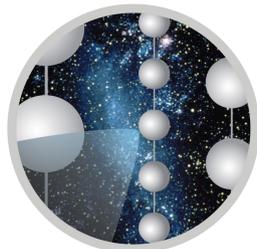# IceCube Data Filtering Review and Future Plans
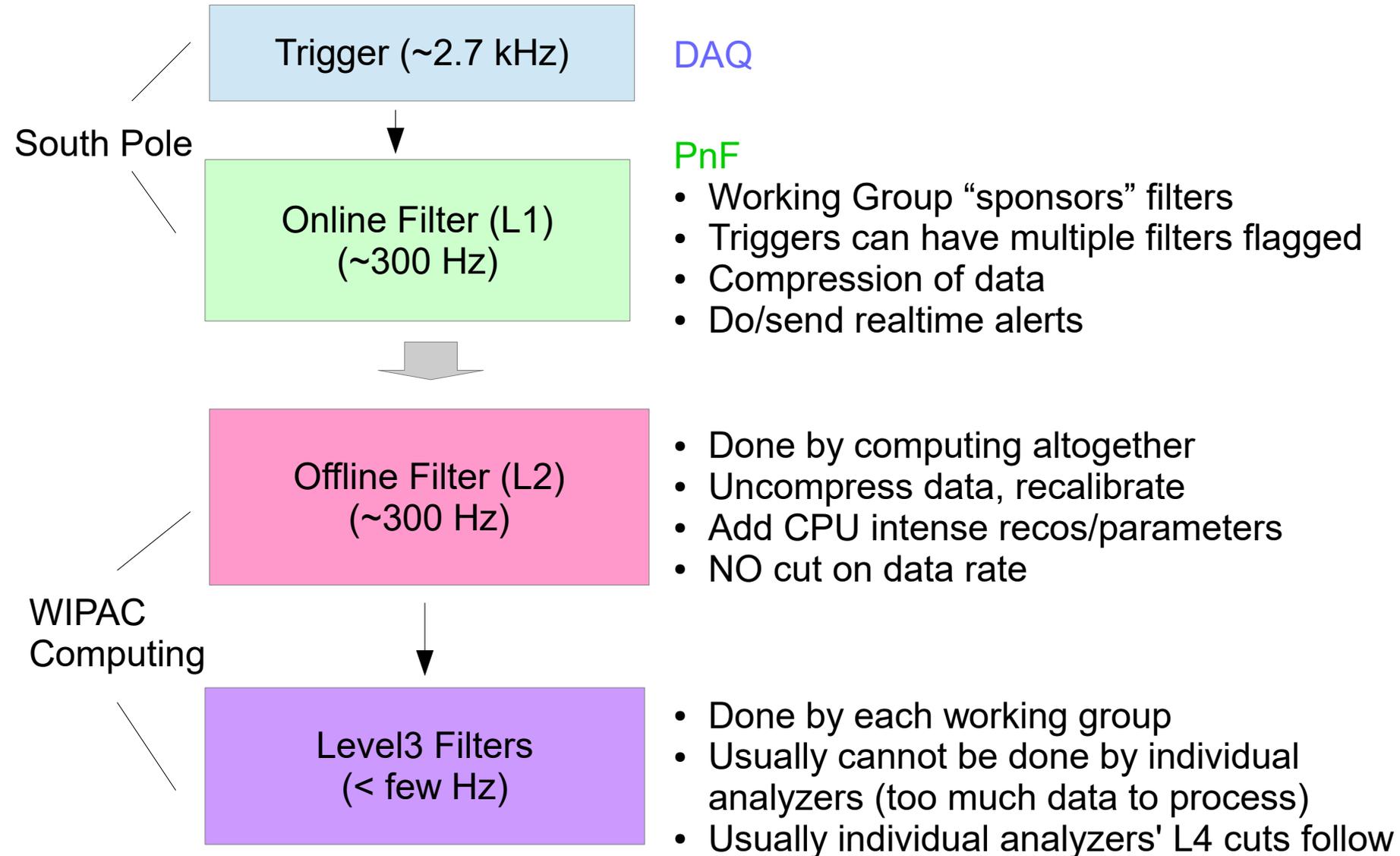
Naoko Kurahashi Neilson
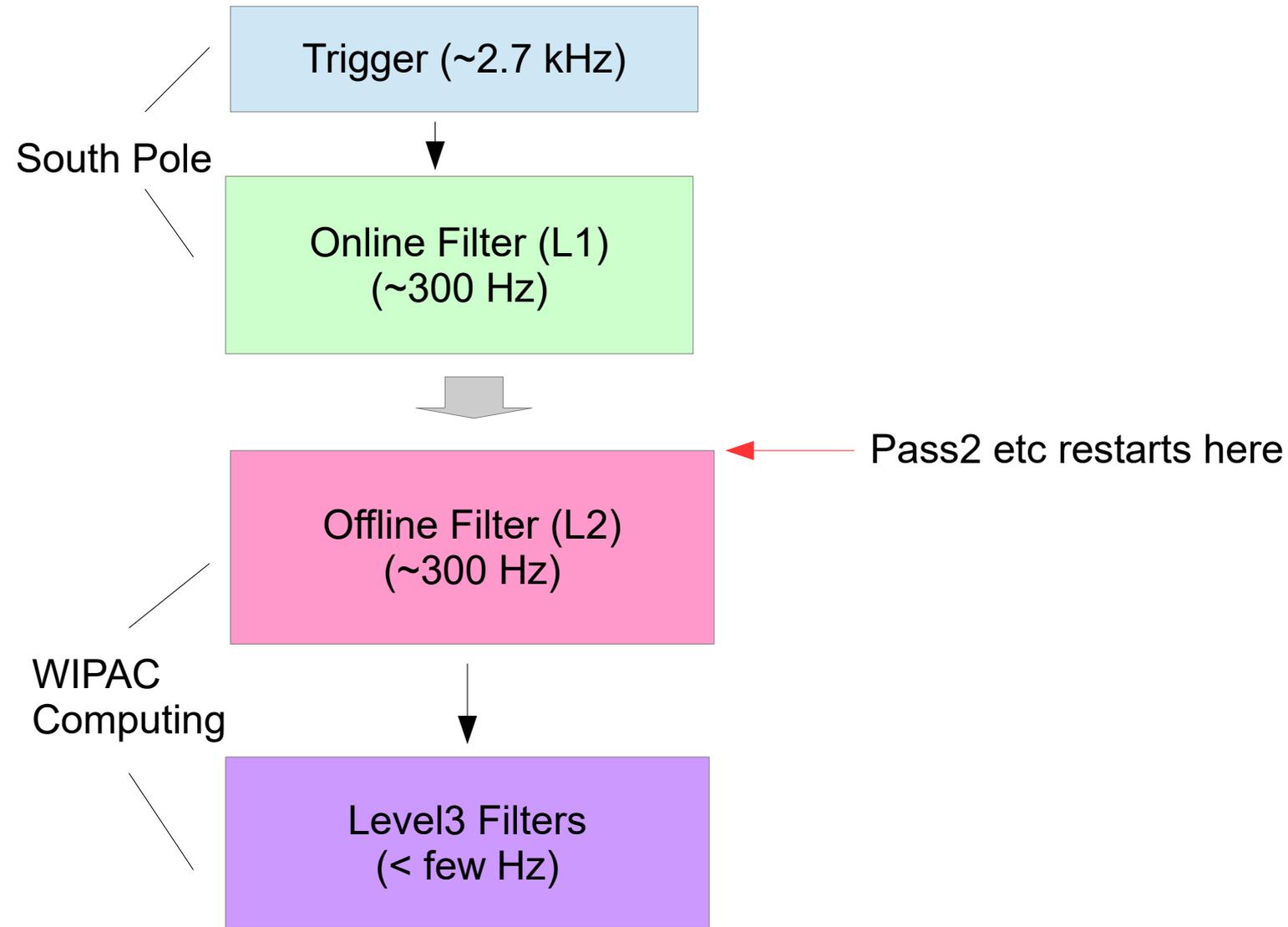Drexel University
IceCube TFT Co-Chair

IceCube

Drexel
UNIVERSITY

SCAP 2021, 1/7/2021

Overview

Trigger (~2.7 kHz)

DAQ

Online Filter (L1)
(~300 Hz)

PnF
- Working Group "sponsors" filters
- Triggers can have multiple filters flagged
- Compression of data
- Do/send realtime alerts

Offline Filter (L2)
(~300 Hz)

WIPAC
Computing

- Done by computing altogether
- Uncompress data, recalibrate
- Add CPU intense recos/parameters
- NO cut on data rate

Level3 Filters
(< few Hz)

- Done by each working group
- Usually cannot be done by individual analyzers (too much data to process)
- Usually individual analyzers' L4 cuts follow

Overview

Trigger (~2.7 kHz)

Online Filter (L1)
(~300 Hz)

South Pole

Offline Filter (L2)
(~300 Hz)

Pass2 etc restarts here

WIPAC
Computing

Level3 Filters
(< few Hz)

Pass 2

Motivation: Re-process all IceCube data with uniform reconstructions & filters. Apply improved SPE corrections.

Seasons 2010-2015, half of 2016: Processing complete at time of previous SCAP

Second half of season 2016 & L3 for all seasons: Processed in Aug-Sep 2018, up to 3k simultaneous jobs on NPX (1 job per input file)

1. DST production from Raw (NPX)
2. L2 production from DST (NPX)
3. L3 production (NPX & Grid): Muon, Cascade, Low Energy

Issues: slow job submission due to checksum calculation solved by running parallel scripts

Resources: ~13M CPU hours, 251 TB (L2)

Pass2a

Processed all Level2 Pass2 data (seasons 2010-2016) with Pass2a which fixes:

- Leap-second bug: 3 years of data are one second behind UTC time due to a master clock firmware bug

- Missing reconstructions (LineFit, SPE, SPE2, MPE) if only Moon filter is on

- MuEx reconstruction in legacy mode for GFU filter (does not affect L3)

Estimate ~1M CPU hours, actual: 933k CPU hours

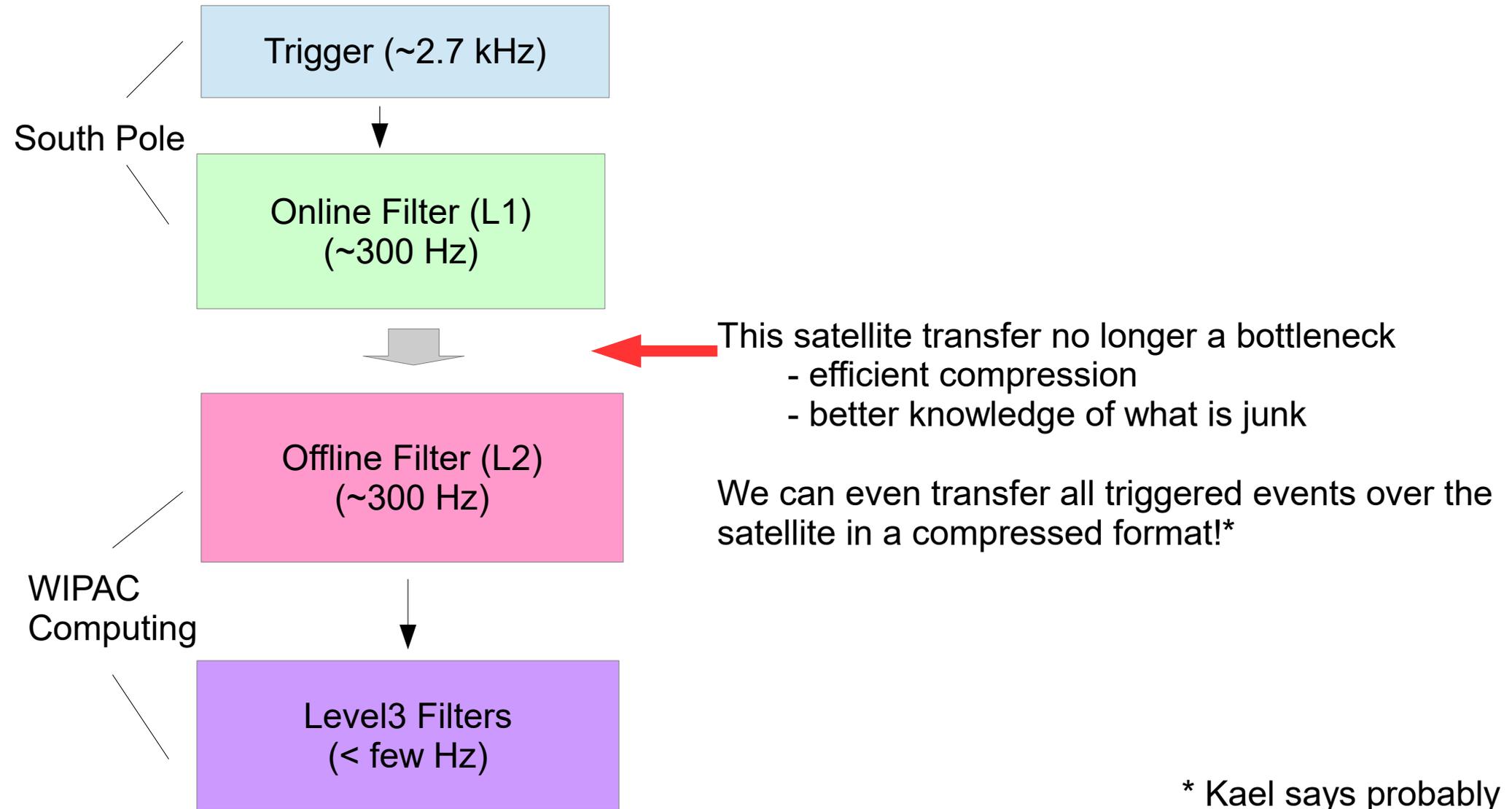6 weeks on NPX with ~500 simultaneous jobs for 7 years of data

1.7M jobs in 36 Condor DAGs
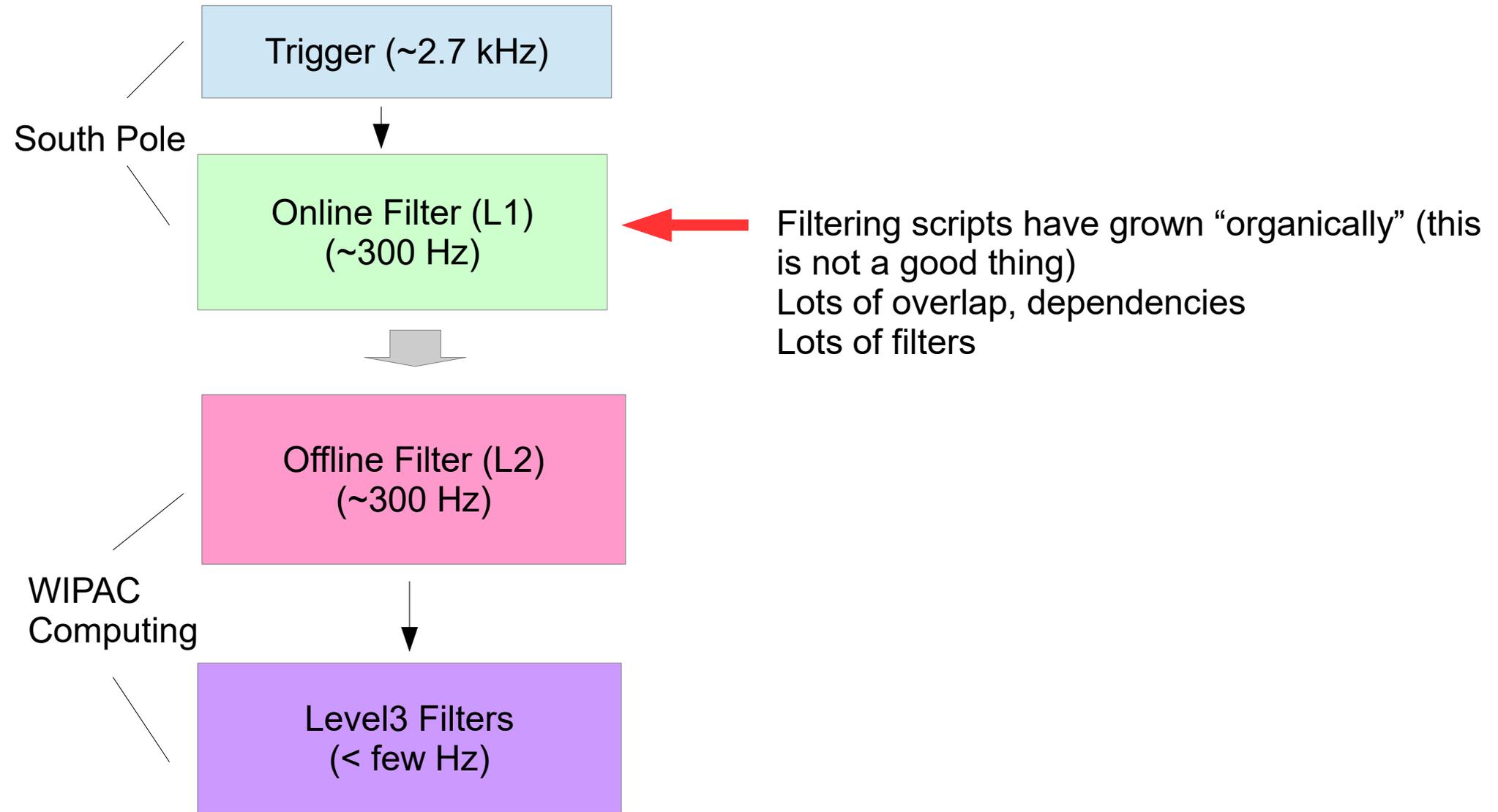
Complete pass2a data set: 271 TB

Jobs for 2013 Jan-Jun DAG

IceProd2

Transition to IceProd2 with new season start: June 2020

IceProd2 operations: generally more errors on Grid vs. NPX, but smooth running with # retries = 10

Other issues:

- Fixed common error with missing output files:
  Create tasks in IceProd2 with status = "idle", add files, change status to "waiting"
- Submission scripts: cron & dryrun options now working; resubmission needs fixing
- Zstd check in post-processing scripts ⇒ Dataio scan in grid jobs
- CalendarView : fixed failing calls
- Log files now captured & stored on disk in post-processing step

Mad dash automatic histograms considered for future monitoring

L2 Jobs by site from July-Dec 2020

Good Updates!

Trigger (~2.7 kHz)

South Pole

Online Filter (L1)
(~300 Hz)

This satellite transfer no longer a bottleneck
- efficient compression
- better knowledge of what is junk

Offline Filter (L2)
(~300 Hz)

WIPAC
Computing

We can even transfer all triggered events over the
satellite in a compressed format!*

Level3 Filters
(< few Hz)

* Kael says probably

Issues That We Can Work On 1

Trigger (~2.7 kHz)

South Pole

Online Filter (L1)
(~300 Hz) ← Filtering scripts have grown "organically" (this is not a good thing)
Lots of overlap, dependencies
Lots of filters

Offline Filter (L2)
(~300 Hz)

WIPAC
Computing

Level3 Filters
(< few Hz)

Lets look at some details...

Current Filters

Neutrino Sources

- GFUFilter_17
- GRECOOnlineFilter_19
- MuonFilter_13
- OnlineL2Filter_17

Cosmic Ray

- IceActTrigFilter_18
- IceTop_InFill_STA2_17
- IceTop_InFill_STA3_13
- IceTopSTA3_13
- IceTopSTA5_13
- InIceSMT_IceTopCoincidence_13
- MoonFilter_13
- ScintMinBias_16
- SDST_IceTop_InFill_STA3_13
- SDSTIceTopSTA3_13
- SDST_InIceSMT_IceTopCoincidence_13
- SunFilter_13
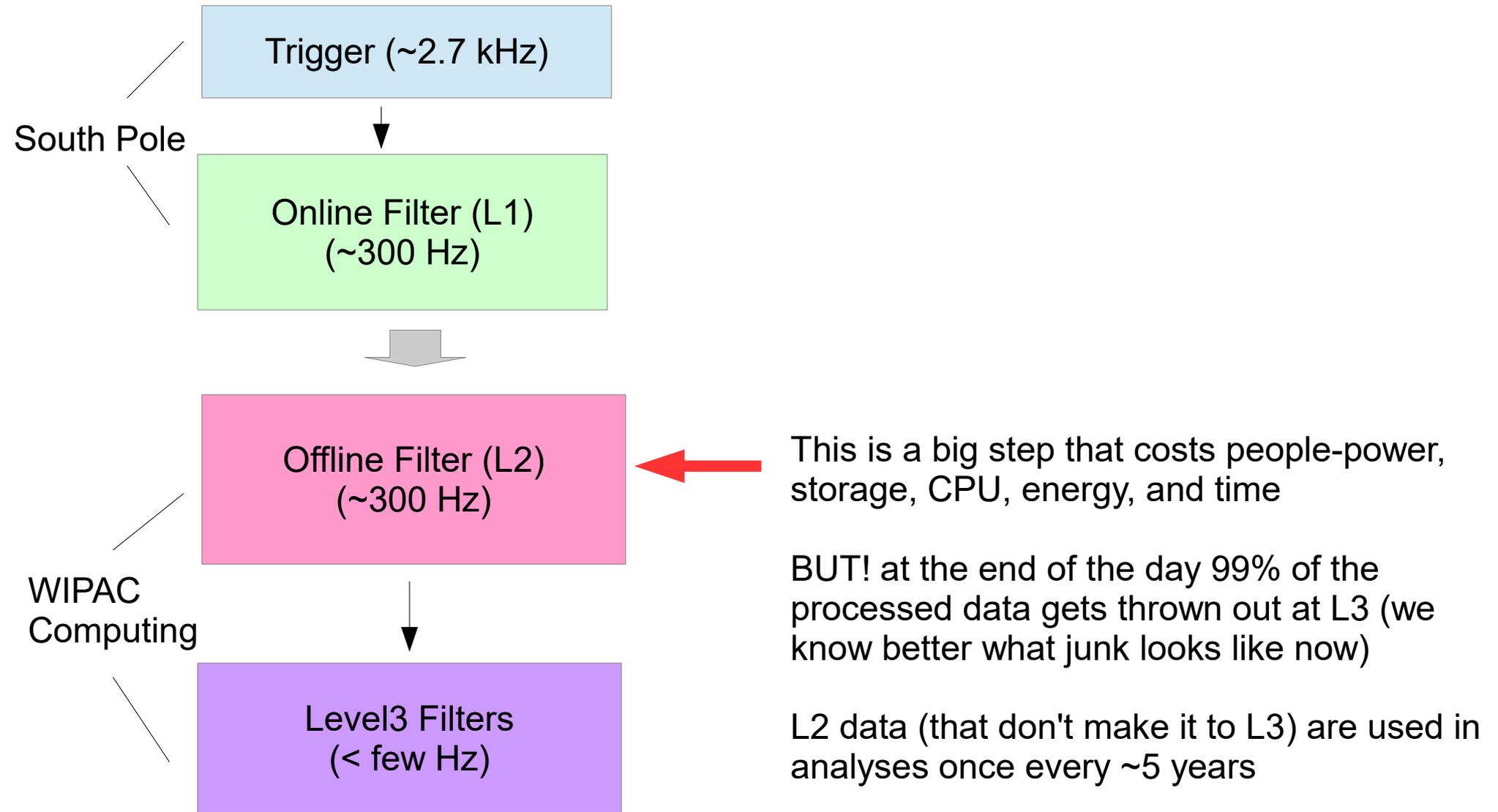
Diffuse

- CascadeFilter_13
- EHEAlertFilter_15
- EHEAlertFilterHB_15
- EstresAlertFilter_18
- HESEFilter_15
- HighQFilter_17
- MESEFilter_15

Beyond Standard Model

- FSSCandidate_13
- FSSFilter_13
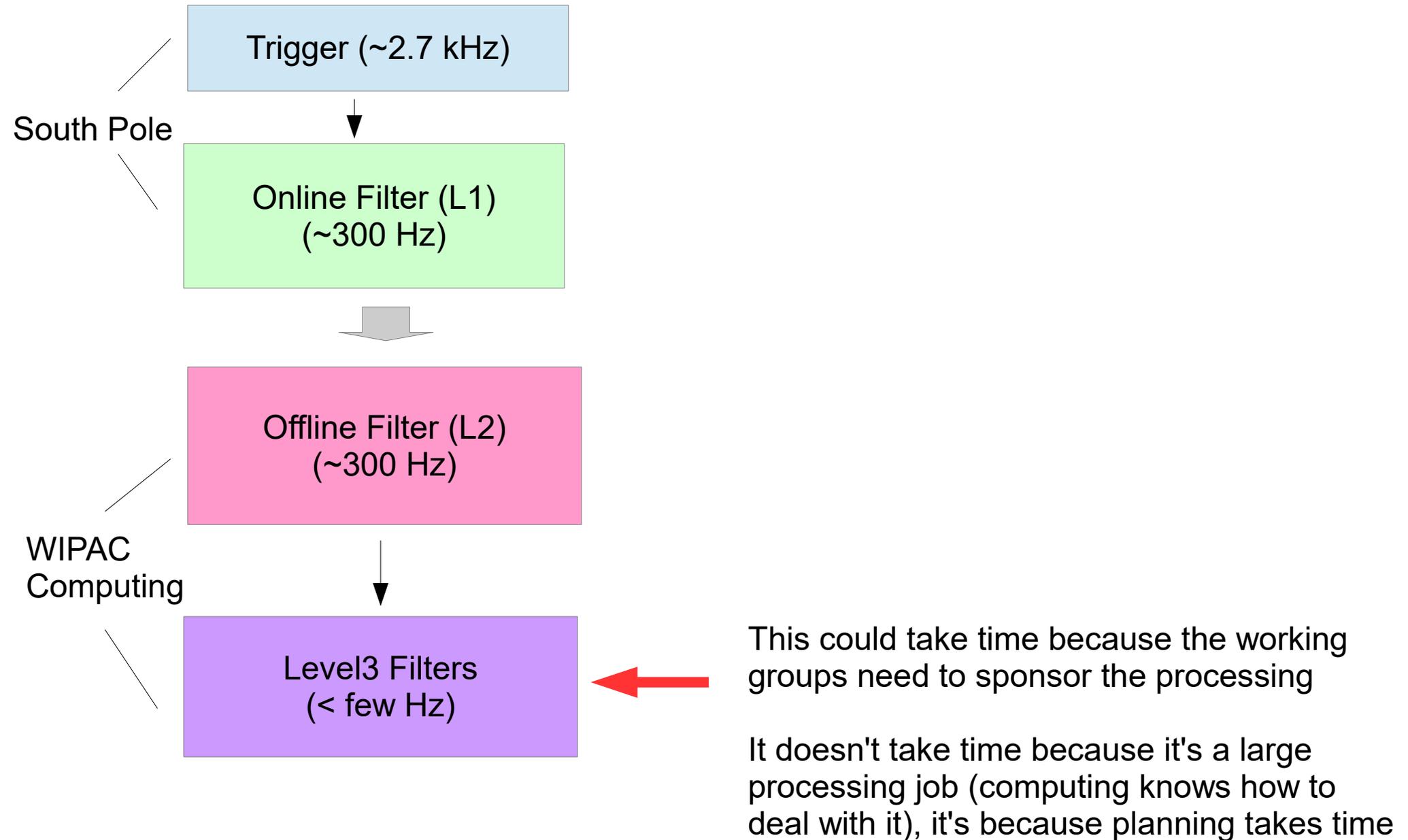- LowUp_13
- MonopoleFilter_16
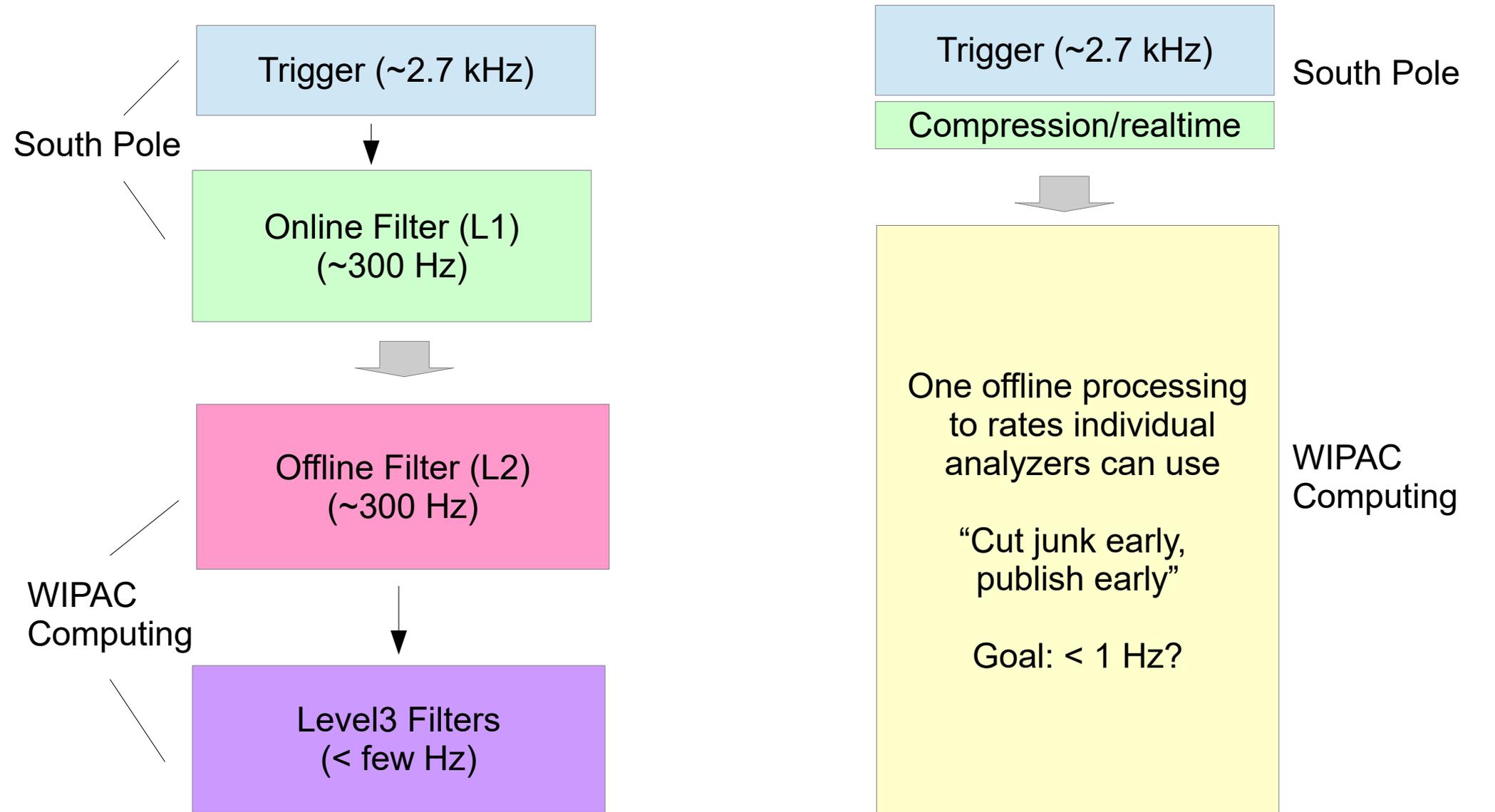- SlopFilter_13
- VEF_13

Issues That We Can Work On 2

Trigger (~2.7 kHz)

↓

Online Filter (L1)
(~300 Hz)

⬇

Offline Filter (L2)
(~300 Hz)

↓

Level3 Filters
(< few Hz)

South Pole

WIPAC
Computing

← This is a big step that costs people-power, storage, CPU, energy, and time

BUT! at the end of the day 99% of the processed data gets thrown out at L3 (we know better what junk looks like now)

L2 data (that don't make it to L3) are used in analyses once every ~5 years

Filter	2020 test run [Hz]	2020 test run [GB/day]
FSSFilter	187.3	7.3
MoonFilter	100.0	4.3
SDST_InIceSMT_IceTopCoinc	57.6	3.2
HighQFilter	0.8	2.6
SLOPFilter	10.9	2.5
MuonFilter	34.5	1.8
MonopoleFilter_16	30.1	1.4
CascadeFilter	33.2	1.2
IceTopSTA5	1.3	1.1
LowUp	27.6	0.9
FilterMinBias	2.7	0.8
MESEFilter_2015	9.4	0.6
DeepCore	16.2	0.5
InIceSMT_IceTopCoinc	0.7	0.4
ICOnlineL2Filter	5.2	0.4
VEFFilter	12.0	0.4
IceTopSTA3	0.6	0.4
SDST_IceTopSTA3	5.9	0.2
FixedRateFilter	0.0	0.1
ScintMinBias_16	2.2	0.1
IceTop_InFillSTA3	0.1	0.1
IceTop_InFill_STA2_17	2.4	0.1
SDST_IceTop_InFill_STA3	1.2	0.0

Lets look at some details....

Issues That We Can Work On 3

Trigger (~2.7 kHz)

Online Filter (L1)
(~300 Hz)

South Pole

Offline Filter (L2)
(~300 Hz)

WIPAC
Computing

Level3 Filters
(< few Hz)

This could take time because the working groups need to sponsor the processing

It doesn't take time because it's a large processing job (computing knows how to deal with it), it's because planning takes time

New Way Forward?

Trigger (~2.7 kHz)

South Pole

Online Filter (L1)
(~300 Hz)

Offline Filter (L2)
(~300 Hz)

WIPAC
Computing

Level3 Filters
(< few Hz)

Trigger (~2.7 kHz)

South Pole

Compression/realtime

One offline processing
to rates individual
analyzers can use
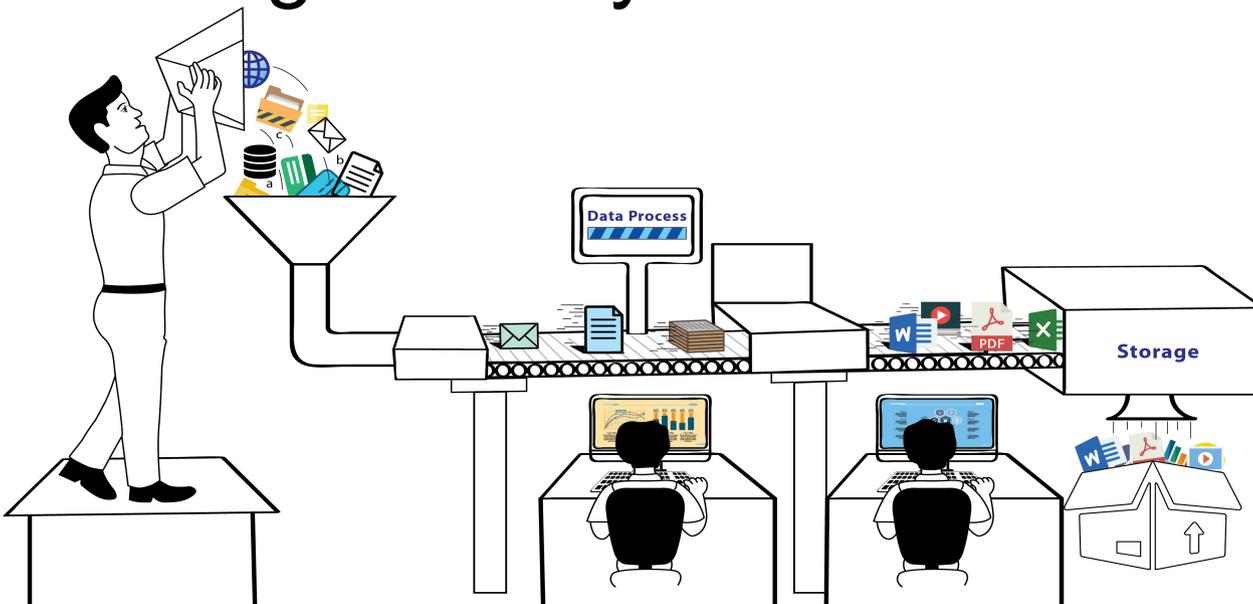
"Cut junk early,
publish early"

Goal: < 1 Hz?

WIPAC
Computing

* Exceptions will happen

Other Benefits

- Chance to re-write filters cleanly and clearly so that even new students can follow

- Clean up dependence on old software that's obsolete (stop supporting extra stuff we don't really need)

- Free up computing resources (both people and CPU time) for future detector data streams

Conclusion

- Data Processing stable, familiar → "If it's not broken, don't touch it"

- Maybe coming to a point where the downside of "leaving it alone" is too great to ignore

- Time may be right, in many ways, to rethink and redesign cleanly with all that we've learned