Software Optimization To Exploit Flash Memory

→ Tightly Integrated Balanced Systems

Dr John R. Busch
CTO and Founder
Schooner Information Technology, Inc

Abstract

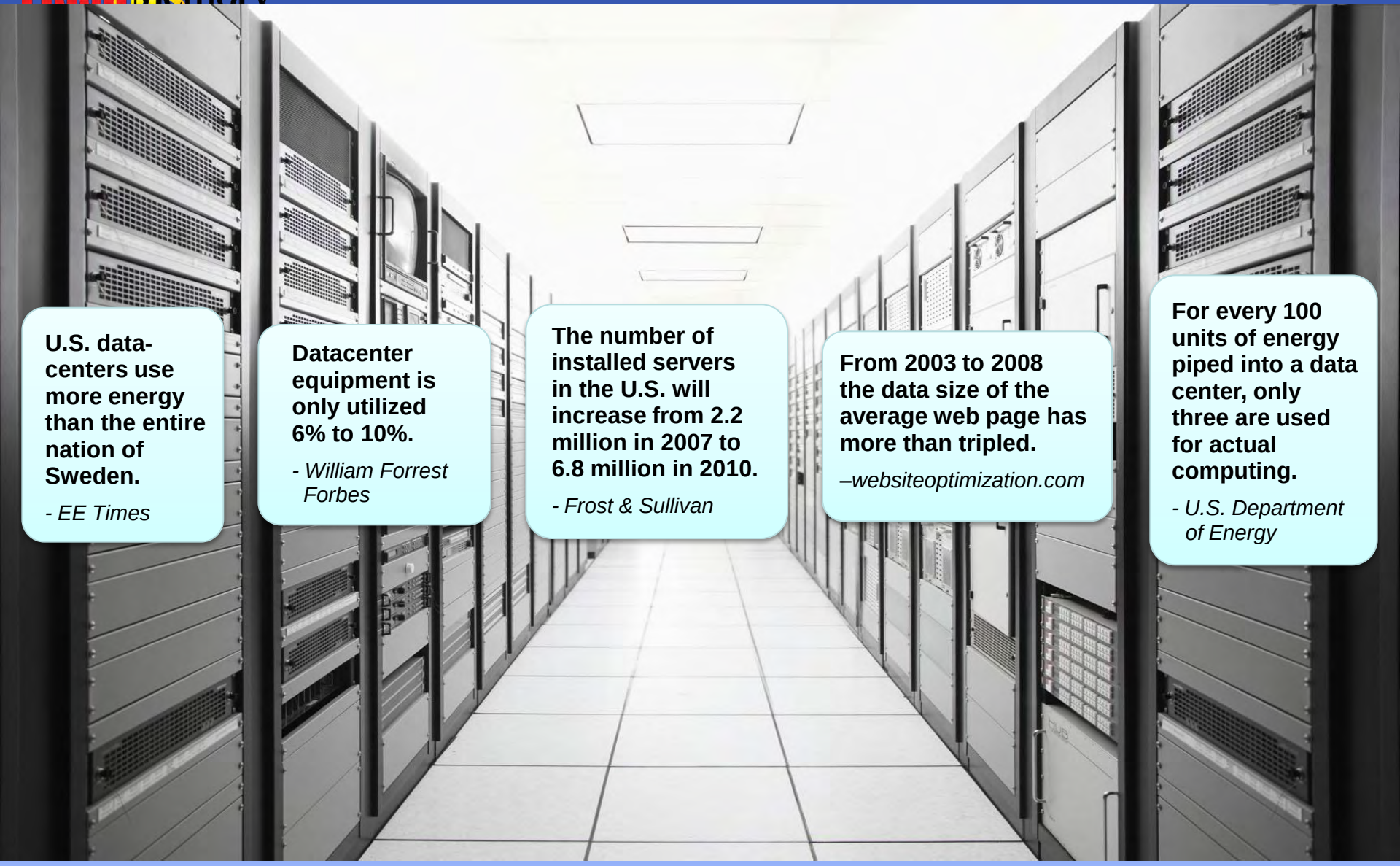
Software Optimization to Exploit Flash Memory
Dr John R. Busch
CTO and Founder
Schooner Information Technology, Inc
John.Busch@SchoonerInfoTech.com

Data center architectures based on servers with large DRAM caches and hard drive storage are highly inefficient. Flash memory offers the potential for order of magnitude improvements in data center performance, power consumption, and space usage. However, realizing this potential requires balanced system architecture, not just assembling locally optimized pieces. To create effectively balanced flash-based systems, software must be optimized for flash memory and for processor core scaling, with high levels of parallelism, granular concurrency control, intelligent memory hierarchy management, and specific consistency, balancing, and fault management algorithms tailored to flash characteristics. This talk focuses on balanced system architecture, design, and real world case studies of flash-based database, caching, and key-value store servers.

Agenda

- Datacenter trends and challenges
- Opportunity for flash
- Balanced Systems and Integrated Software
- Database case studies
- NoSQL case studies
- Flash and Cloud Computing

>> Rack, Power, Pipe, Complexity



U.S. data-centers use more energy than the entire nation of Sweden.

- EE Times

Datacenter equipment is only utilized 6% to 10%.

- William Forrest Forbes

The number of installed servers in the U.S. will increase from 2.2 million in 2007 to 6.8 million in 2010.

- Frost & Sullivan

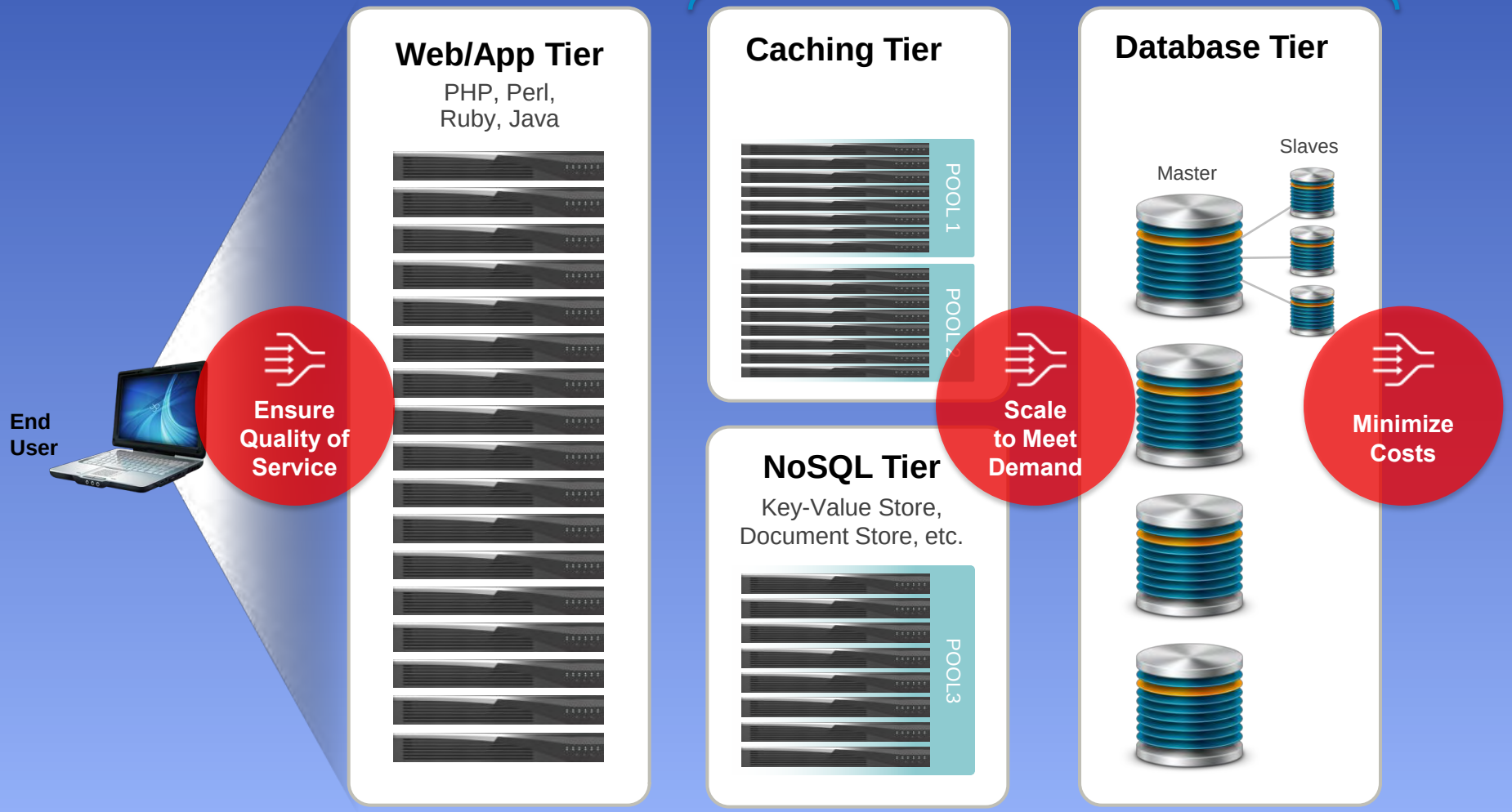
From 2003 to 2008 the data size of the average web page has more than tripled.

–websiteoptimization.com

For every 100 units of energy piped into a data center, only three are used for actual computing.

- U.S. Department of Energy

Typical Scale-Out Datacenter Deployment



Key Challenges

Scale to Meet Demand

- Pain of sharding and re-sharding
- Complexity of adding new servers
- Headache of managing server sprawl

Ensure Quality of Service

- Poor response time and availability
- Trade-offs in consistency models
- Complex, defensive app development

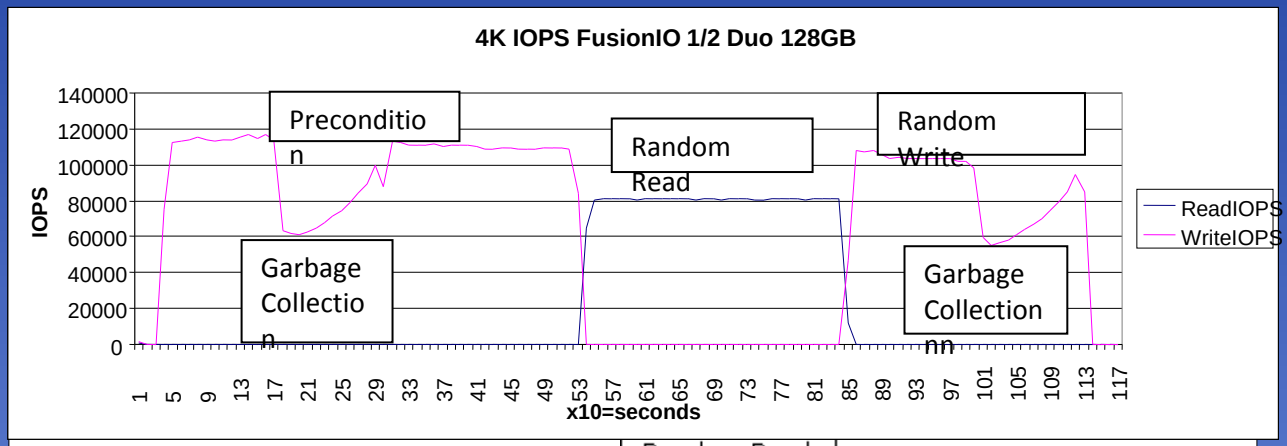
Minimize Costs

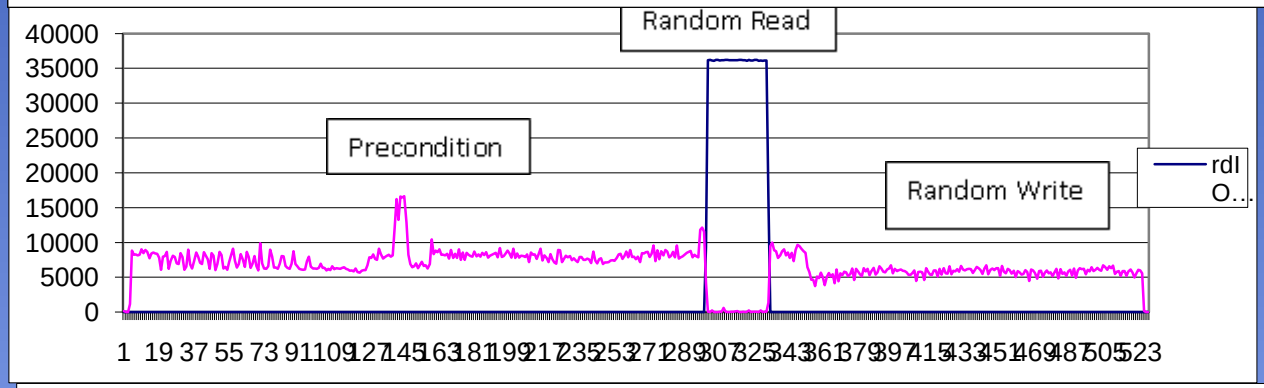
- Too much underutilized hardware
- Wasted power, pipe, and cooling
- Integration & management burden

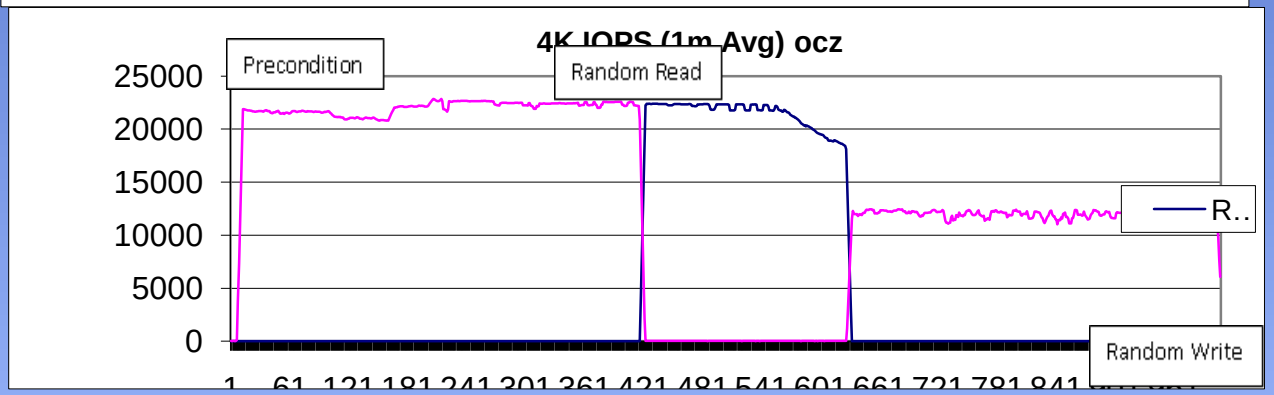
Flash : Performance Potential

- CPU L1 cache reference 0.5 ns
- CPU Branch mispredict 5 ns
- CPU L2 cache reference 7 ns
- Mutex lock/unlock 25 ns
- Main memory reference 100 ns
- Send 2K bytes over 1 Gbps network 20,000 ns
- Read from solid state media (SSD) 70,000 ns
- Read 1 MB sequentially from memory 250,000 ns
- Round trip within same datacenter 500,000 ns
- Disk seek (15000 rpm) 4,000,000 ns
- Disk seek (7200 rpm) 10,000,000 ns
- Read 1 MB sequentially from disk 20,000,000 ns
- Send packet US->NL->US 150,000,000 ns

Flash Drives: iozone micro-benchmarks







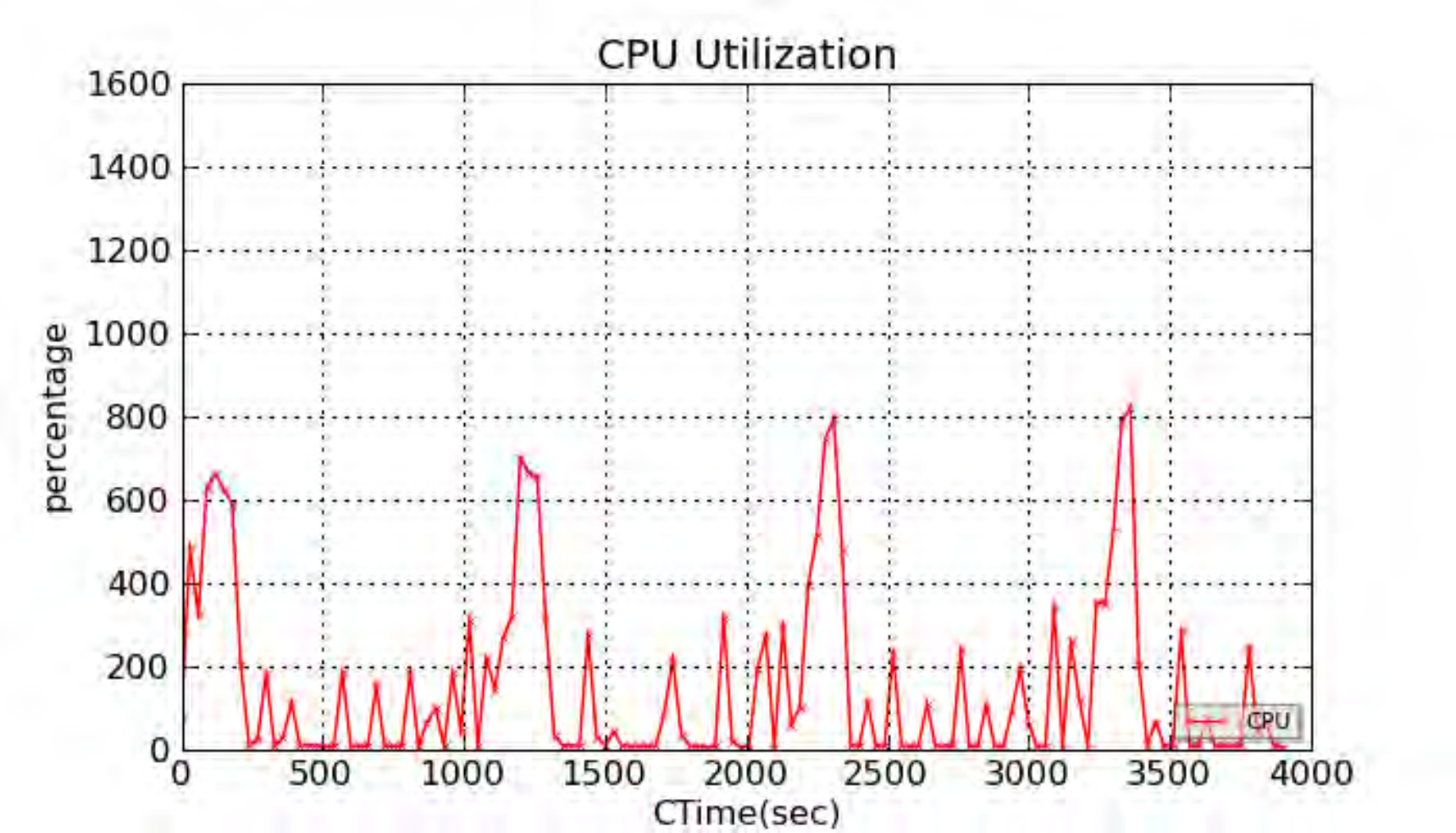
http://www.schoonerinfotech.com/blog/2010/06/28/schooner-labs-evaluation-and-optimization-of-database-technologies-flash-memory-multi-core-processors-software/

2U Server Flash Subsystem

	FusionIO Duo	Intel X25E	SandForce OCZDeneva
Capacity GB	440	488	1456
4KB Read IOPS	324,000	288,000	160,000
4KB Write IOPS	220,000	29,600	88,000
CPU time per I/O	19uS	7uS	7uS
Subsystem $/GB	$60	$14	$7
Storage Subsystem Cost	$26,360	$6,792	$10,600

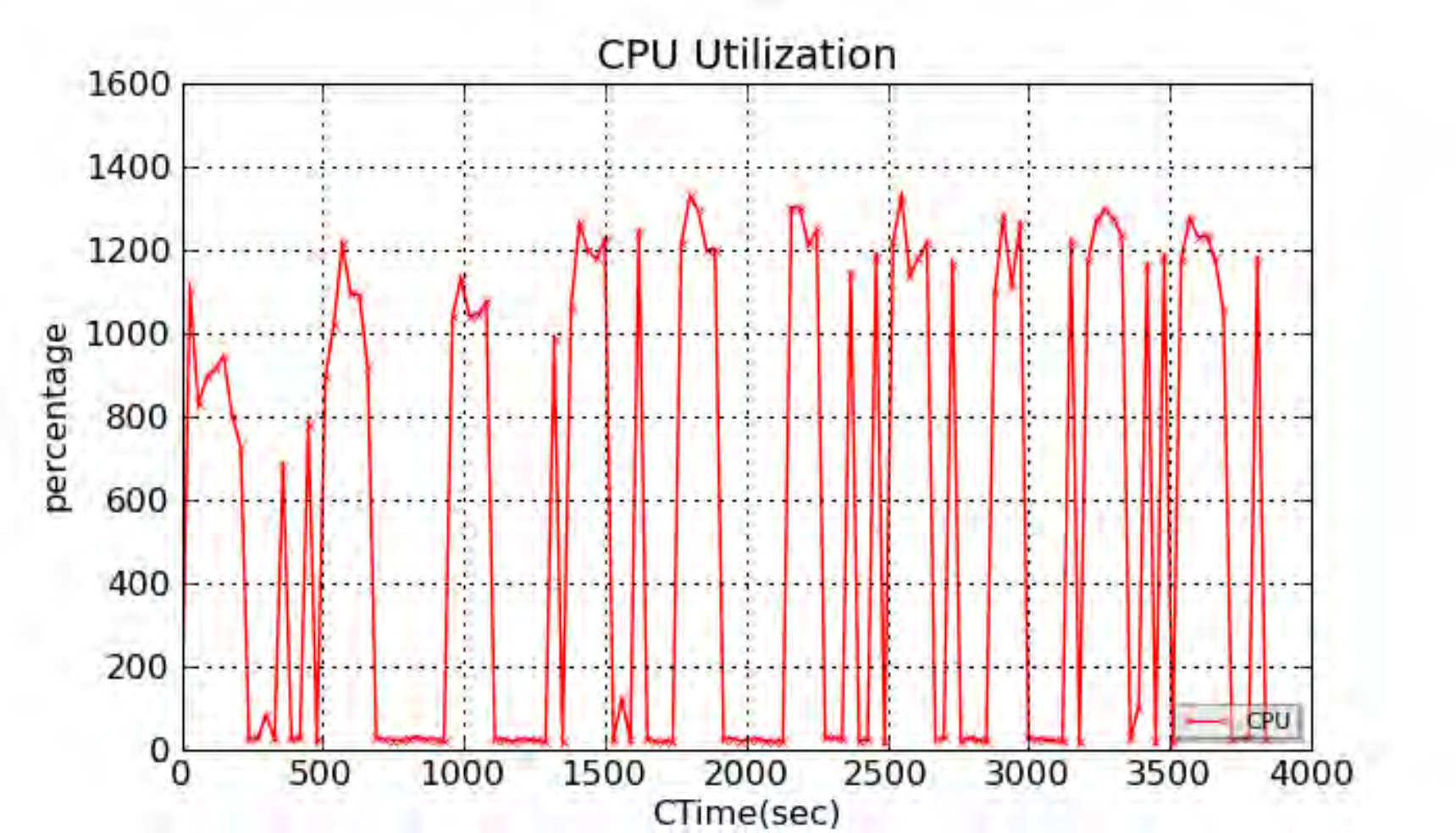
http://www.schoonerinfotech.com/blog/2010/06/28/schooner-labs-evaluation-and-optimization-of-database-technologies-flash-memory-multi-core-processors-software/

MySQL 5.1.44 on x8 Intel X25E SSDs (DBT2)



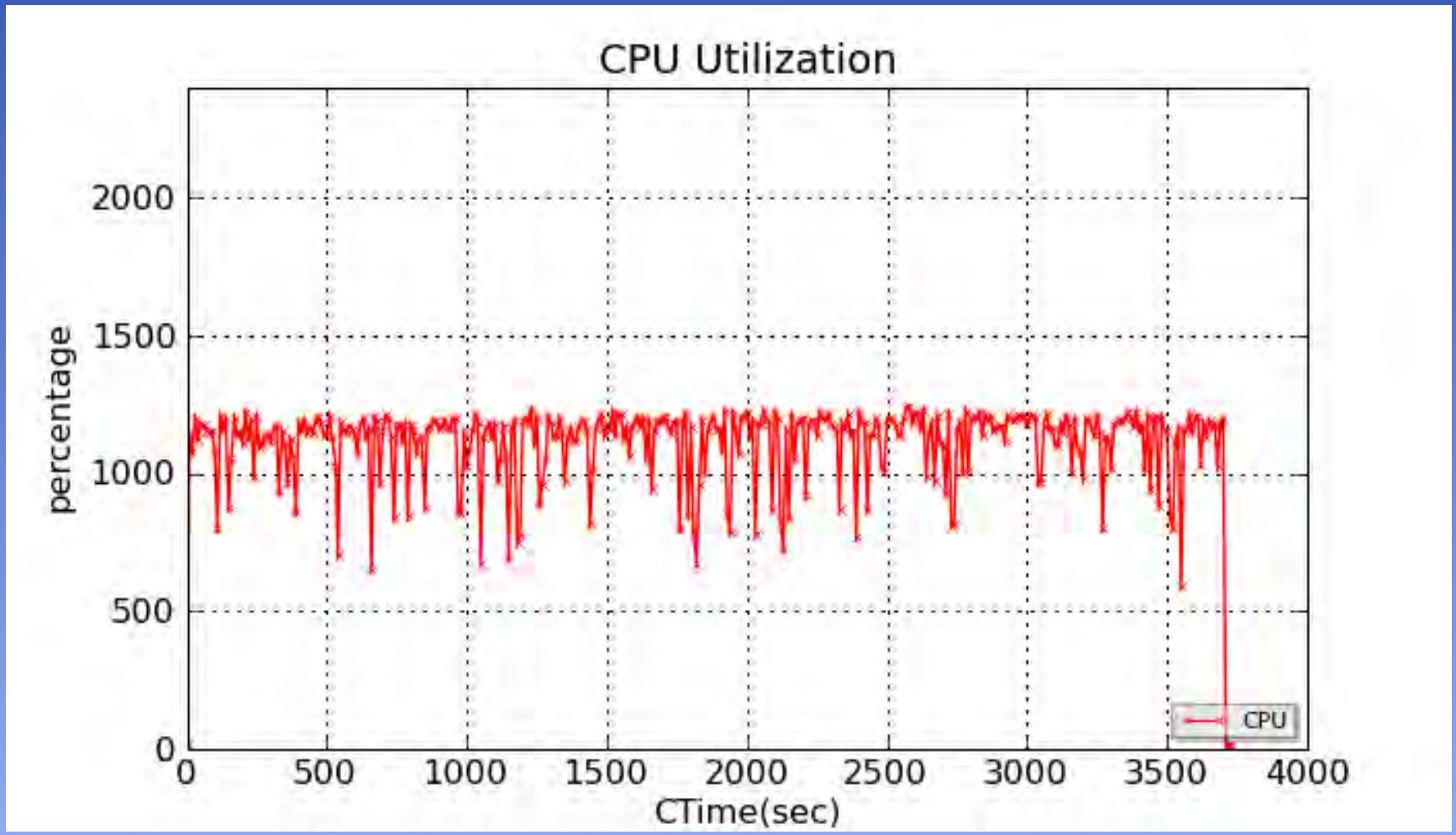
13.4k TPM

MySQL 5.5.4m3 on 2x Fusion-io Duo 320s (DBT2)



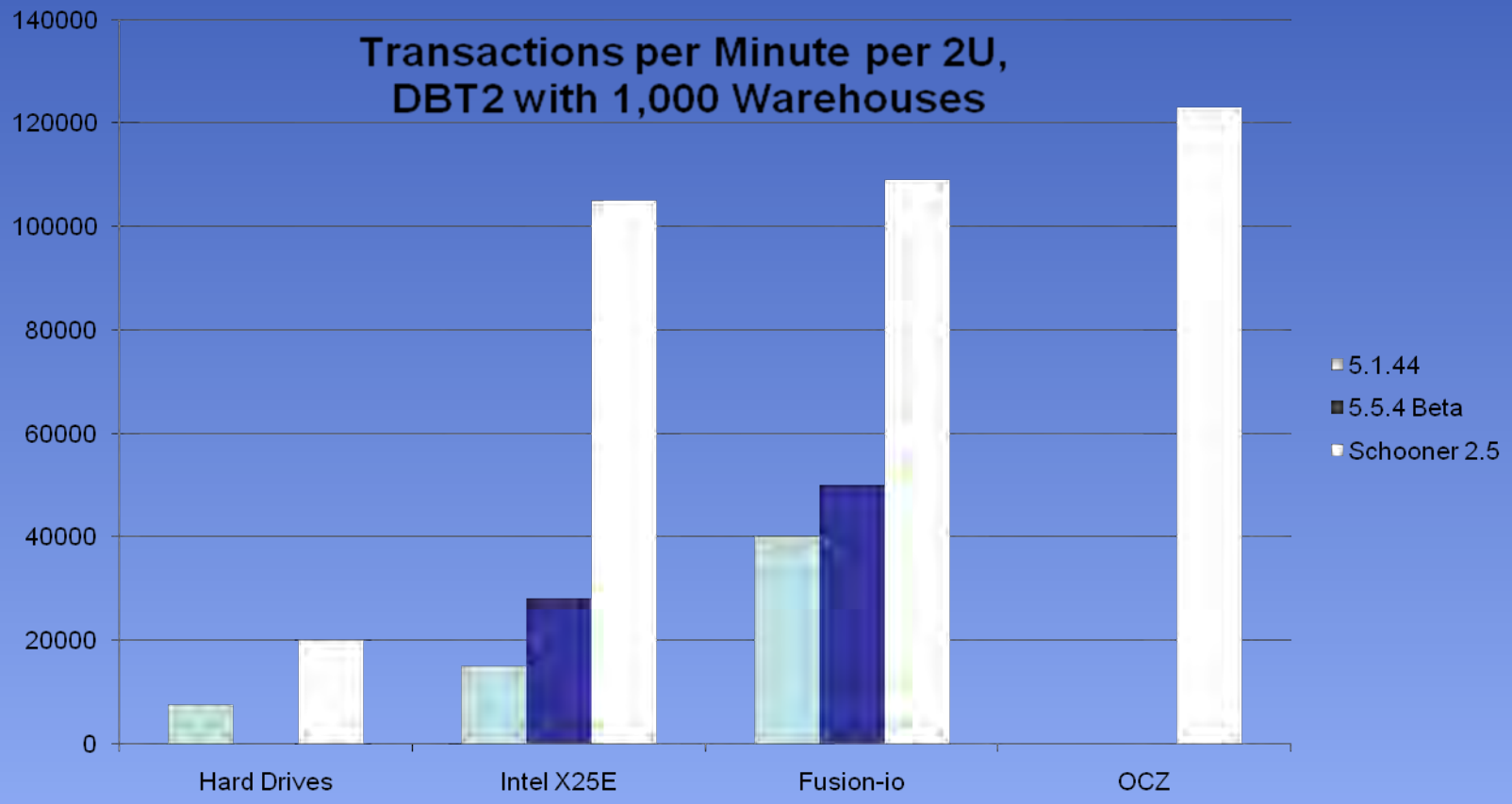
49.3k TPM

Schooner MySQLEnterprise on x8 Intel X25E SSDs (DBT2)



118.1k TPM

Tightly Integrated, Balanced Flash Based Database Systems : Performance Potential



IBM 3650 dual quadcore with 64GB DRAM + Hard Drives 12 hard disk drives (15 k RPM) configured in RAID 5.

•Intel SSDs : 8 Intel X25E solid state drives (SSDs) configured in RAID 5. Fusion-io 2 Fusion-io ioDrive Duo 320s configured in RAID 10

Opportunity : Tightly Coupled, Scalable Data Access Building Blocks

Integrated, optimized, scalable solutions:

- Effectively leverage flash memory, multi-core processors, high-speed networking, scalable data access software
- Incorporate highly optimized, balanced hardware platform, operating environment, integrated data access applications
- Provide efficient, higher level scalable building blocks
- Eliminate complex integration projects and leverage out of the box performance, scalability and availability
- Deliver enterprise class reliability

Tightly Integrated SW +HW Architecture

100% Client Compatibility

MySQL | **Memcached** | **Key-Value Store**

Schooner Operating Environment

Intelligent Caching Hierarchy

- Optimized buffer mgmt & scan-resistant algorithms
- Write-through and write-back caching
- Adaptive fine-grained memory management
- Efficient object meta-data for persistence

Optimized Flash-Memory Access

- Highly parallel read- and write-access
- Intelligent flash-wear algorithms
- Durability with high performance
- High-performance, integrated RAID

Multi-Core Scalability

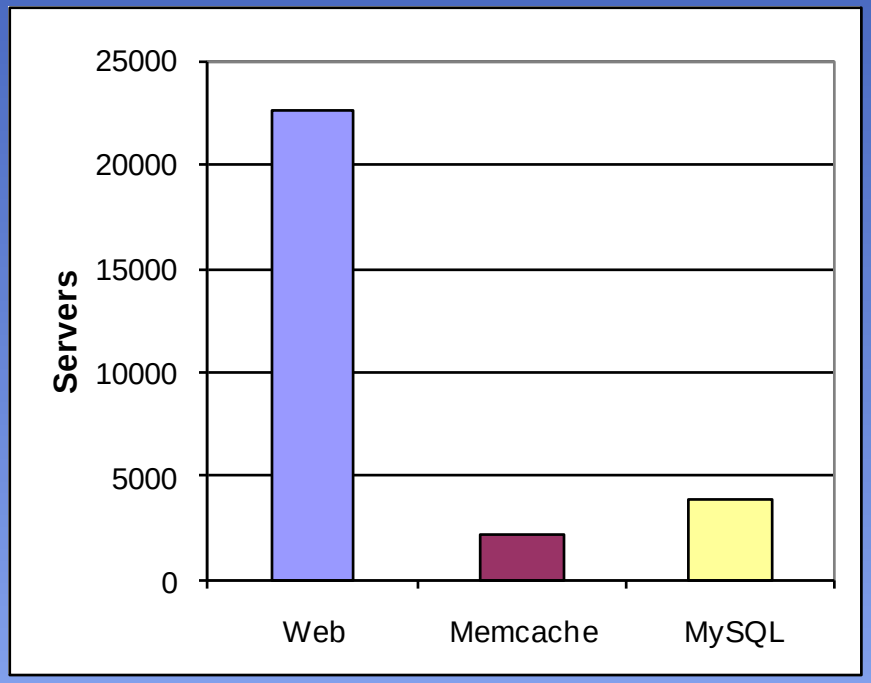
- Fine-grain locking
- Scalable and concurrent data structures
- Optimized thread-to-core allocation
- Efficient handling of network interrupts

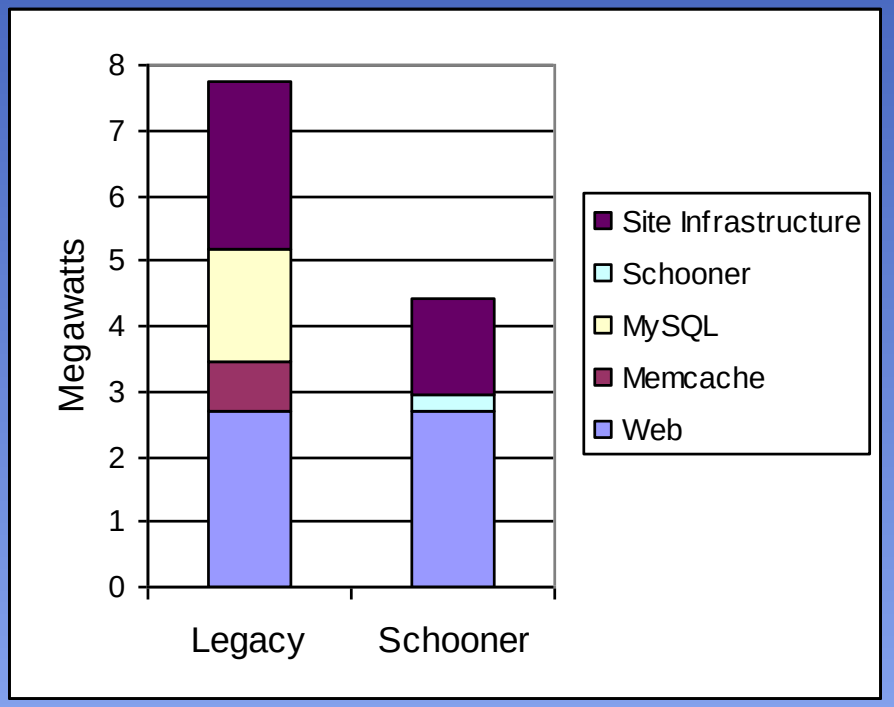
Transparent HA/DR

- Synchronous and asynchronous replication
- Failure detection and automated VIP failover
- Fast, incremental data recovery
- Incremental/full online backup and restore

Optimized, Balanced Hardware Platform

Tightly Integrated, Balanced Flash Based Systems : Data Center Power Reduction





Database Case Study

Schooner Appliance for MySQL Enterprise™ with InnoDB

MySQL Flash Memory SUMMIT

High Performance

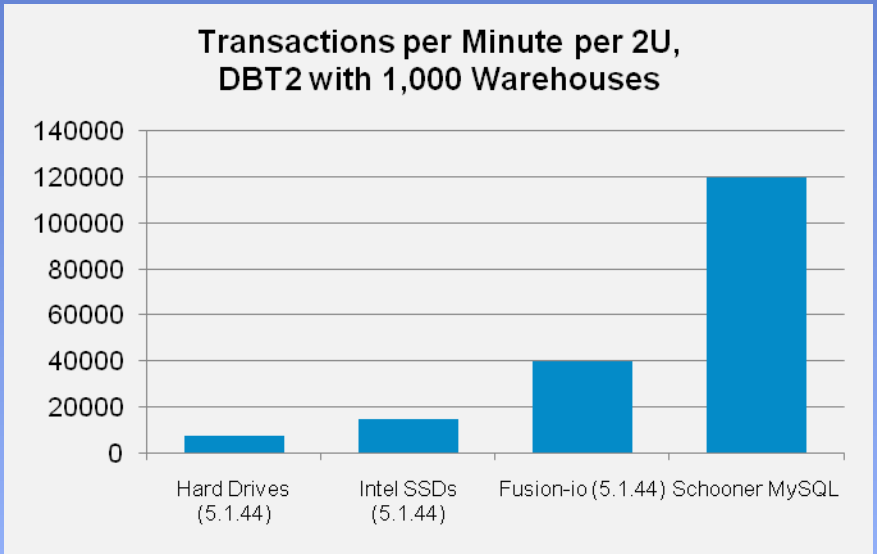
- Highly-parallel optimized flash-memory access
- Advanced buffer-pool caching algorithms
- Multi-core scalability with fine-grained locking
- Delivered on a proven IBM server with ½ TB of flash

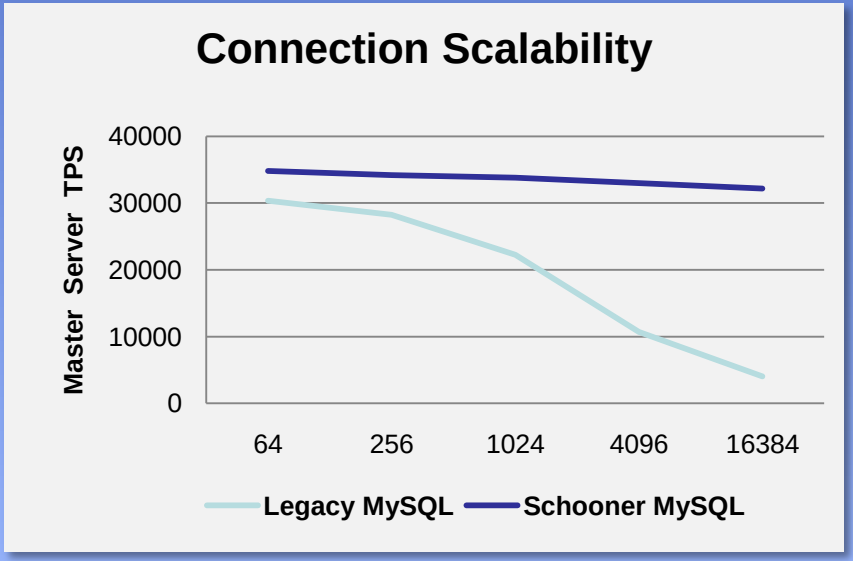
High Availability

- Fully ACID-compliant with data durability
- Integrated replication and automated failover
- Integrated high-performance backup and restore
- RAID across SSDs and HDDs

Turnkey Appliance

- Multi-instance consolidation on single appliance
- Web-based GUI/CLI for centralized management
- Integration with 3rd-party mgmt & monitoring tools
- 100% compatible and fully certified by Oracle/MySQL

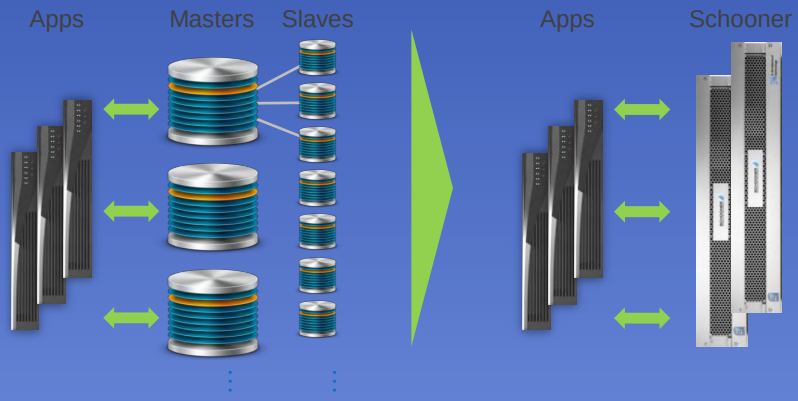




MySQL Flash Memory SUMMIT

What Can I Do With It?

Reduce sharding and consolidate slaves



Scale performance and process queries faster

70K TPM (DBT2)
20K Connections

½ TB or 1TB Flash

Reduce planned and unplanned downtime

Replication + transparent Failover

SSD RAID & Data Durability

SSD RAID & Data Durability

SSD RAID & Data Durability

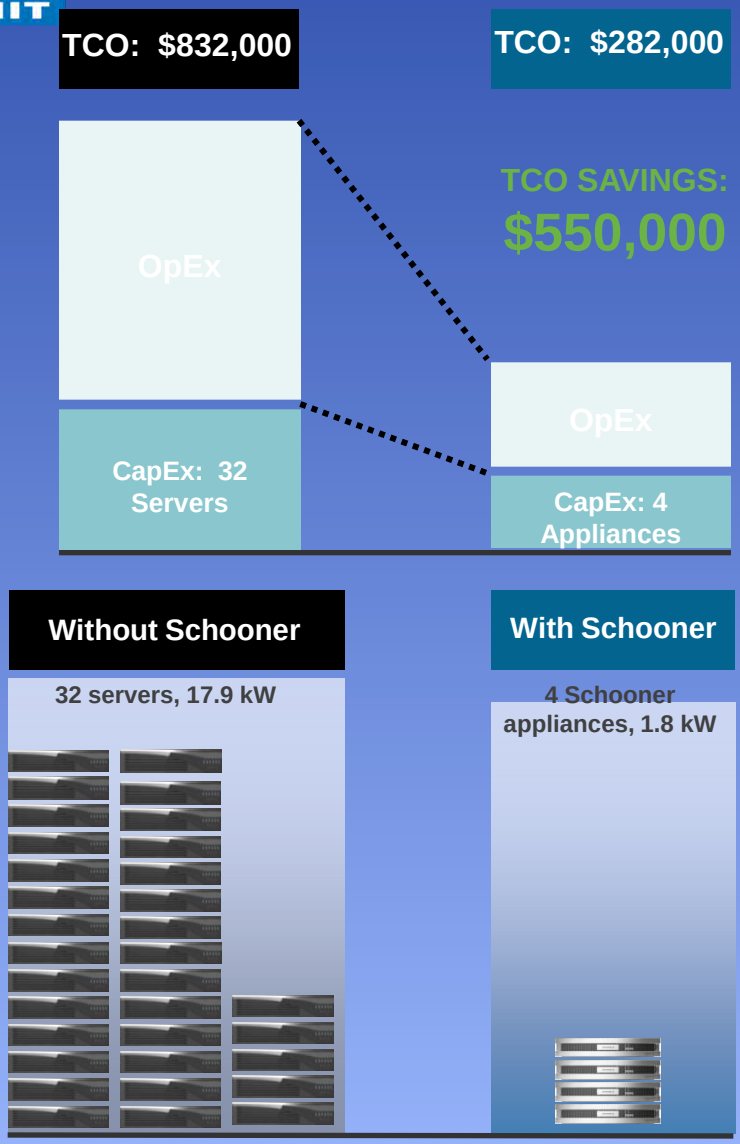
Eliminate integration and optimization headaches

- Software
- Hardware
- Support
- Certified
- Complete

MySQL Database Consolidation and Cost Savings

Flash Memory SUMMIT

3 Year TCO (2 TB MySQL)



THE BOTTOM LINE

- Immediate capex savings
- 66 % TCO savings ($550,000) over 3 years
- Power & space reductions enable green datacenter

MySQL Flash Memory SUMMIT

MySQL Customer Examples

"Schooner is the perfect solution for any MySQL enterprise whose business success requires great performance, exceptional reliability, and the ability to smoothly scale the datacenter as demand increases. Schooner helps us create a new wave of social networks, bringing technology that helps us create and sustain social communities like never before — efficiently, effectively, and effortlessly."

– Rayes Lemmens, CEO at MyLivePage

"In our business, Website performance and efficiency is key to the success of our Web properties. The Schooner MySQL Appliances have significantly helped GuteFrage improve their overall Website response time while at the same time allowing them to consolidate their database slaves onto a single Schooner appliance, dramatically reducing the time necessary for database administration."

– Frank Penning, CTO of Holzbrinck Digital

"Our ad hoc MySQL queries run at least five times faster after installing the Schooner Appliances. They deliver a huge performance benefit and are a breeze to install and manage."

– Darryl Weatherspoon, VP of Eng at Xoom

We explored a variety of options from commodity SSD drives to PCI-express based flash memory cards. We decided to purchase Schooner MySQL appliancess They produce an awesome appliance and the performance has been great..

- Mark Imbriaco DBA 37signals

NoSQL Case Study

Memcached / NoSQL

Schooner Appliance for Memcached/NoSQL

Performance

- Highly-parallel, optimized flash-memory access
- Fast, efficient DRAM-to-flash caching algorithms
- Multi-core scalability with parallel thread allocation
- Delivered on a proven IBM server with ½ TB of flash

Extreme Availability

- Modes for pure cache and persistent key-value store
- Transparent replication and automated failover
- Non-disruptive, rolling upgrades
- RAID & full/incremental backup and restore

Turnkey Appliance

- Dynamic containers for consolidation & multi-tenancy
- Web-based GUI/CLI for centralized management
- Integration with 3rd-party mgmt & monitoring tools
- 100% compatible and fully memcapable compliant

Memcached Caching Service
- 144 GB capacity, 250K TPS
- Replication & Failover
- 100% Memcapable

+

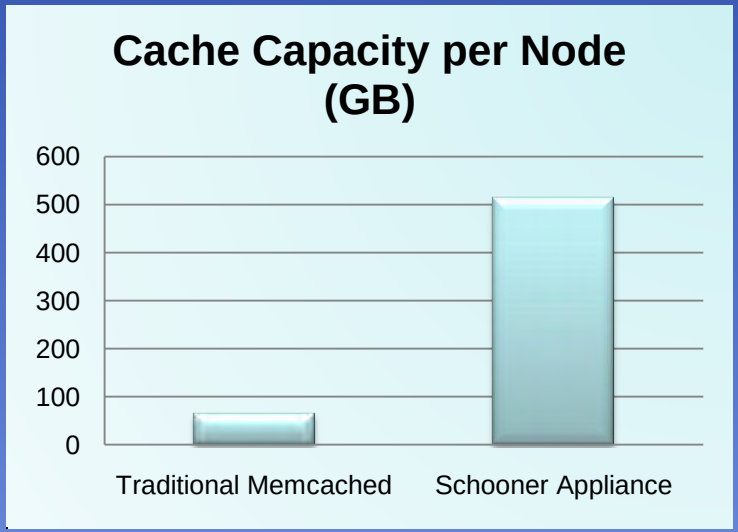
Key-Value Store Service
- 368 GB capacity, 150K TPS
- Replication & Failover
- Persistence, RAID, and Backup
- Memcached-based key-value API

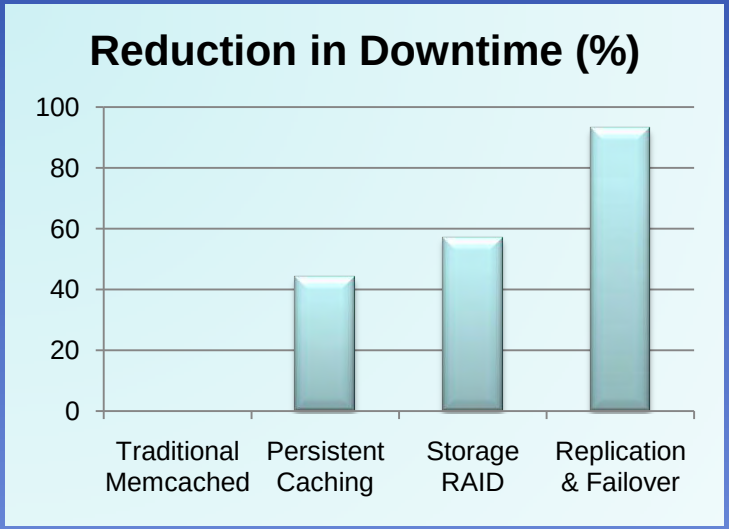
Memcached / NoSQL

Flash Memory SUMMIT

Performance, Capacity, and Availability vs. NoSQL Alternatives

Schooner vs. Traditional Memcached





Schooner vs. NoSQL Alternatives

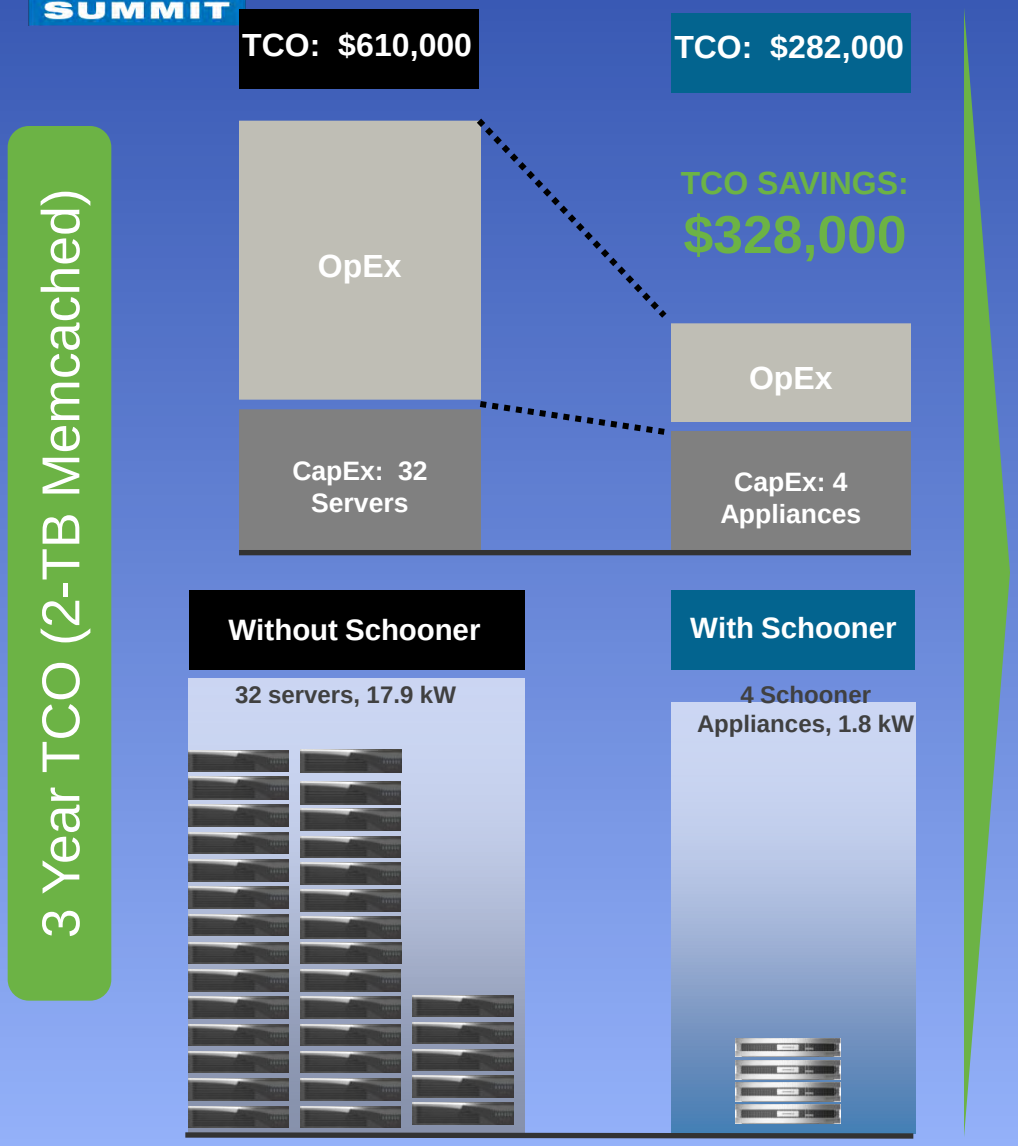
TPS/Node, Random Queries	In DRAM	In Flash
CouchDB	1,000	1,100
Cassandra	10,500	1,790
MongoDB	49,000	4,000
Schooner MySQL	115,000	101,000
Schooner NoSQL	310,000	160,000

Note: NoSQL benchmark is a key-value random query of 32M and 64M 1kByte items, on the same hardware (dual quad-core Intel Nehalem processors with 64 GB of DRAM and 8 parallel Intel X25E flash drives).

Typical Memcached/NoSQL Consolidation and Cost Savings

Memcached / NoSQL

Flash Memory SUMMIT



THE BOTTOM LINE

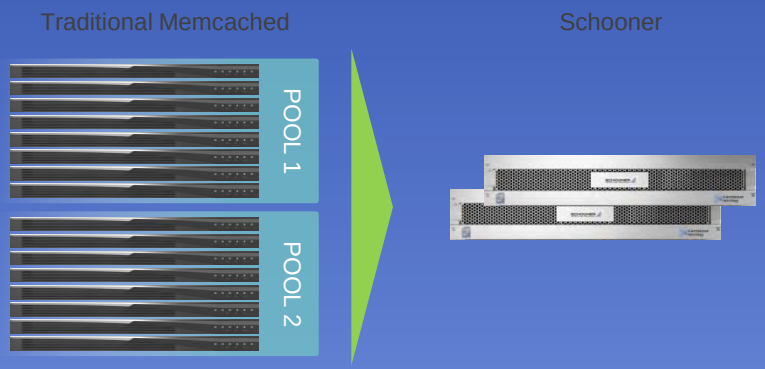
- Immediate capex savings
- 54% TCO savings ($328,000) over 3 years
- Power & space reductions enable green datacenters

Memcached/
NoSQL

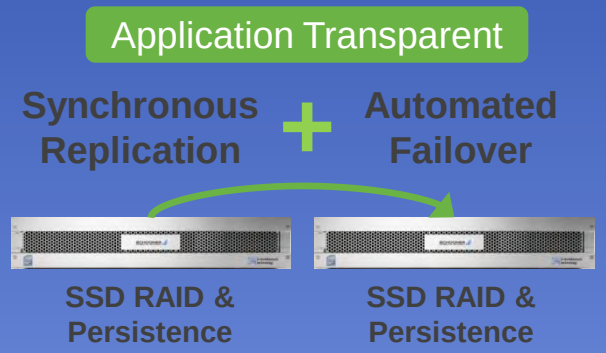
Flash Memory SUMMIT

What Can I Do With It?

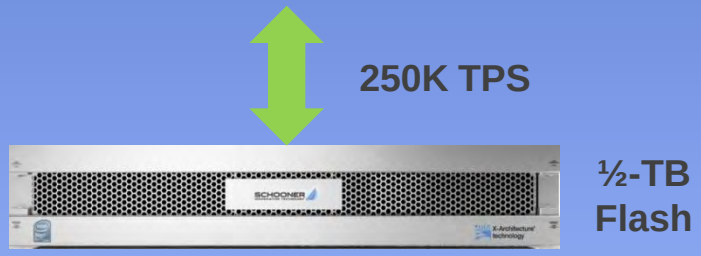
Consolidate to reduce server sprawl



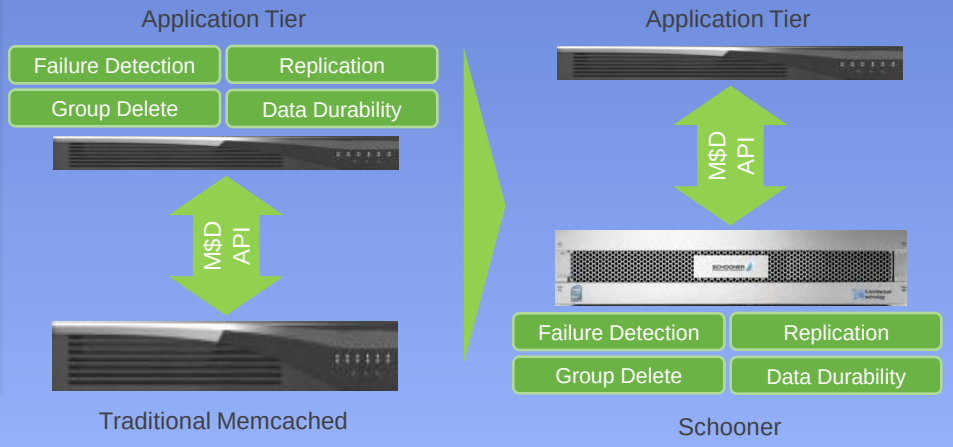
Reduce planned and unplanned downtime



Scale cache capacity and process requests faster



Reduce complexity of application development



NoSQL Customer Examples

Memcached / NoSQL

Flash Memory SUMMIT

CCP

Caching backbone of Eve and future gaming platform

Large capacity persistent, available scalable caching service insures performance + eases application development

Flixster
Watch movies. Tell friends.

"Scaling the data tier is a common challenge, and Schooner is helping us do just that. Power is the big constraint right now, so anything we can do to reduce that footprint right now is helpful. From an administrative perspective, fewer machines is always better, and it also means reductions in potential failures due to fewer boxes."
– Saran Chari, CTO and Founder at Flixster

plaxo

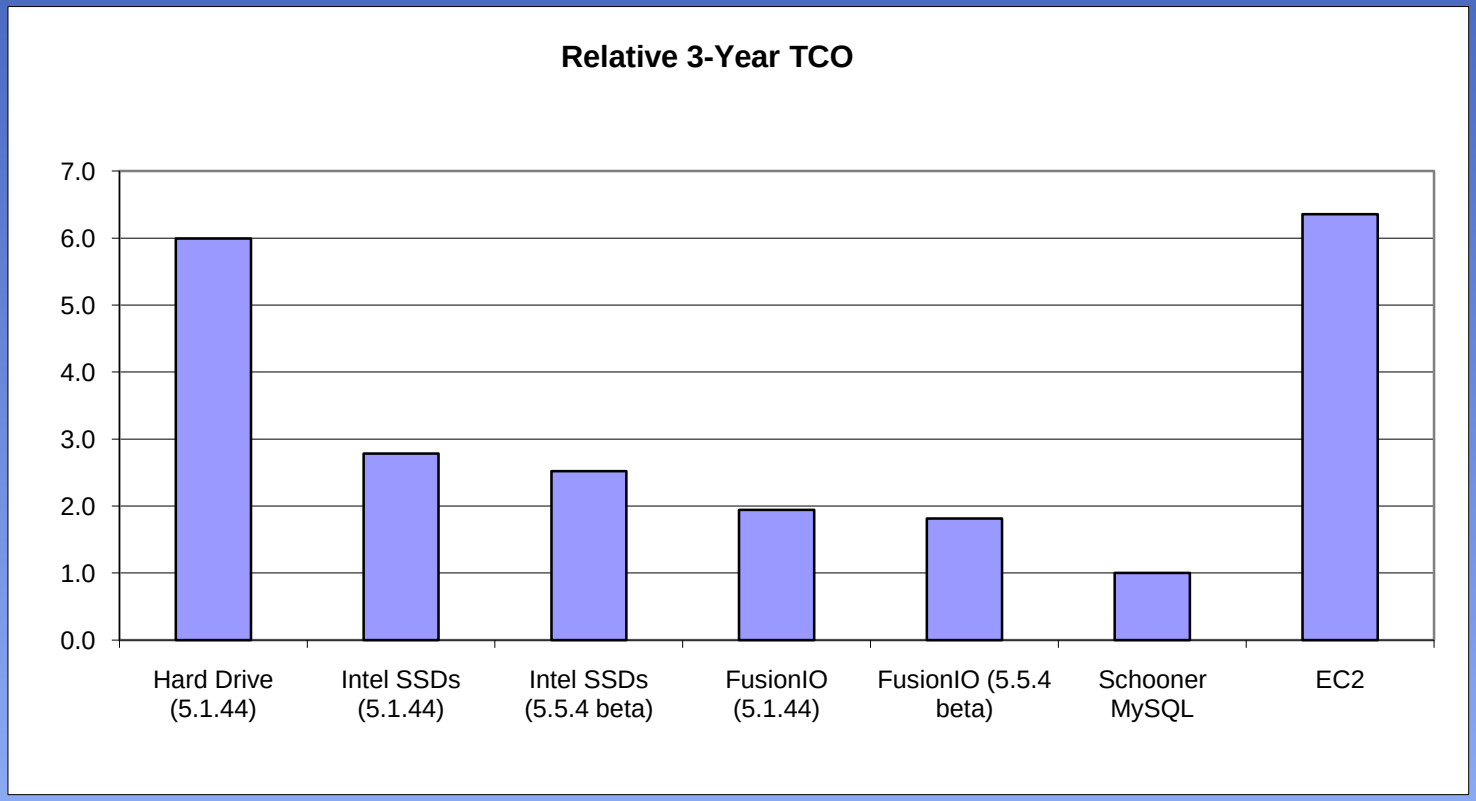
In the wonderful Schooner world, failovers go away. Schooner replication means that you're sure that what you have on one node will also be on the other. Our developers don't have to worry about cache coherency. They can plan on the data being available, so they don't have to program defensively.
– Ethan Erchinger, Director of Ops at Plaxo

iStockphoto

10:1 Consolidation

Performance AND Availability
Synchronous Replication/Transparent Failover are key

Cloud
→ Hybrid



Q&A

Benchmarking Configuration

- Hardware
  - 2 Quad Core 5560 Xeons (with 2 HT per core, 16 CPUs via Linux)
  - 8 Intel X25E SSDs, 2 Fusion-io Duo 320, 8 SF OCZ
  - 64GB Memory
- DBT2 benchmark – osdldbt.sf.net
  - 1000 warehouses – 100G data
  - 32 connections
  - Zero think time
- MySQL Configuration
  - innodb_buffer_pool_size = 48G
  - innodb_flush_log_at_trx_commit = 1

About the DBT2 benchmark

- open-source database benchmark ~TPCC™
  - Focuses on OLTP (online-transaction processing)
  - Scales with data-size, includes ramp-up and steady state
  - High write rate, requires good locality in buffer pool
- Transactions with Select, Update, Insert, Delete
- Throughput metric: TPM (New Order)
  - Transaction Ratios: New Order 45% (with 1% rollback), Payment 43%, Stock Level 4%, Order Status 4%, Delivery 4%
- Results include TPM, response time (avg and 90th %ile), CPU, iostat, etc.