



"Ask PanDA" AI project

Paul Nilsson (BNL)

July 2, 2025

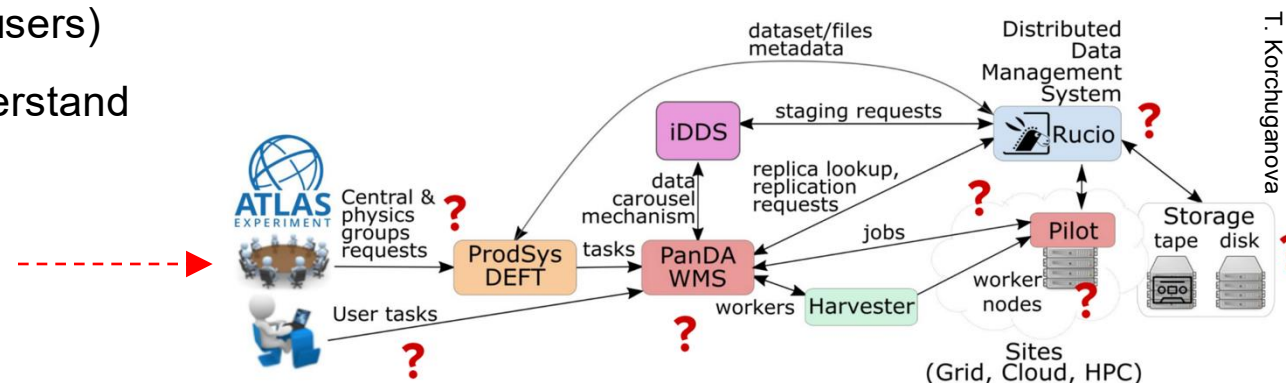
NPPS meeting



Explaining Errors



- The plan is to improve the situation with hard to understand errors in the system
 - Users and operators often ask why grid jobs failed
- Failed jobs are presented with error code + diagnostics in job monitor
 - Error messages are often static/generic
 - Can be difficult to understand (especially for new users)
 - Job logs are available, but - again - difficult to understand
- Many components = Many sources of errors
 - Most of the time, cannot fix the errors directly
 - But we can identify and explain the errors
- **Idea:** Use LLMs (and possibly ML techniques) on multiple levels to alleviate the situation



Approach: AI vs Errors

- **Error categorization**
 - Plan: Start with sorting error codes into different categories
 - E.g. many error types are site related and not related to the payload, but many users don't know that
 - Also, many user jobs are retried in vain and will never finish (actual problem is not getting fixed)
 - If a certain error was classified properly, the number of retries could be optimized
 - Use LLM to do error categorization by error code analysis (there are too many error codes to do it manually)
- **Improved and explained error messages**
 - There are many different components, several hundred error codes reflecting different failure scenarios
 - Some error codes are in components/tools outside of the PanDA ecosystem
 - Transform, bash and setup scripts, apptainer/singularity, ..
 - Use LLM to analyze source code (all components) for each error code, ask it to provide better diagnostics
- **Log analyzer**
 - When a job has failed, explain to user why exactly that particular job failed
 - Use LLM to explain failure as shown in log file
- **Chatbot**
 - Allow user to ask questions about the system (and possibly also about job failures, etc)
 - Use LLM aided by RAG* using our own documentation

* RAG = Retrieval-Augmented Generation

Pilot Error Categories

- Currently focusing on pilot error codes
 - Other components are more specific, and each have fewer error codes
 - Categorization might not be necessary
- All pilot error codes (145) have been sorted into six categories to be used in the error classification and analysis
- ChatGPT was used for the sorting which worked rather well

Category	Explanation
1. File and Storage Issues	Errors related to missing files, file transfers, storage space, file integrity, and checksum mismatches.
2. Execution and Payload Failures	Errors related to payload setup/execution, transform/download failures, timeouts during execution, etc.
3. Network and Communication Errors	Issues related to connectivity, authentication, and server availability.
4. Job Termination and Kill Signals	Errors caused by jobs being manually or automatically killed.
5. Software and Environment Issues	Errors due to missing software, misconfigurations, or resource constraints.
6. Internal and Unknown Errors	Catch-all category for unclassified failures, exceptions, and debugging-related issues.

Error categorization

Sorted Pilot Error Codes

Category 1: File and Storage Issues

Error Code	Error Acronym	Error Message
1098	NOLOCALSPACE	Not enough local space
1099	STAGEINFAILED	Failed to stage-in file
1100	REPLICANOTFOUND	Replica not found
1103	NOSUCHFILE	No such file or directory
1104	USERDIRTOOLARGE	User work directory too large
1106	STDOUTTOOBIG	Payload log or stdout file too big
1115	NFSSQLITE	NFS SQLite locking problems
1124	OUTPUTFILETOOLARGE	Output file too large
1133	NOSTORAGE	Fetching default storage failed: no activity related storage defined
1137	STAGEOUTFAILED	Failed to stage-out file
1141	PUTMD5MISMATCH	md5sum mismatch on output file
1143	CHMODTRF	Failed to chmod transform
1145	GETMD5MISMATCH	md5sum mismatch on input file
1151	STAGEINTIMEOUT	File transfer timed out during stage-in
1152	STAGEOUTTIMEOUT	File transfer timed out during stage-out
1165	MISSINGOUTPUTFILE	Local output file is missing
1168	SIZETOOLARGE	Total file size too large
1171	GETADMISMATCH	adler32 mismatch on input file
1172	PUTADMISMATCH	adler32 mismatch on output file
1190	LFNTOOLONG	LFN too long (exceeding limit of 255 characters)
1191	ZEROFILESIZE	File size cannot be zero
1199	MKDIR	Failed to create directory
1221	FILEEXISTS	File already exists
1242	CHKSUMNOTSUP	Query checksum is not supported
1303	FILEHANDLINGFAILURE	Failed during file handling
1313	NOSTORAGEPROTOCOL	No protocol defined for storage endpoint
1320	LOGFILECREATIONFAILURE	Failed during creation of log file
1326	NOREPLICAS	No matching replicas were found in list_replicas() output
1331	MISSINGINPUTFILE	Input file is missing in storage element
1332	BLACKHOLE	Black hole detected in file system (consult Pilot log)
1333	NOREMOTESPACE	No space left on device
1350	EMPTYOUTPUTFILE	Empty output file detected
1366	CHECKSUMCALCFAILURE	Failure during checksum calculation
1371	BADOUTPUTFILENAME	Output file name contains illegal characters
1376	LOGCREATIONTIMEOUT	Log file creation timed out

Category 2: Execution and Payload Failures

Error Code	Error Acronym	Error Message
1110	SETUPFAILURE	Failed during payload setup
1149	TRFDOWNLOADFAILURE	Transform could not be downloaded
1187	NOPAYLOADMETADATA	Payload metadata does not exist
1212	PAYLOADOUTOFMEMORY	Payload ran out of memory
1213	REACHEDMAXTIME	Reached batch system time limit
1223	BADALLOC	Transform failed due to bad_alloc
1235	PAYLOADEXCEEDMAXMEM	Payload exceeded maximum allowed memory
1240	MESSAGEHANDLINGFAILURE	Failed to handle message from payload
1305	PAYLOADEXECUTIONFAILURE	Failed to execute payload
1310	PAYLOADEXECUTIONEXCEPTION	Exception caught during payload execution
1317	EXCEEDEDMAXWAITTIME	Exceeded maximum waiting time
1334	SETUPFATAL	Setup failed with a fatal exception (consult Payload log)
1335	MISSINGUSERCODE	User code not available on PanDA server (resubmit task with --useNewCode)
1336	JOBALREADYRUNNING	Job is already running elsewhere
1337	BADMEMORYMONITORJSON	Memory monitor produced bad output
1346	TRANSFORMNOTFOUND	Transform not found
1349	UNRECOGNIZEDTRFARGUMENTS	Unrecognized transform arguments
1351	UNRECOGNIZEDTRFSTDERR	Unrecognized fatal error in transform stderr
1356	PREPROCESSFAILURE	Pre-process command failed
1357	POSTPROCESSFAILURE	Post-process command failed
1367	COMMANDTIMEDOUT	Command timed out

Category 3: Network and Communication Errors

Error Code	Error Acronym	Error Message
1116	QUEUEDATA	Pilot could not download queuedata
1180	GETGLOBUSSYSERR	Globus system error during stage-in
1181	PUTGLOBUSSYSERR	Globus system error during stage-out
1246	NOUSERARTBALL	User tarball could not be downloaded from PanDA server
1316	RUCIOSERVICEUNAVAILABLE	Rucio: Service unavailable
1318	COMMUNICATIONFAILURE	Failed to communicate with server
1321	RUCIOLOCATIONFAILED	Failed to get client location for Rucio
1322	RUCIOLISTREPLICASFAILED	Failed to get replicas from Rucio
1324	SERVICEUNAVAILABLE	Service not available at the moment
1327	UNREACHABLENETWORK	Unable to stage-in file since network is unreachable
1330	JSONRETRIEVALTIMEOUT	JSON retrieval timed out
1361	REMOTEFILECOULDNOTBEOPENED	Remote file could not be opened
1362	XRDCEPERROR	Xrdcp was unable to open file
1368	REMOTEFILEOPENTIMEDOUT	Remote file open timed out
1369	FRONTIER	Frontier error
1374	REMOTEFILEDICTDOESNOTEXIST	Remote file open dictionary does not exist
1378	LSETUPTIMEDOUT	Lsetup command timed out during remote file open

Category 4: Job Termination and Kill Signals

Error Code	Error Acronym	Error Message
1144	PANDAKILL	This job was killed by panda server
1150	LOOPINGJOB	Looping job killed by pilot
1200	KILLSIGNAL	Job terminated by unknown kill signal
1201	SIGTERM	Job killed by signal: SIGTERM
1202	SIGQUIT	Job killed by signal: SIGQUIT
1203	SIGSEGV	Job killed by signal: SIGSEGV
1204	SIGXCPU	Job killed by signal: SIGXCPU
1205	USERKILL	Job killed by user
1206	SIGBUS	Job killed by signal: SIGBUS
1207	SIGUSR1	Job killed by signal: SIGUSR1
1208	SIGINT	Job killed by signal: SIGINT
1328	PAYLOADSIGSEGV	SIGSEGV: Invalid memory reference or a segmentation fault
1363	KILLPAYLOAD	Raythena has decided to kill payload
1379	PREEMPTION	Job was preempted

Category 5: Software and Environment Issues

Error Code	Error Acronym	Error Message
1163	NOPROXY	Grid proxy not valid
1177	NOVOMSPROXY	Voms proxy not valid
1186	NOSOFTWAREDIR	Software directory does not exist
1211	MISSINGINSTALLATION	Missing installation
1306	SINGULARITYGENERALFAILURE	Singularity(Apptainer): general failure
1307	SINGULARITYNOLOOPDEVICES	Singularity(Apptainer): No more available loop devices
1308	SINGULARITYBINDPOINTFAILURE	Singularity(Apptainer): Not mounting requested bind point
1309	SINGULARITYIMAGEMOUNTFAILURE	Singularity(Apptainer): Failed to mount image
1325	SINGULARITYNOTINSTALLED	Singularity: not installed
1338	STAGEINAUTHENTICATIONFAILURE	Authentication failure during stage-in
1340	SINGULARITYNEWUSERNAMESPACE	Singularity(Apptainer): Failed invoking the NEWUSER namespace runtime
1341	BADQUEUECONFIGURATION	Bad queue configuration detected
1342	MIDDLEWAREIMPORTFAILURE	Failed to import middleware (consult Pilot log)
1345	SINGULARITYFAILEDUSERNAMESPACE	Singularity(Apptainer): Failed to create user namespace
1347	UNSUPPORTEDSLSOS	Unsupported SL6 OS
1348	SINGULARITYRESOURCEUNAVAILABLE	Singularity(Apptainer): Resource temporarily unavailable
1360	IMAGENOTFOUND	Image not found
1364	MISSINGCREDENTIALS	Unable to locate credentials for S3 transfer
1365	NOCTYPES	Python module ctypes not available on worker node
1370	VOMSPROXYABOUTTOEXPIRE	VOMS proxy is about to expire
1372	APTAINERNOTINSTALLED	Apptainer: not installed
1373	CERTIFICATEHASEXPIRED	Certificate has expired
1377	CVMFSSNOTALIVE	CVMFS is not responding
1380	ARCProxyFAILURE	General arcproxy failure
1381	ARCProxyLIBFAILURE	Arcproxy failure while loading shared libraries
1382	PROXYTOOSHORT	Proxy is too short
1383	STAGEOUTAUTHENTICATIONFAILURE	Authentication failure during stage-out

Category 6: Internal and Unknown Errors

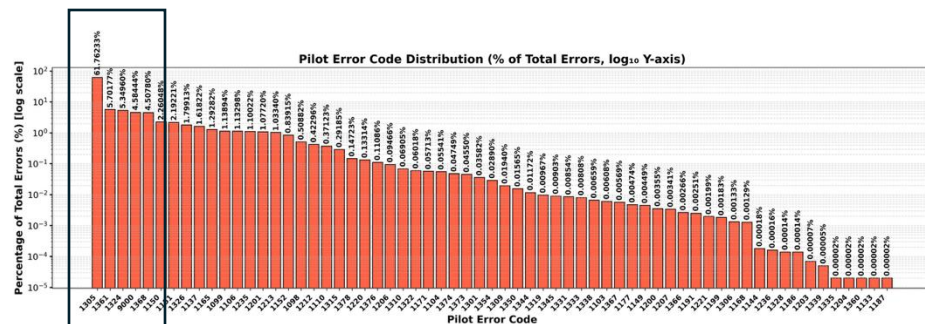
Error Code	Error Acronym	Error Message
1008	GENERALERROR	General pilot error, consult batch log
1117	QUEUEDATANOTOK	Pilot found non-valid queuedata
1220	UNKNOWNPAYLOADFAILURE	Job failed due to unknown reason (consult log file)
1224	ESRECOVERABLE	Event service: recoverable error
1228	ESFATAL	Event service: fatal error
1234	EXECUTEDCLONEJOB	Clone job is already executed
1236	FAILEDBYSERVER	Failed by server
1238	ESNOEVENTS	Event service: no events
1244	NORELEASEFOUND	No release candidates found
1247	BADXML	Badly formed XML
1300	NOTIMPLEMENTED	The class or function is not implemented
1301	UNKNOWNEXCEPTION	An unknown pilot exception has occurred
1302	CONVERSIONFAILURE	Failed to convert object data
1311	NOTDEFINED	Not defined
1312	NOTSAMELENGTH	Not same length
1314	UNKNOWNCHECKSUMTYPE	Unknown checksum type
1315	UNKNOWNTRFFFAILURE	Unknown transform failure
1319	INTERNALPILOTPROBLEM	An internal Pilot problem has occurred (consult Pilot log)
1323	UNKNOWNCOPYTOOL	Unknown copy tool
1329	NONDETERMINISTICDDM	Failed to construct SURL for non-deterministic ddm (update CRIC)
1339	DBRELEASEFAILURE	Local DBRelease handling failed (consult Pilot log)
1343	NOOUTPUTINJOBREPORT	Found no output in job report
1344	RESOURCEUNAVAILABLE	Resource temporarily unavailable
1352	STATFILEPROBLEM	Failed to stat proc file for CPU consumption calculation
1353	NOSUCHPROCESS	CPU consumption calculation failed: No such process
1354	GENERALCPUCALCPROBLEM	General CPU consumption calculation problem (consult Pilot log)
1355	COREDUMP	Core dump detected
1358	MISSINGRELEASEUNPACKED	Missing release setup in unpacked container
1359	PANDAQUEUENOTACTIVE	PanDA queue is not active
1375	LEASETIME	Lease time is up

Error categorization

Failed jobs January-March 2025

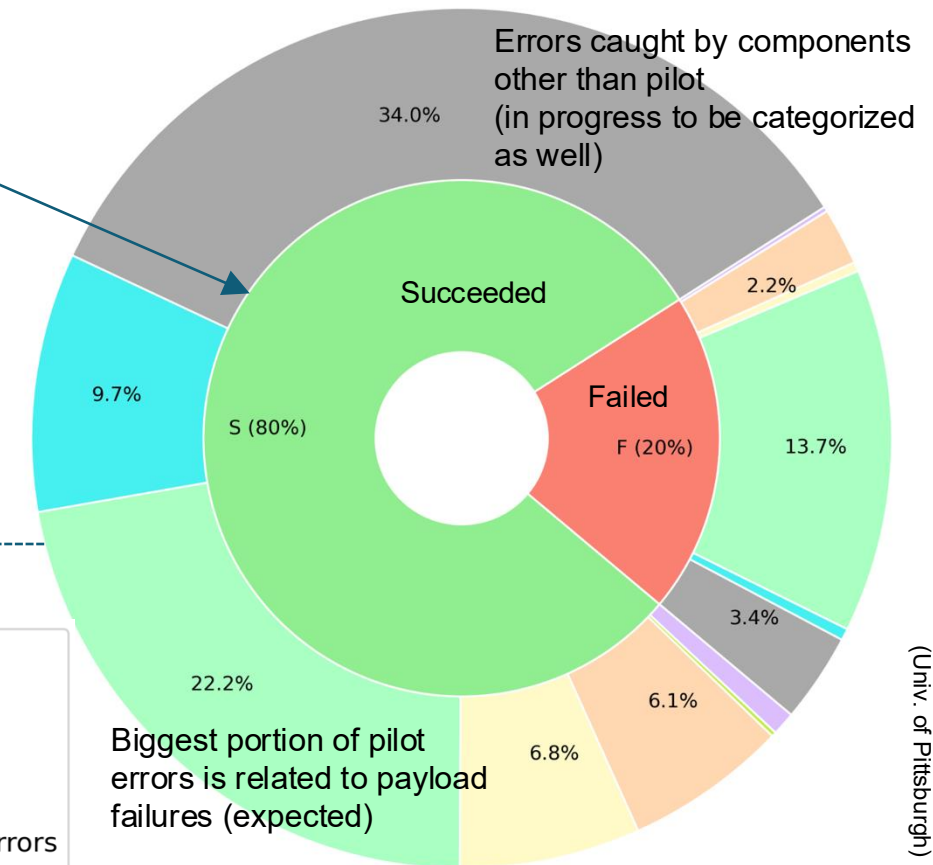
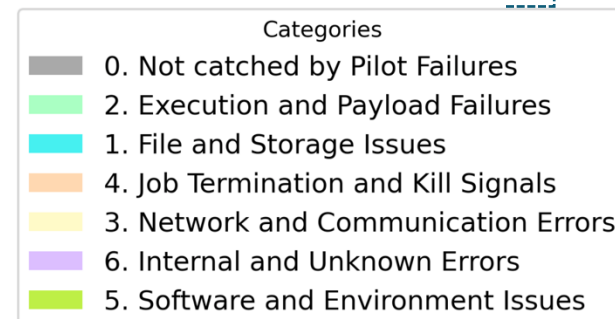
is_succeeded_eventually vs Error category

- 8.4 million failed jobs were collected from tasks that reached their final state
- Based on error code information, these failures were categorized into the six defined categories
- 80% of jobs that initially failed, eventually finished (inner layer)
 - Of the **20% jobs** that never succeeded, the biggest error category was payload failures at 70% (13.7/20)
 - Of the **80% jobs** that did succeed eventually, payload failures only corresponded to 30% -> optimization possible
 - If the errors are related to transient problems like storage and network problems, PanDA retries them to some other site and the job likely finishes successfully in the end
 - Studying errors in the various categories, should enable us to reduce number of retries which would be beneficial for improving resilience



Payload + staging errors

S. Dutta (BNL)



Error Codes in Numbers

PanDA: Server error codes

- 38 (task buffer)
- 8 (data service)
- 6 (job dispatcher)
- 3 (executor)
- **Total: 55**

PanDA: WNScripts error codes

- runAthena: 9
- runHPO: 8
- runGen: 7
- runMerge: 7
- buildJob: 4
- preGoodRunList: 3
- buildGen: 2
- **Total: 70**

PanDA: Pilot error codes

- **Total: 145**

Transform (payload) error codes

- **Total: 70**

Others

- Nordugrid error code 9000 to be split up
- Other tools and scripts
- Needs special treatment (currently leads to generic error code se by pilot)
- **Total: ~30**

Total all components: 370+

Improved and explained error messages

What Do All Those 370+ Errors Mean?

- All source codes have been analyzed by ChatGPT to provide better/longer explanations
 - Being reviewed by experts, but explanations appear reasonable – good enough as a start
- A new DB table was created (and populated) with extended error diagnostics
- The more detailed explanations are displayed on failed job pages in PanDA Monitor
 - *Also useful as a starting point/hint for LLM analysis to explain errors in logs*

Job error summary:			
Component	Code	Message	Description ⓘ
Pilot error	1150	Looping job killed by pilot	The LOOPINGJOB error is used to detect and handle payloads that appear to be 'looping' ↴ meaning they are running without producing any output for an extended time. This is a typical symptom of jobs that are stuck in an infinite loop or have stalled in a way that is not immediately fatal but effectively unproductive.

Transform errors

Error code	Detailed explanation
1	The NEEDCHECK error means that the job requires manual intervention. This typically indicates that something went wrong during the job's execution, and it cannot proceed without someone checking and resolving the issue. It is a signal to the user or operator that they need to look into the job's details, logs, or configuration to understand and fix the problem before the job can continue or be rerun.
3	The TRF_SETUP exit code is used to indicate a "Transform setup error". This means that there was an issue during the initial setup phase of the transformation process. This could involve problems with configuration files, environment settings, or other prerequisites that need to be in place before the transformation can proceed.
4	The TRF_ARG_CONV_FAIL error represents a failure to convert transform arguments to the correct type. This error occurs when the transformation process encounters an issue while trying to convert the provided arguments into the expected data types.
5	The TRF_ARG_OUT_OF_RANGE error is an exit code used to indicate that an argument provided to a transform is outside the allowed range of values. This error helps identify when an input value is not within the expected limits, which can prevent the transform from executing correctly.
7	The TRF_ARG_ERROR error indicates that there is an issue with one of the arguments provided to the transform. This could be due to various reasons such as: The argument is not recognized by the transform. The argument is in an incorrect format. The argument is missing required components. The argument does not meet the expected criteria or constraints.
8	The TRF_ARG_DATASET exit code represents an error that occurs when there is an inconsistent dataset value in the input file list. This means that the dataset values provided in the input files do not match the expected or required format, causing the transform process to fail.
9	The TRF_INPUT_FILE_ERROR error is an exit code used to indicate that there is a problem with one of the input files that the transformation process is trying to use. This could be due to various reasons such as the file being missing, corrupted, or not in the expected format. When this error is encountered, it means that the transformation process cannot proceed correctly due to issues with the input file.
11	The TRF_OUTPUT_FILE_ERROR error indicates that there was a problem when handling a transform output file. This could mean that the program encountered an issue while trying to create, write to, or manage an output file during its execution.
12	The TRF_GRAPH_ERROR error code indicates a problem in setting up the substep graph. This error occurs when there is an issue with configuring the sequence of steps (substeps) that the transformation process needs to follow. The substep graph is essentially a representation of the workflow or the sequence of operations that need to be executed. If there is a problem in defining or connecting these steps correctly, the TRF_GRAPH_ERROR is raised.
13	The TRF_EXECUTION_PATH_ERROR error indicates that there is a problem trying to generate the requested outputs from the given inputs because the graph does not provide a path to do this. In simpler terms, it means that the system cannot figure out how to process the input data to produce the desired output, likely due to a configuration or setup issue in the processing steps.
14	The TRF_KEY_ERROR error represents a key error in a transform's internal dictionary. This error occurs when the code tries to access a key in a dictionary that does not exist. In Python, dictionaries are collections of key-value pairs, and attempting to access a non-existent

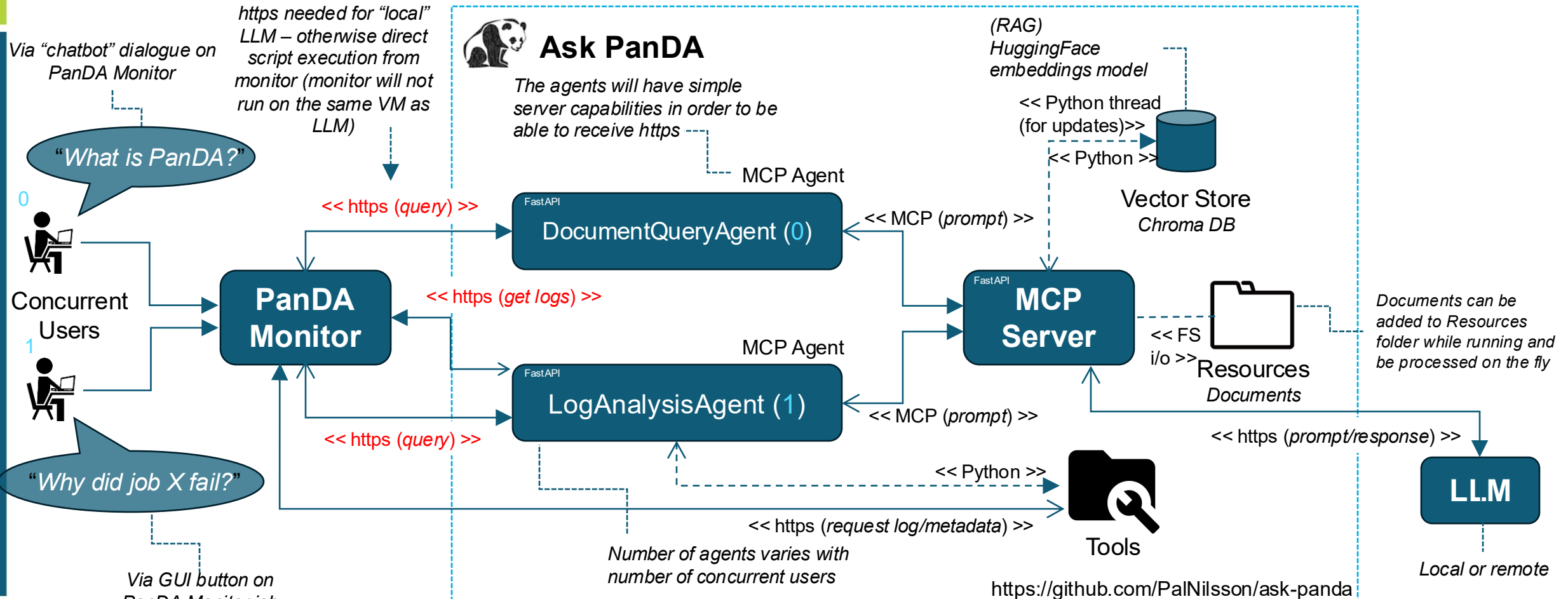
Task buffer errors

Error Code	Explanation
100	The EC_Kill error means that a task or job was terminated or killed. This could happen for various reasons, such as manual intervention by an administrator, a system decision to stop the task, or other conditions that necessitate stopping the task before it completes naturally.
101	The EC_Transfer error typically indicates a transfer timeout. This means that a data transfer operation has taken too long to complete and has timed out. This could happen due to network issues, server problems, or other interruptions that prevent the data from being transferred within the expected time frame.
102	The EC_Expire error indicates that a task or job has expired. This typically means that the task did not complete within a specified time limit and was therefore terminated. Expiration can occur due to various reasons, such as system constraints, resource limitations, or predefined timeout settings.
103	The EC_Aborted error means that a task or job was aborted. This typically means that the process was intentionally stopped before it could complete. This could happen for various reasons, such as a manual intervention by an operator, a system decision to halt the task due to certain conditions, or an error that necessitated stopping the process to prevent further issues.
105	The EC_Reassigned error indicates that a task or job has been reassigned due to re-brokerage. Re-brokerage typically means that the job was moved to a different resource or broker for execution, possibly due to issues or inefficiencies with the original assignment. This error code helps in identifying and handling such scenarios in the system.
106	The EC_Retried error means that a task or job has been retried by the server. This typically happens when the initial attempt to complete the task was unsuccessful, and the server decides to retry the task in hopes of a successful completion on the subsequent attempt. This error code helps in tracking and managing retries within the system.
107	The EC_PilotRetried error indicates that a job was retried by the pilot. In distributed computing, a pilot is a lightweight job that manages the execution of tasks on a computing resource. When a task fails or encounters an issue, the pilot can decide to retry the task in an attempt to complete it successfully. This error code is used to signify that such a retry attempt has been made by the pilot.
111	The EC_EventServiceRetried error means that an event service job has been retried. This typically means that the job encountered an issue during its execution and was automatically retried by the system in an attempt to complete it successfully. This error code helps in identifying and handling scenarios where retries are necessary for event service jobs.
112	The EC_EventServiceMerge error indicates that there was an issue related to merging for the event service. This typically means that an event service job encountered a problem during the merge process, which is a step where individual event outputs are combined.
113	The EC_MergeFailed error means that a merge job has failed. This typically means that an operation intended to combine or merge data, files, or tasks did not complete successfully. The specific context of the failure would depend on the broader application logic and what the merge operation was intended to accomplish.
114	The EC_EventServiceMaxAttempt error indicates that the maximum number of attempts for an Event Service operation has been reached. This typically means that the system has tried to process an event multiple times but has failed to do so within the allowed number of retries. This error is used to signal that no further attempts will be made for this particular event, and it should be handled accordingly.
115	The EC_EventServiceWaitOthers error means that the system is doing

Ask PanDA

- "Ask PanDA" virtual assistant in development
 - Originally intended to improve how we access project documentation but has grown into something more interesting (and useful)
 - Can in principle answer both **static** and **dynamic** questions, internally different workflows
 - "What is PanDA?" (static) [user query can be sent directly to LLM]
 - "Why did job XYZ fail?" (dynamic) [question must be interpreted before prompt can be sent to LLM]
 - Tested with OpenAI, Gemini and Anthropic APIs (and soon Llama)
- Original "Ask PanDA" tool was rewritten as an agentic workflow using Model Context Protocol (MCP) technology
 - MCP hides many implementation details and allows for simple interfaces to different LLMs
- Tests currently done manually (*Thanks Torre for Anthropic+OpenAI+Gemini tests*)
- All code and explanations in personal GitHub repo (<https://github.com/PalNilsson/ask-panda>)
 - To be moved to NPPS GitHub asap
 - Can be used with free Gemini API key if commercial key not available

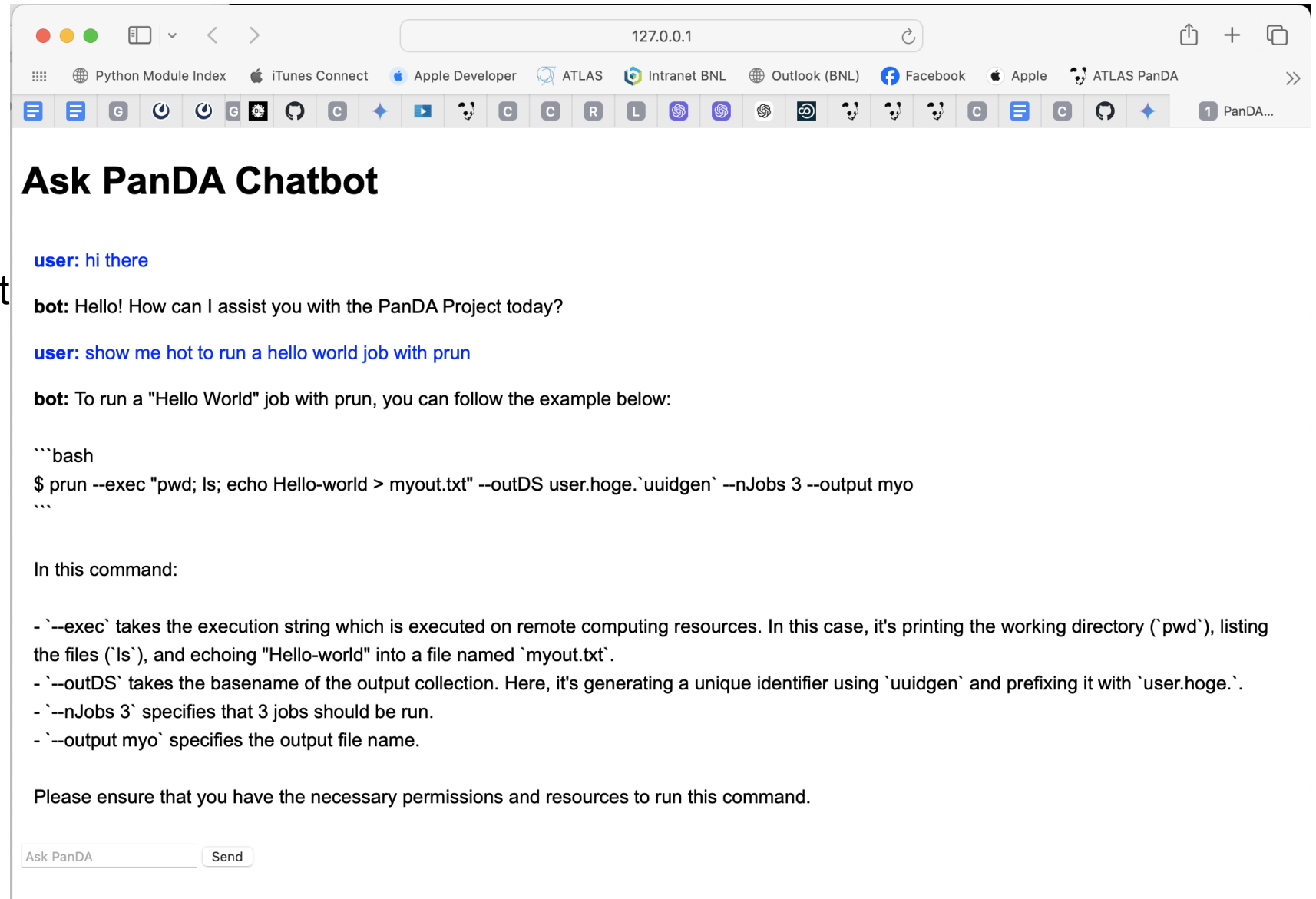
Ask PanDA MCP Implementation



- The dashed "Ask PanDA" box represents the planned implementation
- As of July 2, 2025, agents are not yet capable of receiving https messages
- Both "DocumentQueryAgent" and "LogAnalysisAgent" can be executed manually; MCP server run by uvicorn
- Asynchronous calls to agents / server to allow for concurrency
- MCP server health checks may pause user query if busy with e.g. vector store updates and restarts

Example

- Original Ask PanDA implementation equipped with a chatbot
- Currently no “chatbot” per-se in MCP implementation
 - But questions are asked via agent, currently manually executed
 - Proper chatbot to be restored at a later time



Example

- *Preliminary results from new MCP implementation in development*
- More advanced prompt with request for both user and expert explanations
- Gemini response (free 2.5-flash model – currently favored model since it is .. free) for a job with error code: 1368, message: Remote file open timed out
 - Response 1 (for users)

1) Explanations and suggestions for non-expert users:

- * **What happened?** The job tried to access a file located on a remote server, but it took too long, and the connection timed out. Think of it like trying to load a website that never finishes loading.
- * **Why did it happen?** There are a few possibilities:
 - * **The remote server was busy or slow:** The server hosting the file might have been overloaded with other requests, causing delays.
 - * **Network problems:** There might have been issues with the network connection between the worker node and the remote server. This could involve temporary network outages, high network traffic, or firewall issues.
 - * **File system issues:** The file may have been stored on a filesystem with underlying issues that caused it to become inaccessible.
 - * **Configuration issues:** The error message "Info: /cvmfs mounted; do 'setupATLAS -d -c ...' to skip default mounts. Info: \$HOME mounted; do 'setupATLAS -d -c ...' to skip default mounts." indicates a potential problem with how the ATLAS software environment is set up.
- * **What can you do?**
 - * **Try re-running the job:** Often, these types of transient network or server issues resolve themselves. A simple retry might be enough.
 - * **Check the PanDA monitoring dashboard:** Look for any widespread problems or warnings related to the storage system or network.
 - * **Contact support:** If the problem persists or if you see other jobs failing with similar errors, contact the PanDA support team or the resource provider. Provide them with the job ID and the error message.
 - * **Check to see if there are any known issues with network latency to the LPC site:** Use standard tools (e.g. `ping`, `traceroute`, `mtr`) to probe network latency to the LPC site.
 - * **Check to see if the /cvmfs filesystems are healthy:** Problems with the CVMFS client can cause hangs in remote file access.

Example (continued)

- Gemini response for same job with error code: 1368, message: Remote file open timed out
 - Response 2 (for experts / operators)
- Plan: user could select to display “User” or “Expert” response in PanDA monitor
 - Prompt will be generated, LLM response back to monitor in JSON

2) Explanations and suggestions for experts and/or system admins:

* **Root Cause Analysis:** The error message "PilotException: error code: 1368, message: Remote file open timed out" indicates a timeout during remote file access. The accompanying details "Info: /cvmfs mounted; do 'setupATLAS -d -c ...' to skip default mounts. Info: \$HOME mounted; do 'setupATLAS -d -c ...' to skip default mounts." are crucial.

* **CVMFS and '\$HOME' Mounts:** The message suggests that the default CVMFS and '\$HOME' mounts might be interfering with the remote file access, especially within a containerized environment (Apptainer). The 'setupATLAS -d -c ...' command disables these default mounts. This implies that the job might be trying to access files that are shadowed by the default CVMFS or '\$HOME' mounts, leading to resolution problems or conflicts that cause delays and timeouts.

* Possible Causes and Solutions:

1. **Incorrect ATLAS Set up Inside Container:** The container might not be correctly configured to access the necessary ATLAS software and libraries.

* **Solution:** Ensure the 'setupATLAS' script (or equivalent environment setup) is executed "within" the container with the '-d -c ...' flags to avoid default mounts that can conflict with the intended file access. This might require modifying the payload execution script or the container image itself.

2. **Network Latency to IN2P3-LPC Storage:** The logs show the job attempting to access a file at 'root://d1cgse01.in2p3.fr:1094//dpm/in2p3.fr/home/atlas/atlasdatadisk/rucio/data18_13TeV/a74b/DAOD_PHYS.43636385_000195.pod.root.1'. This points to a specific storage endpoint (IN2P3-LPC). High latency or intermittent network connectivity issues to this site will cause timeouts.

* **Solution:**

- * **Network Monitoring:** Monitor network performance between the worker node and 'd1cgse01.in2p3.fr'. Use tools like 'ping', 'traceroute', 'mtr', and 'iperf' to identify latency, packet loss, or routing problems.
- * **Firewall Rules:** Verify that firewall rules are not blocking or throttling traffic between the worker node and the IN2P3-LPC storage endpoint. Pay attention to port 1094, which is used by the ROOT protocol.
- * **Storage System Health:** Contact the IN2P3-LPC storage administrators to check the health and performance of their storage system (dCache) and the ROOT service.
- * **Routing:** Check routing and ensure it is optimized for high performance.

3. **ROOT Protocol Issues:** The job uses the ROOT protocol ('root://'). There might be problems with the ROOT client configuration or version compatibility on the worker node or within the container.

* **Solution:**

- * **ROOT Version:** Ensure that the ROOT version used in the job and the container is compatible with the storage server.
- * **ROOT Configuration:** Check the ROOT configuration files (e.g., '.rootc') for any settings that might be causing connection problems or timeouts. Pay particular attention to settings related to network buffering and connection timeouts.

4. **CVMFS interference with file access:** There may be issues with the worker node accessing CVMFS.

* **Solution:**

- * **Check CVMFS client:** Verify that the CVMFS client is properly configured and running on the worker node.
- * **CVMFS Version:** Ensure the CVMFS version is up-to-date and compatible with the ATLAS software stack.

5. **Rucio Service Availability:** Although less likely given the detailed error, verify the Rucio service ('rucio-bprod.cern.ch') is healthy. While the logs show successful Rucio interaction for trading, issues could still arise later.

* **Solution:** Check the Rucio status pages and dashboards for any reported incidents.

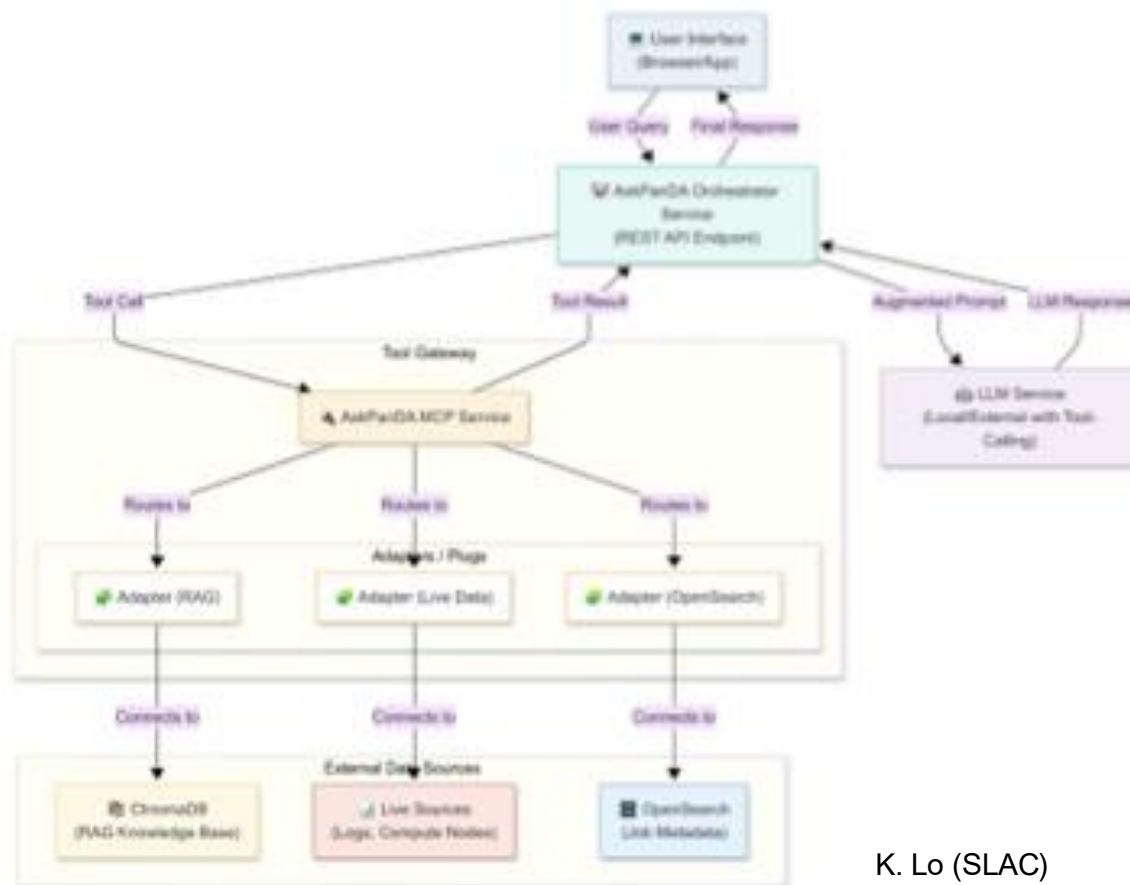
* Debugging Steps:

1. **Reproduce the Problem:** Try to reproduce the problem manually on the worker node or within a similar container environment. This will allow you to isolate the issue and test potential solutions.
2. **Increase Verbosity:** Increase the logging verbosity of the Pilot, ROOT, and any other relevant software components to gather more detailed information about the error.
3. **Packet Capture:** Use 'tcpdump' or 'Wireshark' to capture network traffic between the worker node and the storage server. This can help identify network latency, dropped packets, or other network-related problems.
4. **Consult Pilot and ATLAS Documentation:** Review the Pilot and ATLAS documentation for troubleshooting tips and known issues related to remote file access and containerization.

By systematically investigating these areas, you should be able to pinpoint the root cause of the "Remote file open timed out" error and implement the appropriate solution.

Local LLM at SLAC (K. Lo, W. Yang)

AskPanDA: A Scalable Agent with a Hybrid Tool Gateway



K. Lo (SLAC)



AskPanDA System Architecture

- Ask PanDA MCP Service plans in Vera C. Rubin Experiment
 - Acts as a Hybrid Tool Gateway, routing tool calls to different adapters
 - RAG adapter (ChromaDB with embedded Rubin job summaries)
 - OpenSearch adapter (Rubin job metadata)
 - .. and a planned Live Data adapter (for real-time logs and compute node state)
- This setup allows answering operational questions like:
 - “What is job X doing?” (real-time)
 - “How many jobs are we running?”
- The LLM runs locally via Ollama and calls tools via augmented prompts

POC MCP Server for Claude Desktop

- Ask PanDA partly ported as MCP server(s) for Claude Desktop
 - One server for log analysis
 - Async downloads metadata from CRIC, log from PanDA, constructs prompt and returns to LLM for interpretation
 - Another server for questions about site information
 - Server downloads metadata from CRIC, returns to LLM
 - (Actually standard Ask PanDA doesn't know how to do this yet)
- Can currently handle prompts like
 - *“Show me all the queues that are using rucio as copytool”* [metadata]
 - *“Display the corepower for all queues”* [metadata]
 - *“What copytool does BNL use?”* [metadata]
 - *“What is the status of job 6709690286?”* [metadata]
 - *“Why did job 6715819744 fail?”* [log analysis]
- Different types of prompts have to be supported individually
- ***Slow*** development using free Claude account
 - Either runs out of credits or dialogs become too long too quickly
 - But it is .. free

PN Why did job 6715819744 fail?

I'll analyze job 6715819744 to determine why it failed. Let me download the metadata and examine the failure details.

G panda_job_analyze_failure

Based on the failure analysis, job 6715819744 failed due to a **file transfer timeout during stage-out**. Here are the key details:

Failure Summary

- **Job Status:** Failed
- **Primary Cause:** File transfer timeout (Pilot Error Code 1152)
- **Site:** UKI-NORTHGRID-SHEF-HEP (Sheffield, UK)

Specific Error Details

The job failed during the stage-out phase when trying to transfer the output file `output.1.2abc361d-8752-479b-a4ea-303d72221559_61142.pool.root` from the computing site to the storage destination `UKI-NORTHGRID-MAN-HEP-CEPH_DATADISK`.

Error Message: "File transfer timed out during stage-out: copy command timed out: TimeoutException: Timeout reached, timeout=310 seconds"

Job Execution Details

- The job itself executed successfully (transformation exit code: 0)
- CPU consumption was only 19 seconds, indicating the actual processing completed quickly
- The job ran for about 58 minutes total, but most of this time was spent waiting or transferring files
- Memory usage was normal (max RSS: 67MB)

What Happened

1. The job successfully processed the input file (2.5GB DAOD_PHYS file)
2. The actual computation completed without errors
3. During the final stage-out phase, the file transfer to storage timed out after 310 seconds
4. The Rucio copy tool failed to complete the transfer within the timeout period

This is typically a network or storage system issue rather than a problem with the job's code or 15 configuration. The job could be retried, as it has 2 more attempts remaining (currently on attempt 1 of 3).

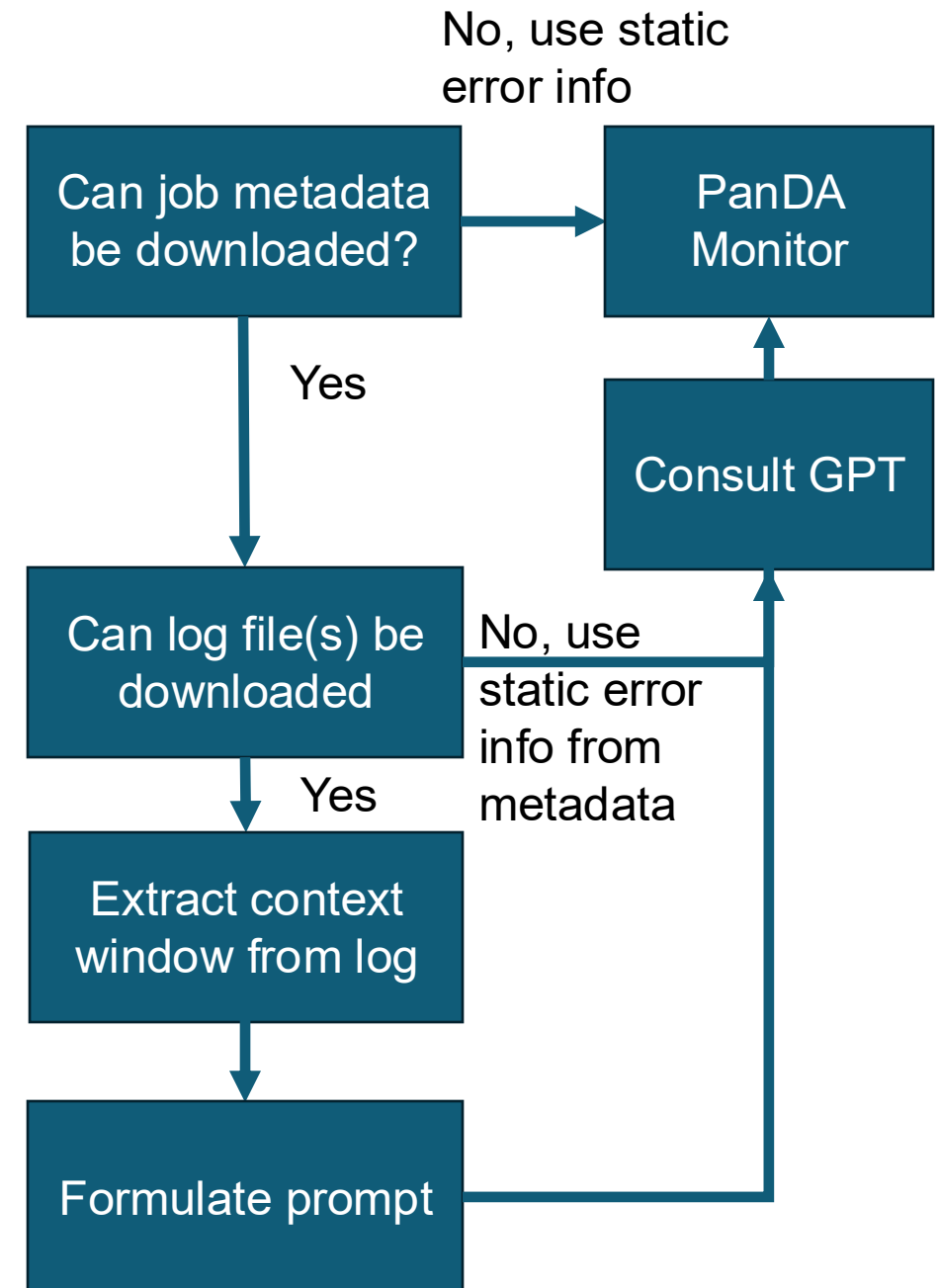
Next Steps

- Update error descriptions for recent modifications in panda-wnscripts
- Refine and improve (if necessary) error categories
 - Incl. re-categorization using LLM generated detailed error diagnostics (need model with larger context window)
- Test MCP agents with new version of PanDA Monitor as soon as it is available
- Initiate testing with local LLM at SLAC (existing log analysis + “chatbot”)
 - Investigate and plan for adapters for Rubin
- Development and testing with Claude Desktop
 - Ask PanDA has a single MCP server for all use cases
 - Claude [Desktop] favors having multiple MCP servers for different use cases
 - This makes it easier for Claude to select the correct tool to use but has several drawbacks (increased complexity, duplication of effort, multiple servers = wasting system resources, etc)
 - Import as many tools and functions as possible from main Ask PanDA in MCP servers for Claude

Additional slides ...

Log Analysis Agent

- For questions like “why did my job fail?”, LLMs will of course not know anything unless prompt is carefully prepared
- Log analyzer script performs multiple steps before asking LLM what went wrong
 - Script first downloads job metadata to learn general details about job (esp. state, error info)
 - Relevant log will be downloaded from PanDA
 - Extract context window (e.g. log messages immediately prior to failure)
 - Could use ML model for this step at a later time
 - Construct detailed prompt (can be enhanced with detailed error diagnostics to aid analysis) and send to LLM
- Response is returned from model and presented to user



RAG Approach for Static Queries

- **Retrieval-Augmented Generation (RAG)** is a better approach for our (static) purposes than training a new model
 - E.g. PanDA documentation (< 1MB) is essentially the only source of knowledge needed for our chatbot (and thus has too little information to train an LLM from scratch)
 - (Throwing in other relevant documents as well; e.g. Rucio documentation)
- RAG stores the input document in an **embedding-based database**
 - When a user asks a question, retrieve the most relevant sections and pass them to an existing LLM
 - RAG does not extend the LLM model but augments the answers
 - Works well for long documents since it avoids input size limitations
 - Originally Langchain OpenAI embeddings, just moved to HuggingFace (free)
 - Everything stored in local Chroma DB vector store