File Systems for the World's Fastest Computer*

Gary Grider, James Nunez, John Bent, Meghan Wingate, Alfred Torrez, HB Chen, Aaron Torres, Cody Scott

World's Fastest Computer*

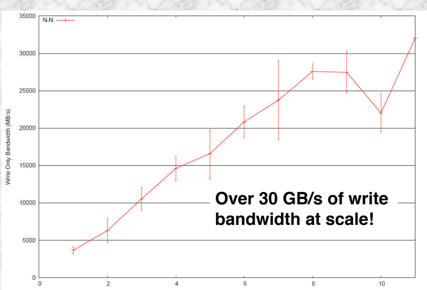
Roadrunner, the first computer to run the LINPAC benchmark at a PetaFLOP/s, runs lots of good science:
- Origins of the unseen universe
- The largest HIV evolutionary tree
- Nanowire stretching
- Laser plasma interaction
- Structural failure due to shock waves
- and lots more

Roadrunner is a *complex* system:
- Over 6000 AMD opterons
- Over 12,000 attached cell accelerators
- Over 30,000 total processing units
- Hundreds of switches, over five miles and tons of cables

Problem? Something is always about to fail. Large parallel jobs are interrupted when any one component fails.

Approach? Checkpoint-restart which requires a very fast parallel storage system.

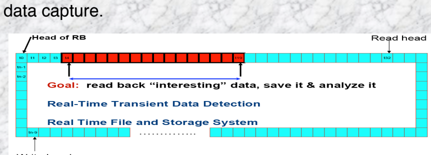
Parallel storage system on Roadrunner:
- Panasas ActiveScale File System (PanFS)
- Two petabytes of storage
- Over 100 shelves of devices
- Over 2000 hard drives



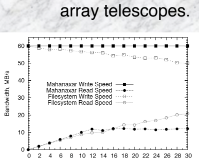
Ring Buffer File Systems

What: File systems for storing streaming transient data

Problem: Many different sources of data produce high bandwidth streams. Much of the data can be discarded but periodic events of interest need to be retained. Additionally, there needs to be quality of service guarantees so that data analysis does not interfere with data capture.



Motivating example: Telescope data from multiple instruments. This map of New Mexico shows the location of light wavelength array telescopes.



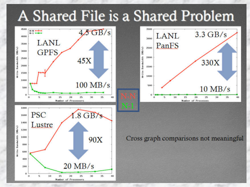
Preliminary QoS Results: This graph shows how our system, Mahanaxar, protects write bandwidth from greedy readers.

External: David Bigelow, Scott Brandt, Santa Cruz

* According to Top 500: http://www.top500.org/

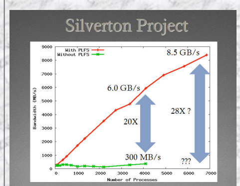
Parallel Log-Structure File System

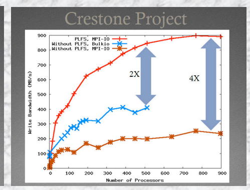
Problem: Some parallel IO patterns (i.e. N-1) preferred by users perform horribly on parallel file systems

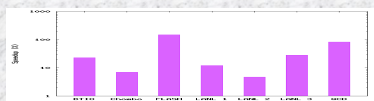


Solution: Use a virtual parallel file system, PLFS, to turn bad I/O access patterns into access patterns that parallel file systems are optimized for

Result: Orders of magnitude improvement for seven different applications and benchmarks





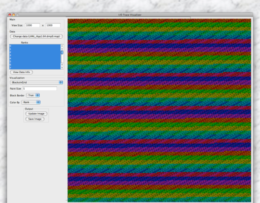


External: Milo Polte, Garth Gibson, Paul Nowoczynski

SC09 paper: http://institutes.lanl.gov/plfs/plfs.pdf

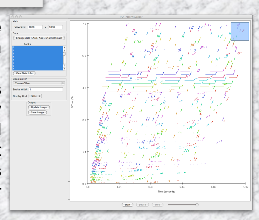
Visualization Projects

Ninjat: IO patterns in concurrent file access from an IO trace

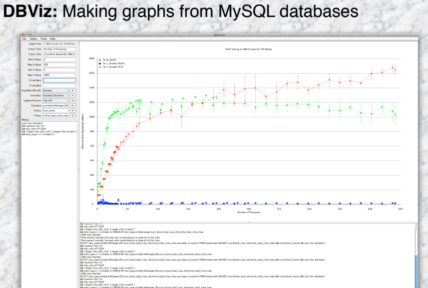


The file is drawn as a linear array of bytes wrapped in a rectangle. Each color corresponds to a different writer. This image shows that this file was written in a hybrid N-1 strided and nonstrided pattern.

Another view of the same trace showing when each offset was accessed. Notice the movie controls at the bottom which allow each image to be animated across time showing not just where each access occurs but when they occur relative to each other.



DBViz: Making graphs from MySQL databases



Queries any MySQL data and makes graphs. User specifies axes, lines, and filter.

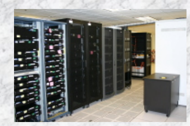
External: Calvin Loncaric, Harvey Mudd, Ryan Kroiss, UWisc

Data-Intensive Super Computing

Problem: As datasets grow, more HPC applications are migrating from computationally intensive to data intensive. Current LANL supercomputers have remote storage systems and low latency, expensive interconnect networks that are inappropriate for data-intensive computing.

Solution: Build a data-intensive supercomputer with local storage consolidated with a data-intensive file system such as Hadoop DFS.

Current Status: Small prototype cluster built and doing analysis of user workloads to determine if existing data intensive tools need to be modified for HPC applications. Data ingest is a large unsolved problem. Working with users in cosmology, cyber security, and image processing. Also investigating how to and whether to mix traditional computationally intensive HPC workloads with emerging data intensive HPC workloads. Analyzing whether any existing file systems are well suited for both workloads.

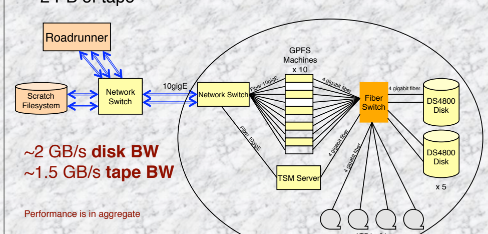


External: Christopher Mitchell and Grant Mackey, Central Florida, Katherine Nyquist, UNM-LA, Esteban Molina-Estolano, Scott Brandt, Carlos Maltzhan, Santa Cruz, Maya Gokhale, John May, Livermore.

Parallel Archival Storage System

Problem: Roadrunner needed a new archival storage system for long term storage of checkpoint and output data.

Design:
- 10 GPFS nodes
- 100 TB of fast disk
- 100 TB of slow disk
- 2 PB of tape



Approach: Efficient, smart scheduling of tape, exploit parallelism as much as possible, minimize creation of new code.
- Tape scheduling
  - Database tracking of file location (i.e. tape or disk)
  - Users run in restricted sand-box
  - No functionality lost; sand-box ensures efficiency
- Parallelism when possible
  - Parallel copy to/from GPFS
  - Parallel tape operations using TSM
  - Chunking of huge files
- Only about 25K new lines of code

Los Alamos NATIONAL LABORATORY EST. 1943