NVM - Refocusing on Memory

Not Storage, Hierarchies

David Nellans

dnellans@fusionio.com

What is NAND-Flash

Volatile Memory

Non-Volatile Memory

?

Volatile-Storage

Non-Volatile Storage

What is NAND-Flash

Von Neumann legacy means we have two classes

Volatile Memory

DRAM

Non-Volatile Memory

Hot OS research topic

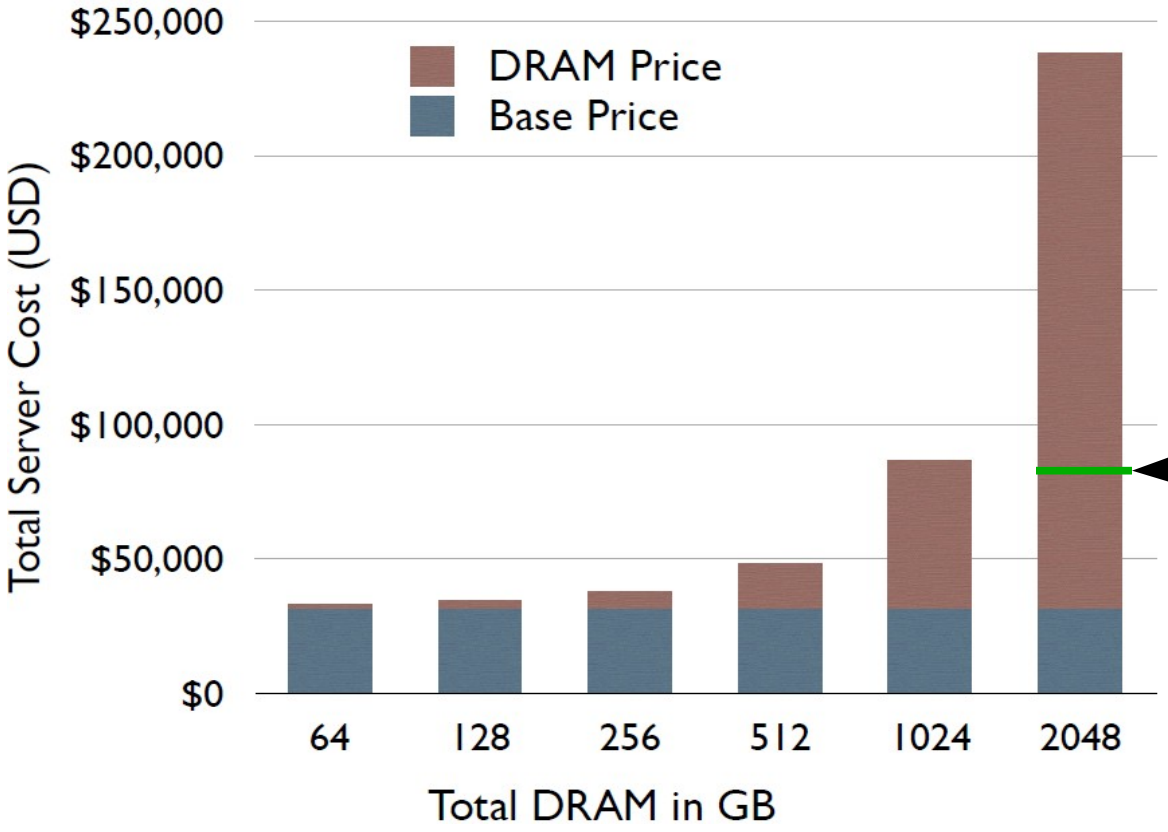
Volatile-Storage

Time frame
defines volatility

Non-Volatile Storage

Magnetic Disks
NAND based Disks
Tape Drives
Optical Media

Why Use NAND as Memory?



High density DRAM costs 10x high performance flash devices

5 TB of NAND

4TB is maximum capacity DRAM in single systems

Why Use NAND as Memory?

High Density PCIe NAND-flash

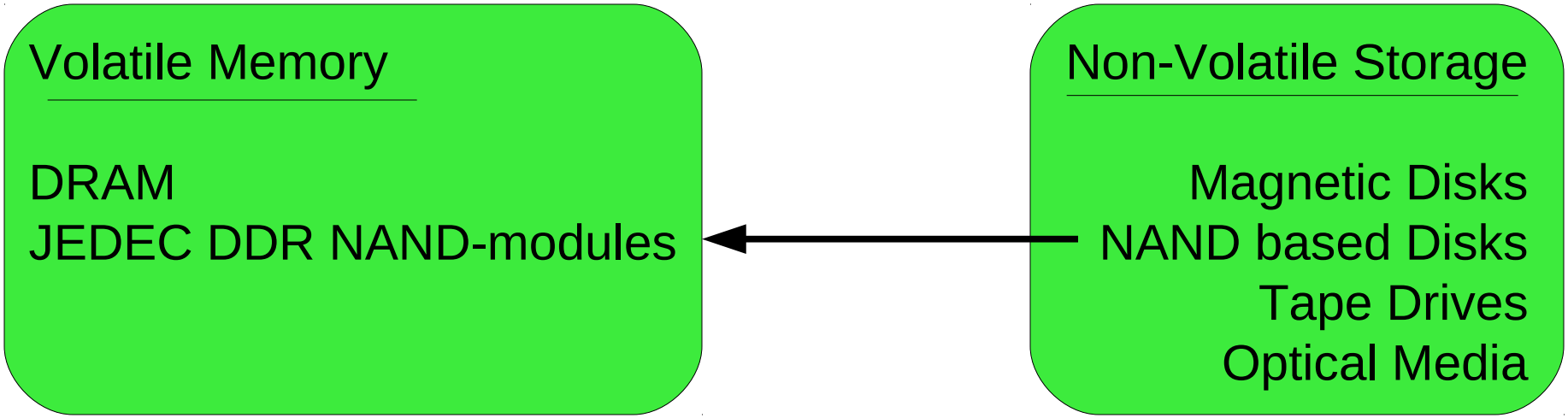
5 rack units, 45TB capacity, 1.2kW power consumption

High Density DRAM

DRAM (GB)	128	256	512	1024	2048	4096
Space (RU)	6	6	10	40	80	80
Power (kW)	1.1	1.4	2.7	6.5	7.3	14.4

NAND-Flash as Memory

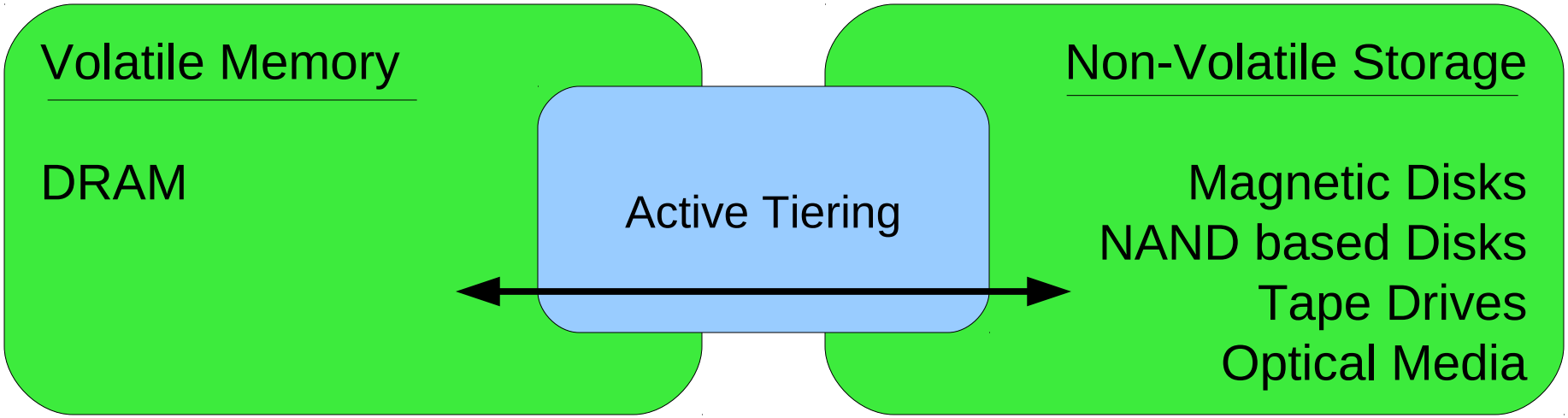
Approach 1: Move NAND-flash onto the memory bus



Advantages: Application Simplicity, Latency, Bandwidth
Disadvantages: FTL Handling, Engineering Effort, Standards

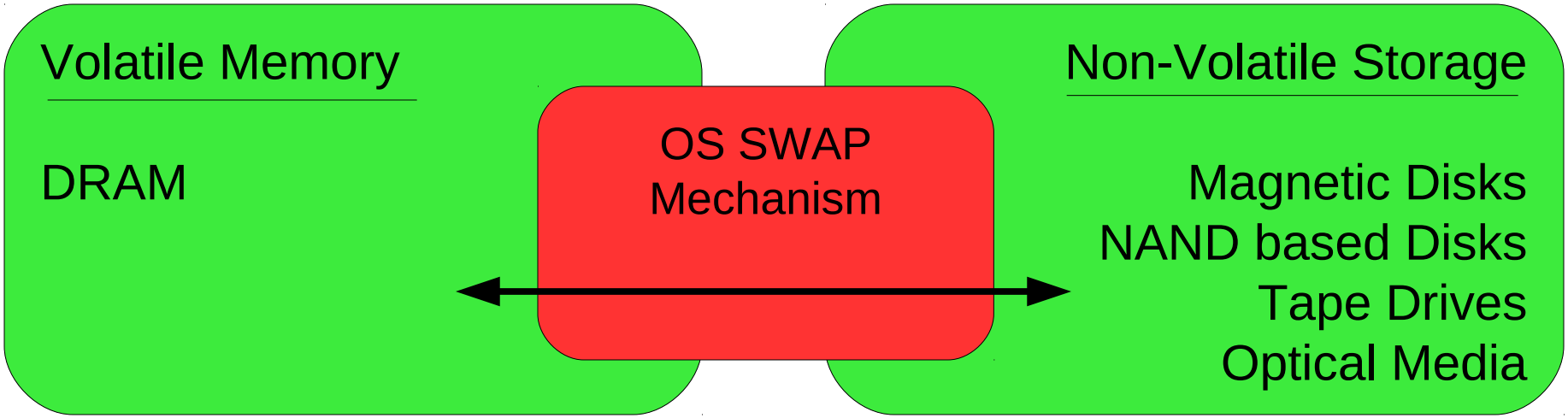
NAND-Flash as Memory

Approach 2: Allow tiering between DRAM and NAND



Advantages: Application Simplicity, No Device Engineering
Leverage Faster Storage Advancements
Disadvantages: Implementations Optimized for Magnetic Disk

DRAM to NAND Tiering



Traditional SWAP: "Last resort" - before OOM
<= 30MB/s throughput
10-100ms software overhead

DRAM and NAND Tiering

Transparent Expansion of Application Memory *

Application Transparency: No source code modification!

Unhindered Access to DRAM

Low overhead tiering: Must not inhibit flash performance

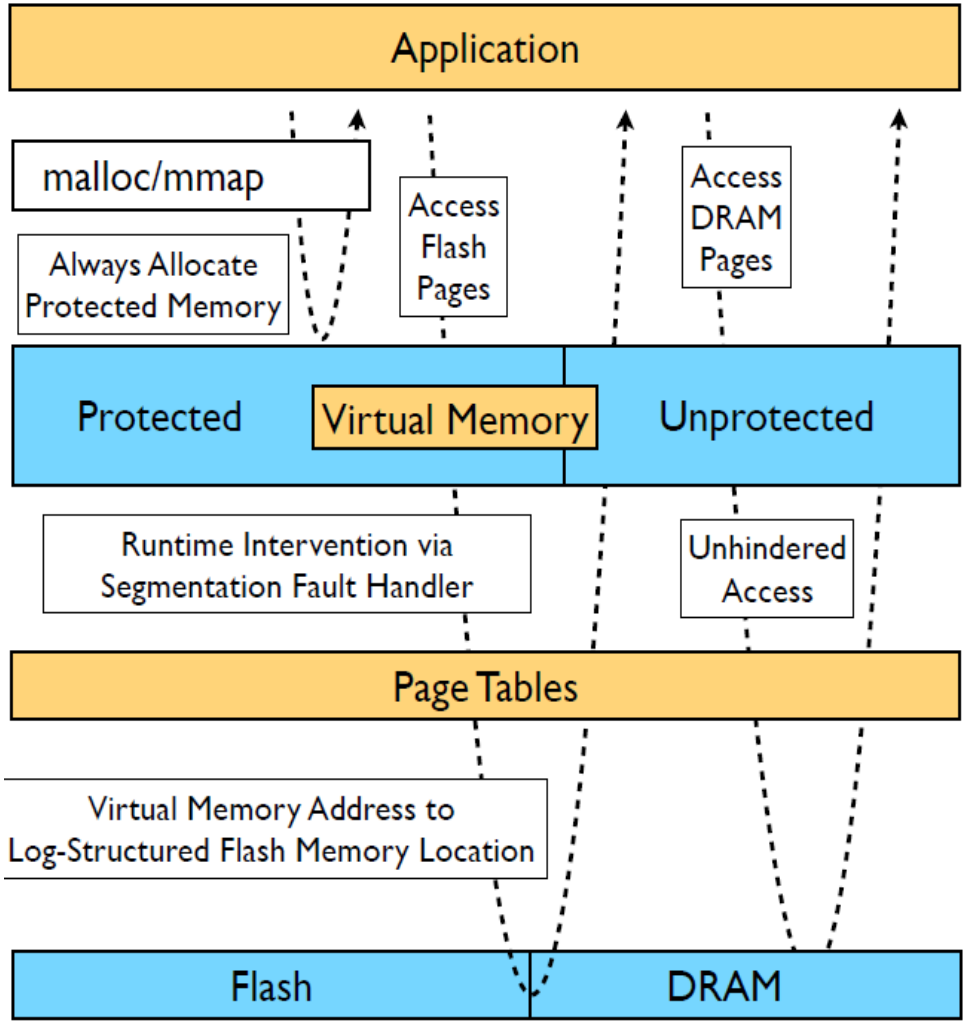
Intelligent paging decisions including application hints

* Fusion-io and Princeton University Collaboration

DRAM and NAND Tiering

Builds on historical distributed shared memory concepts.

Instead of distributed DRAM, use locally or remote attached NAND-flash.



Key Implementation Details

1. Allow thread-level paging: avoid process locking
2. Optimize for page-in operation to reduce latency
3. Intelligent utilization of flash-devices (wear-out aware)
4. Optional application hints to intelligently page data

Memory Latency

Xeon 3.43Ghz with DDR3 1333Mhz running Linux

- 10,900,000 Random 64Byte Memory IOPS
- 120,000 Random 512B NAND-flash IOPS

- Linux SWAP: 11k Random NAND-memory IOPS
- TEAM Tiering: 93k Random NAND-memory IOPS

How does latency affect application performance?

Questions to be Answered

Can we provide near-native application support?

Is transparency goal performance hindering?

Is NAND fast enough be a main memory replacement?

Is NAND fast enough to be used for tiered main memory?

Database Benchmarking

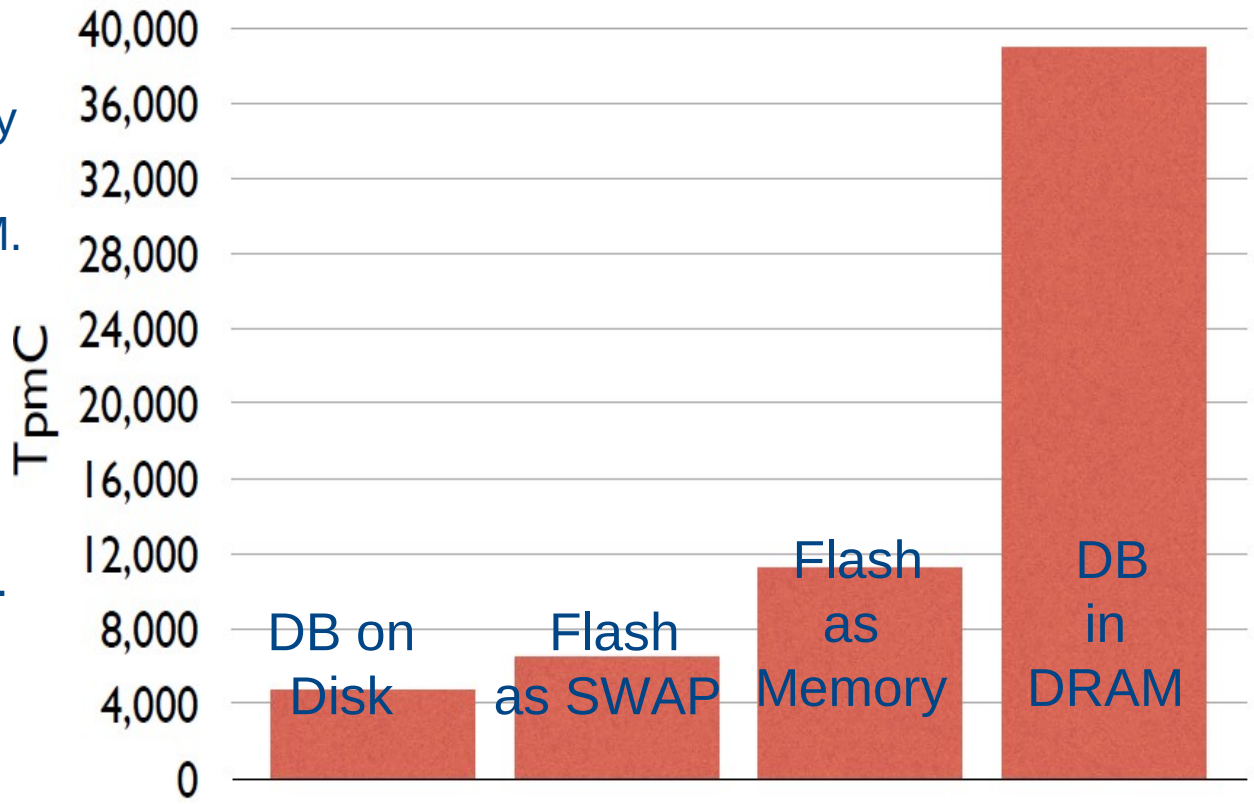
Choose Percona MySQL 5.5 running TPC-C

- Good native tiering support between DRAM and disk
- More memory is always better
- Not cost effective to put 100% of data on flash

Small Database Workloads

Flash as main memory achieves 33% the performance of DRAM.

At high densities DRAM is 10x more expensive than DRAM.

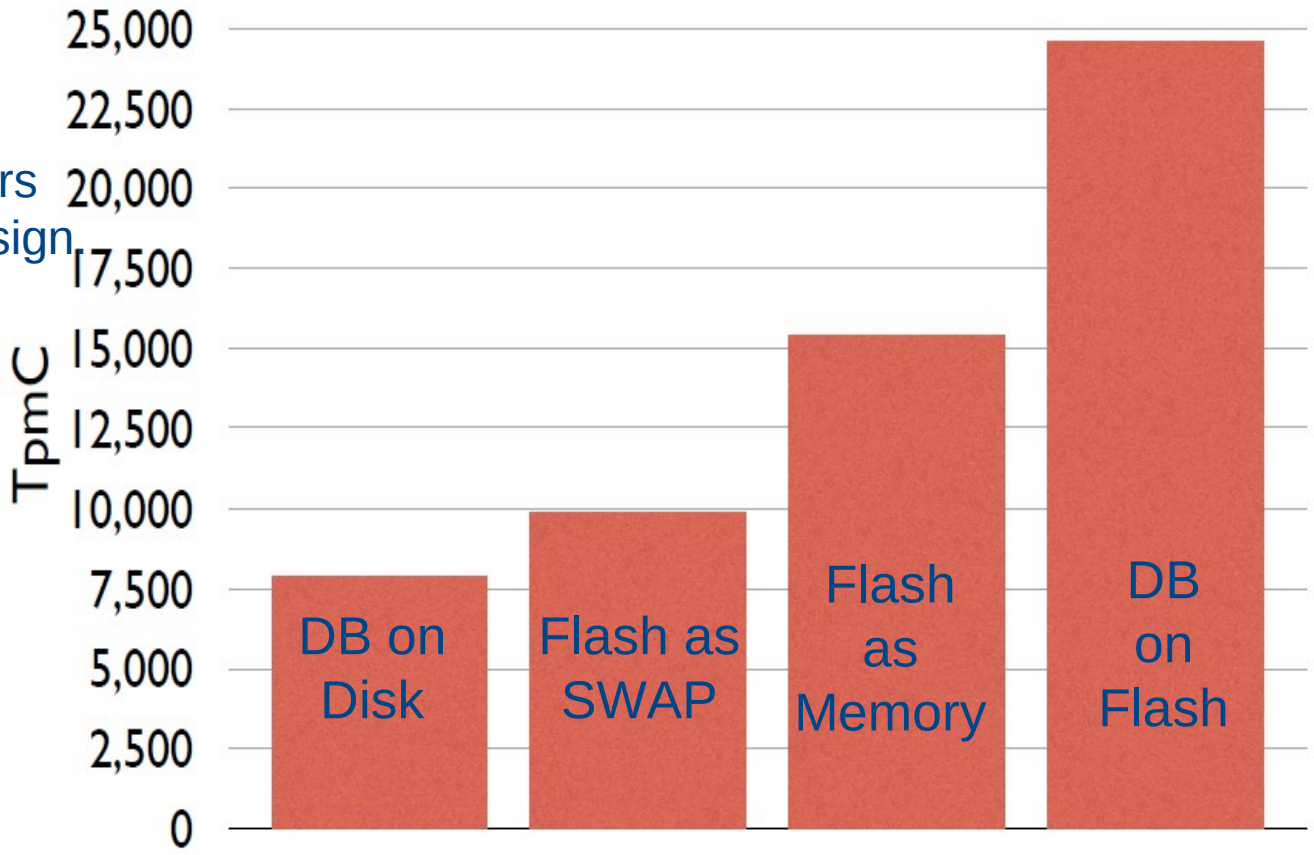


24 core Xeon, 40G DRAM, 140G Fusion-io NAND-flash: 40G DB size

Medium Database Workloads

Flash as memory is 66% as good as years of MySQL tiering design

No substitute for lots of DRAM and lots of flash.
If you can afford it.

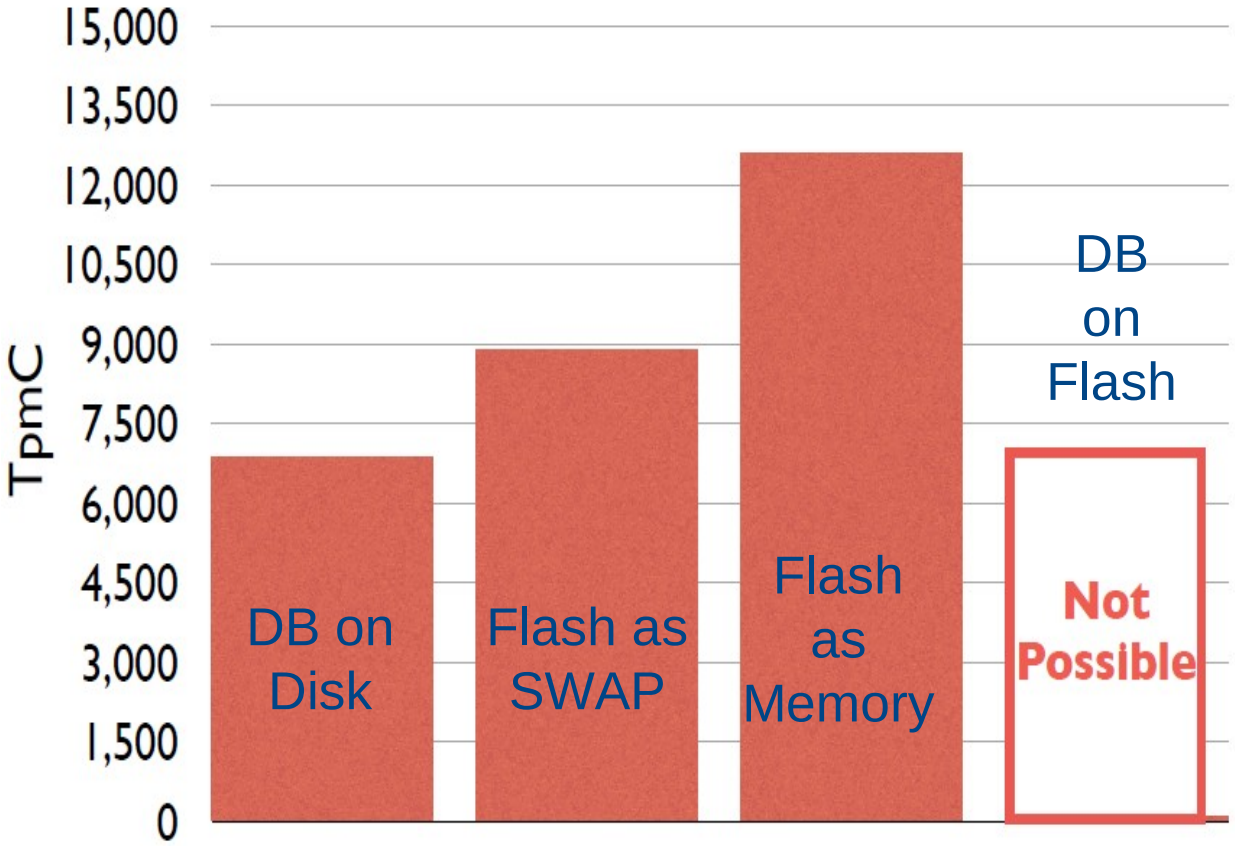


24 core Xeon, 40G DRAM, 140G Fusion-io NAND-flash: 140G DB size

Large Database Workloads

Using flash as memory can double application throughput without needing application re-write.

Increases single system scaling without sharding data sets.



24 core Xeon, 40G DRAM, 140G Fusion-io NAND-flash: 400G DB size

Main Memory Scaling Trends

Legacy applications may require 10's of TB of main memory.

Scaling nodes up has been seen as non cost-effective.

Continued sharding of data makes locality hard to maintain.

A low-power, high density replacement for DRAM is needed.

Conclusions

NAND-flash is ***not ready*** as a wholesale DRAM replacment.
Dense, power efficient, cheap. Too slow.

NAND-flash + DRAM tiering can provide:

66% the performance of an application re-write for tiering
33% the performance of all DRAM, 8% the TCO, and
5% power consumption.

NAND-flash is a ***cost effective*** way to build large memory systems.

Questions and Comments

Thank you!

David Nellans

dnellans@fusionio.com