Developments on the ALFA Parameter manager

Charis Kouzinopoulos, CERN

ALICE Offline Week 01.04.2016

Motivation

A new Parameter manager is under development for ALFA - the Condition and Calibration Data Base (CCDB)

Intended as a replacement for the Offline Conditions Database (OCDB) that is currently (Run 1 and 2) used to store calibration and alignment data

Why a new solution – requirements for the new DB

- Scalability - High number of processes/growing size of the DB
- Low latency - Faster storage/retrieval
- Redundancy - Data loss avoidance

OCDB is not a database – it is a set of entries in the AliEn file catalog that point to the ROOT files stored in the Grid that contain the actual data

Parameter Manager

A continuation of the work that started by Tom Van Steenkiste Tom Van Steenkiste 2016

An abstraction layer with different modules: data loading, serialization and DB communication

The modules offer a simple interface:

Data loading	*Serialization*	*DB communication*
Load object Create object	Serialize De-serialize	PUT GET
OCDB AliCDBEntry ROOT files	*messages*	*key/value pairs*

Additional modules are considered!

protobuf Protocol Buffers

riak

Parameter Manager

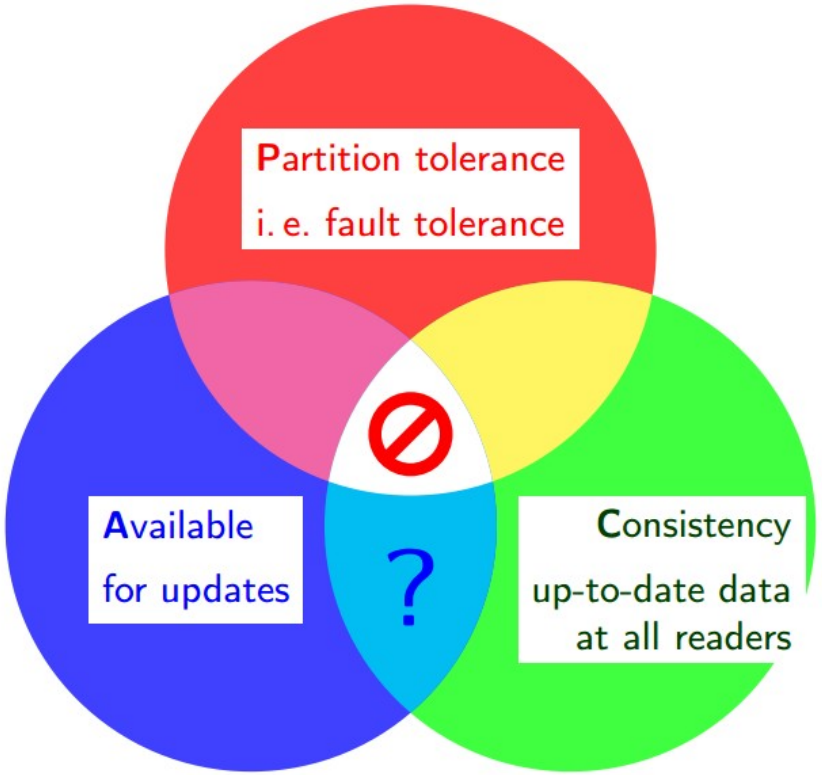
Why Riak?

A popular key/value *distributed* database

Offers data availability: distributes data across multiple nodes

It handles synchronization between the nodes internally in a transparent way

It is an *eventually consistent* system – in a failure scenario data *can* be available but *potentially* not up to date



A distributed storage system can have at most two out of three desirable properties

See Jakob Blomer's primer on key/value databases

Jakob Blomer 2015

Parameter Manager

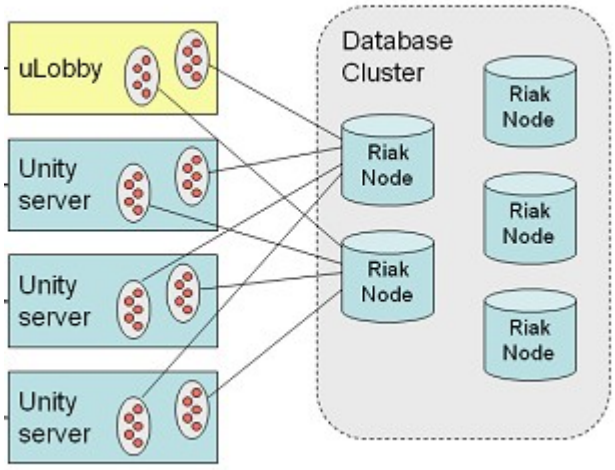
Data organization

Riak organizes data into Buckets, Keys and Values

Values are identified by a unique Key

Each Key/Value pair is stored in a bucket

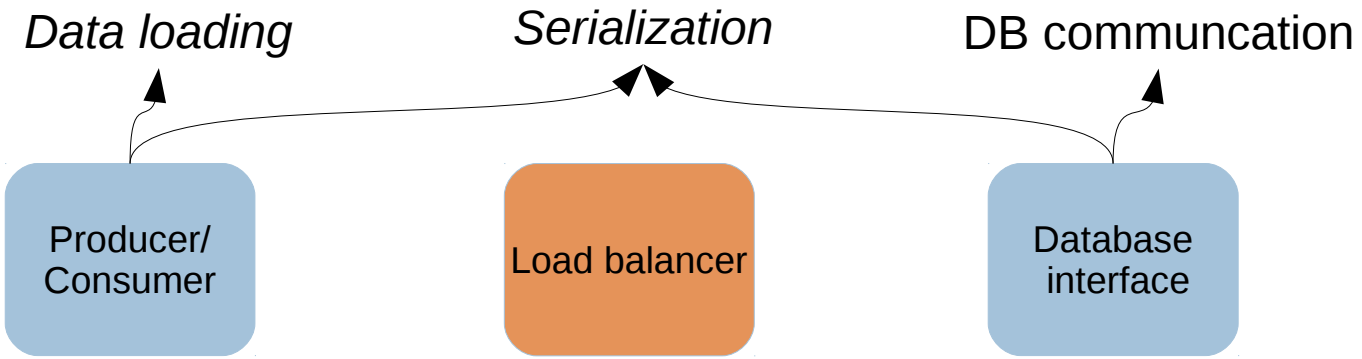
Buckets offer a flat namespace - allow multiple keys with the same name to exist in the database and some per bucket configurability

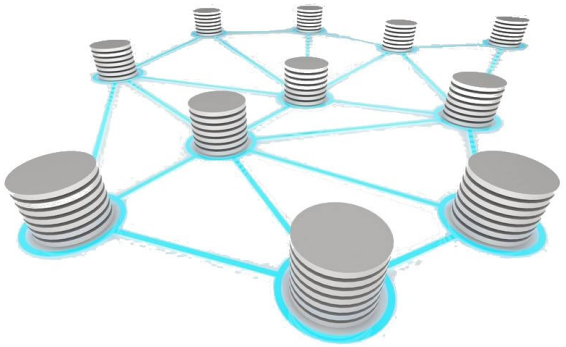


See Jakob Blomer's primer on key/value databases Jakob Blomer 2015

System design

Components of the Parameter Manager:

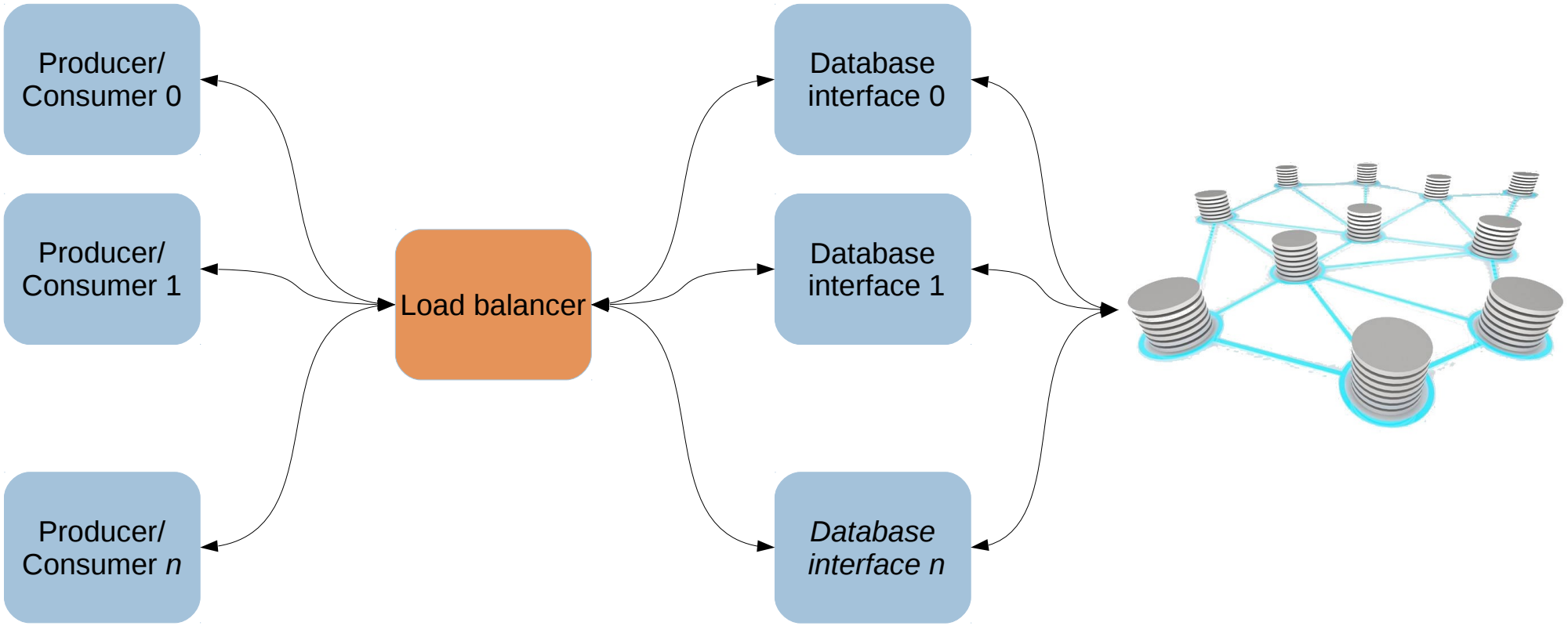




C++ repository: https://github.com/kouzinopoulos/KeyValueClusterPerf
Java repository: https://github.com/kouzinopoulos/RiakJavaC

System design

Processes topology of the Parameter Manager:



Data flow

Producer/
Consumer 0

Producer/
Consumer 1

Producer/
Consumer *n*

Loading of ROOT files containing AliCDBEntry objects

Streaming the objects to memory

Storing the object path

/2011/OCDB/TPC/Calib/TimeDrift/

- Run1_10_v0_s0.root
- Run1_10_v0_s1.root
- Run1_10_v0_s2.root

Data flow

Producer/
Consumer 0

Producer/
Consumer 1

Producer/
Consumer *n*

value

Loading of ROOT files containing AliCDBEntry objects

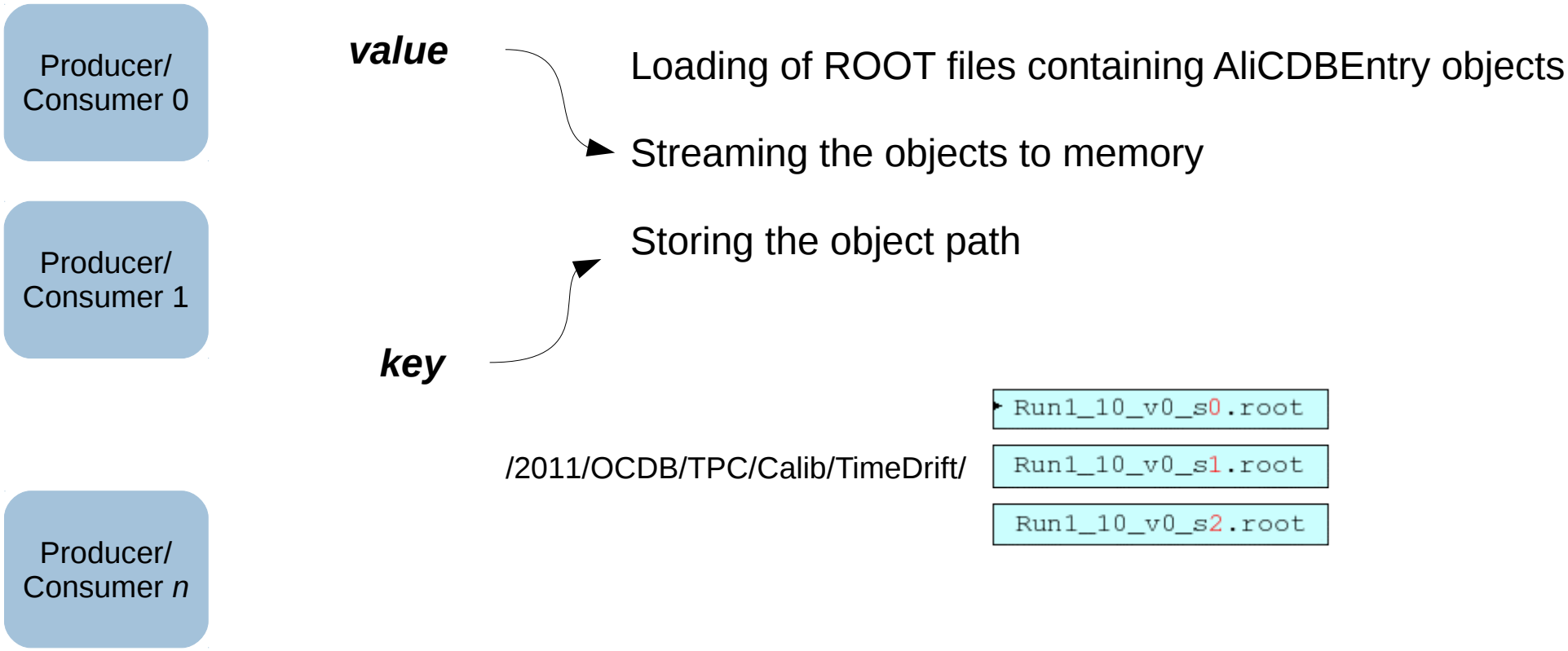
Streaming the objects to memory

Storing the object path

/2011/OCDB/TPC/Calib/TimeDrift/

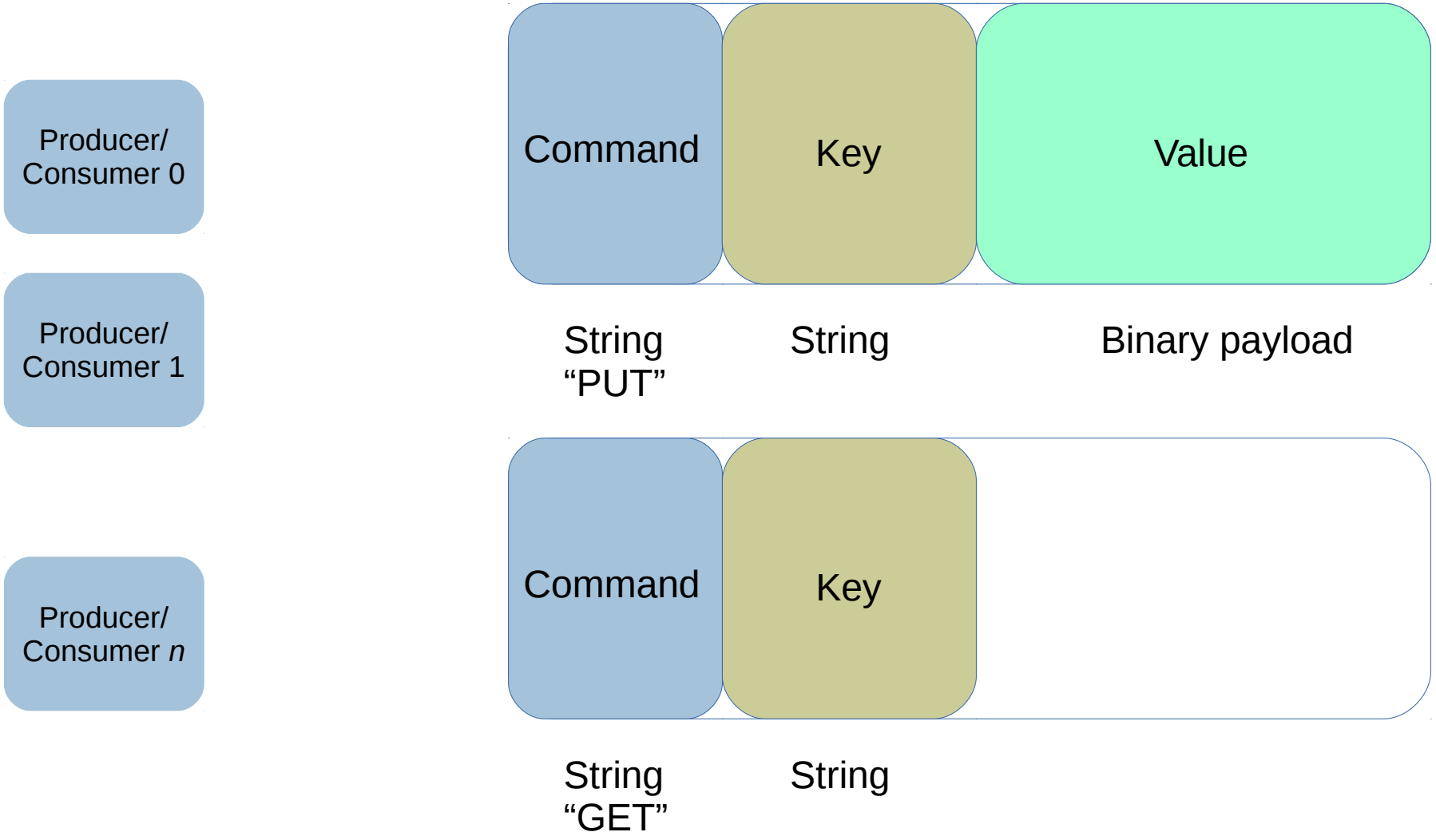
- Run1_10_v0_s0.root
- Run1_10_v0_s1.root
- Run1_10_v0_s2.root

Data flow

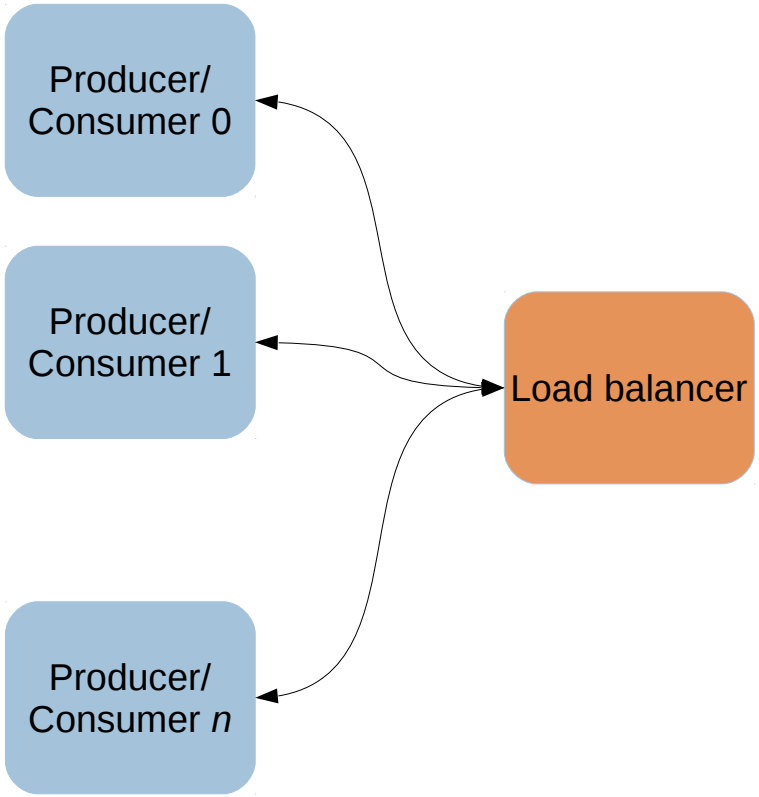

Producer/
Consumer 0
Producer/
Consumer 1
Producer/
Consumer *n*
value
Loading of ROOT files containing AliCDBEntry objects
Streaming the objects to memory
Storing the object path
key
/2011/OCDB/TPC/Calib/TimeDrift/
Run1_10_v0_s0.root
Run1_10_v0_s1.root
Run1_10_v0_s2.root

Data flow

Message crafting using the serialization module:



Data flow



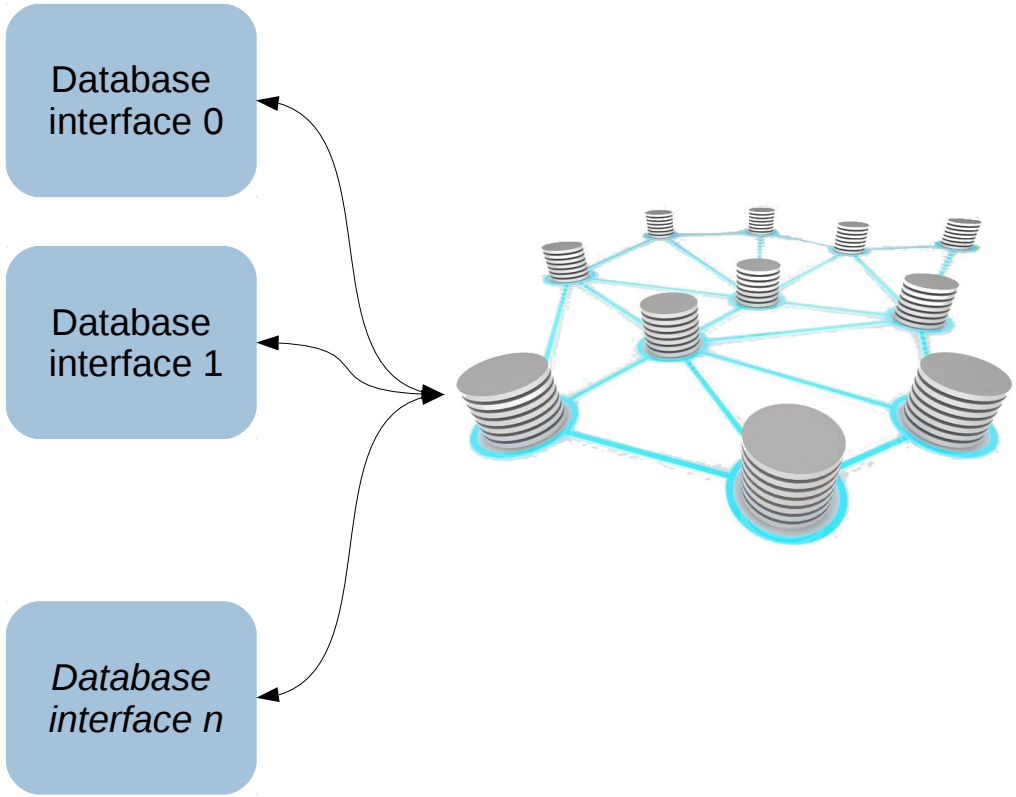
The serialized message is transmitted to the Load balancer node using ZeroMQ (FairMQ)

The Load balancer node forwards the message to a Database interface node on a round robin basis

Data flow

The message is de-serialized using Protocol buffers

A “PUT” or “GET” command is executed to the Riak cluster

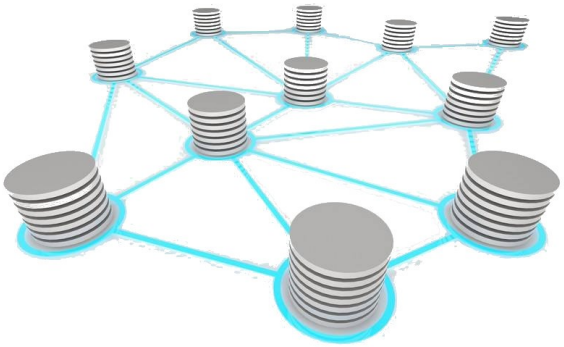


Preliminary results

Hardware of the O^2 test facility:

aido2db DB server	
CPU	Intel Xeon E5-2623 @3GHz 16 cores
Memory	64GB
Disc	Seagate ST91000640NS 1TB SATA 6GB/sec 64MB cache
Network	Mellanox MT27520 40Gb/s adapter

4x aido2qc4 nodes	
CPU	Intel Xeon E5-2640 @2.6GHz 32 cores
Memory	128GB
Disc	Seagate ST9500620NS 500GB SATA 6GB/sec 64MB cache
Network	Mellanox MT27520 40Gb/s adapter



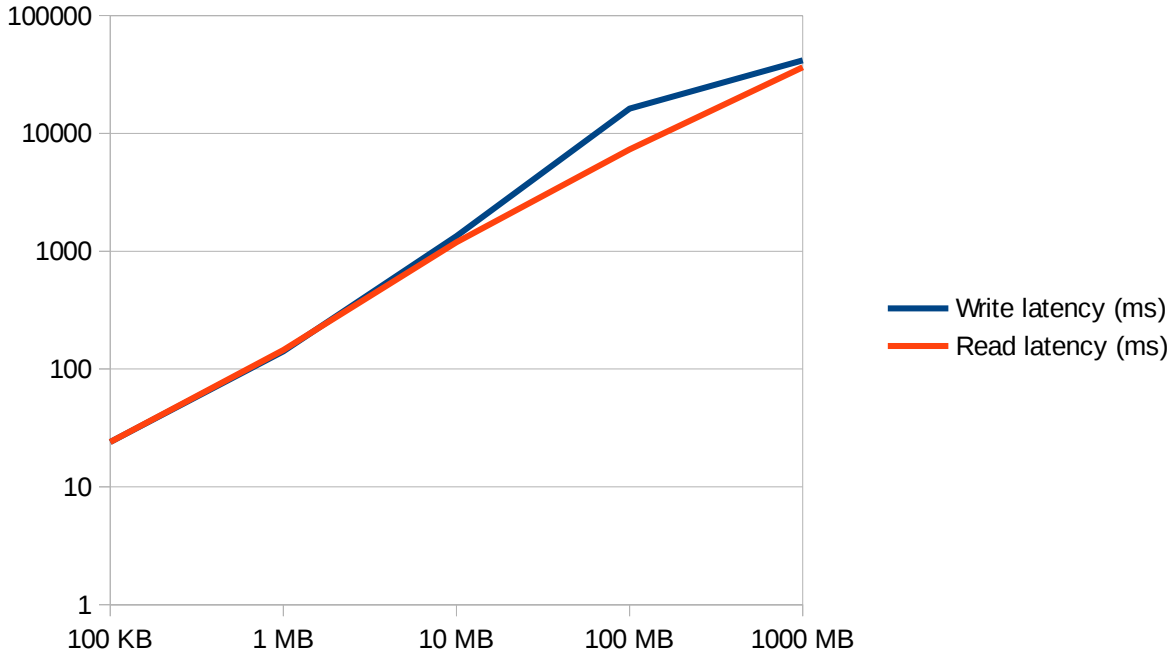
Preliminary results

Data set:

- /2011/OCDB/TPC/Calib/
- 7199 ROOT files
- 574 MB *compressed* size
- 1170 MB *uncompressed* size

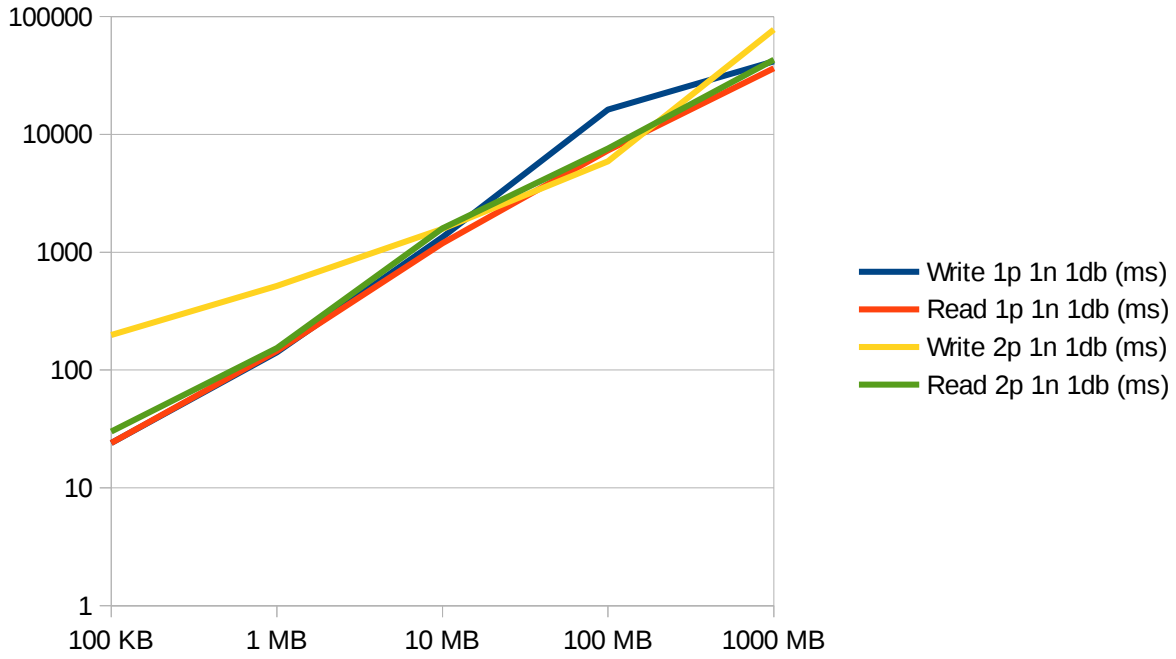
Subdirectory	Number of ROOT objects	Compressed size
Raw	1715	448 MB
RecoParam	14	70 KB
Temperature	2314	72 MB
TimeDrift	1648	33 MB
TimeGain	1503	19 MB

Preliminary results



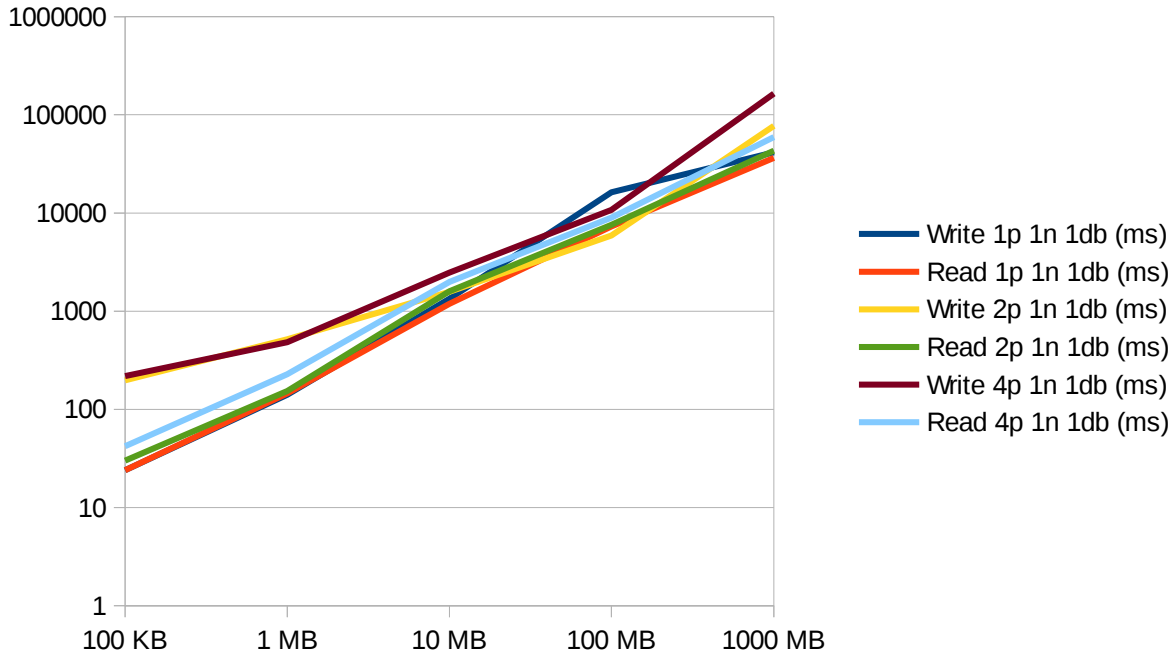
Size	Write 1p 1n 1db (ms)	Read 1p 1n 1db (ms)	Number of ROOT objects
100 KB	24	24	6
1 MB	141	145	40
10 MB	1343	1191	543
100 MB	16226	7324	1676
1000 MB	41676	36436	5230

Preliminary results



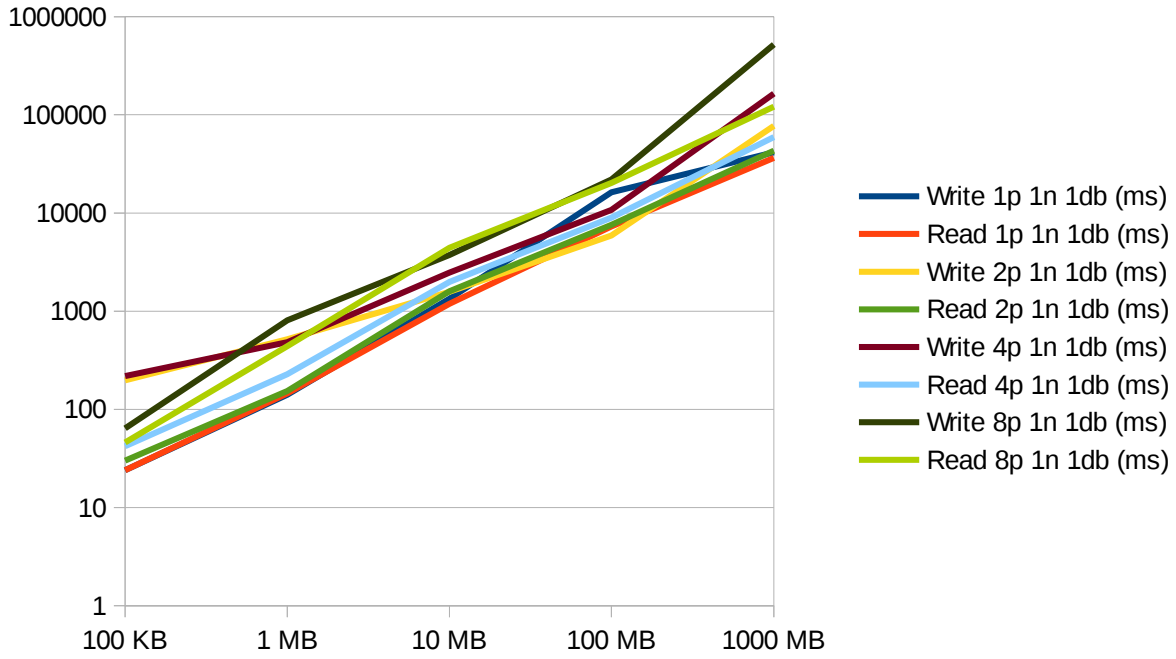
Size	Write 1p 1n 1db (ms)	Read 1p 1n 1db (ms)	Write 2p 1n 1db (ms)	Read 2p 1n 1db (ms)
100 KB	24	24	198	30
1 MB	141	145	519	154
10 MB	1343	1191	1580	1598
100 MB	16226	7324	5907	7575
1000 MB	41676	36436	77411	43016

Preliminary results



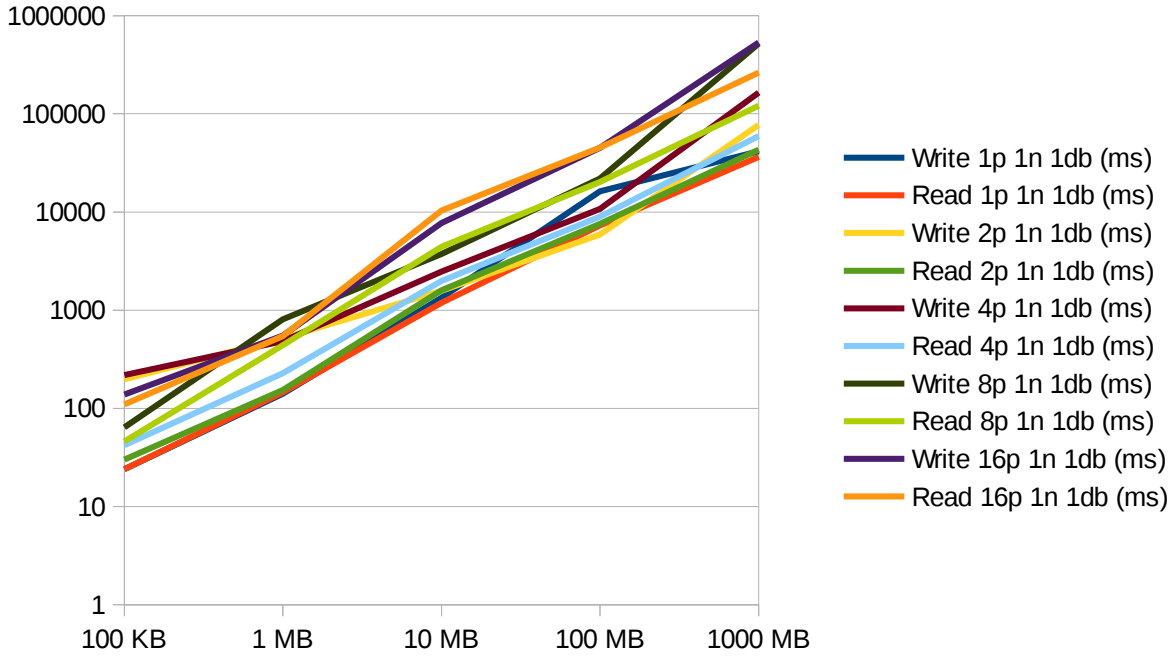
Size	Write 1p 1n 1db (ms)	Read 1p 1n 1db (ms)	Write 2p 1n 1db (ms)	Read 2p 1n 1db (ms)	Write 4p 1n 1db (ms)	Read 4p 1n 1db (ms)
100 KB	24	24	198	30	218	42
1 MB	141	145	519	154	484	228
10 MB	1343	1191	1580	1598	2480	1989
100 MB	16226	7324	5907	7575	10733	8970
1000 MB	41676	36436	77411	43016	164496	59380

Preliminary results



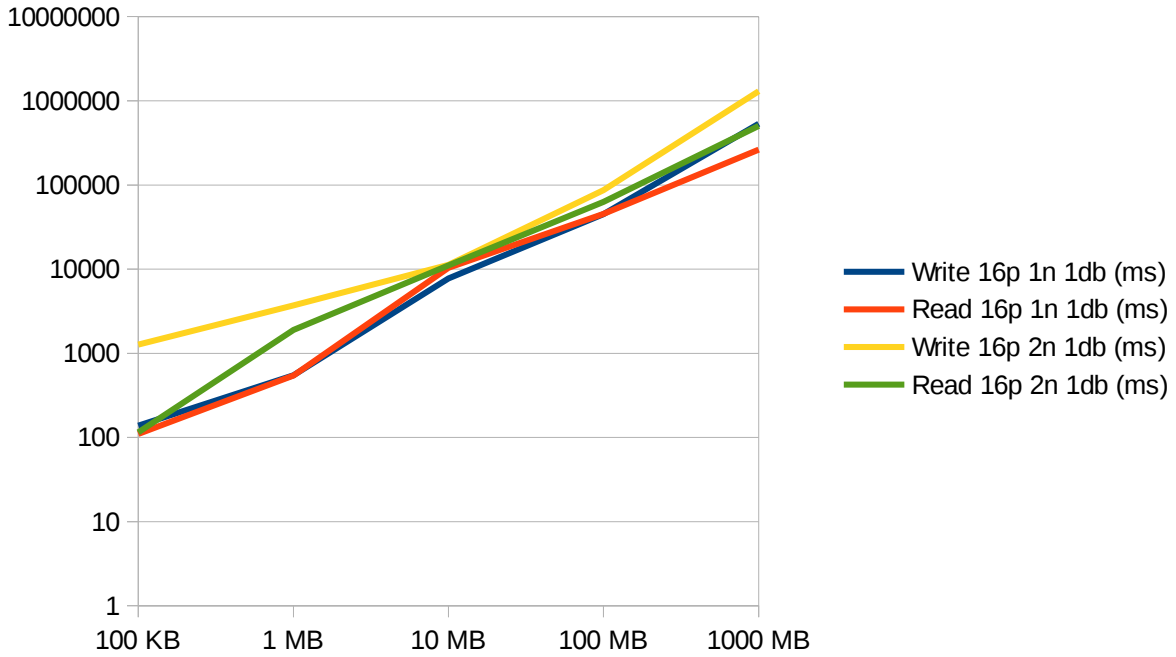
Size	Write 1p 1n 1db (ms)	Read 1p 1n 1db (ms)	Write 2p 1n 1db (ms)	Read 2p 1n 1db (ms)	Write 4p 1n 1db (ms)	Read 4p 1n 1db (ms)	Write 8p 1n 1db (ms)	Read 8p 1n 1db (ms)
100 KB	24	24	198	30	218	42	64	46
1 MB	141	145	519	154	484	228	808	440
10 MB	1343	1191	1580	1598	2480	1989	3744	4422
100 MB	16226	7324	5907	7575	10733	8970	21945	20228
1000 MB	41676	36436	77411	43016	164496	59380	520835	121399

Preliminary results



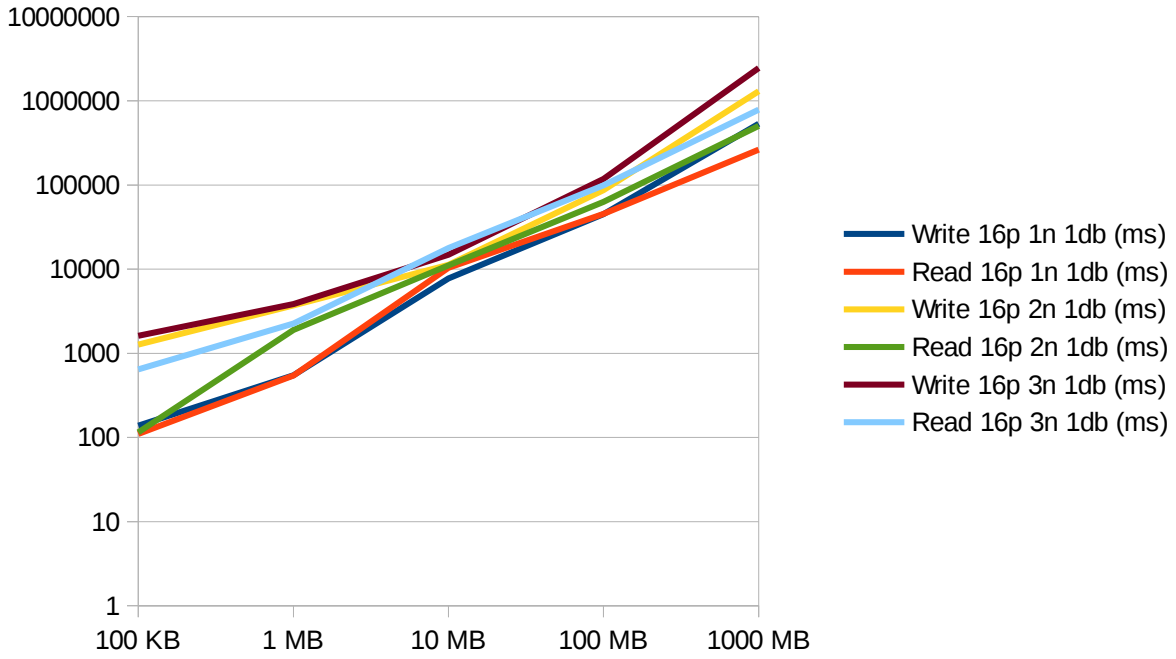
Size	Write 1p 1n 1db (ms)	Read 1p 1n 1db (ms)	Write 2p 1n 1db (ms)	Read 2p 1n 1db (ms)	Write 4p 1n 1db (ms)	Read 4p 1n 1db (ms)	Write 8p 1n 1db (ms)	Read 8p 1n 1db (ms)	Write 16p 1n 1db (ms)	Read 16p 1n 1db (ms)
100 KB	24	24	198	30	218	42	64	46	138	110
1 MB	141	145	519	154	484	228	808	440	549	545
10 MB	1343	1191	1580	1598	2480	1989	3744	4422	7745	10381
100 MB	16226	7324	5907	7575	10733	8970	21945	20228	45024	45475
1000 MB	41676	36436	77411	43016	164496	59380	520835	121399	532899	262491

Preliminary results



Size	Write 16p 1n 1db (ms)	Read 16p 1n 1db (ms)	Write 16p 2n 1db (ms)	Read 16p 2n 1db (ms)
100 KB	138	110	1268	115
1 MB	549	545	3722	1892
10 MB	7745	10381	11268	11086
100 MB	45024	45475	86935	62905
1000 MB	532899	262491	1300222	505033

Preliminary results



Size	Write 16p 1n 1db (ms)	Read 16p 1n 1db (ms)	Write 16p 2n 1db (ms)	Read 16p 2n 1db (ms)	Write 16p 3n 1db (ms)	Read 16p 3n 1db (ms)
100 KB	138	110	1268	115	1608	642
1 MB	549	545	3722	1892	3847	2267
10 MB	7745	10381	11268	11086	14836	17853
100 MB	45024	45475	86935	62905	118150	98857
1000 MB	532899	262491	1300222	505033	2448355	784717

Questions?