Instrumental
HIGH PERFORMANCE INNOVATION

Storage Requirements for the Future: It's the applications that matters

Henry Newman
CEO/CTO
Instrumental, Inc.
April 9, 2014

What Matters and Why

Instrumental
HIGH PERFORMANCE INNOVATION

- Weather workflow example
- Impact to storage stack
- What needs to get rethought for big data analysis

Weather Example

It impacts us every day everywhere we live except maybe San Diego ☺

Data Collected Today

Instrumental
HIGH PERFORMANCE INNOVATION

- Satellite data
  - Raw images
  - Multi-spectral
- Fixed location data
  - Buoys
  - Land observations
  - Radar around the world
- Other collections
  - Planes
  - Ships
- From around the world

Data Collected Tomorrow

Instrumental
HIGH PERFORMANCE INNOVATION

- Home weather stations
- Cell phones
- New satellites with huge improvements in resolution
- New radars with huge improvements in resolution
- Way more data

Data Assimilation

Instrumental
HIGH PERFORMANCE INNOVATION

- Take all of the inputs and put them into a database or database like structure
  - Data for each grid point (latitude/longitude by x Km Sq x altitude)
    - Pressures, temperatures, moisture, wind velocities and direction, for each altitude
    - Combine with the other observations
    - Also might combine with multiple observations and changes for 4th dimension
- All of this is small block I/O and small amounts of data
  - Pressure, temperature, and wind velocity for example are small and can be 32 bit values
- All of this is not aligned on any file system, RAID or disk boundary
  - Bigger and bigger problem for weather
- You can align or waste space but very few people know how to align
  - Therefore the I/O is inefficient given how storage hardware and software work

Data Assimilation Conundrum

Instrumental
HIGH PERFORMANCE INNOVATION

- If you cannot assimilate all of the most recent data very quickly before you run, or then you are running your weather forecast against old data
  - Well-known fact that you might have the best model, but if you start with out-of-date initial conditions you might predict something a bit off
    - Hurricane hits 60 miles in the wrong place
    - Tornado warnings off 40 miles in the wrong direction
  - This could and has cost big $$ and cost lives
- Data assimilation performance is one of the 1st, but not the only bottleneck in weather forecasting
  - But it is critical step

Applications Requirements

Instrumental
HIGH PERFORMANCE INNOVATION

- Step 1 is to read the data in for the latitude, longitude and altitude that is assigned to that thread
- Lots of methods are used to get the data into the model and it really depends on the model
  - MPI-IO
    - Small reads from disk and from many threads
  - Rank 0
    - Large reads from disk but data must be distributed to all nodes
  - POSIX reads
    - Very small reads from disk and very large number of threads
- None of methods are all that fast and there are no really good answers
  - But this is not unlike many other big data applications

Now run the model

Instrumental
HIGH PERFORMANCE INNOVATION

- Weather models write out data regularly which is forecast
- Some examples include:
  - 1 code writes 6 GB state every 10 seconds
    - Writes are asynchronous and several state writes can themselves be running in parallel
      - Historically this has been a problem with device contention and inconsistent performance
  - Another code does 2 GB every 10 seconds but runs 21 copies at a time
    - The I/O is FORTRAN writes and NetCDF I/O
      - NetCDF historically without lots of tuning has poor performance given the lack of alignment and internal I/O methods

Weather File System Requirements

Instrumental
HIGH PERFORMANCE INNOVATION

- There is a general need across all the industry for a parallel file system with:
  - Fast metadata transaction rates
  - High aggregate rates from many independent parallel threads are more important than high single thread rates
- Starting point of aggregate number is 50,000 metadata transactions/second (*open*/*create* and *rmdir*)
  - Very fast metadata sampling rates (*stat*() and etc) separate from the transactions needed
    - Aggregate number required by most is 200,000 samples/second.
- Directory based quotas help for some environments
  - They allow allocation of disk space by project on a single file system.
- Delivery of a large fraction of file system performance metrics, particular transactions and bandwidth, when disks are rebuilding
  - On very large file systems rebuilds and eventually multiple rebuilds will be nearly continuously present.
  - Either disks should not fail or declustered RAID has to be very fast
    - It is nice to want disks not to fail ☺

Future Methods

Instrumental
HIGH PERFORMANCE INNOVATION

- 4D-Var is a simple generalization of 3D-Var for observations that are distributed in time
- Data input is much more I/O intensive than current and inferior to the 3D approach
  - 1 example, separate from assimilation, is very, very metadata intensive, done with some database (often home grown), and is done on a separate small block optimized parallel file system file system
    - This design is also far from optimal but because the size of the file system and range of accessed blocks is surprisingly small, some implementations caches well
      - Often bound by things like directory search times
      - Often have directories have grown to over 100,000 tiny files
      - Solution is to throw hardware at it and move to SSDs
        - » Still will have alignment issues

Impact to storage stack

What does all this mean

POSIX Stack is not Going to Change

Instrumental
HIGH PERFORMANCE INNOVATION

- No change is planned nothing is going to happen
  - REST access which will not work well (at least in current state) is taking over market
- Small block I/O does not do well on any file system
  - Some are better than others
- Hardware alignment issues are a big problem for storage access
  - Nothing has changed here either and I am aware of no plan to fix this and in some ways SSDs are worse given the cost and performance hit

What needs to get rethought for big data analysis

We need to do a complete rethink

Understanding Weather: What has to happen

Instrumental
HIGH PERFORMANCE INNOVATION

- We have lots of data coming in
  - More every year
- Flash is not going to solve all of the problems performance issues for small block
- Data alignment does to match flash or disk hardware
- Only solution I see is people need to restructure workflow
  - That takes money

Thoughts on restructuring

Instrumental
HIGH PERFORMANCE INNOVATION

- Hardware architectural understanding of required hardware
  - Disk, SSD, data flows, IOPS rebuilding etc
- System software architectural understanding of mapping
  - Allocation sizes, access patterns, contention
- Applications software re-design
  - POSIX asynchronous I/O, larger requests, MPI-IO, and future REST for archives

Lots of working being done

Instrumental
HIGH PERFORMANCE INNOVATION

- Prediction of impact of rain, temperature and growing season on crops is worth trillions of $
  - This has not be lost on organization around the world
- On the other hand how do you turn data into a prediction that can be actionable?
  - Lots of cause effect correlations analysis that takes years and
    - Many PFLOPS, PB of storage, 100 of PB of data movement, and these numbers are likely low

Lots of Big Data Analysis Needed

Instrumental
HIGH PERFORMANCE INNOVATION

- If for example you see snow pack early in XYZ mountain range, how does that historically correlate to rainfall in the spring somewhere else, or tornadoes or whatever?
  - There is the potential for lots of cause and effect for many weather/climate events
- This can be used from everyone to farmers to commodities traders to shippers
  - As Sir Francis Bacon said, knowledge is power
  - Will power be given to Governments or commodities brokerage houses?

Thank you

Thanks for listening