panasas

SOS10 Future Directions Panel

Garth Gibson
CTO, Panasas, and Assoc. Prof, CMU
ggibson@panasas.com

March 9, 2006

Areas of Systems Research?

- As a CMU storage researcher, not so hard to answer
  - Protocol/API standardization for deployment of what works
    - OSDv2, NFSv4.1/pNFS, neoPOSIX, extended attributes, neo-ISAM?
    - HSM & ILM -- will slow disk/fast disk migration tools meet HPC HSM needs
  - Indepth analysis of failure; design for failure (unrecoverable read error)
  - Indepth understanding of apps; benchmarking for HEC IO
  - Automation of tuning/healing for HEC sizes; diagnosis accelerators
  - Formal testing methods; “a logic for file systems”; model checking

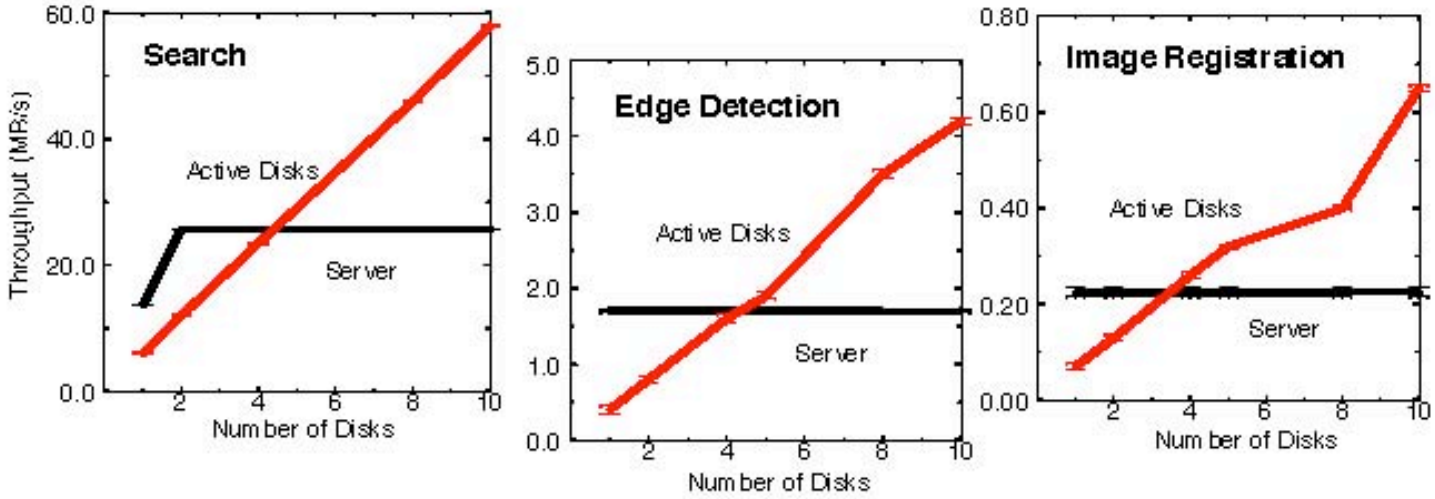
- Ie., What might be the new function in storage
  - Search, index, scan, filter, ….
  - Transactions? Microsoft is saying FS API should have begin/commit/abort

Object Storage & Active Disks

- Object storage directions
  - T10 OSDv2 – useability (error cases, COW, obj RAID, embedded indexing)
  - Richer metadata/extended attributes – uses and mechanisms
  - HSM under the covers of the OSD
  - Support for Parallel NFS (pNFS) for the broader NFS market
  - Support for XAM/reference data for the broader write-once market
- Active Disks
  - Can be powerful for scan, search and even join
  - Application crash and reboot the disk? Data integrity? Debugging?
  - In the short term, safe bet is integration with DB (e.g. Netezza)
  - To show how it can work, treat as a heterogeneous parallel programming
    - User level FS interface (XAM?), XML attribute on file that is applet overriding “read”, ship at user level to IO node, run applet at IO node using any FS & local disk

Performance with Active Disks

application	input	computation (inst/byte)	throughput (MB/s)	memory (KB)	selectivity (factor)	bandwidth (KB/s)
Select	m=1%	7	28.6	-	100	300
Search	k=10	7	28.6	72	80,500	0.1
Frequent Sets	s=0.25%	16	12.5	620	15,000	1
Edge Detection	t=75	303	0.67	1776	110	2
Image Registration	-	4740	0.04	672	150	2



Carnegie Mellon

Parallel Data Laboratory

http://www.pdl.cs.cmu.edu

Active Disks

for Data Mining

Opportunity - Active Disks

Basic advantages of an Active Disks system

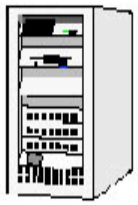
- **parallel processing** - lots of disks
- **bandwidth reduction** - filtering operations common
- **scheduling** - little bit of computation can go a long way

Appropriate applications

- execution time dominated by data-intensive core
- allows parallel implementation
- small memory footprint
- small number of cycles per byte of data processed

NASD Followon Work: Active Disks

Database Server



Digital AlphaServer 8400
- 12 x 612 MHz 21164
- 8 GB memory
- 3 64-bit PCI busses
- 29 FWD SCSI controllers

= 7,344 total MHz

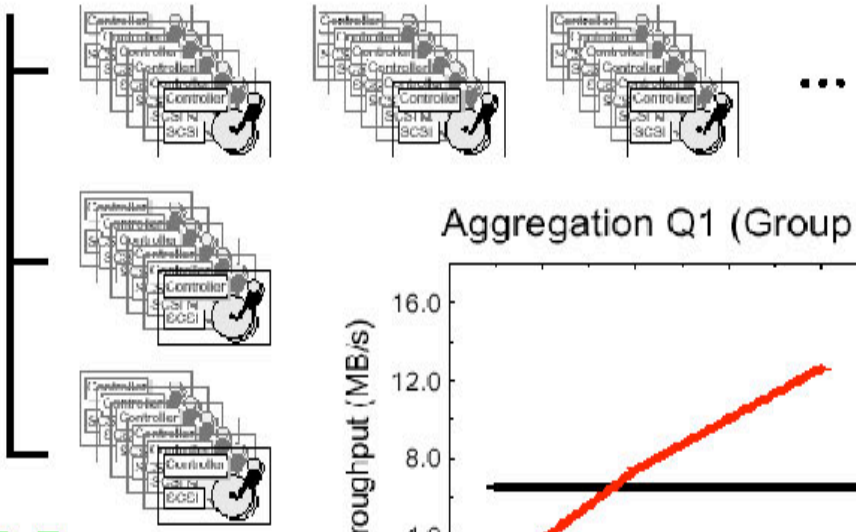
3 x 266 = 798 MB/s

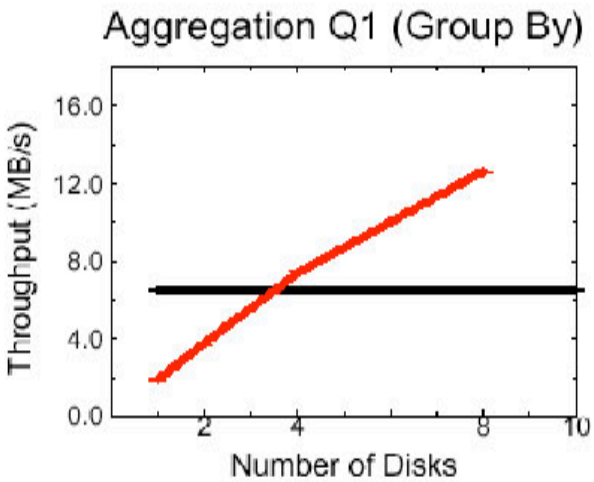
29 x 40 = 1,160 MB/s

Storage
- 520 rz29 disks
- 4.3 GB each
- 2.2 TB total

= 104,000 total MHz
(with 200 MHz drive chips)

= 5,200 total MB/s
(at 10 MB/s per disk)





PostgreSQL 6.5

- drives (133MHz, 64MB)
  server (500MHz, 256MB)

Questions 3 & 4

- 3) Programming for Storage?
  - Huh? Isn't SQL explicitly programming for storage?
  - Data representation, ie., mesh designs, not all application specific, but often very unique to a research group
    - More benefit from careful out-of-core than function shipping closer to media?

- 4) HPC Specialized Storage
  - So far, Object Storage is seen from the mainstream as HPC specialized
  - Heard this morning that SLAC is going to fund FS-sized solid-state-disk
  - Would love to see MRAM (DRAM cell with magnetic material embedded)
    - "fill the access gap"

The Ultimate Earth Mover