



Stainless Steel Elephant

Standing on the shoulders of giants

Alex Chistyakov, Principal Engineer at Git in Sky
Feb 05 2016, PgConf.Russia 2016

Who we are

- A small consulting company from SPb., Russia
- Web ops engineers, performance engineers
- Automation engineers
- PostgreSQL fans



Who are you?

- DBAs?
- DBDs?
- DB(.*)s?
- PostgreSQL fans?
- Performance engineering, anyone?



OK, what is going on here?

- FreeBSD!



OK, what is going on here?

- FreeBSD!
- Not scared? Let's try again!



OK, what is going on here?

- FreeBSD!
- Not scared? Let's try again!
- ZFS!



OK, what is going on here?

- FreeBSD!
- Not scared? Let's try again!
- ZFS!
- Hmm, you still there?



OK, what is going on here?

- FreeBSD!
- Not scared? Let's try again!
- ZFS!
- Hmm, you still there?
- DragonFly BSD! HAMMER!



OK, what is going on here?

- FreeBSD!
- Not scared? Let's try again!
- ZFS!
- Hmm, you still there?
- DragonFly BSD! HAMMER!
- Okay, enough for you



Hardware configuration

- Dell R430
- 32Gb RAM
- PERC H730 mini
- Two Edge Boost Pro P SSDs in RAID0
- 2 x Xeon(R) CPU E5-2630 v3 @ 2.40GHz



Ready, set, go!

- BTW, the plan is:
- Install something very elite
- Get great results
- Install something less elite
- Get not so great results
- Compare and swap
- **PROFIT!**



DragonFly BSD 4.4.1 (the latest)

- It's dead, Jim!
- Was not able to even install it

```
idrac, PowerEdge R430, User: root, 1.4 fps
File View Macros Tools Power Next Boot Virtual Media Help

perfbias31: <CPU perf-energy bias> on cpu31
coretemp31: <CPU On-Die Thermal Sensors> on cpu31
ACPI: Enabled 2 GPEs in block 00 to 3F
orm0: <ISA Option ROMs> at iomem 0xc0000-0xc7fff,0xed800-0xf17ff on isa0
pmtimer0 on isa0
vga0: <Generic ISA UGA> at port 0x3c0-0x3df iomem 0xa0000-0xbffff on isa0
sc0: <System console> at flags 0x100 on isa0
sc0: UGA <16 virtual consoles, flags=0x300>
hpt27xx: no controller detected.
CAM: Configuring 13 busses
panic: Bad link elm 0xffffffffe06a6ae0 prev->next != elm
cpuid = 3
Trace beginning at frame 0xffffffffe3ba01b370
panic() at panic+0x267 0xffffffff805f7b93
panic() at panic+0x267 0xffffffff805f7b93
camperiphfree() at camperiphfree+0xa3 0xffffffff802a7309
cam_periph_release() at cam_periph_release+0x67 0xffffffff802a82b8
probedone() at probedone+0xa01 0xffffffff802a64d2
camisr_runqueue() at camisr_runqueue+0x3f3 0xffffffff802a2719
Debugger("panic")

CPU3 stopping CPUs: 0xffffffff7
stopped
Stopped at          Debugger+0x38:  movb    $0,0x13352bb(%rrip)
db> █
```

- DOA too

```
idrac, PowerEdge R430, User: root, 1.4 fps
File View Macros Tools Power Next Boot Virtual Media Help

perfbias31: <CPU perf-energy bias> on cpu31
coretemp31: <CPU On-Die Thermal Sensors> on cpu31
ACPI: Enabled 2 GPEs in block 00 to 3F
orm0: <Option ROMs> at iomem 0xc0000-0xc7fff,0xed800-0xf17ff on isa0
pmtimer0 on isa0
vga0: <Generic ISA UGA> at port 0x3c0-0x3df iomem 0xa0000-0xbffff on isa0
sc0: <System console> at flags 0x100 on isa0
sc0: UGA <16 virtual consoles, flags=0x300>
hpt27xx: no controller detected.
CAM: Configuring 13 busses
panic: Bad link elm 0xffffffffe06a6b7670 prev->next != elm
cpuid = 0
Trace beginning at frame 0xffffffff81cd2800
panic() at panic+0x267 0xffffffff805f0d10
panic() at panic+0x267 0xffffffff805f0d10
cam_periph_alloc() at cam_periph_alloc+0x479 0xffffffff802a7d1a
xpt_scan_lun() at xpt_scan_lun+0x1d9 0xffffffff802a3ebe
xpt_action() at xpt_action+0xb36 0xffffffff802a1fc4
xpt_scan_bus() at xpt_scan_bus+0x1ba 0xffffffff802a1190
Debugger("panic")

CPU0 stopping CPUs: 0xffffffffe
stopped
Stopped at          Debugger+0x38:  movb    $0,0x12f311b(%rip)
db> 5
```



DragonFly BSD 4.0.6

- Hoorah!
- Well, that's elite enough

- [root@dfbsd /usr]# uname -a

```
DragonFly dfbsd.gitinsky.com 4.0-RELEASE DragonFly v4.0.6-RELEASE #0: Fri Jun 12  
19:57:23 EDT 2015
```

```
root@www.shiningsilence.com:/usr/obj/home/justin/release/4_0/sys/X86_64_GENERIC x86_64
```

```
[root@dfbsd /usr]#
```

Now you have two problems

- Okay, what's next?
- “`pgbench -i -s 1000 --foreign-keys pgbench`” to load some data
- “`pgbench -T 300 -P 10 -c N -j N -r pgbench`” to run some tests
- A carefully trained monkey to interpret the results

- `date; pgbench -i -s 1000 --foreign-keys pgbench; date`

Fri Feb 5 03:33:28 MSK 2016

creating tables...

100000000 of 100000000 tuples (100%) done (elapsed 257.56 s, remaining 0.00 s)

vacuum...

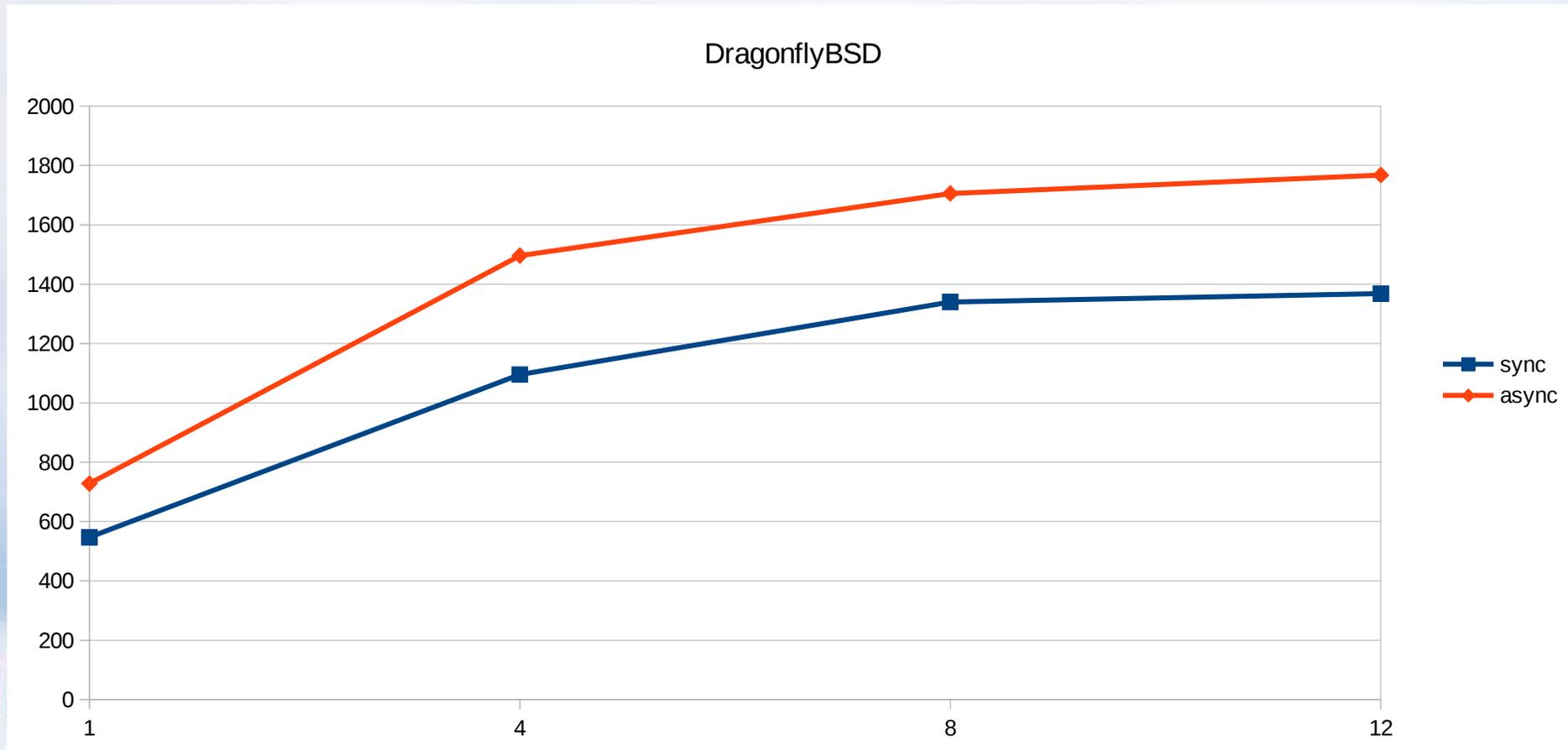
set primary keys...

set foreign keys...

done.

Fri Feb 5 03:43:30 MSK 2016

- X – num of clients/threads, Y – transactions per second



A note on observability tools

- I'm in your gdb killing your backendz!
- [root@dfbsd /home/chistyakov]# gdb -ex "set pagination 0" -ex "thread apply all bt" -batch -p 436012

Couldn't get registers: Device busy.

Couldn't get registers: Device busy.

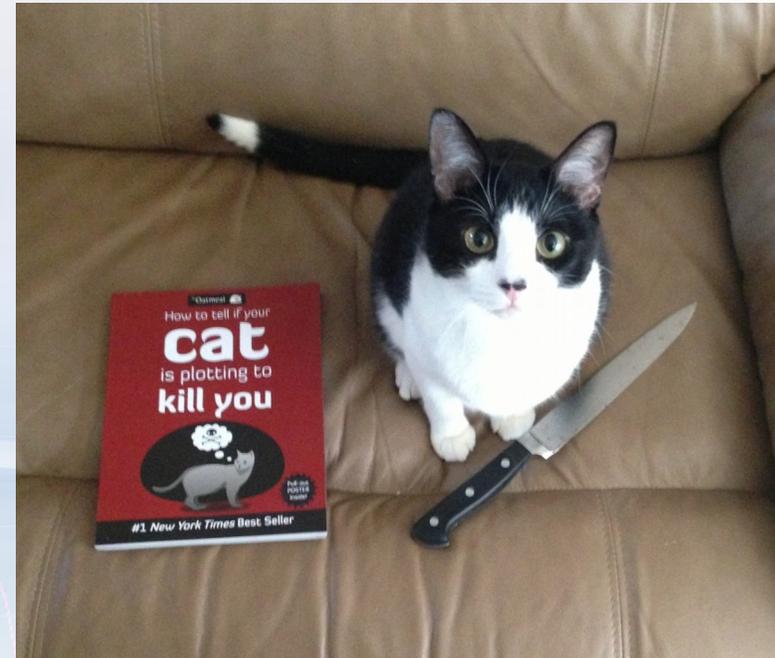
Couldn't get registers: Device busy.

Thread 1 (process 436012):

Couldn't get registers: Device busy.

Quitting: ptrace: Device busy.

[root@dfbsd /home/chistyakov]#



- BSD guys,
what is wrong
with you?

```
idrac, PowerEdge R430, User: root, 2.2 fps
File View Macros Tools Power Next Boot Virtual Media Help
mfi0: COMMAND 0xfffffe00010b2198 TIMEOUT AFTER 90 SECONDS
mfi0: COMMAND 0xfffffe00010b2ff0 TIMEOUT AFTER 90 SECONDS
mfi0: COMMAND 0xfffffe00010b3a90 TIMEOUT AFTER 90 SECONDS
mfi0: COMMAND 0xfffffe00010b2d48 TIMEOUT AFTER 90 SECONDS
mfi0: COMMAND 0xfffffe00010b3320 TIMEOUT AFTER 90 SECONDS
mfi0: COMMAND 0xfffffe00010b2dd0 TIMEOUT AFTER 90 SECONDS
mfi0: COMMAND 0xfffffe00010b36d8 TIMEOUT AFTER 90 SECONDS
mfi0: COMMAND 0xfffffe00010b3100 TIMEOUT AFTER 90 SECONDS
mfi0: COMMAND 0xfffffe00010b6378 TIMEOUT AFTER 90 SECONDS
mfi0: COMMAND 0xfffffe00010b2e58 TIMEOUT AFTER 90 SECONDS
mfi0: COMMAND 0xfffffe00010b3980 TIMEOUT AFTER 90 SECONDS
mfi0: COMMAND 0xfffffe00010b3540 TIMEOUT AFTER 90 SECONDS
mfi0: COMMAND 0xfffffe00010b38f8 TIMEOUT AFTER 90 SECONDS
mfi0: COMMAND 0xfffffe00010b3b18 TIMEOUT AFTER 90 SECONDS
mfi0: COMMAND 0xfffffe00010b35c8 TIMEOUT AFTER 90 SECONDS
mfi0: COMMAND 0xfffffe00010b23b8 TIMEOUT AFTER 90 SECONDS
mfi0: COMMAND 0xfffffe00010b40f0 TIMEOUT AFTER 90 SECONDS
mfi0: COMMAND 0xfffffe00010b3ed0 TIMEOUT AFTER 90 SECONDS
mfi0: COMMAND 0xfffffe00010b4398 TIMEOUT AFTER 90 SECONDS
mfi0: COMMAND 0xfffffe00010b3dc0 TIMEOUT AFTER 90 SECONDS
mfi0: COMMAND 0xfffffe00010b4640 TIMEOUT AFTER 90 SECONDS
mfi0: COMMAND 0xfffffe00010b3760 TIMEOUT AFTER 90 SECONDS
mfi0: COMMAND 0xfffffe00010b4178 TIMEOUT AFTER 90 SECONDS
mfi0: COMMAND 0xfffffe00010b46c8 TIMEOUT AFTER 90 SECONDS
```

- This is how a deferred success looks like

```
idrac, PowerEdge R430, User: root, 1.4 fps
File View Macros Tools Power Next Boot Virtual Media Help
kbd0 at ukbd0
ums0: <Mouse> on usb0
ums0: 3 buttons and [Z] coordinates ID=0
ums1: <Mouse REL> on usb0
ums1: 3 buttons and [XYZ] coordinates ID=0
Root mount waiting for: usb0
ugen0.5: <Avocent> at usb0
umass0: <SCSI Transparent Interface 0> on usb0
umass0: SCSI over Bulk-Only; quirks = 0x0000
umass0:10:0:-1: Attached to scbus10
Trying to mount root from ufs:/dev/mfid0s1a [rw]...
cd0 at umass-sim0 bus 0 scbus10 target 0 lun 0
cd0: <iDRAC Virtual CD 0329> Removable CD-ROM SCSI-0 device
cd0: 40.000MB/s transfers
cd0: cd present [327711 x 2048 byte records]
cd0: quirks=0x10<10_BYTE_ONLY>
da0 at umass-sim0 bus 0 scbus10 target 0 lun 1
da0: <iDRAC Virtual Floppy 0329> Removable Direct Access SCSI-0 device
da0: 40.000MB/s transfers
da0: Attempt to query device size failed: NOT READY, Medium not present
da0: quirks=0x2<NO_6_BYTE>
pid 17 (sh), uid 0: exited on signal 10
Feb 5 08:04:01 init: /bin/sh on /etc/rc terminated abnormally, going to single
user mode
Enter full pathname of shell or RETURN for /bin/sh: █
```



Okay, SmartOS then

- 100000000 of 1000000000 tuples (100%) done (elapsed 107.03 s, remaining 0.00 s).

vacuum...

set primary keys...

set foreign keys...

done.

real 4m27.237s

user 0m23.381s

sys 0m2.118s

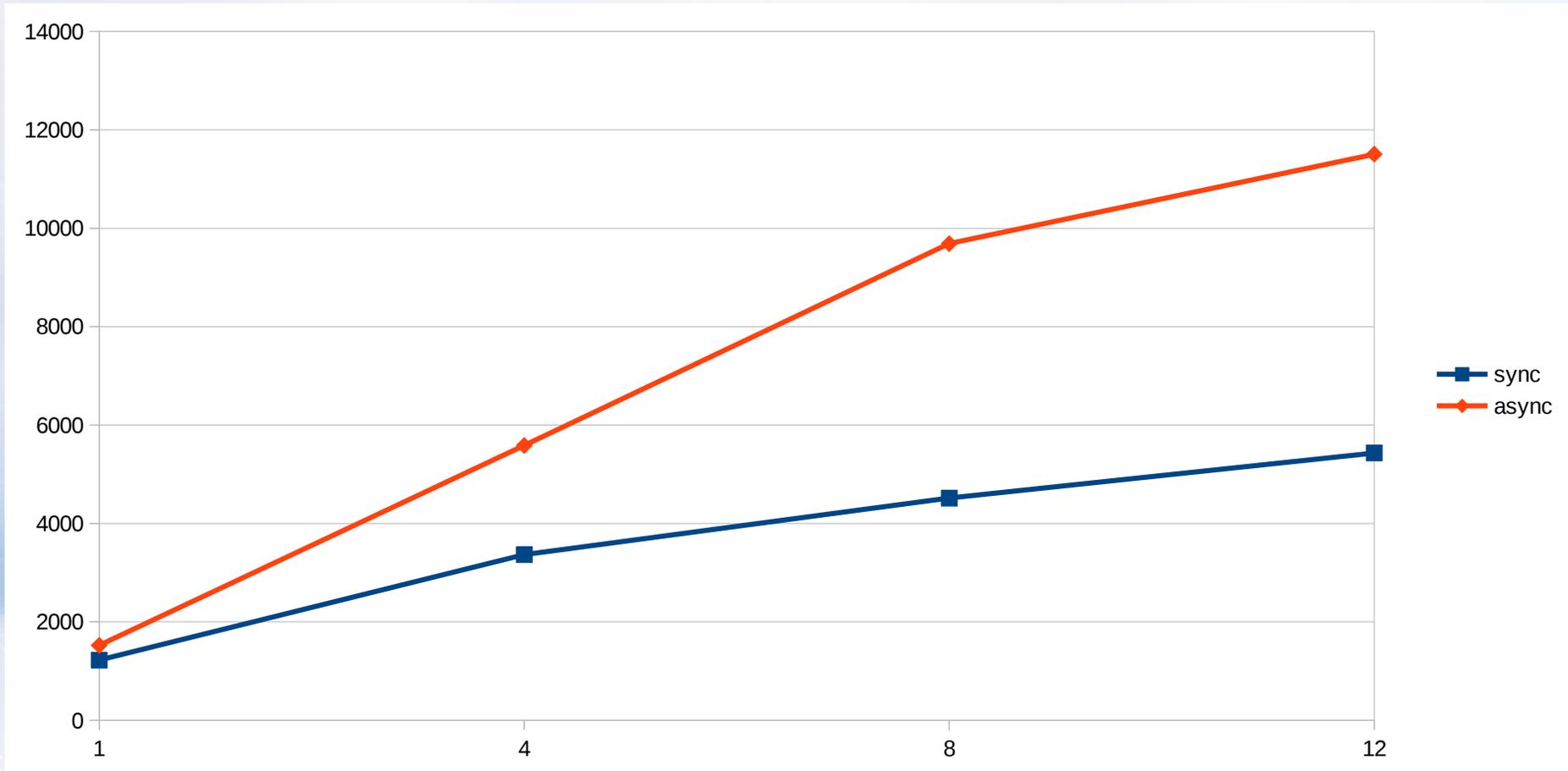
bash-4.1\$

A note on observability tools

- zpool iostat 1

zones	25,3G	419G	0	4,83K	0	417M
zones	25,3G	419G	0	4,63K	0	400M
zones	25,3G	419G	0	4,58K	0	386M
zones	25,4G	419G	0	4,55K	0	366M
zones	25,4G	419G	0	4,57K	0	364M
zones	25,4G	419G	0	4,51K	0	378M
zones	25,4G	419G	0	4,20K	0	358M
zones	25,4G	419G	0	4,83K	0	422M
zones	25,5G	419G	0	4,32K	0	360M

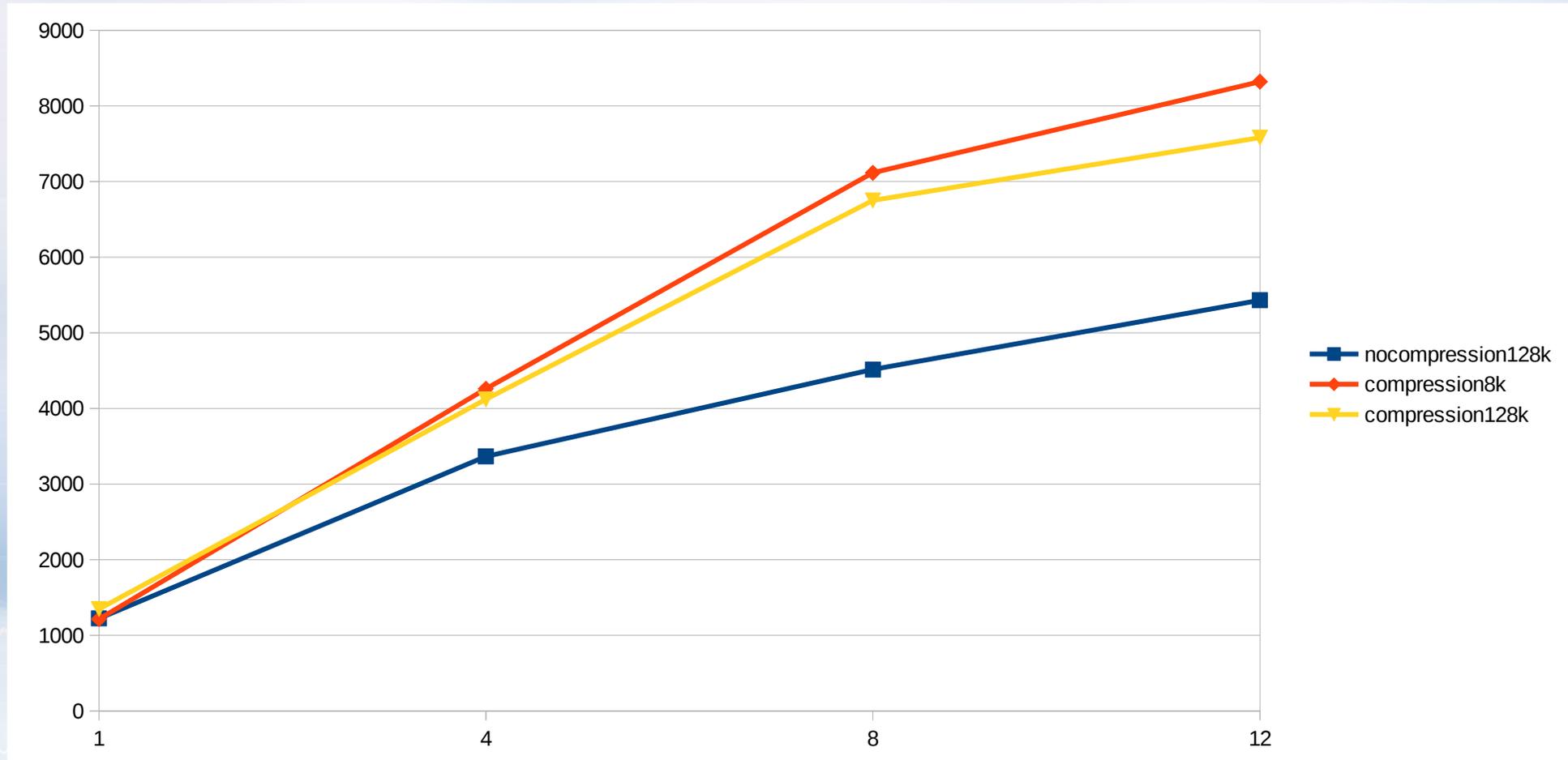
- * This chart is a lie



Lies, damned lies and LZO

- Compression MUST be turned on, period
- ZFS record size: 128K by default
- PostgreSQL block: 8K
- Compressing 128K blocks is more efficient
- Random reads: 8K → 128K amplification
- Fight!

- * This chart is probably a lie



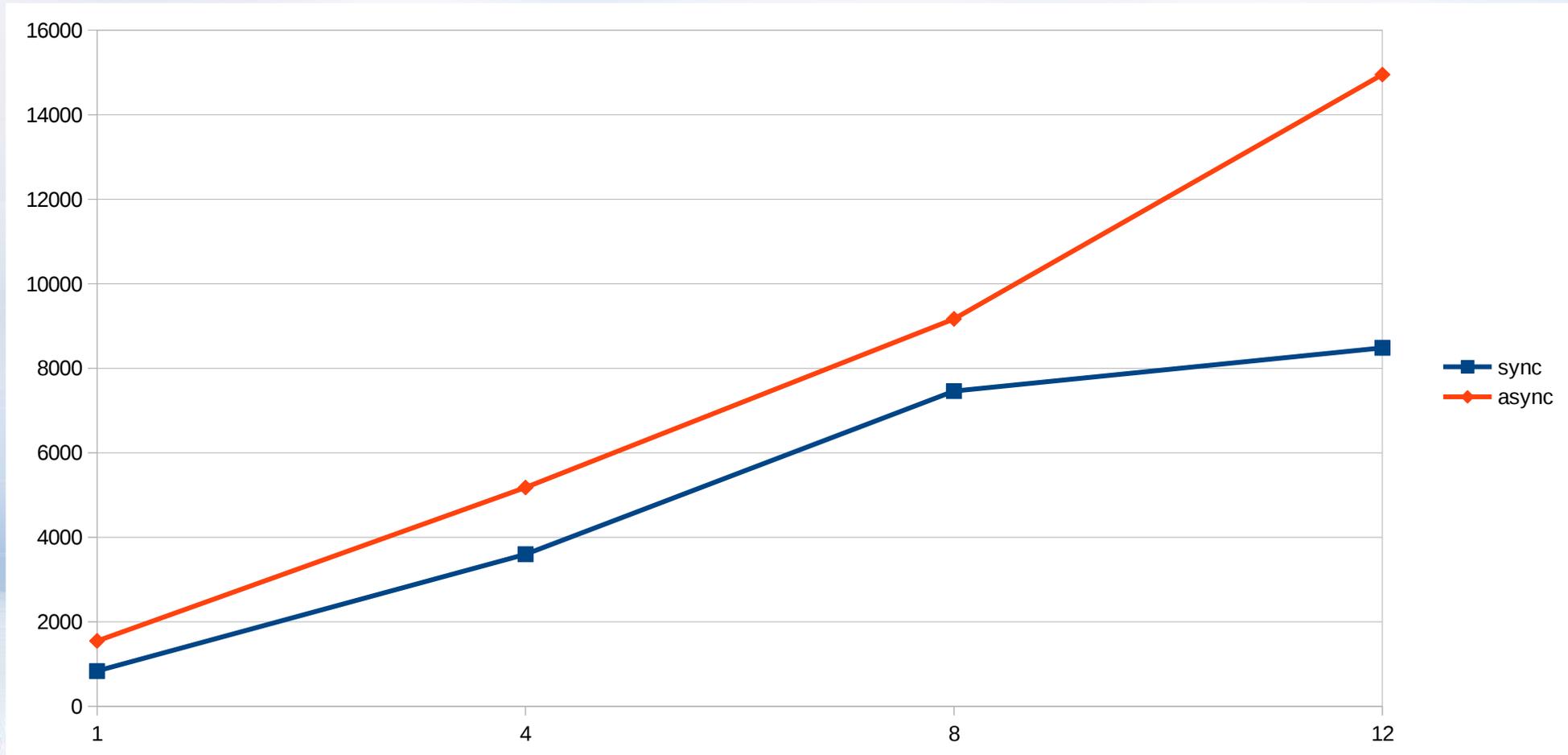
Guess which one is 128K?

- zones/var refcompressratio 5.20x -
zones/var written 3,31G -
zones/var logicalused 16,8G -
zones/var logicalreferenced 16,8G -
- zones/var refcompressratio 11.64x -
zones/var written 1,27G -
zones/var logicalused 14,7G -
zones/var logicalreferenced 14,7G -

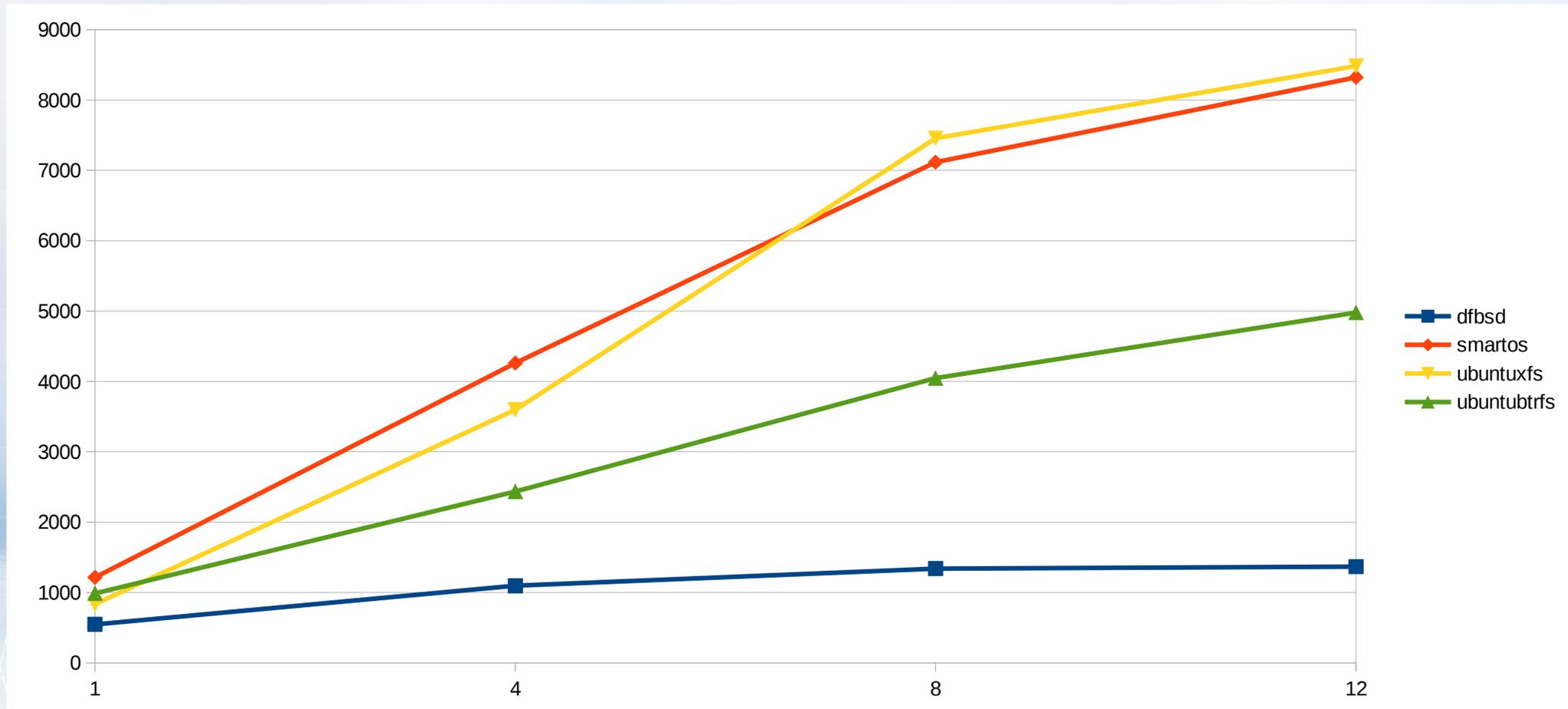
- 100000000 of 1000000000 tuples (100%) done (elapsed 65.91 s, remaining 0.00 s)
vacuum...
set primary keys...
set foreign keys...
done.

```
real 2m54.867s  
user 0m27.372s  
sys 0m0.872s  
postgres@ubuntu:~$
```

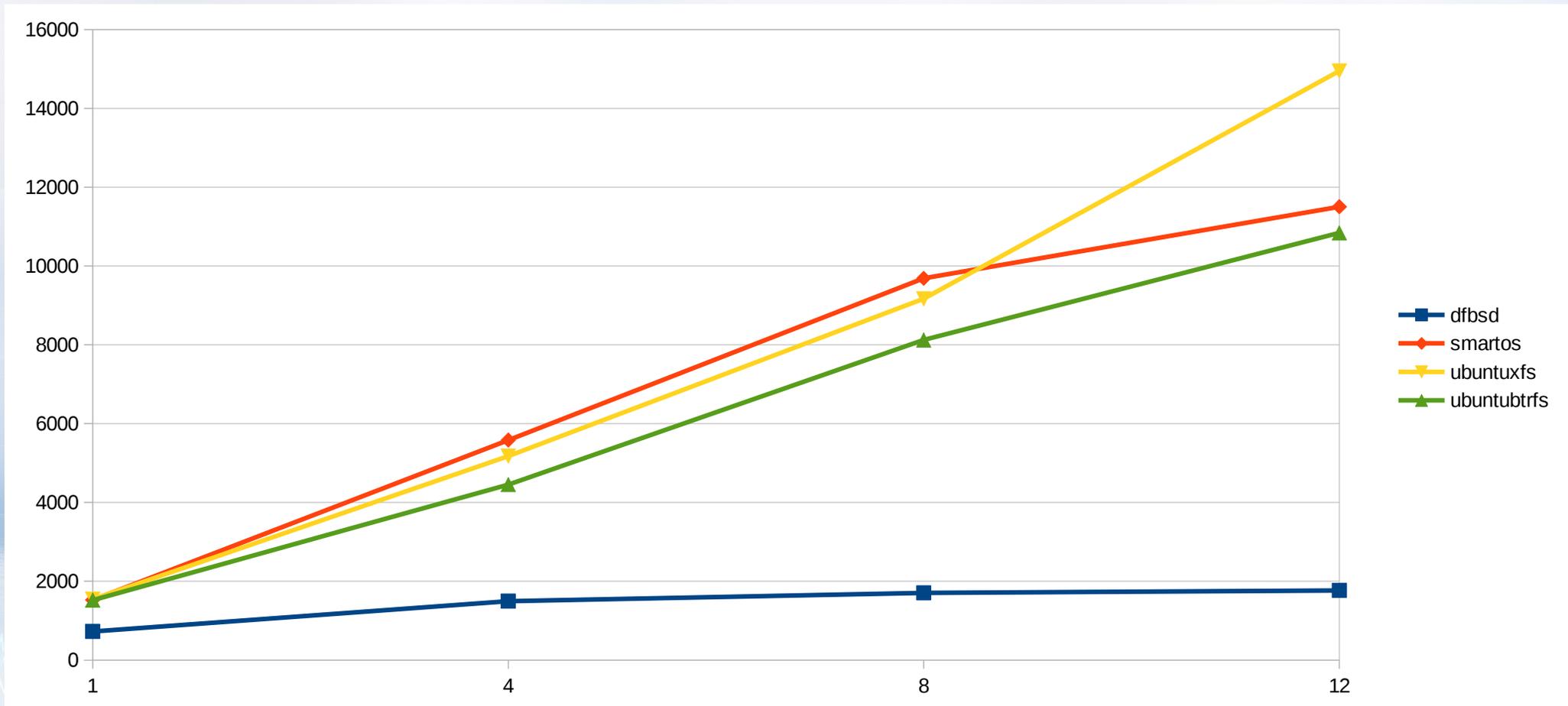
- This chart is not a lie



- SmartOS/ZFS (compression, 8K block) – Ubuntu/XFS : 1 - 1



- And, finally, we have a winner! Ubuntu/XFS



- Hammers don't fly at decent speed
- FreeBSD is very elite
- But you still can use ZFS if you are not elite enough for FreeBSD
- Ubuntu and XFS are not even close to be elite but are very fast
- SmartOS? Never heard of it!

- - Why did you tweak ZFS but did not tweak BTRFS and HAMMER?
- - Because I'm a ZFS zealot, you heretics!



That's all folks!

- Thank you!
- alex@gitinsky.com
- <http://gitinsky.com>
- <http://meetup.com/DevOps-40>