EXPLORING HADOOP MAPREDUCE ON NETWORK ATTACHED STORAGE

Ellis Wilson (Penn State), Garth Gibson (CMU)

OVERVIEW

- MapReduce (MR) framework is heavily used for analytics
- Coupled with HDFS to co-locate data and computation
- However most organizations already use NAS
  - Complete disk management solutions in place
  - Advanced RAID in NAS may be more cost effective
  - Flexibility to grow compute and storage independently
  - Higher quality parts in specialized systems
  - Ability to share among multiple disks
  - Can disaggregate compute nodes without risking data loss

Traditional HDFS to HDD
- Specify 1 or more local paths to local HDDs on each node
- Pros: Simplicity, replication
- Cons: (see above)

HDFS to NAS:
- Give local mount path of NAS to HDFS instead of HDD
- Pros: Simplicity, HDFS replication as well as NAS reliability
- Con: HDFS replication may provide no benefit if all copies in same NAS RAID set
- Con: Overheads from HDFS, namespace not useful through NAS

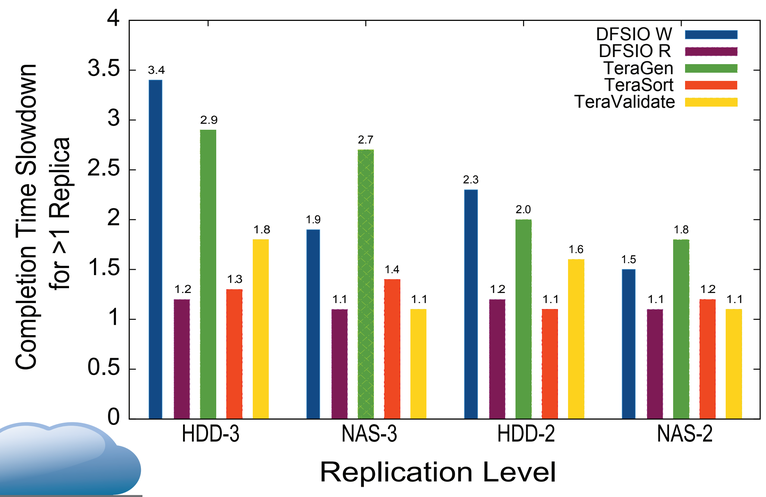
Direct:
- Disable HDFS, provide MR with mount paths to NAS system
- Pros: Direct data access, namespace equiv. between NAS and MR
- Cons: Requires new FileSystem type, no additional replication

EXPERIMENTAL FRAMEWORK

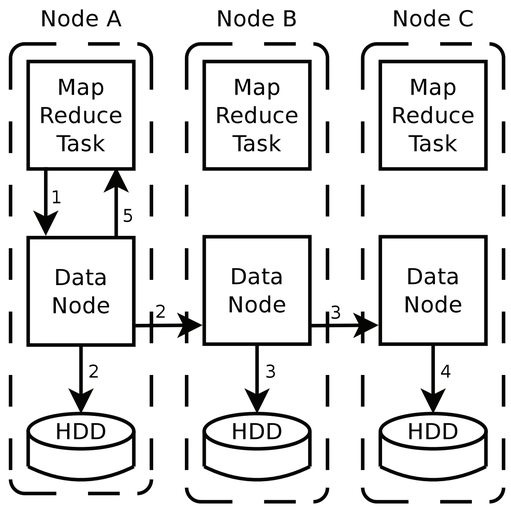
- 50 VMs, 1 per machine, on OpenCirrus (1Gbps)
  - 2 cores (of 8), 3.8GB (of 16), 200GB (of 1TB)
- 5 Panasas ActiveStor 12 shelves (Ver 4.0.2)
  - 20Gbps per shelf, 40TB per shelf (20 disks)
- Apache Hadoop Tests (DFSIO-W, DFSIO-R)
- Apache Terasort (TeraGen, TeraSort, TeraValidate)
- MapReduce: 4 maps, 1 reduce slot per node
- Blocksize: 512 MB (input,output), 32 MB (intermediate files)

IMPACT OF REPLICATIONS

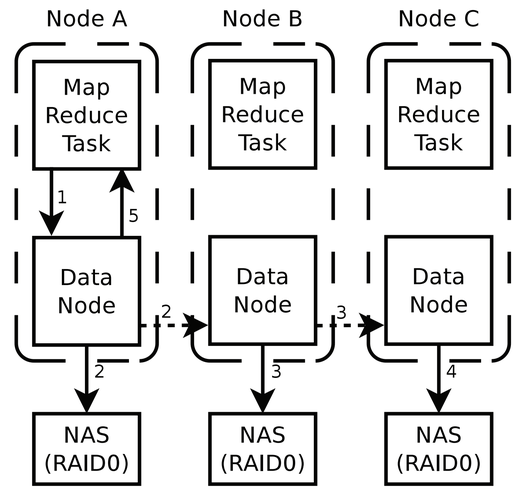
- Compare HDFS replication levels 2 & 3 to 1 HDFS replica
- NAS is Panasas RAID5 using DirectFlow
- Best comparison is HDD-3 and NAS-2 (triple-failure tolerant)
- Hardware differs, cost differs, so we show relative slowdown for extra replication vs 1 replica



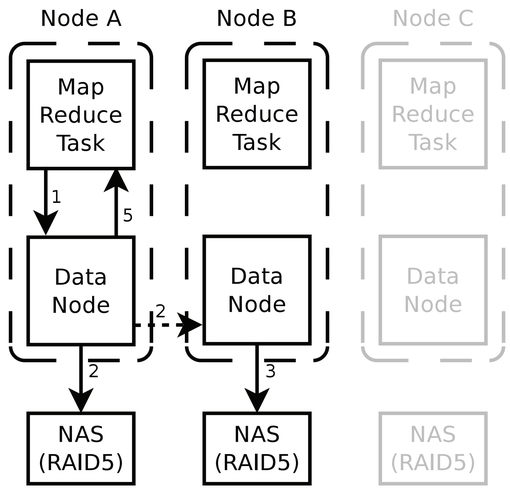
I/O FLOW MODELS



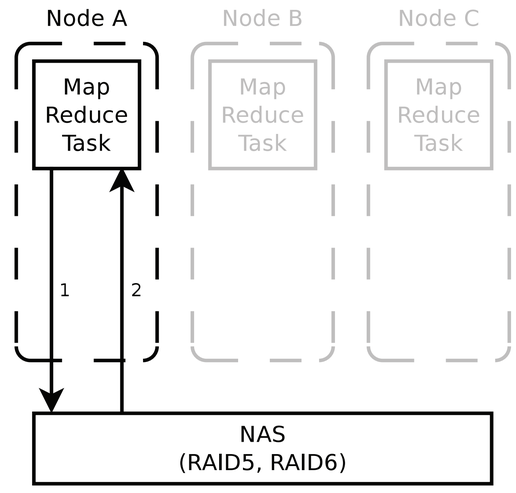
Traditional HDFS to HDD
- 2 network hops per block write
- Double-disk failure tolerance



HDFS to NAS – Replication 3
- 5 network hops per block write
- Double-disk failure tolerance (RAID0 NAS)



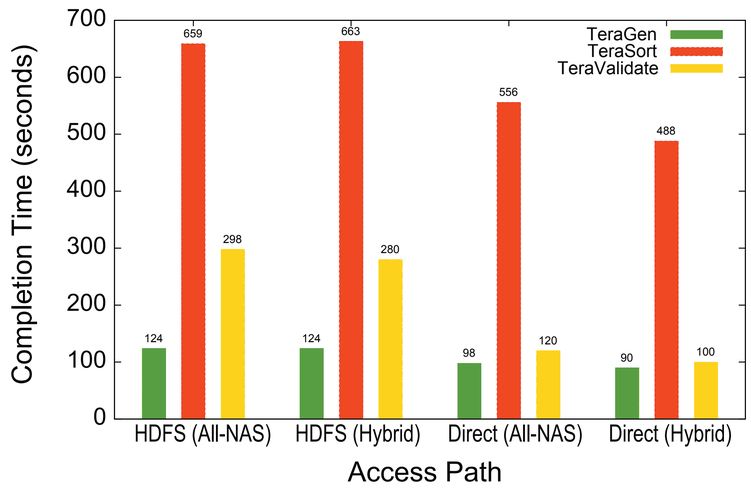
HDFS to NAS - Replication 2
- 3 network hops per block write
- Triple-disk failure tolerance (double RAID5 domains)



Direct to NAS - Replication 1
- 1 network hop per block write
- Single or double-disk failure tolerance (RAID5 or RAID6)
- HDFS (rep=1) to NAS is same

IMPACT OF ACCESS PATHS

- Compare use of NAS (Panasas RAID5 DirectFlow)
  - All NAS: input/output and intermediate files on NAS
  - Hybrid: input/output on NAS, intermediate files on local disk
  - HDFS: data flows through HDFS to NAS (rep=1)
  - Direct: data bypasses HDFS to NAS (rep=1)
- Significant cost for flowing data through HDFS



Intel Science & Technology Center for Cloud Computing
Carnegie Mellon University
Georgia Tech
intel
PRINCETON UNIVERSITY
UC Berkeley