Intro

- I work for IT databases group at CERN

- We are involved in running Hadoop service
- My responsibilities include
  - Exploring potential service evolution
  - Providing consultancy to the users
  - Day-to-day administration

Motivations for using VMs

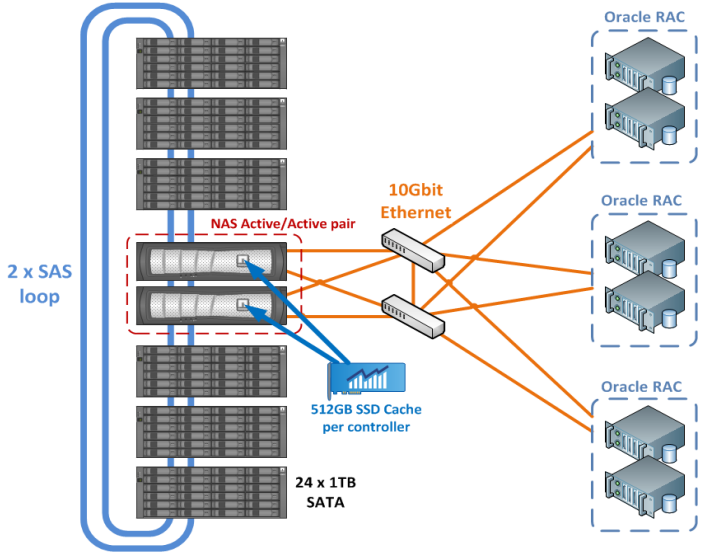
- Easy scaling up or down of the service
- Unified procurement procedure
- Clusters with configurations optimized for individual needs
- Potential user separation

Generic VM characteristics

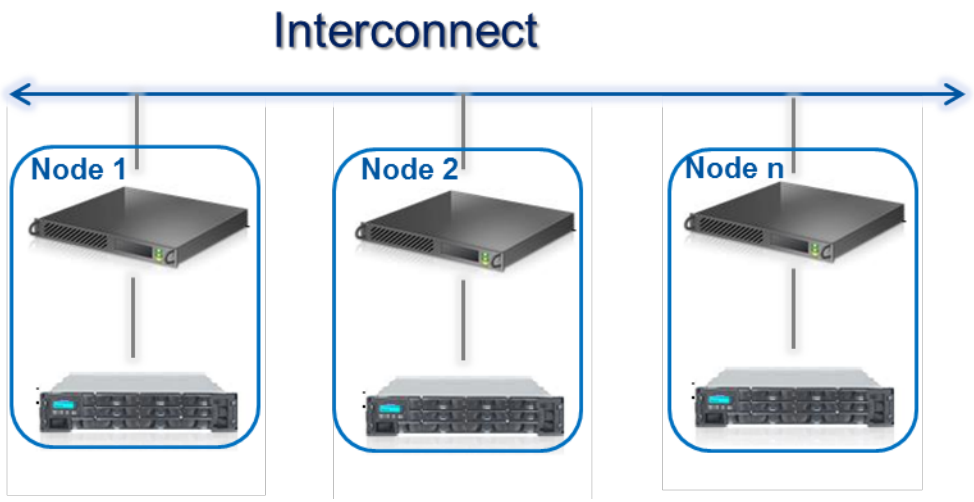
- Fast CPU and memory
- SSDs available

- More constraints on storage,
  we have to choose between
  - Local disks shared with other VMs running on the same hypervisor
  - Network attached volumes
  - External HDFS installation

What about data locality and emphasis on high throughput?



shared storage



shared nothing

Particular, but popular, applications

- For some workloads, distributed processing is more desired than data locality
  - Analytics (Spark, Impala, MapReduce)
  - NoSQL databases (HBase)
  - Indexing engines (Solr)

- Data scanning is CPU intensive
  - Parsing data while reading / writing
  - Data decoding and decompression

Benchmarks

- Sequential reads
  - Custom queries on Atlas Event Index data
  - Impala – SQL query engine for Hadoop
  - Stressing throughput
  - Smart data scanning

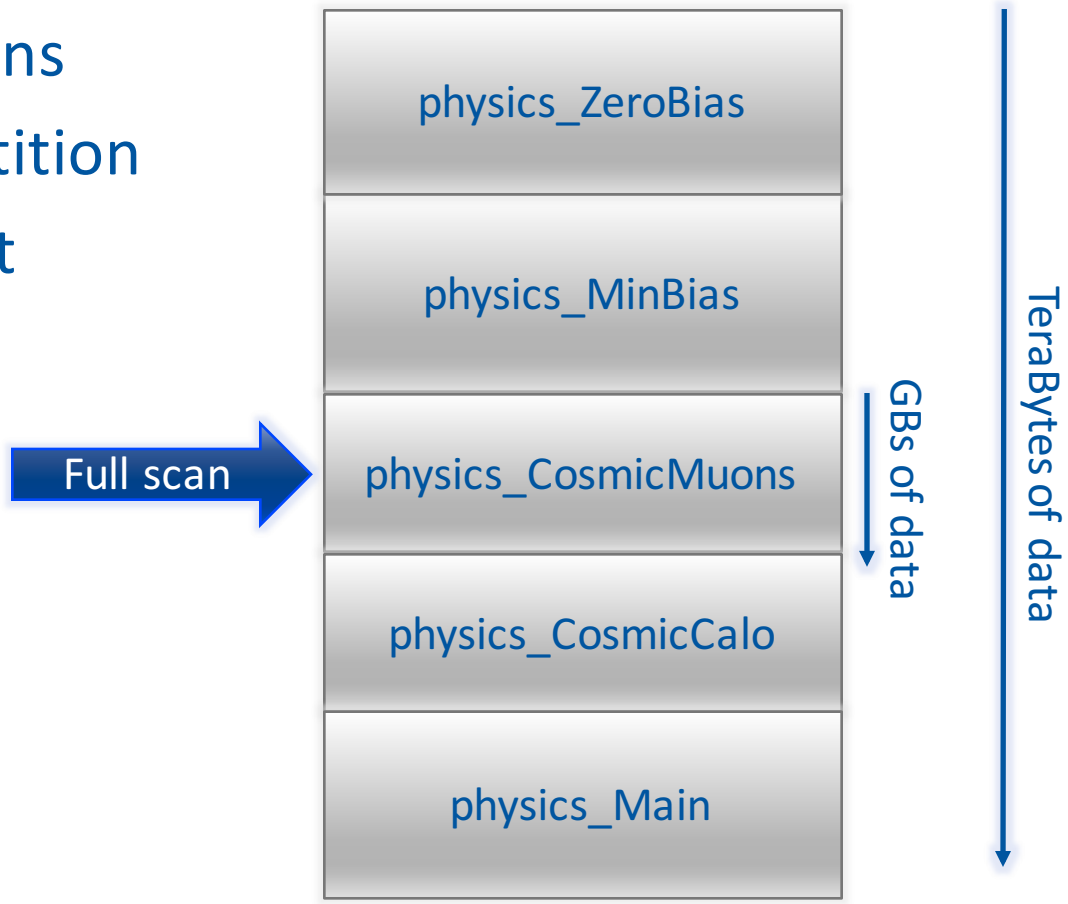
- Key-value store workloads
  - Yahoo! Cloud Serving Benchmark (YCSB)
  - HBase – non relational, key-value data store on HDFS
  - Potential caching and in-memory operations

Test clusters

- VM cluster on CERN Openstack with HDDs
  - 4 nodes, m1.large flavour, per node:
    - 4 VCPUs (Nehalem-C 2.26 GHz)
    - 8 GB RAM
    - 80 GB disk
- VM cluster on CERN Openstack with SSDs
  - 4 nodes, m2.large flavour, per node:
    - 4 VCPUs (Haswell 2.4 GHz)
    - 7.3 GB RAM
    - 40 GB disk
- Physical production cluster
  - 14 nodes, per node:
    - 16 cores (2x Ivy Bridge-EP 2.6 GHz)
    - 64 GB RAM
    - 173 TiB (48x 4 TB disks)

Big Data processing case

- Generating a report for given collection
  1) Prune the partitions
  2) Fully scan the partition
  3) Present the report



Sequential scanning of a dataset

- Looking for events of interest in certain dataset by performing full scan

- Input data set (after preselection)
  - 40M of records
  - Data highly compressed in volume of 12GB
    - in CSV format it occupies 74GB

Scanning performance results: Cluster on virtual machines with HDDs

	VMs on HDDs	VMs with HDFS over network	VMs on SSDs	Physical
scan time	59 sec	43 sec	9 sec	6 sec
scan speed	660K row/s	916K row/s	4.3M row/s	6.6M row/s
Throughput (total)	210 MB/s	287 MB/s	1.3 GB/s	2.05 GB/s
Throughput per machine	52 MB/s	71 MB/s	343 MB/s	147 MB/s

- Processing is throttled by IO performance
  - 50MB/s per VM
  - CPU utilized in 50%

Scanning performance results: VM cluster with remote HDFS attached

	VMs on HDDs	VMs with HDFS over network	VMs on SSDs	Physical
scan time	59 sec	43 sec	9 sec	6 sec
scan speed	660K row/s	916K row/s	4.3M row/s	6.6M row/s
Throughput (total)	210 MB/s	287 MB/s	1.3 GB/s	2.05 GB/s
Throughput per machine	52 MB/s	71 MB/s	343 MB/s	147 MB/s

- Processing throttled by Network performance
  - 70MB/s per VM
  - CPU utilized in 60%

Scanning performance results: Cluster on virtual machines with SSDs

	VMs on HDDs	VMs with HDFS over network	VMs on SSDs	Physical
scan time	59 sec	43 sec	9 sec	6 sec
scan speed	660K row/s	916K row/s	4.3M row/s	6.6M row/s
Throughput (total)	210 MB/s	287 MB/s	1.3 GB/s	2.05 GB/s
Throughput per machine	52 MB/s	71 MB/s	343 MB/s	147 MB/s

- Processing throttled by CPU
  - All 4 VM cores utilised
  - IO throughput per VM 343MB/s

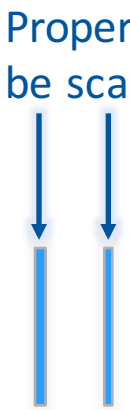
Scanning performance results: Physical cluster was underutilised

	VMs on HDDs	VMs with HDFS over network	VMs on SSDs	Physical
scan time	59 sec	43 sec	9 sec	6 sec
scan speed	660K row/s	916K row/s	4.3M row/s	6.6M row/s
Throughput (total)	210 MB/s	287 MB/s	1.3 GB/s	2.05 GB/s
Throughput per machine	52 MB/s	71 MB/s	343 MB/s	147 MB/s

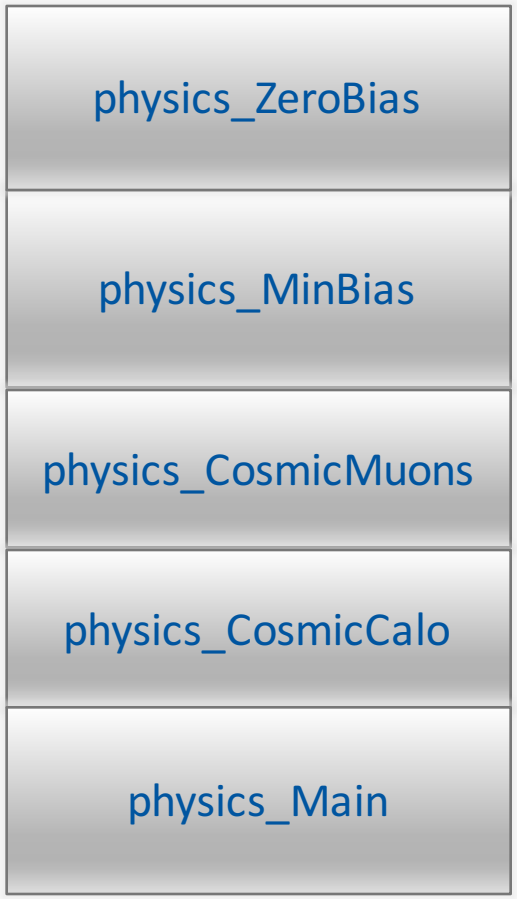
- The test dataset was too small to make use of all resources available
  - Processing 76 files occupied 76 cores, 1/3 of total

Column-wise analytics

- There are data file formats that stores the data in a column-oriented way (e.g Parquet)

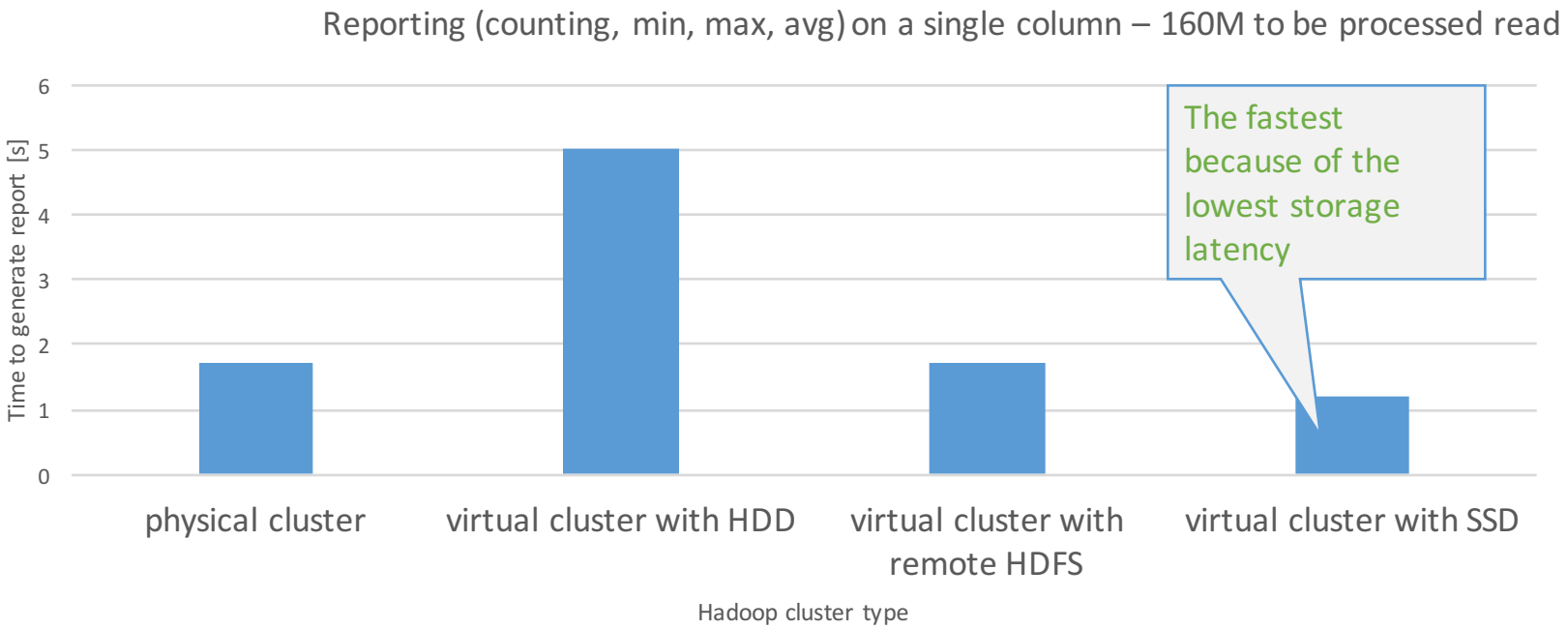


- This reduces amount of data to be read to the subsets of interest



Column-wise analytics - performance

- Reporting workload (counting, averaging, min, max, stdev)
- Single columns values extractions



Recap on analytic workloads

- Throughput of sequential scanning on the virtualized clusters was close to the expectations
  - Bound by performance of underlying hypervisor storage – typically 50MB/s per VM with HDD
  - Accessing the data over network improved IO performance to 70MB/s per VM and offers additional capacity
  - Notably high throughputs (400 MB/s per VM) can be obtain when hypervisors are equipped with SSD

- Satisfying performance has been observed for reporting workloads when input data can be pruned to the level of megabytes

Key-value data store case

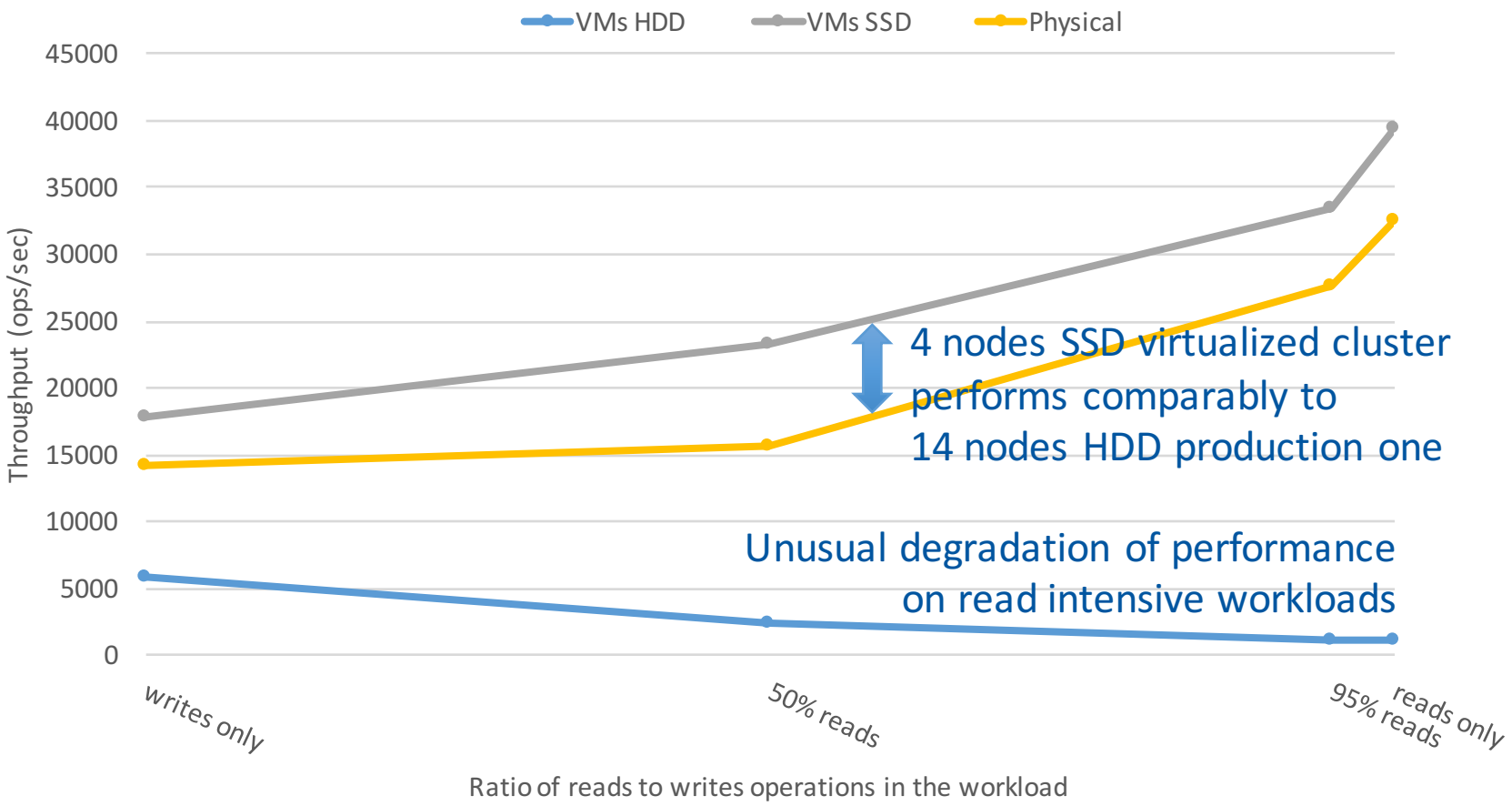
- Different than analytics/reporting workload
  - Only small amount of data is written or retrieved at a time
- HBase used for tests
  - Widely used and strongly linked to Hadoop environment
  - Additional processing on data structures before writing to HDFS
    - Sorting, compression, additional metadata
    - Caching

Yahoo! Cloud Serving Benchmark

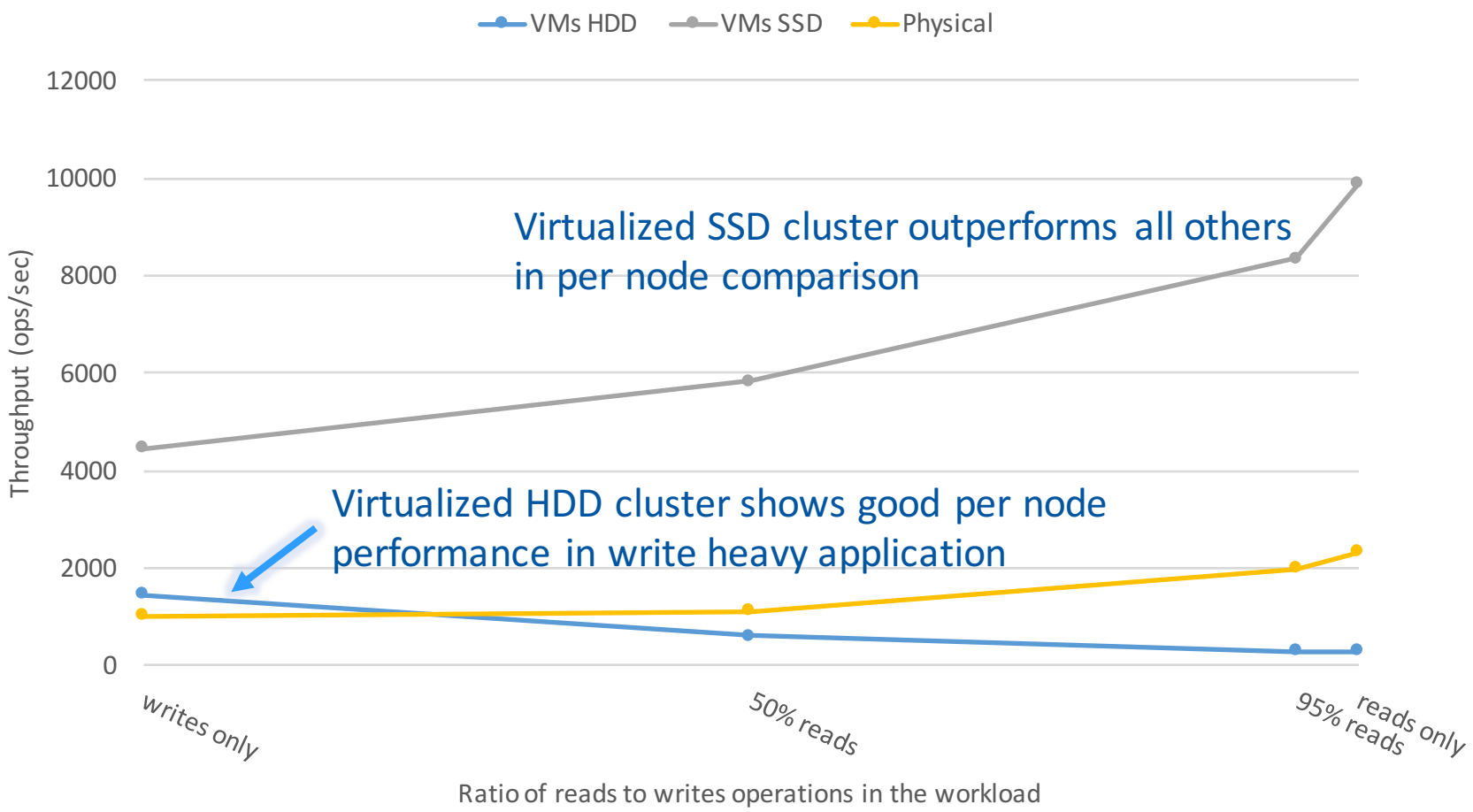
- Popular “Big Data” benchmarking suite

- Tests dataset
  - 10 millions records
  - 14.6 GB on HDFS
- Test workloads
  - Operations on 1 million records
  - Varying ratio of read to write operations

YCSB results – total throughput per cluster



YCSB results – throughput per node



HBase: conclusions

- A small virtual installation on SSDs outperforms our current production cluster
  - Obviously having orders of magnitude smaller capacity

- Virtual installation on HDDs can be suitable for write intensive workloads
  - Per node performance comparable to the physical cluster
  - Could be suitable for write-mostly workloads
  - Can serve up to 5500 write ops/sec
    while only up to 1000 read ops/sec

Summary of use cases for Hadoop on VMs

- VMs on hypervisors with HDDs
  - Development, and functional testing
  - CPU-intensive reporting workloads on highly compressed data
  - NoSQL databases not needing high performance

- VMs on hypervisors SSDs
  - Performance testing of new components
  - Small but high throughput systems

Future work

- Run similar tests on network attached Cinder volumes
- Tests with other workloads and components
  - Streaming data (Spark)
  - Machine Learning (Spark)
- Invite Hadoop users to try out the virtual Hadoop infrastructure
- We plan to use this infrastructure to evaluate new processing framework for Hadoop
  - Apache Flink, Cloudera Kudu

CERN

Acknowledgments

- Many thanks to Bernd Panzer-Steindel and Arne Wiebalck for help with allocating the test HW

- Hadoop Service at CERN
  - Hadoop users' forum and analytics working group https://indico.cern.ch/category/5894/

www.cern.ch

General Observations

- Out of single VM we could get throughput of
  - 60 MB/s reading from local HDD
  - 500 MB/s reading from local SSD
  - 80-90MB/s reading from a network

- VM in CERN Openstack demonstrated very stable performance