NEC'2017

STAR's Approach to Highly Efficient End-to-end GRID Production

Levente Hajdu (BNL Presenter) , Yury Panebrattsev (JINR)
Jérôme Lauret (BNL), Evgeniy Kuznetsov (JINR LIT)
Valery Mitsyn (JINR), Lidia Didenko (BNL), Wayne Betts (BNL),
Nikita Balashov (JINR), Geydar Agakishiev (JINR),
Vladimir Korenkov (JINR)

NEC'2017 ОИЯИ BROOKHAVEN NATIONAL LABORATORY STAR U.S. DEPARTMENT OF ENERGY Office of Science

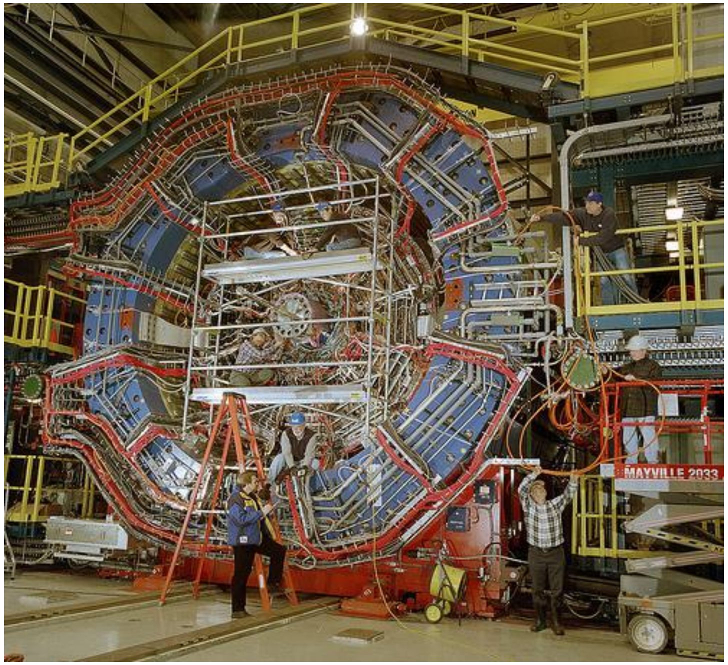
Introduction

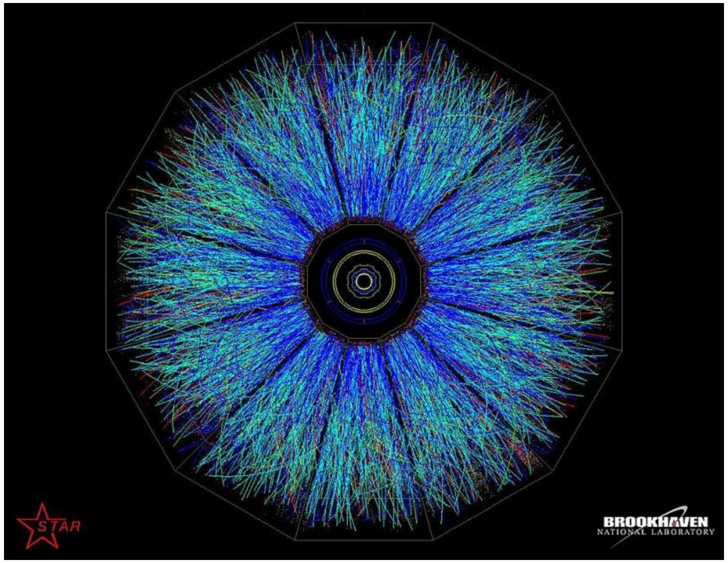
- Introduction to the STAR Experiment
  - Data processing demands
  - STAR’s Production sites
    - Parallel Processing Paradigms
- Introduction to STAR’s GRID Production System
  - Overview
  - Stages, Dataflow, States
- Basic features of a production system:
  - Automated resubmission
  - Multi Site Submission
  - Job feeding with feedback
  - Site selection logic
- Efficiency and statistics

STAR

- **STAR (Solenoidal Tracker At RHIC) is a detector located in one of the interaction regions of the RHIC (Relativistic Heavy Ion Collider)**
  - **second-highest-energy heavy-ion collider in the world**
  - **2.4 miles (3.9 Km) circumference**
- **Took its first data in year 2000 - currently on our 17th physics run (year of data taking).**
- **Very versatile machine 7.7 GeV to 510 GeV wide particle species range from protons-uranium**
- **Able to collide HI and polarized protons**
  - **Heavy-flavor and quarkonia measurement**
  - **Jet measurements**
  - **Chiral magnetic effect, chiral magnetic wave and chiral vortical**
  - **Phase structure of QCD matter – Beam Energy Scan**
  - **Understanding of the nature of the pomeron and potentially discovering the odderon**
  - **Single spin asymmetries in W+/-, Z, direct photon and Drell-Yan production**
- **STAR Virtual tour page:**

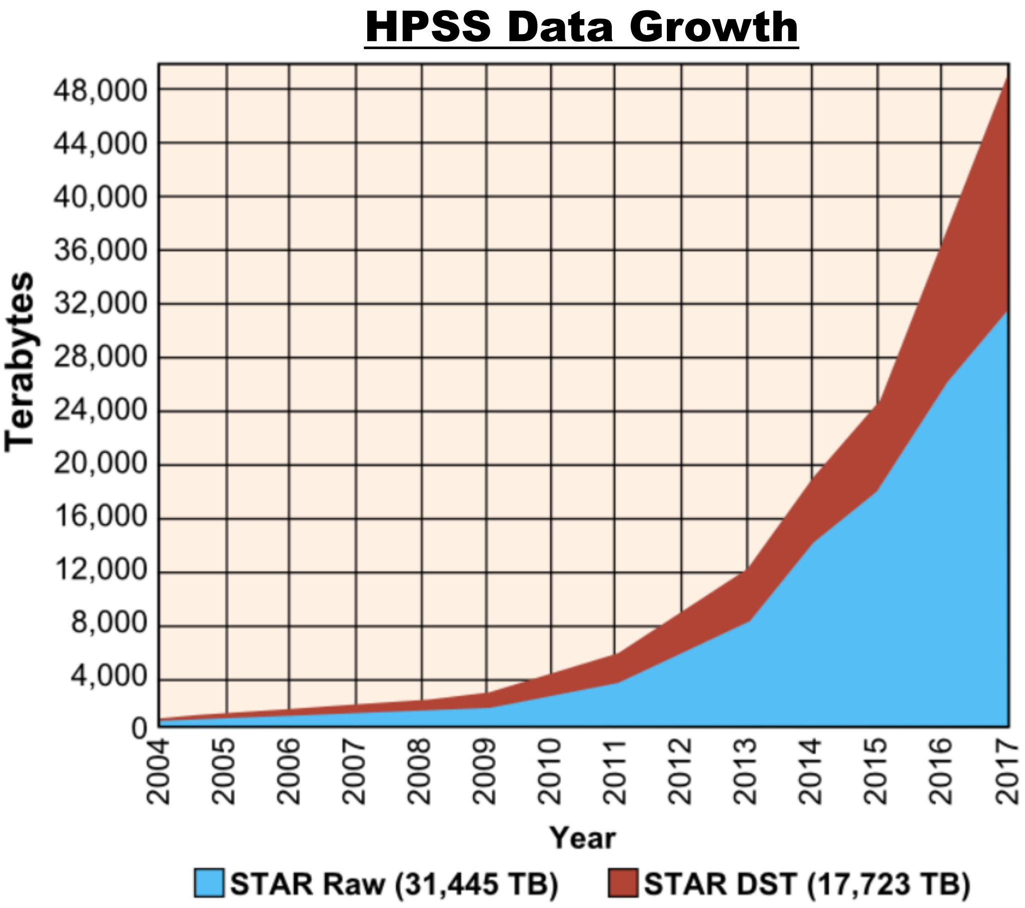
  http://www.star.bnl.gov/public/imagelib/v_tour/tour.html



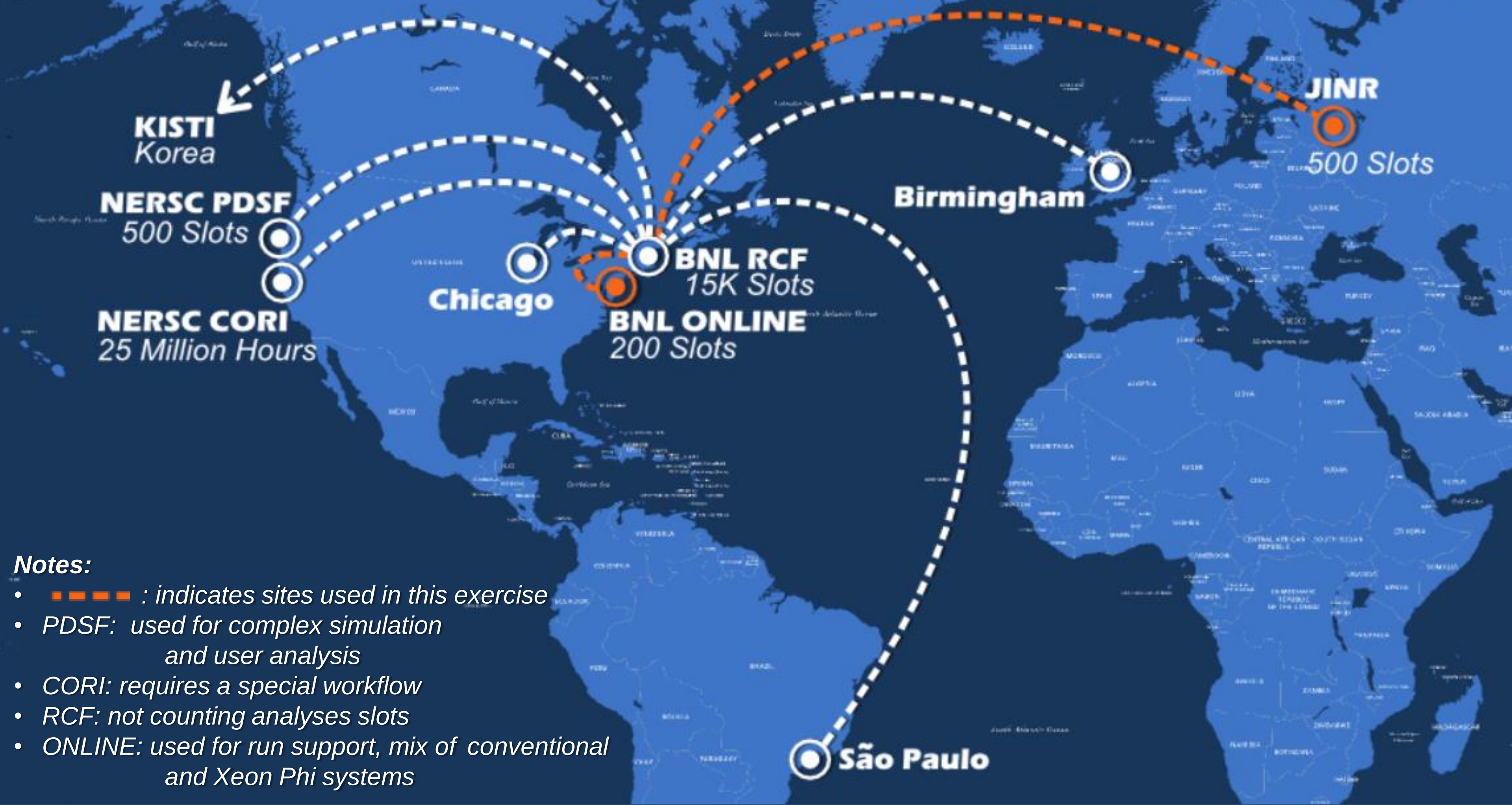


Data Processing Demands

- There is one data-taking run every year; with upgrades, the size of datasets taken each year tend to increase
- Each run produces many datasets
- ~15,000 slots are used for data production at BNL
  - This **ONLY** allows for 1.2-1.4 passes of data reconstruction of a current year
  - In contrast, typical HEP experiments have > 5 passes
- Huge dataset challenges - we seek additional resources to speed up scientific discoveries
- Started using GRID in 2001 for simulation requests and scaled up to different classes of production



STAR'S CURRENT AND FORMER PRODUCTION SITES



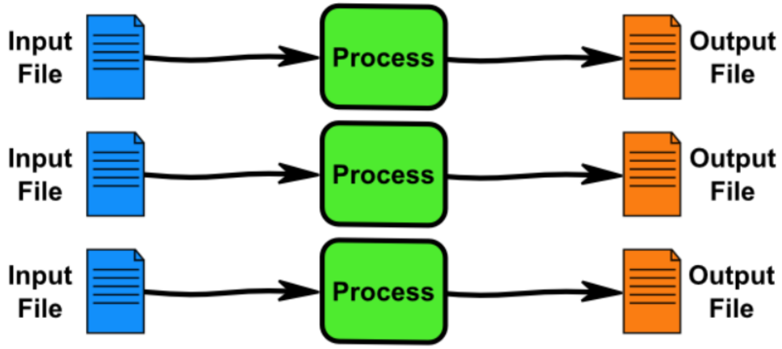
Notes:

- ▬ ▬ ▬ ▬ *: indicates sites used in this exercise*
- *PDSF: used for complex simulation and user analysis*
- *CORI: requires a special workflow*
- *RCF: not counting analyses slots*
- *ONLINE: used for run support, mix of conventional and Xeon Phi systems*

Types of Parallel Computing

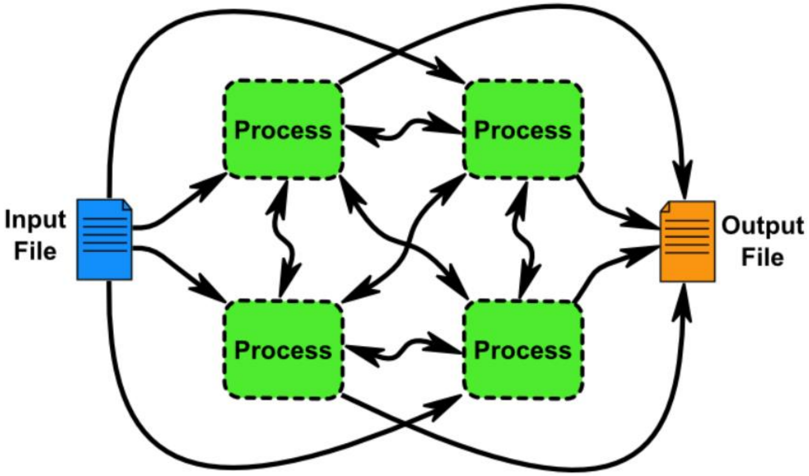
(a hint about our on-going work on Cori, and up-coming talks)

"Pleasantly Parallel"

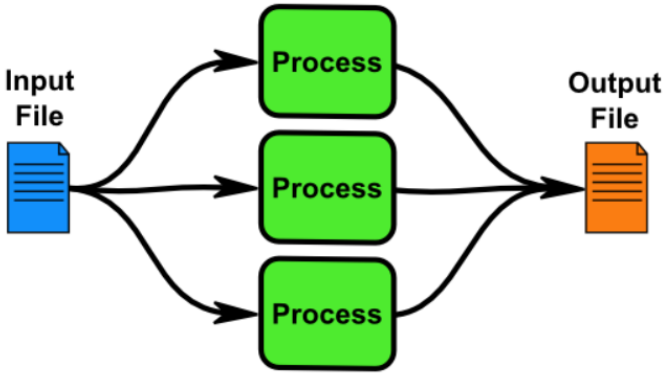


- STAR is traditionally optimized for the "pleasantly parallel" computing model
- Commodity hardware, geographically separated
- No inter-process communication
- One core = One job

- Recent trend – governments and academic institutions are building facilities to solve problems with massive process inter communication
- Massive processors per-slot
- Limited memory, and external I/O
- Can we utilize these systems when not working on this type of workload?



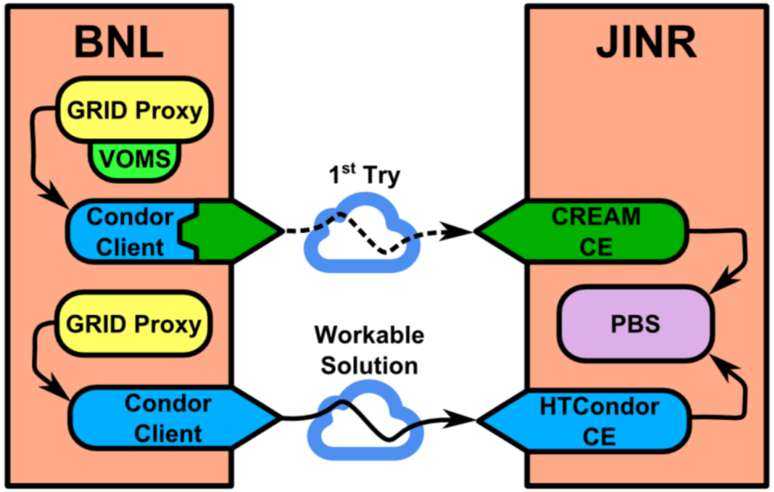
- Event level parallelization - split one file into blocks or ranges of events assigned to individual processor cores and remerge output at the end.
  - The input file contains an array of events, independent of one an other, an analog would be like a PDF file contains different pages.
  - Requires a buffer to rejoin the output.



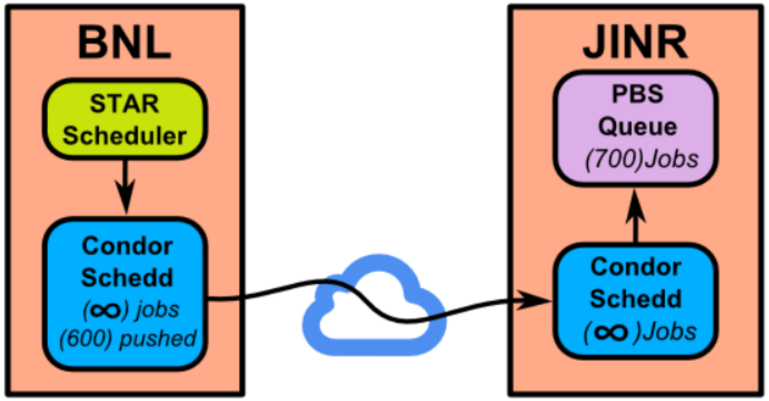
JINR Lessons Learned

- Benchmark bandwidth and submission efficiency in advance (old lesson)
- Initial setup was using local condor client to submit to CREAM authenticating with GRID cert. with VOMS extension, this worked to first order (jobs run) but was not viable for production (Efficiency < 90%).
  - Jobs die as soon as VOMS proxy 3-day extension dies
  - Password-less renew of VOMS extension not working
- JINR setup a CondorCE authenticating with long-lived GRID proxy.
  - HTCondor connector to PBS is not well polished but functions usably:
    - Losing track of some jobs reported as held but still running
    - Network connections transients cause incorrect reporting of runtime
    - Error messages from PBS differ from batch system actual problem
- STAR and JINR negotiated resource allocation
  - 500 running jobs and 700 queued, max runtime of 5 days
    - Over-submission would result in removed jobs, this is prevented by limiting the number of jobs pushed over to the site in the condor schedd.
  - No local staging buffers are available so we will use GridFTP (globus-url-copy) within the jobs runtime to stage input and output files.
    - No event level splitting with local remerge
- We'll move up to 1k jobs Q4 of 2017

Grid Stack

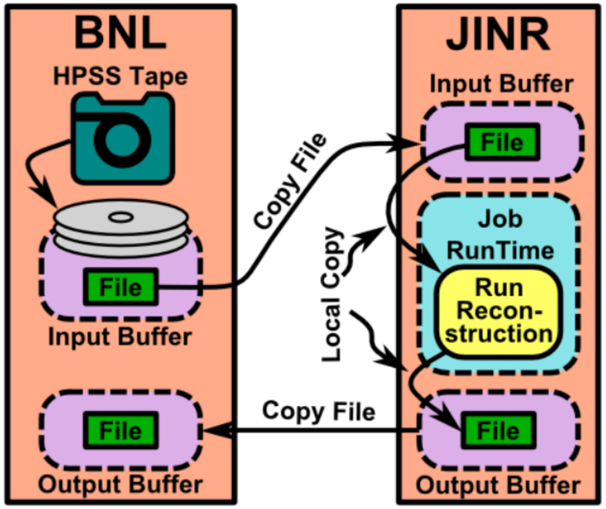


Queue Limits

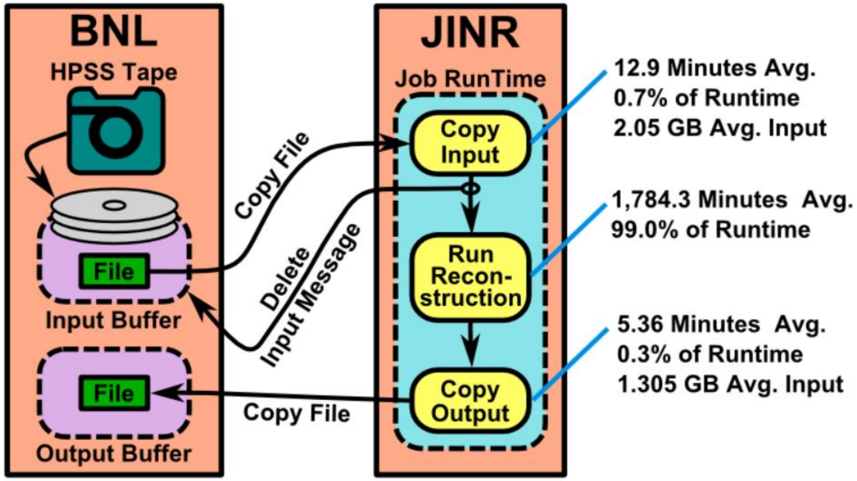


Data Transfer Modes

Outside of Jobs Runtime Using Buffers



Inside Of Runtime (Unreliable)



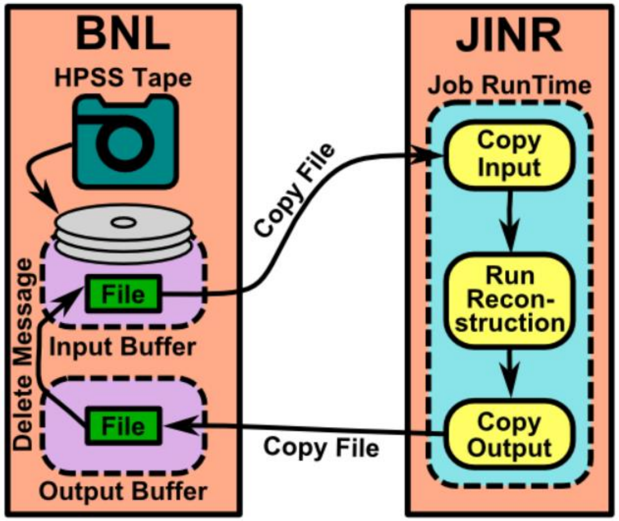
12.9 Minutes Avg.
0.7% of Runtime
2.05 GB Avg. Input

1,784.3 Minutes Avg.
99.0% of Runtime

5.36 Minutes Avg.
0.3% of Runtime
1.305 GB Avg. Input

- Data Transfer Outside of the job's runtime
  - acknowledged as most efficient, transfer can be asynchronous (serialized), but requires large I/O buffers unavailable at JINR but used at Cori NERSC.
  - was tested using the Condor transfer mechanism to JINR but it used the mapped user's $HOME as buffer which was insufficient
- Data Transfer Inside of the job's runtime
  - 1% or less of the jobs total runtime; simplified workflow; no need for host site buffers; used at JINR
  - "Unreliable" mode requires the input files to be restaged from tape if the job fails, but allows more jobs submitted without a bigger buffer
- **In all cases site-to-site copies are done via globus-url-copy, we are being asked if this tool should be phase-out. No replacement exists to transfer files site-to-site with no buffer.**

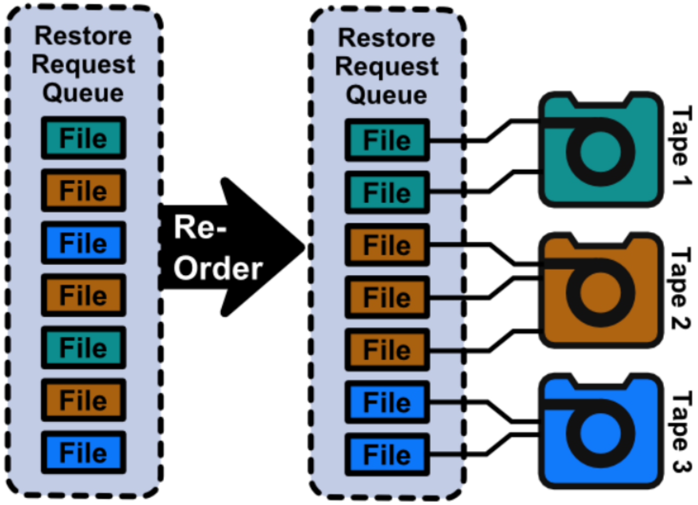
Inside Of Runtime (Reliable)



Existing STAR Tools Reused in the Grid Production Framework

Reuse of long established and well debugged STAR tools minimized development time and provides good reliability and high efficiency.

Data Carousel

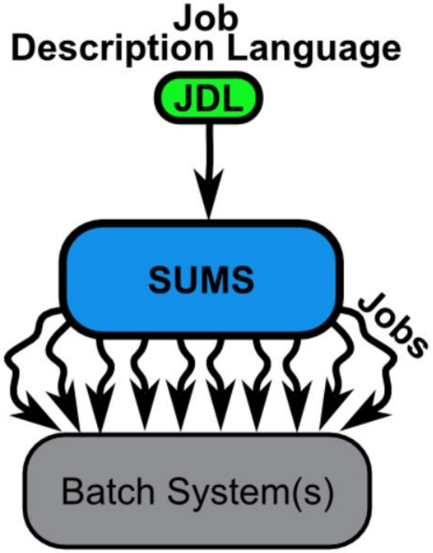


The STAR Data Carousel is a tool for queueing and optimizing requests for the restoration of files from tape by minimizing mount and dismount cycles through reordering.

Link: ACAT 2011 Data Carousel Paper

Link: CHEP 2010 Data Carousel Paper

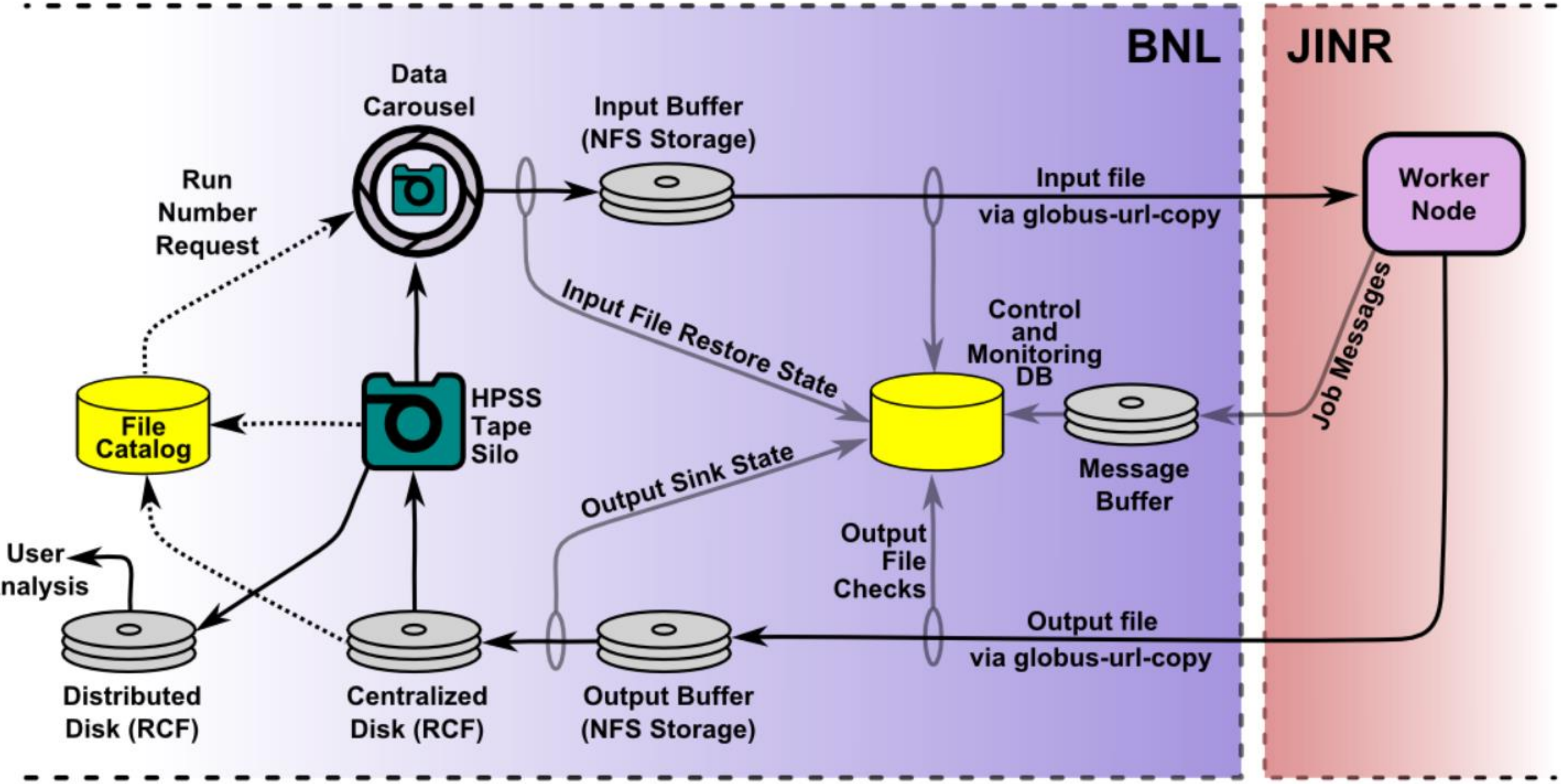
STAR Unified Meta Scheduler



The STAR Unified Meta Scheduler (SUMS), first deployed in 2002, provides a unified interface for submitting jobs to sites and wrapping of the input file and user executable into jobs.

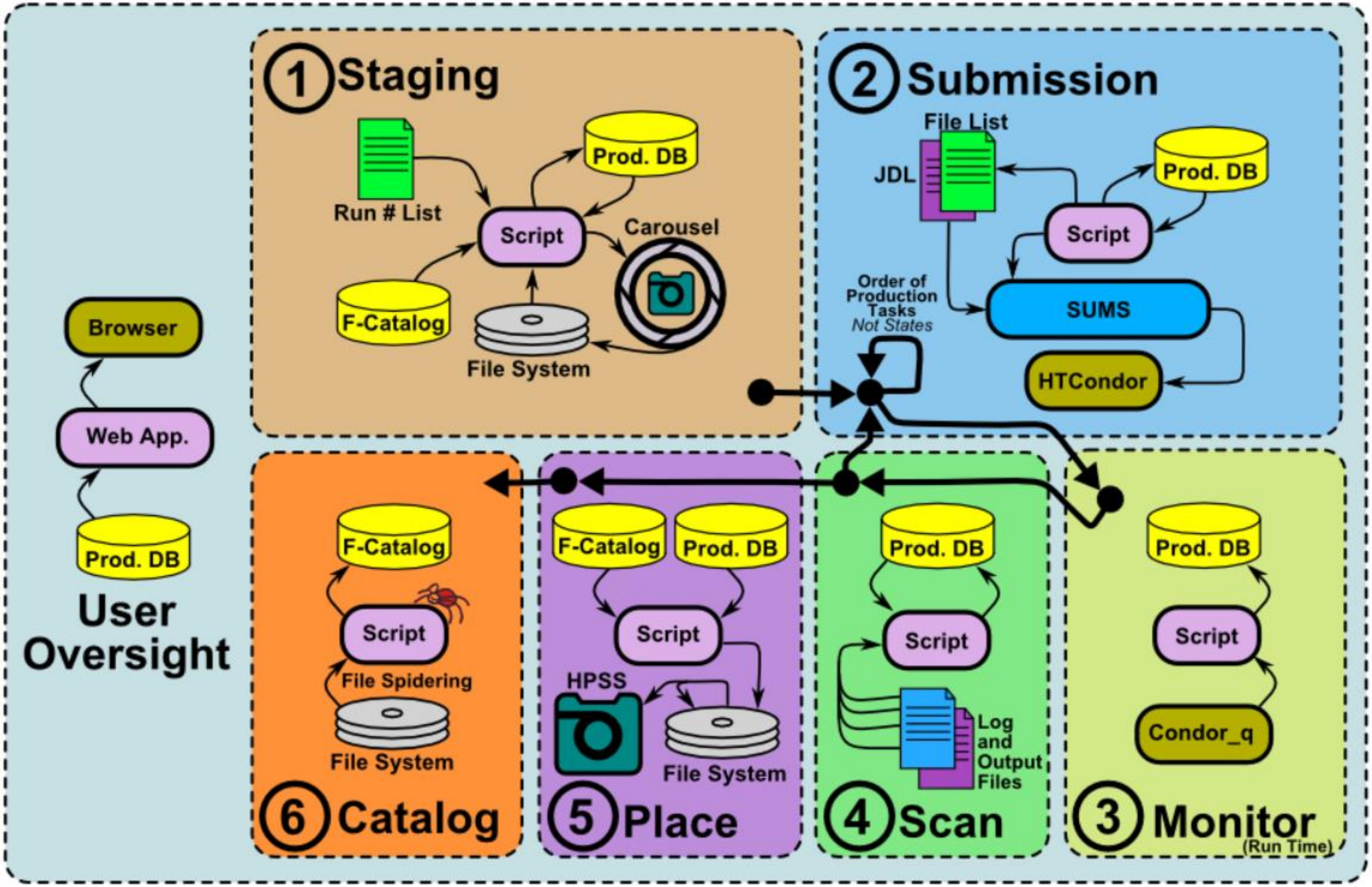
Link: ACAT 2006 SUMS Paper

STAR GRID Production System Data Flow



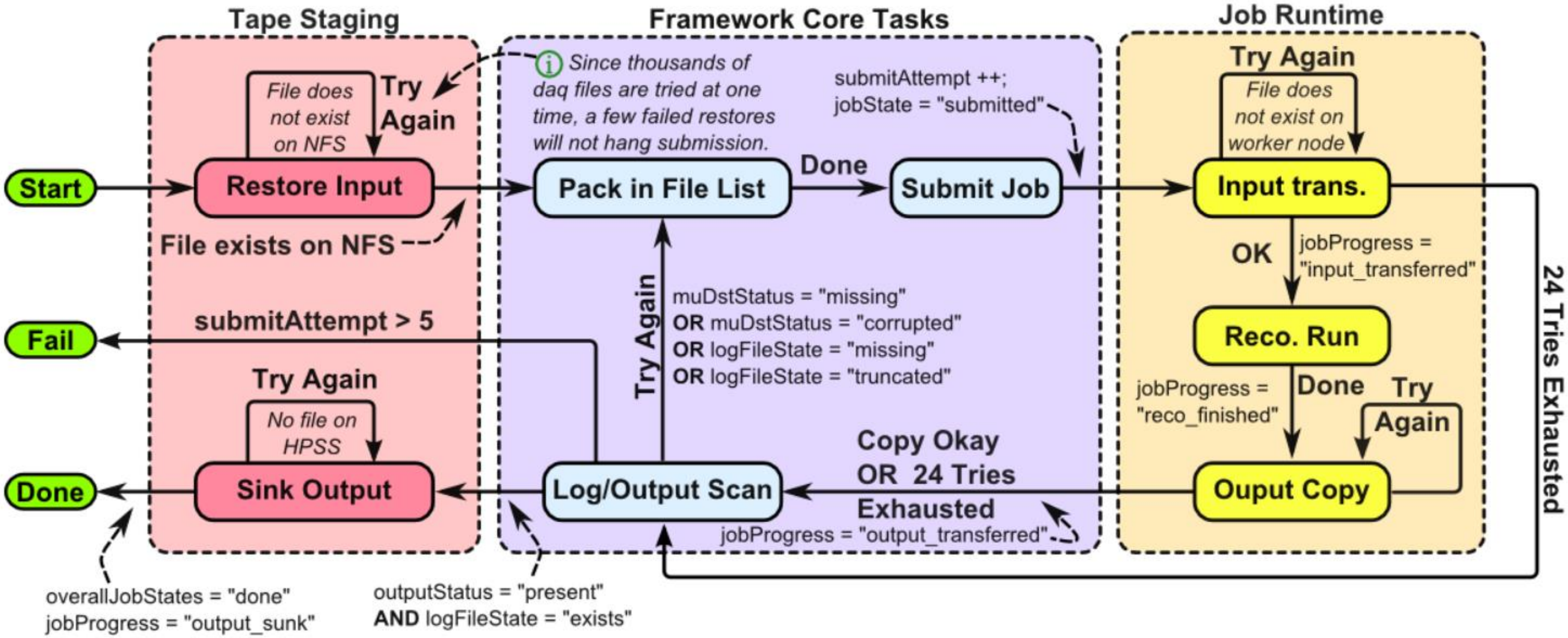
The central database holds the state of the system.

Production System Processes and Steps



STAR GRID Production Finite State Diagram

- Finite state checking exists to verify each stage of the production
- Central DB at BNL holds each job’s state
  - Each job is associated with: One Input file, Batch System ID, Output file(s), Event processing log, Batch System log
  - System gathers information from: log file scans, batch system poll, messages sent from job, file sizes checks on both sides of a transfer

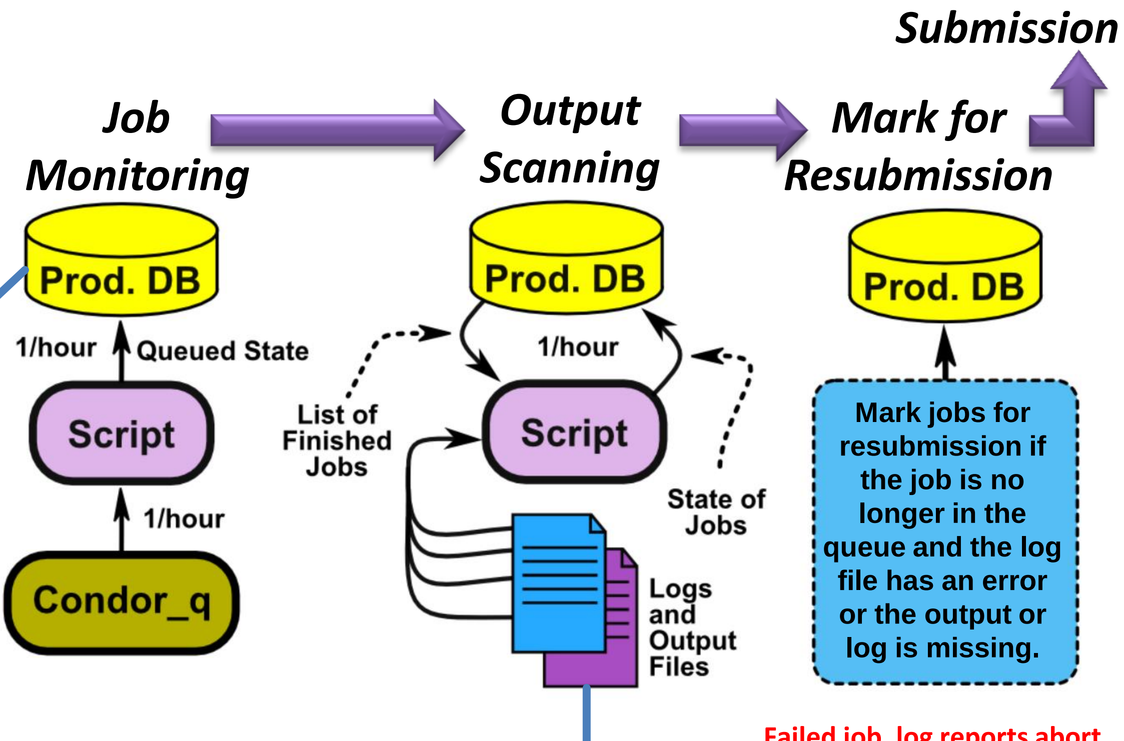

Tape Staging
Framework Core Tasks
Job Runtime
Start
Restore Input
File does not exist on NFS
Try Again
File exists on NFS
Since thousands of daq files are tried at one time, a few failed restores will not hang submission.
Pack in File List
Done
Submit Job
submitAttempt ++;
jobState = "submitted"
Input trans.
Try Again
File does not exist on worker node
OK
jobProgress = "input_transferred"
Reco. Run
jobProgress = "reco_finished"
Done
Try Again
Ouput Copy
24 Tries Exhausted
Fail
submitAttempt > 5
Try Again
muDstStatus = "missing"
OR muDstStatus = "corrupted"
OR logFileState = "missing"
OR logFileState = "truncated"
Try Again
No file on HPSS
Done
Sink Output
Log/Output Scan
Copy Okay
OR 24 Tries
Exhausted
jobProgress = "output_transferred"
overallJobStates = "done"
jobProgress = "output_sunk"
outputStatus = "present"
AND logFileState = "exists"

Some Basic Features of a Production System

And How We Have Implemented Them

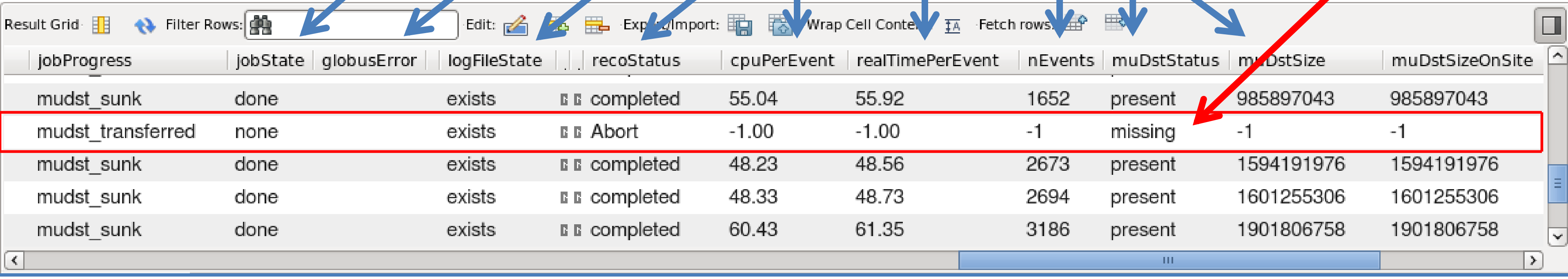
Automated Resubmission of Failed Jobs

- Any number of jobs can be quickly checked and marked for resubmission.
- Finite state model requires output file returned with size check, and log file returned and scanned free of errors else the job is marked for resubmission
- The DB keeps track of the number of times a job is resubmitted to prevent permanent recirculation of non-viable jobs.
  - Limit is four retries



Database View:

Failed job, log reports abort error and no output

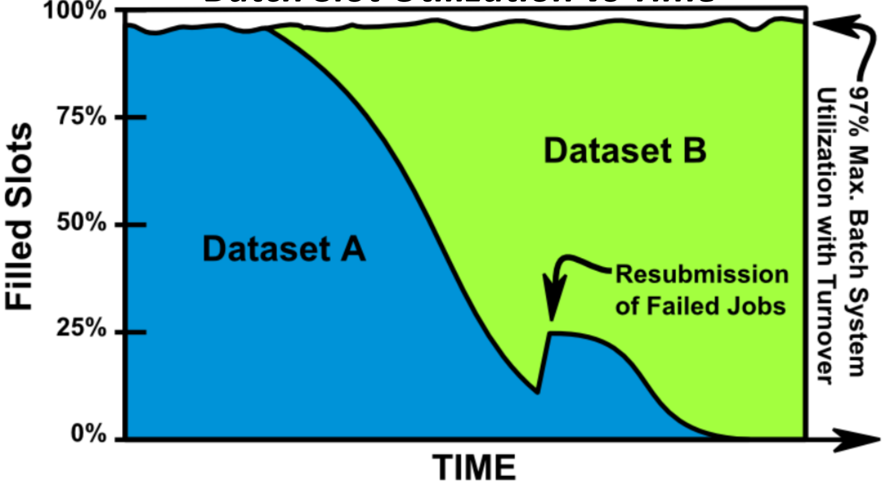


jobProgress	jobState	globusError	logFileState		recoStatus	cpuPerEvent	realTimePerEvent	nEvents	muDstStatus	muDstSize	muDstSizeOnSite
mudst_sunk	done		exists	▯▯	completed	55.04	55.92	1652	present	985897043	985897043
mudst_transferred	none		exists	▯▯	Abort	-1.00	-1.00	-1	missing	-1	-1
mudst_sunk	done		exists	▯▯	completed	48.23	48.56	2673	present	1594191976	1594191976
mudst_sunk	done		exists	▯▯	completed	48.33	48.73	2694	present	1601255306	1601255306
mudst_sunk	done		exists	▯▯	completed	60.43	61.35	3186	present	1901806758	1901806758

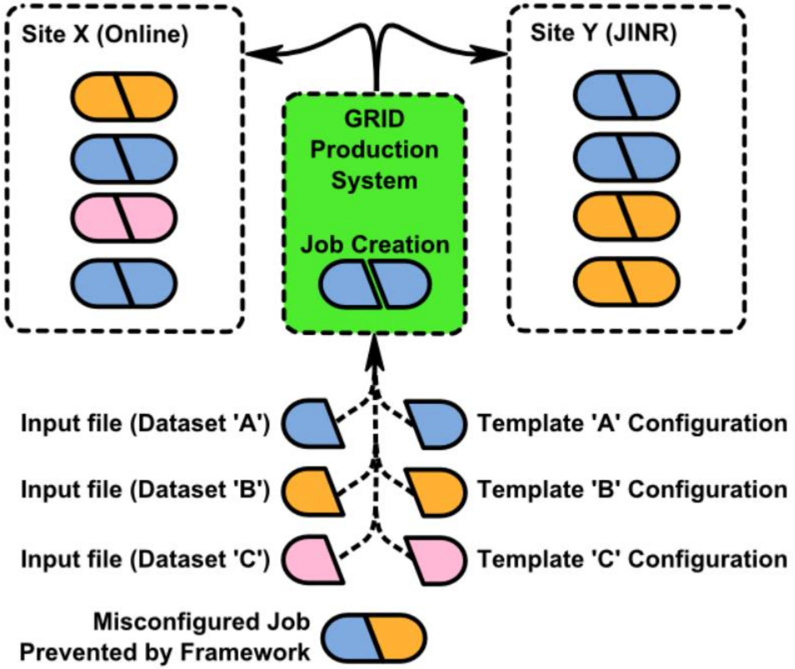
Parallel Submission of Multiple Datasets

- Utilization efficiency is the percent of available slots filled over time. Submission of parallel datasets is a minimum requirement to hold utilization efficiency high.
  - In local production up to 5 datasets are run at once.
- The job consists of two parts
  - Input file
  - Reconstruction parameters (configuration) : Production Tag, Library Version, and Chain Options (Time Stamp, Geometry, Calibration parameters, Selection of tracking algorithms).
- It is the job of the (GRID) production system to correctly associate the correct input file with the correct configuration for that file.
- Site assignment need not be related to dataset type, it could be another parameter such as event count (runtime).
- Misconfigured jobs would be very dangerous as they may return data that appear valid.

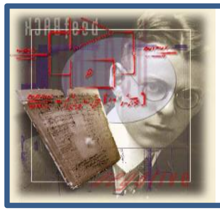
Batch Slot Utilization vs Time



Production System Feeding Three Different Datasets



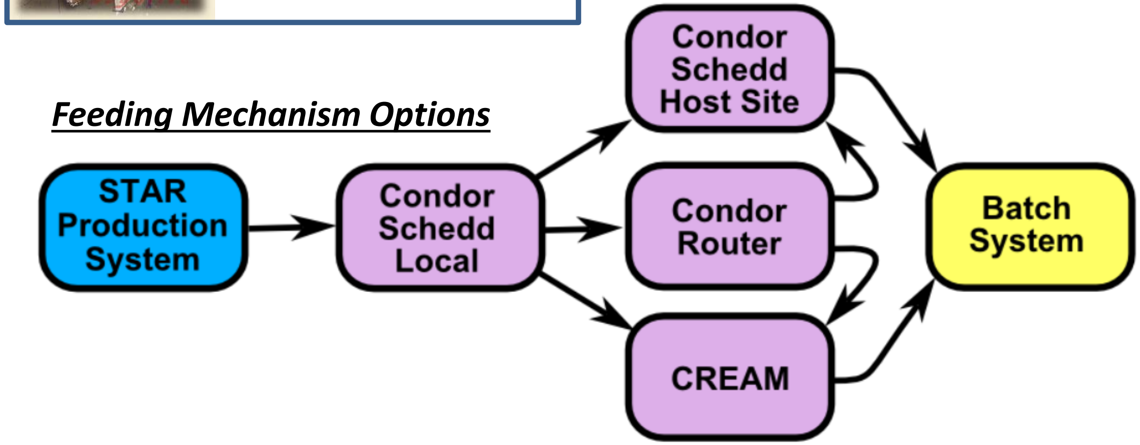
Job Feeding



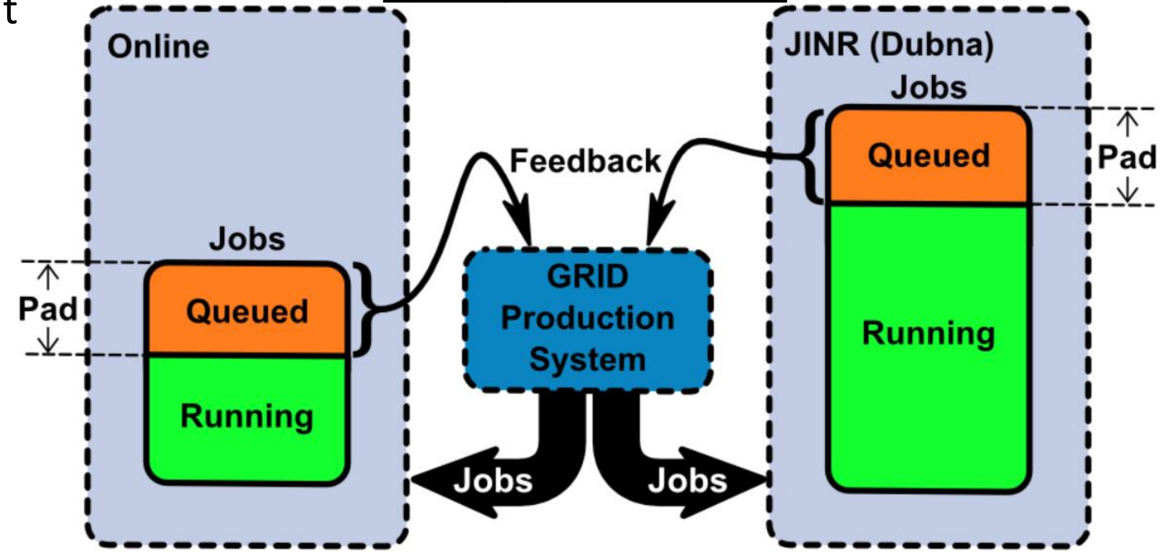
Harold Stephen Black
Inventor of the negative
feedback loop in 1934

- Condor is polled once per hour for idle jobs, if idle jobs per site drops below a set level the system checks if there are more input files to submit in order to keep a **pad** of idle jobs on each running site at all times.
- All viable slots should be filled without any propagation delay from the framework.
- We look at the decay rate of running jobs and tune to insure that in one feeding cycle there are still idle jobs.
- Sometimes no feeding is needed because of other natural limits like limited input buffer size.
- Advanced site pre-assignment of too many jobs can lead to one site finishing all queued jobs and emptying before the other sites.
- Can be done on HTCondor level or before.

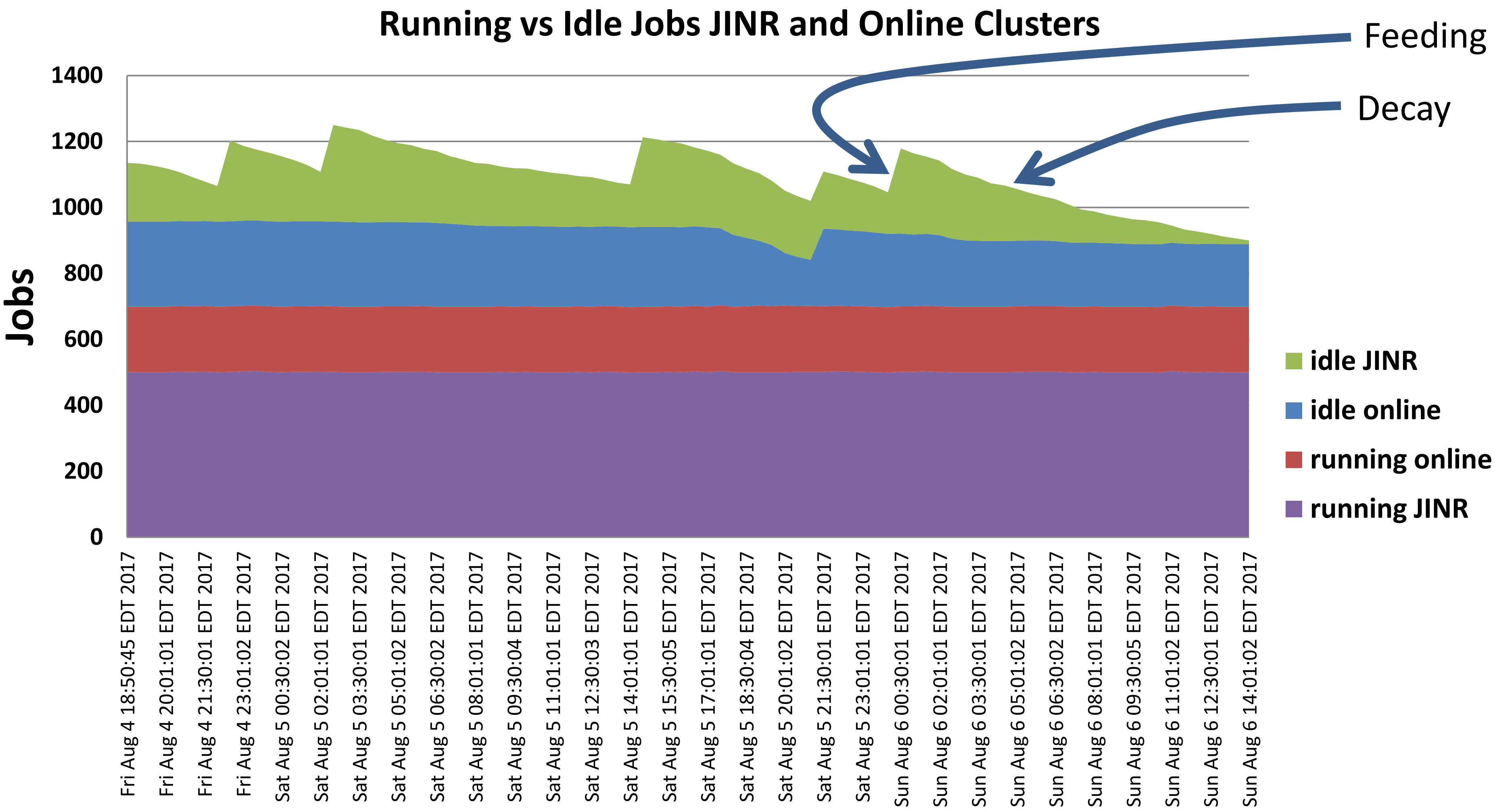
Feeding Mechanism Options



Feeding with Feedback



Job Feeding



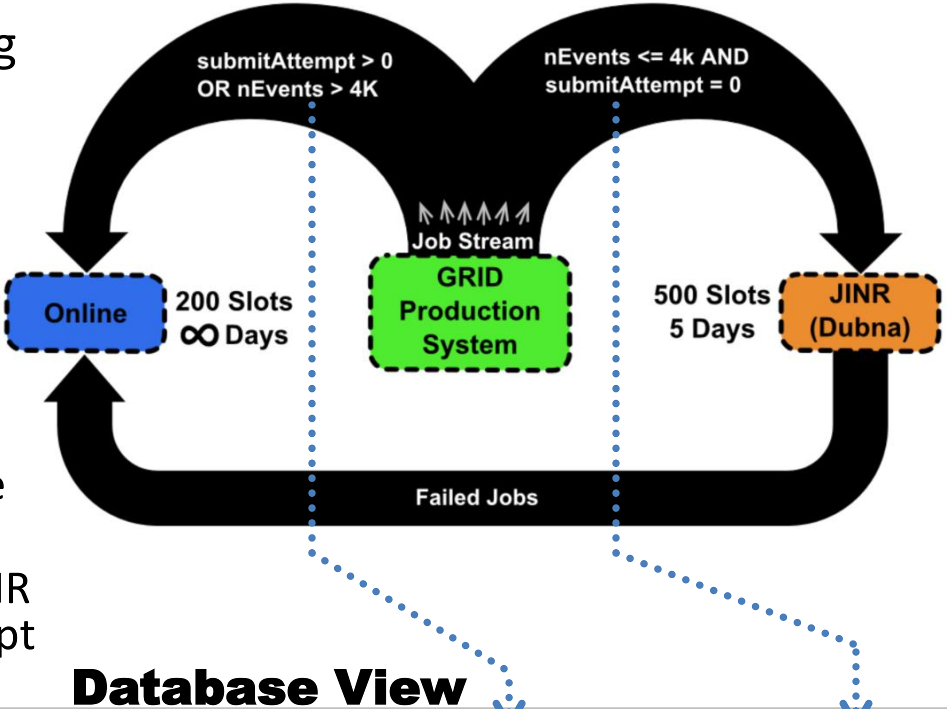
Feeding

Decay

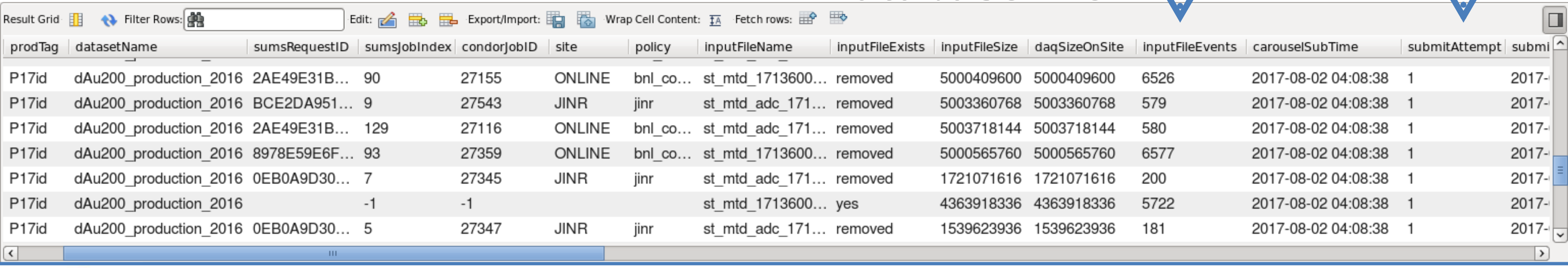
Site Selection Logic

- Allows rules for matching jobs to specific sites, for optimized efficiency
- Can create imbalances of jobs lowering utilization
  - However there is little point to submitting a job to a site where it is unlikely to succeed
- Flexible, can adjust for changing conditions or datasets
- We can make rules, real life examples:
  - Send jobs bigger then 4K events to the Online farm
  - If a job failed in the first attempt at JINR resubmit to Online farm in next attempt

Current Job Flow To Sites



Database View



prodTag	datasetName	sumsRequestID	sumsJobIndex	condorJobID	site	policy	inputFileName	inputFileExists	inputFileSize	daqSizeOnSite	inputFileEvents	carouselSubTime	submitAttempt	submi
P17id	dAu200_production_2016	2AE49E31B...	90	27155	ONLINE	bnl_co...	st_mtd_1713600...	removed	5000409600	5000409600	6526	2017-08-02 04:08:38	1	2017-
P17id	dAu200_production_2016	BCE2DA951...	9	27543	JINR	jinr	st_mtd_adc_171...	removed	5003360768	5003360768	579	2017-08-02 04:08:38	1	2017-
P17id	dAu200_production_2016	2AE49E31B...	129	27116	ONLINE	bnl_co...	st_mtd_adc_171...	removed	5003718144	5003718144	580	2017-08-02 04:08:38	1	2017-
P17id	dAu200_production_2016	8978E59E6F...	93	27359	ONLINE	bnl_co...	st_mtd_1713600...	removed	5000565760	5000565760	6577	2017-08-02 04:08:38	1	2017-
P17id	dAu200_production_2016	0EB0A9D30...	7	27345	JINR	jinr	st_mtd_adc_171...	removed	1721071616	1721071616	200	2017-08-02 04:08:38	1	2017-
P17id	dAu200_production_2016		-1	-1			st_mtd_1713600...	yes	4363918336	4363918336	5722	2017-08-02 04:08:38	1	2017-
P17id	dAu200_production_2016	0EB0A9D30...	5	27347	JINR	jinr	st_mtd_adc_171...	removed	1539623936	1539623936	181	2017-08-02 04:08:38	1	2017-

Conclusion and Statistics

Site	Files	Events	Runtime (Hours)	Dataset Size GB
Online:	2,419	12M	152,392	23,878
JINR:	20,780	138M	534,324	6,488
Total:	23,199	151M	686,716	30,367

- Scavenging additional resources allows for the reconstruction of a few additional small datasets per year.
- 1st Pass Efficiency is **92.8%** and well above other experiments, especially for scavenged, heterogeneous resources
  - slightly below local efficiency (98%) because of added GRID infrastructure overhead
  - Sources of inefficiency: Queue runtime limits, AFS errors (we are investigating CVMFS) , Condor to PBS interface, ‘globus_gsi_callback_module’ copy error 0.505%, Node and batch system testing, farm power outage (mouse got into substation(online))
- System is automated and robust with a robust set of features and finite state workflow:
  - Job tracking, feeding, failure detection and resubmission, site selection logic
  - Reuse of lots of existing STAR software but still dependent on HTCondor and Globus-URL-Copy

Questions ?

nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE

First observation of fluid vortices formed by heavy-ion collisions

PAGES 34 & 62

SUBATOMIC SWIRLS

CLIMATE CHANGE

PARIS AGREEMENT

Time for nations to match words with deeds

PAGE 25

BOOKS

SUMMER SELECTION

Recommended reading for the holiday season

PAGE 28

STEM CELLS

YOUTHFUL SECRETS

How the hypothalamus helps to control the ageing process

PAGE 52

NATURE.COM/NATURE

3 August 2017

Vol. 548, No. 7665