gratka technologie

www.gratka-technologie.pl

gratka technologie

Wdrożenie skalowalnego systemu plików **GlusterFS** w serwisach Gratka.pl

Łukasz Jagiełło
l.jagiello@gratka-technologie.pl

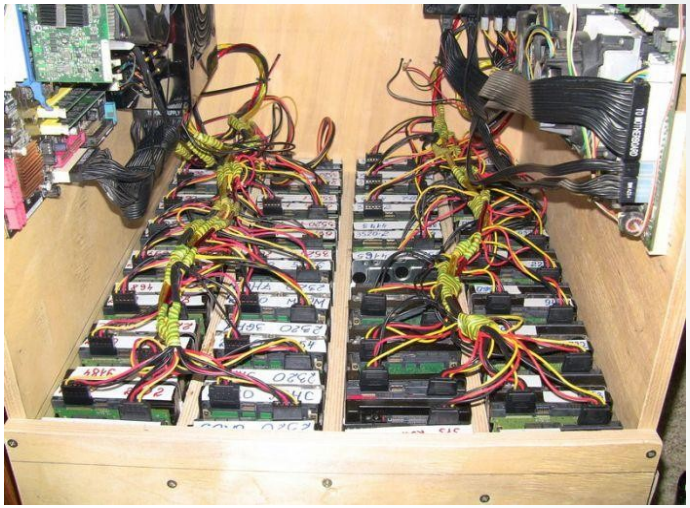
Po co nam storage?

gratka technologie

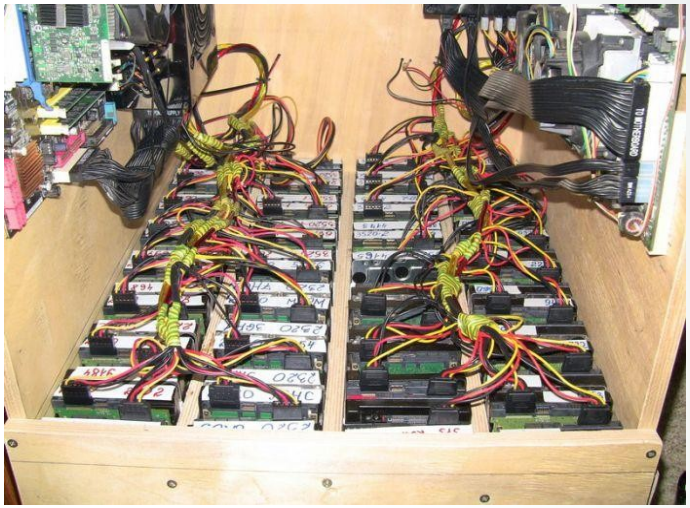
Po co nam storage?

Co mamy do dyspozycji?

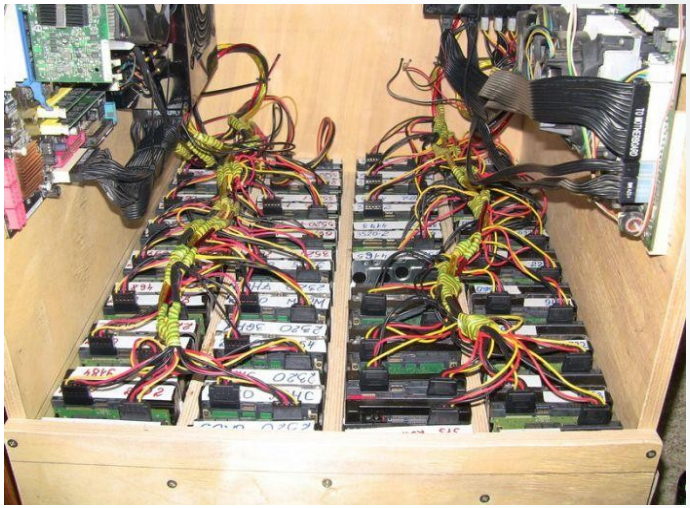
Co mamy do dyspozycji?

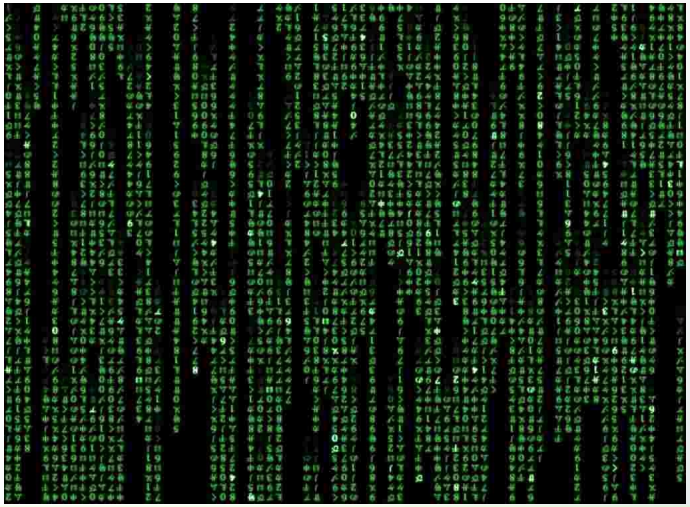


Co mamy do dyspozycji?



Co mamy do dyspozycji?





gratka technologie

Dlaczego wielu ludzi kupuje macierze, chociaż ich nie potrzebuje?


ORACLE
e
EVIL EMPIRE

No to może software?

Przegląd rynku:

- DRBD
- ZFS (w różnych wydaniach)
- Hadoop
- Ceph
- GridFS
- Lustre
- **GlusterFS**
- i wiele innych...

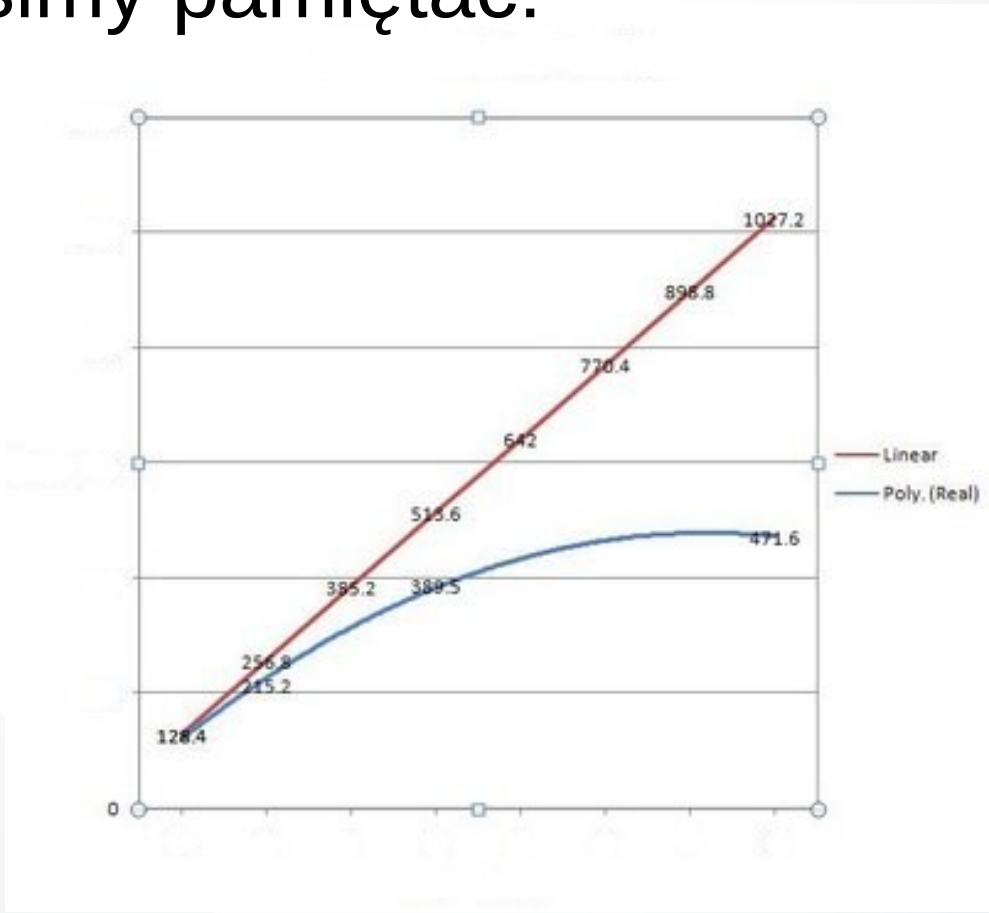
Co to jest Gluster

- open source software (GPLv3)
- clustered file system
- scal-out (several petabytes)
- working at user space
- high-performance
- Infiniband RDMA or TCP/IP

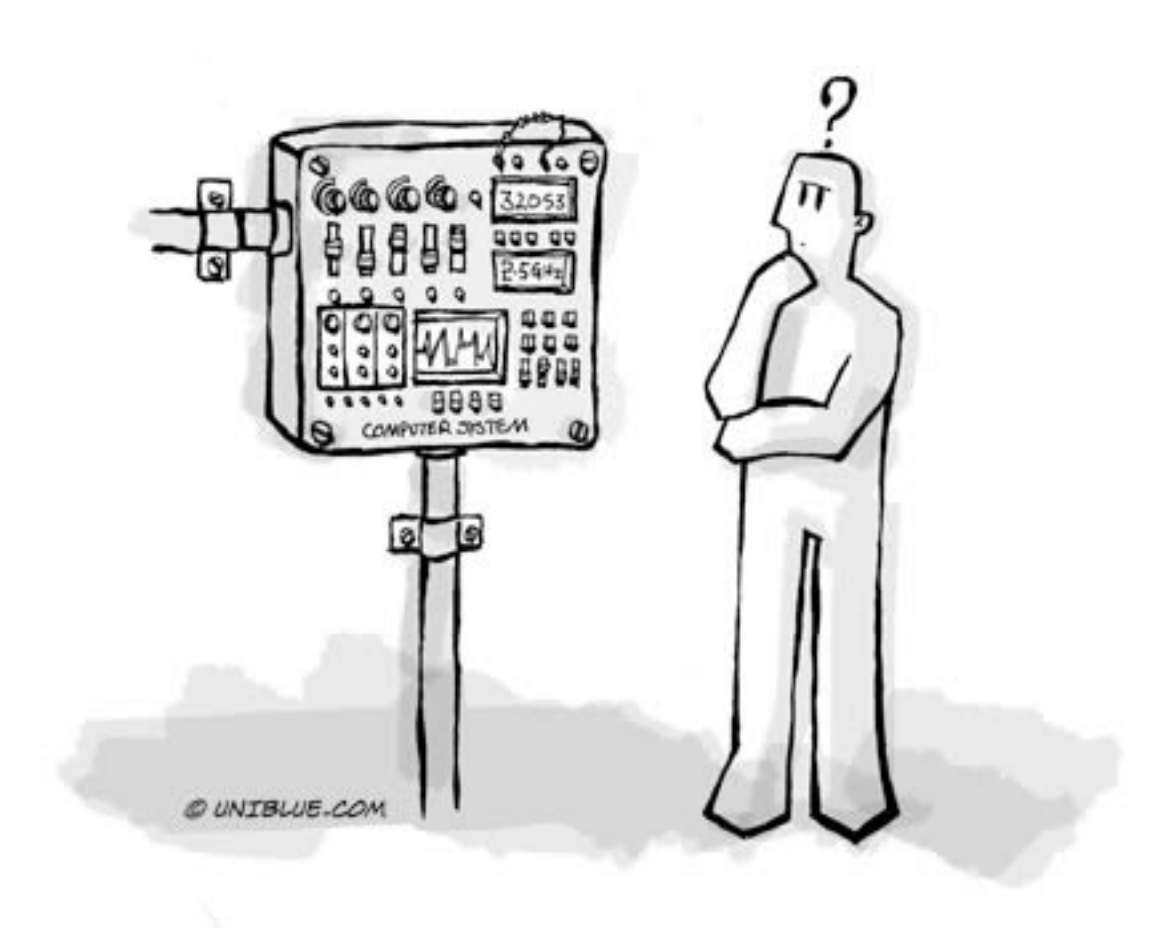
Liniowa Skalowalność?

O czym musimy pamiętać:

- HDD
- system
- CPU
- filesystem
- metadata
- network



Jak to działa?



Jak to działa...

1. Eliminacja synchronizacji i aktualizacji metadanych

Jak to działa...

1. Eliminacja synchronizacji i aktualizacji metadanych
2. Efektywna dystrybucja danych w celu zapewnienia skalowalności i niezawodności

Jak to działa...

1. Eliminacja synchronizacji i aktualizacji metadanych
2. Efektywna dystrybucja danych w celu zapewnienia skalowalności i niezawodności
3. Stosowanie dostępu równoległego w celu zmaksymalizowania wydajności

Jak to brak metadanych?

- Wszystkie dane na zwykłych systemach plików (np. Ext3/4, ReiserFS, ZFS, itd.)

Jak to brak metadanych?

- Wszystkie dane na zwykłych systemach plików (np. Ext3/4, ReiserFS, ZFS, itd.)
- Dzielenie plików z użyciem „split”

Jak to brak metadanych?

- Wszystkie dane na zwykłych systemach plików (np. Ext3/4, ReiserFS, ZFS, itd.)
- Dzielenie plików z użyciem „split”
- Mirror w trybie „active-active”

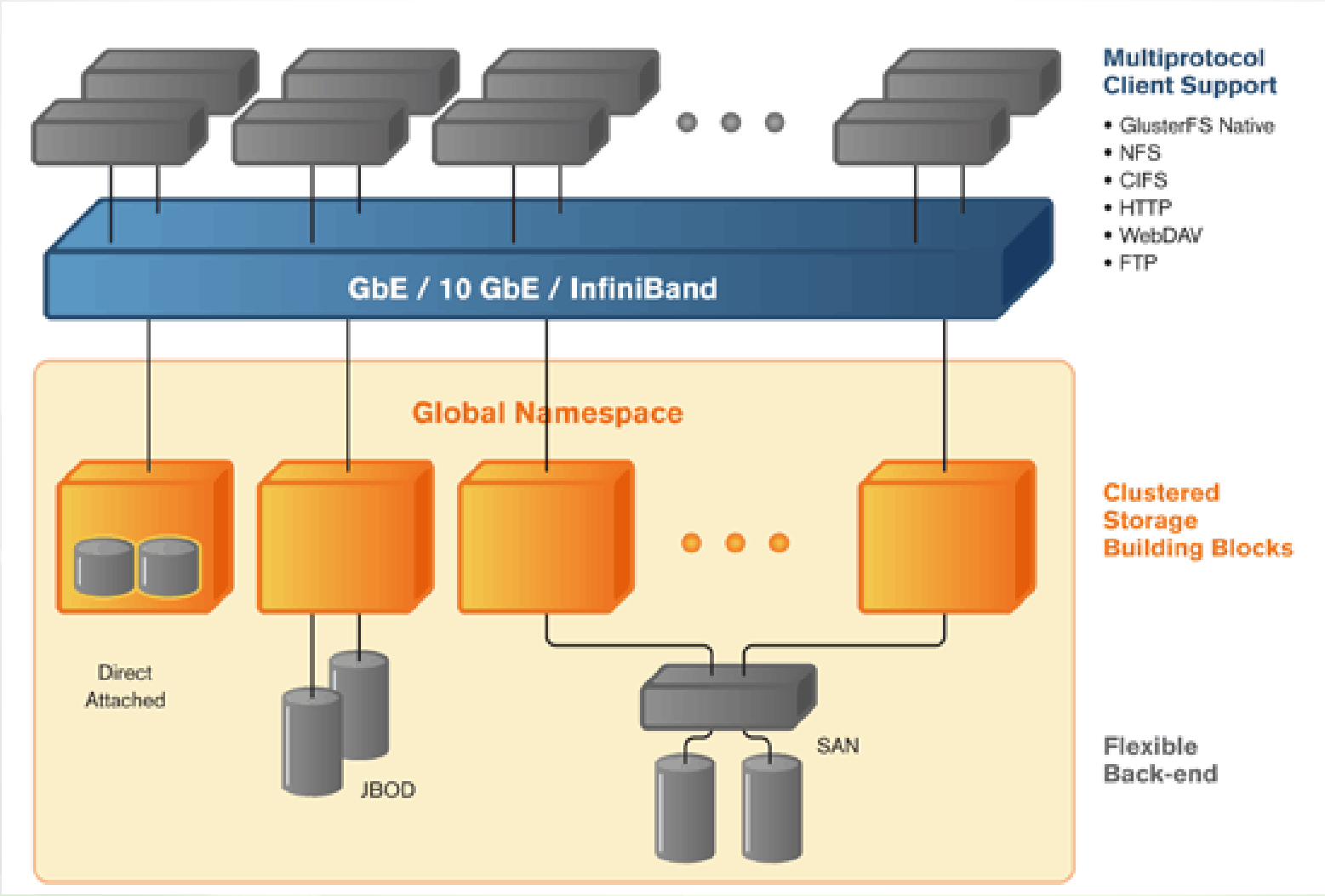
Jak to brak metadanych?

- Wszystkie dane na zwykłych systemach plików (np. Ext3/4, ReiserFS, ZFS, itd.)
- Dzielenie plików z użyciem „split”
- Mirror w trybie „active-active”
- Lokalizowanie/rozrzucanie plików z użyciem podrasowanego algorytmu Daviesa-Meyera*****

Dostęp do danych:

- **GlusterFS Native**
- NFS
- CIFS
- WebDAV
- FTP

Schemat działania



Trochę praktyki...

Tryby działania:

- Distributed
- Distributed Replicated
- Distributed Striped

Banalna instalacja

```
rpm -Uhv glusterfs-core-3.1.0-1 glusterfs-fuse-3.1.0-1
chkconfig glusterd on
service glusterd start


iptables -A RH-Firewall-1-INPUT -m state --state NEW -m tcp -p tcp --dport 24007 -j ACCEPT
iptables -A RH-Firewall-1-INPUT -m state --state NEW -m tcp -p tcp --dport 38465:38485 -j ACCEPT
service iptables save
service iptables restart
```

Distributed Replicated

```
gluster> volume create test replica 2 transport tcp 172.17.100.1:/d0 172.17.100.2:/d0
Creation of volume test has been successful. Please start the volume to access data.
gluster> volume start test
Starting volume test has been successful
gluster> volume info

Volume Name: test
Type: Replicate
Status: Started
Number of Bricks: 2
Transport-type: tcp
Bricks:
Brick1: 172.17.100.1:/d0
Brick2: 172.17.100.2:/d0
```

Distributed Replicated (2)

```
[root@node1 ~]# df -h
System plików          rozm. użyte dost. %uż. zamont. na
/dev/mapper/VolGroup  6,5G  932M  5,3G  15% /
tmpfs                 247M     0  247M   0% /dev/shm
/dev/sda1             485M   27M  433M   6% /boot
```

```
[root@p0x ~]# mount -t glusterfs 172.17.100.1:/test /media/cos2/
[root@p0x ~]# df -h
System plików          rozm. użyte dost. %uż. zamont. na
/dev/sda1             228G  106G  111G  49% /
tmpfs                 2,0G   15M  2,0G   1% /dev/shm
172.17.100.1:/test    6,5G  932M  5,3G  15% /media/cos2
```

Distributed Replicated (3)

```
gluster> volume info
Volume Name: test
Type: Replicate
Status: Started
Number of Bricks: 2
Transport-type: tcp
Bricks:
Brick1: 172.17.100.1:/d0
Brick2: 172.17.100.2:/d0

gluster> volume replace-brick test 172.17.100.1:/d0 172.17.100.3:/d0 start
replace-brick started successfully

gluster> volume replace-brick test 172.17.100.1:/d0 172.17.100.3:/d0 commit
replace-brick commit successful
```

Co jeszcze ?

```
volume info [all|<VOLNAME>] - list information of all volumes

volume create <NEW-VOLNAME> [stripe <COUNT>] [replica <COUNT>] [transport <tcp|rdma>] <NEW-BRICK>

volume delete <VOLNAME>


volume start|stop <VOLNAME>


volume add-brick|remove-brick <VOLNAME> <NEW-BRICK>


volume rebalance <VOLNAME> start|stop|status


volume replace-brick <VOLNAME> (<BRICK> <NEW-BRICK>) start|pause|abort|status


volume set <VOLNAME> <KEY> <VALUE>


volume reset <VOLNAME>
```

Dlaczego GlusterFS?

- dowolny sprzęt kompatybilny z Fedora 11
- wygodne skalowanie
- elastyczne volumeny
- NFS, native client, CIFS, HTTP, FTP
- zawsze możliwy dostęp do danych
- brak centralnego serwera z metadanymi
- POSIX

Wydajność

Gluster vs NFS

System	I/O rate	Resp time
Random Read (4K) threads=32		
Gluster	892,91	35,835
NFS	13257,66	2,406
Random Write (4K) threads=32		
Gluster	934,23	34,249
NFS	214,05	149,496
50-50 Read/Write (4K) threads=32		
Gluster	875,27	36,558
NFS	512,19	62,473

Gluster vs NFS (2)

System	MB/sec	Resp time
Sequential Read (MB/sec) threads=32 512KB block		
Gluster	107,90	148,276
NFS	111,69	143,259
Sequential Write (MB/sec) threads=32 512KB block		
Gluster	49,97	321,204
NFS	25,16	636,010

Wdrożenie...

... czyli nie zawsze jest tak pięknie

Wdrożenie - porażki

- Sprzęt podobnej klasy
- Wydajność małych plików
- Wbudowany NFS = FAIL
- Support
- Wersja stabilna != stabilna
- Replace-brick działa, ale w teorii :)

Wrożenie - plusy

- Odseparowanie storage-a
- Duże tempo rozwoju (może jutro będzie patch)
- Wydajność wystarczająca w Gratce
- Migracja na EXT4 :)
- Mimo wszystko stabilne działanie
- Prostota użytkowania

gratka technologie

Pytania?