

The Future of PanDA in ATLAS Distributed Computing

K. De, A. Klimentov, <u>T. Maeno</u>, P. Nilsson, D. Oleynik, S. Panitkin, A. Petrosyan, J. Schovancova, A. Vaniachine, T. Wenaus on behalf of ATLAS collaboration University of Texas at Arlington, USA Brookhaven National Laboratory, USA Joint Inst. for Nuclear Research, RU Argonne National Laboratory, USA CHEP2015, Okinawa, Japan, April 13-17 2015

Introduction



- PanDA = Production and Distributed Analysis System
 - Designed to meet ATLAS production/analysis requirements for a data-driven workload management system capable of operating at LHC data processing scale
- PanDA has performed well for ATLAS including the LHC Run1 data taking period
 - Producing high volume Monte-Carlo samples and making huge computing resources available for individual analysis
 - Running ~150K jobs concurrently
 - Processing ~0.7 million (~1.5 million at peak) jobs per day
 - Being actively evolved to meet the rapidly changing requirements for analysis use cases
 - No significant service disruptions

> New developments for Run 2 and beyond

Tadashi Maeno, CHEP2015, Okinawa, Japan April 13-17 2015



Motivation for New Developments

- More efficient usage of pledged resources
- Partitioning of workload suitable for opportunistic resources based on their dynamic characteristics
- Handling of the workflow and bookkeeping both with coarse and fine granularities
- > Integration of network awareness
- > Improvement of visualization



Major System Evolution for Run 2 and Beyond

Tadashi Maeno, CHEP2015, Okinawa, Japan April 13-17 2015

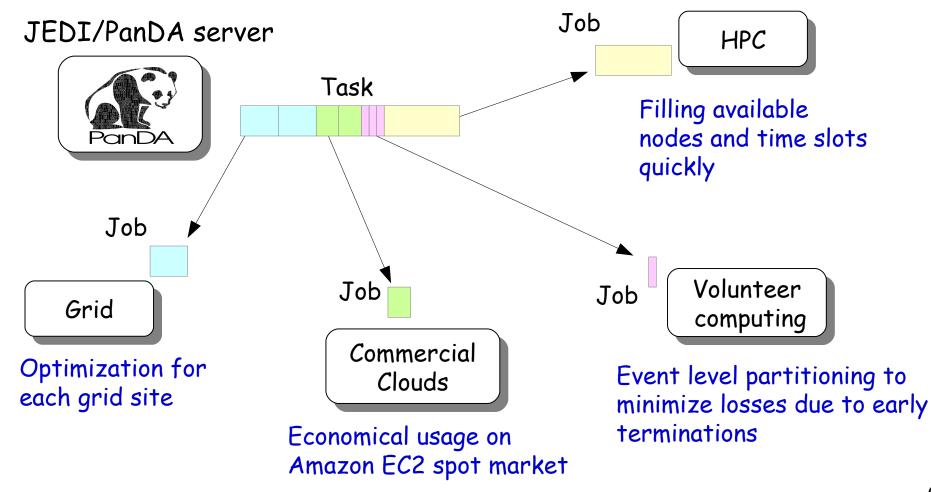
4

Dynamic Job Definition 1/3 BROOKHAVE

- > Two new components
 - DEFT (Database Engine for Task)
 - Handles production requests and tasks
 - JEDI (Job Execution and Definition Interface)
 - Dynamically splits workload for optimal usage of resources
 - Manages workload at task, job, file, and event level
 - Automatically merges outputs
- Tasks are accepted to be partitioned to jobs based on the dynamic state of available resources
 - Jobs are an implementation detail of getting tasks done

Dynamic Job Definition 2/3 BROOKHAVEN NATIONAL LABORATOR

Workload partitioning for traditional and opportunistic resources



Dynamic Job Definition 3/3 BROOKHAVEN > Benefits

- Excluding requirements from users of detailed knowledge on computing resources
 - Especially for heterogeneous resources, e.g., many CPU cores, very short walltime limit, etc
- Self-optimization of job parameters
 - Real job metrics are collected using scout jobs
 - A small number (~10) of jobs (= scout jobs) are generated for each task with minimum input chunks
 - Job parameters are optimized using job metrics for the rest of input
- Simplification of client tools and centralization of user functions



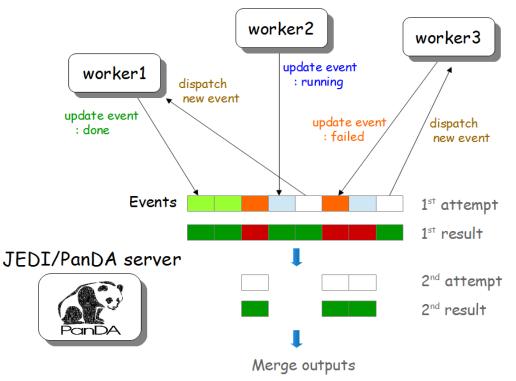
Integration of Network Awareness

> Usage of WAN data access for user jobs

- Job brokerage taking costs for WAN data access into account
- Slightly relaxing the ATLAS computing model
 - Sending a fraction of user jobs to sites which don't locally have data but have good network connection to remote data
- Throttle mechanism to protect SE
- Dynamic coupling of Tier1 and Tier2 sites based on network performance and data locality
 - Files are transferred to Tier2 site from Tier1 site via good network without multi-hop



Event Service



The fine grained partitioning of processing

- Allowing workloads to be tailored dynamically to resources currently available
- Minimizing losses when opportunistic processing slots are abruptly revoked

> HPC : validated on NERSC Edison, being ported to other platforms

- Commercial cloud : validating on Amazon EC2 spot market
- Volunteer computing : being ported to ATLAS@Home
- More details in CHEP15 talks

#140, V. Tsulaia : Fine grained event processing on HPCs with the ATLAS Yoda system #183, T. Wenaus : The ATLAS Event Service: A new approach to event processing Tadashi Maeno, CHEP2015, Okinawa, Japan April 13-17 2015



Evolving PanDA Pilot 1/2

- Refactorization to core modules and experimentspecific plugins
- Supporting new workflows for Event Service and HPCs
 - HPC plug-ins have been developed for Titan (OLCF, US), Edison/Hopper (NERSC, US), Mira (ALCF, US) and Anselm (Ostrava, CZ)
 - Event Service on HPCs using newly developed Yoda software suite
 - Yoda acts as an intermediary layer between the PanDA Server and the PanDA Pilot which does not have access to outside connections on HPCs
 - Successfully validated on NERSC sites and is currently being extended to Titan
 - In development for Volunteer Computing ATLAS@Home project



Evolving PanDA Pilot 2/2

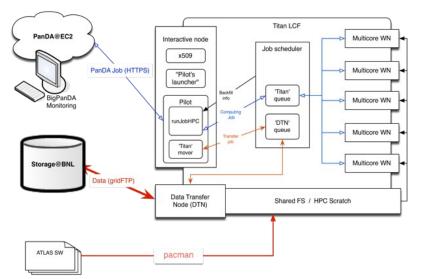
> Using object stores as temporary storage

- Highly useful for event service jobs that produce many small output files that are merged after the event service job has finished
- Also in testing for log files
- Ceph based object stores are available at BNL and CERN, and soon as RAL (UK)
- Support for gLExec
 - Pilot was also refactored to enable gLExec integration, the ability to dynamically switch the identity from the pilot to the user when executing the payload
 - More details in CHEP15 poster

#155, E. Karavakis : gLExec Integration with ATLAS PanDA Workload Management System

PanDA on Titan at OLCF





More details in CHEP15 talk

#152, S. Panitkin : Integration of PanDA workload management system with Titan supercomputer at OLCF Work on integration of Titan machines with PanDA

NATIONAL LABORATOR

- Modified PanDA pilot to run on Titan's front-end nodes with backfill mode
 - Collecting information about free resources in quasirealtime
 - Submits jobs to Titan's scheduler based on the info
- Successfully demonstrated steady operations for continuous PanDA job submission in backfill mode





- Based on Django framework
- Clear separation between data access and visualization
- REST APIs to access object information
- Provide task-oriented view

ATLAS PanDA monitor home		alpanda053 @ 13:14,	cache 3' produce -	Convision -	10 -	Itala																							
		albanaacce @ 10111	Prodsys •	Services -	VU V	нер	Ť																				ices - VO		
obal concurrent running job core counts,	all alter all ich fr	man by cloud loot 1 7 20 d															PanDA task 4881914: mc14_13TeV.206547.PowHegPythia8_AU2CT10_ggH1000_w30_tautauhh.recon.e3540_s1982_s2008_r5787 alpanda056 @ 15:13, cache :												
obal concurrent running job core counts,	an sites, an job ty	pes, by cloud, last 1, 7, 50 da	iys														Task 4881914: mc14_13TeV.206547.PowHegPythia8_AU2CT10_ggH1000_w30_tautauhh.recon.e3540_s1982_s2008_r5787												
States of Recording Joles 31 Along the 2015/21/21 12/01 p.2015/21/21 22/01 p.2015/21/21 22/01 p.2015/21/21 22/01 p.2015/21/21 22/01 p.2015	Sourcesul	Slots of Plurining Subs 7 Days from 2025-00-34 St 2025-00-45	Marcard Marcard	Stats of Asseming Jul 27 Days from 2015-01-05 01 2	15-42-05		1										Task ID	Request	Туре	WorkingGrou	p User	Campai	gn Ta: sta	SK L fini		Created	Modified	Core	
	200			.		1											4881914	1782	prod	AP_HIGG	mehlhas	e MC14a	do	80 (100	80	2015- 02-23 10:13	03-01 07:18	8	
ATLAS PanDA monitor Desh * Tesks * Jobs * Errors * Users * Site								s ⊤ Ir	nciden																				
PanDA production task dashboard, last 7 days. Params: mode=task												States of jobs in this task (merge jobs excluded) Show all jobs																	
	produc	production task summary by cloud, last 7 days Hover over state name to see full name. Task str												defined	waiting	pending	assigned	throttled	activated	sent s	arting r	unning	holding	transferrin	ng fi				
E E E E E	E E	E E E	E E	-							1		1															80	
				Cloud	-		def	assgn	rdy	pend	scout	sctd	run	prep	done	fail	Jump to	job param	neters, 1	ask paramete	rs								
obal concurrent running job core counts,	all sites, all job ty	pes, by activity, last 1, 7, 30 o	days		2035	69	2	5	0	1245	0	0	0	0	0	0													
				CA	4494	0	0	0	0	56	31	0	209	0	4052	19		job list (acce: and logs)	ss to job	details pare 4881	nt task 913		prodsys page		okerage gger	JEDI a logger		error summa	
Slob of Family Jobs 24 Algun ton 2015 (1994) 1138 (1994) (1994)	Soon and New Control	Slots of Running Jobs 7 Zays Run 2023 do Ja & 3003 do as	Name I Control Name	CERN	5058	0	0	0	1	15	44	0	692	0	4275	0									_				
				DE	2259	0	0	0	4	18	13	0	333	0	1780	5	6 datasets, show/hide by ty												
		ه کا کا کا ک		ES	2222	0	0	0	67	10	20	0	277	0	1795	1	Dataset,	container na	ame T	ype Stream	tream State Status Nfi			es Created Modi					
		FR	2835	0	0	0		40	58		166	0	2490	17	Job para	ob parameters													
						U	U	U	1			U		U		17		-											
		And Add Add Add Add		Т	4664	0	1	0	897	13	59	0	340	0	3303	7	13	0 2	2	1 0	0	0	8	0	0	0			
Reserves Bill Reserve States States	NC Section NC Account of the Action Exclusion	Terrer 1000 Annua 2000	K baselan K kasalan Dashazay Basas	ND	2476	0	0	0	5	30	69	0	203	0	2119	2	20	1 1	4	0 0	0	0	13	0	0	0			
				NL	2670	0	0	0	69	9	33	0	216	0	2273	6	36	0 2	0	1 0	0	0	7	0	0	0			
				тw	3465	0	1	0	1	9	47	0	156	0	3215	3	14	0 1	4	0 0	0	0	5	0	0	0			
				UK	2945	0	0	0	24	22	13	0	388	0	2392	9	25	1 5	5	2 0	0	0	14	0	0	0			

Tadashi Maeno, CHEP2015, Okinawa, Japan April 13-17 2015



Future Plans

- More intelligence to workload partitioning and brokerage
- Proactive control of the network to optimize workflows and dataflows
- New computing resources in production more efficiently and economically
- Lightweight tools for users, who are not fully integrated to the grid, to leverage PanDA for utilization of local beyondpledge resources



Conclusions

- PanDA has performed well for ATLAS including the LHC Run 1 data taking period
- New components and features have been delivered to ATLAS before LHC Run 2
- Many developments and challenges to come while steadily running for LHC Run 2