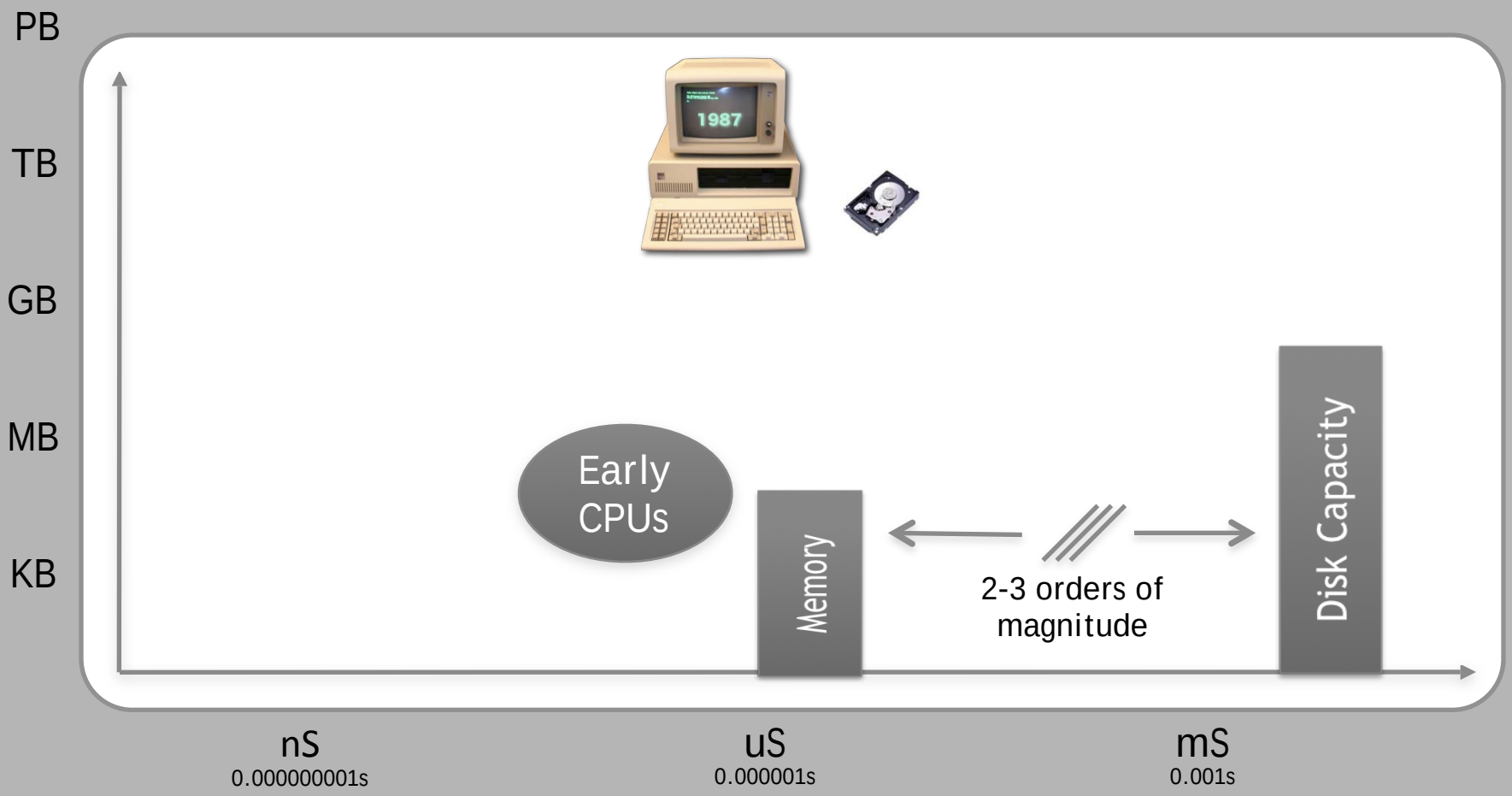
NAND Flash in the Enterprise

David Flynn
CTO Fusion-io

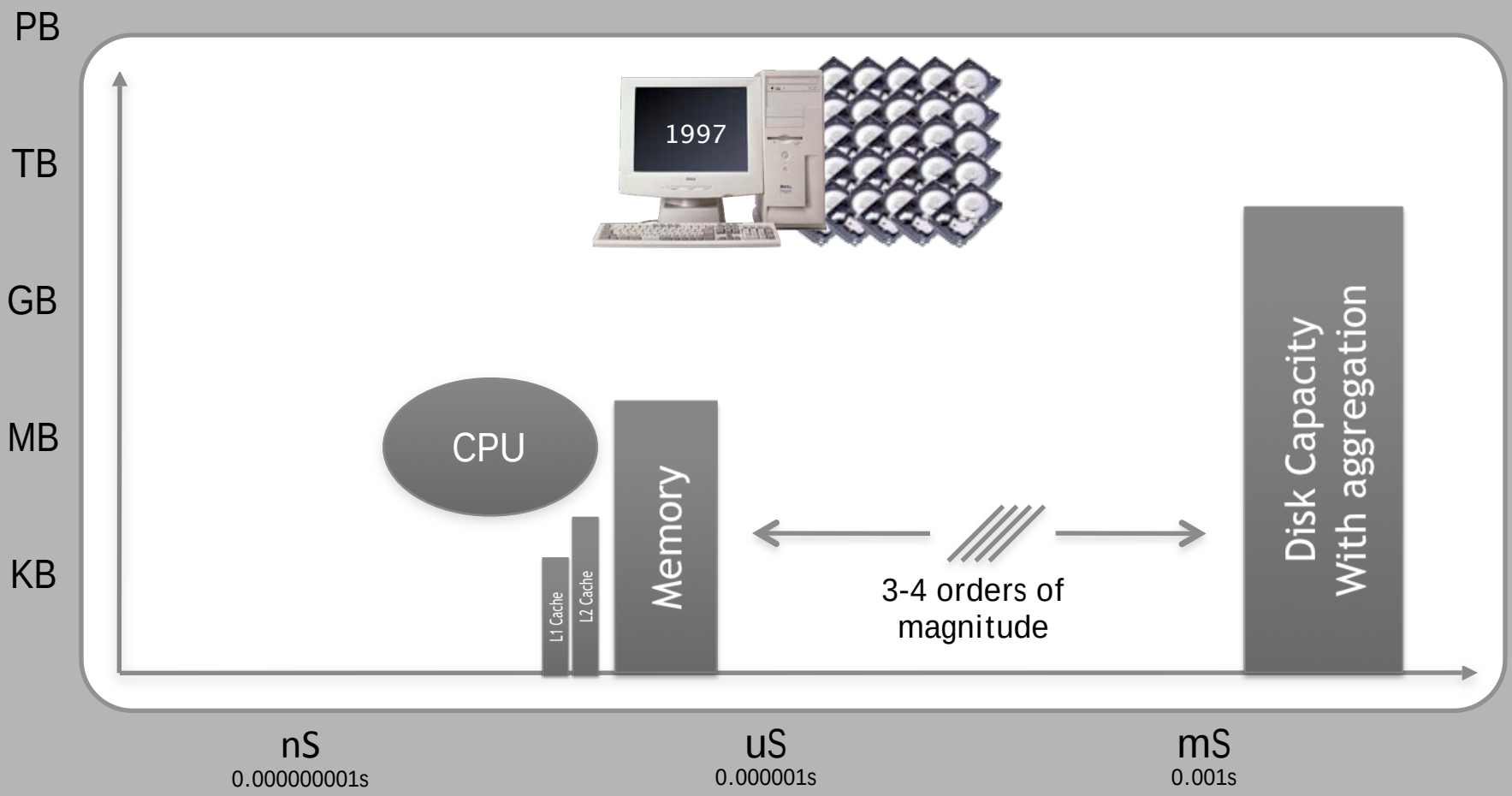
February 2009

Where we started



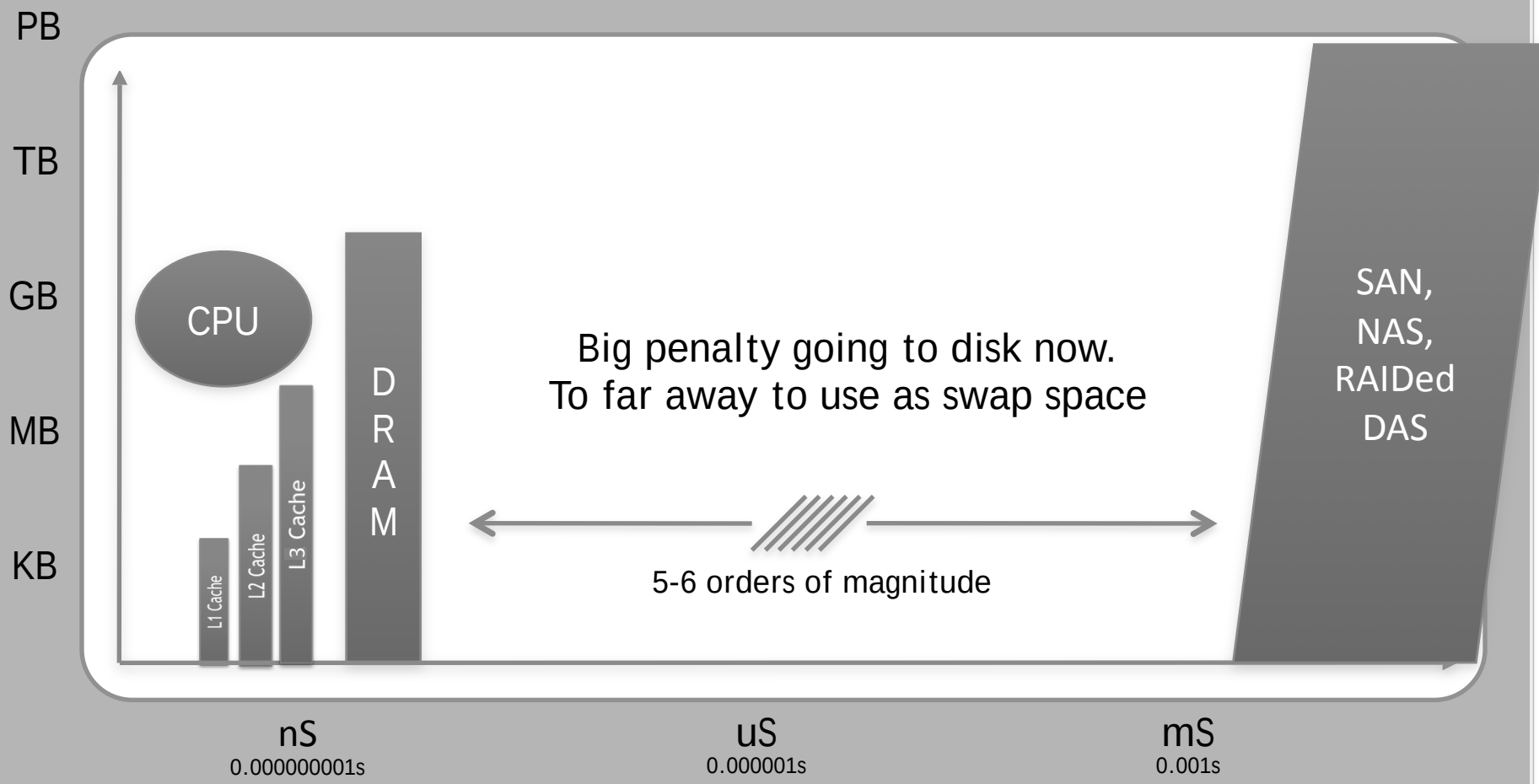
Access delay in time

Where we went



Access delay in time

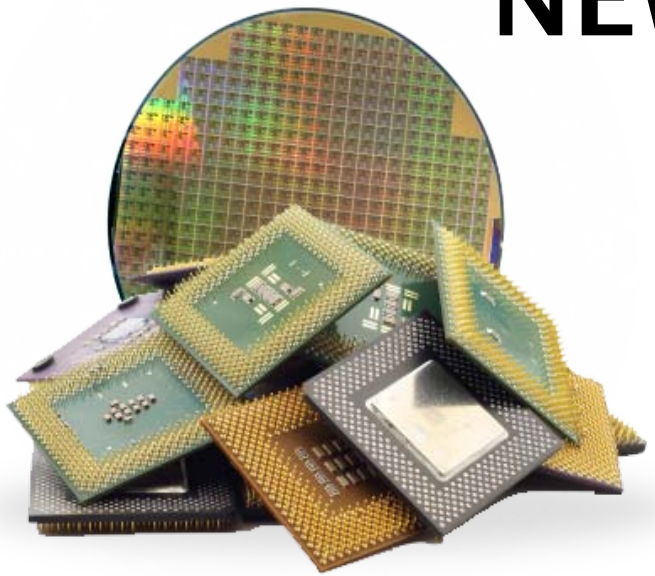
Where we are today



Access delay in time

it's

MOORE's LAW vs.
NEWTON's LAWS



today processors are

2,000,000 TIMES FASTER

disk seek time is only

12 TIMES FASTER

if 20 years ago

it was like going a

FEW MILES

to a **7-ELEVEN**

today it's like going

240 THOUSAND MILES

to the **MOON**

Newton lost

CPU cores sit idle

We need a

NEW CORNER MARKET

We need a

NEW MEMORY TIER

one that follows

MOORE's LAW

That **NEW MEMORY TIER**

is **NAND FLASH**

Why **NOW**

NAND has been around forever

Why Now

- Market Drivers
  - Thumb drives, cameras, MP3 players drove volumes
  - Cell phones and laptops now accelerating adoption
  - Each year more bits of NAND ship than DRAM ever has
  - Each year more than twice as many NAND bits ship
- Results
  - Price dropped by 60% each of the last three years
  - Price expected to continue drop 50% per year
  - Capacity will continue to double each year

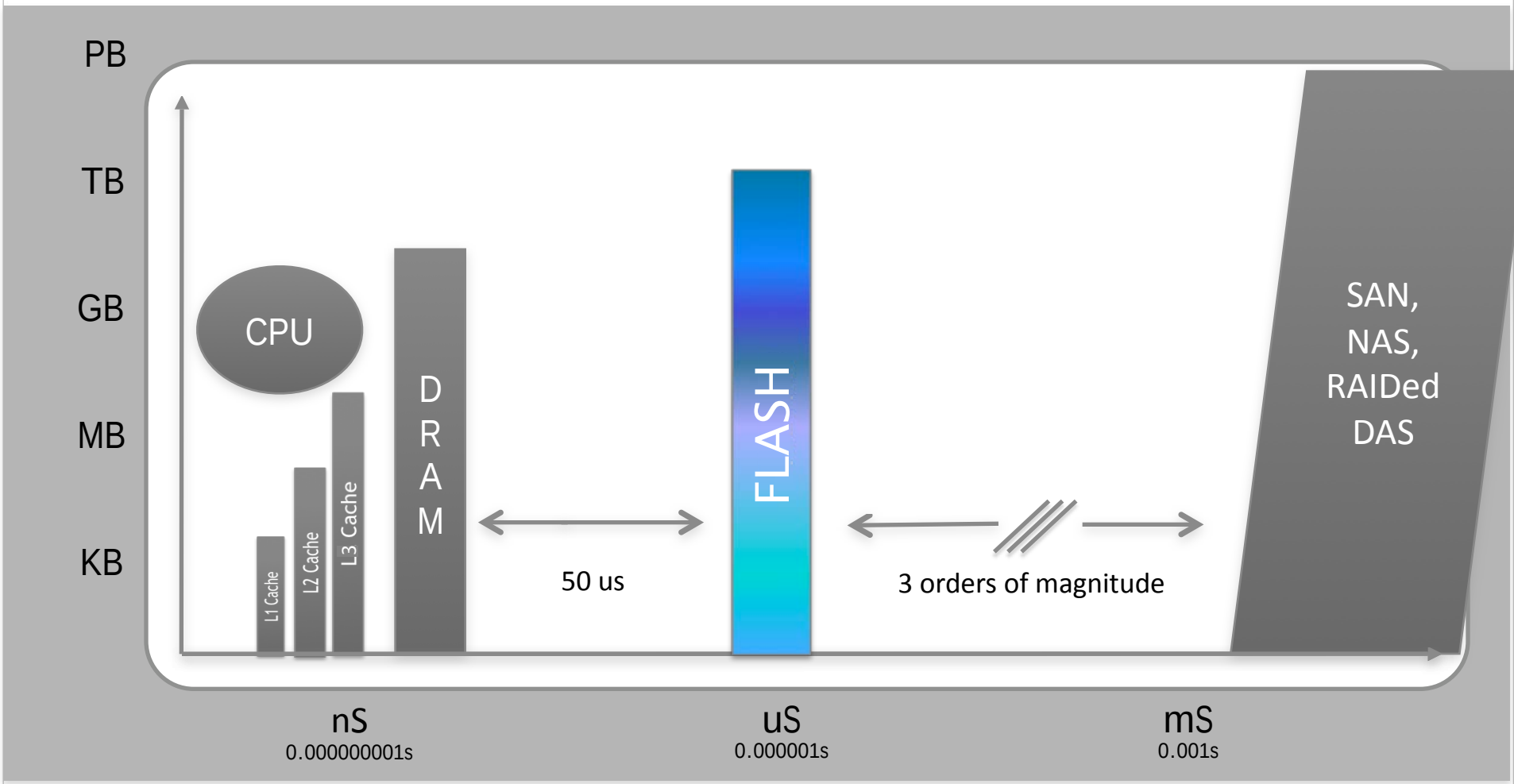
Flash Compared to DRAM – Strengths

- Non-volatile
- Similar bandwidth
- 10x Less expensive per GB
- 100x less power & heat
- 100x capacity per module
  - 1.5x cell density (simpler design)
  - 12 to 18 months ahead on manufacturing processes
  - Multiple bits per cell (with MLC)
  - Die stacking within chip (quad/octal die pack)
  - Chip stacking on module (dual chip stacks)

Flash Compared to DRAM – Weaknesses

- Higher latency read access (25us)
- Bulk write required
  - Erase required before program
  - Program takes 200us
  - Erase takes 2,000us
- Wear-out
  - SLC 100,000 to 500,000 cycles per cell
  - MLC 10,000 to 50,000 cycles per cell
- Failures too probable
  - Newest semiconductor fab process
  - Smallest feature sizes
  - Shared control lines
  - 20V internal
- Indirection required (Management)

A New Memory Tier



Access delay in time

how to integrate

FLASH

into the

MEMORY HIERARCHY?

put it close to the CPU on the

SURFACE STREETS

not into **ORBIT**

on the

SYSTEM BUS

not into

HDD infrastructure

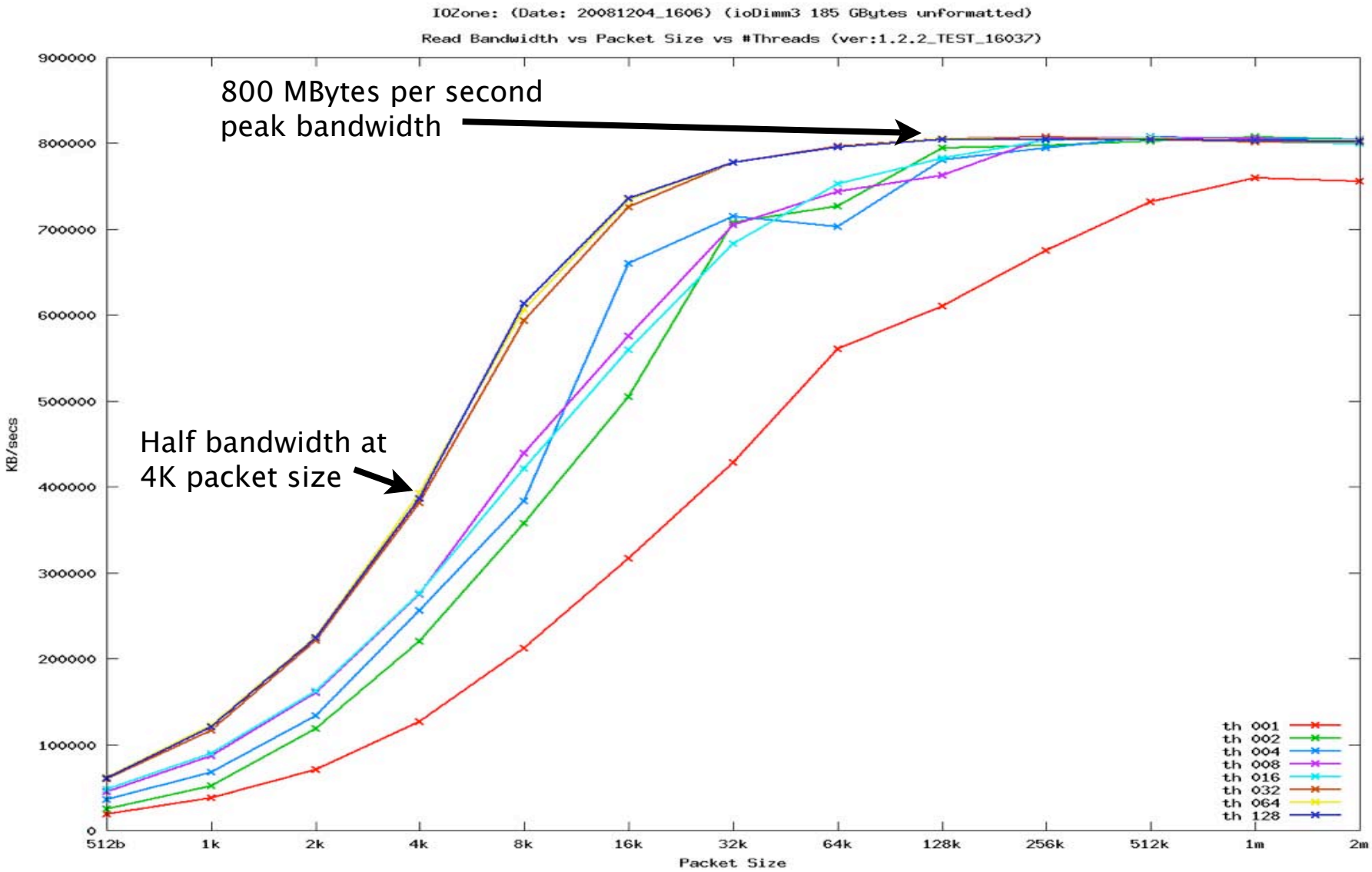
because, from

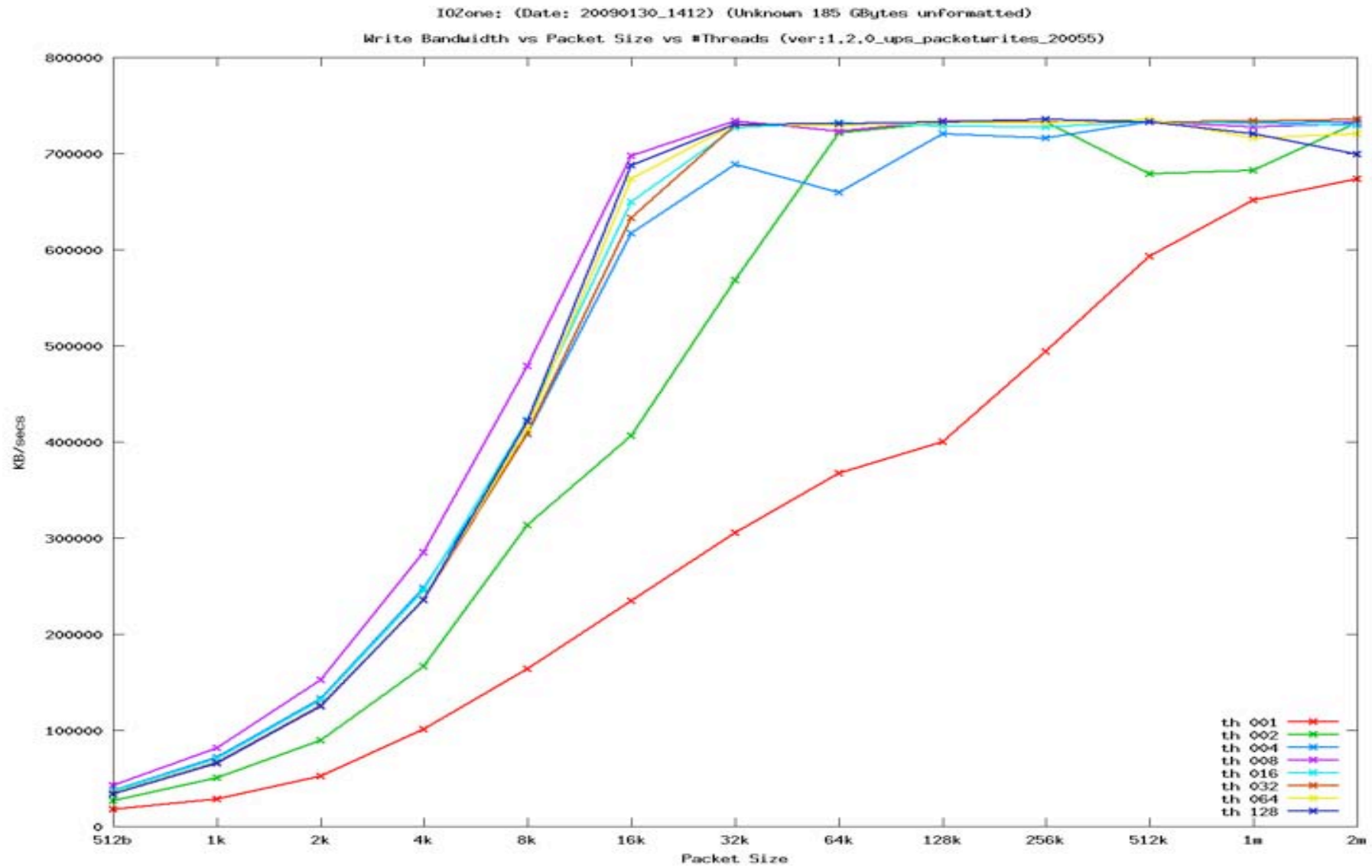
SURFACE STREETS

it doesn't take a **SATURN-V**

NAND on PCIe – Strengths

- Higher performance
  - Lower latency (25us)
  - Higher IOPS (120,000)
  - Higher bandwidth (800 MB/s)
  - No write performance drop
  - No read / write mix performance drop


IOZone: (Date: 20081204_1606) (ioDimm3 185 GBytes unformatted)
Read Bandwidth vs Packet Size vs #Threads (ver:1.2.2_TEST_16037)
800 MBytes per second
peak bandwidth
Half bandwidth at
4K packet size
KB/secs
900000
800000
700000
600000
500000
400000
300000
200000
100000
0
512b
1k
2k
4k
8k
16k
32k
64k
128k
256k
512k
1m
2m
Packet Size
th 001
th 002
th 004
th 008
th 016
th 032
th 064
th 128


IOZone: (Date: 20090130_1412) (Unknown 185 GBytes unformatted)
Write Bandwidth vs Packet Size vs #Threads (ver:1.2.0_ups_packetwrites_20055)
KB/secs
800000
700000
600000
500000
400000
300000
200000
100000
0
512b
1k
2k
4k
8k
16k
32k
64k
128k
256k
512k
1m
2m
Packet Size
th 001
th 002
th 004
th 008
th 016
th 032
th 064
th 128

NAND on PCIe – Strengths

- Higher performance
  - Lower latency (25us)
  - Higher IOPS (120,000)
  - Higher bandwidth (800 MB/s)
  - No write performance drop
  - No read / write mix performance drop
- Better RASM
  - Self-healing N+1 internal redundancy
  - Meta-data rebuild from scratch & hardware validated lookups
  - Data always protected in-flight (parity) and at-rest (11 bit BCH)
  - No potential for in-flight data loss on power cut
  - SNMP, SMIS, extensible SDK, java GUI
- Higher capacity
  - Redundancy allows for more components
  - 640 GB today, 1.3 TB 2nd half
- Lower cost per GB
  - Lower fixed costs - no HDD packaging
  - Fixed costs amortized over larger capacity

NAND on PCIe – Strengths Continued

- Longer endurance
  - More physical capacity to spread wear
  - Endurance monitoring and longevity projection
  - End-of-life data-loss protection
- Enterprise quality MLC
  - Usable for all but most write intensive workloads
  - Better parts availability
  - Lower cost structure
  - Higher peak capacity
- Efficient scale-up
  - PCIe goes direct into northbridge - no RAID controller necessary
  - No drive bays consumed
- Efficient scale-out
  - PCIe goes direct into network bridges (Ethernet, Infiniband, FC)
  - Split control-path from data-path
  - Off-the-shelf software control path (iSCSI or other)
  - Hardware accelerated data-path (iSER - iSCSI Extended for RDMA)
  - Ethernet & Infiniband networks

1U Server with (4) ioDriveDuos

- 8 ioMemory 320 MLC
- 2.56 TB Capacity
- 5.6 GBytes/s read
- 4 GBytes/s write
- 800K IOPS

Scale-up: 4U server with (16) ioDriveDuo

- 32 ioMemory 320 MLC
- 10 TB Capacity
- 22 GBytes/s read
- 16.0 GBytes/s write
- 3.2M IOPS

Scale-out: 1 Rack (36) Infiniband Attached Servers

- 72 ioDriveDuo's (2 per server)
- 72 ioSAN's (2 per server)
- 288 ioMemory 320 MLC
- 92 TB Capacity
- 144 ports of 40 Gbps QDR Infiniband
- 200 GBytes/s read
- 144 GBytes/s write
- 28M IOPS

What are enterprises using it for?

Solving Application Throughput

- Excessive RAM to avoid IO at any cost
  - Load servers / workstation with 64GB+ of DRAM to get most out of DB license
  - Expensive DRAM appliance (TMS, Violin, etc)
  - High density DRAM gets very expensive
- Excessive Spindles to aggregate performance
  - High RPM, Low capacity short stroked drives
  - Poor capacity utilization
  - Already poor HDD latency gets much worse
  - Expensive and inefficient
- Scale-out server farms
  - Add many boxes to get DRAM and DAS spindle count
  - Poor CPU utilization - cores sit idle
  - Power consumption
- Expert Man hours (talented staff)
  - Years to optimize application
  - Apps become inflexible unable to adapt to new technology

With the Fusion-io™

- Hill AFB takes NASTRAN from 3 days to 6 hours
- NYSE market maker doubles performance of trading systems
- Online retailer Wine.com shows 12x transaction rate

Wine.com Original Configuration

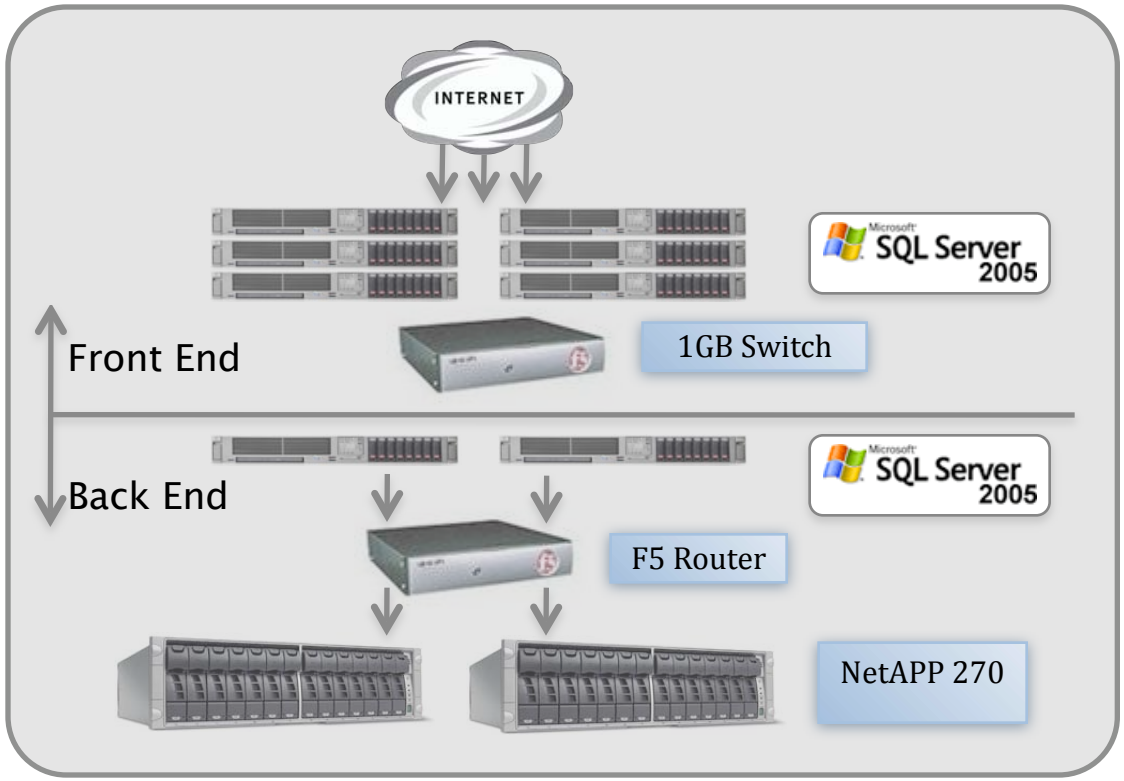
Problem

- Running at capacity
- 3 million new customers

Back-end Solution

- NetAPP 3140 (100 drives)
- = $150K +
  - — Cage Relocation (size)
  - — Larger Cage cost
  - — Larger Power cost

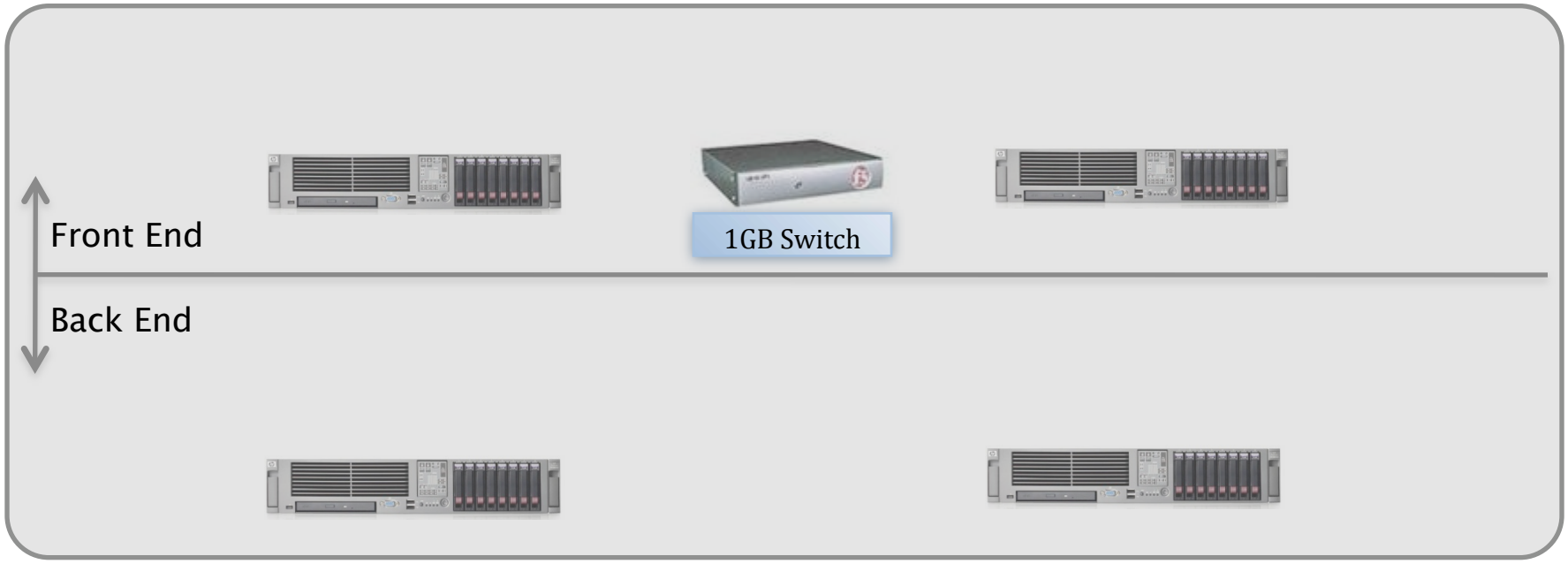
No budget left to address Front end shortcomings



Database approx 80gig

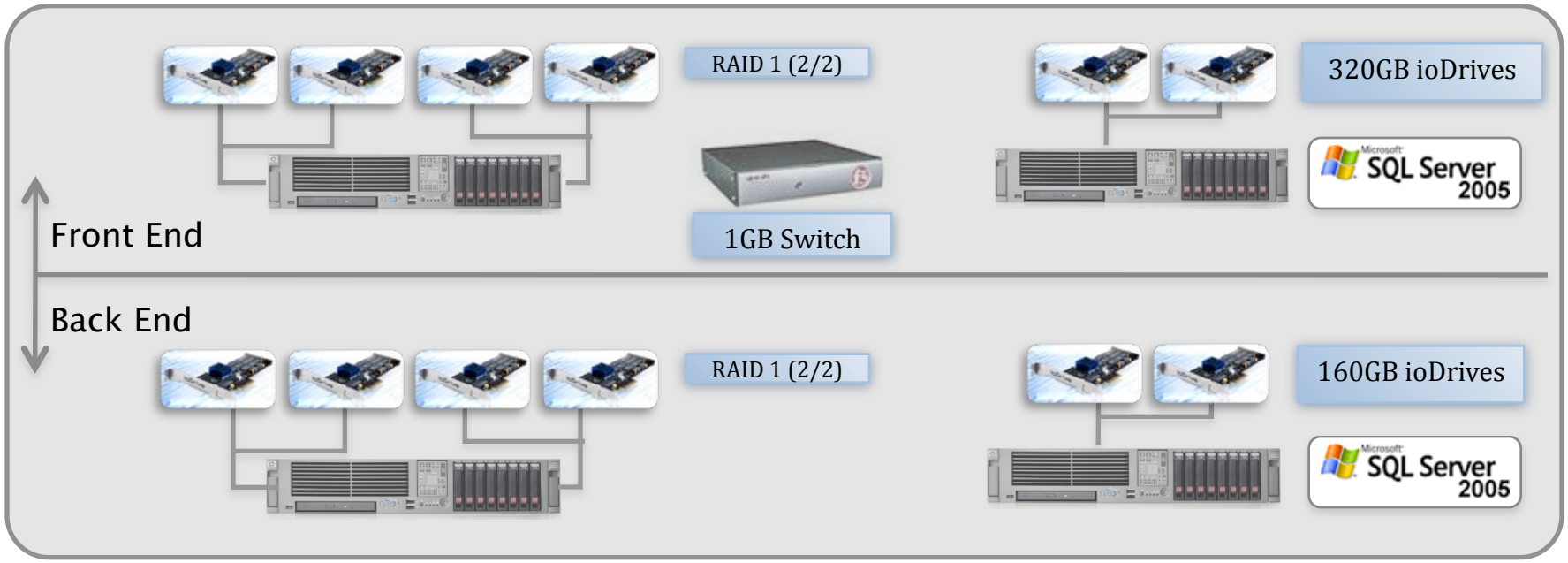
wine.com

Now



wine.com

Now — Enough capacity for 2 years



2x Customer growth capacity (future proof)

- Reduced cage cost
- Reduced power budget

Customer Challenge:

SQL Server 2005 running on NetApp appliance, poor performance in terms both latency and search queries. Average reads and writes were too slow.

Fusion-io Solution:

- 4 x 160GB ioDrives™, RAID 1 in primary server, 2 x 160GB in secondary sever
- Entire SQL database was moved from NetApp to ioDrive™

ioDrive™ Advantage:

- Dramatic performance Improvement over existing NetApp solution
- 1,200% improvement on average WRITE
- 1,400% improvement on average READ
- Average latency on WRITE: Down from 4 ms to 1 ms on ioDrive™
- Average latency on READ: Down from 12 ms to 1 ms on ioDrive™

Wine.com post holiday summary (Source: CTO – Wine.com)

Metric	Pre Fusion-io	Post Fusion-io	Improvement	Customer facing improvement
Average duration of a SQL transaction	345 milliseconds	88 milliseconds	**300%**	Website pages faster, each page has multiple DB requests. Reducing Time fetching data improves customer experience, leads to better conversion.
Time taken to take a full backup of the largest database	2 Hours	6 minutes	**1,900%**	During backups, Customer experience is hindered as customers compete for I/O with backup routine.
Time taken to restore a full backup of the largest database	3 hours	15 minutes	**1,100%**	Faster time to recovery, less loss exposure in major outage.
Time taken to post a batch of 100 invoices	2 minutes	10 seconds	**1,100%**	financial team could work through the holidays, allowing for faster analysis of the year and the health of the company (inventory, AP, and AR)
Average number of read/write operations waiting in a queue to complete	**0.4**	**0.008**	**4,900%**	Less time for customer to wait on another customers long running operation
Number of transactions in 1 hour window that took more than 500 milliseconds	**3011**	**163**	**1,700%**	Website pages faster, each page has multiple DB requests. Reducing Time fetching data improves customer experience, leads to better conversion. More cart transactions per second.

With the Fusion-io™

- Hill AFB takes NASTRAN from 3 days to 6 hours
- NYSE market maker doubles performance of trading systems
- Online retailer Wine.com shows 12x transaction rate
- Oracle shows 35x performance of unstructured search

Open World 2008: Flash Presentation

Storage Micro-Benchmarks

- Index Scan (10k actual queries, 2 million docs-40GB, text index size of 7.7GB, random read-only workload
  - 3,700% improvement on IOPS
  - 5,600% improvement on IO latencies
  - 500% improvement on IO bandwidth
  - 3,500% improvement on elapsed time on queries
- External Sort (ORDER BY query on 3.2 million rows)
  - 500% improvement with sequential IO bandwidth
  - 250% faster
- ioDrive/disk hybrid – OTLP Performance
  - 300% improvement on transmit time
  - 300% fewer Oracle foregrounds
  - 130% improvement on IOPs

With the Fusion-io™

- Hill AFB takes NASTRAN from 3 days to 6 hours
- NYSE market maker doubles performance of trading systems
- Online retailer Wine.com shows 12x transaction rate
- Oracle shows 35x performance of unstructured search
- IBM shows 1M IOPS & 5x performance improvement of Cognos on DB2
- Microsoft shows NAV has 4x performance improvement
- Shipping giant shows 30 to 1 box reduction for reliable messaging
- Medical records data warehouser shows two ioDriveDuo = 800 HDD's
- Social networking site shows 3 to 1 mysql box reduction
- Oil and gas company shows geologist workstation 5x to 20x less wait time

3D Seismic interpretation software challenge

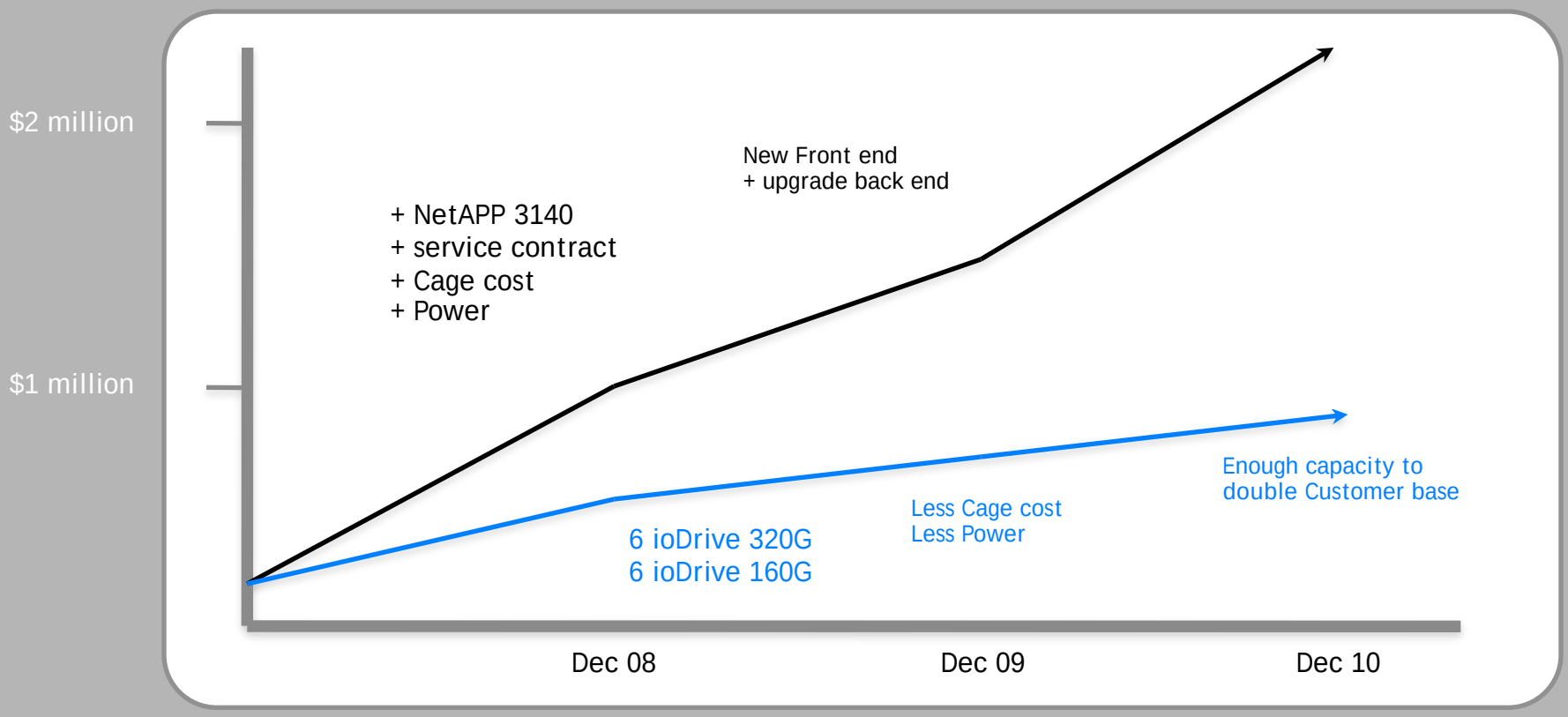
Graphics Rendering Engine

Dell Precision 690 with 80G ioDrives dual 600G SATA 300 7200RPM RAID0

- Simple 30.2GB file copy (dataset)
  - 2:02 minutes vs 7:48 (3,800%)
- Time slice on 3D dataset
  - 17 minutes vs 28 (1,600%)
- Crossline display of dataset
  - 1.3 seconds vs 12 (1,000%)
- Ran WinXP virtual inside the Win2008 w/HyperV and loaded project directly into this server
  - 10 minutes clean vs 30 minutes with server locked up

Rendering engine technology is common across Seismic, Military, CGI and Animation verticals

Cost Effective Application Throughput Scaling



Fusion-io solution addressed both front and back end capacity problems and limited incremental costs

October 2008

"Seldom have I seen technology advances that win in almost every way at the same time, in terms of speed, capacity, reliability, endurance, power usage, and simplicity."

– Steve Wozniak

CPU PERFORMANCE
continues to DOUBLE

NAND COST

continues to HALVE

BENEFIT / COST ratio
improves by
MOORE's LAW SQUARED

Thank You