

# Greenplum 集群安装配置及最佳实践

## 目录

Greenplum 集群安装配置及最佳实践.....	1
目录.....	1
1 总体介绍.....	2
1.1 硬件平衡.....	2
1.2 高可用.....	2
1.3 部署方案.....	2
1.3.1 Group Mirroring 部署方案.....	3
1.3.2 Spread Mirroring 部署方案.....	3
1.3.3 Group + Spread Mirroring 部署方案.....	4
2 硬件选型.....	4
2.1 Master 节点.....	4
2.1.1 网卡.....	4
2.1.2 内存.....	4
2.1.3 磁盘.....	4
2.1.4 CPU.....	5
2.2 Segment 节点.....	5
2.2.1 网卡.....	5
2.2.2 内存.....	5
2.2.3 磁盘.....	5
2.2.4 CPU.....	5
2.3 节点配置实例.....	6
2.4 硬件配置经验总结.....	6
3 硬件配置.....	7
1 主机配置实例.....	7
2 磁盘的配置.....	7
3 网络配置.....	8
4 交换机配置.....	8
4 储存规划.....	9
4.1 磁盘可用空间.....	9
4.2 用户数据容量.....	9
4.3 系统数据容量.....	10
5 集群的配置.....	10
5.1 最低系统要求.....	10
5.2 系统设置.....	11
5.3 操作系统参数设置.....	11
5.4 I/O 设置.....	12

5.5 其他设置.....	13
6 集群安装.....	13
6.1 安装方式.....	13
6.2 数据库目录.....	14
7 集群验证.....	14
7.1 硬件性能验证.....	14
7.2 集群初始化.....	15
7.3 配置文件.....	15
7.4 数据目录.....	16
7.5 pg_hba.conf 配置文件.....	16
7.6 安装完后的环境变量.....	17
8 可选组件安装.....	18
8.1 安装外部支持的语言.....	18
8.2 故障诊断.....	19

# 1 总体介绍

## 1.1 硬件平衡

- 1、性能
- 2、容量
- 3、成本

## 1.2 高可用

- 1、节点
- 2、网络
- 3、磁盘

## 1.3 部署方案

- 1、Master 与 Standby Master 分级部署
- 2、Primary Segment 与 Mirror Segment 分机部署
- 3、Segment Mirroring 部署方案
  - 3.1、Group Mirroring
  - 3.2、Spread Mirroring
  - 3.3、Group + Spread Mirroring
- 4、Pivotal Supported Greenplum 必须部署 Mirroring Segment
- 5、铜一主机 Segment 个数

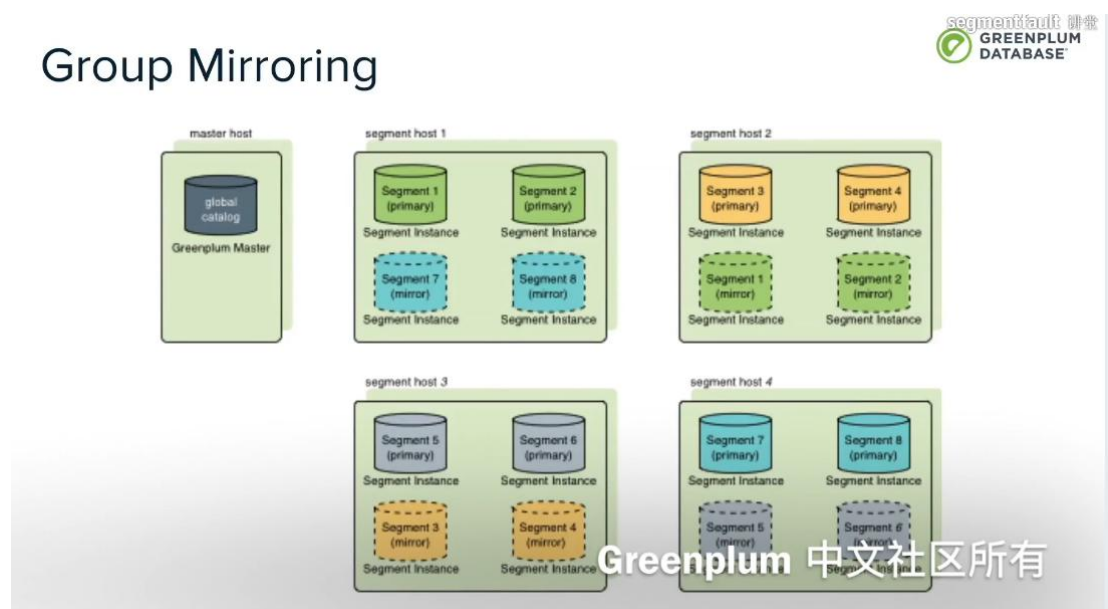
- 5.1 CPU/Core 数据
- 5.2 查询并发数
- 5.3 查询复杂度
- 5.3 单机 Primary Segment 总数不能超过

### 1.3.1 Group Mirroring 部署方案

按照以下 4 台机器 Group Mirroring 的部署方案总结

缺点: 一台机器 down 掉后, 会把流量全部放在下一个节点, 下一个节点的流量会变成 2 倍的流量

优点: down 掉一台机器后, 集群能正常的提供服务, 如果再 down 掉第二台集群就不可用



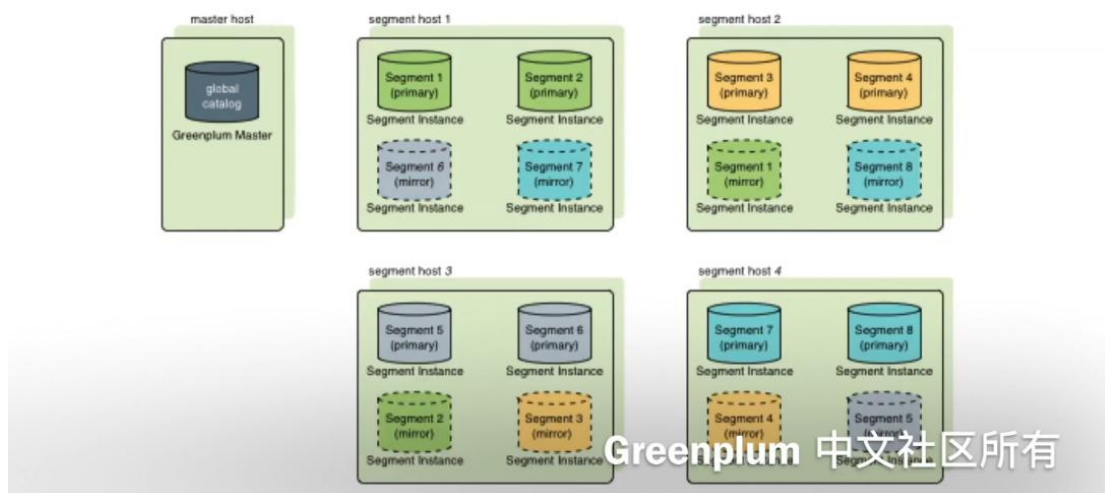
### 1.3.2 Spread Mirroring 部署方案

按照以下 4 台机器 Spread Mirroring 的部署方案总结

缺点: 一台机器 down 掉后, 会把流量全部放在下两个节点

优点: down 掉一台机器后, 集群能正常的提供服务, 如果再 down 掉第二台集群就不可用

# Spread Mirroring



## 1.3.3 Group + Spread Mirroring 部署方案

如果集群比较大建议使用 Group + Spread Mirroring 部署方案，如果集群由 down 流量会分流道其他的机器上,集群不可用的几率比较小。

## 2 硬件选型

### 2.1 Master 节点

#### 2.1.1 网卡

- 1、2 块万兆网卡内部互联
- 2、1-2 块千兆网卡带外管理及接入客户网络

#### 2.1.2 内存

DDR4 64GB 以上，建议 256G

#### 2.1.3 磁盘

- 1、6 块 600G/900G 10k RPM SAS 盘

- 2、采用 RAID5 或 RAID10
- 3、单独预留 hotspare 盘
- 4、1 块 RAID 卡，cache 1GB 以上，带有掉电保护功能

## **2.1.4 CPU**

- 1、2 路 8 核及以上
- 2、主频 2.5G HZ 以上

## **2.2 Segment 节点**

### **2.2.1 网卡**

- 1、2 块万兆网卡内部互联
- 2、1-2 块千兆网卡带外管理及接入客户网络

### **2.2.2 内存**

DDR4 64GB 以上，建议 256G

### **2.2.3 磁盘**

- 1、24 块 600G/900G 10k RPM SAS 盘
- 2、采用 RAID5 或 RAID10
- 3、单独预留 hotspare 盘
- 4、1-2 块 RAID 卡，cache 1GB 以上，带有掉电保护功能

### **2.2.4 CPU**

- 1、2 路 8 核及以上
- 2、主频 2.5G HZ 以上

## 2.3 节点配置实例

### 节点配置示例



Component	Masters	Segment Servers
Processor	2 E5-2680v3	2 E5-2680v3
Memory	256GB of DDR4 @ 2133MHz (8x32GB DIMMs)	256GB of DDR4 @ 2133MHz (8x32GB DIMMs),
Size	1U	2U
Sockets	Dual Socket	Dual Socket
Cores	24	24
Network	Dual (one internal, one for customer network)	Single 10G Dual Port NIC for internal connection
Controller	12Gb x8 SAS controller (LSI base, Intel branded)	1 x 12Gb x8 SAS controller (LSI based, Intel branded) w/ 12Gb SAS expander
Disks	(6) 1.8Tb 10k RPM SAS	(24) 1.8Tb 10k RPM SAS

## 2.4 硬件配置经验总结

- 1、磁盘故障时 Greenplum 集群最常见的故障
  - 1.1 分析性查询: SAS 盘 > SATA 盘
  - 1.2 高并发小 IO 查询: 优先 SSD 或 NVMe
- 2、RAID 级别
  - 2.1 RAID-5 VS RAID-10
- 3、RAID 卡一定带 Cache 功能，能提高磁盘的读写性能。
- 4、硬件监控
- 5、预留灾备机

RAID-5 VS RAID-10 的区别:

RAID-5:在容量上会比较大，储存会达到 90%的利用率，在读写的性能上比 RAID10 会好一点,RAID5 的可靠性会差很多。

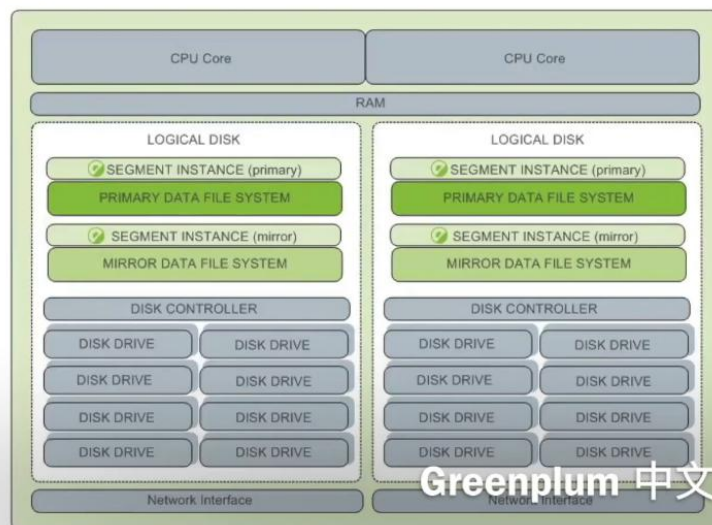
RAID-10:冗余级别更高，储存会达到 50%的利用率，当有的盘出现坏的情况下，磁盘的可靠性有保证。

## 3 硬件配置

### 1 主机配置实例

一下的配置两个 CPU 主机的例子,两个 segment 两个 primary 两个 mirror 以及两个网卡的例子

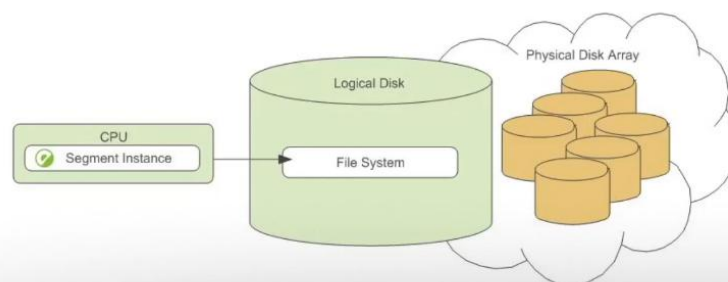
#### 主机配置



Greenplum 中文社区所有

### 2 磁盘的配置

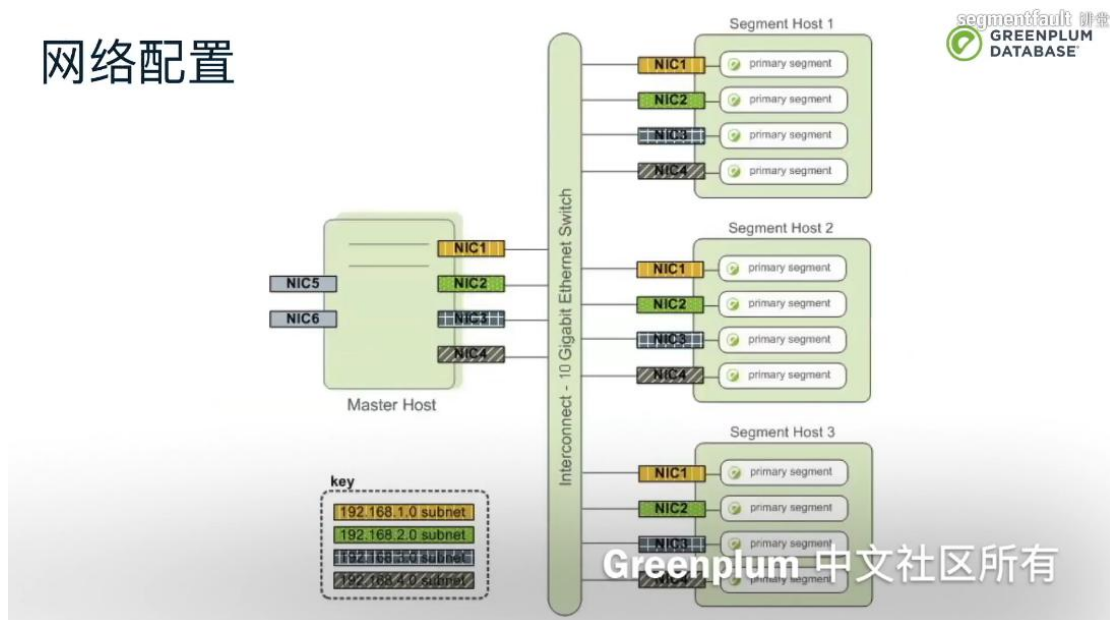
#### 磁盘配置



Greenplum 中文社区所有

### 3 网络配置

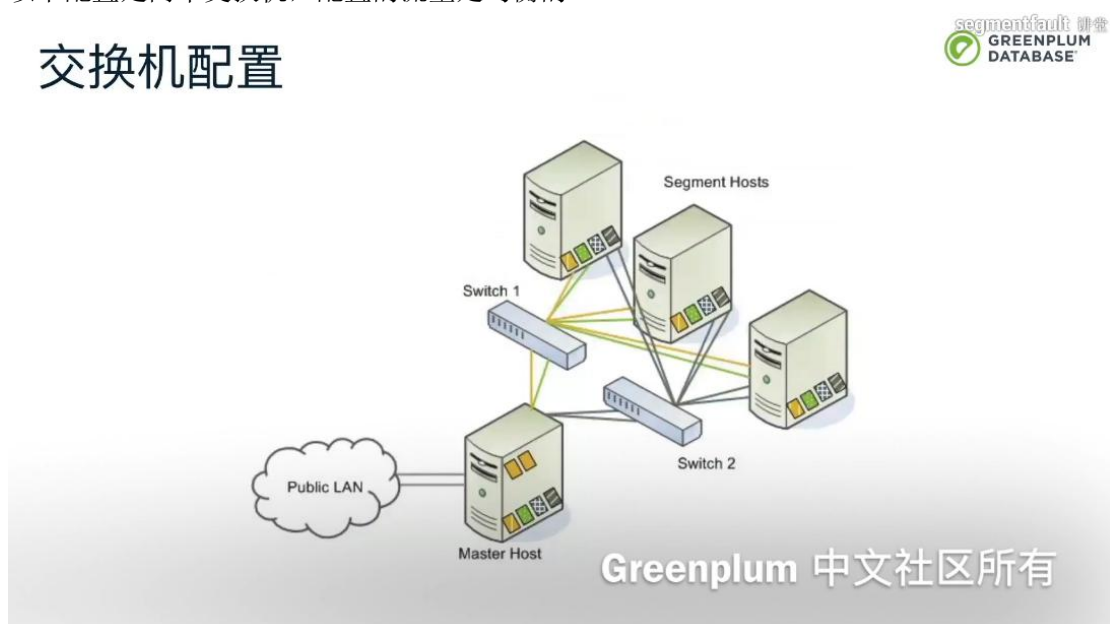
#### 网络配置



### 4 交换机配置

以下配置是两个交换机，配置的流量是均衡的。

#### 交换机配置





## 4 储存规划

### 4.1 磁盘可用空间

#### 磁盘可用空间



- 原始磁盘容量
  - $\text{raw\_capacity} = \text{disk\_capacity} * \text{number\_of\_disks}$
- 文件系统开销
  - 文件系统格式 (10%)
  - RAID级别
    - RAID-5 10%
    - RAID-10 50%
  - $\text{formatted\_disk\_space} = (\text{raw\_capacity} * 0.9) / 2$
- 性能因素
  - 磁盘容量保持在70%最佳
  - $\text{usable\_disk\_space} = \text{formatted\_disk\_space} * 0.7$
- 用户数据空间及临时空间
  - 用户数据容量U, 临时空间容量U/3
  - With Mirror:  $\text{usable\_disk\_space} = (2 * U) + U / 3$
  - Without Mirror:  $\text{usable\_disk\_space} = U + U / 3$

Greenplum 中文社区所有

### 4.2 用户数据容量

#### 用户数据容量



- 数据库用户数据 = 原始用户数据 \* 1.4
- Page开销
  - 每32KB的Page需要20字节Header开销
- Tuple开销
  - 每个Heap Tuple需要24字节Header开销
  - 每个AO Tuple需要4字节Header开销
- Attribute开销
  - 类型有关
- Index开销
  - BTree索引与唯一值数目有关

Greenplum 中文社区所有

## 4.3 系统数据容量

### 系统数据容量



- Catalog
  - 每个Segment约300MB
- Write Ahead Log
  - WAL分为多个64MB的段文件
  - 段文件数目最多为  $2 * \text{checkpoint\_segment} + 1$  (默认checkpoint\_segment为8)
  - 每个Segment上WAL最多  $(2 * 8 + 1) * 64\text{MB} = 1088\text{MB}$
- Database Log文件
  - Log Rotation机制

Greenplum 中文社区所有

## 5 集群的配置

### 5.1 最低系统要求

以下是系统的最低的配置

#### 系统要求



操作系统	<ul style="list-style-type: none"><li>• SUSE Linux Enterprise Server 64-bit 11 SP4, 12 SP2 and SP3 with kernel version greater than 4.4.73-5.</li><li>• CentOS 64-bit 6.x or 7.x</li><li>• Red Hat Enterprise Linux (RHEL) 64-bit 6.x or 7.x</li><li>• Oracle Linux 64-bit 7.4, using the Red Hat Compatible Kernel (RHCK)</li></ul>
文件系统	<ul style="list-style-type: none"><li>• SUSE Linux和RedHat上XFS用于数据存储</li><li>• EXT3用于root文件系统, EXT3生产环境中比XFS慢60%</li></ul>
CPU最低要求	<ul style="list-style-type: none"><li>• Pentium Pro compatible (P3/Athlon及以上)</li></ul>
内存最低要求	<ul style="list-style-type: none"><li>• 最低: 1 GB RAM per CPU core / 16 GB RAM per Server</li><li>• 推荐: 8-16 GB RAM per CPU core</li></ul>
磁盘	<ul style="list-style-type: none"><li>• 安装空间需要150 MB</li><li>• 推荐高速本地存储</li></ul>
网络	<ul style="list-style-type: none"><li>• 万兆网卡</li><li>• non-blocking配置交换机</li><li>• 多网卡绑定bonding</li></ul>
系统工具	<ul style="list-style-type: none"><li>• zlib, bash shell, perl, security shell</li><li>• GNU tars, GNU gzip, GNU sed</li></ul>

Greenplum 中文社区所有

## 5.2 系统设置

### 系统设置



- 必须禁用SELinux
  - 检查SELinux状态

```
# sestatus
SELinuxstatus: disabled
```

- /etc/selinux/config文件中配置SELinux

```
SELINUX=disabled
```

- 重启系统，配置生效

- 建议禁用Firewall
  - RHEL6.X和CentOS 6.X: iptables

```
# /sbin/chkconfig --list iptables
iptables 0:off 1:off 2:off 3:off
4:off 5:off 6:off
# /sbin/chkconfig iptables off
```

- RHEL 7.X和CentOS 7.X: firewalld

```
# systemctl status firewalld
* firewalld.service - firewalld -
dynamic firewall daemon Loaded:loaded
(/usr/lib/systemd/system/firewalld.se
rvice; disabled; vendor preset:
enabled)Active: inactive (dead)
# systemctl stop firewalld.service
# systemctl disable firewalld.service
```

Greenplum 中文社区所有

## 5.3 操作系统参数设置

### 操作系统参数设置



- /etc/hosts
  - 配置Greenplum集群中所有主机名及网络地址
- /etc/security/limits.conf
  - 配置用户Limit参数

```
* soft nfile 524288
* hard nfile 524288
* soft nproc 131072
* hard nproc 131072
```

- /etc/sysctl.conf
  - 配置Linux系统参数，sysctl -p生效

```
kernel.shmmax = 500000000
kernel.shmmni = 4096
kernel.shmall = 4000000000
kernel.sem = 500 2048000 200 40960
kernel.sysrq = 1
kernel.core_uses_pid = 1
kernel.msgmnb = 65536
kernel.msgmax = 65536
kernel.msgmni = 2048
net.ipv4.tcp_syncookies = 1
net.ipv4.conf.default.accept_source_route = 0
net.ipv4.tcp_max_syn_backlog = 4096
net.ipv4.conf.all.arp_filter = 1
net.ipv4.ip_local_port_range = 10000 65535
net.core.netdev_max_backlog = 10000
net.core.rmem_max = 2097152
net.core.wmem_max = 2097152
vm.overcommit_memory = 2
vm.swappiness = 10
vm.zone_reclaim_mode = 0
vm.dirty_expire_centisecs = 500
vm.dirty_writeback_centisecs = 100
vm.dirty_background_bytes = 1610612736
vm.dirty_bytes = 4294967296
```

Greenplum 中文社区所有

## 5.4 I/O 设置

### I/O设置



- XFS挂载选项

```
rw,nodev,noatime,nobarrier,inode64
```

- 设置blockdev预读尺寸

```
# /sbin/blockdev --setra 16384 /dev/sdb
```

- 设置磁盘I/O调度器为deadline

```
# echo deadline > /sys/block/sdb/queue/scheduler
```

Greenplum 中文社区所有

### I/O设置



- 禁用Transparent Huge Page (THP)
  - RHEL 6.X配置/boot/grub/grub.conf

```
kernel /vmlinuz-2.6.18-274.3.1.el5 ro root=LABEL=/  
elevator=deadline crashkernel=128M@16M quiet console=tty1  
console=ttyS1,115200 panic=30 transparent_hugepage=never  
initrd /initrd-2.6.18-274.3.1.el5.img
```

- RHEL 7.X运行grub2命令

```
# grubby --update-kernel=ALL --args="transparent_hugepage=never"
```

- /etc/systemd/logind.conf中禁用RemoveIPC

```
RemoveIPC=no  
# service systemd-logind restart
```

Greenplum 中文社区所有

## 5.5 其他设置

### 其他设置



- 数据库管理员账户
  - gpadmin
- 系统时钟同步
  - NTP

Greenplum 中文社区所有

## 6 集群安装

### 6.1 安装方式

#### 安装方式



- RPM安装
  - 所有节点上安装RPM
  - 用户创建gpadmin
  - 用户设置节点间无密码SSH访问

- Binary安装
  - 只在Master上安装，然后运行gpsegininstall
  - gpsegininstall负责
    - 集群间拷贝Binary
    - 创建gpadmin
    - 设置节点间无密码SSH访问

```
$ unzip greenplum-db-<version>-<platform>.zip
$ bin/bash greenplum-db-<version>-<platform>.bin
$ source /usr/local/greenplum-db/greenplum_path.sh
$ gpsegininstall -f hostfile_all
```

Greenplum 中文社区所有

## 6.2 数据库目录

### 数据库目录



bin	postgres程序，数据库管理工具等
sbin	内部工具
lib	库文件，Extension等
include	头文件
share	内部共享文件
ext	内置程序，如Python
docs	命令行工具帮助文件，集群配置文件示例
greenplum_path.sh	环境变量文件

Greenplum 中文社区所有

## 7 集群验证

### 7.1 硬件性能验证

#### 硬件性能验证



- gpcheckperf
  - 网络性能
  - 磁盘IO
  - 内存带宽

```
$ gpcheckperf -f hostfile_gpchecknet_icl -r N -d /tmp > subnet1.out
```

```
$ gpcheckperf -f hostfile_gpcheckperf -r ds -D -d /data/primary -d /data/mirror
```

Greenplum 中文社区所有

## 7.2 集群初始化

### 集群初始化



- 创建数据目录

```
$ mkdir /data/master      # Master节点和Standby Master节点上创建数据目录
$ mkdir /data/primary     # 所有Primary Segment节点上创建数据目录
$ mkdir /data/mirror      # 所有Mirror Segment节点上创建数据目录
```

- 集群初始化命令

```
$ gpinitssystem -c gpinitssystem_config      # 集群初始化配置文件
               -h hostfile_gpinitssystem    # 集群所有segment地址名
               -s standby_master_hostname    # 初始化Standby Master (可选)
               -S                             # Mirror配置为Spread (可选, 默认Group)
```

- 集群初始化日志位置
  - \$HOME/gpAdminLogs
- 集群初始化撤销脚本
  - \$HOME/gpAdminLogs/backout\_gpinitssystem.sh

Greenplum 中文社区所有

## 7.3 配置文件

### 配置文件



```
#### Cluster
ARRAY_NAME="Greenplum Data Platform"
SEG_PREFIX=gpseg
TRUSTED_SHELL=ssh
CHECK_POINT_SEGMENTS=8
ENCODING=UNICODE

#### Master
MASTER_HOSTNAME=mdw
MASTER_DIRECTORY=/data/master
MASTER_PORT=5432

#### Primary Segment
PORT_BASE=6000
declare -a DATA_DIRECTORY=(/data1/primary /data1/primary /data1/primary /data2/primary /data2/primary /data2/primary)

#### Mirror Segment
MIRROR_PORT_BASE=7000
declare -a MIRROR_DATA_DIRECTORY=(/data1/mirror /data1/mirror /data1/mirror /data2/mirror /data2/mirror /data2/mirror)
```

Greenplum 中文社区所有



## 7.4 数据目录

### 数据目录



base	数据库数据文件
global	全局数据表文件，如pg_database
pg_log	数据库查询日志文件
pg_xlog	数据库WAL日志文件
pg_clog	事务提交信息日志文件
pg_tblspc	数据库所有tablespace的符号链接
postgresql.conf	数据库主配置文件
postmaster.opts	数据库postmaster进程启动命令
pg_hba.conf	客户端连接认证配置

Greenplum 中文社区所有

## 7.5 pg\_hba.conf 配置文件



### pg\_hba.conf

数据库名 & 用户名

- 逗号分隔多个数据库名或用户名
- all用于所有数据库或所有用户名

网络地址

- 仅用于host, hostssl, hostnossl连接类型
- 0.0.0.0/0用于Full Network

local	database	user	CIDR-address	auth-method	[auth-options]
host	database	user	CIDR-address	auth-method	[auth-options]
hostssl	database	user	CIDR-address	auth-method	[auth-options]
hostnossl	database	user	CIDR-address	auth-method	[auth-options]
host	database	user	IP-address IP-mask	auth-method	[auth-options]
hostssl	database	user	IP-address IP-mask	auth-method	[auth-options]
hostnossl	database	user	IP-address IP-mask	auth-method	[auth-options]

连接类型

- local: Unix Domain Socket连接
- host: Network连接, 连接可选加密
- hostssl: Network连接, 连接必须加密
- hostnossl: Network连接, 连接无需加密
- .....

认证方式

- trust: 无需密码, 认证通过
- reject: 认证拒绝
- md5: 认证需要密码, 密码用md5取值
- password: 认证需要密码, 密码用明文
- ldap: ldap认证
- gss: GSSAPI和Kerberos认证
- cert: SSL证书认证

```
host all all 0.0.0.0/0 trust # 任意用户从任何主机可自由访问所有数据库
```

Greenplum 中文社区所有



## pg\_hba.conf

- 处理客户端连接认证
- 采用First Matching策略
- 默认客户端连接策略为Deny
- 策略更改后需要重启集群或者gpstop -u重新加载

Greenplum 中文社区所有

## 7.6 安装完后的环境变量

### 环境变量

变量名	变量值示例	变量含义
GPHOME	/usr/local/greenplum-db	Greenplum安装主目录
MASTER_DATA_DIRECTORY	/data/master/gpseg-1	Master节点数据目录
PGHOST	mdw	Master节点主机地址
PGPORT	5432	Master数据库主进程端口号
PGDATABASE	postgres	默认数据库
PGUSER	gpadmin	默认数据库用户名
PGPASSWORD	*****	用户对应密码
PGOPTIONS	'-c gp_session_role=utility'	数据库连接选项

Greenplum 中文社区所有

## 8 可选组件安装

### 8.1 安装外部支持的语言

#### Extension

- Procedural Language
  - PL/Java
  - PL/Python
  - PL/R
  - PL/Perl
- Machine Learning
  - MADlib
- Geospatial
  - PostGIS



Greenplum 中文社区所有

### 8.2 使用 gppkg 命令安装

#### gppkg

- Greenplum包管理器
  - 集群间安装Greenplum Extension及其依赖库
  - 支持集群扩容
  - 集群升级后需要重新下载及安装Extension
- 安装Extension示例

```
$ gppkg -i ./madlib-1.16-gp5-rhel7-x86_64/madlib-1.16-gp5-rhel7-x86_64.gppkg
```



Greenplum 中文社区所有

## 8.2 故障诊断

### 常见故障



- 集群初始化失败
  - 检查\$HOME/gpAdminLogs
  - 系统参数设置不正确或没有重新加载生效
  - 检查数据目录不存在，不为空，权限不正确
- 远程客户端无法连接数据库
  - 检查pg\_hba.conf设置
  - 检查postgresql.conf设置
- 查询性能问题
  - 使用gpcheckperf逐个排查网卡，磁盘，内存等硬件问题

Greenplum 中文社区所有