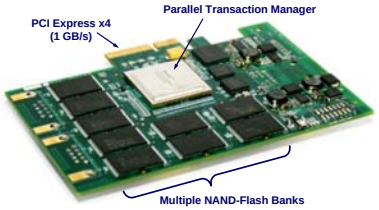
Threading Opportunities in High-Performance Flash-Memory Storage

Craig Ulmer
Sandia National Laboratories, California
cdulmer@sandia.gov

Maya Gokhale
Lawrence Livermore National Laboratory
maya@llnl.gov

Data-Intensive Applications: Storage is the Bottleneck

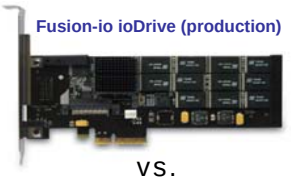
- Storage-Intensive Supercomputing (SISC) at LLNL
  - System architectures for applications with ***massive datasets***
  - New technologies: processing elements, networks, and storage
- NAND-Flash storage in high-performance computing
  - Flash chips have great potential
  - 100x better access times, 10x better bandwidth
  - However, few commercial products have delivered performance
- Exception: Fusion-io's ioDrive
  - PCIe x4 card with 80-320 GB of flash
  - Theoretical read speed of 700 MB/s
  - Hardware allows many IOPs to be in-flight concurrently



Anatomy of a Flash-Memory Storage Device
In 2007, Fusion-io delivered its first PCIe prototype for a NAND-Flash storage device. This prototype (above) employed 16 parallel flash chips, a hardware controller, and a PCIe x4 interface to the host. The 2008 production model (below) improved performance by expanding the number of parallel flash chip pads to 20 and improving the speed and capabilities of the hardware controller.

Investigating Threading Opportunities with Flash Memory

- Observation: Increasing simultaneous IOPs improves performance
  - Opposite of what we expect from hard drives
  - Due to flash memory packaging: chip is actually a die stack
- Implemented a set of I/O microbenchmarks to investigate
  - Threaded with mixed I/O characteristics
  - Vector-based computations (32 components, single-precision floating point)
  - Currently: Block transfer, kNN, external sort, binary search, k-means
- Compare threading performance between flash and hard drives
  - Dual quad-core CPUs (2.33MHz Intel E5345s)
  - 2 GB of memory
  - One 80 GB Fusion-io ioDrive
  - Three SATA drives in software-based RAID0



vs.

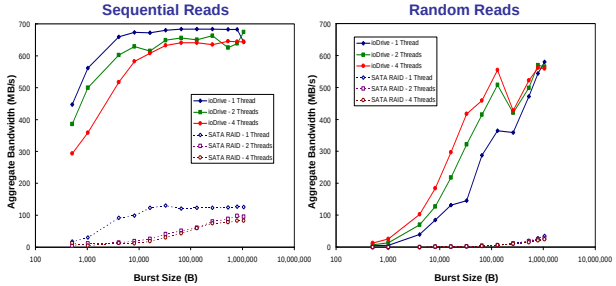
Results: Threading can Improve Flash Memory Performance

- Threading can improve performance when using flash-memory storage
  - Small number of threads allows hardware to overlap transactions
  - Performance gain more dramatic with random access patterns
  - Important, given multicore computing environment
- Fusion-io's ioDrive is substantially faster than hard drive RAID
  - More than an incremental performance gain: **3x** for Sequential, **>200x** Random
  - Note: Also possible to make a RAID of ioDrive cards
- Future directions
  - Continue developing microbenchmarks relevant to data analysis problems
  - Focus on improving performance when handling many small files
  - Compare to new flash-memory SATA drives

Block Transfer

- **Task**
  - Access different blocks of data on disk
  - Sequential and Random, Reads and Writes
- **Caveats**
  - Access patterns picked to minimize overlap and caching
- **Approach**
  - Multiple threads issue requests as fast as possible
  - Bandwidth reported as an aggregate
- **Observations**
  - Sequential reads work best with one thread for both devices
  - Threading improves ioDrive's random reads
  - ioDrive's dip at 256 KB is due to its internal block size
- **Speed comparison (Peak Performance)**

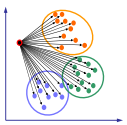
Test	SATA RAID	ioDrive	Speedup
SeqRead	125 MB/s	683 MB/s	5x
SeqWrite	139 MB/s	661 MB/s	4x
RandRead	34 MB/s	580 MB/s	17x
RandWrite	46 MB/s	658 MB/s	14x

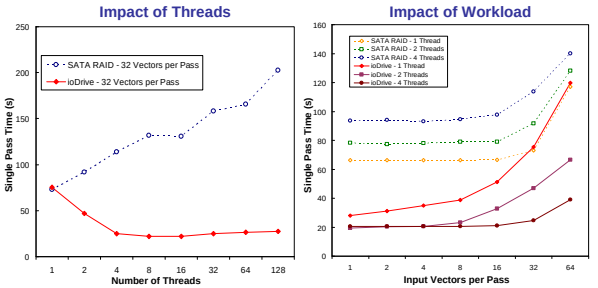


k-Nearest Neighbors

- **Task**
  - Find k training vectors in a large file that are the most similar to an input vector
- **Caveats**
  - Must read all training vectors
- **Approach**
  - Use multiple threads that operate on different portions of training vector file
  - Combine threads' results after all complete
  - Process multiple input vectors at a time
- **Observations**
  - Purely streaming I/O behavior
  - Threading helps ioDrive but hurts SATA RAID
  - Transitions from I/O-bound to Compute-bound workload
- **Speed comparison (Peak Performance)**

Test	SATA RAID	ioDrive	Speedup
16 inputs per pass	66 s	21 s	3x
32 inputs per pass	73 s	22 s	3x

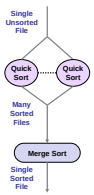


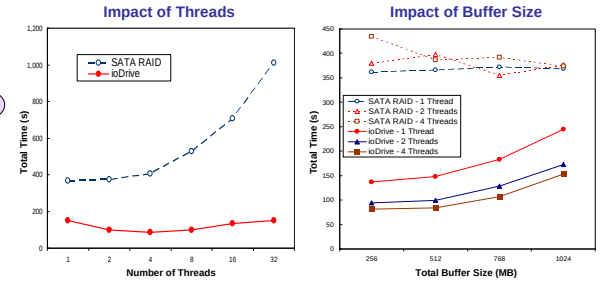


External Sort

- **Task**
  - Sort a large file of vectors
- **Caveats**
  - File is larger than main memory
  - Fixed amount of buffer space available
- **Approach**
  - Many threads quick sort different regions of input
  - Intermediate results written back to disk
  - Single thread merges all intermediate files together
- **Observations**
  - Threading degraded SATA RAID performance
  - ioDrive had best performance with
    - Small number of threads (four)
    - Small buffer size (256 MB)
- **Speed comparison (Peak Performance)**

Test	SATA RAID	ioDrive	Speedup
Total Time	361 s	81 s	4x

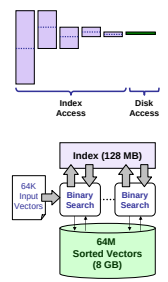


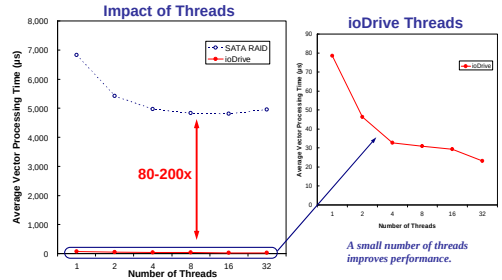


Binary Search

- **Task**
  - Locate input vectors in a large, sorted file
- **Caveats**
  - Sorted file is larger than main memory
  - Requires log(n) vector comparisons
- **Approach**
  - Index the sorted file at start time
  - Use index to minimize disk reads
  - Use threads to process multiple inputs at a time
- **Observations**
  - Purely random I/O behavior
  - Threading helps both SATA RAID and ioDrive
  - Performance depends on disk access time
- **Speed comparison (Peak Performance)**

Test	SATA RAID	ioDrive	Speedup
Create 128 MB index	387 s	38 s	10x
Process 64k inputs	315 s	1.5 s	> 200x





National Nuclear Security Administration

Sandia is a multiprogram laboratory operated by Sandia Corporation, a LockheedMartin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Sandia National Laboratories