Flash Optimized Databases and Data Stores

Dr. Brian O'Krafka
San Disk Corporation Fellow

Tutorial G-31: Enterprise Applications Part 3
Database/Data Management Acceleration
Thursday, August 23
From 8:30 to 10:50 am

Abstract

Tremendous technology advances have been made in recent years that offer the potential for major data center improvements. Commodity hardware industry advances in enterprise flash memory and multi-core processors offer vast improvements in performance while reducing power and space consumption.

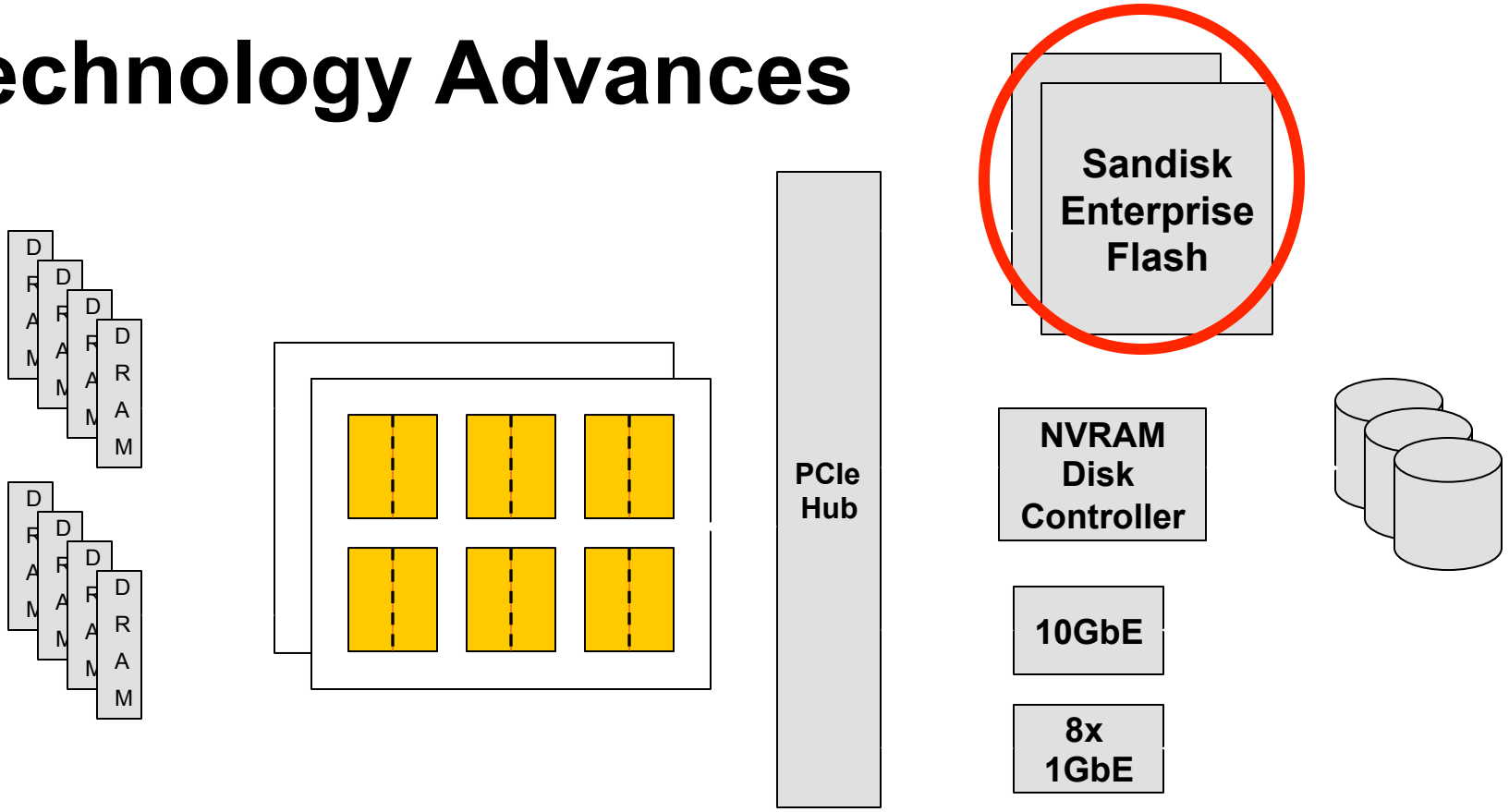
New database and datastore software architectures are required to exploit these hardware advances to provide commensurate improvements in data center performance, scalability, availability and cost structure.

In this presentation we will discuss these technologies along with resulting data center and IT transformations. We will provide several case studies of mission critical enterprise deployments that demonstrate the use cases and the benefits.

Flash Optimized Databases and Data Stores

- Flash breaks IOP bottleneck, but whole system must be optimized
- Example 1: Membrain Key/Value Cache/Store
- Example 2: SchoonerSQL Database
- Conclusion

Technology Advances



100's to 1000's of GB of Flash Holds Database, Provides 10,000's of Write IOPS, 10,000's of Read IOPS at 10's of microseconds latency

Diminishing Returns Without System Optimization

TPS/ Node	In DRAM	In Flash
Cassandra	10,500	1,790
MongoDB	49,000	4,000

TPS/ Node	In Hard Drives	In Flash
MySQL 5.5	7,500 TPM	20,000 TPM

Benchmark

- Random key-value query of 32M (fits in DRAM) and 64M (fits in Flash) 1kB objects
- dual quad-core Intel Nehalem processors
- 64 GB of DRAM
- ½ TB of flash

DBT2 open-source
OLTP version of TPC-C

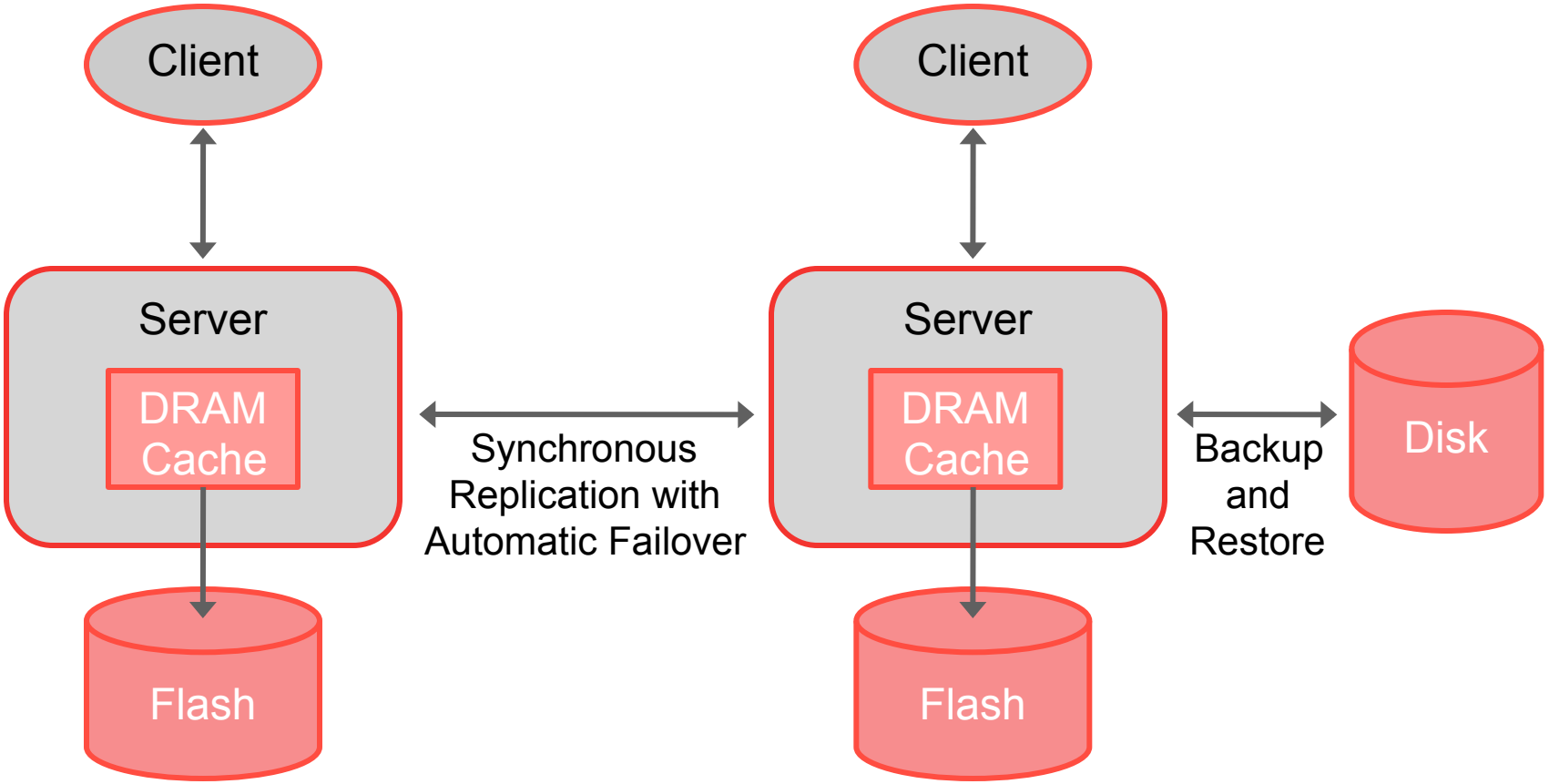
- 1000 warehouses, 32 connections
- 0 think-time
- Result metric: TPM (new order)

Measurement Configuration

- 2 node Master-Slave configuration
- 2 socket Westmere
- 72GB DRAM

Example 1: Membrain HA Cache and Key-Value Store

100% memcached compatible

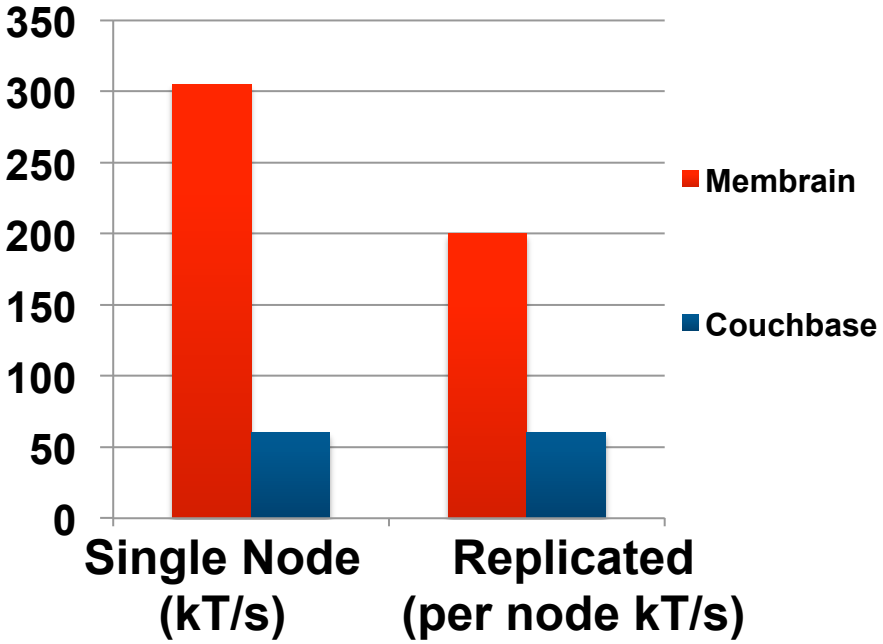


Membrain Versus NoSQL Alternatives

TPS/ Node	In DRAM	In Flash
Cassandra	10,500	1,790
MongoDB	49,000	4,000
Membrain	310,000	160,000

Benchmark

- Random key-value query of 32M (fits in DRAM) and 64M (fits in Flash) 1kB objects
- dual quad-core Intel Nehalem processors
- 64 GB of DRAM
- ½ TB of flash

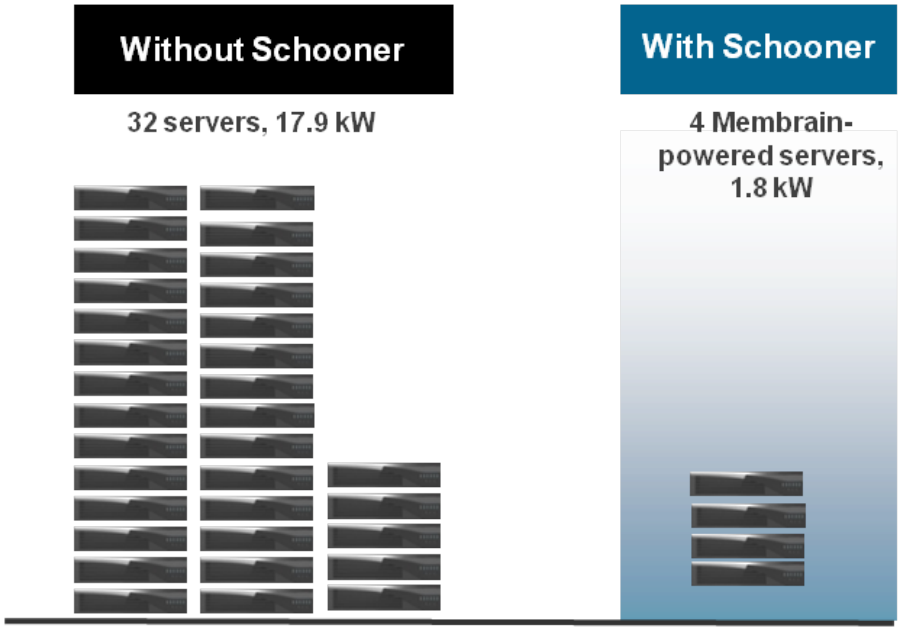


Benchmark

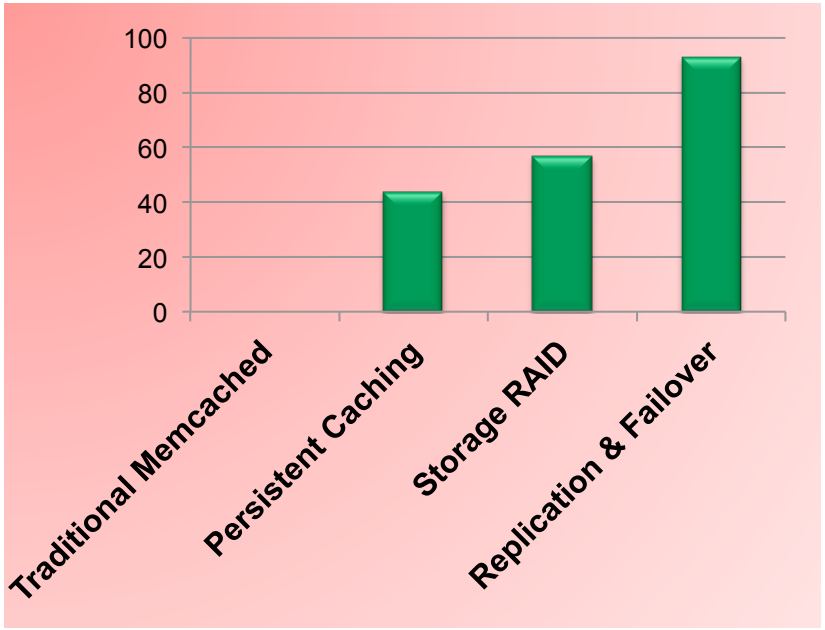
- Random gets (95%) and puts (5%) to 2kB (avg) objects (35% DRAM miss rate)
- Dual Intel Westmere processors
- 64 GB of DRAM
- ½ TB of flash

Membrain Customer Deployment Example

- 8:1 server consolidation + 10.1 power reduction
- 95% availability improvement
  - With transparent synchronous replications and automated failover



Membrain's Consolidation and Power Savings



Cumulative % reduction in downtime

What Was Required to Exploit Flash?

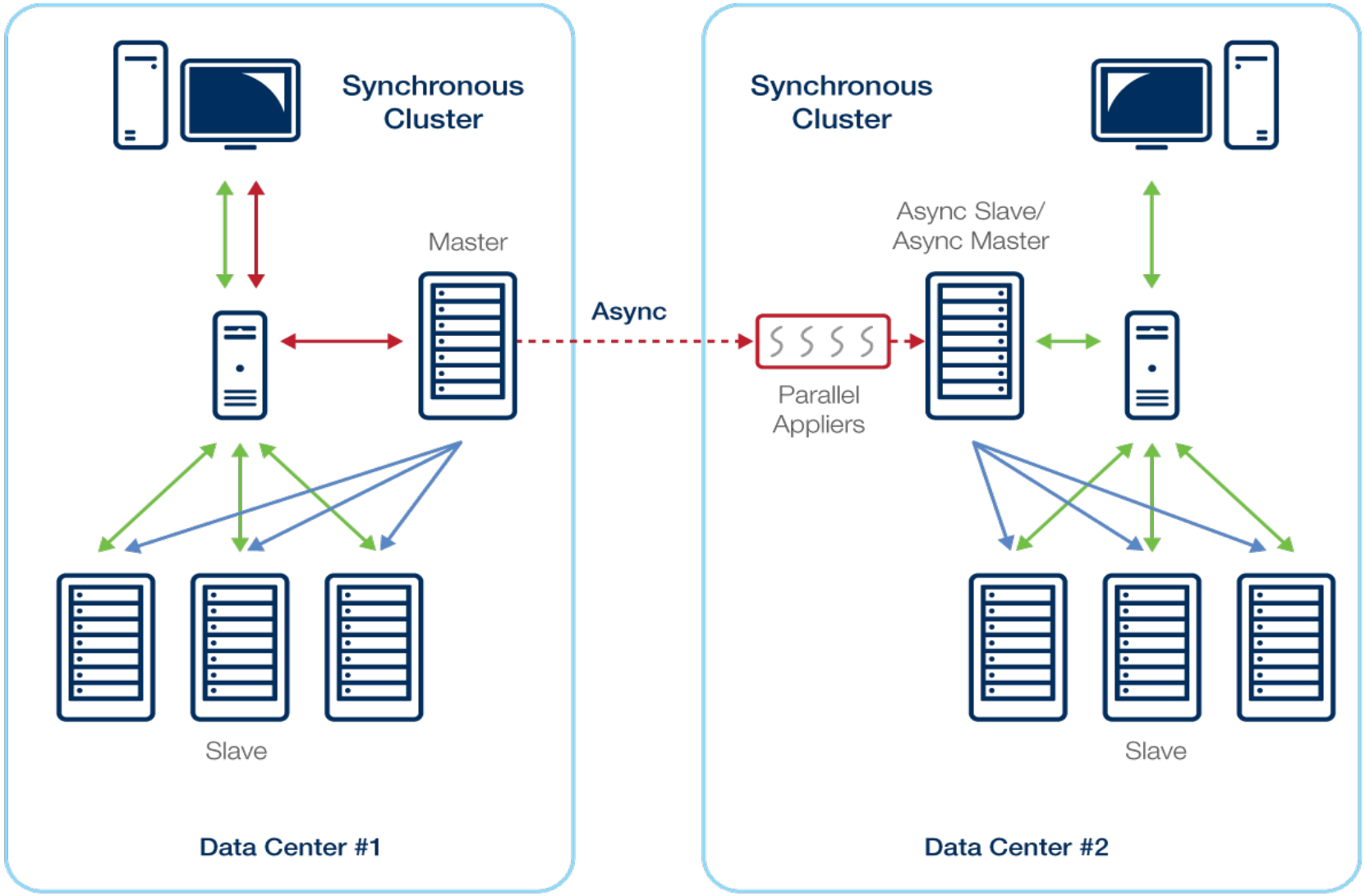
- **Intelligent DRAM caching:**
  - Most recently used objects are cached in DRAM
  - Highly space efficient, even for very small objects (~100B)
  - Self-balancing as object sizes vary over time
  - DRAM cache is shared by all containers
  - Writeback or write-through (customizeable to workload)

- **Custom lightweight threading to maximize concurrency and minimize response time**
  - User-space thread scheduler uses polling to avoid cost of OS context switches

- **Configurable flash management algorithms to optimize different workloads:**
  - FIFO mode: multiple updates are accumulated in a FIFO buffer and written as a batch
    - Good for caching when overwrite rates are low
  - Slab mode: storage is managed as power-of-2 slabs
    - Good for persistence and when overwrite rates are high

- **High Performance Replication with fully automatic failover and failback:**
  - High throughput, low latency synchronous replication
  - Fully automatic failover using VIP's
  - Fully automatic failback when a failed node comes back online
  - Fast data transfer to recovering node

Example 2: SchoonerSQL Database



Benefits of Flash Optimizations in SchoonerSQL

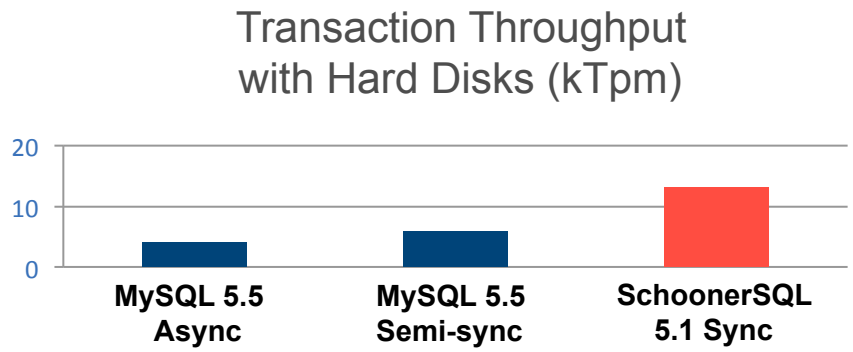
DBT2 open-source

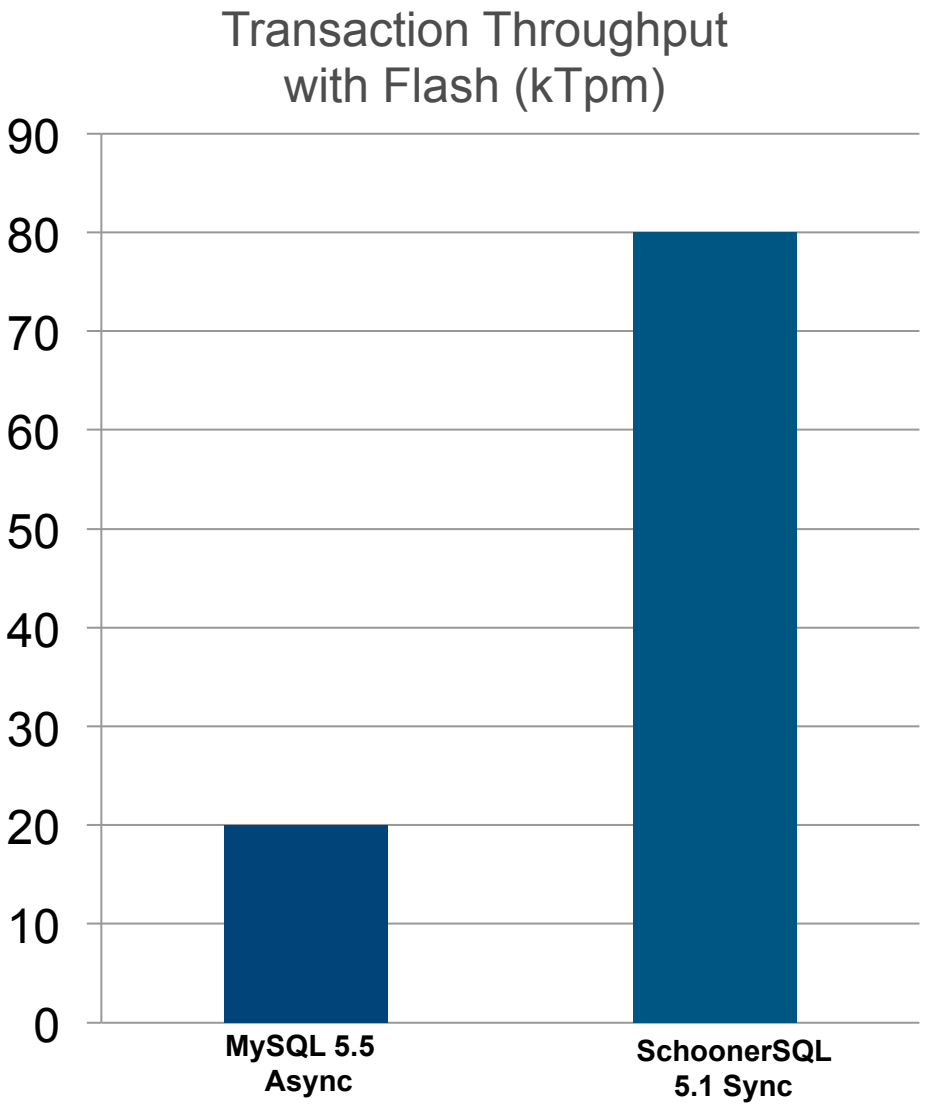
OLTP version of TPC-C

- 1000 warehouses, 32 connections
- 0 think-time
- Result metric: TPM (new order)

Measurement Configuration

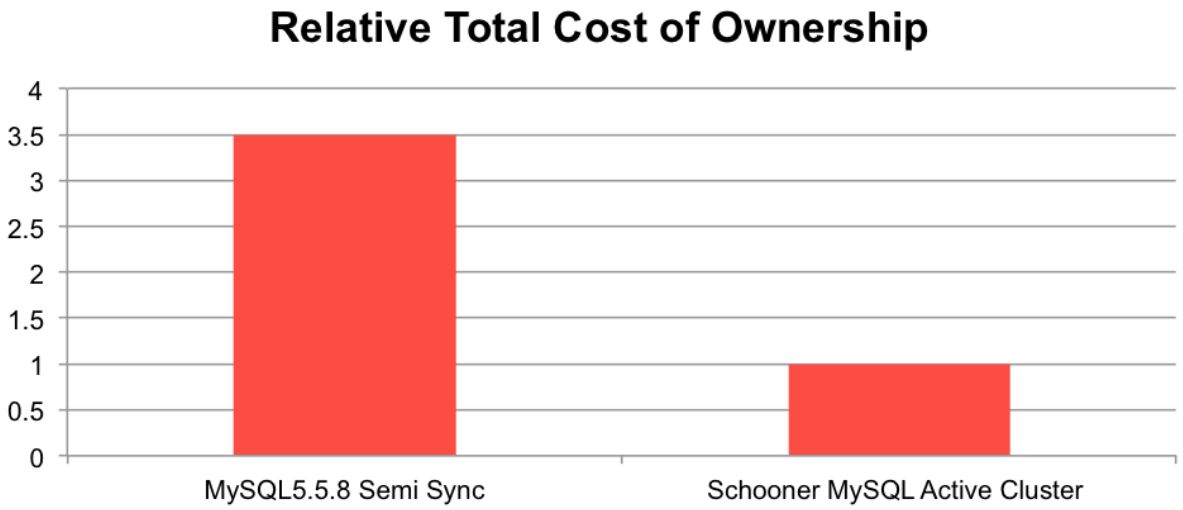
- 2 node Master-Slave configuration
- 2 socket Westmere
- 72GB DRAM





SchoonerSQL: Higher Performance Means Lower Cost of Ownership

- Reduced capital and operating costs through reduction in servers, power, space, admin



- TCO and ROI models are customer and workload specific
- Function (throughput/server; server, rack, and network costs, software license and support costs, admin costs; space and power costs; cost of downtime)

What Was Required to Exploit Flash?

- **Intelligent DRAM caching:**
  - Optimized buffer pool management algorithms
  - Scan resistance
  - Optimized double write buffer management
- **Locking optimizations to reduce lock granularity**
- **Efficient file syncing**
- **High Performance Synchronous Replication with fully automatic failover and failback:**
  - High throughput, low latency synchronous replication
  - Fully automatic failover using VIP's
  - Fully automatic failback when a failed node comes back online
  - Fast data transfer to recovering node
- **High Performance Asynchronous Replication**

Conclusion

- Commodity servers + flash + optimized software
  → more performance per watt, cubic foot, $$$
- Consolidation made possible with flash requires robust HA:
  - Need high performance replication
- Package common optimizations in an open SDK to simplify flash optimization of new applications

SanDisk Flash-Accelerated Products

Thank-you!