LDRD Streaming Readout Testing Plan

Marco Battaglieri, Markus Diefenthaler, Vardan Gyujyan, and **David Lawrence**[*]

May 15, 2024

Abstract

Testing plan for the Jefferson Lab LDRD project to develop a Real-Time Development Platform (RTDP) for Streaming Readout Data Acquisition (SRO).

1 Overview

Development of *The SRO Real-Time Testing and Development Platform* (RTDP) will include multiple testing campaigns requiring collections of resources from the experimental program which include front-end electronics, high speed networks, storage systems, and compute systems. The testing campaigns will be performed in phases and done to minimize interference with existing operations. One major campaign is planned as specified in the proposal as the fifth major objective of the project. The text describing it is copied here for convenience:

> **Objective 5:** High Bandwidth Test: Configuring a full scale system that includes both real and proxy components and testing it at high bandwidth is necessary to demonstrate the platform's core functionality. Current expectations are to have a 400Gbps link available between the Hall A,B,C counting house and the Computer Center in CEBAF Center sometime in FY2024. The high speed testing will be coordinated to occur when the beam is down so that the full bandwidth will be available for the testing periods. The SoLID experiment serves as an example of the type of high bandwidth experiments being anticipated to run at JLab in the future. It will then serve as a useful guide for the testing configuration, even if the configuration is not an exact match for SoLID. There are currently eight U280 FPGA cards in the Computer Center purchased for use with the EJFAT project which would be available to use for these tests. Similarly, the Scientific Computing farm will have a few dozen GPUs (mostly Tesla T4's) available that could also be utilized for these tests. Utilizing real hardware components will be an important part of the platform and so will need to be included for the full scale configuration testing. We will utilize existing components developed outside of this project to exercise the heterogeneous components. For example, PHASM, CLAS12 tracking, and the EIC R&D project: ML4FPGA [1].

There is also some additional text in section 3 of the proposal alluding to testing with direct beam that says:

> ... Moreover, considering the current physics program and the number of scheduled experiments, JLab represents the ideal test bed for the final system in real conditions. It is noted that CLAS12 already has a full set of VTP hardware and upcoming network upgrades will provide for the full infrastructure to stream data. They are also considering a GPU-based tracking system as part of a level-3 trigger design. This could be incorporated as a component of the proposed platform.

[*] davidl@jlab.org

Several major testing events are planned throughout the two-year project. These are listed in table 1 where the test corresponding to Objective 5 is highlighted in yellow. Details on the following tests can be found in the following sections.

Date	Event	beam	Hall A	Hall B	Hall D	EJFAT
Dec 2023	CLAS12 Partial Data Capture: Hall-B	✓		✓		
Mar 2024	CLAS12 Partial Data playback: Hall-B to NERSC			✓		✓
May 2024	CLAS12 Partial Data Stream: JLab farm	✓				✓
Jul 2024	GlueX Triggered Data Stream: Hall-D to NERSC				✓	✓
Oct 2024	BDX Single Crate Stream: JLab farm					
Jan 2025	SBS Stream: Hall-A to NERSC		✓			✓
Apr 2025	ePIC Simulated Stream from BNL to JLab					✓
Jul 2025	**SoLID-inspired Configuration Stream: JLab***		✓			✓
Sep 2025	Stream from NERSC (*TBD*): NERSC to JLab					✓

Table 1: Major Testing Events. The Jul 2025 item is specifically called out as corresponding to Objective 5 in the LDRD proposal.

2 CLAS12 Partial Data Capture: Hall-B

NOTE: This has been updated from the v0.7 version that was used for the Dec. 2023 packet capture exercise. Please see that version for details specific to that exercise.

Details on the Run Group-E plans can be found here:

https://wiki.jlab.org/clas12-run/index.php/Run_Group_E#tab=Short_Term_Schedule

Quick Notes:

- FADC250 streaming mode currently only supports mode 7 = Pulse integral
- Previous CLAS12 TriDAS tests had .pt file modification times/sizes indicating 0.37MB/s for the 3x3 EIC calorimater in 2023. A few up to 13 MB/s for the CLAS12 FT in 2022. The highest rates (13MB/s) were for the FT run 157.
- Estimate of the maximum disk space that will be needed is given by:
(30MB/s)(3hrs)(3600s/hr) = 324k MB = 324GB

The nvme disk on ejfat-fs can be written to at about 1.6GB/s:

```
$ cd /nvme/proj/RTDP/2023.12.17.CLAS12
$ dd if=/dev/zero of=disk_speed_test.dat bs=1G count=100 oflag=dsync
100+0 records in
100+0 records out
107374182400 bytes (107 GB, 100 GiB) copied, 66.1002 s, 1.6 GB/s
```

A RAM disk on ejfat-fs can be written to at about 2.5GB/s:

```
$ sudo mkdir /mnt/ramdisk
$ sudo mount -t tmpfs -o size=20G tmpfs /mnt/ramdisk
$ cd /mnt/ramdisk/
$ dd if=/dev/zero of=disk_speed_test.dat bs=1G count=15 oflag=dsync 15+0 records in 15+0
records out 16106127360 bytes (16 GB, 15 GiB) copied, 6.52641 s, 2.5 GB/s
```

The home directory on ejfat-fs can be written to at about 0.39GB/s:

```
$ dd if=/dev/zero of=disk_speed_test.dat bs=1G count=10 oflag=dsync
10+0 records in
10+0 records out
10737418240 bytes (11 GB, 10 GiB) copied, 27.6037 s, 389 MB/s
```

tcpdump can handle about 9.4Gbps from 18 streams without packet loss:

On ejfat-fs run this:

```
$ cd /nvme/proj/RTDP/2023.12.17.CLAS12
$ sudo tcpdump -i enp193s0f1np1 -j adapter_unsynced –time-stamp-precision=nano -s 0 -w
junk.pcap portrange 5001-5018 -B 10480000
```

On ejfat-2 run this:

```
$ iperf -c ejfat-fs-daq -t 1000 -i 3 -P 18 –bandwidth 500M
```

n.b. This works by having tcpdump allocate a 10GB buffer and limiting the bandwidth of each of the streams to 500Mbps(x18=9.44Gbps total).

Checklist:

□ *tcpdump* installed and able to capture packets from multiple streams from 18 crates (ensure metadata is included so tcpreplay can reproduce time structure)

□ Streaming configuration for CLAS12 detector (18 crates)

□ FADC250 thresholds and window sizes configured

□ Format verification tool to ensure data is valid and readable

□ Disk location in Hall-B and transfer mechanism to tape confirmed. Location, file naming scheme, and format recorded on RTDP wiki.

2.1 Running the CLAS12 DAQ System

The DAQ will be the standard Hall-B/CLAS12 system, but with additional programs run on the ejfat-fs-daq computer to receive the streams and record them into files.

IMPORTANT: When starting a new run ””'ALWAYS””' take CODA all the back to the ”Configure” stage. If it is not reconfigured for each run then the firmware is known to report incorrect channel numbers!

To start the capture processes do this (note the complete start_capture.sh script can be seen below):

1. log into ejfat-fs
2. cd /nvme/proj/RTDP/2024.05.17.CLAS12
3. **./start_capture.sh**

The above will create a file in the **"/nvme/proj/RTDP/2024.05.17.CLAS12/files"** directory with a name that includes the current time. This ensures we don't accidentally overwrite a file. An example file name is:

```
CLAS12_ECAL_PCAL_DC_2024-05-17_12-00-00.pcap
```

The DAQ system should use the following:

- CODA config: zzz1
- Trigger config: STREAMING/fc_streaming.trg

Table 2 lists the run plan for the May 2024 data capture exercise.

Run Plan for CLAS12 SRO beam-on partial data Capture May 2024	
step	**Task**
1	Ensure adequate disk space for storing stream file on ejfat-fs (=350GB)
2	Ensure $LD_2 + Al$ target is full, Drift Chambers are on, and Forward Carriage detectors are on
3	Start data capture processes on ejfat-fs-daq (cd /nvme/proj/RTDP/2024.05.17.CLAS12; ./start_capture.sh)
4	Set CODA configuration (see text). Transition through Download and Prestart phases
5	Request **5nA** beam
6	Wait for beam to stabilize
7	Start run
8	Capture data for up to 10 minutes ensuring at least 50% of time beam is on and at least 1 beam trip occurs
9	End run and reset CODA back to configure stage.
10	Ctl-C out of start_capture.sh script on ejfat-fs-daq and restart it.
11	Request **70nA** beam
12	Transition through CODA Download and Prestart phases
13	Wait for beam to stabilize
14	Start run
15	Capture data for up to 10 minutes ensuring at least 50% of time beam is on and at least 1 beam trip
16	Repeat steps 9-15 for additional beam currents: 40nA, 60nA, 25nA
17	Time permitting, Repeat steps 9-15 for additional beam currents: 80nA, 50nA, 15nA(if permitted)

Table 2: Specific steps in the run plan for beam-on data capture of the CLAS12 detector.

For completeness, the start_capture script.sh has contents:

```
#!/bin/bash

start_port=7001  # Replace with your start port
end_port=7018    # Replace with your end port
export CAPTURE_DIR="/nvme/proj/RTDP/2024.05.17.CLAS12/files"

# Define routine to kill any netcat instances. This will
# be run at the begining and end of this script as well
# as when a SIGINT (Ctrl-C) is received.
killall_nc() {
echo "Killing all netcat listeners ..."
killall nc
set +x
}

# Trap SIGINT so we can stop all netcat programs
trap killall_nc, SIGINT

killall_nc

# Start up a netcat (nc) instance for each port we are capturing.
for ((port=$start_port; port<=$end_port; port++))
do
   nc -l -p $port > /dev/null &
done
echo "Netcat is listening on ports from $start_port to $end_port"


# Define the capture file name.
datetime=$(date +"%Y-%m-%d_%H-%M-%S")
export CAPTURE_FILE="${CAPTURE_DIR}/CLAS12_ECAL_PCAL_DC_${datetime}.pcap"

# Start tcpdump to capture all streams into a single file.
# This will block until the user stops it with Ctl-C
echo "Capturing TCP packets from ports: ${start_port}-${end_port}"
echo "Capturing into file: ${CAPTURE_FILE}"
set -x
sudo tcpdump -i enp193s0f1np1 -j adapter_unsynced --time-stamp-precision=nano \
  -s 0 -w ${CAPTURE_FILE}  portrange ${start_port}-${end_port} -B 10480000
set +x

echo "Capture stopped."
killall_nc

exit 0
```

3 INDRA Data Capture

This exercise will capture data from the SAMPA system in the INDRA lab by sending it to an EJFAT computer in the Data Center across the hall. The SAMPA system is currently setup to read 2 of the 5 cards. These have pulser signals being fed into them for testing purposes. The goals of this exercise are:

1. Capture streaming data that can be used to help with RTDP development offline. (This will be in a different format than what was captured from CLAS12.)
2. Exercise the EJFAT Load Balancer to redirect traffic and establish a working testbed for RTDP that utilizes EJFAT.

The SAMPA chips are able to send data in DAS or DSP mode. The DAS mode includes all samples and requires more bandwidth than is available in the current system. The DSP mode does zero suppression on the front end reducing the bandwidth requirement significantly.

The exercise will consist of 3 parts illustrated in figure 1. The first part will simply send data directly to the ejfat-fs-daq NIC and capture the packets with hardware timestamps as they arrive. This is very similar to the CLAS12 Partial Data Capture exercise described in section 2. Part II will repeat this, but with the data packets directed through the EJFAT Load Balancer installed in ejfat-fs. Part III will repeat part II, but instead of capturing the packets with tcpdump, the packets will be directed into a reassembly engine before being passed into an analyzer program that will perform some minimial interpretation of the data.

Several things are needed before the exercise can be attempted.

1. Verify that the system still functions and the software to capture data into files on alkaid still operates.
2. Either Document or reference existing documentation on how to operate this using the RTDP wiki
3. Document the bandwidth limits of the SAMPA to alkaid connection, the maximum bandwidth of DAS mode, and the maximum bandwidth of DSP mode.
4. Test the network connection from alkaid to ejfat-fs-daq using iperf2. Record the maximum achievable bandwidth with a single stream and with multiple streams.
5. Obtain the modified ALICE code Vardan created for reading the SAMPA data from the DMA pool and modify it to send over the network to a specified UDP host/port
6. Configure the EJFAT LB on ejfat-fs. Document how this is done on the RTDP wiki.
7. Add link to documentation on SAMPA data format to RTDP wiki.
8. Identify existing code for parsing and processing data from the SAMPA system. Modify it to accept data from a UDP port as the input.

Control Plane Command (run on ejfat-2):

```
docker compose -f /home/goodrich/esnet/esnet-smartnic-fw/sn-stack/docker-compose.yml -f
/home/goodrich/esnet/test-scenarios/docker-compose.yml exec udplbd udplbd start
```

4 Collection of Streaming Data from CLAS12

Development of the platform will benefit the most with a complete set of stream data from the full CLAS12 detector. This can be done by simply dumping the outputs of all VTP modules while operating in streaming mode to individual files while beam is on. It will be important for the streams to contain accurate time information for synchronization. It will also be important to capture meta data on the more macro time structure that will allow the streams to be played back with similar “burstiness” as when running in real time.

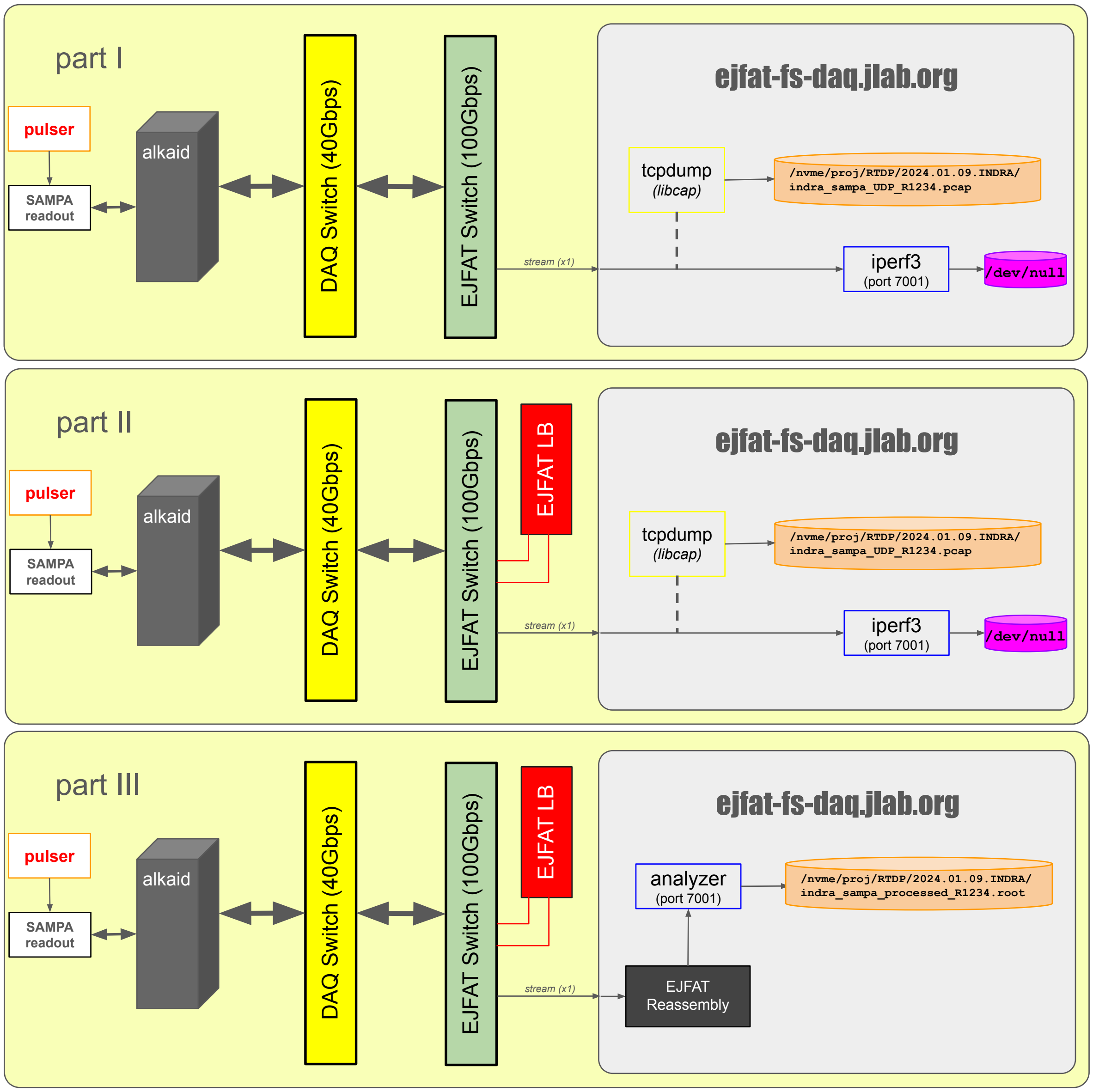

Figure 1: Configurations for the 3 parts of the INDRA capture test using the SAMPA readout.

The best opportunities for collecting full streaming data from the CLAS12 detector will be near configuration changes. Periods near the configuration changes will have some time when the beam is off allowing for setup and DAQ configuration changes. This will include changing the firmware on the VTP modules. Also, beam has historically been more likely to be of lower quality after configuration changes to the accelerator. This tends to make it less valuable to the running experiment and therefore, a good option for this type of beam test.

start date	end date		target
2023-09-30	2023-12-17	Run Group D	-
2023-12-15		pass change Halls B	
2023-12-18	2023-12-19	Run Group K	LH_2
2023-12-20	2023-12-21	Beam Off	
2023-12-22	2024-01-01	Winter Break	
2024-01-02	2024-01-05	Beam Off	
2024-01-06	2024-01-09	Restore	
2024-01-10	2024-03-10	Run Group K	LH_2
2024-02-05		pass change Halls B	
2024-01-10	2024-03-10	Run Group K	LH_2
2024-03-11	2024-03-14	Reconfigure	
2024-03-15	2024-05-19	Run Group E	Nuclear
2024-05-20	2024-09-05	SAD	

Table 3: Hall-B Schedule for FY24[2][3]. The green lines are planned configuration changes adjacent to beam-on days. The yellow lines are periods when beam will be off. White lines are beam-on production periods. The target information comes from the Hall-B Run Groups page[4].

Table 3 shows the latest schedule for Hall-B running in FY24[2]. A good opportunity will be on Dec. 15, 2023 when the conditions are changed from Run Group D to Run Group K accompanied by a corresponding pass change for Hall-B. Up to two hours should be dedicated for the exercise with a goal of capturing at least 10min of beam on data with a typical trip rate during that time.

Additional opportunities should be tentatively scheduled for Dec. 19, 2023 (just before beam goes off for the winter break). Other opportunities can be found during target filling and emptying exercises when beam may be available, but not as useful to the running experiment. These will need to be coordinated with the Run Group K Run Coordinator(s).

Specific steps in the data capture run plan for this exercise can be seen in table 4

5 High Bandwidth Test

High Bandwidth testing will occur mainly during the FY24 and FY25 (Schedule Accelerator Down) SAD periods. Currently, the FY24 SAD is scheduled to begin May 20, 2024 (see table 3). The FY25 SAD will coincide with the Y2Q3 milestones listed in the proposal. Specifically:

Y2Q3
M22: Establish working test of system that transfers >=100Gbps from CH to compute center
M23: Establish working test of system that includes GPU component for portion of stream
M24: Establish working test of system that includes FPGA component for portion of stream
M25: Test system with remote compute facility (e.g. BNL or NERSC) at limits of available resources

Table 5 list specific tasks that would be done during the Y2Q3 High Bandwidth test. As a reminder, the goal of the test is not simply to prove operability at high bandwidth but to show that the platform can be used to test complex, high bandwidth designs containing multiple components which may interact in unforeseen ways once inserted into a larger ecosystem.

Run Plan for CLAS12 SRO beam-on data Capture	
step	**Task**
1	Ensure adequate disk space for storing stream file on CLON cluster
2	Update VTP firmware with SRO version
3	Load CODA configuration for SRO test
4	Ensure LH_2 target is full
5	Request **10nA** beam
6	Wait for beam to stabilize
7	Start run
8	Capture data for up to 10 minutes ensuring at least 50% of time beam is on and at least 1 beam trip
9	Request **50nA** beam
10	Wait for beam to stabilize
11	Start run
12	Capture data for up to 10 minutes ensuring at least 50% of time beam is on and at least 1 beam trip
13	Request **75nA** beam
14	Wait for beam to stabilize
15	Start run
16	Capture data for up to 10 minutes ensuring at least 50% of time beam is on and at least 1 beam trip
17	Restore VTP firmware
18	Restore CODA configuration

Table 4: Specific steps in the run plan for beam-on data capture of the CLAS12 detector.

High Bandwidth Test	
step	**Task**
0	Reserve *Nnodes* on SciComp farm + 1 SciML node with GPU
0	Prepare EJFAT node(s) and configure to use reserved pool
0	Start idle data movement, calibration, and reconstruction jobs manually on CC nodes
1	Update VTP firmware with SRO version with testing support
2	Stage prepared data files from earlier beam-on capture
3	Activate RTDP monitoring
4	Begin synchronized playback at low rate (5% of estimated maximum)
5	Monitor operation. Identify any bottlenecks
6	Increase stream rate to 10% of estimated maximum and repeat
7	Increase stream rate to 50% of estimated maximum and repeat
8	Increase stream rate to 100% of estimated maximum and repeat
9	Remove limits on stream rate and observe response to backpressure
10	Change configuration to pass Drift Chamber Streams through GPU node
11	Observe performance for minimum of 1hr
12	Change configuration to include FPGA modules
13	Observe performance for minimum of 1hr
14	Run failure mode tests (TDB)
15	Stop streams and RTDP programs.
16	Clear temporary disk space. Release CC resources.
17	Restore VTP firmware

Table 5: Specific steps in the run plan for beam-on data capture of the CLAS12 detector.

Required Resources FY24	
Exclusive use period (beam off)	14 days
Storage Hall-B CH	40TB
Storage Bandwidth Hall-B CH	30Gbps
Networking	- VTP to switch 10Gbps optical links - 100Gbps switch to CH

Table 6: Specific steps in the run plan for beam-on data capture of the CLAS12 detector.

Required Resources FY25	
Exclusive use period (beam off)	14 days
Disk Storage Computer Center	200TB
Storage Bandwidth	200Gbps
Networking	- 400Gbps switch to CH - 400Gbps Hall-B CH to CC

Table 7: Specific steps in the run plan for beam-on data capture of the CLAS12 detector.

References

[1] F. Barbosa, L. Belfore, N. Branson, C. Dickover, C. Fanelli, D. Furletov, S. Furletov, L. Jokhovets, D. Lawrence, and D. Romanov. Development of ml fpga filter for particle identification and tracking in real time. *IEEE Transactions on Nuclear Science*, 70(6):960–965, 2023.

[2] Beam Time Manager Version 4 (Published 23 ott 2023). https://ace.jlab.org/btm/schedule/october-2023.

[3] CURRENT SCHEDULE: October 2023 - May 2025. https://www.jlab.org/sites/default/files/user-liaison/files/Acc%20Schedule%202023-10-23.xlsx.

[4] Hall B – Run Groups. https://userweb.jlab.org/~doug/Schedule/2022/HallB_RunGroup_20222102.pdf.