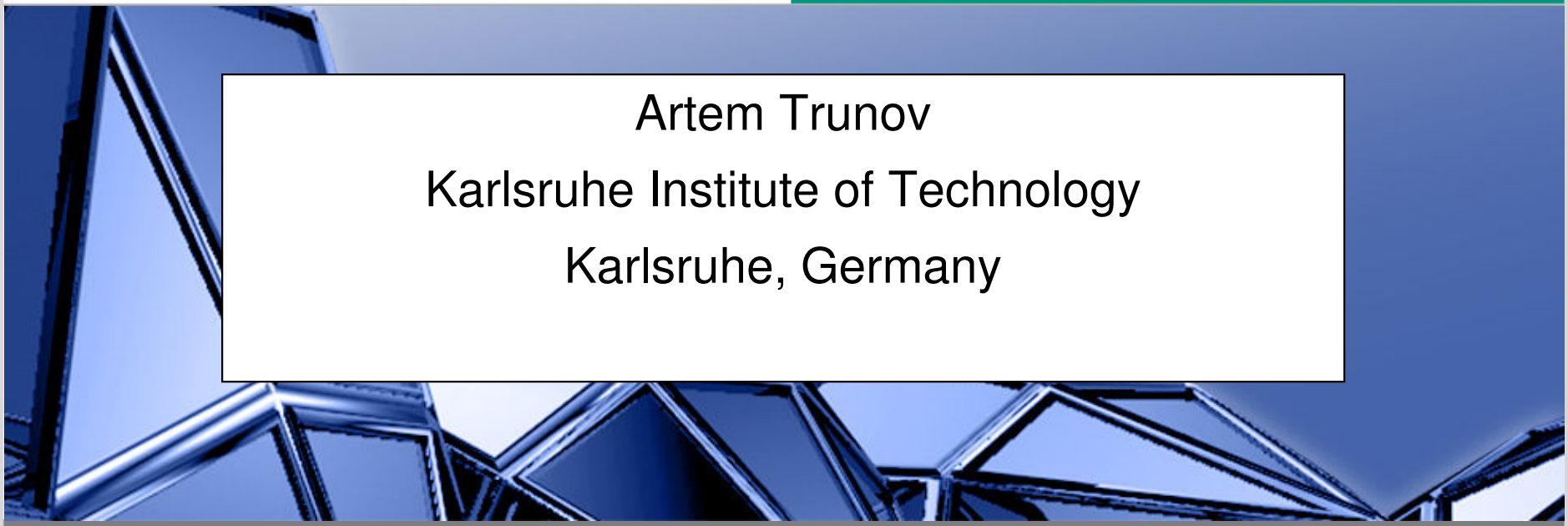
Karlsruhe Institute of Technology

Hadoop on HEPiX storage test bed at FZK

Forschungszentrum Karlsruhe
in der Helmholtz-Gemeinschaft

Universität Karlsruhe (TH)
Forschungsuniversität • gegründet 1825



Artem Trunov

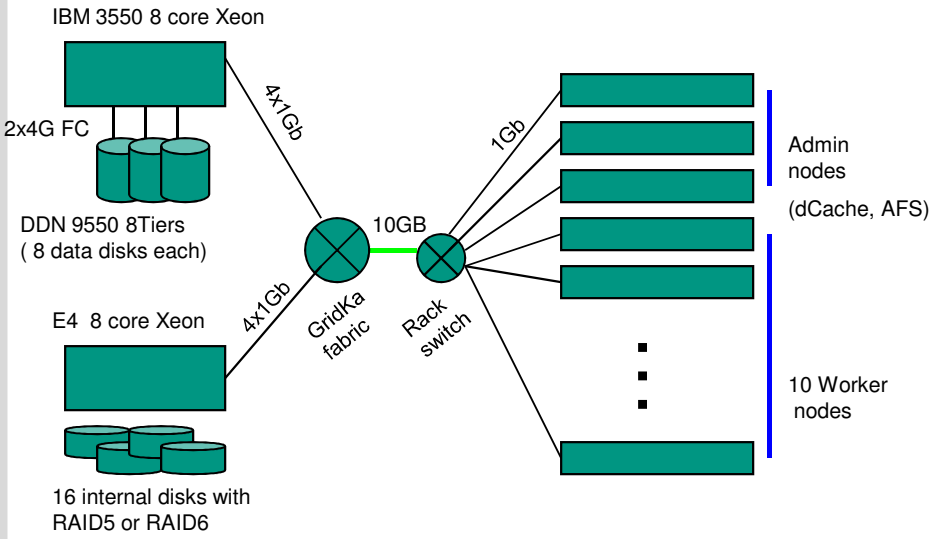
Karlsruhe Institute of Technology

Karlsruhe, Germany

Motivation

- Hadoop is a distributed file system with map/reduce framework, designed to run on commodity cluster and make use of their local hard drives
- Has potential for high parallel performance, scalable with the number of nodes
- Extremely fault tolerant against loss of cluster nodes
- Already in use at a number of OSG site in production
- Packaged in OSG, supported. Reference installation at FNAL
  - OSG Hadoop is packaged into rpms for SL4, SL5 by Caltech
  - BeStMan, gridftp backend
- There is a native ROOT I/O plugin for Hadoop
  - Exists as a patch, by Brian Bockelman
- HEPiX storage WG has a test bed and reference data for performance comparison

Test bed at FZK and previous tests



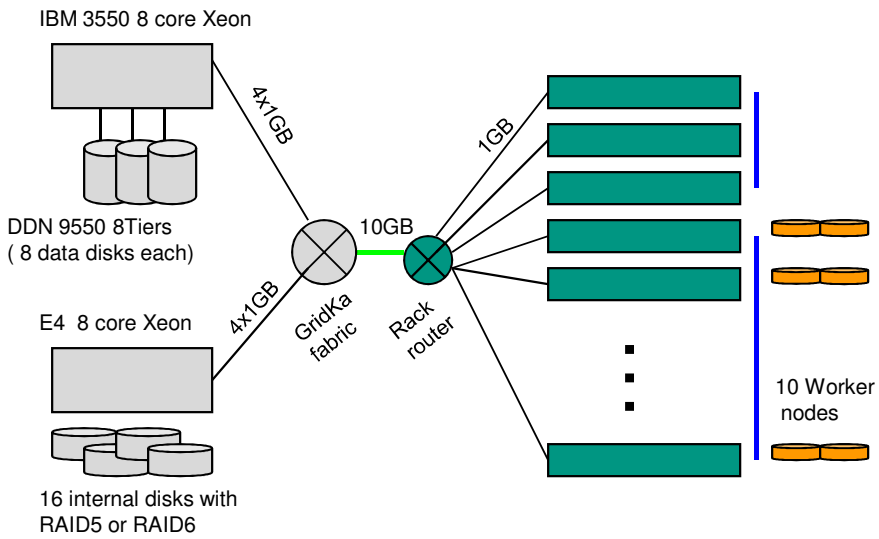
Best Results with DDN disks

AFS/XFS	73 MB/sec	116 MB/sec	116 MB/sec	113 MB/sec
NATIVE	145647 evs	229958 evs	225194 evs	214410 evs
GPFS	171 MB/sec	307 MB/sec	398 MB/sec	439 MB/sec
	168381 evs	297807 evs	380273 evs	420741 evs
AFS/LU	134 MB/sec	251 MB/sec	341 MB/sec	394 MB/sec
VIA VICE	161485 evs	297370 evs	401725 evs	445878 evs
LUSTRE	146 MB/sec	256 MB/sec	358 MB/sec	399 MB/sec
NATIVE	174865 evs	308427 evs	423087 evs	470939 evs

Basic setup of tests with DDN (Andrei Maslennikov)

- Server
  - 8 cores E5335 @ 2.00GHz
  - RAM: 16 GB
  - FC: 2 x Qlogic 2462 dual 4Gb
  - Network: Quad GE card in ALB bonding
    - mode=6. miimon=100
    - measured 450 MB/sec memory-memory in both directions
  - OS: SuSE SLES10 SP2 64 bit
    - Kernel: 2.6.16.60-0.27_lustre.1.6.6-smp
- Disk system: DDN 9550
  - Direct attach with 2 x 4 Gb FC
  - 4 Luns each composed of two tiers
  - 16 data disks per Lun
  - Lun block size - 4MB
  - Cache: disabled
- Clients
  - 10 x 8 cores E5430 @ 2.66GHz
  - 16 GB RAM
  - OS: RHEL4 64 bit
  - Kernel: 2.6.9-78.0.13.ELsmp non-modified + Lustre modules
- Test jobs
  - CMSSW 1_6_6
  - 2,4,6,8 jobs per node, 10 nodes
  - 40 minutes run

Testbed for Hadoop

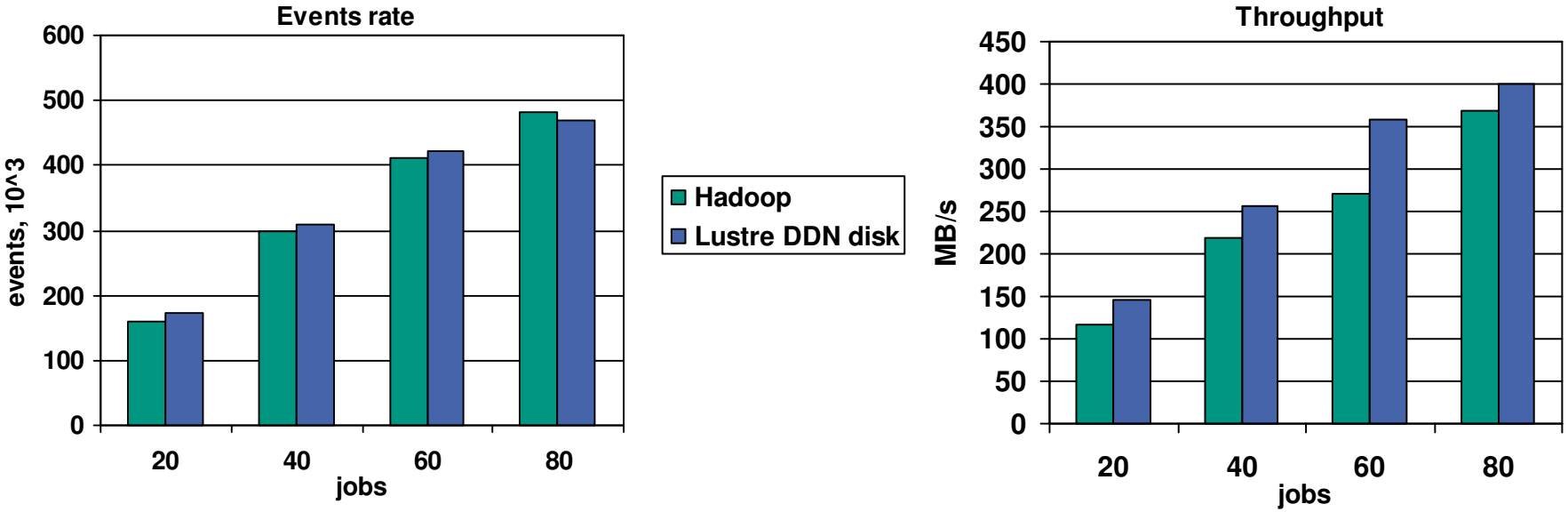


- *External servers and storage are not used*
- *Make use of worker nodes' two internal 250 GB SATA hard drives*
  - *On each allocate ~200MB partitions, format with ext3*
  - *Total of ~4TB of free space*
- Hadoop setup
  - Version 0.19 SL4 x86_64 from Caltech repo
  - 10 datanodes + 1 namenode
  - Test Jobs were run on 9 data nodes and the name node
  - Fuse interface to HDFS, mounted on each node
- Slight complication: due to high sensitivity of Hadoop to performance of hard drives, had to reject one data node and use one of admin nodes as data nodes
  - This had little impact on the test result.

- Hadoop settings:
  - block size 64M
  - replication factor 1
  - java heap size 512MB
- fuse settings:
  - 'ro,rdbuffer=65536,allow_other'
- network settings (pretty standard):
  - net.core.netdev_max_backlog = 30000
  - net.core.rmem_max = 16777216
  - net.core.wmem_max = 16777216
  - net.ipv4.tcp_rmem = 4096 87380 16777216
  - net.ipv4.tcp_wmem = 4096 65536 16777216
- block device settings:
  - echo 32767 > /sys/block/sd${dr}/queue/max_sectors_kb
  - echo 16384 > /sys/block/sd${dr}/queue/read_ahead_kb
  - echo 32 > /sys/block/sd${dr}/queue/iosched/quantum
  - echo 128 > /sys/block/sd${dr}/queue/nr_requests
- Vary during the tests
  - block devise settings:
    - read_ahead_kb: from 1 to 32 MB
    - nr_requests from 32 to 512
  - fuse read ahead buffer:
    - rdbuffer: from 16k to 256k
- Optimum was found at the following:
  - read_ahead_kb: 16 MB
  - nr_requests: 128
  - fuse rdbuffer: 128k
- Measure
  - Total event count for 40 minute test jobs
  - Read rate from disk

Best results

	20 threads	40 threads	60 threads	80 threads
Hadoop	116 MB/sec	218 MB/sec	270 MB/sec	369 MB/sec
	161103 evs	298243 evs	412836 evs	481841 evs
LUSTRE DDN DISK	146 MB/sec	256 MB/sec	358 MB/sec	399 MB/sec
	174865 evs	308427 evs	423087 evs	470939 evs



Discussion

- Hadoop in this test bed is close to Lustre and outperforms it in the maximum load test.
  - 8 jobs on 8 core machine is a standard batch setup
- Some other considerations are also taken into account when selecting storage
  - Cost of administration, ease of deployment, capacity scaling, support for large name spaces
- Hard to think it's a main HEP T1 storage solution, or it needs a lot of additional testing and careful deployment.
- As T2/T3 storage should be very interesting to WLCG sites
  - Cost and maintenance factor is very favorable to small sites

Future plans

- HEPiX test bed at FZK moved to a dedicated rack, RH5
- Hadoop 0.20 or 0.21, all 64bit
- Newer CMS and Atlas software
- Check performance with replication factor >1
- Check with various chunk sizes
- Test on a high end storage?

Acknowledgments

- Andrei Maslennikov
  - test suite setup, valuable guidance and comments
- Brian Bockelman, Terrence Martin (OSG)
  - Hadoop wiki, tuning tips
- FZK team
  - Jos van Wezel, Manfred Alef, Bruno Hoeft, Marco Stanossek, Bernhard Verstege