COMPASS Production System: Processing on HPC

Artem Petrosyan, JINR

GRID 2018, Dubna, Russia

COMPASS collaboration

Common **M**uon and **P**roton **A**pparatus for **S**tructure and **S**pectroscopy

24 institutions from 13 countries
– nearly 250 physicists

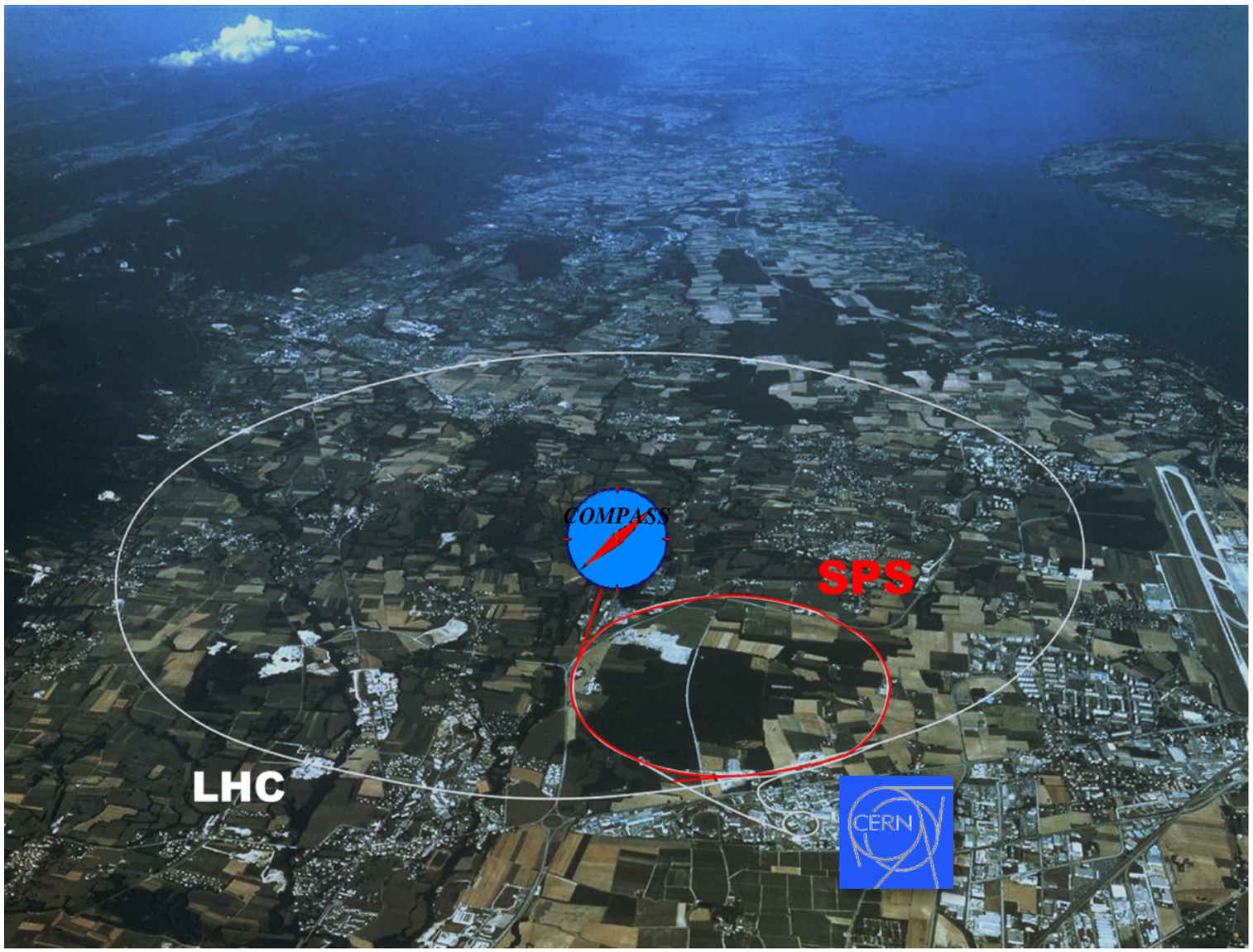
- **CERN SPS north area**
- **Fixed target experiment**
- **Approved in 1997 (20 years)**
- **Taking data since 2002**

Wide physics program

COMPASS-I
- **Data taking 2002-2011**
- **Muon and hadron beams**
- **Nucleon spin structure**
- **Spectroscopy**

COMPASS-II
- **Data taking 2012-2018 (2021?)**
- **Primakoff**
- **DVCS (GPD+SIDIS)**
- **Polarized Drell-Yan**
- **Transverse deuteron SIDIS**

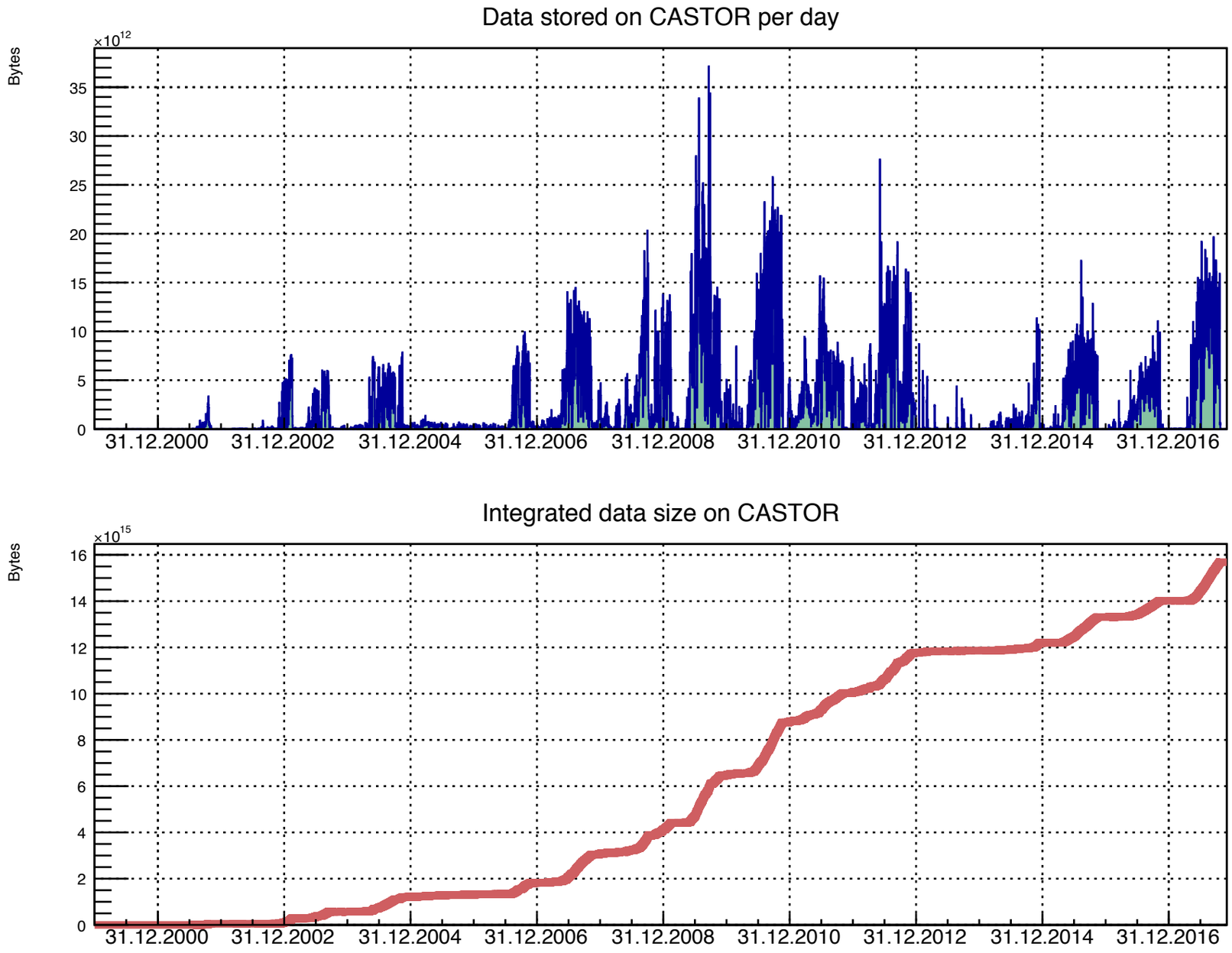
Many "beyond 2021" ideas



COMPASS web page: http://wwwcompass.cern.ch

Raw data

2001 - 13 TB
2002 - 196
2003 - 230
2004 - 496
2006 - 390
2007 - 912
2008 - 523
2009 - 1223
2010 - 1740
2011 - 518
2012 - 878
2015 - 801
2016 - 571
2017 - 1391



ProdSys components

1. Task requests layer: Web UI
2. Job definition layer
3. Job execution layer: PanDA
4. Workflow management
5. Data management
6. Monitoring

Stats and performance

- Since August 2017
  - ~2 500 000 chunks of raw data processed
  - ~70 000 000 of events processed
  - ~500TB of merged data produced and migrated to Castor
  - ~5 000 000 jobs processed since August: reco, ddd filtering, merging of mDST, hist and event dumps
- Up to 20 000 of jobs being processed simultaneously

Processing on Blue Waters

- Allocation: 9M node hours per year
- Raw data delivered to BW manually via Globus Online
- Production software installed on local file system
- Calibration db runs on each computing node, i.e. per each 32 jobs, first job on the node starts new db instance
- PanDA Multi-Job Pilot is used, extended by COMPASS logic
  - Submission size: each Pilot can run up to 512 jobs on16 nodes
- Task submission, management and monitoring fully integrated into ProdSys UI and PanDA monitoring
- Processing 25-50K jobs, 500-1000 nodes, target is to process 100-150k of jobs

Blue Waters System Summary

- The Blue Waters system is a Cray XE/XK hybrid machine composed of AMD 6276 "Interlagos" processors (nominal clock speed of at least 2.3 GHz) and NVIDIA GK110 (K20X) "Kepler" accelerators all connected by the Cray Gemini torus interconnect.
- Total Peak Performance: 13.34 PF
- Total System Memory: 1.634 PB
- Total Usable Storage: 26.4 PB

Setup overview

- PanDA server over MySQL, Production System and Monitoring deployed at JINR Cloud service
- Data delivered to BW manually via Globus Online
- No CVMFS, software installed in the project directory
- Pilots run on dedicated node with 32 CPUs: up to 100 processes, pilots execution and amount controlled by Python daemon
- X509 proxy delivered each 24 hours from PanDA server at JINR because there is no VOMS clients on BW

Solved issues

- PanDA server was upgraded in order to increase jobs dispatch rate from 1 per minute to 500 per minute in bulk mode

- Pilots are consuming CPU resources and, when run on login node, being removed by watcher. In order to get rid of that, pilots are now run on a MOM node, shared node for submissions management

- Archiving of logs at Pilot side was removed in order to reduce CPU consumption

- COMPASS calibration database has to run with jobs on the same node since there is no commutation between worker nodes during execution

Jobs submission tuning

- Pilot can work stable with 512 jobs
- If PanDA server replies that there is no jobs, smaller submission is prepared
- Production jobs run up to 18 hours, depending on number of events in the raw file
- Merging of production job results run 1 hour
- Merging of histograms runs 30 hour
- Merging of event dumps runs less than 30 minutes
- In order to avoid requesting excessive resources, three queues were defined: long for processing, shorter for merging of job results and short for histogram and event dumps merging

System performance

Job attribute summary Sort by count, alpha	
attemptnr (8)	1 (18) 4 (1913) 5 (2823) 6 (7831) 7 (11104) 8 (10595) 9 (3343) 10 (708)
computingsite (1)	BW_COMPASS_MCORE (38335)
destinationse (1)	local (38335)
jobstatus (7)	activated (4292) failed (4) finished (6679) holding (65) running (25201) starting (2093) transferring (1)
minramcount (1)	0-1GB (38335)
priorityrange (2)	1000:1099 (18) 2000:2099 (38317)
prodsourcelabel (1)	prod_test (38335)
production (1)	dy2015W07t5BW (38317)

System performance

3885773.bw	petrosya	normal	SAGA-Python-PBSJ	29778	16	512	--	24:00:00	R	07:40:07
3885779.bw	petrosya	normal	SAGA-Python-PBSJ	17154	16	512	--	24:00:00	R	02:02:20
3887209.bw	petrosya	normal	SAGA-Python-PBSJ	22097	16	512	--	18:00:00	R	15:51:38
3888162.bw	petrosya	normal	SAGA-Python-PBSJ	32692	16	512	--	18:00:00	R	05:18:11
3888276.bw	petrosya	normal	SAGA-Python-PBSJ	--	16	512	--	18:00:00	Q	--
3888278.bw	petrosya	normal	SAGA-Python-PBSJ	--	16	512	--	18:00:00	Q	--
3888281.bw	petrosya	normal	SAGA-Python-PBSJ	--	16	512	--	18:00:00	Q	--
3888282.bw	petrosya	normal	SAGA-Python-PBSJ	--	16	512	--	18:00:00	Q	--
3888283.bw	petrosya	normal	SAGA-Python-PBSJ	--	16	512	--	18:00:00	Q	--
3888286.bw	petrosya	normal	SAGA-Python-PBSJ	--	16	512	--	18:00:00	Q	--
3888289.bw	petrosya	normal	SAGA-Python-PBSJ	--	16	512	--	18:00:00	Q	--
3888290.bw	petrosya	normal	SAGA-Python-PBSJ	--	16	512	--	18:00:00	Q	--
3888291.bw	petrosya	normal	SAGA-Python-PBSJ	--	16	512	--	18:00:00	Q	--
3888292.bw	petrosya	normal	SAGA-Python-PBSJ	--	16	512	--	18:00:00	Q	--
3888295.bw	petrosya	normal	SAGA-Python-PBSJ	--	16	512	--	18:00:00	Q	--
3888297.bw	petrosya	normal	SAGA-Python-PBSJ	--	16	512	--	18:00:00	Q	--
3888299.bw	petrosya	normal	SAGA-Python-PBSJ	--	16	512	--	18:00:00	Q	--
3888300.bw	petrosya	normal	SAGA-Python-PBSJ	--	16	512	--	18:00:00	Q	--
3888301.bw	petrosya	normal	SAGA-Python-PBSJ	--	16	512	--	18:00:00	Q	--
3888304.bw	petrosya	normal	SAGA-Python-PBSJ	--	16	512	--	18:00:00	Q	--
3888307.bw	petrosya	normal	SAGA-Python-PBSJ	--	16	512	--	18:00:00	Q	--
3888308.bw	petrosya	normal	SAGA-Python-PBSJ	--	16	512	--	18:00:00	Q	--

System performance

```
top - 03:26:03 up 21 days, 11:27,  1 user,  load average: 16.98, 18.98, 19.47
Tasks: 677 total,  19 running, 658 sleeping,   0 stopped,   0 zombie
Cpu(s): 45.1%us,  9.8%sy,  0.0%ni, 44.9%id,  0.2%wa,  0.0%hi,  0.1%si,  0.0%st
Mem:     64624M total,    46480M used,    18143M free,       13M buffers
Swap:        0M total,        0M used,        0M free,    28803M cached

  PID USER      PR  NI  VIRT  RES  SHR S   %CPU %MEM    TIME+  COMMAND
22317 petrosya  20   0 1239m  60m 6520 R    101  0.1  16:51.14 python
28517 petrosya  20   0 1239m  60m 6516 R    100  0.1  31:27.62 python
30344 petrosya  20   0 1239m  60m 6516 R    100  0.1  36:54.66 python
 3184 petrosya  20   0 1239m  60m 6520 R    100  0.1  41:00.07 python
28519 petrosya  20   0 1240m  60m 6536 R    100  0.1  17:54.20 python
30225 petrosya  20   0 1303m  59m 6520 R    100  0.1  67:38.51 python
  594 petrosya  20   0 1302m  60m 6516 R    100  0.1 135:29.71 python
 6902 petrosya  20   0 1239m  61m 6520 R    100  0.1 134:55.77 python
17423 petrosya  20   0 1303m  59m 6520 R    100  0.1 100:24.62 python
17962 petrosya  20   0 1239m  60m 6524 R    100  0.1  45:07.43 python
20174 petrosya  20   0 1239m  60m 6516 R    100  0.1  31:30.95 python
16537 petrosya  20   0 1240m  59m 6520 R     99  0.1  15:34.36 python
 6937 petrosya  20   0 1240m  60m 6520 R     92  0.1  58:02.83 python
 7532 petrosya  20   0 1231m  54m 6496 R     57  0.1   9:18.13 python
 4950 petrosya  20   0 1293m  51m 6496 R     34  0.1  62:21.43 python
 6906 petrosya  20   0 1229m  52m 6500 R     33  0.1  57:43.53 python
 7609 petrosya  20   0 1293m  51m 6496 S     30  0.1  53:56.00 python
 5813 petrosya  20   0 1290m  48m 6508 S     20  0.1   5:24.34 python
31766 petrosya  20   0 1293m  52m 6496 S     12  0.1  55:38.07 python
10889 petrosya  20   0 1293m  51m 6496 S     11  0.1  58:02.88 python
23805 petrosya  20   0 1293m  51m 6496 S     10  0.1  60:55.16 python
28951 petrosya  20   0 1293m  51m 6500 S     10  0.1  59:03.84 python
22460 petrosya  20   0 1293m  51m 6496 R      8  0.1  64:52.34 python
18594 root       0 -20     0    0    0 S      7  0.0 148:34.66 kgnilnd_sd_00
18595 root       0 -20     0    0    0 S      7  0.0 147:25.54 kgnilnd_sd_01
31351 petrosya  20   0 1290m  48m 6508 S      7  0.1   5:45.84 python
18596 root       0 -20     0    0    0 S      7  0.0 148:21.28 kgnilnd_sd_02
27177 petrosya  20   0 1226m  49m 6508 S      7  0.1   5:24.06 python
25964 petrosya  20   0 1291m  49m 6508 S      6  0.1   5:44.03 python
32322 petrosya  20   0 1290m  49m 6508 S      6  0.1   6:00.73 python
13006 petrosya  20   0 1291m  50m 6508 R      6  0.1   5:25.04 python
14155 petrosya  20   0 1291m  49m 6512 S      5  0.1   5:07.45 python
20503 petrosya  20   0 1293m  52m 6496 S      5  0.1  56:00.92 python
```

Summary

- ProdSys runs COMPASS production jobs via PanDA on Blue Waters
- Environment for automated data processing on BW was prepared and runs reliably in daemon mode
- Further development
  - Upgrade to PanDA Harvester will allow to consume more resources with higher level of stability and efficiency
  - Enable automatic data stage in and stage out to and from BW