Making Elephants Fly

Jez Wain, Escala Competency Centre
Atos, Grenoble
Not so long ago

Power

x86

image: www.thegreenhead.com
RISC vs. CISC
Data lock-in
Power

x86

Common data format
Power x86

Common application data format

Linux-el
Common application data format

Power

Database Files

x86
Cold Standby
Cold Standby
Cold Standby

Power

Database Files

x86
Cold Standby - Fail-back
Cold Standby

Power → Database Files → AWS
Improving the recovery time and recovery point
PostgreSQL Operation
Hot/Warm Standby

Ubuntu/Power

Ubuntu/x86

Transaction Stream

Log Shipping

Archived WAL

WAL Directory
INSERT INTO rtdata (ts, ip, value) VALUES(now(), inet_server_addr(), 0)
<table>
<thead>
<tr>
<th>ID</th>
<th>Timestamp</th>
<th>DB IP</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1630</td>
<td>2018-05-14 19:37:16</td>
<td>192.160.160.170</td>
<td>0</td>
</tr>
<tr>
<td>1628</td>
<td>2018-05-14 19:37:01</td>
<td>192.160.160.170</td>
<td>0</td>
</tr>
</tbody>
</table>

Both servers running

- The web-server is writing on main database
- The main database is replicated on the hot standby server.
- The web-server can read from both databases.
Primary DB halted

- Data still available on secondary (read-only)
- Server cannot write until secondary promoted to primary

2018-05-14 19:44:17.390369+02 Cannot execute query
INSERT INTO rtdata (ts, ip, value)
VALUES(now(), inet_server_addr(), 0)
Standby promoted to primary

- After a few seconds the hot-standby is promoted as main db server.
- Data are now written through the hot standby server (10.197.160.58)

<table>
<thead>
<tr>
<th>ID</th>
<th>Timestamp</th>
<th>DB IP</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1723</td>
<td>2018-05-14 19:46:19</td>
<td>10.197.160.58</td>
<td>0</td>
</tr>
<tr>
<td>1722</td>
<td>2018-05-14 19:46:14</td>
<td>10.197.160.58</td>
<td>0</td>
</tr>
<tr>
<td>1721</td>
<td>2018-05-14 19:46:09</td>
<td>10.197.160.58</td>
<td>0</td>
</tr>
<tr>
<td>1720</td>
<td>2018-05-14 19:46:04</td>
<td>10.197.160.58</td>
<td>0</td>
</tr>
</tbody>
</table>
Postgres-XL
Postgres-XL: Sharded PostgreSQL

- Scale-out, shared-nothing cluster

Coordinator
Global transaction manager

Postgres-XL

Data nodes
Postgres-XL: Sharded PostgreSQL

- Scale-out, shared-nothing cluster
Postgres-XL: Sharded PostgreSQL

- Scale-out, shared-nothing cluster
PostgreSQL on Power9 Performance Benchmark
- TPC-B style transactional workload

- Five queries:
  SELECT, UPDATE, INSERT
  Read-only (SELECT) option

- Simulates multiple clients
pgbench scaling factor

<table>
<thead>
<tr>
<th>Table</th>
<th># rows</th>
</tr>
</thead>
<tbody>
<tr>
<td>pgbench_branches</td>
<td>1</td>
</tr>
<tr>
<td>pgbench_tellers</td>
<td>10</td>
</tr>
<tr>
<td>pgbench_accounts</td>
<td>100,000</td>
</tr>
<tr>
<td>pgbench_history</td>
<td>0</td>
</tr>
</tbody>
</table>

Scale = 1

<table>
<thead>
<tr>
<th>Scale</th>
<th>Rows in accounts</th>
<th>DB size</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>10,000,000</td>
<td>1.5 GB</td>
</tr>
<tr>
<td>300</td>
<td>30,000,000</td>
<td>4.5 GB</td>
</tr>
<tr>
<td>1,000</td>
<td>100,000,000</td>
<td>15.0 GB</td>
</tr>
<tr>
<td>3,000</td>
<td>300,000,000</td>
<td>45.0 GB</td>
</tr>
</tbody>
</table>
Power9 14 cores, 64GB
RHEL 7.5 4.14

PG Bench Injector
PG Bench Injector
PG Bench Injector

Postgres 10.5 & 11 beta4
NVMe

16GB/s FC
EMC Unity 300 (Flash)

pg_bench: scale 1000 = ~15 GB
Select (read-only)
~700MB/s

8Gb FC Saturated
Transactions per second

# Clients

Power9: 14 cores 40 GB RAM
## PostgreSQL 9.4.4 Results

<table>
<thead>
<tr>
<th>Scale</th>
<th>TPS (incl. conn)</th>
<th>TPS (excl. conn)</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>423,148</td>
<td>446,022</td>
</tr>
<tr>
<td>300</td>
<td>379,866</td>
<td>709,271</td>
</tr>
<tr>
<td>1000</td>
<td>406,589</td>
<td>435,311</td>
</tr>
<tr>
<td>3000</td>
<td>377,233</td>
<td>566,129</td>
</tr>
</tbody>
</table>

Sébastien Chabrolles IBM Montpellier + Splendid Data

Power8: 20 cores 256 GB RAM
Transactions per second

<table>
<thead>
<tr>
<th># Clients</th>
<th>Transactions per second</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>26K</td>
</tr>
<tr>
<td>32</td>
<td>30K</td>
</tr>
<tr>
<td>48</td>
<td>34K</td>
</tr>
<tr>
<td>64</td>
<td>38K</td>
</tr>
<tr>
<td>80</td>
<td>42K</td>
</tr>
<tr>
<td>96</td>
<td>46K</td>
</tr>
<tr>
<td>112</td>
<td>50K</td>
</tr>
<tr>
<td>128</td>
<td>54K</td>
</tr>
<tr>
<td>144</td>
<td></td>
</tr>
<tr>
<td>160</td>
<td></td>
</tr>
</tbody>
</table>

14 cores P9@3.3GHz SMT8, 40GB RAM, shared_buffers=24GB

EXT4 nobarrier pgbench -j8

Read-Write - scale 1,000
SMT4 vs SMT8 on POWER9

E4-900: 14 cores P9@3.3GHz 40GB RAM
PostgreSQL11beta4 pgbench -S with shared_buffers=8GB

Graph showing the comparison between SMT4 and SMT8 in terms of transactions per second for different numbers of clients. The graph indicates that SMT8 generally outperforms SMT4, with SMT8 reaching higher transaction rates at higher client counts.
Leading performance and cost optimized for Big Data and Analytics workloads

Standardize on Linux

Escala supports industry standard Linux distributions from RedHat, Suse and Ubuntu along with applications optimized for emerging business challenges such as Big Data & analytics, mobile and Java platforms. The support for little endian Linux ensures easy portability for thousands of Open Source applications. Leading ISV's such as SAP also support the Linux on Power platform.

OpenStack Cloud enabling / automation

The PowerVC management solution turns Escala VMs, storage and virtual networks into a OpenStack based IAAS cloud solution. PowerVC groups multiple physical servers into managed pools, greatly simplifying the creation and deployment of AIX and Linux VMs. Features like Snapshots and cloning reduce the average deployment time of applications from several hours to just a few minutes. Automation can be integrated or triggered by popular Open Source tools such as Chef or Puppet.

The constant economic pressure to reduce IT infrastructure costs while increasing flexibility through dynamic, on-demand provisioning has driven many organizations towards standardization onto x86 hardware and the Linux operating system. As price/performance has become the key measure for infrastructure purchases, TCO calculations are mainly conducted in isolated domains such as hardware, OS/virtualization and support costs. However, bigger picture aspects such as the cost impact on ISV licenses, platform reliability, OS vulnerability and VM isolation are often not factored in.

Escala delivers key advantages which can have a significant overall TCO impact such as a leading performance per core, the best reliability in the industry, superior security isolation and virtualization efficiency.

Escala Mission Critical Linux servers

Escala Mission Critical Linux servers are designed for commercial workloads requiring performance, reliability and scalability. The Escala L1-800 and L3-800 servers use the latest Power8 processor technology delivering unprecedented performance, scalability, reliability, and manageability, for demanding commercial workloads. They are optimized to help deliver new solutions and services faster, and with higher quality. They are also optimized for big data and analytics, and provide the ideal foundation for scale-out data and cloud environments in a compact 2U package. They deliver superior throughput compared to x86-based offerings for similar workloads.

Escala OpenPower-based Linux servers

Escala OpenPower-based Linux servers are optimized for Big Data and cloud workloads. The Escala L1-OP80 and L3-OP80 Power8 servers are built for most OpenSource workloads, as well as OpenSource databases, due to their cost efficiency and expandability. They are designed to run either Hadoop environments thanks to larger internal storage capacity and powerful SMT8 cores, or in-memory analytics such as Spark, as well as innovative solutions from the OpenPower partner ecosystem such as CAPI. They are also ideal for hyperscale ISP environments where high efficiency and cost-effectiveness are the most important features.

Atos and Linux advantages:

• Market leading memory bandwidth
• Slim packaging with high internal disk expandability
• Ideal for Big Data as a Hadoop node thanks to internal disk expandability and powerful SMT8 cores
• Ideal for Spark, OpenSource databases (PostgresSQL, MariaDB..) and Business applications (SAP)

discover the open alternative: Escala Linux solutions

escala mission critical servers

splendid data
GRIP ON POSTGRES
This is the sign you've been looking for.