Flash-Optimized Database Consistency: *Why Wait? Write Now.*

Michael Slavitch

Senior System Architect

Diablo Technologies

THE SITUATION TODAY

Databases can be divided into two major categories:

Commit-Oriented

(Ex. DB2, MySQL, SQL Server, Oracle RDBMS)

- Historical Prevalence (40+ years)
- General-Purpose Applicability
- Real-Time Data Consistency
- Exploits High End Hardware
- Expectations of high performance

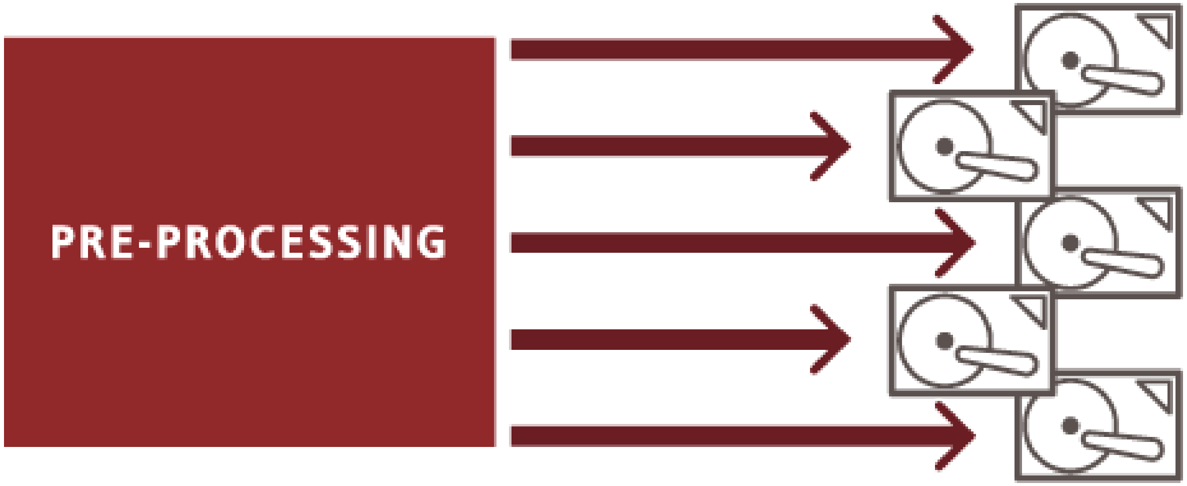
Eventually Consistent

(Ex. Cassandra, RocksDB, NoSQL DBs)

- Evolved over last 15-20 years
- Originally Targeted At Specific Use Cases
- Delayed ("Eventual") Data Consistency
- Designed For Commodity Servers / Storage
- Now being pushed towards performance

EVENTUALLY CONSISTENT ARCHITECTURES: *SHORTCOMINGS ARE NOW APPARENT*

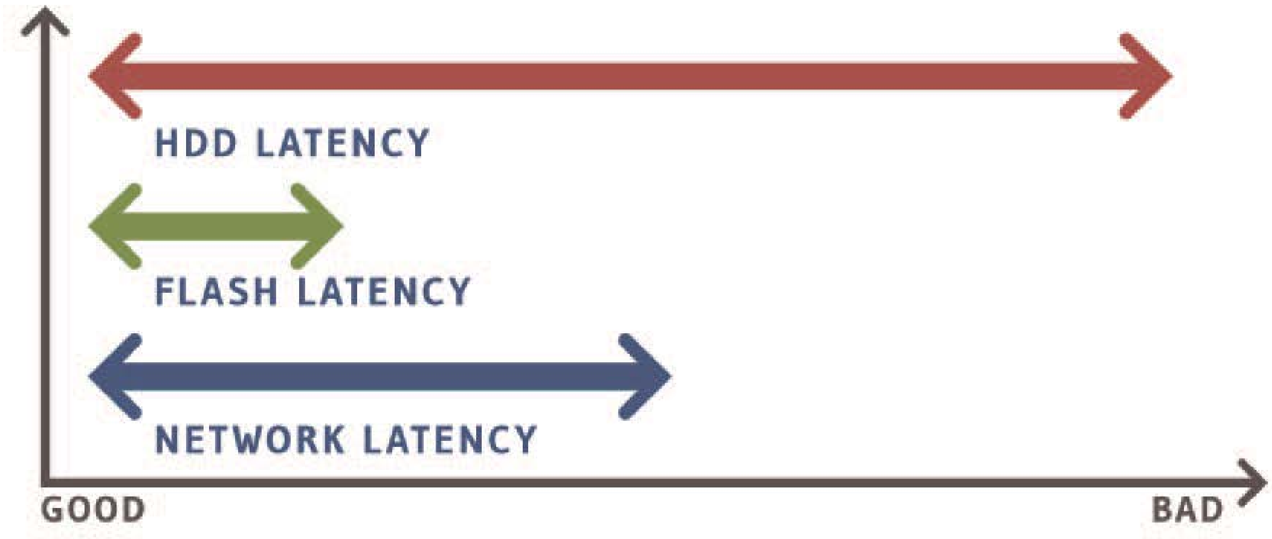
- Designed For Cheap, Low Performance HDDs
  - Hardware with poor random read/write performance
  - Pre-processing incorporated to minimize application wait times



- Resulting Lag (Before Commit) Can Be Significant

EVENTUALLY CONSISTENT ARCHITECTURES: *SHORTCOMINGS ARE NOW APPARENT*

- Flash **Should Be** The Answer

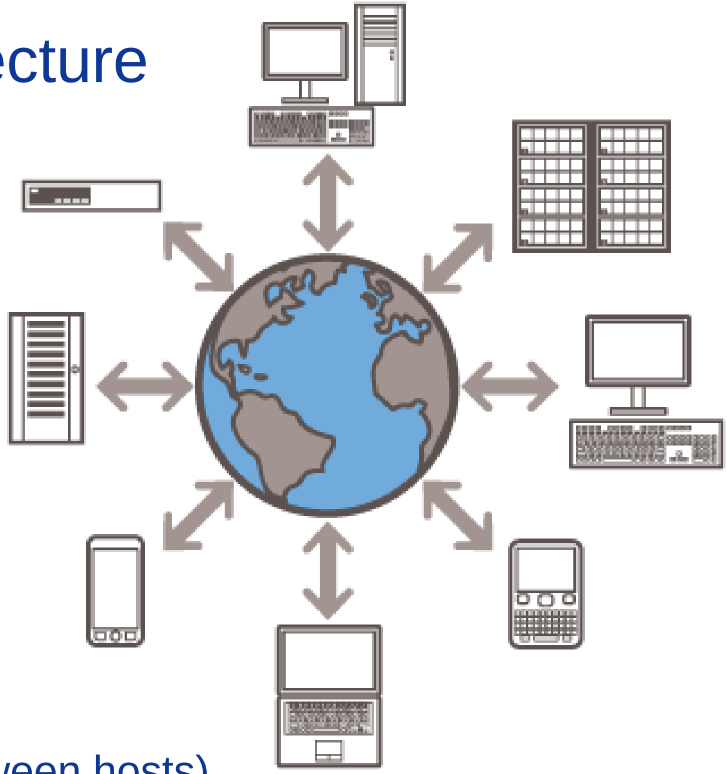


However…

- Simply Replacing HDDs With SSDs ≠ Success

WHERE DIABLO SEES INNOVATION: *NETWORK EQUIPMENT VENDORS*

- No Time For Traditional DB Architecture
  - Need 200-400GB/sec for 100’s of millions of connections
  - Massive concurrency on reads/writes



- Following KISS Approach:
  - Simple algorithms to distribute data
  - Use Flash as object store (both on-host and between hosts)
  - User space drivers to avoid kernel overhead
  - Special stacks for both network and storage
  - Using mathematical models to find and eliminate bottlenecks

WHERE DIABLO SEES INNOVATION: *SURPRISE! RELATIONAL DATABASE VENDORS*

IBM

DB2 with BLU Acceleration

ORACLE

Automatic Storage Management (ASM)

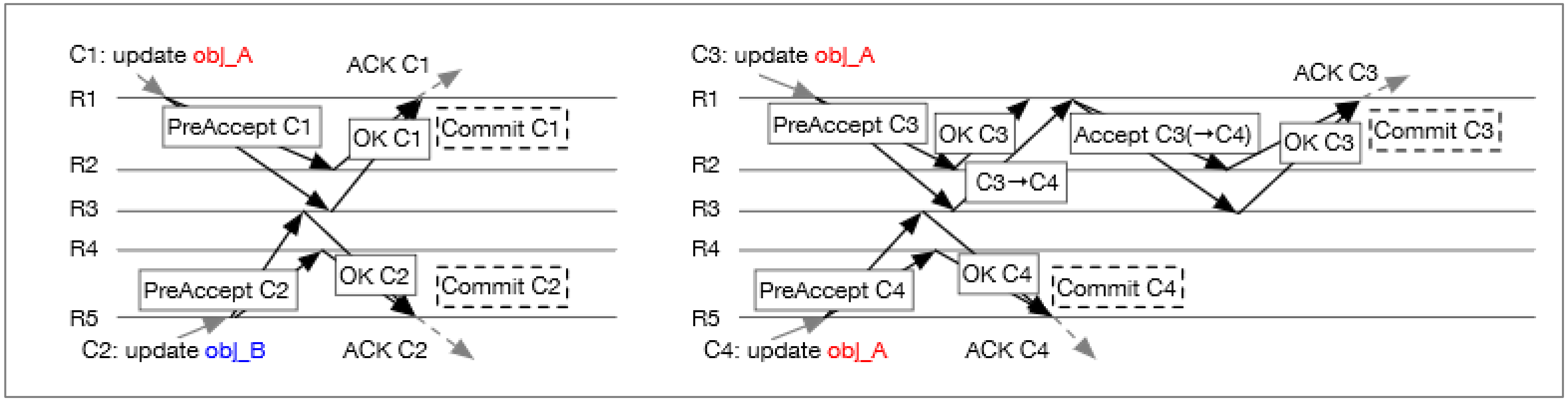
PERCONA

Performance-Optimized MySQL

- Different Approaches, Similar Philosophies:
  - Leverage Flash performance
  - Massive parallelism
  - Enable fast commits
- Looking Ahead:
  - User-space drivers to avoid kernel overhead
  - Data commits at cache line granularity

WHERE DIABLO SEES INNOVATION: *FASTER QUORUM IN DISTRIBUTED COMPUTING*

- New Distributed Consensus Algorithms
  - Egalitarian Paxos (EPaxos), faster distributed commits



  - Quorum rate bottlenecks high performance distributed computing
  - If there is fast quorum, there can be fast commits
  - Still in academic research phase

- Commercial Research in Flash-oriented Databases
  - Proprietary, under NDA

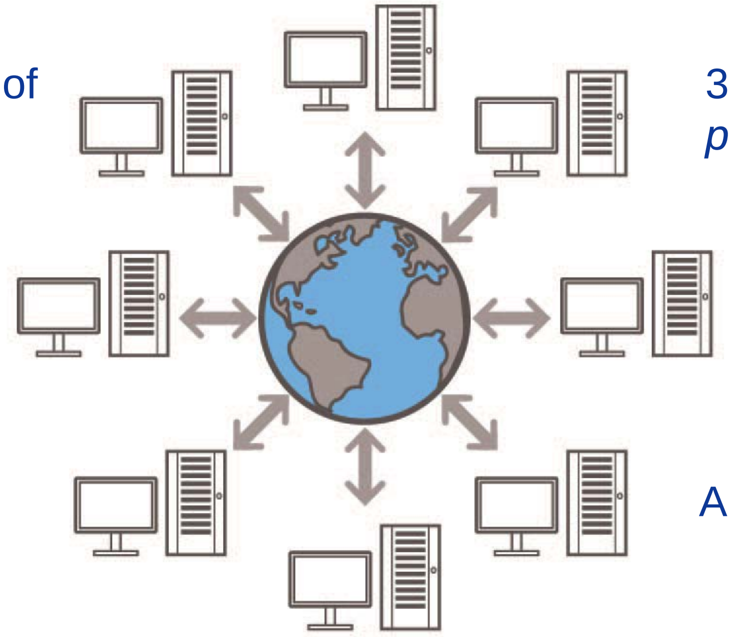
Image Source: *There Is More Consensus in Egalitarian Parliaments*, Iulian Moraru, David G. Andersen, Michael Kaminsky (Carnegie Mellon University and Intel Labs)

WHAT DIABLO PREDICTS: *NEW ENTRANTS MAY CAUSE DISRUPTION*

- Watch Out For The Networking Vendors
  - CDNs and proxy servers are essentially distributed object DBs

Already building millions of concurrent connections

300-400Gbits of payload traffic *per server* on commodity HW



Cloud deployment is familiar territory

Always fast, now getting faster

- May Emerge As Threats To Distributed DB Vendors

WHAT DIABLO RECOMMENDS: *FLASH-ORIENTED DISTRIBUTED COMPUTING*

- Stop Compensating For Spinning Disk:

  - Replace SST files with directly accessed objects
  - Replace SST file hierarchy with tiered DHT's
  - Tiering/distribution automatic and fast on-host
  - Automates redundancy without need for RAID

- Compelling Advantages
  - Frees CPU (no more sorting/merging)
  - Frees RAM (cached objects replace in-RAM tables)
  - Improved Application Performance
  - Increased Software Simplicity

HOW DIABLO CAN HELP:
MEMORY CHANNEL STORAGE™ TECHNOLOGY

- MCS: Placing Flash Within The Memory Subsystem

Lowest storage latency via DDR3 data interface

Leverages NUMA parallelism for ultra-efficient scaling

Block interface (no changes required to existing software)

Powering SanDisk® "ULLtraDIMM™" and IBM® "eXFlash™ DIMM"

- Perfect Fit For Database Acceleration
  - Sub-5µs commit times
  - Inherent parallelism fits distributed architectures
  - Unlocks true potential of flash

CONCLUSIONS

- Traditional RDBMS Vendors Are On The Right Track
  - Focus on efficient utilization of Flash
  - Natural progression for likes of Oracle, IBM, MySQL, MSFT
  - Allows them to revive technologies once seen as obsolete
- NoSQL Vendors Have Catching Up To Do
  - HDD-to-SSD upgrade is only part of the solution
  - Fundamental architectural roadblocks must be addressed
  - New research is promising, but execution is key
- Survival Of The Fittest
  - Networking vendors on COTS hardware will be disruptive
  - Effective utilization of Flash is crucial
  - DB vendors must learn and adapt…..or risk being displaced

THANK YOU