«Каменный век закончился не потому, что закончились камни»

© Ахмед Заки Ямани

Ускорение традиционных баз данных нетрадиционными методами

Михаил Цветков

Если вам сказали, что ваша машина медленная...

Можно было пойти, и купить новую...

Но есть проблемы

Самолет все равно быстрее

Дисковым СУБД нужны «крылья», чтобы догнать In-Memory DB в аналитике

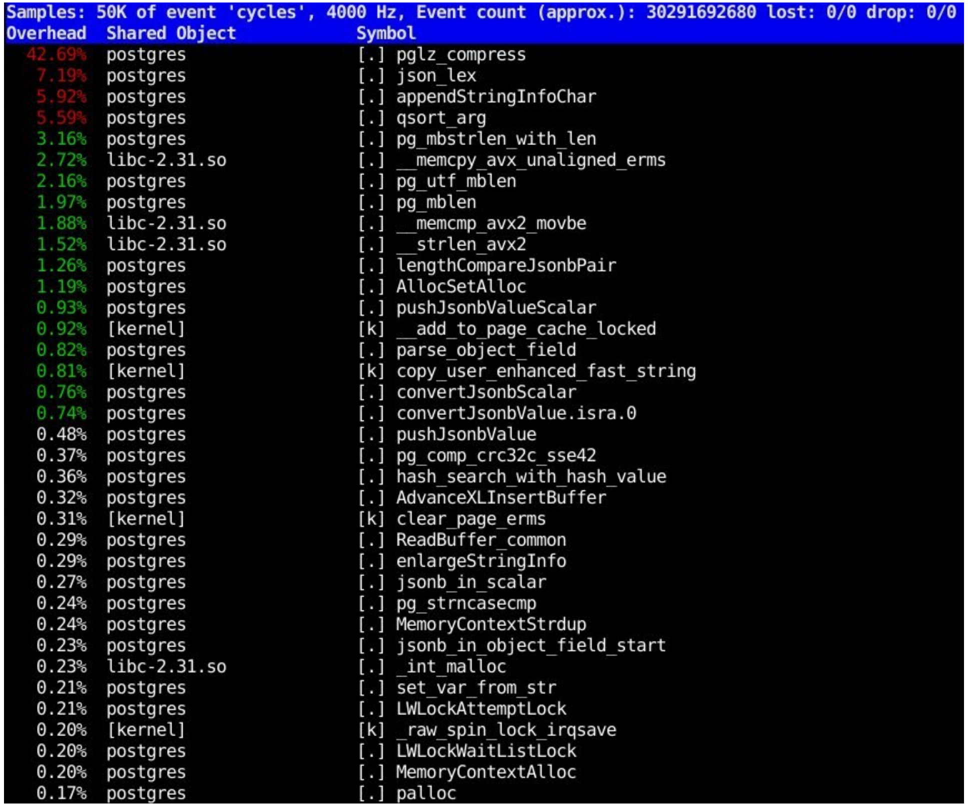
Становятся важны «малотоннажные» специализированные машины

с приходом в датацентры новых архитектур: ARM, RISC-V

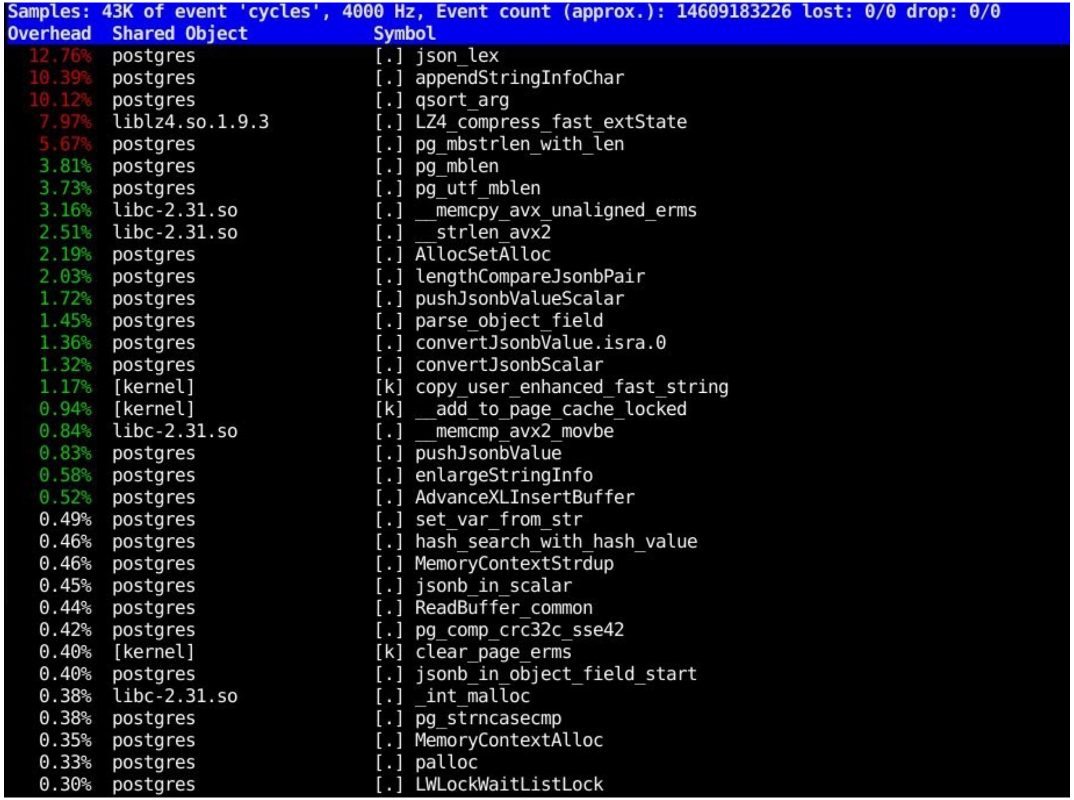
Storage-центричная акселерация

традиционных дисковых СУБД

Диагностика проблемы. Compress.



```
Samples: 50K of event 'cycles', 4000 Hz, Event count (approx.): 30291692680 lost: 0/0 drop: 0/0
Overhead  Shared Object               Symbol
  42.69%  postgres                    [.] pglz_compress
   7.19%  postgres                    [.] json_lex
   5.92%  postgres                    [.] appendStringInfoChar
   5.59%  postgres                    [.] qsort_arg
   3.16%  postgres                    [.] pg_mbstrlen_with_len
   2.72%  libc-2.31.so                [.] __memcpy_avx_unaligned_erms
   2.16%  postgres                    [.] pg_utf_mblen
   1.97%  postgres                    [.] pg_mblen
   1.88%  libc-2.31.so                [.] __memcmp_avx2_movbe
   1.52%  libc-2.31.so                [.] __strlen_avx2
   1.26%  postgres                    [.] lengthCompareJsonbPair
   1.19%  postgres                    [.] AllocSetAlloc
   0.93%  postgres                    [.] pushJsonbValueScalar
   0.92%  [kernel]                    [k] __add_to_page_cache_locked
   0.82%  postgres                    [.] parse_object_field
   0.81%  [kernel]                    [k] copy_user_enhanced_fast_string
   0.76%  postgres                    [.] convertJsonbScalar
   0.74%  postgres                    [.] convertJsonbValue.isra.0
   0.48%  postgres                    [.] pushJsonbValue
   0.37%  postgres                    [.] pg_comp_crc32c_sse42
   0.36%  postgres                    [.] hash_search_with_hash_value
   0.32%  postgres                    [.] AdvanceXLInsertBuffer
   0.31%  [kernel]                    [k] clear_page_erms
   0.29%  postgres                    [.] ReadBuffer_common
   0.29%  postgres                    [.] enlargeStringInfo
   0.27%  postgres                    [.] jsonb_in_scalar
   0.24%  postgres                    [.] pg_strncasecmp
   0.24%  postgres                    [.] MemoryContextStrdup
   0.23%  postgres                    [.] jsonb_in_object_field_start
   0.23%  libc-2.31.so                [.] _int_malloc
   0.21%  postgres                    [.] set_var_from_str
   0.21%  postgres                    [.] LWLockAttemptLock
   0.20%  [kernel]                    [k] _raw_spin_lock_irqsave
   0.20%  postgres                    [.] LWLockWaitListLock
   0.20%  postgres                    [.] MemoryContextAlloc
   0.17%  postgres                    [.] palloc
```



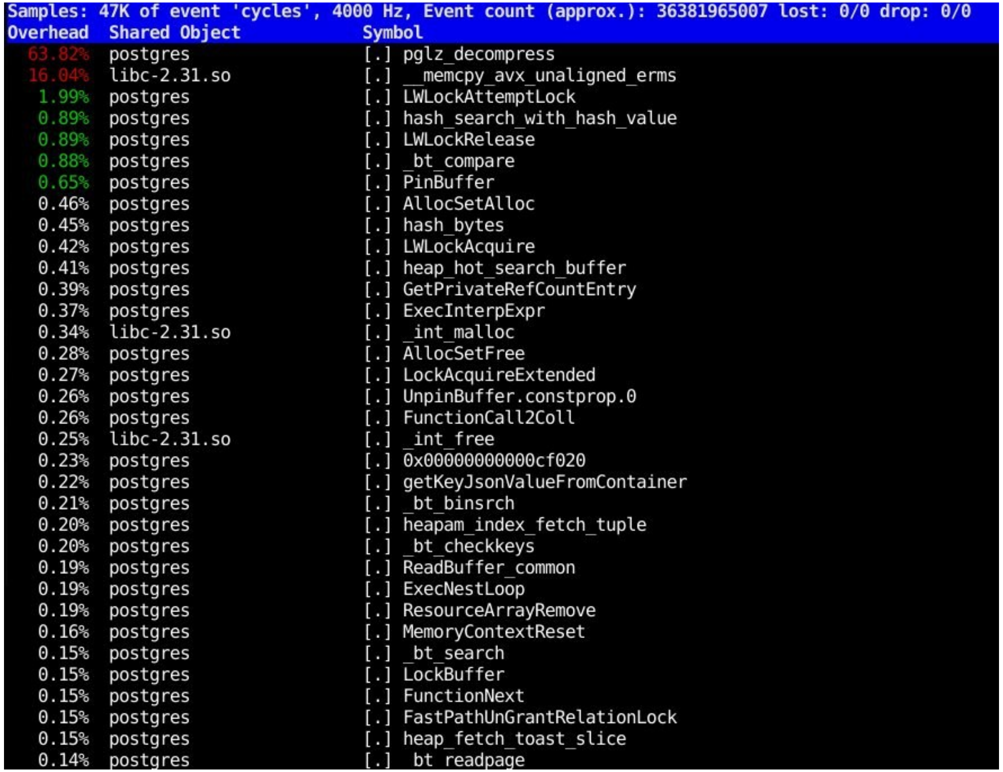
```
Samples: 43K of event 'cycles', 4000 Hz, Event count (approx.): 14609183226 lost: 0/0 drop: 0/0
Overhead  Shared Object               Symbol
  12.76%  postgres                    [.] json_lex
  10.39%  postgres                    [.] appendStringInfoChar
  10.12%  postgres                    [.] qsort_arg
   7.97%  liblz4.so.1.9.3             [.] LZ4_compress_fast_extState
   5.67%  postgres                    [.] pg_mbstrlen_with_len
   3.81%  postgres                    [.] pg_mblen
   3.73%  postgres                    [.] pg_utf_mblen
   3.16%  libc-2.31.so                [.] __memcpy_avx_unaligned_erms
   2.51%  libc-2.31.so                [.] __strlen_avx2
   2.19%  postgres                    [.] AllocSetAlloc
   2.03%  postgres                    [.] lengthCompareJsonbPair
   1.72%  postgres                    [.] pushJsonbValueScalar
   1.45%  postgres                    [.] parse_object_field
   1.36%  postgres                    [.] convertJsonbValue.isra.0
   1.32%  postgres                    [.] convertJsonbScalar
   1.17%  [kernel]                    [k] copy_user_enhanced_fast_string
   0.94%  [kernel]                    [k] __add_to_page_cache_locked
   0.84%  libc-2.31.so                [.] __memcmp_avx2_movbe
   0.83%  postgres                    [.] pushJsonbValue
   0.58%  postgres                    [.] enlargeStringInfo
   0.52%  postgres                    [.] AdvanceXLInsertBuffer
   0.49%  postgres                    [.] set_var_from_str
   0.46%  postgres                    [.] hash_search_with_hash_value
   0.46%  postgres                    [.] MemoryContextStrdup
   0.45%  postgres                    [.] jsonb_in_scalar
   0.44%  postgres                    [.] ReadBuffer_common
   0.42%  postgres                    [.] pg_comp_crc32c_sse42
   0.40%  [kernel]                    [k] clear_page_erms
   0.40%  postgres                    [.] jsonb_in_object_field_start
   0.38%  libc-2.31.so                [.] _int_malloc
   0.38%  postgres                    [.] pg_strncasecmp
   0.35%  postgres                    [.] MemoryContextAlloc
   0.33%  postgres                    [.] palloc
   0.30%  postgres                    [.] LWLockWaitListLock
```

pglz_compress 20s, **43%** CPU; LZ4_compress 13s, **13%** CPU;

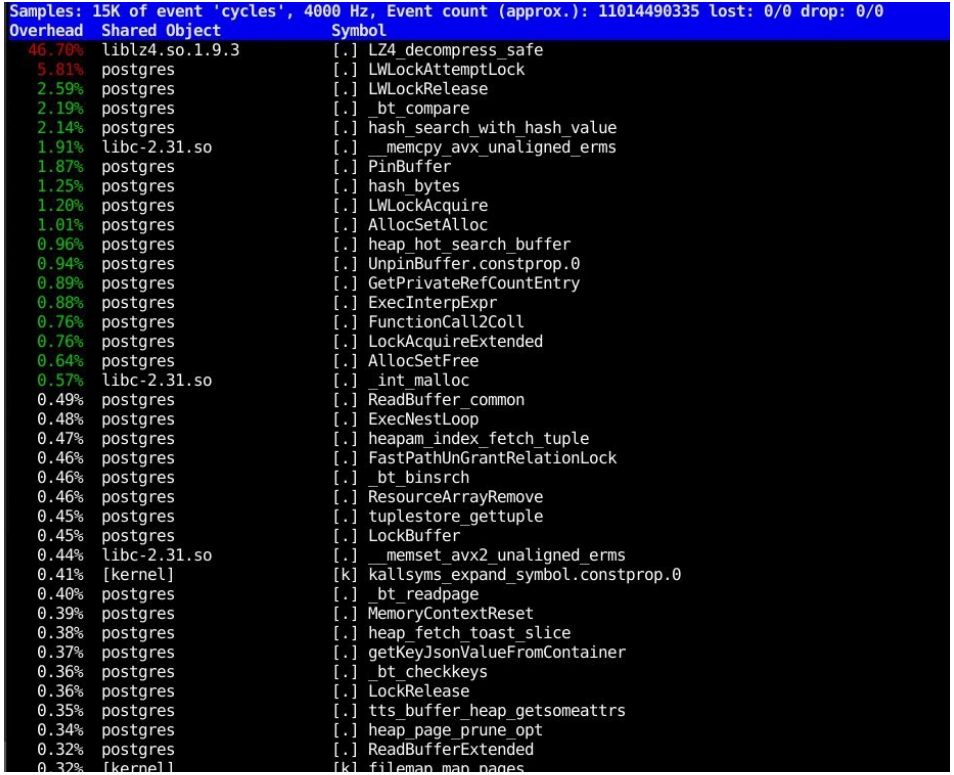
PostgreSQL 14.2 on x86_64-pc-linux-gnu,
compiled by gcc (Debian 10.2.1-6), 64-bit
Xeon 5220R 24C 2.2GHz DDR4 2666, NVMe SSD

INSERT into github_events_raw **SELECT** x::jsonb as val from unnest(string_to_array(pg_read_file(:'FILE'),chr(10))) as x where length(x) > 0;

Диагностика проблемы. Decompres.



```
Samples: 47K of event 'cycles', 4000 Hz, Event count (approx.): 36381965007 lost: 0/0 drop: 0/0
Overhead  Shared Object            Symbol
  63.82%  postgres                 [.] pglz_decompress
  16.04%  libc-2.31.so             [.] __memcpy_avx_unaligned_erms
   1.99%  postgres                 [.] LWLockAttemptLock
   0.89%  postgres                 [.] hash_search_with_hash_value
   0.89%  postgres                 [.] LWLockRelease
   0.88%  postgres                 [.] _bt_compare
   0.65%  postgres                 [.] PinBuffer
   0.46%  postgres                 [.] AllocSetAlloc
   0.45%  postgres                 [.] hash_bytes
   0.42%  postgres                 [.] LWLockAcquire
   0.41%  postgres                 [.] heap_hot_search_buffer
   0.39%  postgres                 [.] GetPrivateRefCountEntry
   0.37%  postgres                 [.] ExecInterpExpr
   0.34%  libc-2.31.so             [.] _int_malloc
   0.28%  postgres                 [.] AllocSetFree
   0.27%  postgres                 [.] LockAcquireExtended
   0.26%  postgres                 [.] UnpinBuffer.constprop.0
   0.26%  postgres                 [.] FunctionCall2Coll
   0.25%  libc-2.31.so             [.] _int_free
   0.23%  postgres                 [.] 0x00000000000cf020
   0.22%  postgres                 [.] getKeyJsonValueFromContainer
   0.21%  postgres                 [.] _bt_binsrch
   0.20%  postgres                 [.] heapam_index_fetch_tuple
   0.20%  postgres                 [.] _bt_checkkeys
   0.19%  postgres                 [.] ReadBuffer_common
   0.19%  postgres                 [.] ExecNestLoop
   0.19%  postgres                 [.] ResourceArrayRemove
   0.16%  postgres                 [.] MemoryContextReset
   0.15%  postgres                 [.] _bt_search
   0.15%  postgres                 [.] LockBuffer
   0.15%  postgres                 [.] FunctionNext
   0.15%  postgres                 [.] FastPathUnGrantRelationLock
   0.15%  postgres                 [.] heap_fetch_toast_slice
   0.14%  postgres                 [.] _bt_readpage
```



```
Samples: 15K of event 'cycles', 4000 Hz, Event count (approx.): 11014490335 lost: 0/0 drop: 0/0
Overhead  Shared Object            Symbol
  46.70%  liblz4.so.1.9.3          [.] LZ4_decompress_safe
   5.81%  postgres                 [.] LWLockAttemptLock
   2.59%  postgres                 [.] LWLockRelease
   2.19%  postgres                 [.] _bt_compare
   2.14%  postgres                 [.] hash_search_with_hash_value
   1.91%  libc-2.31.so             [.] __memcpy_avx_unaligned_erms
   1.87%  postgres                 [.] PinBuffer
   1.25%  postgres                 [.] hash_bytes
   1.20%  postgres                 [.] LWLockAcquire
   1.01%  postgres                 [.] AllocSetAlloc
   0.96%  postgres                 [.] heap_hot_search_buffer
   0.94%  postgres                 [.] UnpinBuffer.constprop.0
   0.89%  postgres                 [.] GetPrivateRefCountEntry
   0.88%  postgres                 [.] ExecInterpExpr
   0.76%  postgres                 [.] FunctionCall2Coll
   0.76%  postgres                 [.] LockAcquireExtended
   0.64%  postgres                 [.] AllocSetFree
   0.57%  libc-2.31.so             [.] _int_malloc
   0.49%  postgres                 [.] ReadBuffer_common
   0.48%  postgres                 [.] ExecNestLoop
   0.47%  postgres                 [.] heapam_index_fetch_tuple
   0.46%  postgres                 [.] FastPathUnGrantRelationLock
   0.46%  postgres                 [.] _bt_binsrch
   0.46%  postgres                 [.] ResourceArrayRemove
   0.45%  postgres                 [.] tuplestore_gettuple
   0.45%  postgres                 [.] LockBuffer
   0.44%  libc-2.31.so             [.] __memset_avx2_unaligned_erms
   0.41%  [kernel]                 [k] kallsyms_expand_symbol.constprop.0
   0.40%  postgres                 [.] _bt_readpage
   0.39%  postgres                 [.] MemoryContextReset
   0.38%  postgres                 [.] heap_fetch_toast_slice
   0.37%  postgres                 [.] getKeyJsonValueFromContainer
   0.36%  postgres                 [.] _bt_checkkeys
   0.36%  postgres                 [.] LockRelease
   0.35%  postgres                 [.] tts_buffer_heap_getsomeattrs
   0.34%  postgres                 [.] heap_page_prune_opt
   0.32%  postgres                 [.] ReadBufferExtended
   0.32%  [kernel]                 [k] filemap_map_pages
```

pglz_decompress 7.3s, **64%** CPU; LZ4_decompress 2.3s, **49%** CPU;

PostgreSQL 14.2 on x86_64-pc-linux-gnu,
compiled by gcc
(Debian 10.2.1-6) 10.2.1 20210110, 64-bit

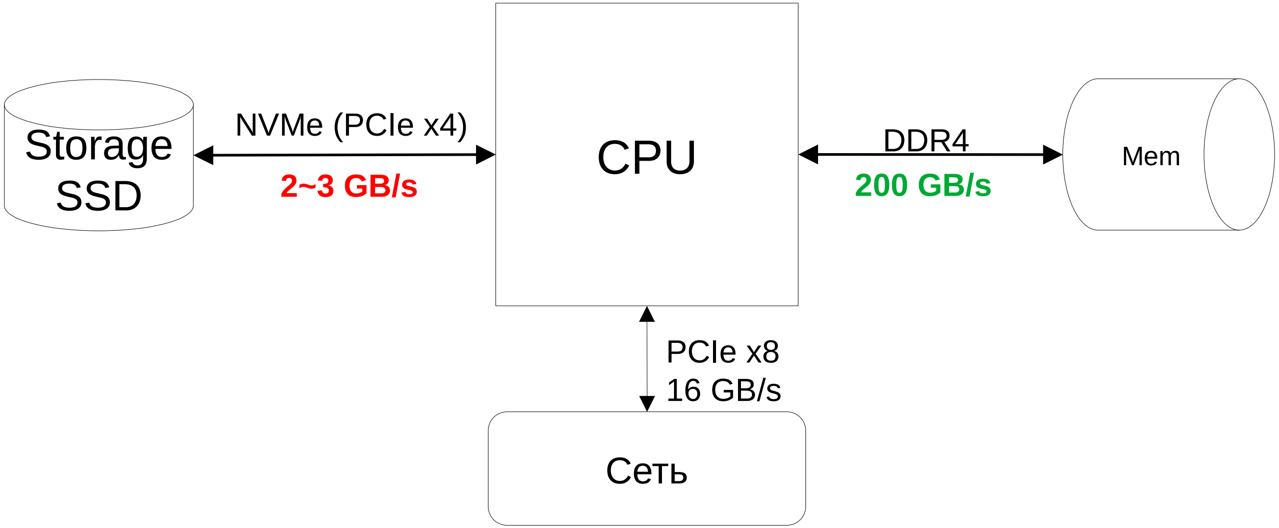
SELECT count(x->>'ref') **FROM** github_events_raw x, generate_series(1,10) y;

Xeon 5220R 24C 2.2GHz DDR4 2666, Samsung 970 NVMe SSD

История аппаратного ускорения PostgreSQL

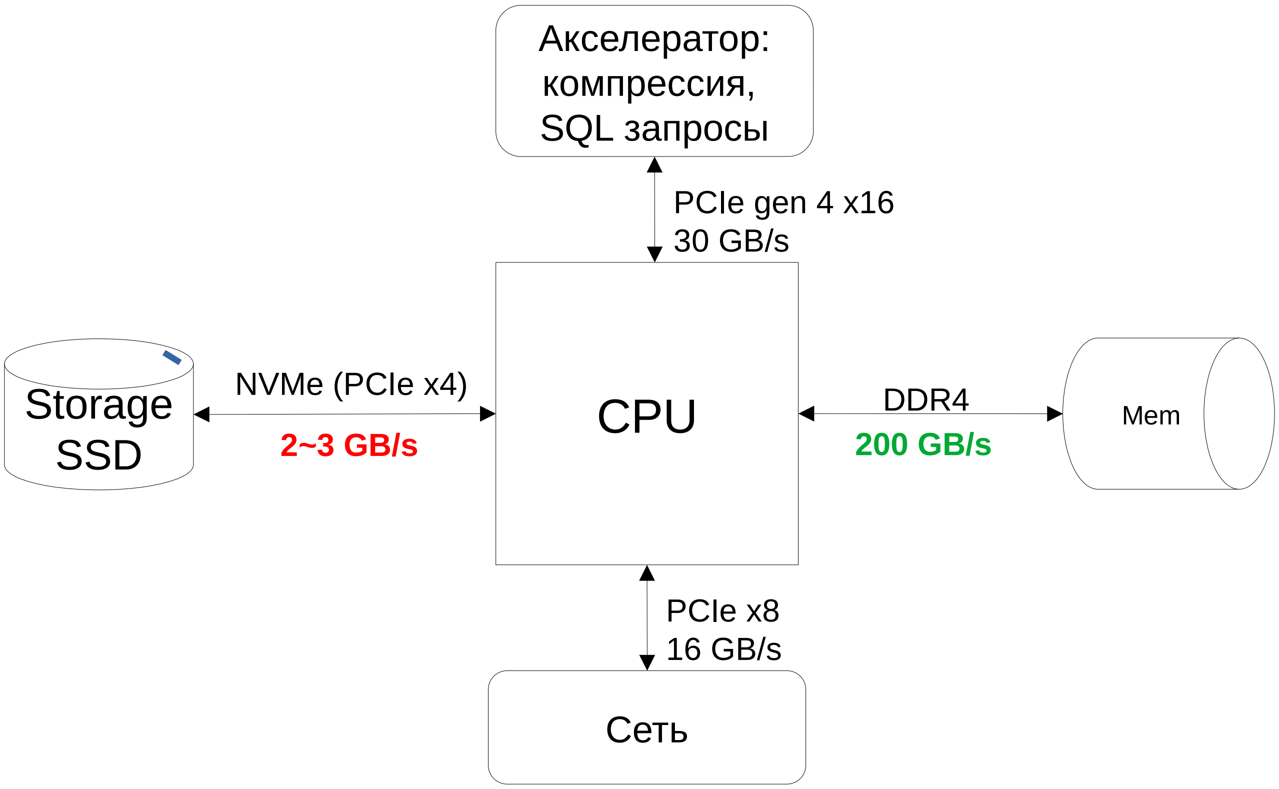
- **Netezza -** ускорители на FPGA, куплена IBM за $1.7B в 2010, 500 клиентов
- **Swarm64** - FPGA+Pmem, куплена ServiceNow в 2021
  - выступали на PGConf.Moscow 2020: https://pgconf.ru/2020/273250)
- **Memhive** — Pmem, переписана IO часть под PMem
  - выступали на PGConf.Moscow 2021: https://pgconf.ru/2021/290403)
- **PG-Strom** — GPU, добавили поддержку NVIDIA GPUDirect Storage

Обычный сервер PostgreSQL



Ключевые ограничения — хранилище и сжатие/шифр

CPU-центричная акселлерация

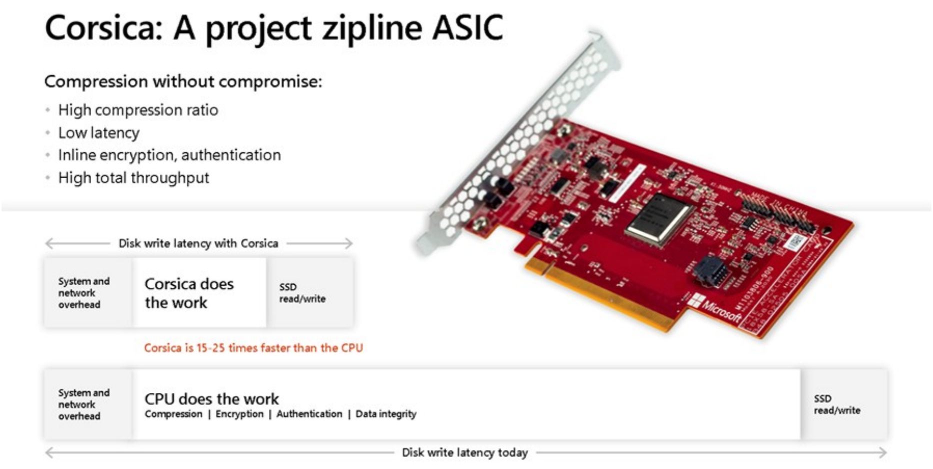


Освобождает ресурсы CPU, но не устраняет избыточного траффика

Примеры аппаратуры

- FPGA AIC, GPU
- Xilinx SQL Accelerator[1]

- Microsoft Project Corsica[2]



Corsica: A project zipline ASIC

Compression without compromise:

- High compression ratio
- Low latency
- Inline encryption, authentication
- High total throughput

Disk write latency with Corsica

System and network overhead	Corsica does the work	SSD read/write

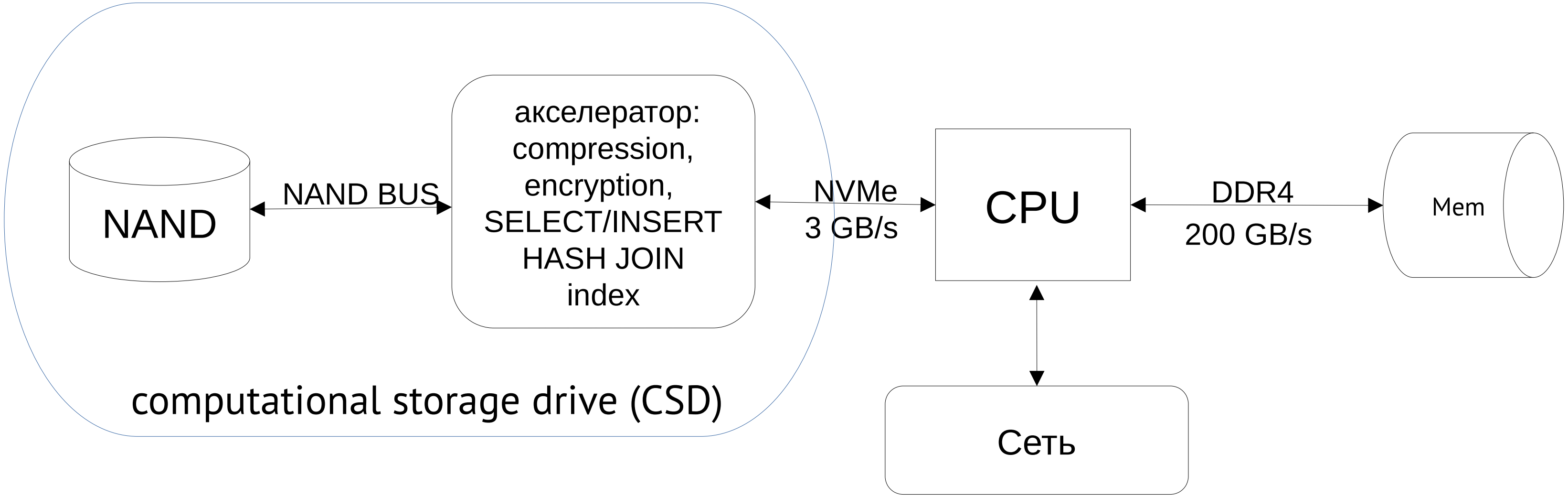
Corsica is 15-25 times faster than the CPU

System and network overhead	CPU does the work Compression | Encryption | Authentication | Data integrity	SSD read/write

Disk write latency today

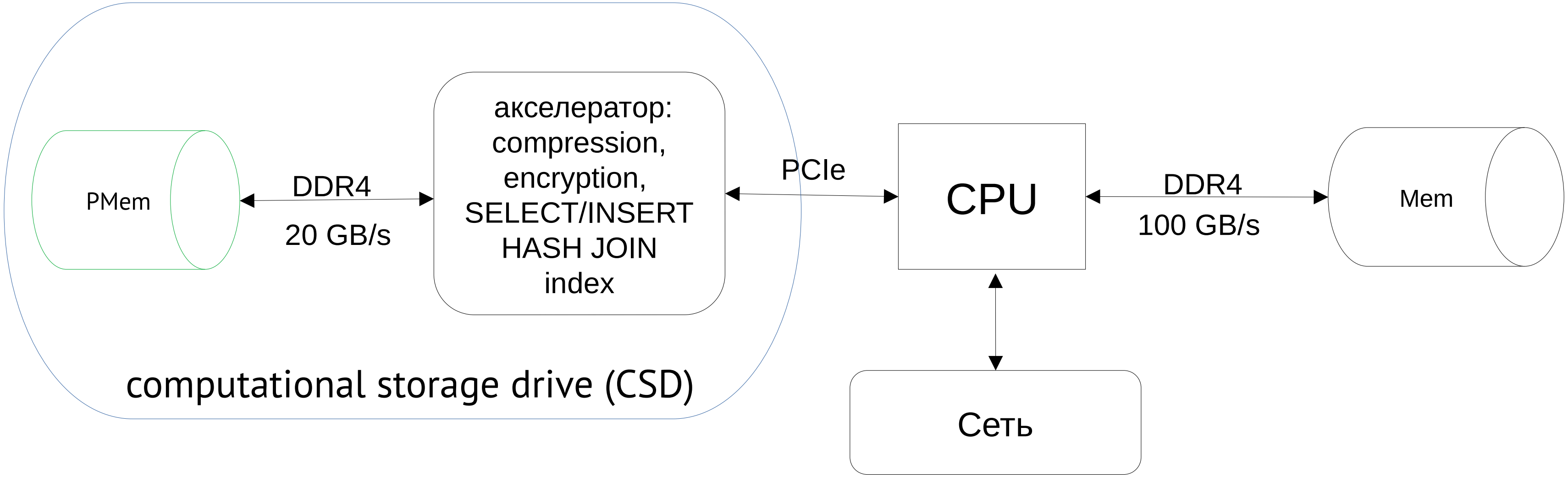
[1] https://github.com/Xilinx/data-analytics
[2] https://azure.microsoft.com/fr-fr/blog/improved-cloud-service-performance-through-asic-acceleration/

Storage-центричная акселлерация


NAND
NAND BUS
акселератор: compression, encryption, SELECT/INSERT HASH JOIN index
computational storage drive (CSD)
NVMe
3 GB/s
CPU
DDR4
200 GB/s
Mem
Сеть

Освобождает ресурсы CPU и уменьшает внутрисистемный траффик

Storage-центричная акселлерация с PMem



Освобождает ресурсы CPU и ускоряет хранилище

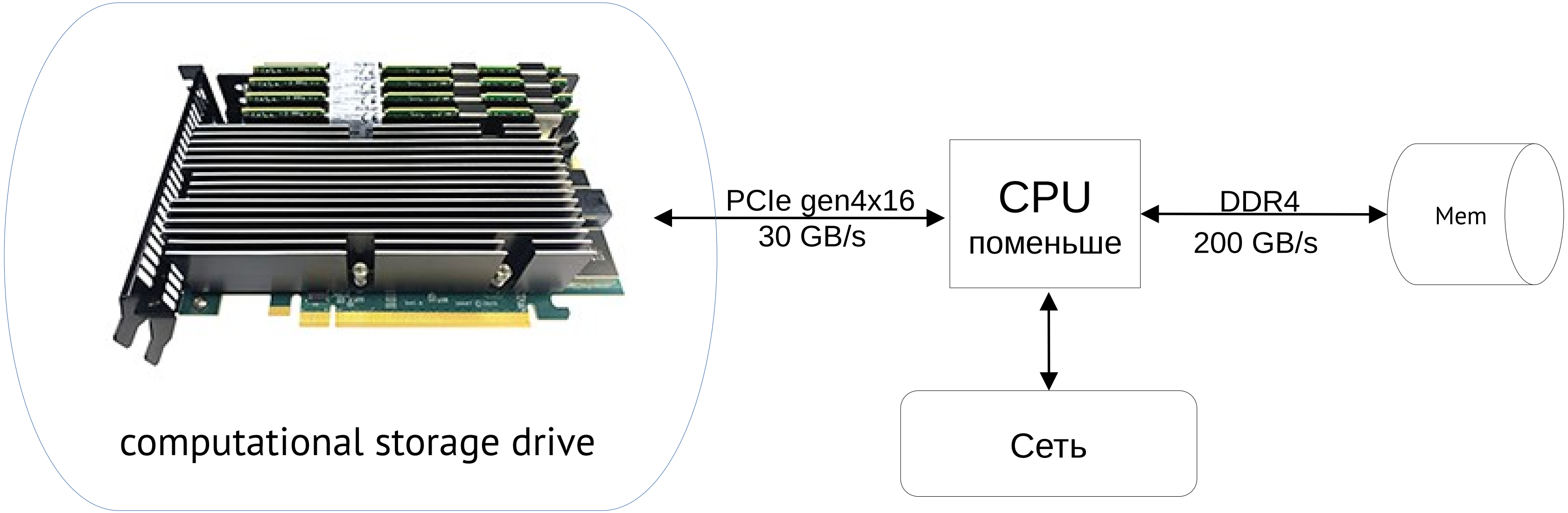
Примеры аппаратуры

- Samsung SmartSSD®[1]

Kestral™ PCIe Optane AIC[2]

[1] https://www.xilinx.com/applications/data-center/computational-storage/smartssd.html
[2] https://www.smartm.com/product/advanced-memory/AIC

Pmem+FPGA = SuperCSD



computational storage drive

PCIe gen4x16

30 GB/s

CPU

поменьше

DDR4

200 GB/s

Mem

Сеть

https://www.smartm.com/product/advanced-memory/AIC

Направления развития

- Декомпозиция функций СУБД и перенос регулярной нагрузки с CPU
- Использование свойств локальности данных, обработка в устройствах хранения
- Новая физика хранения с побайтной адресацией и быстрым доступом — Pmem
- Активное продвижение архитектуры RISC-V в устройствах хранения (WD, Seagate)
- Создание Domain Specific Accelerators (DSA) SoC — тренд в больших облаках