iDDS integration

Wen Guan, Tadashi Maeno, Gancho Dimitrov
Brian Bockelman, Torre Wenaus
Fernando Barreiro, Fahui Lin, Rui Zhang, Misha Borodin, Paul Nilsson

May 28, 2020
WFMS

iDDS Status

- **Main architecture (production)**
  - iDDS database, core, REST API
  - Plugins
  - Agents
  - Watchdogs
- **Documents & monitors**
  - **Home page:** https://idds.cern.ch
  - **Codes:** https://github.com/HSF/iDDS
  - **Documents:** https://idds.readthedocs.io (dev)
  - **ATLAS Monitor:** https://bigpanda.cern.ch/idds/
- **Use cases**
  - Fine-grained data carousel -- ready
  - Hyperparameter tuning -- integrating
  - Decision making for active learning -- developing
  - Other usecases in 2020

iDDS Data Carousel

iDDS data carousel with data15 reprocessing

- **Integration Test Status with data15 reprocessing**
  - 99% staged in 4 days
    - Many files were handled shortly after the request was submitted
    - Few files were stuck on tape, finished some days later.
  - **63 datasets**
  - **426,331 files**
  - **850TB**
  - iDDS was not fully occupied, will test with more tasks

	COUNT(*)	SUM(IN_TOTAL_FILES)	SUM(IN_BYTES)	SUM(OUT_PROCESSED_FILES)
1	63	426331	932290670456064	426331

Requests:

Show 10 entries Search:

request_id	scope	name	status	transform_status	in_status	in_total_files	in_processed_files	out_stat
6	data15_13tev	data15_13tev.00276262.physics_main.daq.raw	Finished	Finished	Closed	5573	5573	Closed
13	data15_13tev	data15_13tev.00276954.physics_main.daq.raw	Finished	Finished	Closed	192	192	Closed
17	data15_13tev	data15_13tev.00279259.physics_main.daq.raw	Finished	Finished	Closed	1524	1524	Closed
24	data15_13tev	data15_13tev.00279984.physics_main.daq.raw	Finished	Finished	Closed	12200	12200	Closed
28	data15_13tev	data15_13tev.00280464.physics_main.daq.raw	Finished	Finished	Closed	8990	8990	Closed
34	data15_13tev	data15_13tev.00283780.physics_main.daq.raw	Finished	Finished	Closed	16679	16679	Closed
39	data15_13tev	data15_13tev.00284427.physics_main.daq.raw	Finished	Finished	Closed	4132	4132	Closed
41	data15_13tev	data15_13tev.00276790.physics_main.daq.raw	Finished	Finished	Closed	408	408	Closed
46	data15_13tev	data15_13tev.00280614.physics_main.daq.raw	Finished	Finished	Closed	3769	3769	Closed
52	data15_13tev	data15_13tev.00282992.physics_main.daq.raw	Finished	Finished	Closed	13100	13100	Closed

Showing 1 to 10 of 64 entries Previous 1 2 3 4 5 6 7 Next

iDDS data carousel with all datasets

- **Data Carousel Integration with all datasets(mainly data15 and zerobias)**
  - 667 datasets
  - 768, 115 files
  - 1.2 PB
  - iDDS was not fully occupied, will test with more tasks, finished zerobias datasets in few days.

COUNT(*)	SUM(IN_TOTAL_FILES)	SUM(IN_BYTES)	SUM(OUT_PROCESSED_FILES)
667	768115	1.2889E+15	767449

request_id	scope	name	status	transform_status	in_status	in_total_files	in_processed_files	out_status	out_total_files	out_processed_files
724	21423307	**21423307**	Finished	Finished	Closed	1	1	Closed	5	5
723	21422500	**21422500**	Transforming	Transforming	Closed	1	1	Processing	4	1
722	21420035	**21420035**	Transforming	Transforming	Closed	1	1	Processing	4	1
721	21407107	**21407107**	Finished	Finished	Closed	1	1	Closed	5	5
720	21405829	**21405829**	Transforming	Transforming	Closed	1	1	Processing	5	3
719	21404562	**21404562**	Transforming	Transforming	Closed	1	1	Processing	5	4
718	21404113	**21404113**	Transforming	Transforming	Closed	1	1	Processing	5	3
717	21391145	**21391145**	Finished	Finished	Closed	1	1	Closed	5	5
716	21391132	**21391132**	Transforming	Transforming	Closed	1	1	Processing	4	1
715	21390625	**21390625**	Transforming	Transforming	Closed	1	1	Processing	4	1

Showing 1 to 10 of 695 entries Previous 1 2 3 4 5 … 70 Next

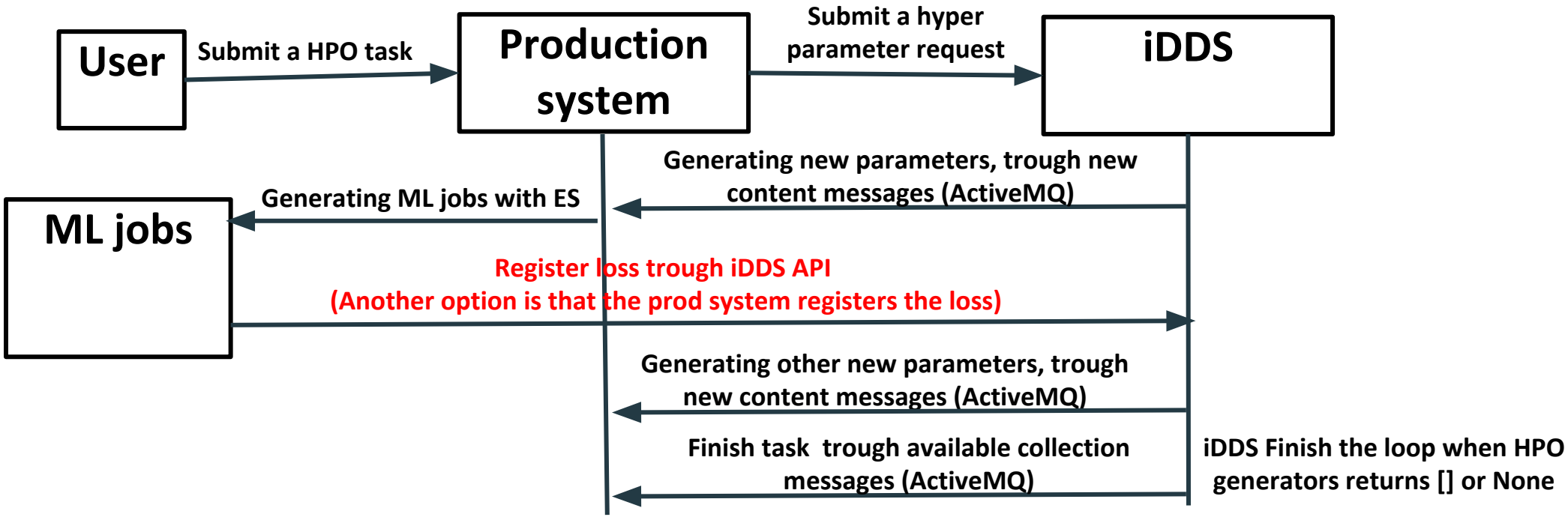
A lot of tasks have finished.

The monitor has increased from 7 pages to 70 pages

iDDS HyperParameterOptimiztion (HPO)

iDDS HPO (Hyper Parameter Optimization)

- **Purpose**
  - Using iDDS to generate hyperparameters and trigger production system to automatically process training with the new hyperparameters.
  - https://idds.readthedocs.io/en/latest/usecases/hyperparemeter_optimization.html
  - Status: Integrating it with JEDI (successful task https://bigpanda.cern.ch/task/21423307/)
- **workflow:**



iDDS HPO (Hyper Parameter Optimization)

- **iDDS Messages**
  - New content: When a new output content(file) is created but not processed.
  - Available content: When an output content(file) is available(successfully processed or evaluated)
  - Available Collection: When all contents(files) in a collection(dataset) are available.
- **iDDS processing:**
  - iDDS runs a hyperparameter generator in a local condor cluster again and again, until the task finishes.
  - Every time the hyperparameter generator will read all evaluated hyper parameters with registered loss, based on them, new hyperparameters are generated for a new loop.
  - iDDS currently developed two predefined hyperparameter generators: bayesian and nevergrad.
  - iDDS also supports docker containers and developed an example for users to define their own generators.
  - Documents are in https://idds.readthedocs.io.

iDDS HPO (Hyper Parameter Optimization)

- **iDDS RESTful client for HPO**
  - get_hyperparameter: to get hyperparameters
  - update_hyperparemeter: to register loss results.
- **iDDS HPO integration with JEDI, Pilot:**
  - **JEDI HPO consumes 'new content' messages:**
    - **New content: for new hyperparameters: to generate event ranges within the ES framework.**
  - **New transform runHPO-00-00-01 running in Pilot:**
    - **Get event range from panda.**
    - **Get hyperparameter from iDDS with event range id(hyper parameter id) through iDDS REST.**
    - **Run ML training**
    - **Register loss results to iDDS through iDDS REST.**
    - **Update event range status**
  - **Finish task**
    - **When all hyperparametes are evaluated, iDDS publish 'Collection available' messages.**
    - **JEDI consumes this message and finishes the task.**

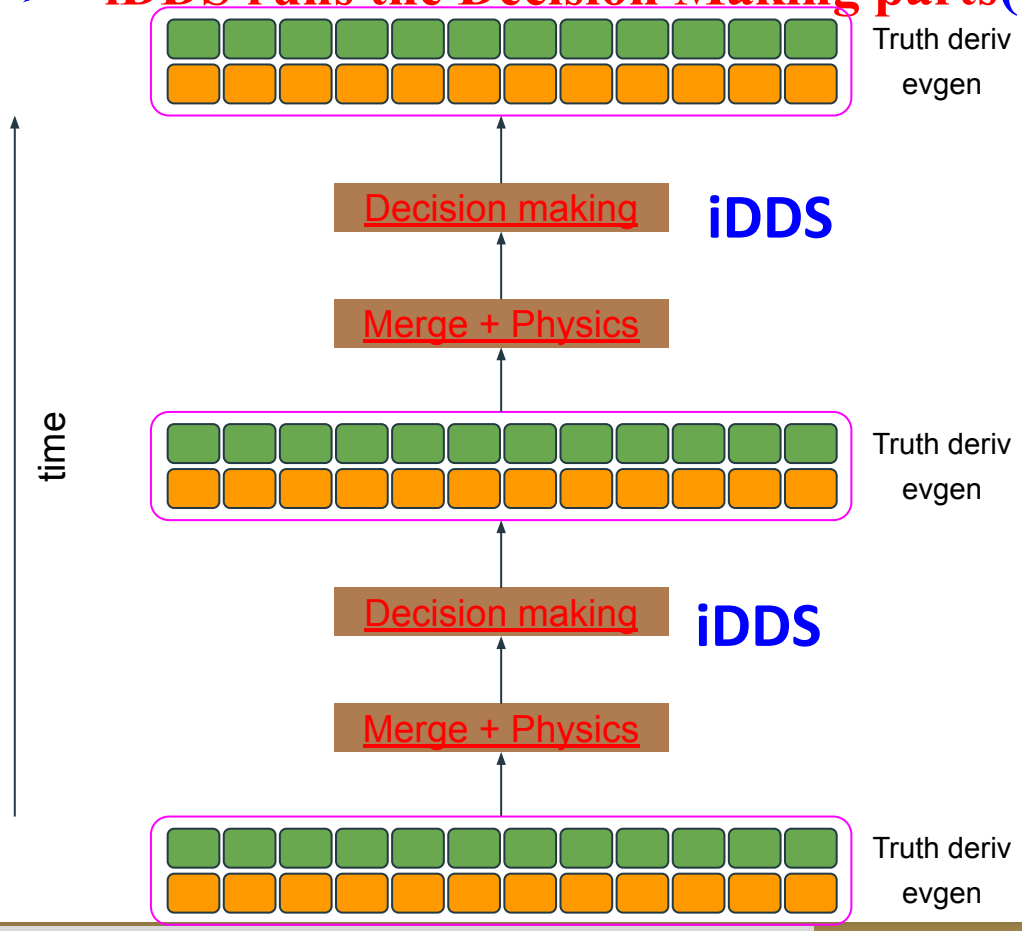
One successful task: https://bigpanda.cern.ch/task/21423307/

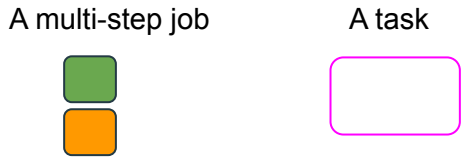
iDDS Active Learning (AL)

To integrate it with Prodsys2

iDDS Active Learning (AL)

- Active learning
  - Running tasks on top of results of old tasks
  - Decision making to generate new tasks from old results
    - Light job, good to execute it immediately and then trigger next step. iDDS can get rid of some latency.
- Workflow with grid entities
  - Production system processes the normal task
  - iDDS runs the Decision Making parts(with/without merge parts)

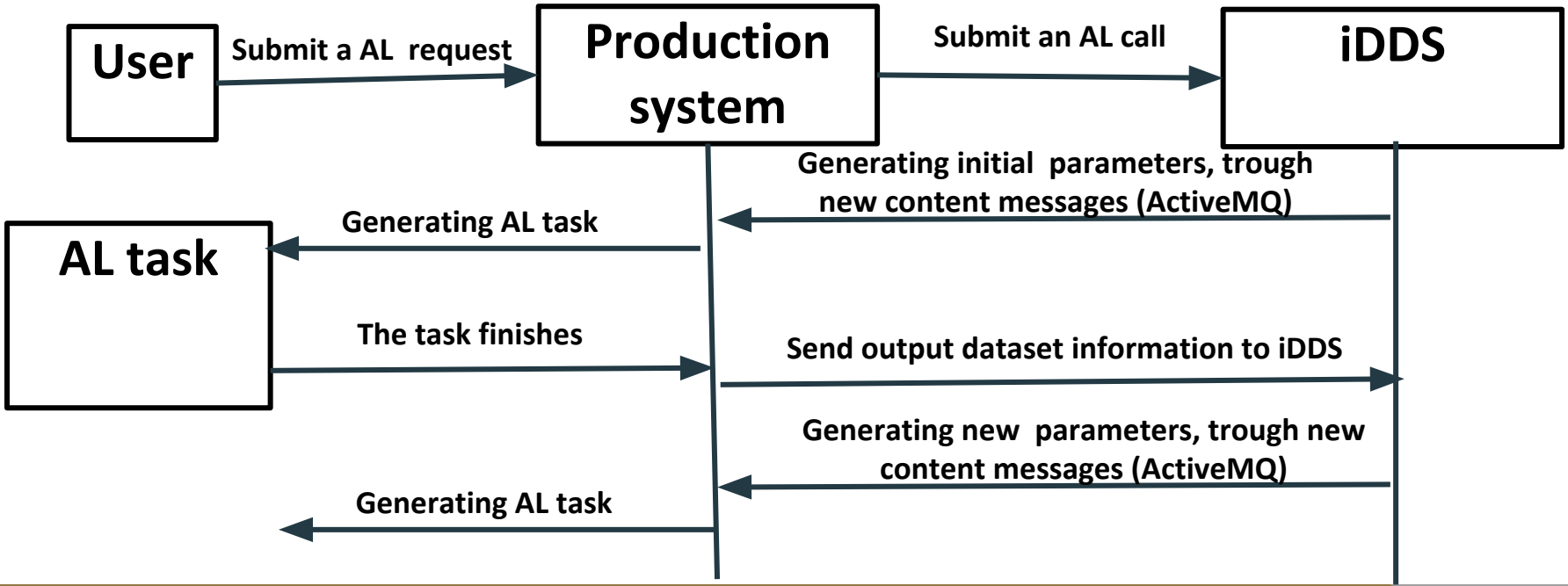




- Each job runs evgen and derivation sequentially
- Each task generates many multi-step jobs
- Once the first task is finished output are merged and some values are calculated
- Parameters for the next task are decided using the values
- Better to get rid of latency in the intermediate steps (merge+physics and decision making)

iDDS Active Learning(AL)

- **Decision making job will run in iDDS**
- **iDDS messaging and iDDS processing will be similar with HPO.**
- **iDDS AL integration with prodsys2 (to discuss)**
  - **Option 1: similar workflow as HPO.**
    - Prodsys2 sends one AL call to iDDS Rest at the beginning of the AL task.
    - iDDS generating parameters with initial parameters.
    - Prodsys2 consumes the 'new content' message to create an AL task.
    - Prodsys2 needs to send information to iDDS when a task finishes, for example, the output dataset. Then iDDS can evaluate the new dataset and generates new parameters.
    - **Here only one call with multiple updates. However, prodsys2 needs to register new output dataset information when it finishes, just like to register HPO loss in HPO workflow.**



iDDS Active Learning(AL)

❖ iDDS AL integration with prodsys2 (to discuss)

- ➢ **Option 2:**
  - Prodsys2 can send another call to iDDS.
  - iDDS runs the Decision making job and returns the outputs.
  - Prodsys2 decides whether to stop or generate another task, for example: stop when the returned outputs is empty.
  - One call per AL task. Totally there will be multiple calls.