Growth and Evolution CMS Offline Computing from Run 1 to HL-LHC

ICHEP 2020

Akanksha Ahuja[1,4], Sharad Agarwal[2,4], David Lange[3,4]

on behalf of CMS Collaboration

University of Sofia[1], University of Wisconsin Madison[2], Princeton University[3], CERN[4]

SOFIA UNIVERSITY "St. Kliment Ohridski"

WISCONSIN UNIVERSITY OF WISCONSIN–MADISON

PRINCETON UNIVERSITY

CERN

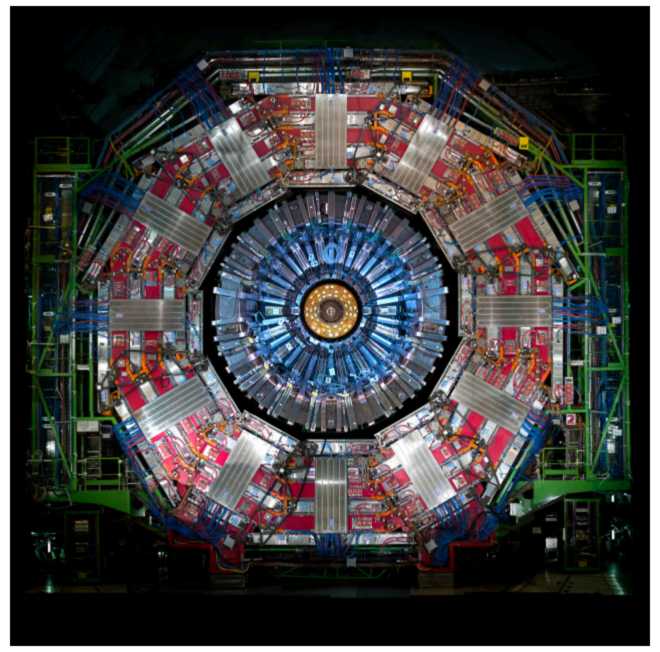
29th July 2020

CMS

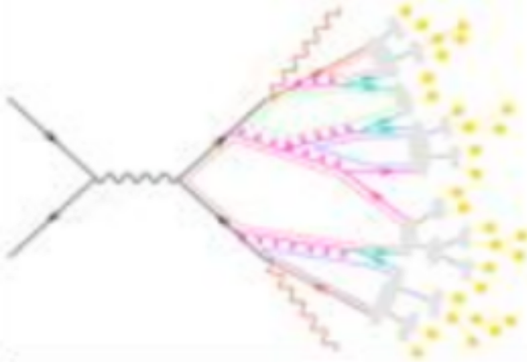
Agenda

- **Introduction to CMS Offline Computing**
- **Growth and Evolution:**
  - **Distributed Grid Computing Infrastructure**
  - **Data Management**
  - **Data Production**

CMS Offline Computing

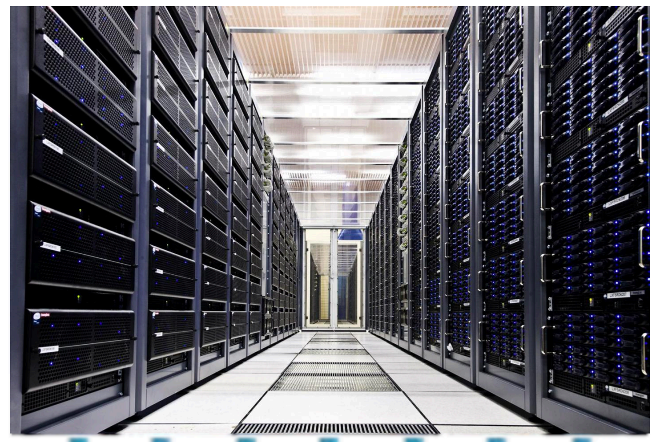


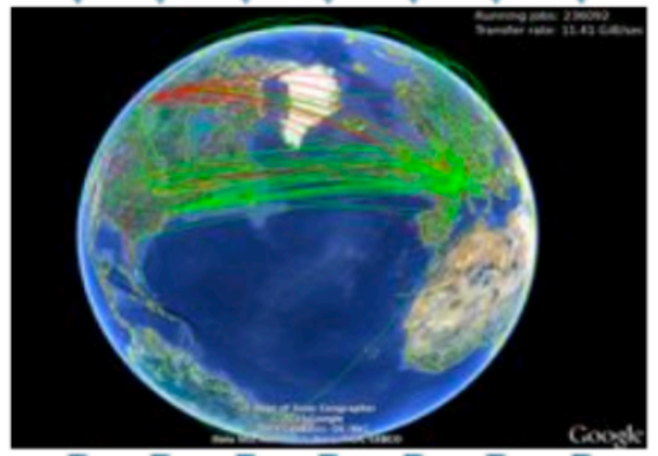
Real Data



Monte Carlo (MC) Simulations

Data Centres

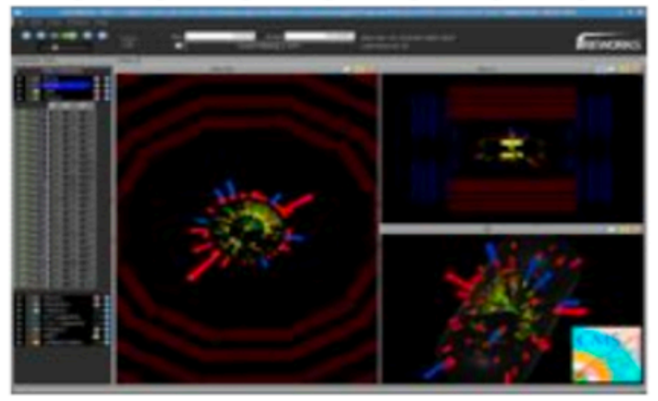




Worldwide LHC Computing Grid (WLCG)

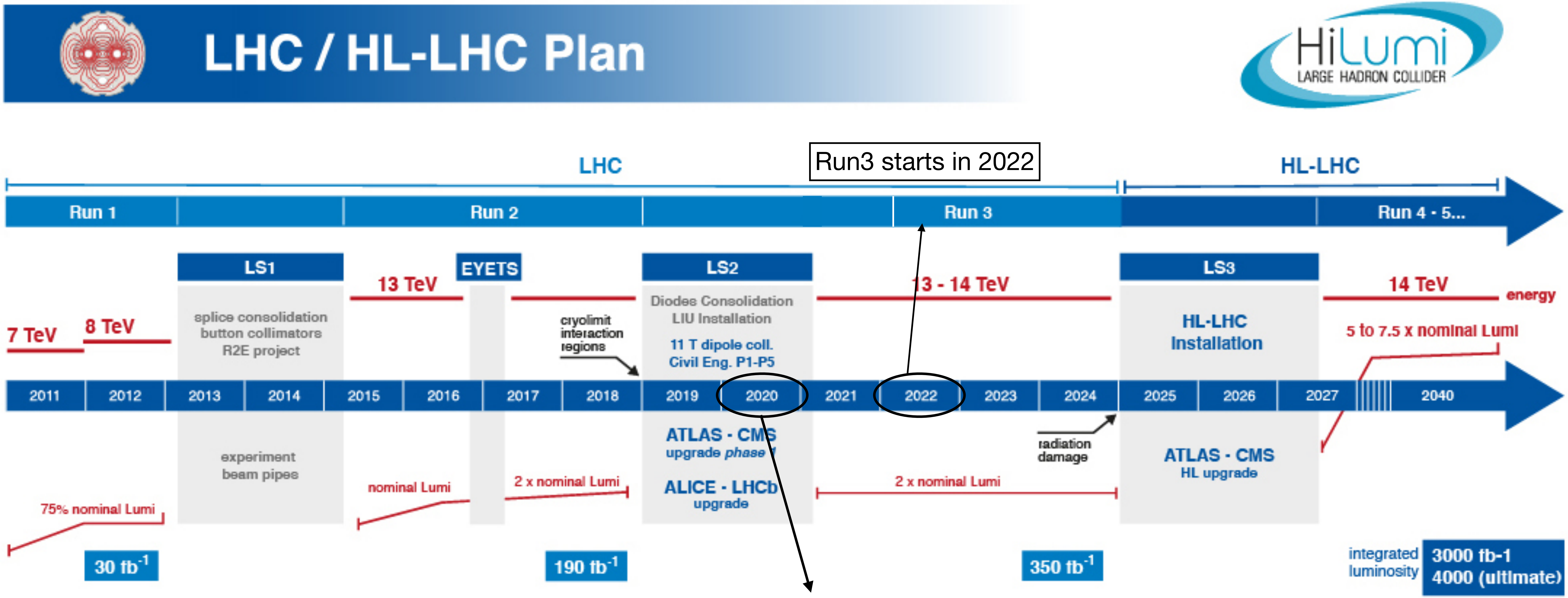
CMS Offline Computing utilises typical 200k cores for MC Simulations and Data Reconstruction

Analysis Results



CMS Offline Computing helps 1000s of CMS users to analyse the CMS data through the WLCG

High Luminosity Large Hadron Collider Plan



Currently we are in LS2

* The timeline has been updated manually as per the latest updates about Run3 which now starts from 2022

Distributed Grid Computing Infrastructure

Run1

gLite, HTCondor_G, GlideinWMS softwares used for scheduling jobs over the WLCG. Started with Grid Mesh Topology.

LS1

Grid Mesh Topology was completed.
XRootD endpoints for AAA.
Introduction to Multi-core pilots.

Run2

Complete Dependency on HTCondor and GlideinWMS.
Switched to Singularity (Decouple from OS at sites).
Advanced xrootd endpoint monitoring.

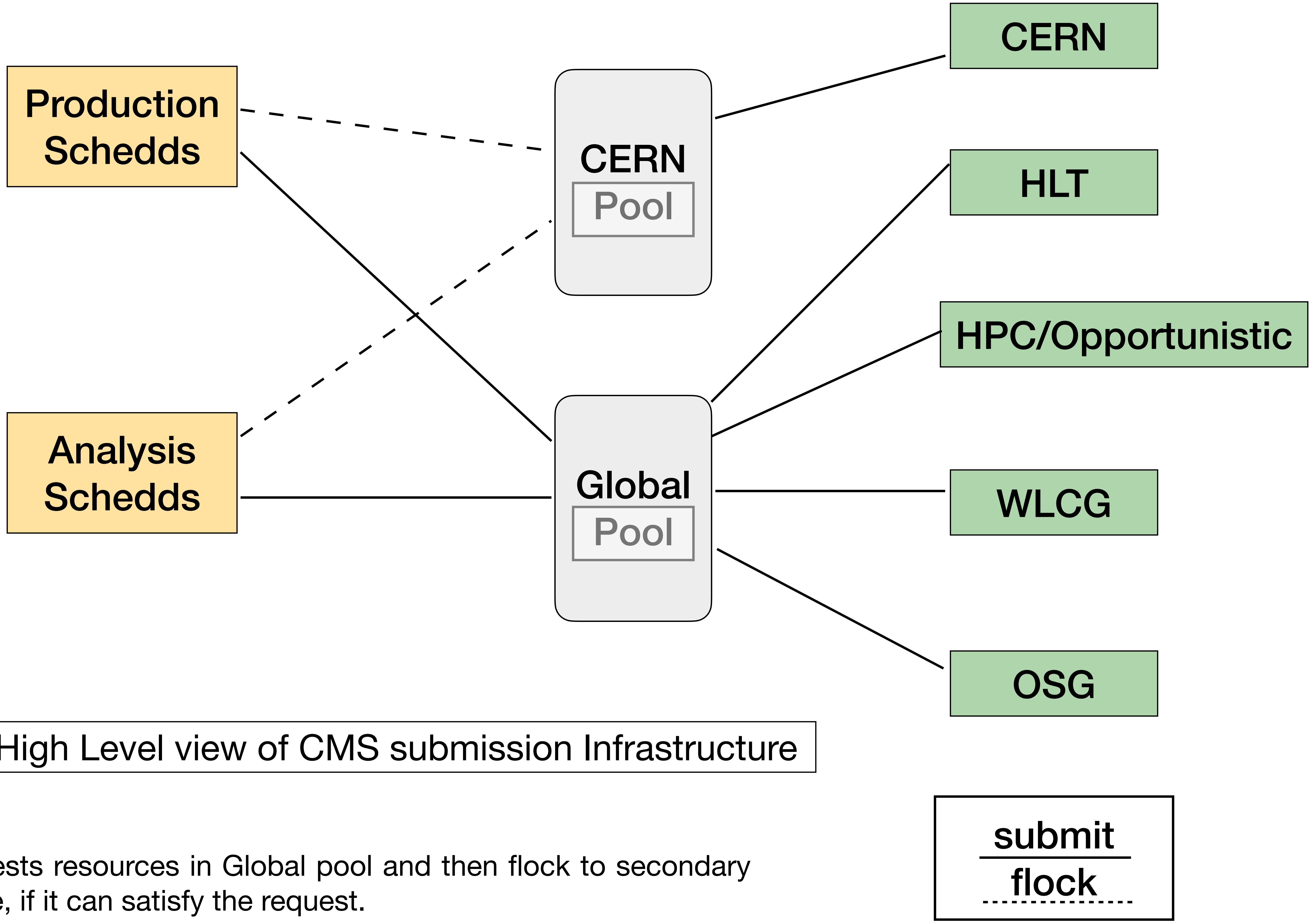
LS2

SiteDB -> CRIC migration.
Advanced monitoring using MonIT/Grafana.
Partitionable slots in Global Pool and start using HPC resources.

Plans for Run3 and HL-LHC

- Improving the usage of HPC resources.
- Heterogenous Computing using hardware accelerators.
- Migration of Certificates to Tokens for authentication.
- Complete the migration of CREAM-CE to HTCONDOR-CE

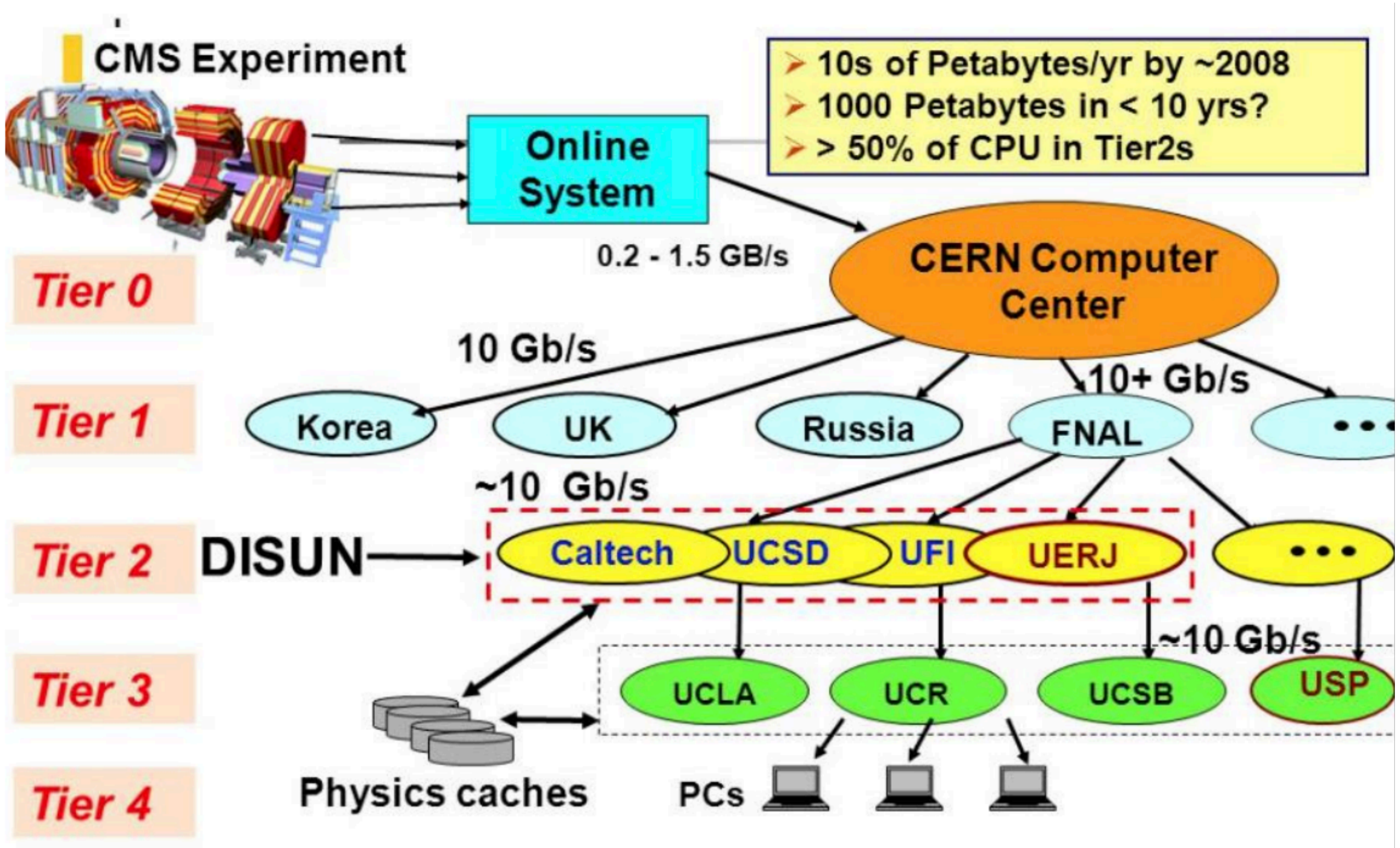
CMS Resource Scheduling



High Level view of CMS submission Infrastructure

* The idle jobs in the Schedds requests resources in Global pool and then flock to secondary resources i.e. CERN pool in this case, if it can satisfy the request.

Hierarchical Model of CMS Grid Sites



MONARCH Model

Original Design Fundamentals

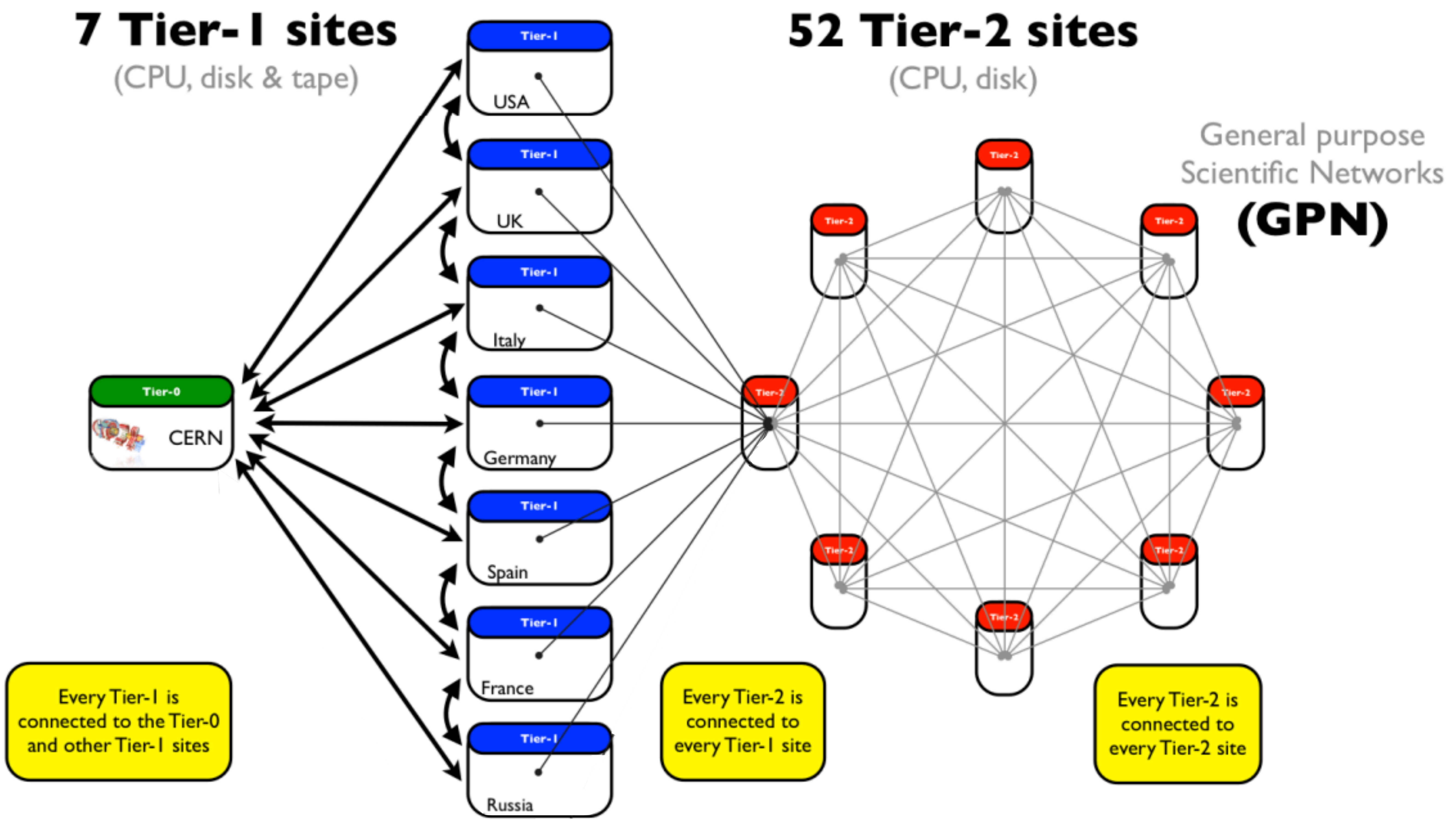
Tier 0: Where the data comes from and is first Reconstructed.

Tier 1s: National Centres, Only for running simulations and data reconstruction

Tier 2s: Regional Centres, Only for analysis

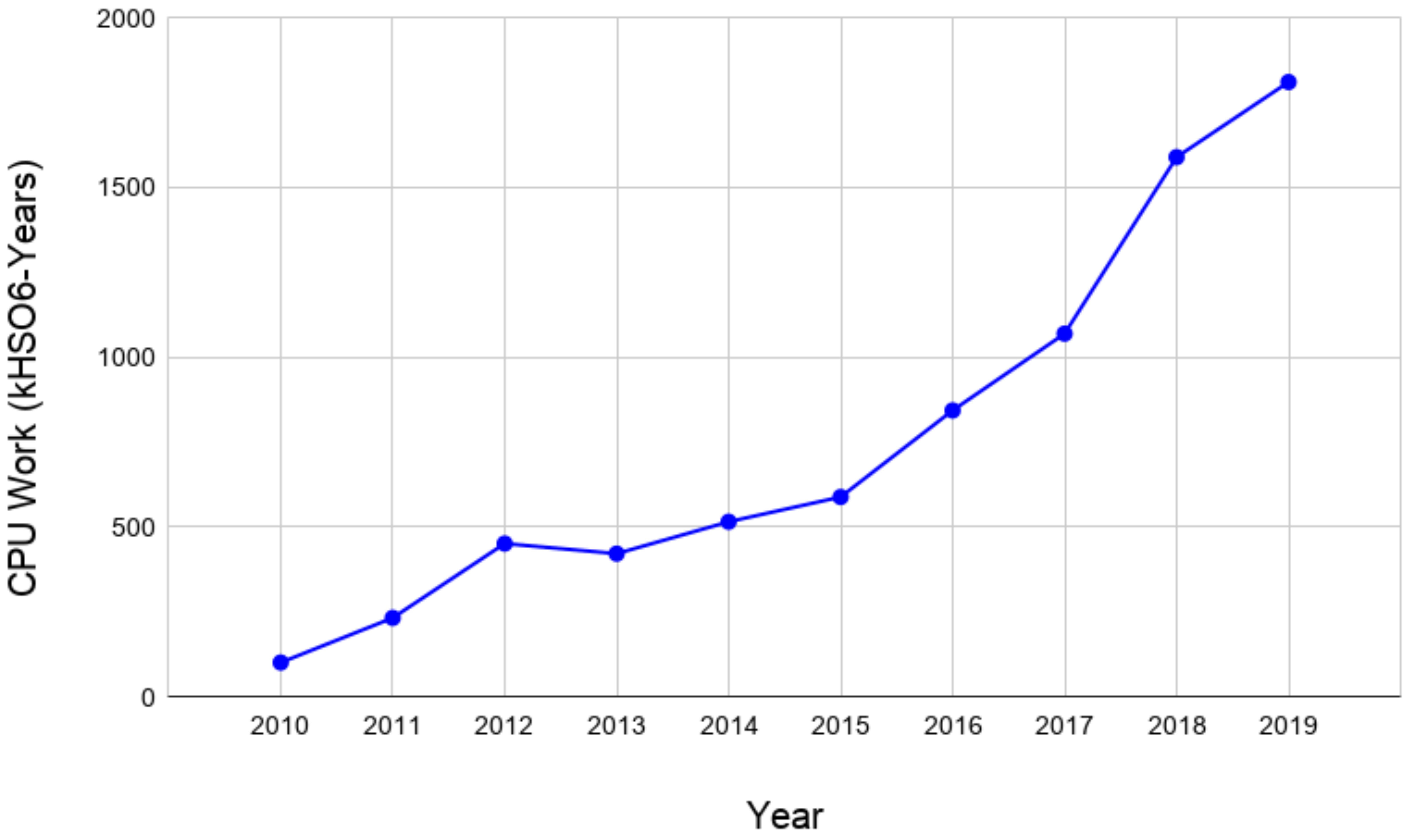
CMS is able to achieve better overall throughput and better resource utilisation with the flexibility in the system.

Mesh Network Topology in CMS



Combined CMS CPU utilisation for T0, T1 and T2 from Run1 to LS2

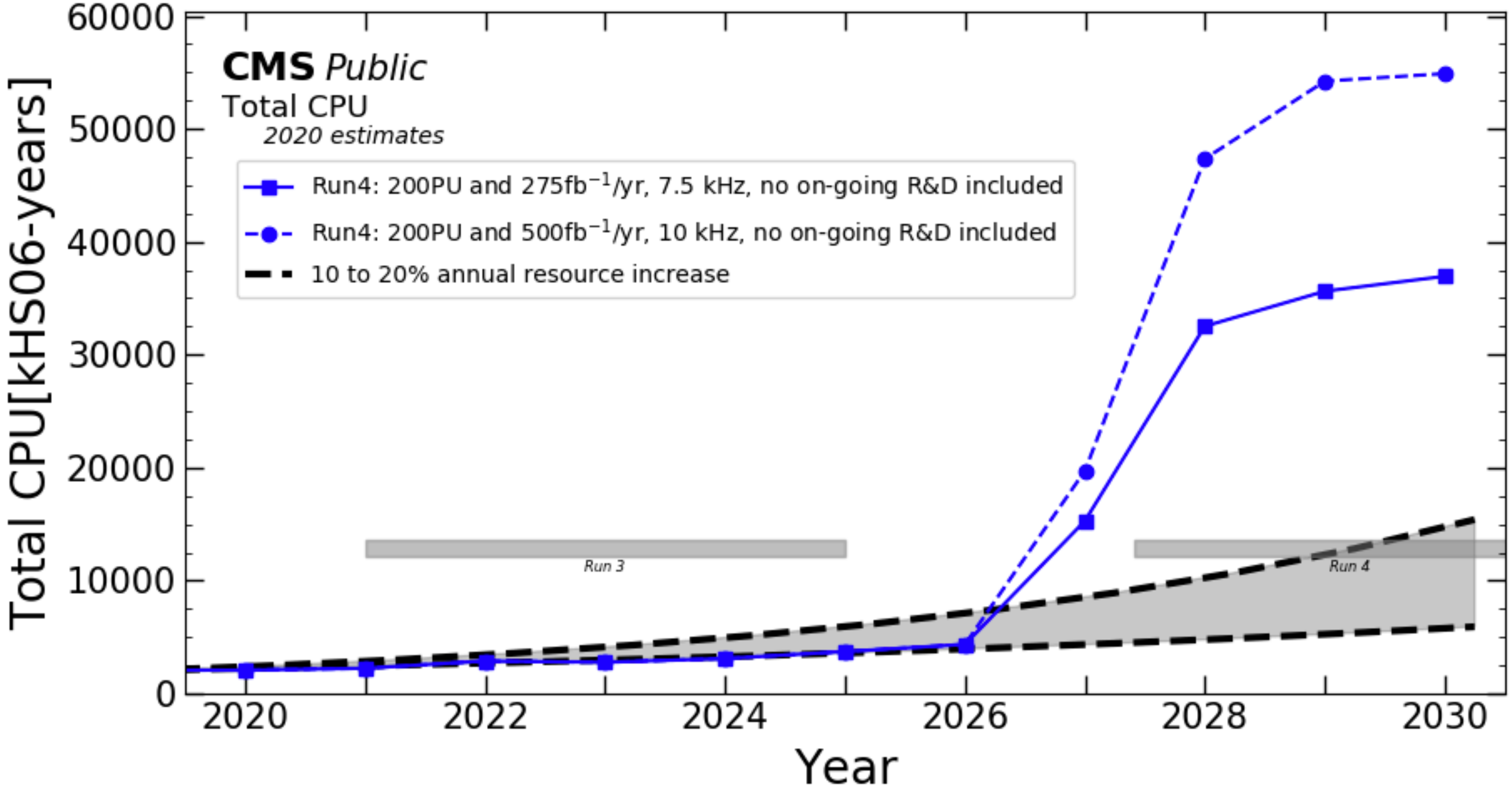
The graph depicts the continuous increase in CMS computing CPU resources.



* The plot has been made using the actual data in the EGI Accounting Portal - https://accounting.egi.eu/

CPU Estimates for Run3 and High Luminosity-Large Hadron Collider (HL-LHC)

The graph estimates the constant increase in CMS CPU resources for Run3 and increases by an order of magnitude for the HL-LHC.



For more deeper insights:

〉 Want to know more?

https://indico.cern.ch/event/868940/contributions/3814459/

ICHEP 2020

Resource provisioning and workload scheduling of CMS Offline Computing

31 Jul 2020, 11:00
20m
virtual conference

Talk | 14. Computing and Da... | Computing and Data Han...

Speaker

Antonio Perez-Calero Yzquierdo (Centro de Investigac...)

Description

The CMS experiment requires vast amounts of computational power in order to generate, process and analyze the data coming from proton-proton collisions at the Large Hadron Collider, as well as Monte Carlo simulations. CMS computing needs have been mostly satisfied up to now by the supporting Worldwide LHC Computing Grid (WLCG), a joint collaboration of more than a hundred computing centers geographically distributed around the world. However, as CMS faces the Run 3 and HL-LHC challenges, with increasing luminosity and event complexity, growing demands for CPU have been estimated. In these future scenarios, additional contributions from more diverse types of resources, such as Cloud and High Performance Computing (HPC) clusters, will be required to complement the limited growth of the capacities of WLCG resources. A number of strategies are being evaluated on how to access and use WLCG and non-WLCG processing capacities as part of a combined infrastructure, successfully exploit an increasingly more heterogeneous pool of resources, efficiently schedule computing workloads according to their requirements and priorities, and timely deliver analysis results to the collaboration, which will be presented in this contribution.

Data Management

Run1

phEDEx for managing data.
Group/Site Managers.

LS1

XRootD

Dynamic Data Management was introduced.
XRootD for AAA.
FTS3 for phEDEx.
Separated Disk and Tape Endpoints for T1s.

Run2

Dynamo - Dynamic Data Management.
Advanced XRootD monitoring.
XRootD write support implemented.

LS2

RUCIO

kubernetes

Grafana

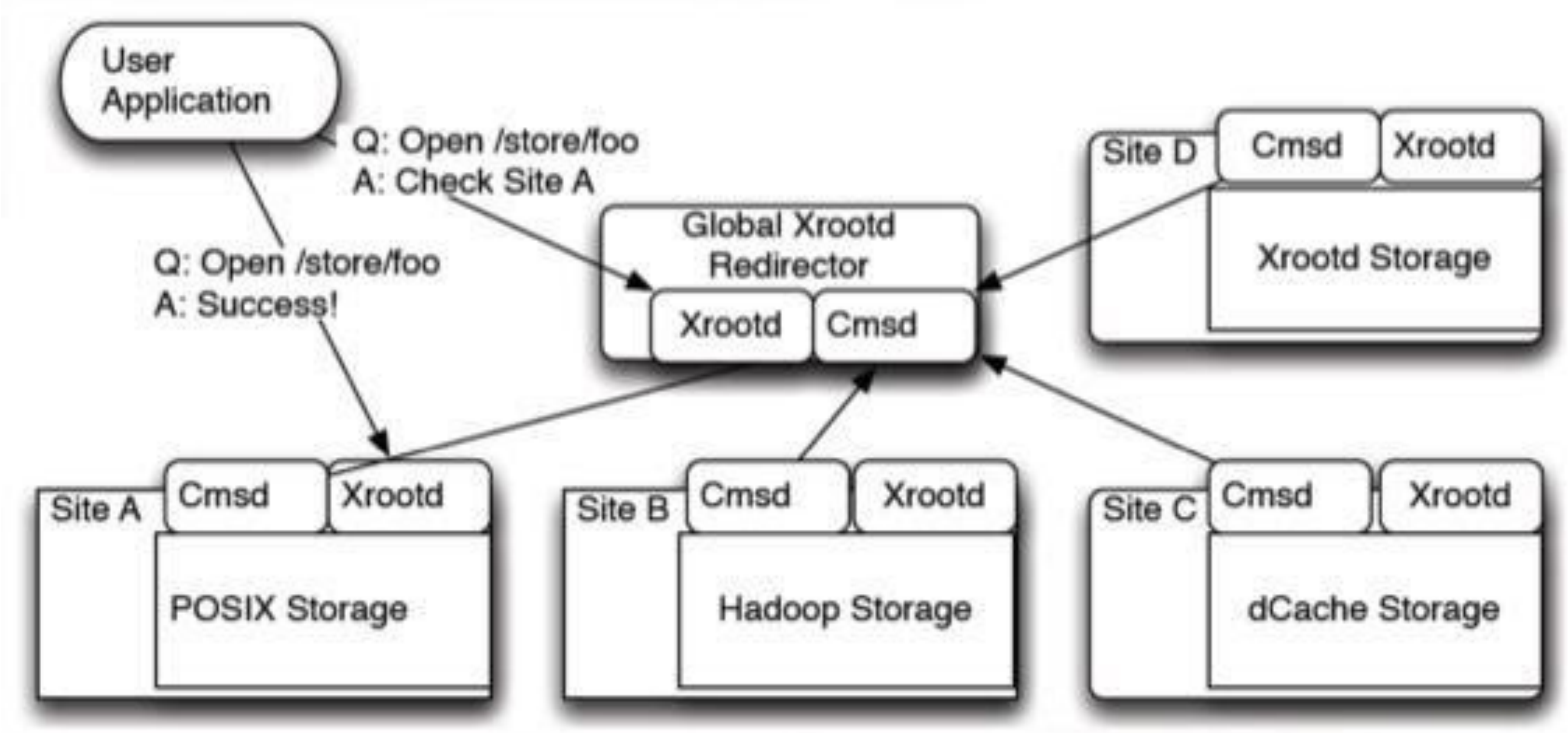
phEDEx -> Rucio Migration & introducing Kubernetes.
Transferring data via XRootD.
Advanced monitoring using MonIT/Grafana.
Started working on operational intelligence.

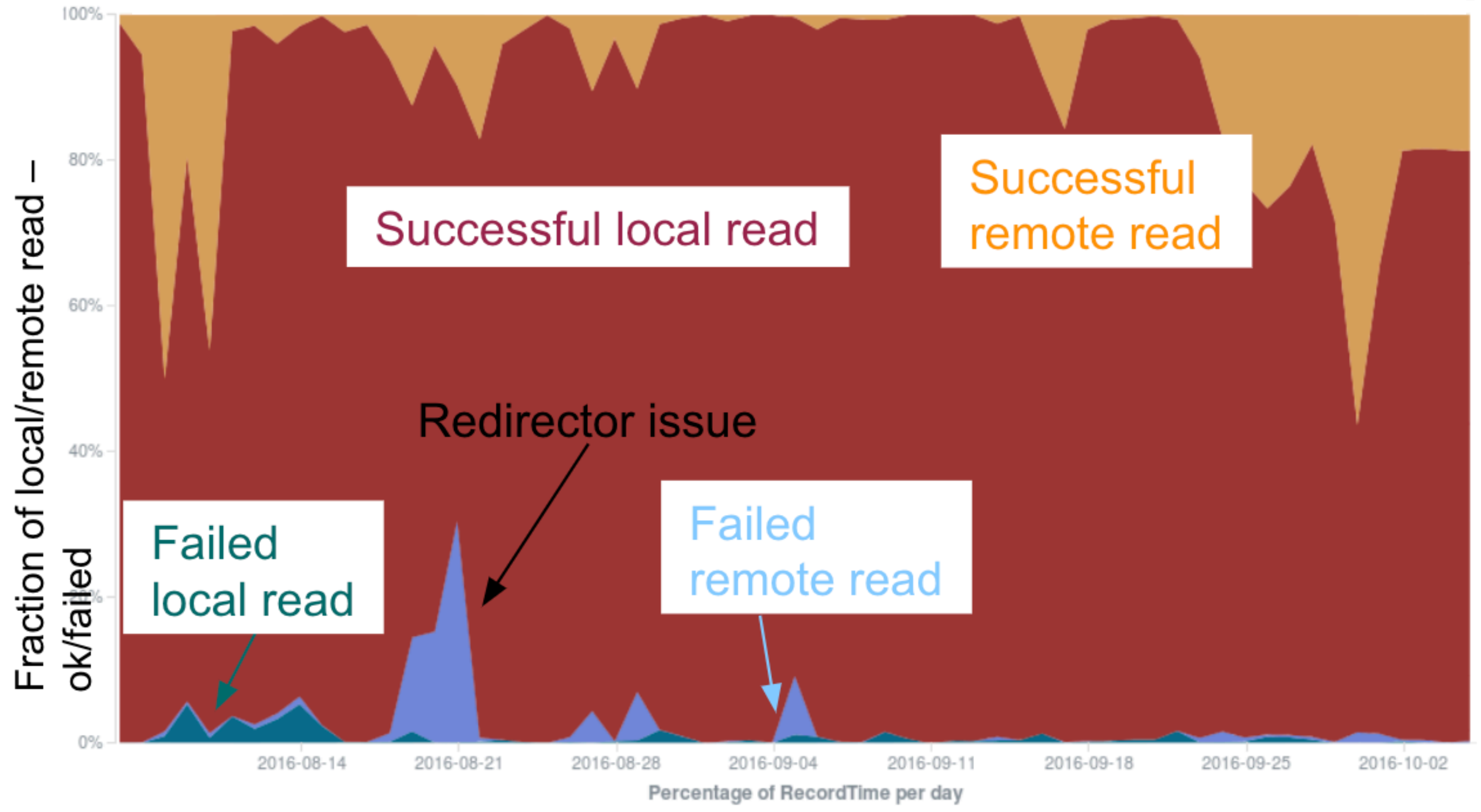
Plans for Run3 and HL-LHC

- Further improve Data Management with more Intelligent algorithms for efficient utilisation.
- Migration of Certificates to Tokens for authentication.
- CMS is working to achieve the same functionality of DDM through Rucio.

Remote Data Access via AAA Storage Federation

- AAA = Any data, Anytime, Anywhere
- Efficient remote data access important for flexibility and increasing throughput.
- CMS application I/O extended to include remote reads.
- Present technology choice
  - XRootD based storage federation
  - Sites “publish” storage inventory to regional re-director
- Central production uses AAA routinely to read input files for Data and MC workflows.
- Physics Analysis, detector commissioning and other users save time.
  - No need to wait for data to be transferred locally before running.





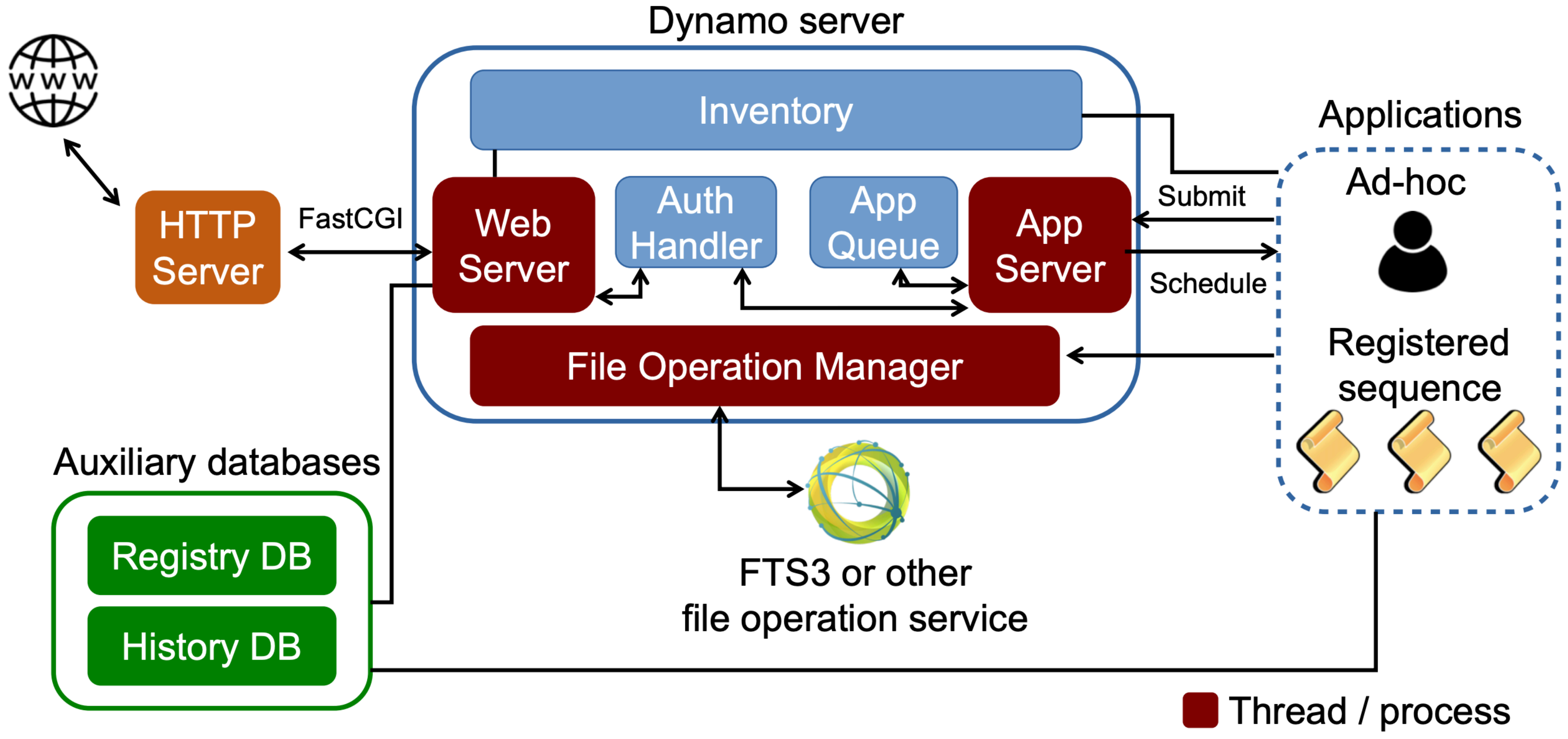
Dynamic Data Management (DDM)

- DDM manages today about 118 PB of disk space
  - All Grid sites (Tier-0, Tier-1s and Tier2s) contribute to the DDM pool

- DDM creates new subscriptions or removes subscriptions based on
  1. Data popularity
     - Access of data is recorded
     - Create more replicas for 'popular' datasets, lower the replication for less popular datasets.

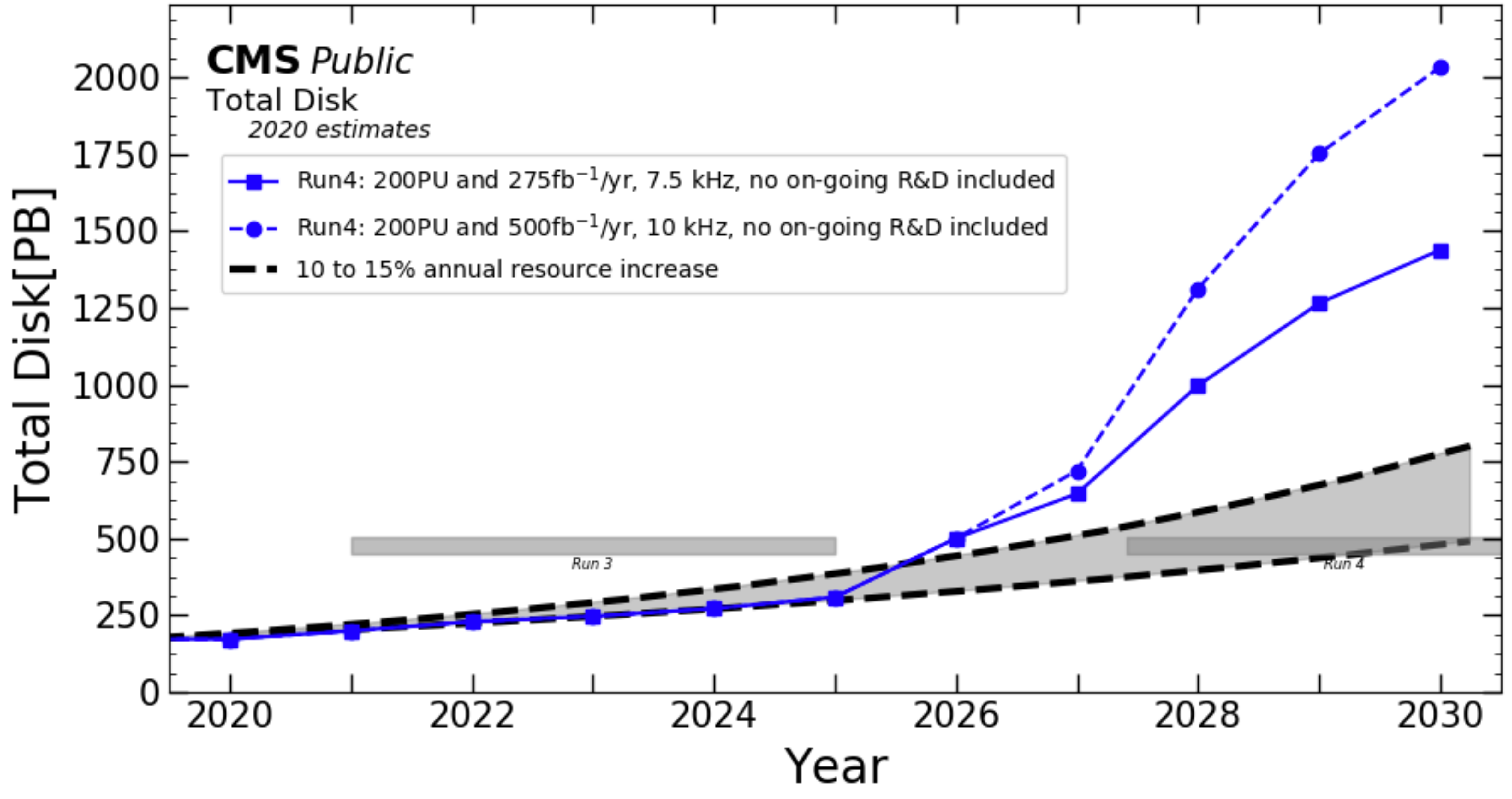
  2. Disk usage level on a given site
     - Keep sites filled at a 'safe' level and always use available disk space.

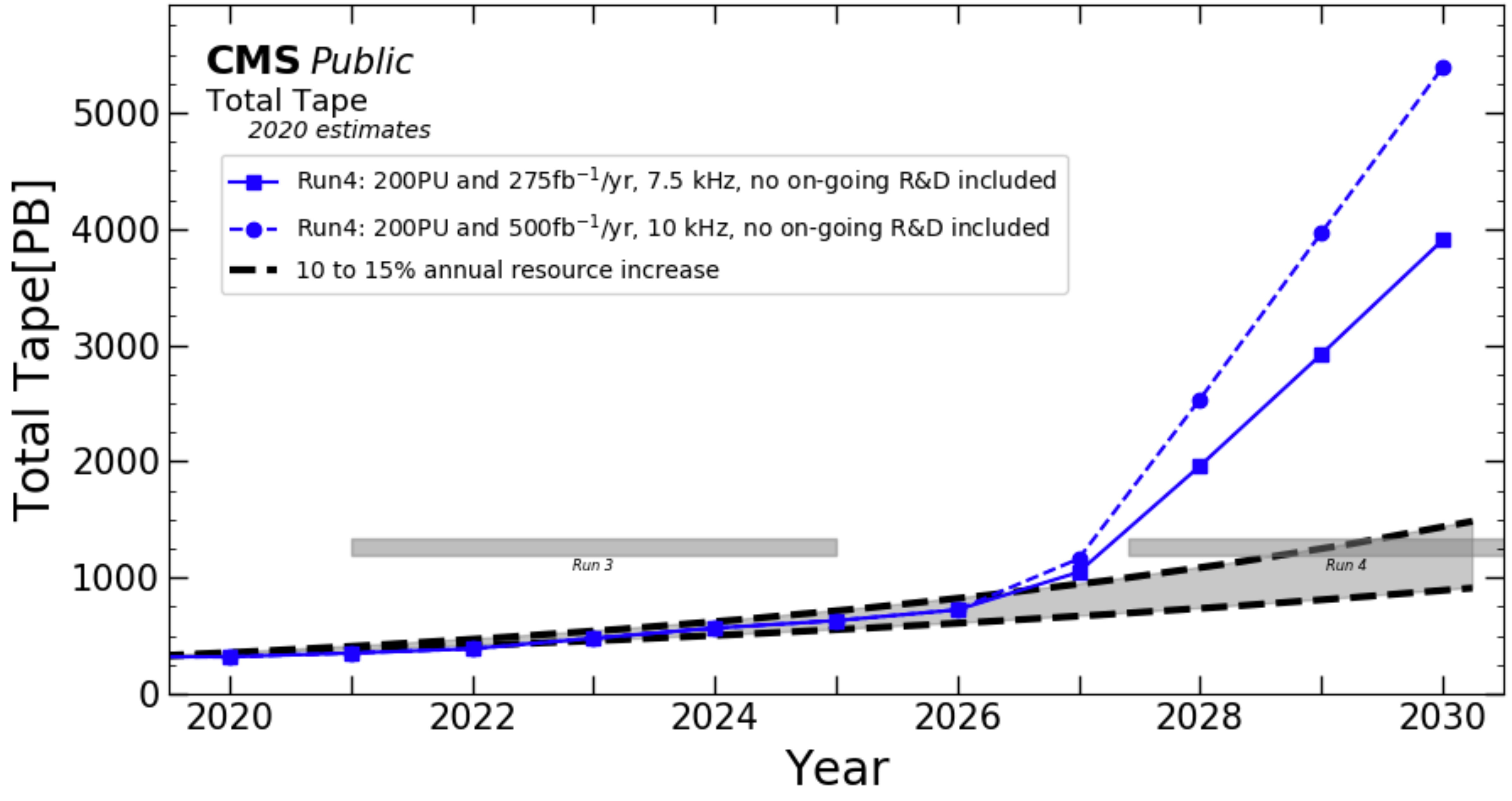
  3. A set of DDM policy rules (examples, actual config my be different!)
     - Keep at least 2 copies of 2016 AOD data.
     - Keep at least 3 copies of MINIAODSIM from main 2016 MC production campaign.
     - Delete RECO datasets from disk after 3 months of lifetime.



Schematic Diagram of Dynamo

Estimates for Run3 and High Luminosity-Large Hadron Collider (HL-LHC)



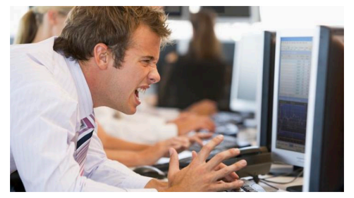


The graphs estimates the constant increase in CMS Storage resources for Run3 and increases by an order of magnitude for the HL-LHC.

Data Production

Run1

Manual Submissions of workflows using Prod Agents.
Data Reconstruction started.



LS1

ProdAgents -> WMAgents.
Remote Data Access allowed decoupling of workflows.

Run2

Recovery system implemented In WMAgents.
Automatic assignment and management of workflows through Unified software.
Introduced StepChain workflows.
Moved towards open source Technologies like MariaDB.

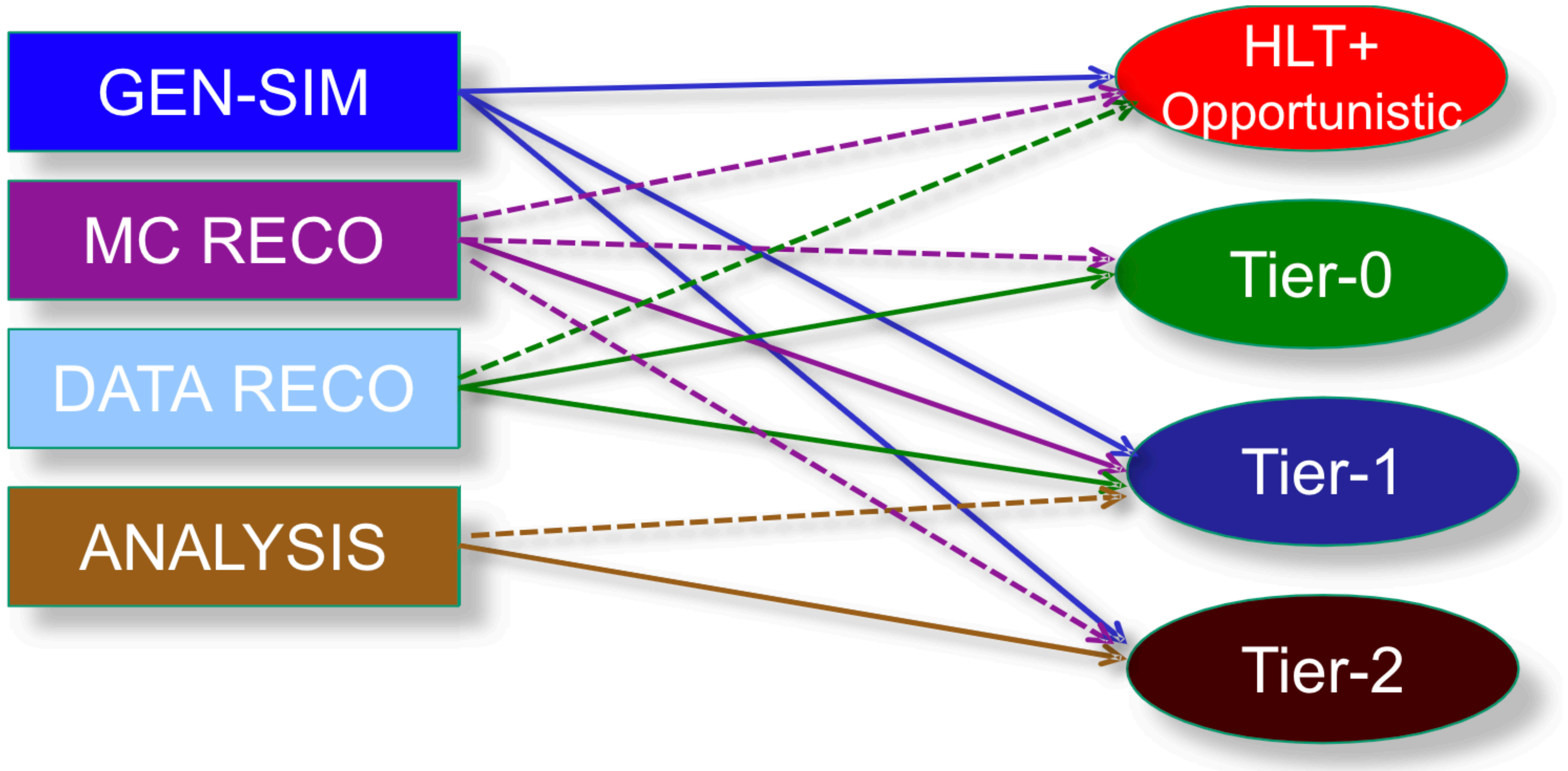
LS2

Automation for resubmissions and log analysis.
Integration to Rucio.
Advanced monitoring using MonIT/Grafana.
Introducing Kubernetes.

Plans for Run3 and HL-LHC

- Further improve Scalability.
- Shift to more community based solutions for Web frameworks and databases.
- Increase Code Concurrency i.e. shift completely to Multithreading and Multiprocessing.
- Horizontal scaling for Kubernetes.
- Better Usage of Data Availability.

Decoupling of Workflows and Resource Types



This graph depicts the decoupling of workflows that was implemented in LS1.
As of LS2, CMS has more flexibility. Everything runs everywhere except the analysis at HLT.

During Run1 ———

New in LS1 - - - - - - - - - -

- Rather tight coupling of workflow types to resources in Run 1
- Big gain in flexibility for Run 2
  - Almost every workflow can run anywhere
  - All CPU joined to one Global HTCondor pool + dedicated Tier-0 pool
  - (Almost) all Tier-1 & Tier-2 disk managed via Dynamic Data Management (DDM)

For more deeper insights on Kubernetes in CMS:

https://indico.cern.ch/event/868940/contributions/3814434/

Want to know more?

kubernetes

Migration of CMSWEB cluster at CERN to Kubernetes

30 Jul 2020, 08:40

20m

virtual conference

Talk | 14. Computing and Da... | Computing and Data Han...

Speaker

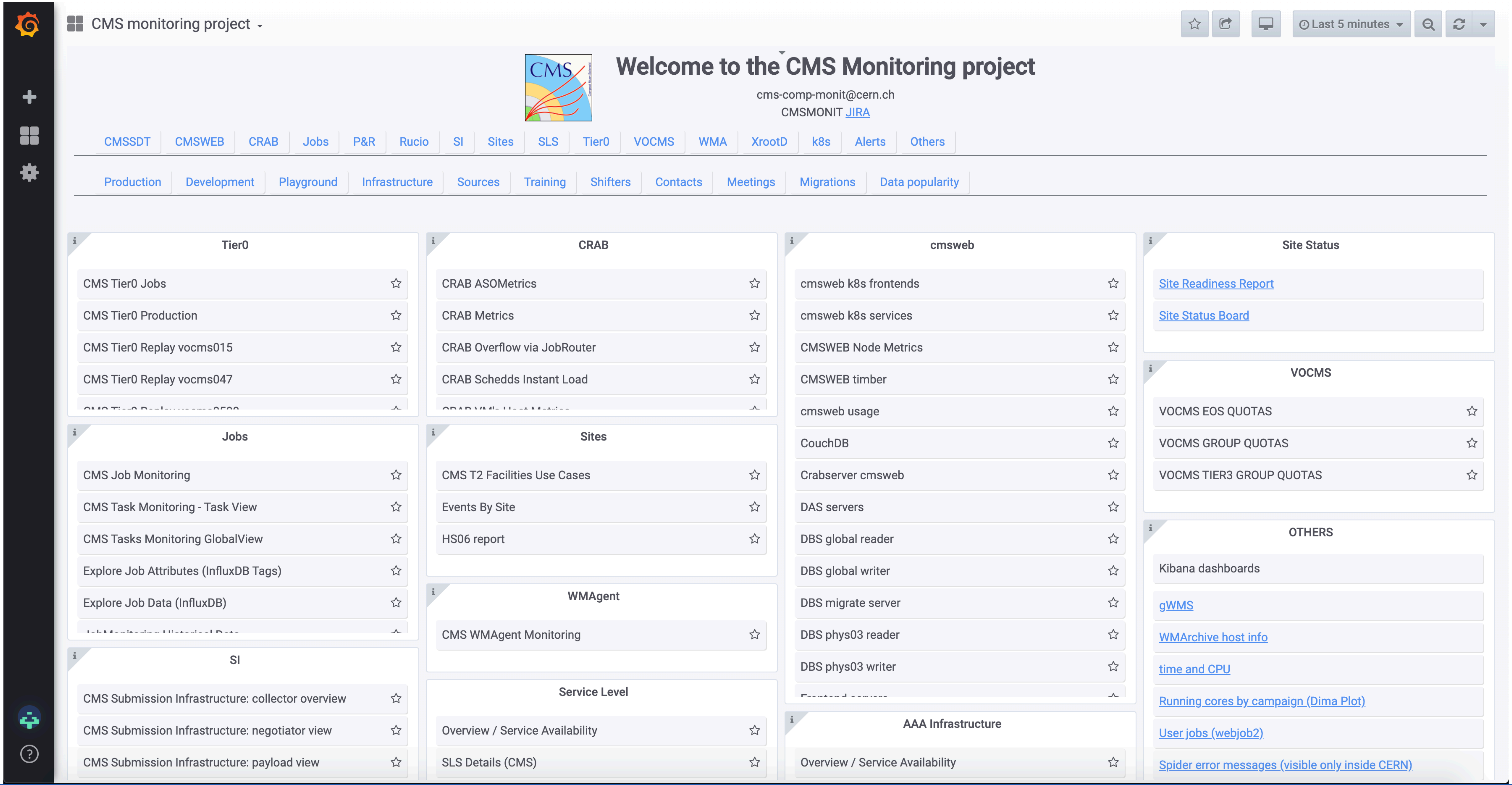
Muhammad Imran (National Centre for P...)

Description

The CMS experiment heavily relies on CMSWEB cluster to host critical services for its operational needs. The cluster is deployed on virtual machines (VMs) from the CERN Openstack cloud and is manually maintained by operator and developers. The release cycle is composed of several steps, from building RPMs, their deployment, validation and coordination tests. To enhance the sustainability of the CMSWEB cluster, CMS decided to migrate it to a containerized solution such as docker, orchestrated with Kubernetes (k8s). This allows us to significantly reduce the release upgrade cycle, follow end-to-end deployment procedure, and reduce operational cost. This contribution gives an overview of the current CMSWEB cluster and its issues. We describe the new architecture of the CMSWEB cluster in k8s and its implementation strategy. We also provide a comparison of VM and k8s deployment approaches, emphasizing pros and cons of the new architecture and report on lessons learned during the migration process.

ICHEP 2020

CMS MonIT-Grafana Dashboard



Challenges for the Future:

- Moving towards Heterogeneous Computing.
- Supporting continuous development and Operations.
- Computing and Storage Resources to meet the needs.
- Developing more intelligent Systems for Operations.

THE END

This is just the beginning!!

We will continue to evolve and provide physics better than ever!

For More Q/As - sharad.agarwal@cern.ch, akanksha.ahuja@cern.ch, david.lange@cern.ch

References

- Y. Iiyama - Dynamo - The dynamic data management system for the distributed CMS computing system
- Y. Iiyama, B. Maier, D. Abercrombie, M. Goncharov, C. Paus - Dynamo - Handling Scientific Data Across Sites and Storage Media
- D. Lange - CMS Full Simulation Status
- J Adelman et al 2014, CMS computing operations during run 1, J. Phys.: Conf. Ser. 513 032040
- https://glideinwms.fnal.gov/doc.prd/download.html
- M. Girone, Offline-Computing Coordination Report, CMS Week 2015,
- Vaandering, Eric. *Transitioning CMS to Rucio Data Management*. United States: N. p., 2019. Web. doi:10.2172/1633740.
- C Wissing and for the CMS Collaboration 2017, Managing the CMS Data and Monte Carlo Processing during LHC Run 2, J. Phys.: Conf. Ser. 898 052012
- D. Piparo, Offline data preparation and computing, CMS Induction Course 2020
- https://twiki.cern.ch/twiki/bin/view/CMSPublic/CMSOfflineComputingResults
- Charpentier, Philippe. (2019). LHC Computing: past, present and future. EPJ Web of Conferences. 214. 09009. 10.1051/epjconf/201921409009.
- C Paus and for the CMS Collaboration 2015, Dynamic Data Management for the Distributed CMS Computing System, CHEP 2015
- https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookGlossary