TUNING LINUX, WINDOWS AND FIREBIRD FOR HEAVY WORKLOAD

Alex Kovyazin,
IBSurgeon

Firebird Tour 2017: Performance Optimization
Prague, Bad Sassendorf, Moscow

Firebird 2017 Tour: Performance Optimization

- Firebird Tour 2017 is organized by Firebird Project, IBSurgeon and IBPhoenix, and devoted to Firebird Performance
- The Platinum sponsor is Moscow Exchange
- Tour's locations and dates:
  - October 3, 2017 – Prague, Czech Republic
  - October 5, 2017 – Bad Sassendorf, Germany
  - November 3, 2017 – Moscow, Russia

- Platinum Sponsor
- Sponsor of
  - «Firebird 2.5 SQL Language Reference»
  - «Firebird 3.0 SQL Language Reference»
  - «Firebird 3.0 Developer Guide»
  - «Firebird 3.0 Operations Guide»
- Sponsor of Firebird 2017 Tour seminars
- www.moex.com

PLATINUM

IBSurgeon

- Replication, Recovery and Optimization for Firebird and InterBase since 2002
- Platinum Sponsor of Firebird Foundation
- Based in Moscow, Russia

www.ib-aid.com

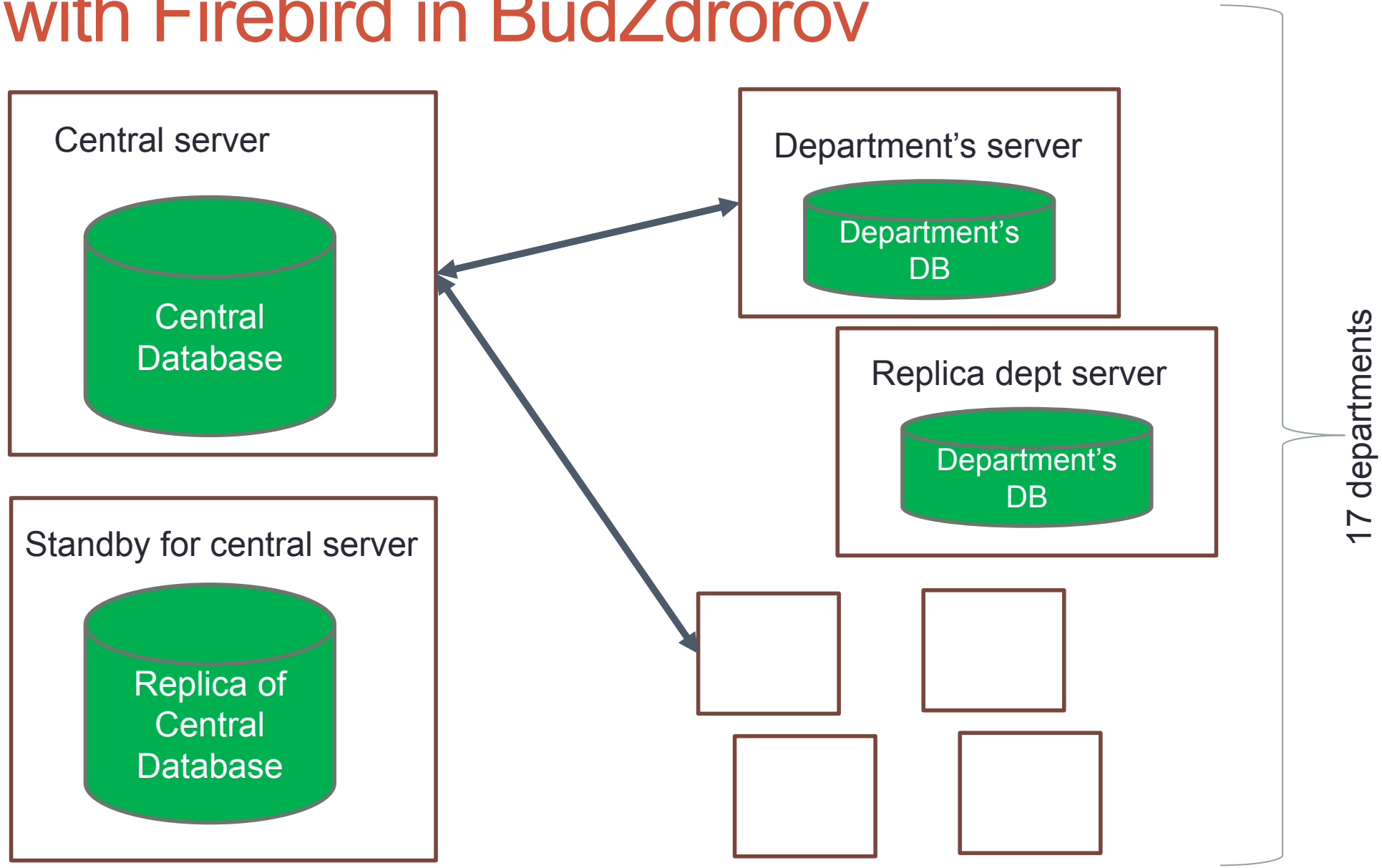
Agenda

- Real customers with big databases
- Hardware they use
- OS tuning
  - CPU
  - RAM
  - IO
  - Network
- Firebird configuration

Customer 1: http://klinikabudzdorov.ru

- BudZdorov
- Medical centers and hospitals in Moscow, Saint-Petersburg and major cities in Russia
- 17 departments
- 365 days per year, from 8-00 to 21-00

ERP with Firebird in BudZdrorov



BudZdorov: Central database

- Size = 453 Gb
- Daily users = from 700 to 1800 (peak)
- Hardware server
- OS = Linux CentOS 6.7
- Firebird 2.5 Classic + HQbird
- Client-server, connected through optic with departments
- With async replica on the separate server

FBDataGuard 5.5

172.16.9.9:8082/static/dashboard.html#hqbirdsrv

FBDataGuard 5.5 - HQbird Enterprise: CENTRAL MASTER

Page refresh in 00:37 | English

Server

Active server

Server: **running**
Version: LI-V2.5.8.27067 Firebird 2.5 HQbird

Temp files

Temp files: **OK** [Last run: 13 min 12 sec ago]
Quantity: **0**, Size: **0 b**

Server log

Server log: **OK** [Last run: 43 min 12 sec ago]

Replication Log

Replication Log: OFF

Server space

Server space: **OK** [Last run: 13 min 12 sec ago]
Size: **157.2 Mb**

Agent space

Agent space: **OK** [Last run: 13 min 11 sec ago]
Agent size: **27.3 Mb**

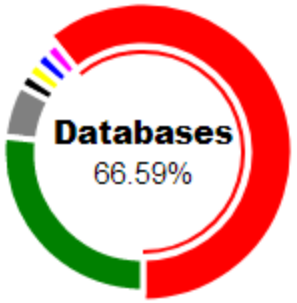
Send logs

Send logs: OFF

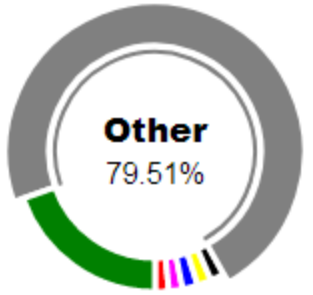
Auto updates

Auto updates: **UNKNOWN**
Installed: **5.5**

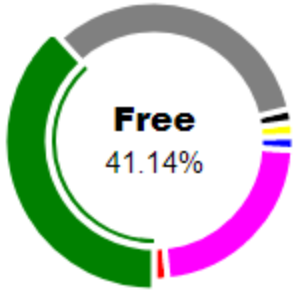
Disk space



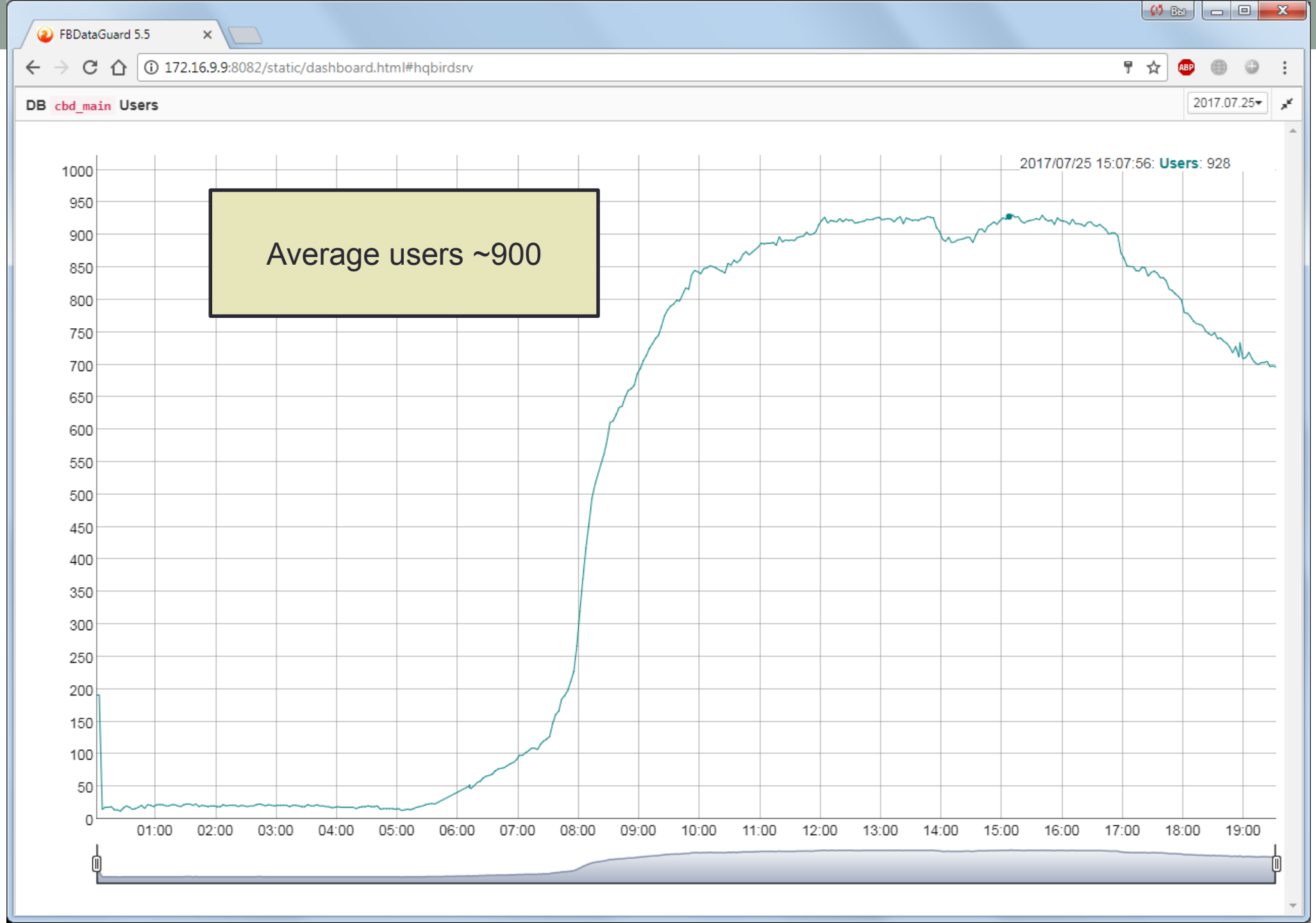
[/home] 679.7 Gb
Databases 66.59 %
Other 5.10 %
Free 28.31 %



[/] 78.6 Gb
Server 0.20 %
Agent 0.03 %
Other 79.51 %
Free 20.26 %



[/3par-vv1] 1.8 Tb
Backups 23.38 %


FBDataGuard 5.5
172.16.9.9:8082/static/dashboard.html#hqbirdsrv
DB cbd_main Users
2017.07.25
2017/07/25 15:07:56: Users: 928
Average users ~900
1000
950
900
850
800
750
700
650
600
550
500
450
400
350
300
250
200
150
100
50
0
01:00
02:00
03:00
04:00
05:00
06:00
07:00
08:00
09:00
10:00
11:00
12:00
13:00
14:00
15:00
16:00
17:00
18:00
19:00

Customer 2: Customer revoked permission to publish information ☺

- Customer #2
- Repair services for xxxxx across Russia
- 365 days per year, 24x7, with 1 hour maintenance every day

Customer #2: Central Database

- Size = 250Gb
- Daily users from 500 to 1000 (peak)
- Hardware server
- Windows 2012R2
- Firebird 3
- Middleware (web)

Performance problems – as usual

- Long running active transactions
  - Garbage collection is blocked for hours and even days
- Badly written SQLs in applications
- Peaks of load
  - People are mostly sick during the winter
  - Railroad has peak of loads
- Anti-failure approach
  - Replica with 1 minute delay

Tuning goals

1. Tune for **throughput** first, then, if possible, for response time
    1. During the day users are Ok with performance
    2. Problems occur only during periods of high load
2. Tune OS to get appropriate results from the powerful hardware

General requirements for high load server

1. Not a Primary/Backup Controller/Small Business Server (Windows)
2. No Exchange (store.exe and MSSQL inside) or Sharepoint (MSSQL inside) or dedicated MSSQL
   - Each MSSQL should be restricted in memory usage
3. Not a File Server/Print Server/Terminal Server/Web server
4. If it is virtual machine, it should be really fast
5. If there is your middleware - does it benefit from being on the same server (i.e., local protocol)?
   1. If not, put it on another server
   2. If yes, make sure to allocate resources

Dedicated server means dedicated!

HARDWARE

Hardware configuration in BudZdorov

- Server model: HP ProLiant DL380p Gen8 2x Xeon(R) CPU E5v2 @ 2.60GHz
  - 2 processors* 6 physical cores * 2 HyperThreading = **24 cores**
- **RAM 384Gb**
- Disks:
  - RAID10 array on SSDs – **680Gb** – for database
  - Tmpfs on SSD -158Gb
  - SAN on SAS15k - 1.8Tb
  - External mounted backup partition for 1.4Tb
- Network
  - BroadCom NetXtreme BCM5719 Gigabit Ethernet PCIe

Hardware configuration in Customer#2

- Server model: Dell PowerEdge R810, 2x Xeon(R) CPU E5-2630 v4
  - **24 cores**
- **RAM 256Gb**
- Disks:
  - RAID1 array on SSDs – **480Gb** – for database
  - OS on SAS15K - 160Gb
- Network
  - Broadcom 57810, 10Gb/sec

TUNING OS/HARDWARE

CPU

- How to improve CPU utilization?
- How can we improve distribution of load between cores?

CPU at Linux

- **irqbalance**
- yum install -y irqbalance && chkconfig irqbalance on && service irqbalance start

- Result: better CPU load distribution, increased throughput

CPU at Windows

- Windows: only CPU_AFFINITY in Firebird configuration

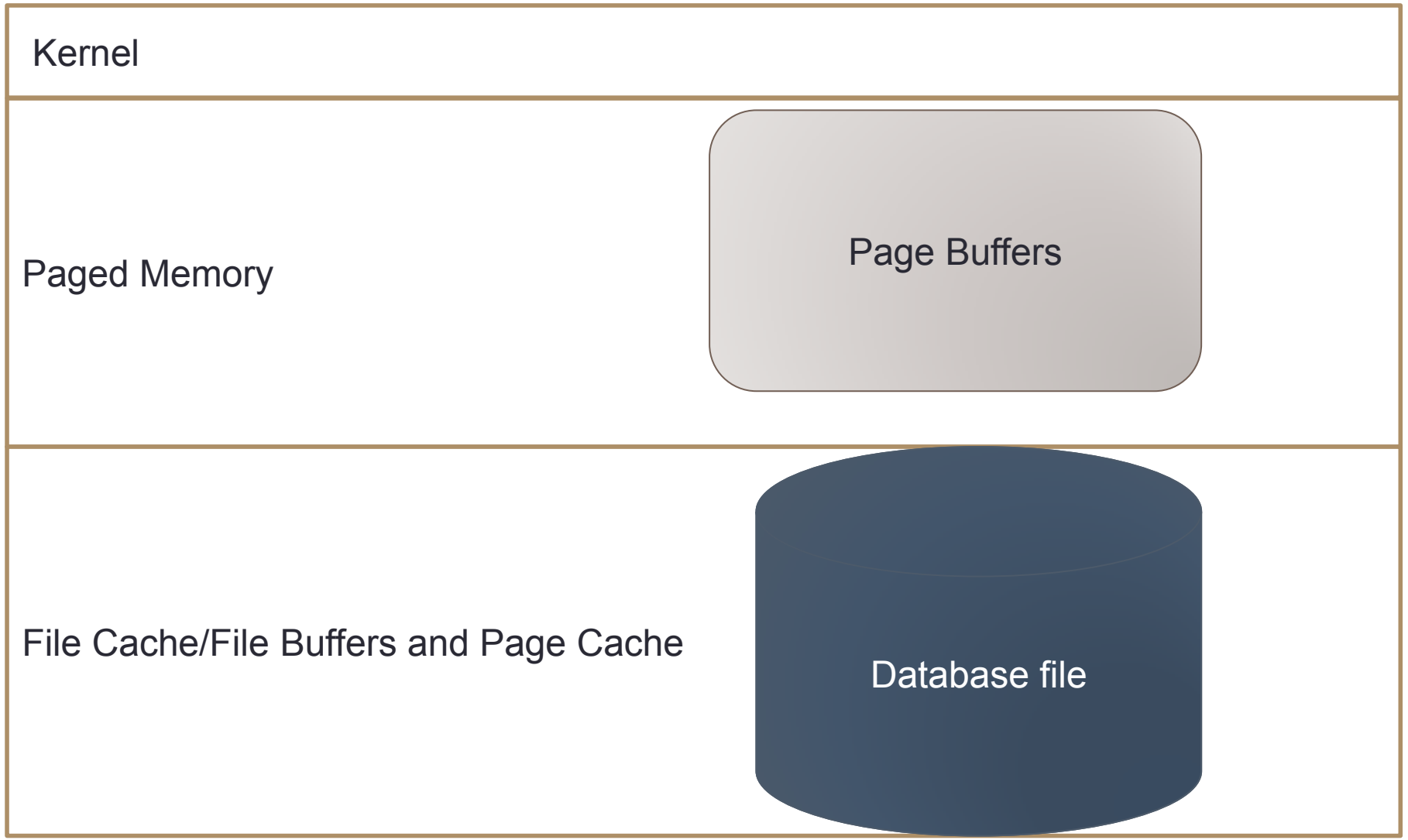
- Result: some cores can be excluded from Firebird usage (reserved for middleware/other services), less conflicts, slightly better throughput

RAM Tuning

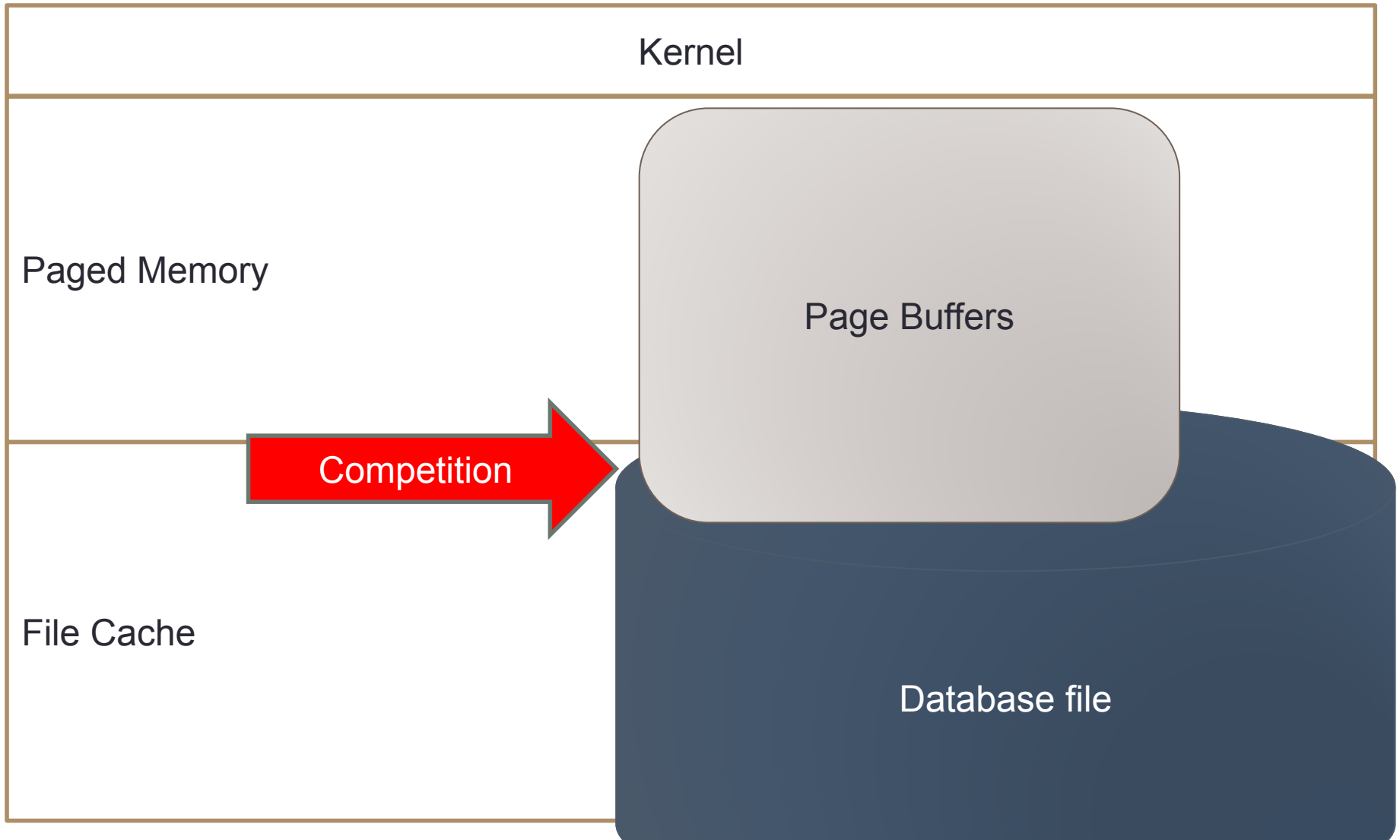
- How to effectively use available RAM?
- How to avoid swapping?

- Firebird settings:
  - DefaultDBCachePages – page buffers cache
  - FileCacheSystemThreshold – limit to use/not use file cache
  - TempCacheLimit – memory space for sorting

Tuning RAM: 3 types of memory



RAM in case of Big Databases and Big Caches



OS Memory Manager vs Firebird

- If Page Buffers is more than Paged Memory, OS Memory Manager tries to send it to swap
- Race for resources between Paged Memory and File Cache leads to swapping

Tuning RAM on Linux

- On Linux RedHat/CentOS file cache is not limited by default

vm.pagecache = 100 #default

- For Classic – it is more or less fine, since it uses file cache heavily
- For SuperServer it is not great, since SS 3.0 can use many page buffers

Recommendation is to limit file cache to 40-50%:

vm.pagecache = 50

Tuning RAM on Linux

- We know that database should be kept in RAM: need to reduce swapping!

- vm.swappiness = 10
- vm.dirty_ratio = 60
- vm.dirty_background_ratio = 2
- vm.min_free_kbytes = 1048576

Tuning RAM at Windows

- Windows Memory Manager has the following default scenario of using RAM:

  50% paged memory

  41% file cache

  9% kernel

Tip: use RAMMap tool to see memory allocation

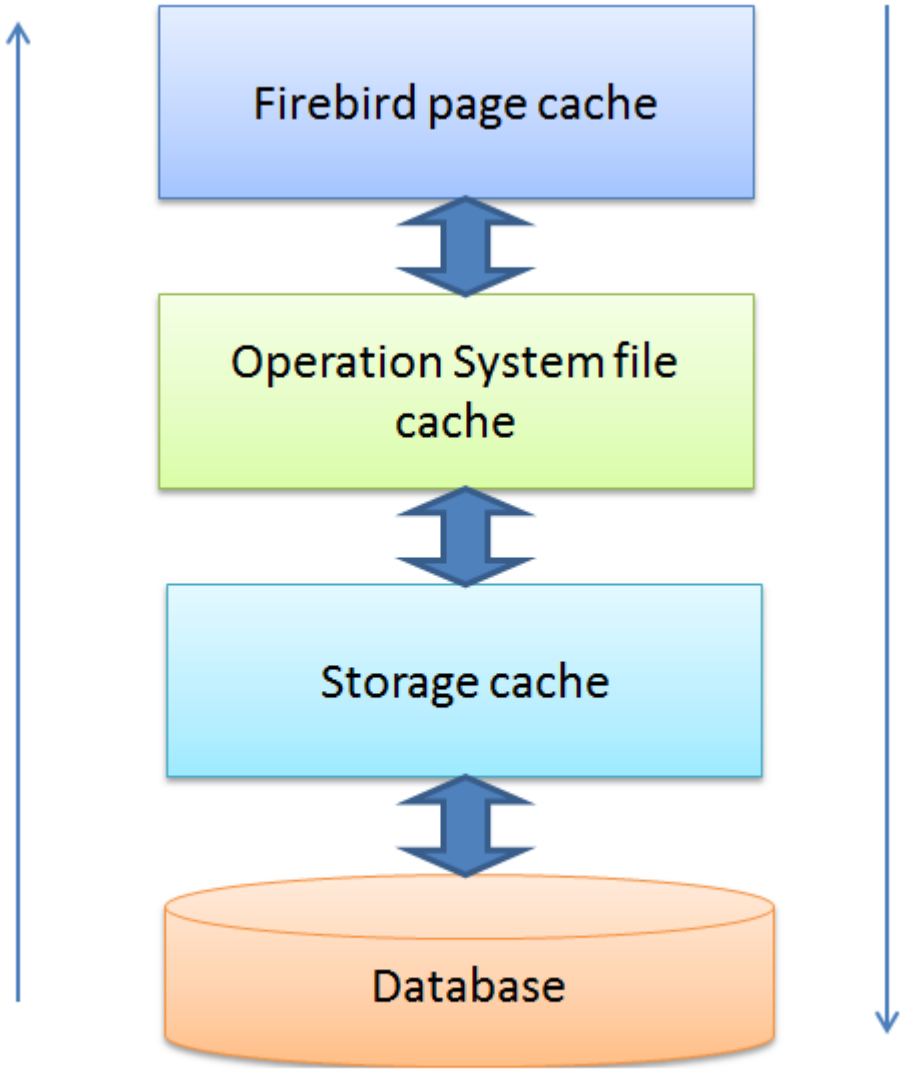
- Memory distribution can be changed in registry/role settings

Recommendations for RAM on Windows

- Page Buffers must be < Paged Memory (50% of RAM by default)
  - %% can be changed on Windows level
- File Cache should be On
  - For Classic and SuperClassic without exceptions
  - For SuperServer with databases with size more than RAM > 2x
- File Cache should be enough to keep frequently requested parts of database
- Firebird by default has file cache enabled: condition is

DefaultDBCachePages < FileSystemCacheThreshold

When can we disable File Cache?



- File Cache can be disabled for SuperServer for
  - Read Only databases
  - For database which fits into Page Buffers with very low % of writes
  - For databases on SSD with small % of writes
- Test it!

Paging file tuning

- In case of balanced settings for Page Buffers and enabled File Cache, and in case of RAM > 32Gb, page file can be limited to 16Gb.
- Page file will work fast on SSD – but not on the SSD with database!
  - Monitor life span of SSD!

Linux: general recommendations

- Centos
- Linux version 2.6.32-642.13.1.el6.x86_64 (mockbuild@c1bm.rdu2.centos.org) (gcc version 4.4.7 20120313 (Red Hat 4.4.7-17) (GCC)) #1 SMP Wed Jan 11 20:56:24 UTC 2017 – not so good, better choose newer OS version

- Use **fresh and popular** Linux distributions: Ubuntu 16+ Server and CentOS 7+
- Use **server version of Linux** distributions – it has already tuned limits for number of open files

Linux: file and process limits

```
# increase max user processes ulimit
(-u) 1291632
# Increase size of file handles and inode
cache
 fs.file-max = 2097152
```

Process forking is set to unlimited

- [root@mskv-cbd-new limits.d]# cat /etc/security/limits.d/90-nproc.conf
- *   soft   nproc   unlimited
- root   soft   nproc   unlimited
- [root@mskv-cbd-new security]# sed -e 's/^[\t]*//' /etc/security/limits.conf | grep "^[^#;]" | sort
- firebird   -   nofile   32768
- *** soft core unlimited**

/etc/xinetd.conf – the most important

```
# cps = 25 30 ==> configures xinetd to allow
#no more than 25 connections PER SECOND to any given
service. If this limit is reached, the service is
retired for 30 seconds.
 cps                 = 1500 10
# Sets the maximum number of requests xinetd can
handle at once.
 instances           = UNLIMITED
# per_source – Defines the maximum number of
#instances for a service per source IP address
 per_source          = UNLIMITED
```

IO

- For RAID
  - Write-Back
  - Enable cache
  - Setup ratio Reads/Writes according your load
  - BBU!
- SSD!

IO on Linux: File System and Barriers

- Ext4

Since we have RAID and Enterprise SSDs with power loss protection(and high quality hardware):

Barrier = 0 (disabled)

Disk IO on BudZdorov

- SSDs deliver high speed: 242Mb/sec

Database info	
Database name	/home/bases/central.fdb
Creation date	25.02.2016 1:01:49
Statistics date	26.07.2017 13:32:07
Page size	16384
Forced Write	ON
Dialect	3
OnDiskStructure	11.2
Implementation	ID24
Attributes	force write
Sweep interval	0
Oldest transaction	290288373
Oldest snapshot	290288374
Oldest active	290288374
Next transaction	290563113
Sweep gap (active - oldest)	1
TIP size	4435 pages, 70954 kilobytes
Snapshot TIP size	274740 transactions, 83 kilobytes
Active transactions	274739, 49% of daily average
Transactions per day	560932, for 518 days
Data versions percent	0,01% - records: 301658 mb, versions: 44 mb, pages 347955 mb, indices 87257 mb
Database Size	463472,00 megabytes, 94% data and indices
Stat speed	242,148 Mb/sec
Fragmented tables	

IO on Windows

- Enable disk cache (it does not work on Primary Disk Controller)

Temp space on RAM/SSD?

- TempCacheLimit – by default it is very low, increase it!
- Temp files are created in %TEMP% or /tmp or in TempDirectories
- Big TempCacheLimit allows to avoid temp files
- However, we still need big TempDirectories to create/restore indices

Network

```
# Increase number of incoming connections
        net.core.somaxconn = 4096
# Increase number of incoming connections backlog
        net.core.netdev_max_backlog = 65536
# Increase the maximum amount of option memory buffers
        net.core.optmem_max = 25165824
# Increase the tcp-time-wait buckets pool to prevent simple DOS attacks
        net.ipv4.tcp_max_tw_buckets = 1440000
        net.ipv4.tcp_tw_recycle = 1
        net.ipv4.tcp_tw_reuse = 1
```

Network

```
#Number of times SYNACKs for passive TCP connection.
        net.ipv4.tcp_synack_retries = 2
#Allowed local port range
        net.ipv4.ip_local_port_range = 2000 65535
#Protect Against TCP Time-Wait
        net.ipv4.tcp_rfc1337 = 1
#Decrease the time default value for tcp_fin_timeout connection
        net.ipv4.tcp_fin_timeout = 15
#Decrease the time default value for connections to keep alive
        net.ipv4.tcp_keepalive_time = 300
        net.ipv4.tcp_keepalive_probes = 5
        net.ipv4.tcp_keepalive_intvl = 15
```

Network

net.ipv4.tcp_congestion_control=htcp

net.ipv4.tcp_no_metrics_save=1

net.ipv4.tcp_moderate_rcvbuf=1

net.ipv4.tcp_slow_start_after_idle=0

net.core.rmem_default = 65536

net.core.wmem_default = 65536

net.core.rmem_max = 16777216

net.core.wmem_max = 16777216

net.ipv4.tcp_mem = 50576 64768 98152

net.ipv4.tcp_rmem = 4096 87380 16777216

net.ipv4.tcp_wmem = 4096 65536 16777216

Network queues

```
# For 24 CPU and 4 RX queues on NIC
    cat > /root/scripts/rps_boot.sh && chmod +x /root/scripts/rps_boot.sh
    bash -c 'echo 00000f > /sys/class/net/eth0/queues/rx-0/rps_cpus'
    bash -c 'echo 0000f0 > /sys/class/net/eth0/queues/rx-1/rps_cpus'
    bash -c 'echo 000f00 > /sys/class/net/eth0/queues/rx-2/rps_cpus'
    bash -c 'echo 00f000 > /sys/class/net/eth0/queues/rx-3/rps_cpus'
    bash -c 'echo 8192 > /sys/class/net/eth0/queues/rx-0/rps_flow_cnt'
    bash -c 'echo 8192 > /sys/class/net/eth0/queues/rx-1/rps_flow_cnt'
    bash -c 'echo 8192 > /sys/class/net/eth0/queues/rx-2/rps_flow_cnt'
    bash -c 'echo 8192 > /sys/class/net/eth0/queues/rx-3/rps_flow_cnt'
    #ethtool -G eth0 rx 2047
```

Network on Windows

- Remove unused network protocols
- Set the correct order of NICs

- Results: well, no big difference

Results from network tuning on Linux

- Much better throughput (users do not claim :)
- Significant decrease of Load Average
- Better distribution of load between CPUs

Conclusion for Linux configuration

- Use server distribution
- Use fresh version (CentOS 7+, Ubuntu Srv 16+)
- xinetd configuration is critical (due to Classic)
- Tune limits for process files, memory, file cache, and network

Conclusion for Windows Tuning

1. Main focus is on RAM tuning
2. CPU tuning is through CPU Affinity restrictions
3. Don’t forget to disable useless services/applications
4. In general Windows has far less parameters to tune, and they are not clear

Misc Windows Tuning tips

- Enable High Performance Power Plan
- Enable background processes priority
- Disable useless services
- Prefetch/Fetch On/Off – no differences
- Desktop Heap for Classic for non Local System account

FIREBIRD CONFIGURATION

Firebird at BudZdorov

- Firebird Classic 2.5
- Why not SuperClassic?
  - It is slow for more than 800 connections
  - No plans to fix it, since Firebird 3 SuperServer must be used

firebird.conf

- [root@mskv-cbd-new ~]# cat /opt/firebird/firebird.conf

DefaultDbCachePages = **1024**

TempCacheLimit = 67108864

TempDirectories = /dev/shm;/3par-vv1/fb_tmp;/tmp

LockHashSlots = **49009**

LockMemSize = 82048576

TcpRemoteBufferSize = 1448

TempCacheLimit tips

- Default firebird.conf
  - TempBlockSize = 1048576
    - May increase to 2 or 3mln bytes, but not to 16mb
  - TempCacheLimit = 67108864
    - SuperServer and SuperClassic. Classic = 8mb.
  - TempDirectories = c:\temp;d:\temp…
  - Increase TempCacheLimit for SuperServer and SuperClassic!

Maintenance and backups

- Automatic sweep is disabled
  - All connections are disconnected at 0-00
  - Manual sweep is at 00-05
- Verified backup (gbak) – every day at 1am
- Replication works as a standby

Summary for 2.5

- 1500 connections and 453Gb is a acceptable load for the Firebird 2.5
- Firebird and Linux should be tuned
- Maintenance is the key: sweep, restart of connections, backups
- Replication is mandatory for protection, since backup/restore takes 18 hours

Firebird at Customer#2

Firebird 3.0.2

DefaultDbCachePages = **2M**

FileCacheSystemThreshold=50M

TempCacheLimit = **9G**

LockHashSlots = **21001**

LockMemSize = 82048576

Summary

- Firebird 3.0.2 get the biggest benefit from huge number of page buffer (properly configured)
- Good design (short write) transactions eliminate need for everyday restarts

Useful links

- Collection of optimized Firebird configuration files

https://ib-aid.com/en/optimized-firebird-configuration/

- Firebird Hardware Guide

https://ib-aid.com/en/articles/firebird-hardware-guide/

- 45 Ways To Speed Up Firebird
- https://ib-aid.com/en/articles/45-ways-to-speed-up-firebird-database/

Thank you!

- Questions?

- www.ib-aid.com
- ak@ib-aid.com