NP04 Online Buffer System Requirements and Options

Brett Viren

Physics Department

BROOKHAVEN
NATIONAL LABORATORY

DUNE Collab
Sep 2016

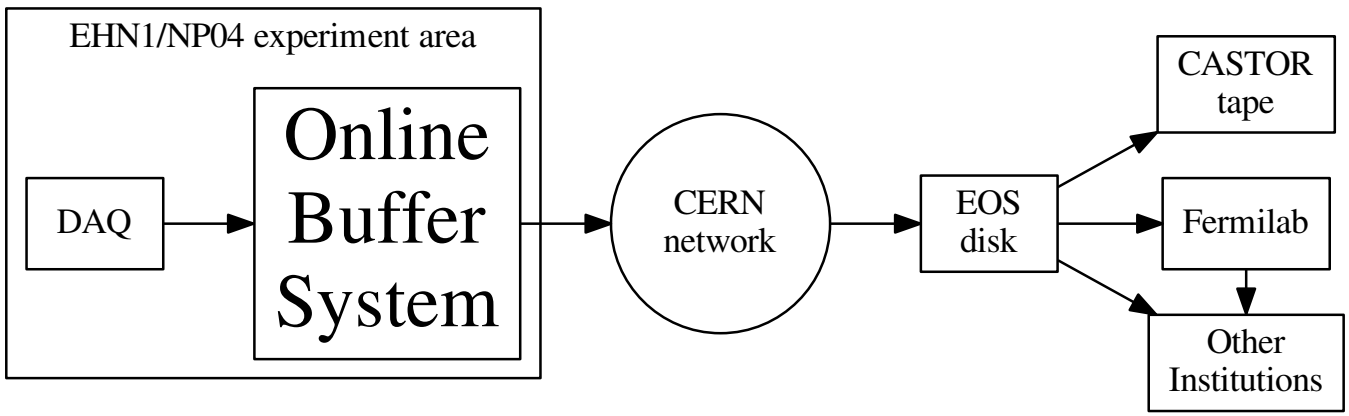
Outline

Requirements

Designs

Discussion

NP04 Online Buffer System Context



The concept: place sufficient storage "close" to DAQ to accept all data files in order to isolate from potential CERN network and/or EOS outage.

High-level Performance Requirements

throughput accept in-spill "high rate" data scenario throughput
→ **3 GByte/s**[1]

capacity cover expected outage duration
→ **1-3 days**[2]

latency small compared to prompt-processing
→ $\lesssim$**1 minute**[3]

[1] Detailed scenarios and estimations in DocDB 1086.
[2] Rumor, need to nail down!
[3] A guess, want 10 minute turn-around for any prompt-processing.

Component Hardware Options

Some simple assumptions:

- **Storage** SATA:50 MB/s, SAS:200 MB/s, SSD:500MB/s, all half-duplex.
- **NICs** 1Gbps or 10Gbps, full-duplex, theoretical max.
- **Switch** match NICs but $\sim \infty$ fabric bandwidth.
- **CPU** various benchmarks (eg, artDAQ, `md5sum`).

Aside: DAQ Minimum Multiplicity

Minimum[4] to keep up with **steady-state** 3 GB/s:

- 4 EB computers with 10 Gbps NICs.
- 2 EB artDAQ processes/comp each pushing 375 MB/s
  - $\rightarrow$ artDAQ benchmark demonstrates 440 MB/s !!
- Need 2 $\times$ 1.5 core for artDAQ + 2 for **external** checksum
  - Do we need this? We expect internal checksums exist as well (at DAQ-level and/or ROOT-level).

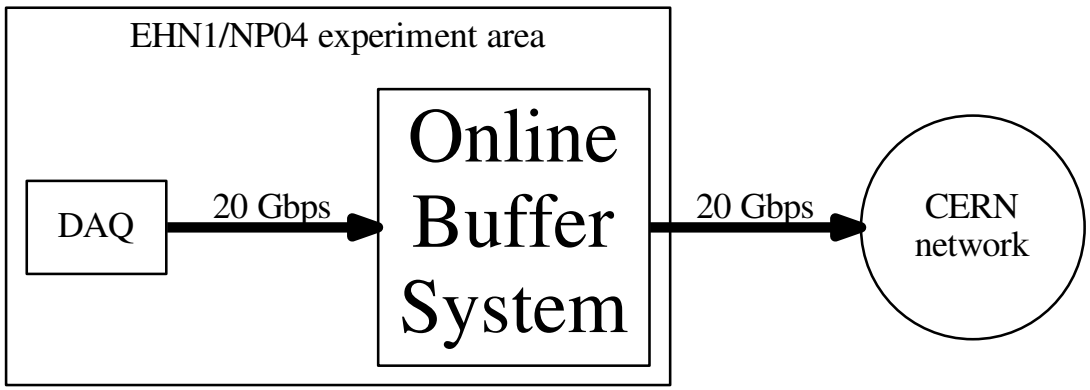
$\Rightarrow$ 5 GB file every $\sim$ 13 seconds from each EB proc.

Two competing consequences of assumptions:

+ Absorb real-world fluctuations require more ($\sim 2\times$?).
- See full 3 GB/s only during the spill ($\sim$20% of the time).

[4] Details in DocDB 1656.

Network Edges of Online Buffer System



Assume 20 Gbps network bandwidth **from DAQ** and **to EOS**.

→ 3 GB/s = 24Gbps, so we are **rounding down** here!

→ As-is, **the network is the bottleneck**.

Note: in "concentrated" design, DAQ→Buffer connection is more intimate.

Transport and Bookkeeping

High level tech decisions: **XRootD** and **F-FTS/SAM**:

1. **DAQ** transfer data into buffer via **XRootD** or **local FS**
   - "**distributed**" buffer: XRootD ingest
   - "**concentrated**" buffer: local FS writes
   - Nominally DAQ also produces & transfers "metadata" file.
2. **F-FTS** learns of newly buffered files:
   - via **push-notify** or **delay-poll**
3. **F-FTS** initiates an **XRootD** transfer to EOS.
   - speak EOS's native tongue: XRootD
4. **F-FTS** registers file location and other metadata with **SAM**'s file catalog.
5. **F-FTS** initiates file purge from buffer.
   - purge cooperates with any other users of buffered files

See Stu's presentation Tomorrow morning in Software/Computing session for more about F-FTS and Onward.

Hardware Spectrum

	Distributed	**Concentrated**
NIC	1 Gbps	10 Gbps
HDD	120 SATA	90-360 SSD / 45 SAS
#nodes	60	~2 FE + 4 storage
tech	commodity	leading edge
h/w	generic/existing	specialized
cost	US$ 25K (+repurpose)	US$ 100K+

"**Distributed**" - flat scaling

- → **pros**: simpler, cheaper/existing systems, fine-grained redundancy, more available idle CPU.
- → **cons**: larger rack footprint, power, cooling.

"**Concentrated**" - hierarchical scaling

- → **pros**: fewer boxes to deal with, higher performance per unit.
- → **cons**: cost, larger unit of redundancy.

Distributed Option: the `neut-buffer` Cluster

The idea:

- Buy 120 SATA disks to upgrade 60 nodes of existing "`neut`" cluster (∼$25k).
- Buy a switch (60 × 1Gbps + 4 × 10Gbps) (∼$5k), or 2 smaller switches.
- Test in-situ, extract nodes and move to NP04 hall for production running.
- 60 × 7.5 cores potentially available (eg, for checksumming).
- Loss of one node represents ∼ 2% throughput degradation.
- Option for DAQ to stream data into buffer or to copy from temp DAQ FS.

Status:

- Nectarios Benekos has large fraction of `neut` online.
  - → Ready for purchasing HDD upgrades.
- Maxim Potekhin has installed and configured XRootD on `neut`.
  - → Ready to start some initial throughput testing.

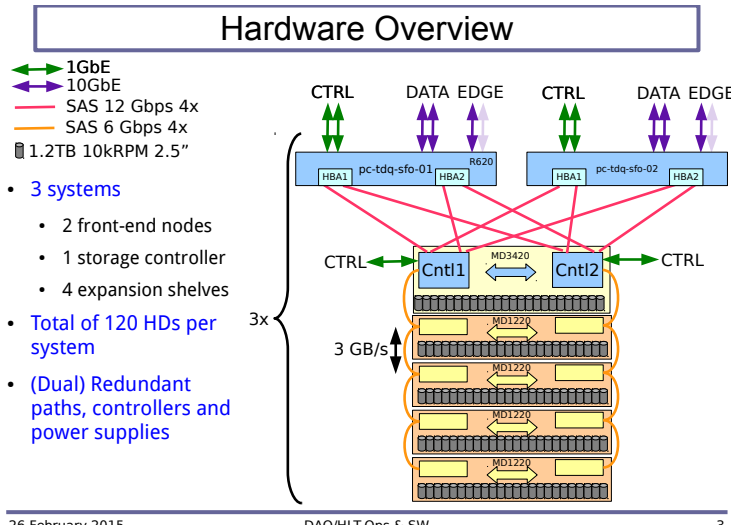
Independent of buffer decision: do we have a need to "fill up" `neut` with some disk?

Open questions for the `neut-buffer` Cluster

- Can we get agreement to use ∼20% of the `neut`[5] cluster hardware for protoDUNE/SP buffer?
- Will we have rack space, power, cooling at NP04 hall for 60 "U" of rack space?

[5] neut twiki for more info

Concentrated option



The idea:[6]

- Buy new, high-end storage hardware.
- Each front end mounts its storage units as “local” FS.
- Used successfully in ATLAS, existing experience.
- Option: use same hardware to host EB (and maybe also BR) artDAQ processes as well as an F-FTS instance.
  - → No distinct buffer “system” per se.

Status:

- Candidate hardware identified.
- Initial understanding of multiplicity and costs.
- Discussions underway to understand the idea more.

[6] Giovanna Lehmann Miotto

Open questions for Concentrated Option

- What testing needs to be done before acceptance of a design?
- How many spares and of which parts are needed to reduce risk of hardware failure or to allow quick fixing?
- Is the DELL proprietary storage management software acceptable?
- Is it acceptable to mix the different DAQ+Buffer layers (BR+EB) on the same hosts?

Important protoDUNE-SP Milestones for Buffer

May 2017 APA 1 integration tests

→ need **some** buffer storage in place prior to this.

Aug 2017 APA 2 integration tests

May 2018 Filling and commissioning

→ full buffer system in place, ready to use before commissioning starts!

Open Questions for Discussion

- How many days must we **really** buffer? 3? 1?
- What end of the hardware spectrum do we take?
  - concentrated vs. distributed
- Who will take charge of hardware procurement and installation?
- Who will provide sysadmin duties?
  - See also Geoff Savage's questions from morning session.
- All estimation based on steady-state throughput, need to think more about fluctuations and back-pressure?
- ...what else?

Let's try to decide **concentrated** vs. **distributed** quickly so we can move on to testing and start thinking about operation.