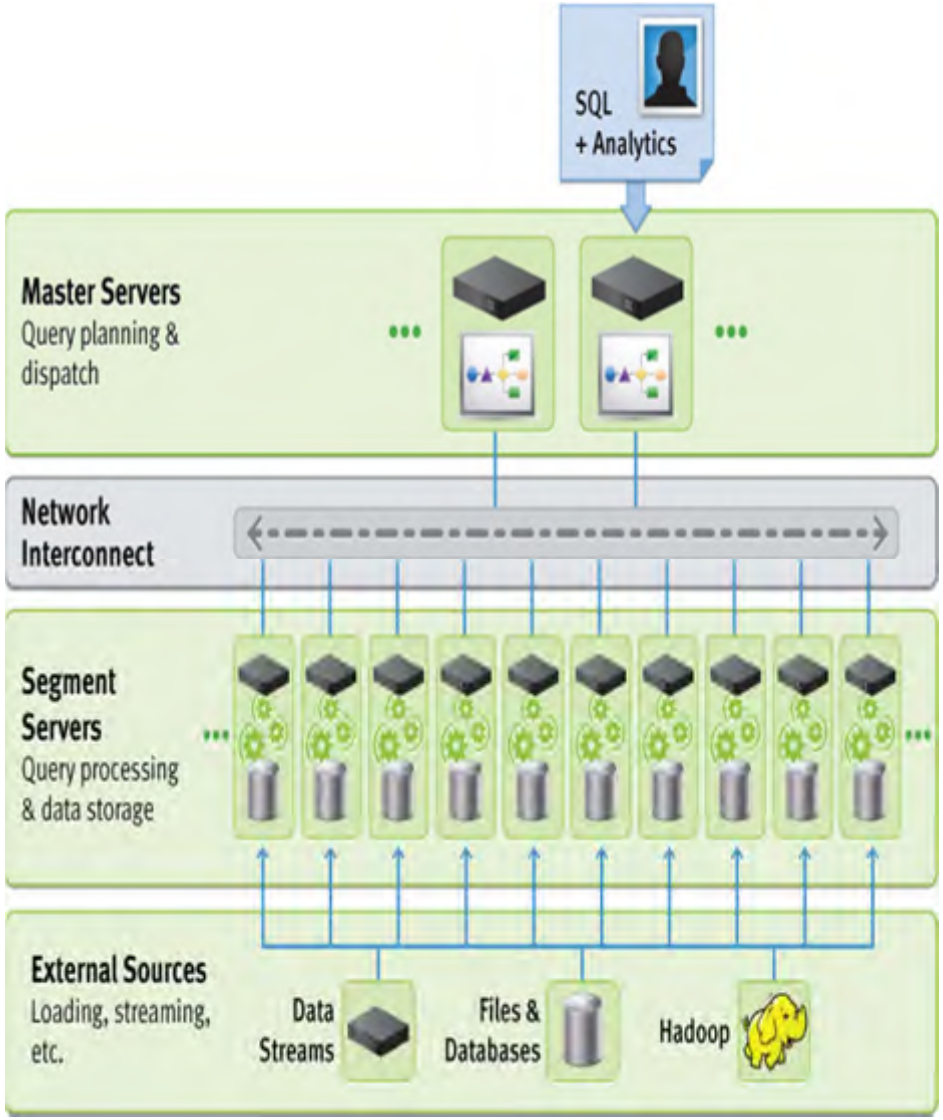
GPDB 2008


SQL
+ Analytics
Master Servers
Query planning &
dispatch
Network
Interconnect
Segment
Servers
Query processing
& data storage
External Sources
Loading, streaming,
etc.
Data
Streams
Files &
Databases
Hadoop

万变不离其宗



Gamma Process Structure
Figure 2



Vitesse Data 简介

- 成立于 9月2014
- 产品
  - vitesse db 9.3, 9.4, 9.5
  - deepgreen db, loft, xdrive
- 产品发布：30+ 次


兄弟就是生产力

性能优化

- LLVM JIT
- Hash table
- Spill Framework
- Planner
- CSV Parser - SIMD
- lz4, zstd 压缩
- approx count distinct

LLVM JIT 黑技术

此 JIT 非彼 JIT

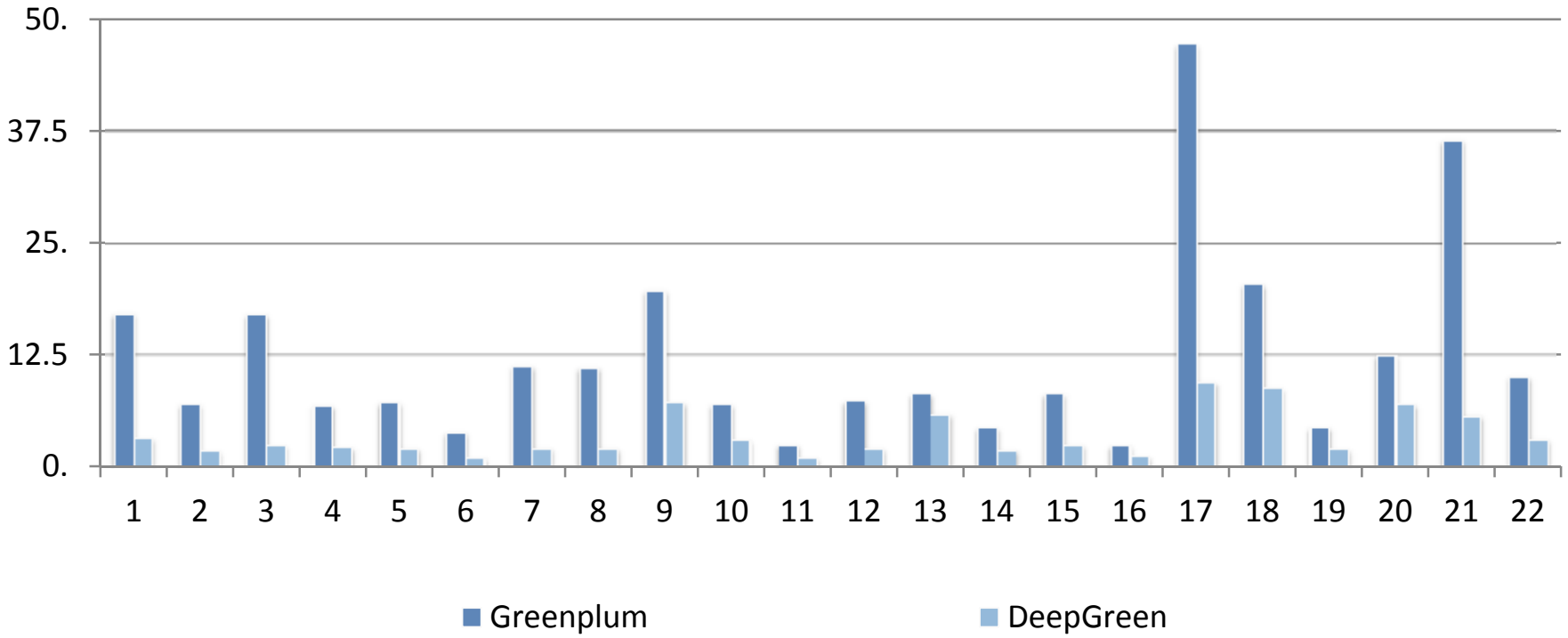
Just-in-time compilation of query plan

- 将整个咨询转换成一个汇编语言程式
- 有效去除 x86 执行器与内存的摩擦

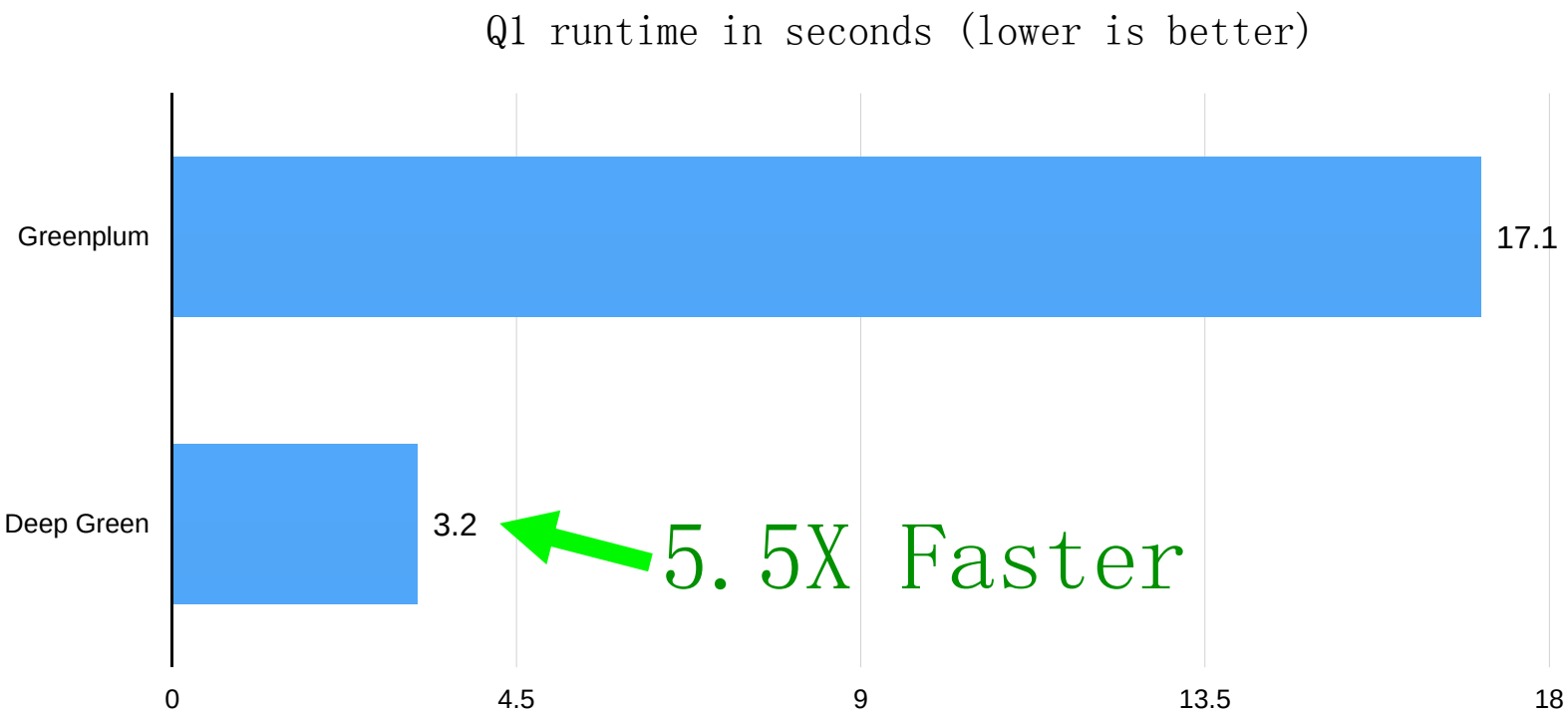
LLVM 只是工具。用法不同，效果各异。

shipping since … 3/2015

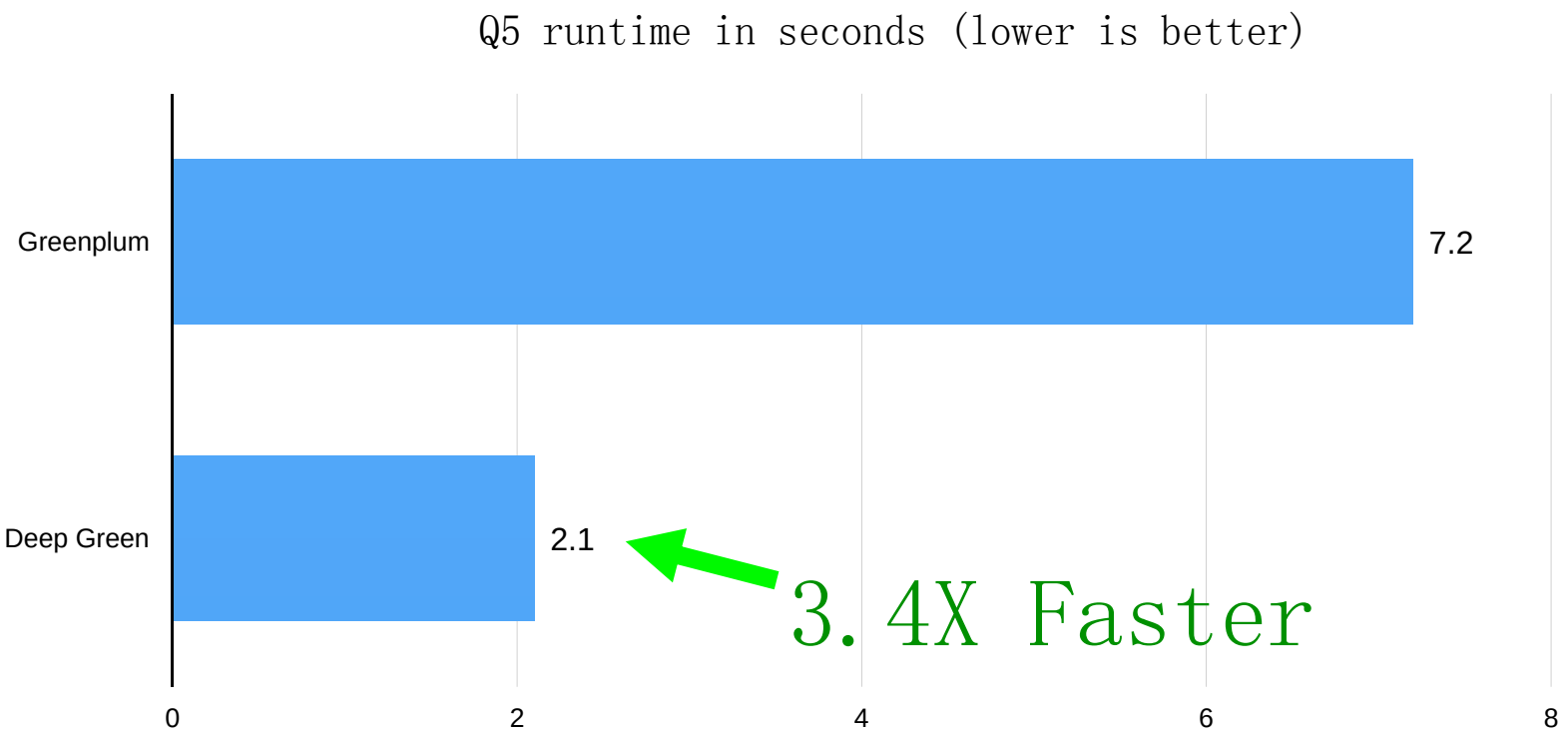
TPCH 10g


50.
37.5
25.
12.5
0.
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Greenplum
DeepGreen

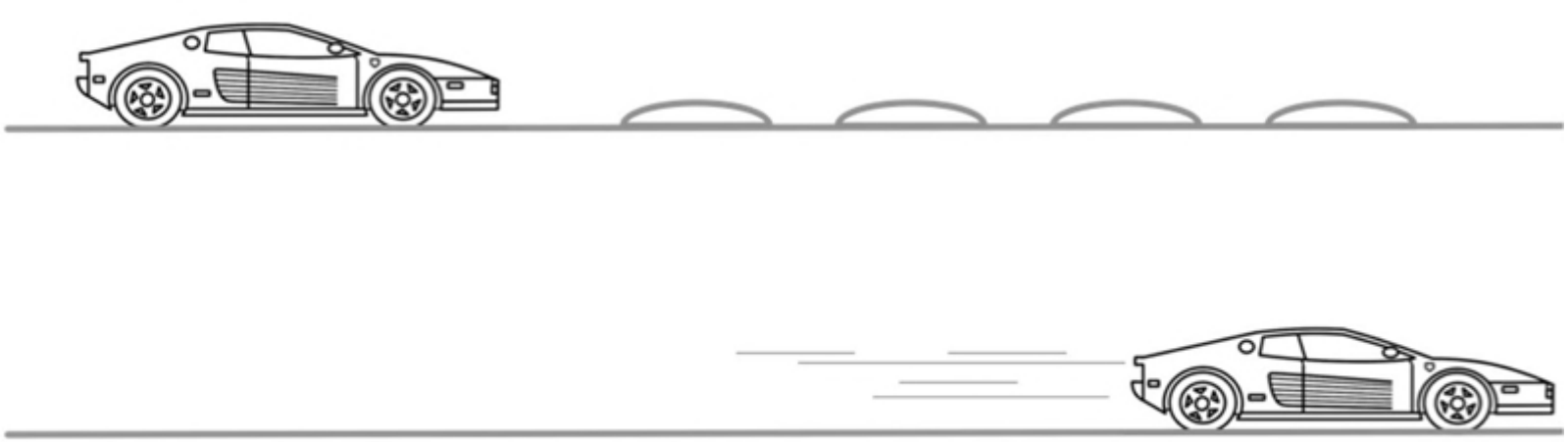
TPCH Q1


Q1 runtime in seconds (lower is better)
Greenplum
17.1
Deep Green
3.2
5.5X Faster
0
4.5
9
13.5
18

TPCH Q5


Q5 runtime in seconds (lower is better)
Greenplum
7.2
Deep Green
2.1
3.4X Faster
0
2
4
6
8

请别如此对待您的x86



CSV 解析器

- 已完成 SIMD: parse 8 bytes at a time.
  - 比 GPDB CSV parser 大约快 2 倍。

- 开发中 AVX-512: parse 64 bytes at a time
  - 估计可比现在加速至少 2 倍。

压缩器：lz4, zstd, zlib

	压缩率	压缩时间	解压缩时间
memcpy	1.00	4200 MB/s	4200 MB/s
lz4	1.61	690 MB/s	2220 MB/s
zstd	2.88	240 MB/s	620 MB/s
zlib -1	2.73	59 MB/s	250 MB/s

10倍

10倍

4倍

2倍

Approximate count distinct

select count(distinct url) from page_view;

- 必须记住每一个URL — 非常耗内存。

select approximate_count_distinct(url) from page_view;

- hyper log log algorithm
- 1% 错误，但快 3 倍。


XDRIVE

xdrive

理解

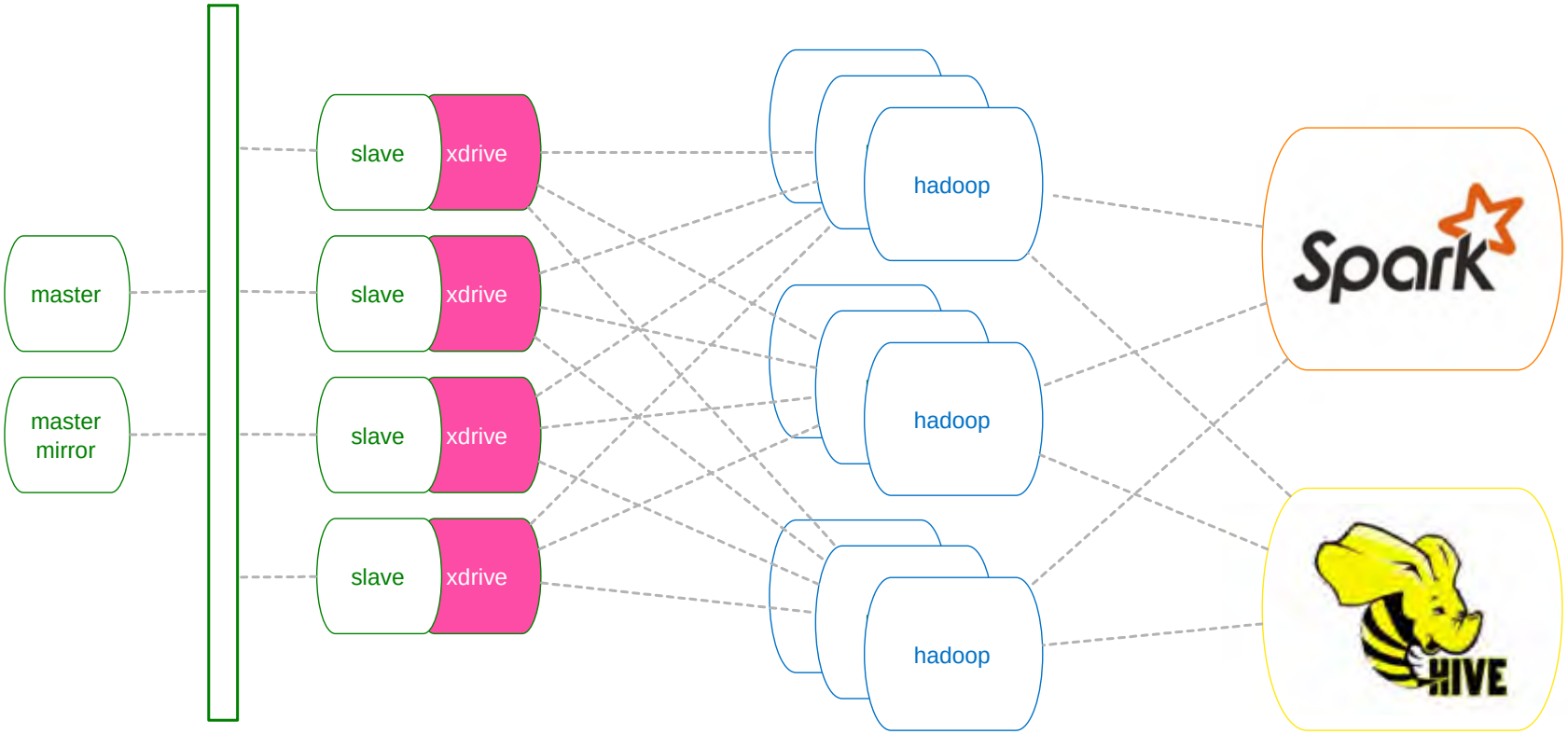
- gpfdist for Hadoop
- DeepGreen DB 上加了 HAWQ 功能

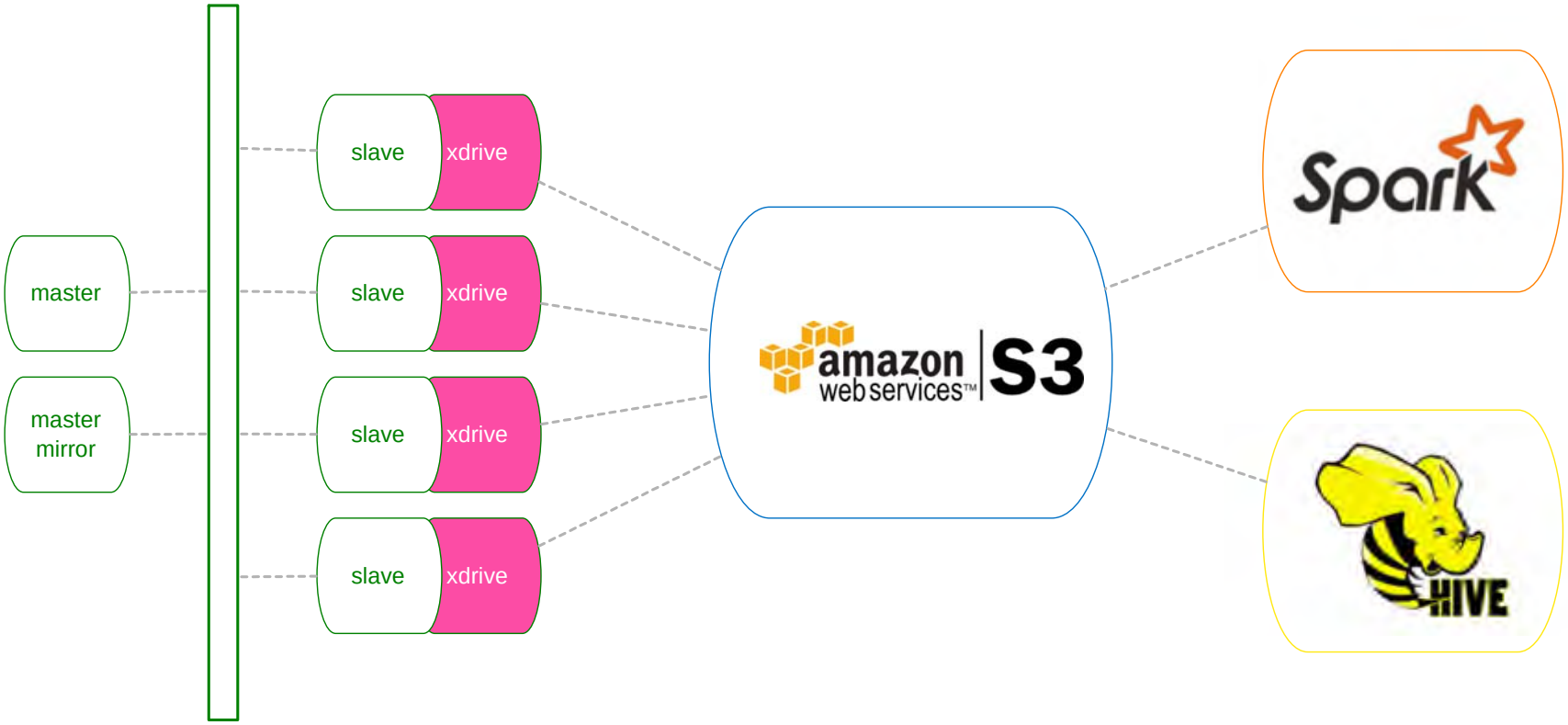
高扩展性

- HDFS, S3, Hive
- NFS, Ceph FS
- Local FS

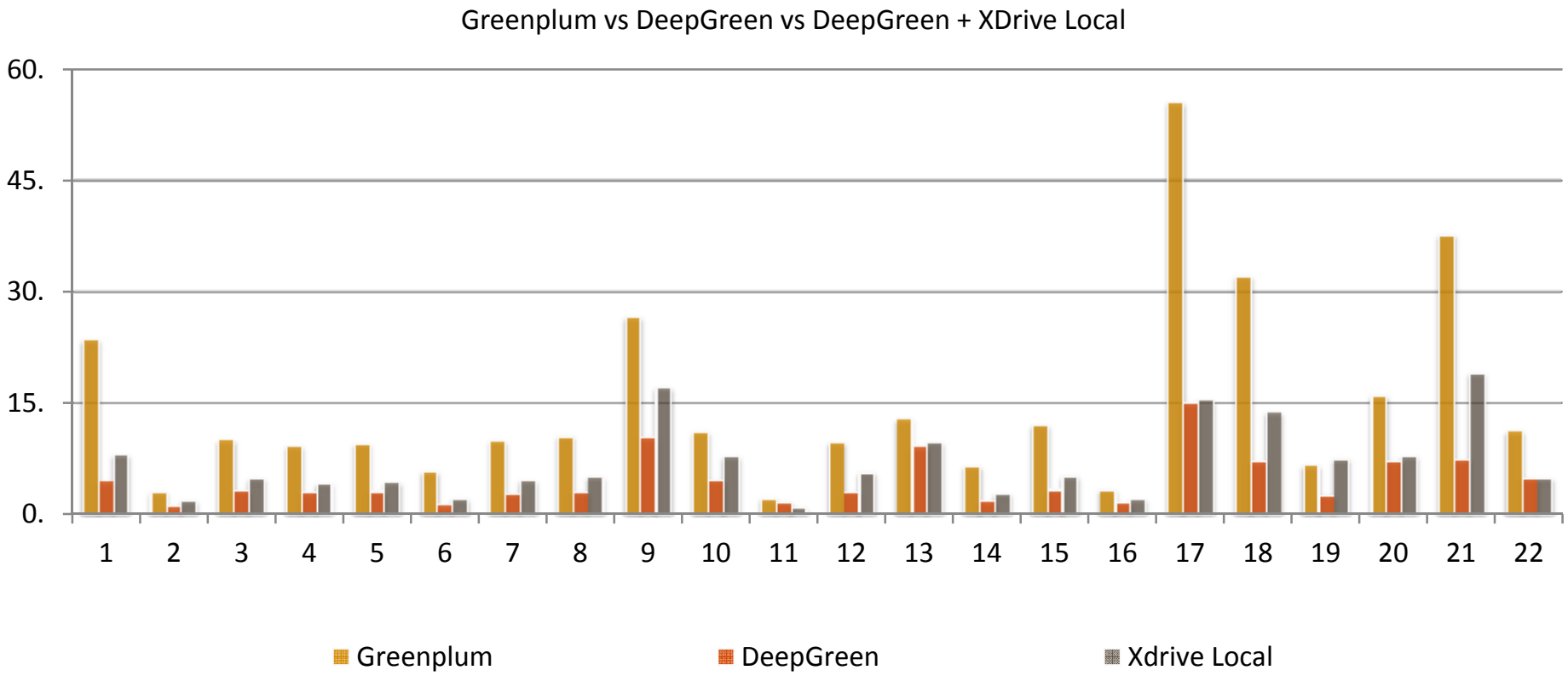
xdrive: 灵活、便捷

- 只需一个配置文件
- 可以在任何地方灵活运行
- 可以连接多个 NFS, Hadoop, S3 等系统
- 可以转换多种文件 csv, parquet, spq, orc
- 嵌入 DeepGreen. 服务 Spark.


master
master
mirror
slave
xdrive
slave
xdrive
slave
xdrive
slave
xdrive
hadoop
hadoop
hadoop
Spark
HIVE


master
master mirror
slave
xdrive
slave
xdrive
slave
xdrive
slave
xdrive
amazon web services S3
Spark
HIVE

xdrive: 性能


Greenplum vs DeepGreen vs DeepGreen + XDrive Local
60.
45.
30.
15.
0.
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Greenplum
DeepGreen
Xdrive Local

xdrive: 性能 [2]

- xdrive 外表性能远高于 GPDB 外表
- xdrive 外表性能高于 GPDB 内建的 Heap表
- xdrive 外表性能略低于 DeepGreen 内建的 Heap表
  - 主要是数据从 HDFS 读出

xdrive: 双活

企业里的两派人马

- SQL 组
- spark / hive

- 通过 xdrive
  - 共享原数据
  - 共享分析结果

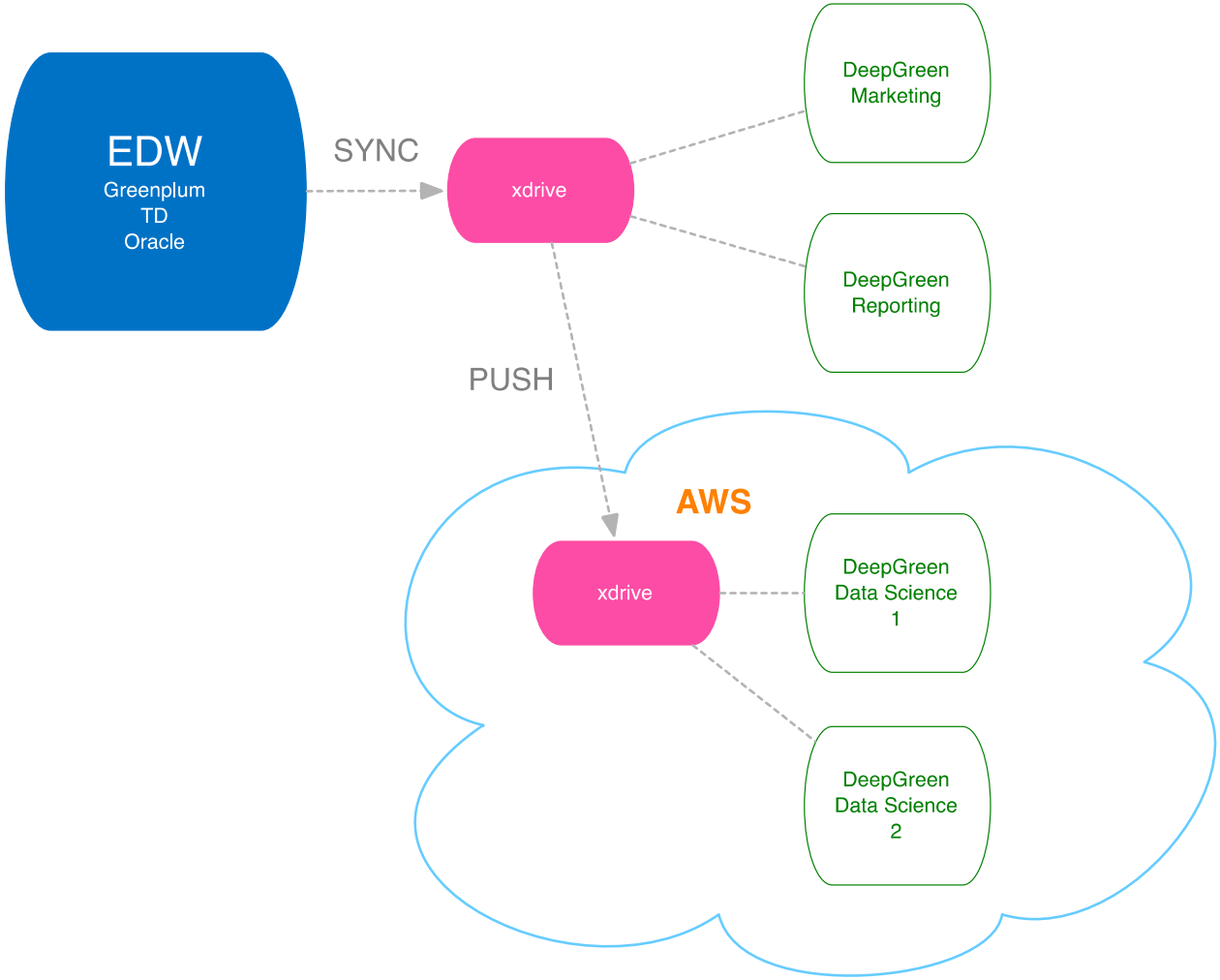
消除对立

xdrive: Data Mart

- EDW 太贵或太忙
- 各个部门有不同的需求
  - 不同数据
  - 不同新鲜度
  - 不同用户群

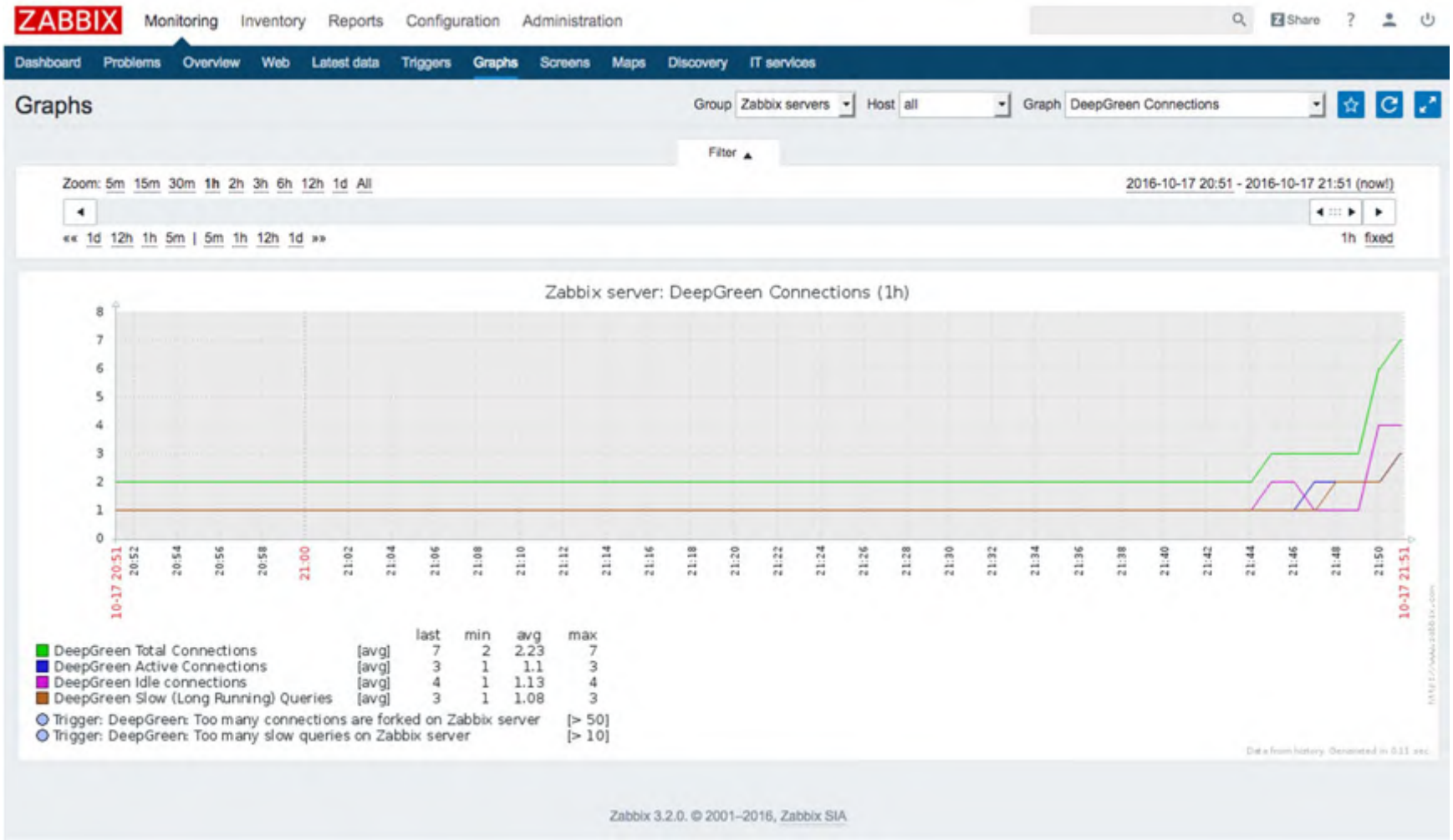
- 复制数个 data mart 可大量减轻 EDW 负荷

xdrive: Data Mart




ZABBIX

DeepGreen DB + Zabbix 监控



DeepGreen 2017 开发方向

- xdrive plugin
- session query monitor
- new utilities with local agent (in GO)
- new interconnect with local hub
- GPU

颠覆性的 PG 9.6

- 一年一版 ＝ GPDB 永远落后 PG 十年
  - 必须考虑非常手段
- multiple backend
  - 足以完全改变 GPDB 的进程架构。
  - 针对 GPDB 的 SLICE，改成 co-backend.
  - 针对 GPDB 的镜子，改成 pg standby.
- 其他

Thanks!

Q & A