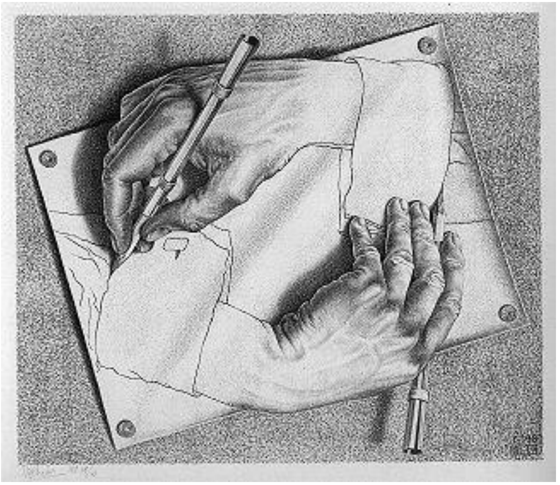
Progress report on STAR's Expansion to JINR via GRID

The 7th International Conference "Distributed Computing and Grid-technologies in Science and Education" (GRID 2016)

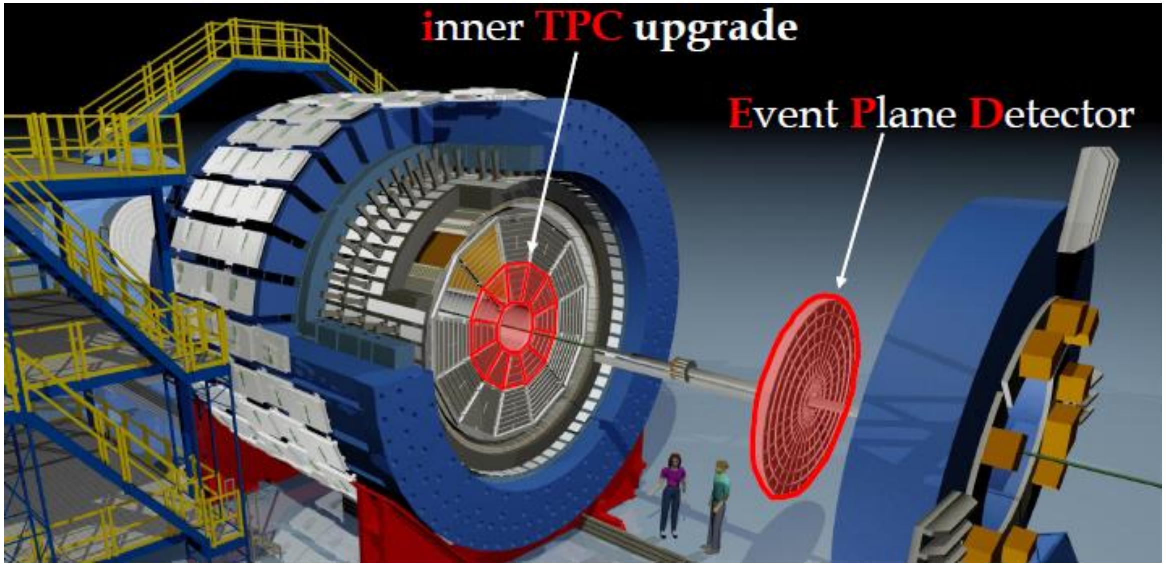
BNL STAR	JINR STAR	JINR LIT
Levente Hajdu	Yuri Panebrattsev	Vladimir Korenkov
Jerome Lauret	Geydar Agakishiev	Valery Mitsin
Lidia Didenko		
Wayne J. Betts		

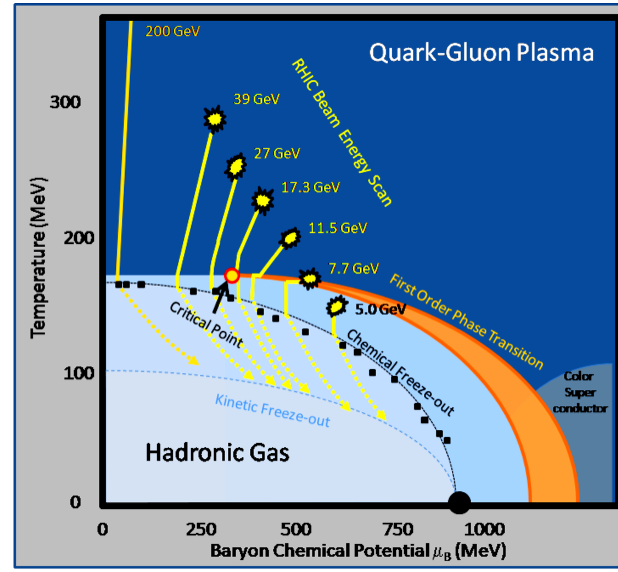
Outline



- About STAR
  - Key Physics Drivers
- Motivations
- Grid Architecture
  - Grid Production System
  - Previous successful productions
- First impressions, findings and refinements at JINR
  - Efficiency Decomposed
- Solutions and Perspective for a Path Forward
  - Proxy Solutions
  - PBS (input and output staging Solutions)

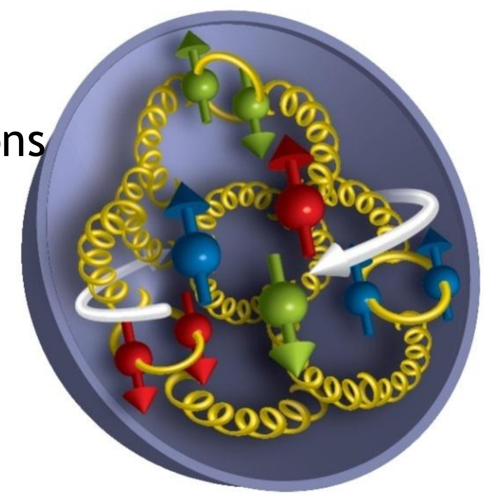
Key Physics Drivers





An experiment located @ RHIC in the USA/NY/LI, dedicated to Heavy Ion and Spin studies & properties of the QGP/QCD

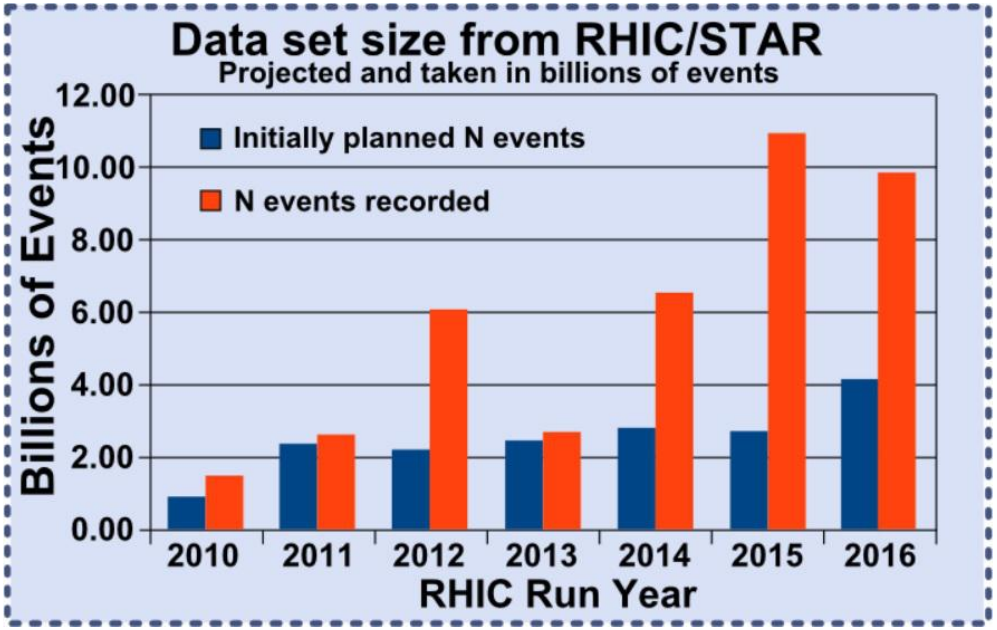
Unique machine able to collide from Heavy Ions and polarized protons



- Heavy-flavor and quarkonia measurement
- Jet measurements
- Chiral magnetic effect, chiral magnetic wave and chiral vortical effect
- 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

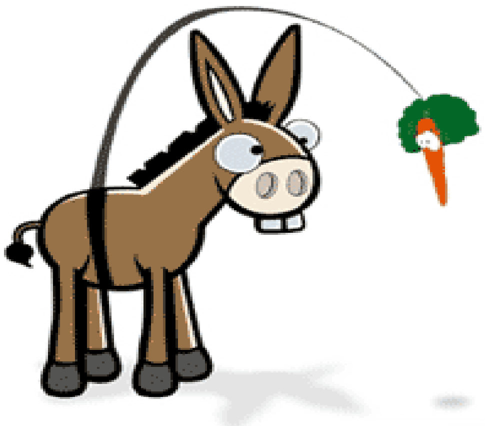
Motivation

- Exciting versatile physics program
- In our 16th year of data taking
- Accumulated 25 petabytes of raw data
  - Projected to reach 40 petabytes
  - CPU needs to double by 2020
- Huge dataset challenges - we seek additional resources to speed up scientific discoveries
  - 13,500 slots are used for data production at BNL
    - This **ONLY** allows for 1.2-1.4 passes of data reconstruction
    - In contrast, typical HEP experiments have > 5 passes
  - Additional resources (+20% minimum) must be found to cover simulations and additional processing passes
    - This translates to ~ 118 Million CPU hours per year at the Tier-0 center and at least 24 Million CPU hours per year from other resources, such as JINR



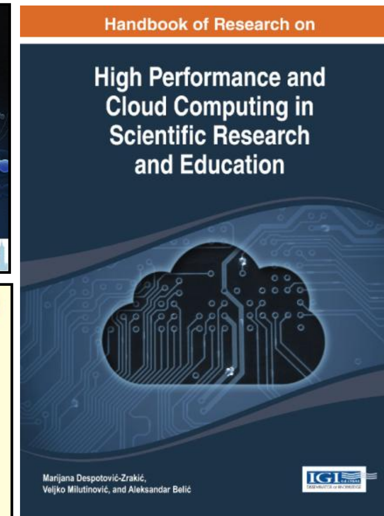
Motivation…

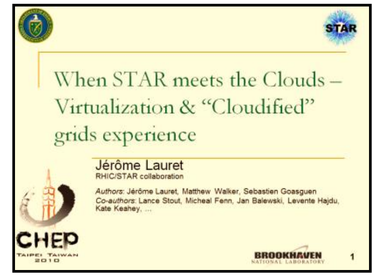
- Growth will unlikely come from the Tier-0 so there is an incentive to find resources "outside"
  - Data production success demonstrated at KISTI …
  - Expertise and software exists but all host sites have differences in:
    - Middleware Stack (grid software)
    - Bandwidth speeds
    - Batch system
    - Resource limits
  - Adjustments to our existing framework, adapting to the new facility "specifics" and tuning are expected to allow smoother, more efficient running.

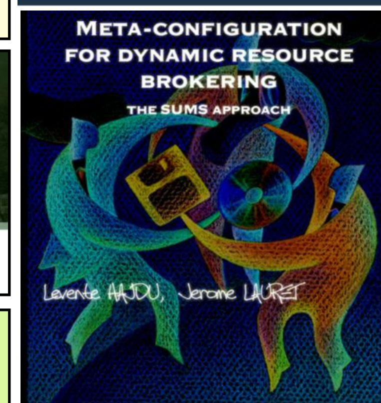


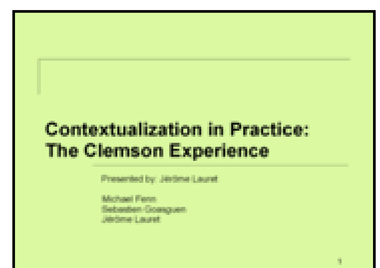
Working on Grids and Clouds for 16 Years

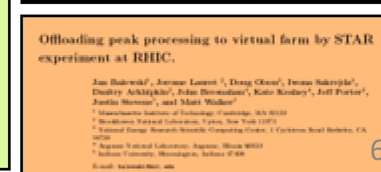
- STAR has a long experience testing and offloading productions to a diversity of platforms (Cloud/Grid) and resources (Amazon, Universities, National Labs, ...) which has made us face different challenges and provided useful lessons learned.
  - "**High Performance and Cloud Computing in Scientific Research and Education**" Chapter 13, IGI Global, Levente Hajdu, Jérôme Lauret, Radomir A. Mihajlovic ISBN13: 9781466657847, ISBN10: 1466657847, EISBN13: 9781466657854.
  - "**Offloading peak processing to virtual farm by STAR experiment at RHIC**", Proc. of the 14th International Workshop on Advanced Computing and Analysis Techniques in Physics Research (ACAT2011), Uxbridge, West London, United Kingdom, September 5-9, 2011, J. Phys. Conf. Ser. 368 (2012) 01211.
  - "**When STAR meets the Clouds: Virtualization & Cloud computing experiences**", Proc. of the 18th International Conference on Computing n High Energy and Nuclear Physics (CHEP2010), Taipei, Taiwan, October 18-22, 2010, J. Phys. Conf. Ser. 331 (2011) 062016.
  - "**Contextualization in practice: The Clemson experience**", Proc. of the 13th International Workshop on Advanced Computing and Analysis Techniques in Physics Research (ACAT2010), Jaipur, India, February 22-27, 2010, Pos ACAT2010 (2010) 027.
    "**Integrating Xgrid into the HENP distributed computing model**", Proc. of the International Conference on Computing in High Energy and Nuclear Physics (CHEP07), Victoria, British Columbia, Canada, September 2-7, 2007, J. Phys. Conf. Ser. 119 (2008) 072018.
  - "**Automated Finite State Workflow for Distributed Data Production**", ACAT 2016 conference proceedings, pending publication".











Efficiency Defined

- **Farm Utilization Efficiency** – The percentage of the farm utilized by successful jobs over all available slots integrated over all retry attempts

$$\boxed{i = try}$$

$$Efficiency_{Farm} = \sum_{i=1 \dots n} \frac{\left(\frac{n\ successful\ jobs}{total\ slots}\right)_i}{try_i}$$

- **CPU Efficiency** – The percentage of a CPU a job (a process) utilizes

$$Efficiency_{CPU} = \left(\frac{CPUTime}{RealTime}\right)$$

- **Compound Efficiency** -

$$= \sqrt{(\textbf{CPU Efficiency}) * (\textbf{Farm Utilization Efficiency})}$$

Previous Production Sites

- STAR uses Grid resources at PDSF/NERSC (Berkeley) and in the past used clusters at Clemson University, Amazon Cloud, Fermilab, University of São Paulo, and others – most of those were used for simulation instead of data production (reconstruction of real data)
- The first site STAR used to run ***data production*** on GRID was KISTI (The Korea Institute of Science and Technology Information)
- STAR saturated 1026 job slots for 9 months processing 105,631 files, generating 213.7 TB of output and consuming 6,179,544 CPU hours
  - **99%** of jobs returned log and output files within 7 attempts
  - **Farm Utilization Efficiency** $Efficiency_{Farm} =$ **97(±2)%**

Automated (Re-)Submission Attempt(s) Per Job at KIST		
Number Of Jobs	**Percent (%)**	**Job Submissions**
102,043	96.6023	1
2,895	2.7406	2
493	0.4667	3
103	0.0975	4
83	0.0785	5
6	0.0056	6
3	0.0028	7
6	0.0056	*Not Runnable

Components of the Grid Production System

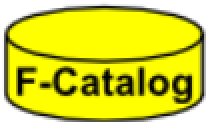
HPSS = Tape silo system at BNL

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

SUMS (STAR Unified Meta Scheduler) = SUMS provides a unified interface for submitting jobs to sites and wrapping of the input file and user executable into a job. Feeding can also be turned on to limit the number of jobs submitted at one time.

HTCondor with Globus = provides authentication of users between sites and a mechanism to interface with the sites local batch system.

Production Database = Database for holding the state of each job

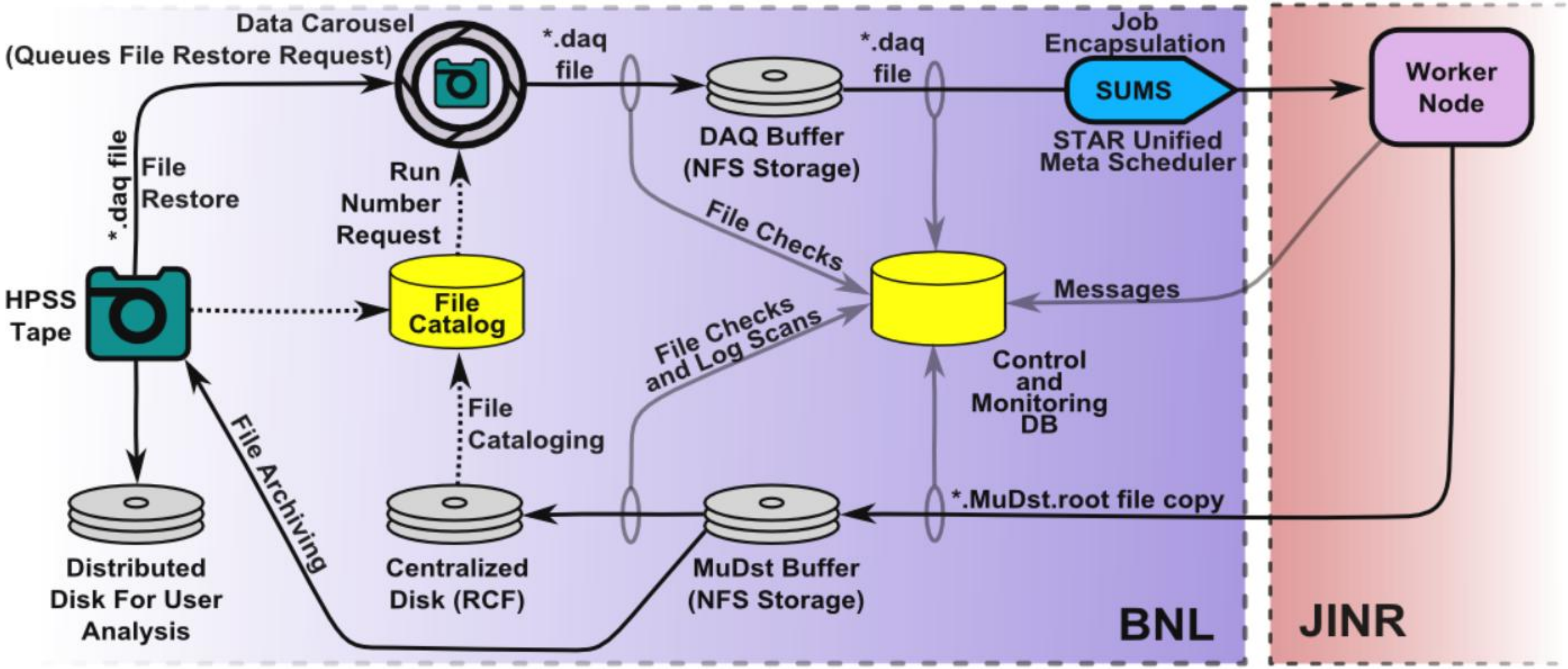


STAR File Catalog = includes PFN, LFN and MetaData

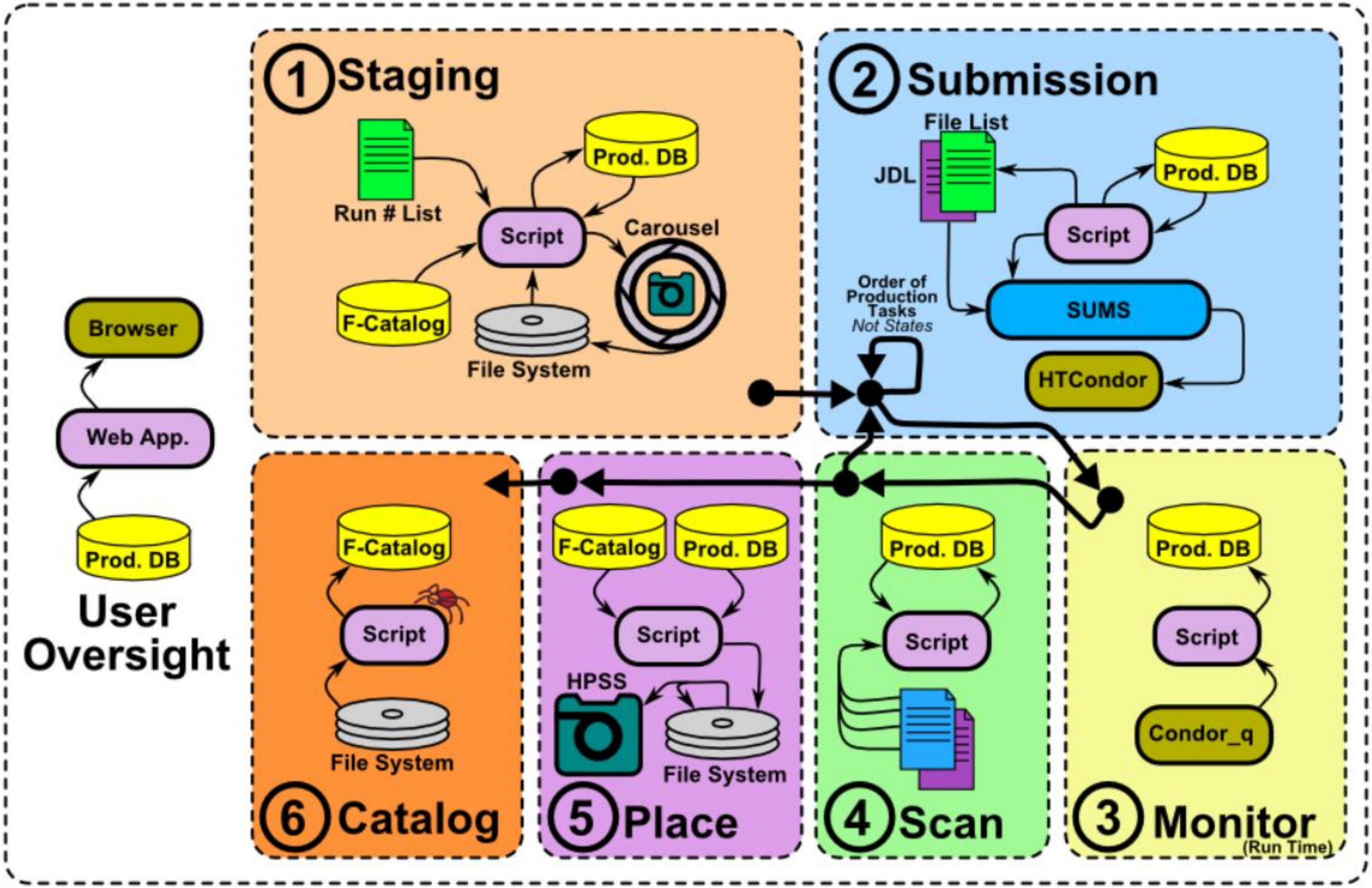
***.Daq Files** = raw detector input files for reconstruction

***.MuDst Files** = reconstruction output files

Current Grid Production Dataflow

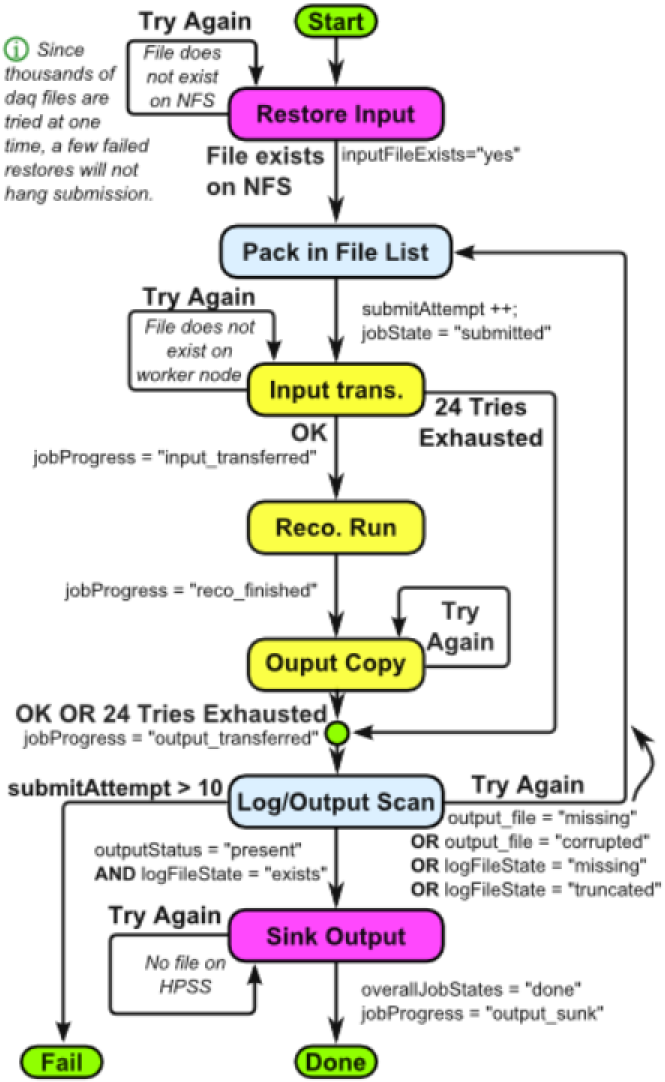


Detailed Stages of Production

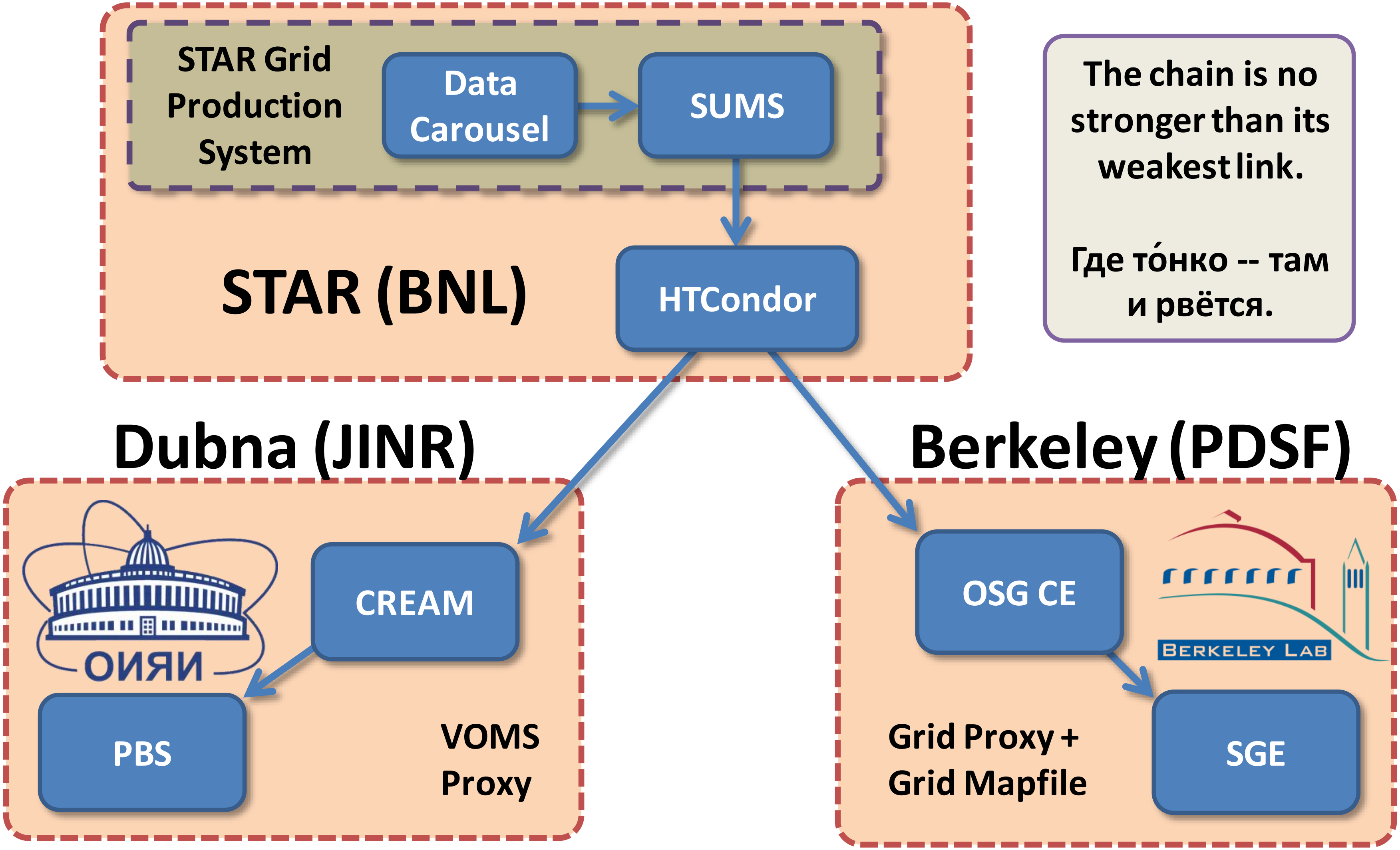


Grid Production Framework State Diagram

- 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
    - Produced Data File
- System gathers information from:
  - File sizes are checked after each transfer
  - Batch system is polled every hour to get the current state of each job
  - Jobs send messages at each stage:
    - job start up (copy input starts), input transfer done (reconstruction starts), reconstruction done (output transfer starts), output transfer done
  - Log files (batch and reconstruction) are scanned for error states



Grid Software Stack Architecture



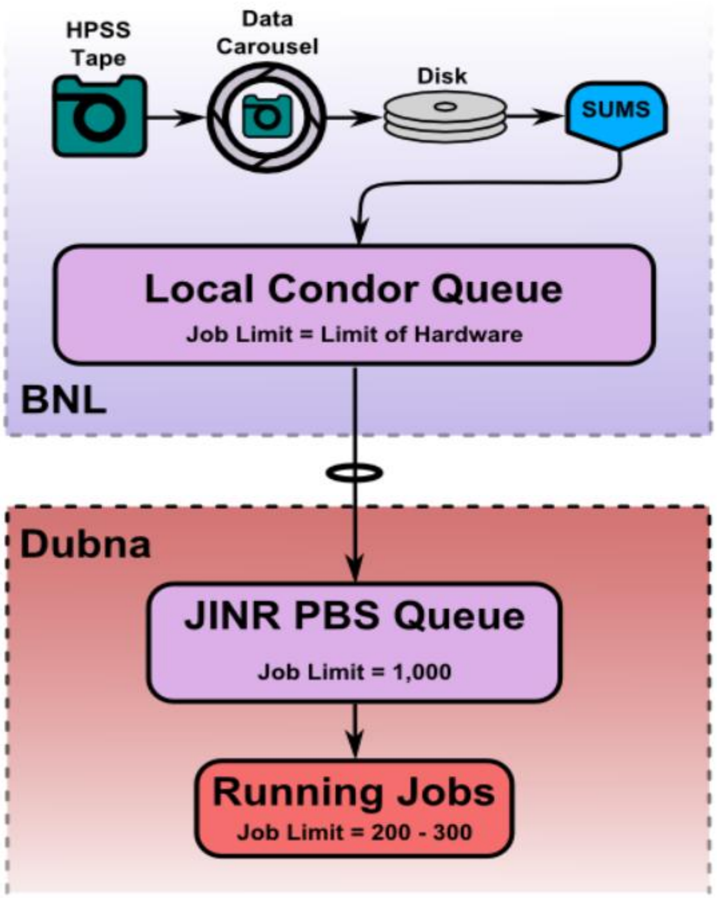
The chain is no stronger than its weakest link.

Где то́нко -- там и рвётся.

First Steps @ Dubna

Initial Tuning

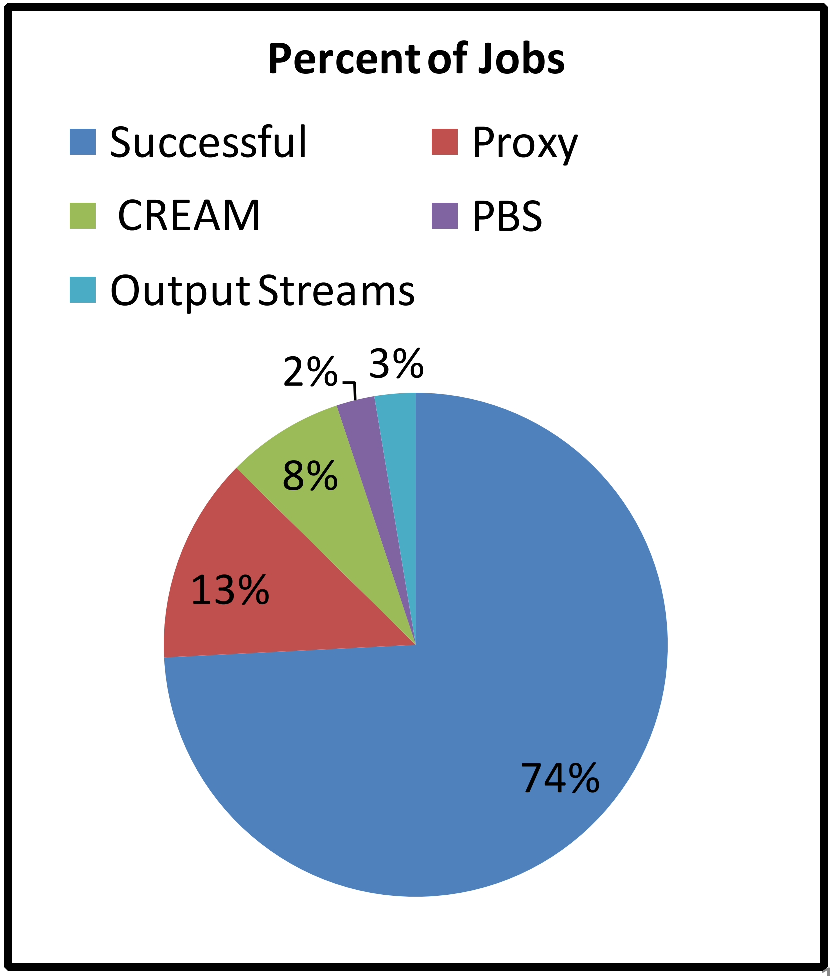
- HTCondor submission to CREAM (Computing Resource Execution And Management) works to first order
  - Forwarding of renewed VOMS proxies is not robust
  - BNL VOMS proxies renewed from GRID proxies don't seem to authenticate
  - BNL VOMS administrators agreed to change limit from 24 hours to 3 days increasing job run time
- AFS allows connection to our software. CPU to wall-time ratio was low likely due to cache thrashing ($Efficiency_{CPU}$=52%)
  - AFS cache raised from 600 MB to 7 GB, improved CPU to wall-time ratio from 52% → **>90%**"
- PBS running job limited to 58 wall hours, 24 CPU hours
  - Increased to one week
- Queue limited to 500 queued jobs
  - Max number of queued jobs raised from 500 to 1k
  - Adjusted Condor configuration parameter to control the number of jobs pushed over at any one time.
- Max Running jobs limited to 100
  - Running jobs raised from 100 max to 200-300 to test scalability
- Thanks to Yuri Panebrattsev and Geydar Agakishiev for help from JINR.
- **Initial tuning on both sides (BNL + JINR) has achieved big improvements and proven viability.**



Current Efficiency Decomposed

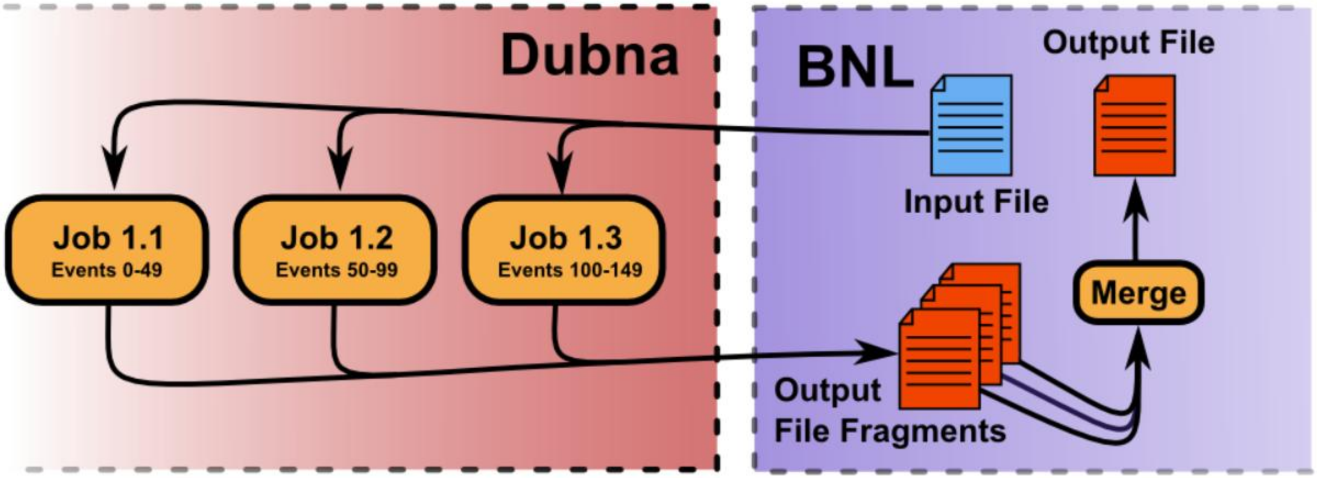
- **Efficiency** in this context is defined as the number of jobs returning output and log files within one submission attempt
- **Proxy** – All jobs running over 3 days are evicted because the new proxy on the submitting end is not forwarded
  - Affected by job runtime which is affected by the number of events in a file and their complexity
- **CREAM Infrastructure Errors** – typically involves CREAM losing track of the job (*"reason=-999"*)
- **PBS** – the batch system kills the job
  - Lack of CPU usage when the job is transferring input and output files
  - Aggravated by bottlenecking (many jobs starting at the same time) and network transients
- **Log Transfers** – The batch system has lost track of where it was writing the output streams from the job's processes

- **Resolution of the top two errors could push us to a comfortable ~95% efficiency**



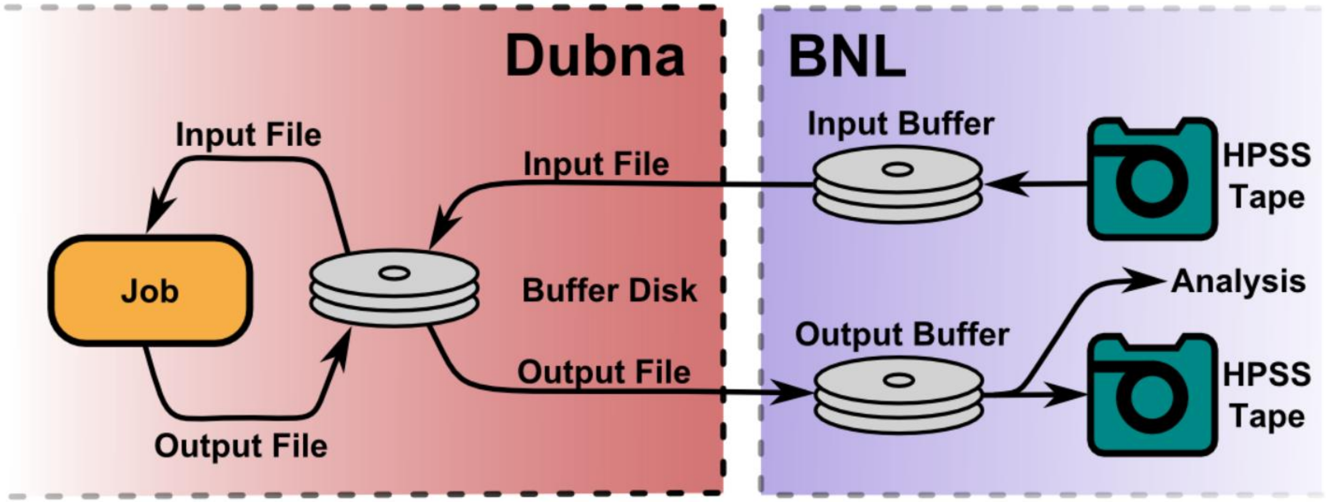
Possible Solution - Proxies

- Test if the existing infrastructure will work with a grid proxy vs. a VOMS proxy
- Install a GLOBUS gatekeeper that will work with a GRID mapfile instead of the CREAM infrastructure (this should also resolve the CREAM errors)
- Explore MyProxy delegation as a possible solution, MyProxy claims to renew the VOMS proxy extension from GRID proxies
  - VOMS extensions automatically renewed from grid proxies have so far failed to authenticate
- Reconstruct files at JINR with a limited number of events
- Breakdown the input files into multiple jobs per input



Possible Changes - I/O Handling

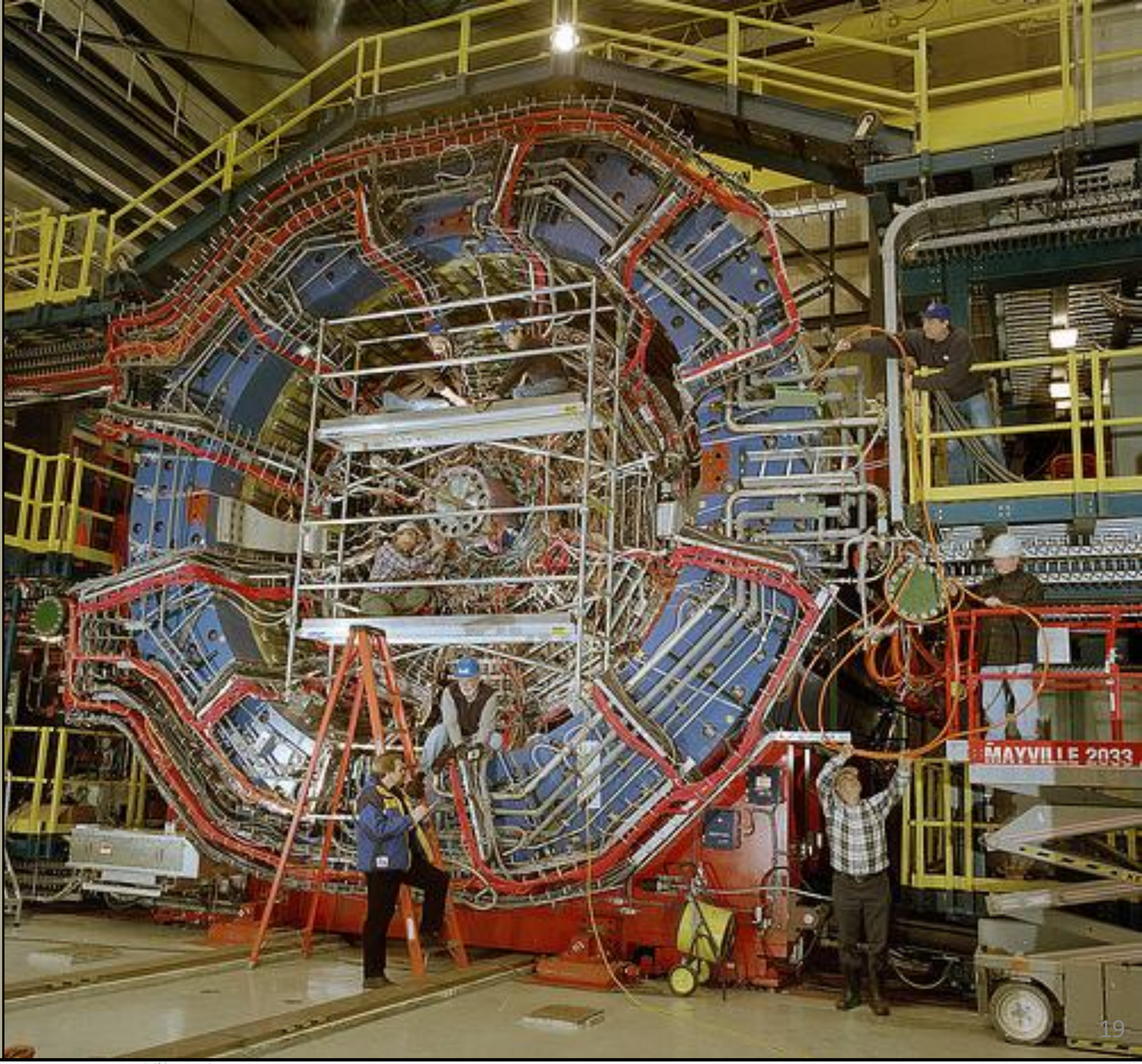
- Do staging of input and output outside of the job's runtime.
  - This solution would require a large (~50 TB (input + output)) local buffer on the host site
  - It can overflow if network interruptions occur
  - It requires changes to our production systems workflow
- Investigate HTCondor's delegated I/O staging
- We have already mitigated jobs getting killed for failure to consume CPU during input copy by adding a 3 minute delay in between each job submission.
  - This reduced bottlenecking during the copy and the corresponding job deaths to 2%



Conclusions

- Cooperative efforts in tackling initial problems & tuning rapidly brought us to a 74% workflow efficiency
- Further tests showed remaining obstacles for which we are investigating solutions
  - Long live proxies + Hand-shake between OSG/CREAM , if resolved, would allow an additional +21% recovery
  - Would bring the production to a stunning 95% efficiency
- Modulo finding solutions for the identified issues, Dubna resources seem to be a good candidate for STAR CPU demands

Questions ?



BROOKHAVEN NATIONAL LABORATORY

STAR

U.S. DEPARTMENT OF ENERGY

Office of Science