DISKREDUCE: RAIDING THE CLOUD - ENCODING OPTIONS AND RELIABILITY

Bin Fan, Wittawat Tantisiriroj, Lin Xiao, Garth Gibson (CMU)

IMMEDIATE VS. BACKGROUND ENCODING

Immediate encoding:

+ **Efficient**
- **Complex: Handling failures on critical path**

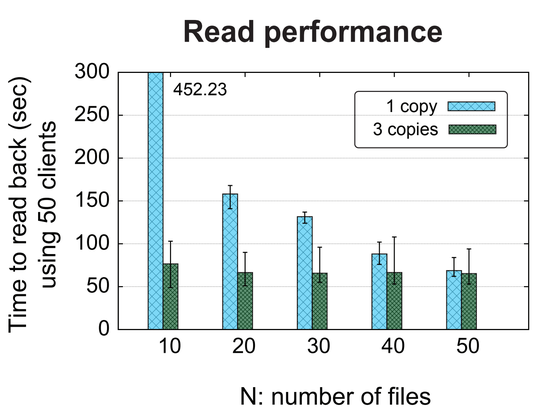
Background encoding:

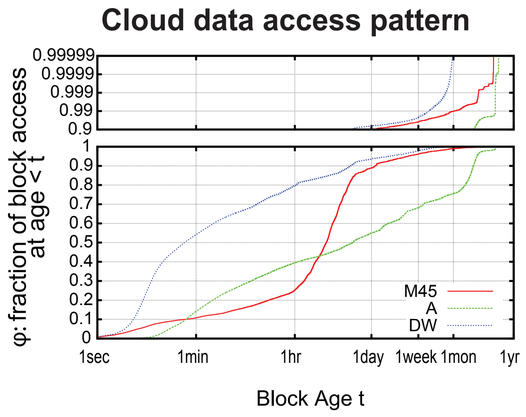
+ **Simple & no change in client code**
+ **Cache young data for higher read bandwidth**
- **Less efficient**

PERFORMANCE IMPLICATION

75GB dataset written by N clients (1 file each) with 1 or 3 copies

- **Note first copy stays in writing node's disk**
- **Hotspots benefit from replication**



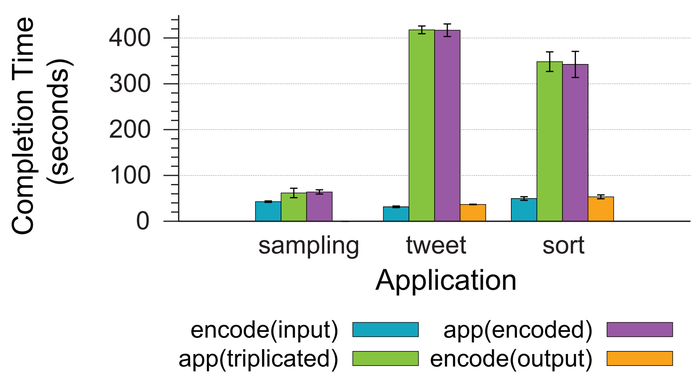


Treat triplicated data as a cache:

- **Locality: about 90% of data blocks are accessed within 1st day after creation in M45 & DW and 50% in cluster A**
- **Even with hotspot, most accesses can get benefit from replication if encoding deferred by 1 day**

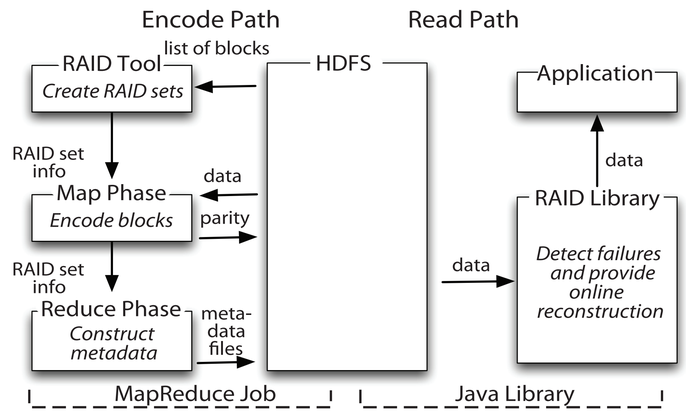
APPLICATION PERFORMANCE

- **Sampling (B. Fu): Read 150GB astronomy data-set**
- **Twitter (B. Meeder): Reformat 24GB dataset to 56GB**
- **Sort: sort 128GB dataset**



- **The time each application takes when input dataset are triplicated or encoded are comparable.**
- **20% slow down if encoding is done during busy time**
- **Or need at least 20% of idle time**

PROTOTYPE



- **The prototype is built as a tool and a client library**
  - **Tool (Mapreduce): encode a directory into RAID sets or repair corrupted files**
  - **Library: detect and correct missing data while reading**
- **Released as Mapreduce-2036 patch for HDFS 0.22.0 @ http://issues.apache.org/jira/browse/MAPREDUCE-2036**
- **60 nodes (two quad-core 2.83GHz Xeon, 16GB memory, four 7200 rpm SATA 1TB disks, 10 Gigabit Ethernet)**
- **Dataset: 240GB (3,840 files, each 64MB in size)**

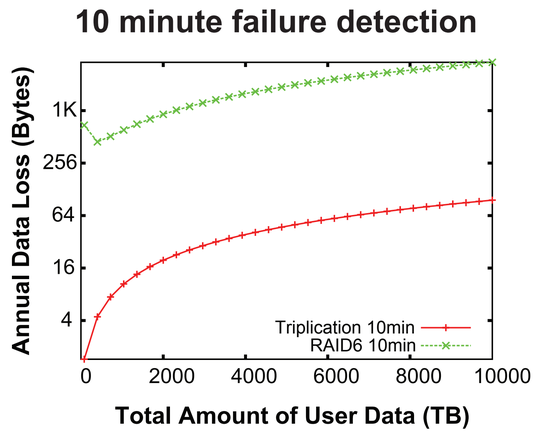
Operation	Throughput GB/s(stdev)	Disk I/O GB/s (stdev)
Write(Triplication)	1.93(0.06)	5.80(0.18)
Encode(RAID6 8+2)	3.69(0.34)	4.61(0.43)
Repair	0.23(0.02)	2.09(0.19)

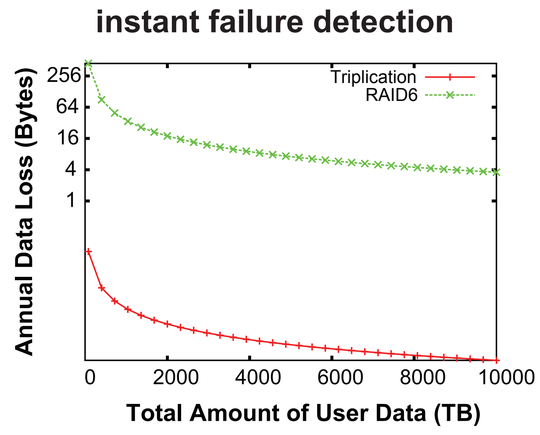
- **Encoding is fast but reconstruction needs improvement**

RELIABILITY MODELING

We expect Triplication to be more reliable than RAID6

- **Time to fail/repair model is exponentially distributed, 2% annual failure rate**
- **1TB/disk 80% full, 64MB/chunk, 8+2, 25MB/s/disk repair**
- **Compare bytes lost per year as a function of total sizes**
- **Orders of magnitude difference is still only a few bytes/year**





Interaction of parallel repair and detection delay

- **With instant detection, surprisingly scalable repair improves reliability with scale**
- **With more common delayed detection, unsurprisingly scalable repair less effective**

Intel Science & Technology Center for Cloud Computing