Massive Data on Ice

Steve Barnet
barnet@icecube.wisc.edu
University of Wisconsin – Madison
IceCube Collaboration
May 2018
Introduction

• Who am I?
• What is IceCube?
• What is a Neutrino
• Why the South Pole?
What is IceCube?

- 300 people
- 51 Institutions
- 12 Countries
- Exotic Locales
The IceCube Neutrino Observatory

- A kilometer scale neutrino detector
- Located at geographic South Pole
- Detects Cherenkov light from neutrino interactions
IceCube Lab

IceTop
81 Stations, each with
2 IceTop Cherenkov detector tanks
2 optical sensors per tank
324 optical sensors

IceCube Array
86 strings including 8 DeepCore strings
60 optical sensors on each string
5160 optical sensors

DeepCore
8 strings-spacing optimized for lower energies
480 optical sensors

Eiffel Tower
324 m

Bedrock

December, 2010: Project completed, 86 strings
What is a Neutrino?

• An elementary particle postulated in 1930 by Wolfgang Pauli
• No charge, only affected by the weak nuclear force
• Matter is almost completely transparent to these particles.
• Existence first confirmed in 1956 – tanks of water by a nuclear reactor
Why the South Pole?
Location, Location, Location

• Lots of ice – a great detection medium
• The ice is very clear
• Thick ice sheet – sensors deep enough to provide significant background reduction
South Pole Quick Facts

- Elevation – 9300 ft base elevation
- 24x7 daylight during summer; 24x7 darkness during winter
- Average summer temp: -18F
- Average winter temp: -76F
- About 45 people remain on station over the winter
Drilling and Deployment

- Hot water drill with ~5 MW output
- Drilling takes 26-30 hours per hole
- Strings deployed in ~ 20 hours
  - DOMs get final test
  - Attached to surface cable
  - Lowered into hole
- About 1 month to completely re-freeze
Construction is complete

- First string deployed in 2004-2005 season.
- Detector completed in 2010-2011 season.
- All science from here on out.
Data - South

- Detector generates ~1TB/day of raw data
- Reduced to about 100Gb/day for transmission via satellite (filtered data)
- 2x copies of raw and filtered data are saved to local media
Compute Systems - South

• 100x custom machines for data readout
  – Commodity computer (SBC)
  – Custom serial card to interface with DOMs

• Onsite compute farm for data reduction

• Data transfer system (JADE) - archive and satellite data movement
Tale of the Tapes

• For many years all data was written to tape
• A source of constant headaches
• Note: all machines are operated out of environmental spec
  – At the edge of elevation
  – 0% humidity
• Most things not a problem tapes, well:
Hi Patrick,

we did some further investigations on the issues with the tape. To check if maybe the low humidity has an influence on the tapes we put a couple of LTO-4 tapes in the greenhouse. After one day in the greenhouse we could write a handful of files to a tape which did not work at all before.
These Days, We Just Use Disks

- Disks are treated like tapes
- 2x copies written (each to a separate disk)
- Shipped back (literally) when the station opens
Data Systems - North

- **Experimental Data**
  - Detector related data (operational state of the detector)
  - Filtered data is not science ready – needs more processing
  - Archive and replication
    - Filtered stream replicated to a collaborating institution (DESY) daily
    - Larger raw data sets replicated to NERSC
Simulation

• Larger than experimental data
  – When a TB/day just isn’t enough
• Analyses are statistical
• Need to understand the properties of the ice – GPUs
• May explore different detector properties
• May want specific physics parameters
Current Online Totals

- Experimental data – 1.9 PB
- Simulation – 4.0 PB
- Users and analysis – 2.0 PB

- 10 Tons of storage

- $1 \text{PB} = 10^{15} \text{Bytes}$
Holding it all – the Evolution
When in Doubt, Start with NFS
Rapidly Runs out of Gas

- Sym links are not a unified namespace
- The free space is never available where you need it
- SAN? – zOMG the $$/TB is too high!
- Enter the Cluster filesystems
Design

• Need to feed a compute cluster
  – several thousand jobs active simultaneously

• Commodity storage – low $$/TB

• Scale out – as storage capacity is added, performance is added

• Separate file meta-data from block storage
First Iteration - IBRIX

- Metadata is distributed across all servers
- Block storage too
- Scaled pretty well – 10-ish Gb/s sustainable
- Until some flaky storage entered the pool
When Storage Arrays Die
Enter Lustre

• File Meta-data lives on one server
• Block storage distributed across multiple servers.
• Clients hit meta-data server first to get the block map
• After that, reads/writes are handled by storage servers.
• Scales *very* well – for some workloads
Lustre – challenges

• Complicated
• It’s in kernel space
  – You can’t let clients and servers get too far apart
  – Patching – yeah, about that …
• Likes big files – small files, not so much
• Not great when storage goes away
Future

• Still lots of Lustre – just deployed 9.6 PB – Thanks DoIT data center team!
• Looking at Object Store (Ceph)
• Cloud – maybe; still pricey at multi-PB scale
Wish List Items

• More timely accounting
  – du takes you only so far
• Better performance analytics
• Fails more elegantly (less impact for single component failures)
• Data management – better automation of data replication
Thanks

• NSF for paying for all of this
• Many great coworkers over the years
• Great teammates of today!
• Most of all:
Questions?
BACKUP SLIDES
Online Processing
Science is Heavy
Science is Heavy