Status of IHEP distributed computing

Xiaomei ZHANG Xianghu Zhao
Institute of High Energy Physics

9th DIRAC User Workshop

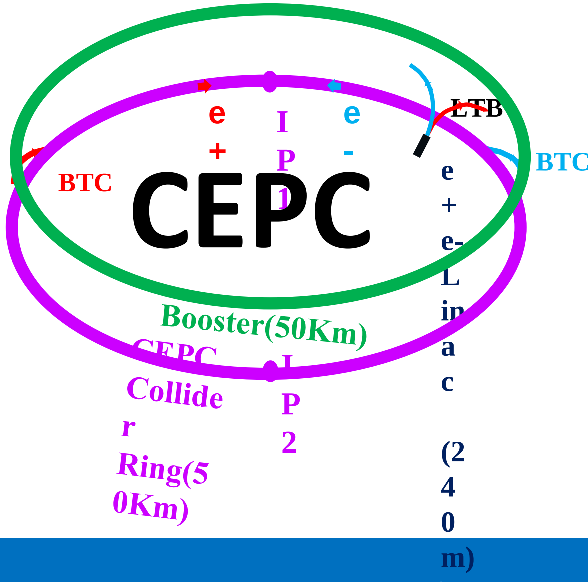
London,U.K.

Reminder

- Distributed Computing (DC) in IHEP was first built for BESIII in 2012
  - Meet peek need of BESIII computing with ~3PB/5year in total
  - Put into production for BESIII in 2014
- Evolve into a general platform for multi experiments in 2016
  - JUNO : operate in prototype
  - CEPC : in production for R&D phase

BESIII (Beijing Spectrometer III at BEPCII)

JUNO (Jiangmen Underground Neutrino Observatory)



DIRAC set-up and upgrade

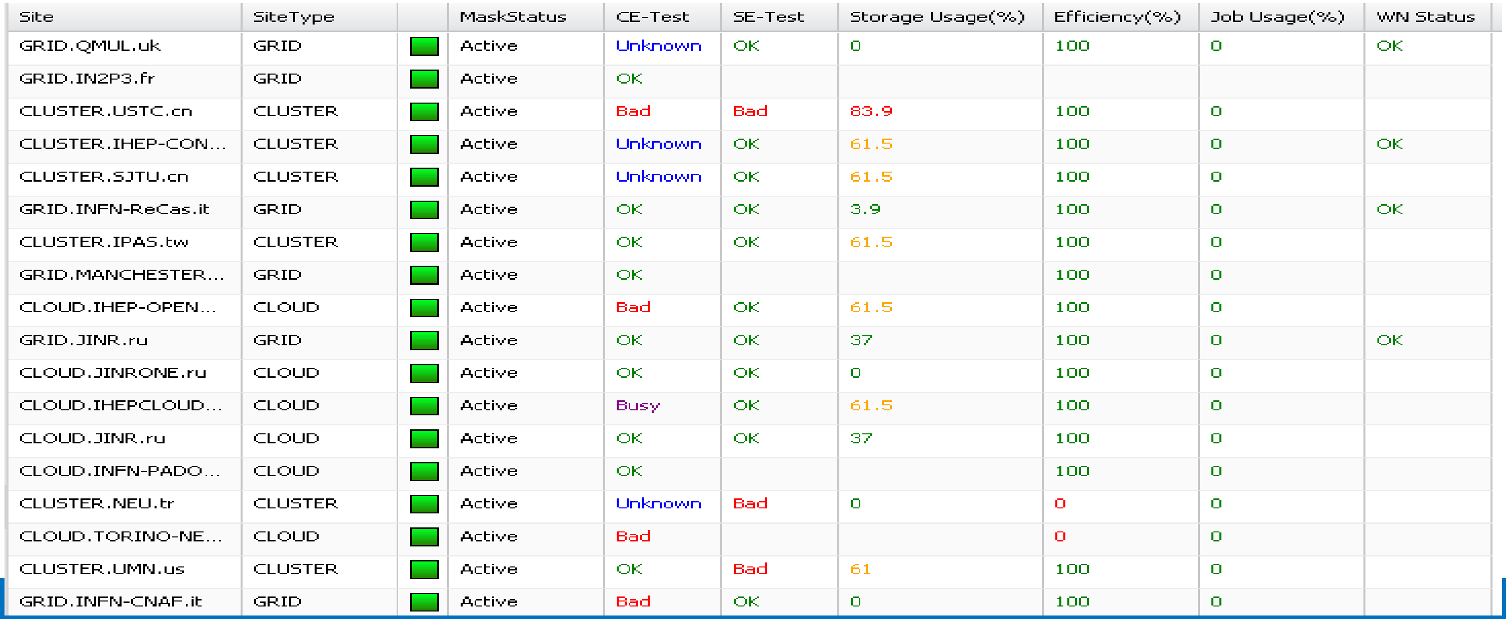
- Three set-up: production, test, development
- Production set-up
  - One main CS server, One DB server
  - Intel(R) Xeon(R) CPU E5-2630 v3 @ 2.40GHz with 16 cores, 64GB Mem with 2.5TB
  - DIRAC version: v6r20p6, VMDIRAC: v2r3-pre1
  - Plan to upgrade to v6r21 soon in new machines
    - Intel Xeon Silver 4116 CPU @ 2.10GHz with 24 cores， 128GB DDR4-2400 Memory，4TB SAS disk
- Development and test set-up
  - Two VMs
  - DIRAC version: v6r21p2

DIRAC functionalities in use

- WMS
  - Cluster: HTCondor, SLURM, PBS
  - Grid: CreamCE
  - Cloud: OpenNebula, OpenStack
- DIRAC File Catalog (DFC)
  - Use as replica and meta-data catalog for besIII
  - Plan to use as replica catalog for juno （juno has developed its own bookkeeping）
- RMS + Transformation (TS)
  - Use as bulk replication/removal for JUNO (in prototype)
- Production system over TS
  - Organize MC Simulation and reconstruction workflow and dataflow for JUNO (in plan)
- Monitoring and Accounting
- Multi-VO

Extensions in use

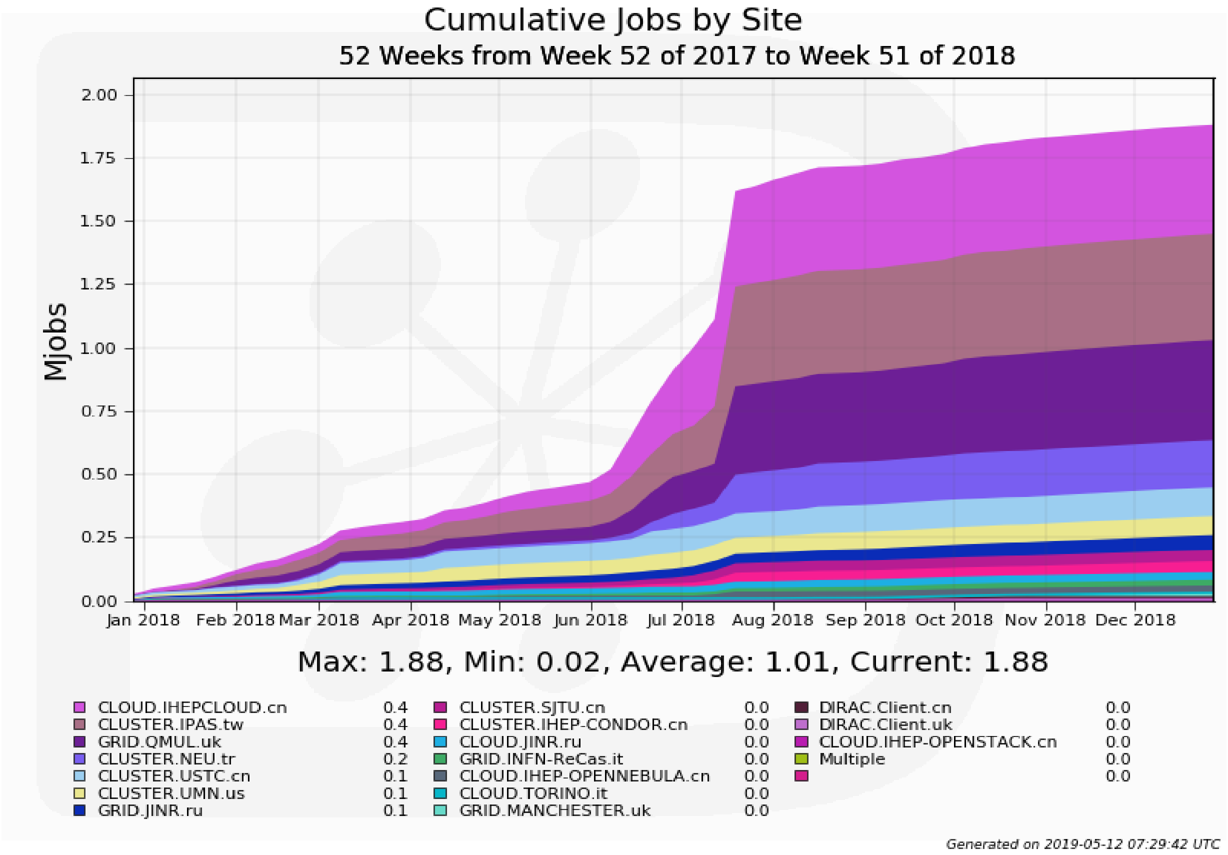
- DIRAC Extensions
  - WebAppDIRAC
  - VMDIRAC to manage cloud from VMDIRACv1.0
- BESIII/JUNO Extensions
  - Task Manager to manage experiment tasks over job monitoring
  - Site Monitoring to monitor site status
    - Send and track SAM tests jobs
    - Summarize user jobs



Site	SiteType		MaskStatus	CE-Test	SE-Test	Storage Usage(%)	Efficiency(%)	Job Usage(%)	WN Status
GRID.QMUL.uk	GRID	🟩	Active	Unknown	OK	0	100	0	OK
GRID.IN2P3.fr	GRID	🟩	Active	OK					
CLUSTER.USTC.cn	CLUSTER	🟩	Active	Bad	Bad	83.9	100	0	
CLUSTER.IHEP-CON...	CLUSTER	🟩	Active	Unknown	OK	61.5	100	0	OK
CLUSTER.SJTU.cn	CLUSTER	🟩	Active	Unknown	OK	61.5	100	0	
GRID.INFN-ReCas.it	GRID	🟩	Active	OK	OK	3.9	100	0	OK
CLUSTER.IPAS.tw	CLUSTER	🟩	Active	OK	OK	61.5	100	0	
GRID.MANCHESTER...	GRID	🟩	Active	OK			100	0	
CLOUD.IHEP-OPEN...	CLOUD	🟩	Active	Bad	OK	61.5	100	0	
GRID.JINR.ru	GRID	🟩	Active	OK	OK	37	100	0	OK
CLOUD.JINRONE.ru	CLOUD	🟩	Active	OK	OK	0	100	0	
CLOUD.IHEPCLOUD...	CLOUD	🟩	Active	Busy	OK	61.5	100	0	
CLOUD.JINR.ru	CLOUD	🟩	Active	OK	OK	37	100	0	
CLOUD.INFN-PADO...	CLOUD	🟩	Active	OK			100	0	
CLUSTER.NEU.tr	CLUSTER	🟩	Active	Unknown	Bad	0	0	0	
CLOUD.TORINO-NE...	CLOUD	🟩	Active	Bad			0	0	
CLUSTER.UMN.us	CLUSTER	🟩	Active	OK	Bad	61	100	0	
GRID.INFN-CNAF.it	GRID	🟩	Active	Bad	OK	0	100	0	

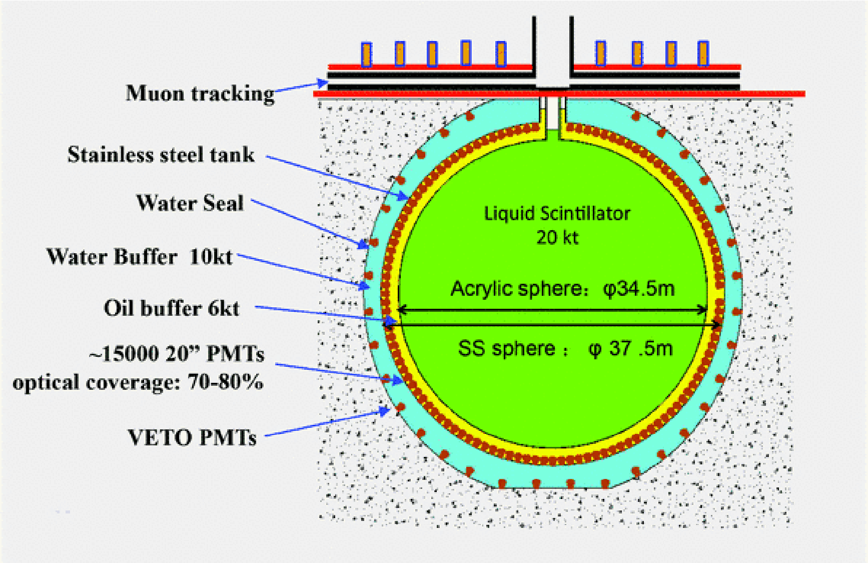
Production status

- About 1.88M jobs running in 2018



Jiangmen Underground Neutrino Observatory

- JUNO, a multi-purpose neutrino experiment designed to measure the neutrino mass hierarchy and mixing parameters
  - Start to build in 2014, operational in 2021, located at Guangzhou province
  - Estimated to produce 2PB data/year for >10 years
  - 20 kt Liquid Scintillator detector, 700m deep underground
  - 2-3% energy resolution

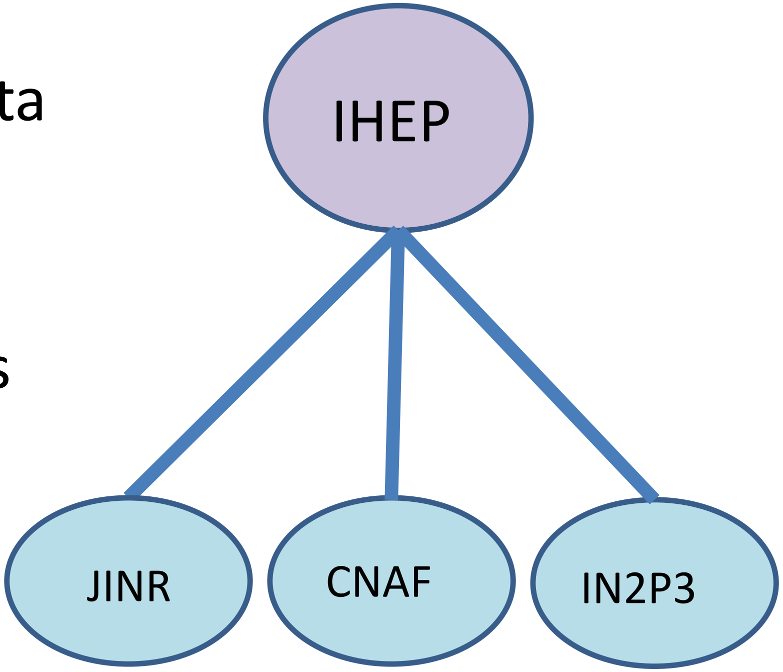


JUNO computing

- No much challenges with normal sim, rec, ana
  - Rec data ~20TB/year , Sim data ~100TB/year, Raw data ~2PB/year
  - ~10000 cores can meet the goal
- The system is planned to be designed in a reliable and simple way
  - Use DIRAC WMS and DMS services as much as possible
- Simulation of optical photons produced by muon is a special one, which poses severe constraints on both CPU time and Memory
  - CPU time >95%, memory ~8GB, > 2 hours/event
- This problem pushes JUNO to explore ways of parallelism
  - Performance optimization using parallelism with Geant4 10.x
  - Massive parallelism with GPU, achieving 1000 times speed-up
  - Fast simulation, and plan to improve it with Machine Learning
- Support of Multi-core and GPU were considered

JUNO computing model

- IHEP as central centre
  - Play major role for raw data reconstruction, calibration, simulation and analysis ….
  - Hold central storage for all the data
- European centers
  - Hold one copy of raw data
  - Mainly for simulation and analysis
  - Possibly share efforts on reconstruction

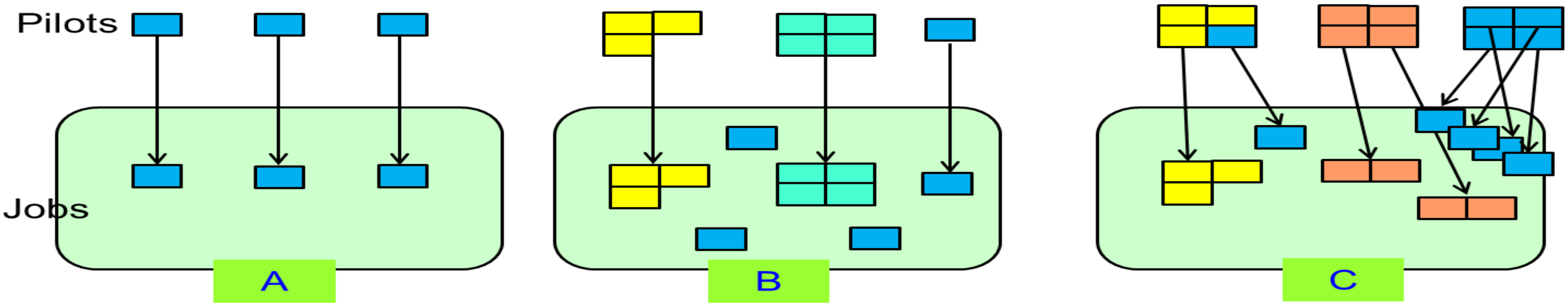


JUNO Data transfer

- Bulk transfers are required among SE of data centers
- Solution based on DIRAC DMS was tried and tested
  - DFC: register files
  - Request manager (RMS): manage queues and interface with FTS
  - Transformation: produce bulk transfers or removals with a list of files
- Testing is fine, but can be improved in some ways
  - Priority control
  - Errors tracking is not so direct
    - There are no direct links to the related FTS logs from RMS
    - Errors seen from TS and RMS is difficult to track
      - Only look into agent log
    - Better to have ways to know more details of files failed to transfer
      - Eg. TS only shows 80% complete, what about 20% failed files? Sometimes they are problematic files

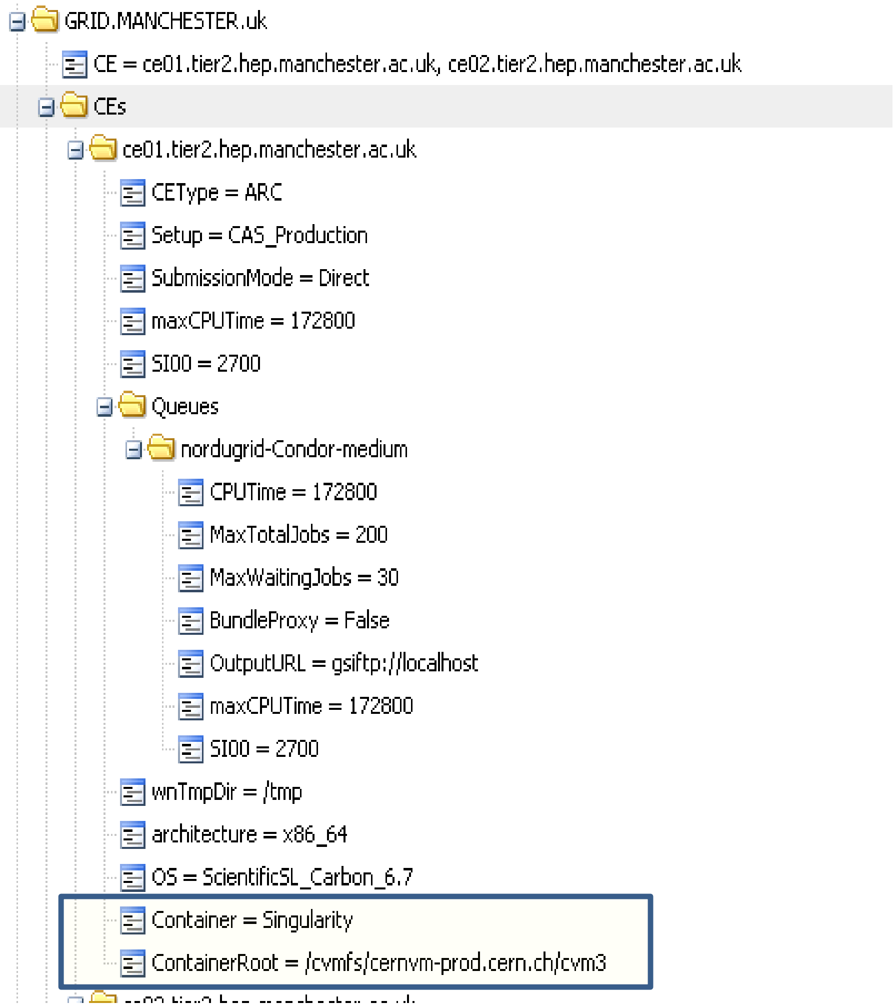
Multicore scheduling

- Juno has tried on multi-core supports
  - Customized and shared partitionable pilot modes
  - More details in Federico's talk "matching jobs to computing resources: MPs + GPU, …"
- The prototype is quite successful, but need more efforts
  - Resource efficiency is concerned if mixture of single-core and multi-core jobs
  - Some changes with core info need to be added to monitoring and accounting
- Next step would like to put it into production if JUNO software is ready for parallelism



Singularity

- Singularity mode was tried and used in production
  - Pilot starts singularity and user jobs runs inside singularity
- Main magic is SingularityCE
  - Developed by Simon Fayer
  - We fixed it with BESIII user cases
- Why need to use Singularity
  - Site OS is not consistent with the OS required by experiment
  - Users ask for special Linux OS to run jobs (not done yet)
- Two sites in BESIII are working well with Singularity
  - Manchester
  - IHEP-HTCondor



```
GRID.MANCHESTER.uk
  CE = ce01.tier2.hep.manchester.ac.uk, ce02.tier2.hep.manchester.ac.uk
  CEs
    ce01.tier2.hep.manchester.ac.uk
      CEType = ARC
      Setup = CAS_Production
      SubmissionMode = Direct
      maxCPUTime = 172800
      SI00 = 2700
      Queues
        nordugrid-Condor-medium
          CPUTime = 172800
          MaxTotalJobs = 200
          MaxWaitingJobs = 30
          BundleProxy = False
          OutputURL = gsiftp://localhost
          maxCPUTime = 172800
          SI00 = 2700
      wnTmpDir = /tmp
      architecture = x86_64
      OS = ScientificSL_Carbon_6.7
      Container = Singularity
      ContainerRoot = /cvmfs/cernvm-prod.cern.ch/cvm3
    ce02.tier2.hep.manchester.ac.uk
```

Use GPU with tags

- Testbed has been setup to try out
  - IHEP GPU farm (SLURM)
  - JINR GPU farm (SLURM)
- With new “Tags” system, CPU and GPU jobs can be successfully sent to right sites
  - “RequiredTag” and “Tag”
- But do we need more tags for matching if jobs required more?
  - GPU jobs have more requirements than CPU jobs
    - GPUModel (CudaDevice), GPUNumber (Request_gpu=1), CudaCapability (GPUVersion >=3), Memory

Summary

- IHEP distributed computing is in good status, but not much challenges in production now
- Recent work would be more focused on JUNO
  - Design of JUNO computing model
  - Production system in plan to manage its workflow and dataflow
  - Multi-core and GPU supports for parallelism of JUNO software