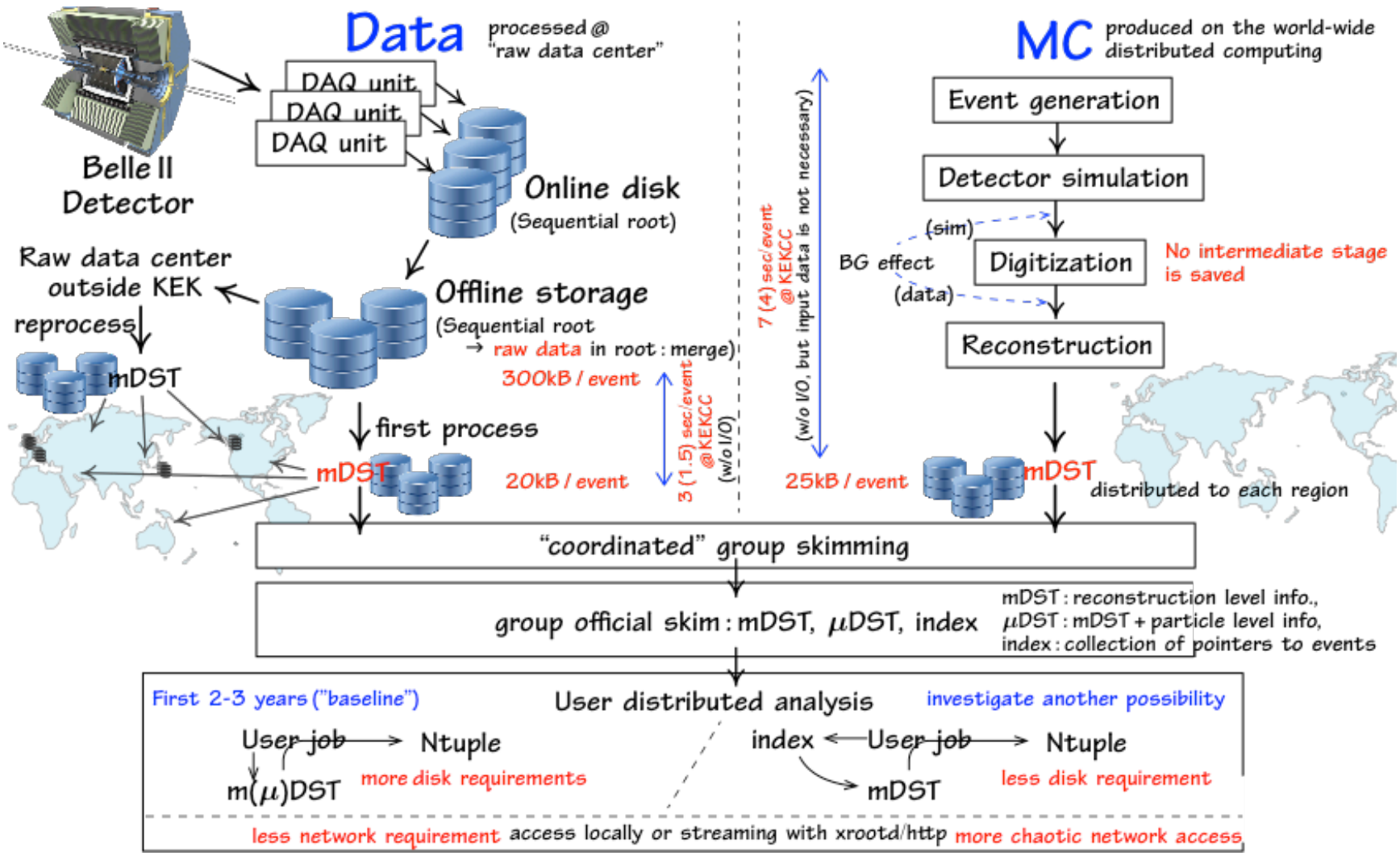
Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

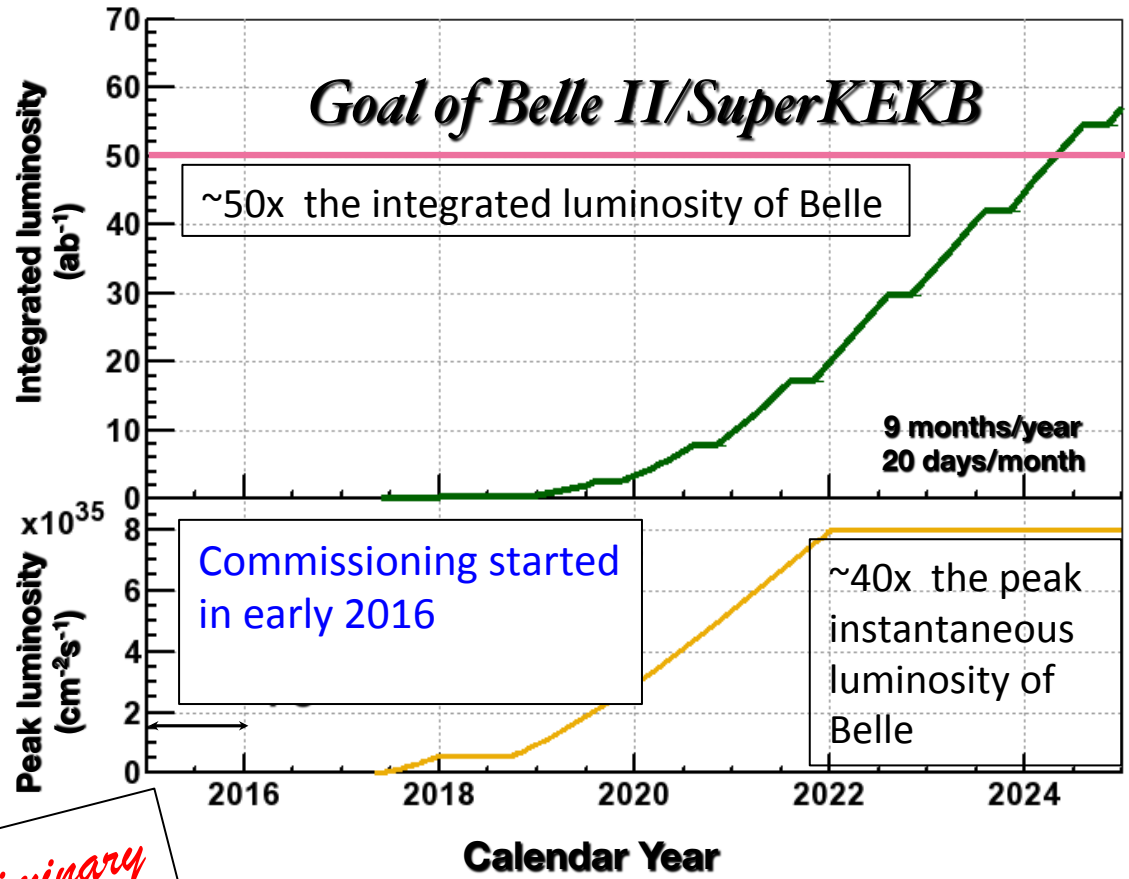
Belle II Computing

MALACHI SCHRAM
PACIFIC NORTHWEST NATIONAL LABORATORY
On Behalf of Belle II collaboration

Computing and Software Workflow



Belle II Data Volume & Resources



- The ***Computing Steering Group*** estimates and defines the expected resources contribution for each country
- Contributions are based on fraction of PhD per country
- The US is expected to provide 15%

Preliminary

Year	2016	2017	2018	2019	2020	2021	2022	2023	2024
Total Storage (PB)	3	8	10	22	60	116	190	265	341
Total CPU (kHEPSpec)	240	340	475	610	710	880	1060	1420	1640

~80-160k cores

US Contribution to Computing

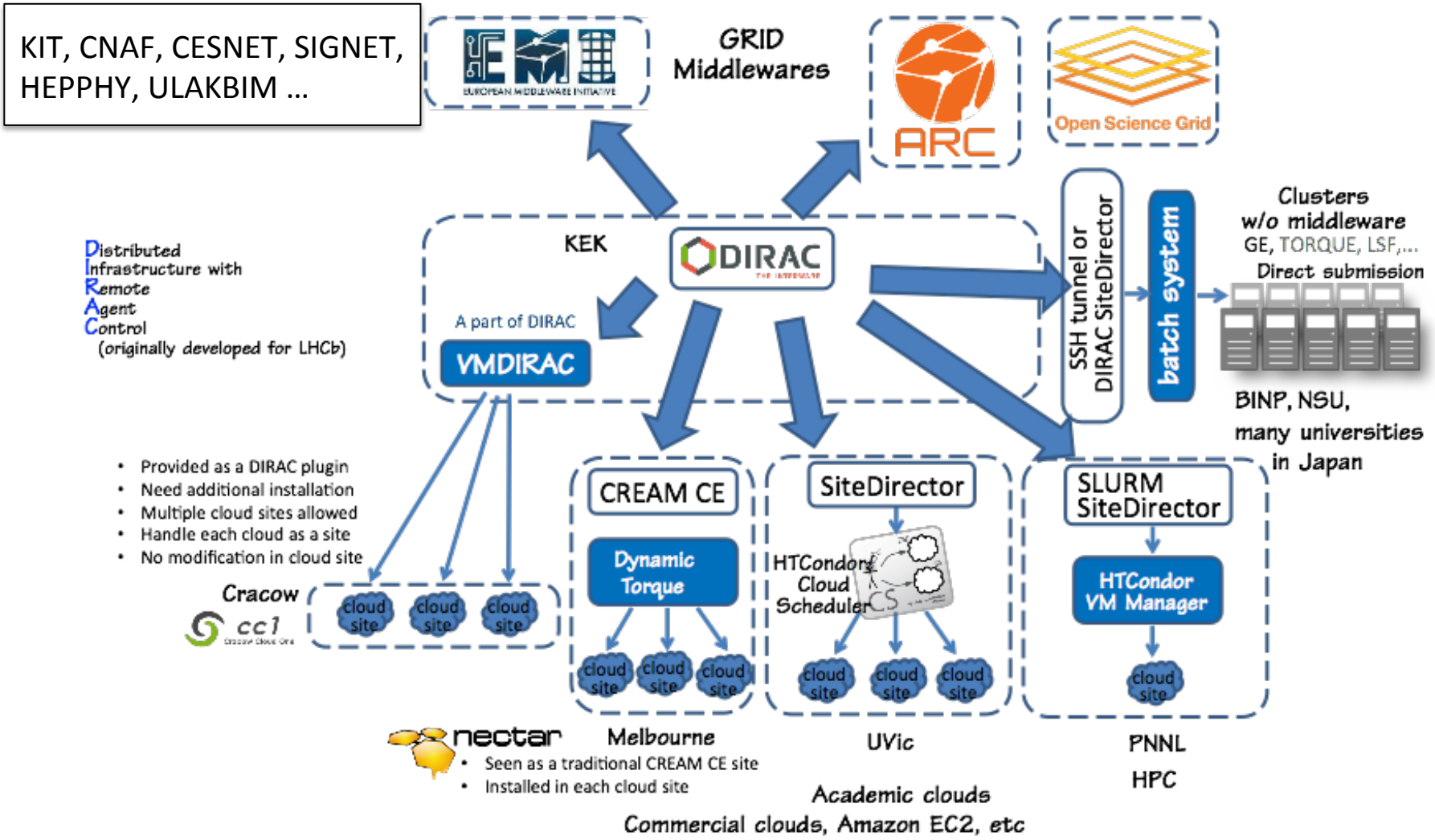
Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

- **Tier-1 Compute and Storage Center**
  - ~4000 cores of compute
  - ~0.5PB of storage
  - Core infrastructure is build using OpenStack and Kubernetes
  - Additional services: VOMS, CVMFS, DIRAC development systems, etc.
  - Investigating using HPC, OSG, University resources
- **Distributed Data Management System and Networking**
  - Coordinator and lead developers (details in slides)
- **Conditions Databases**
  - Coordinator and lead developers (details in slides)

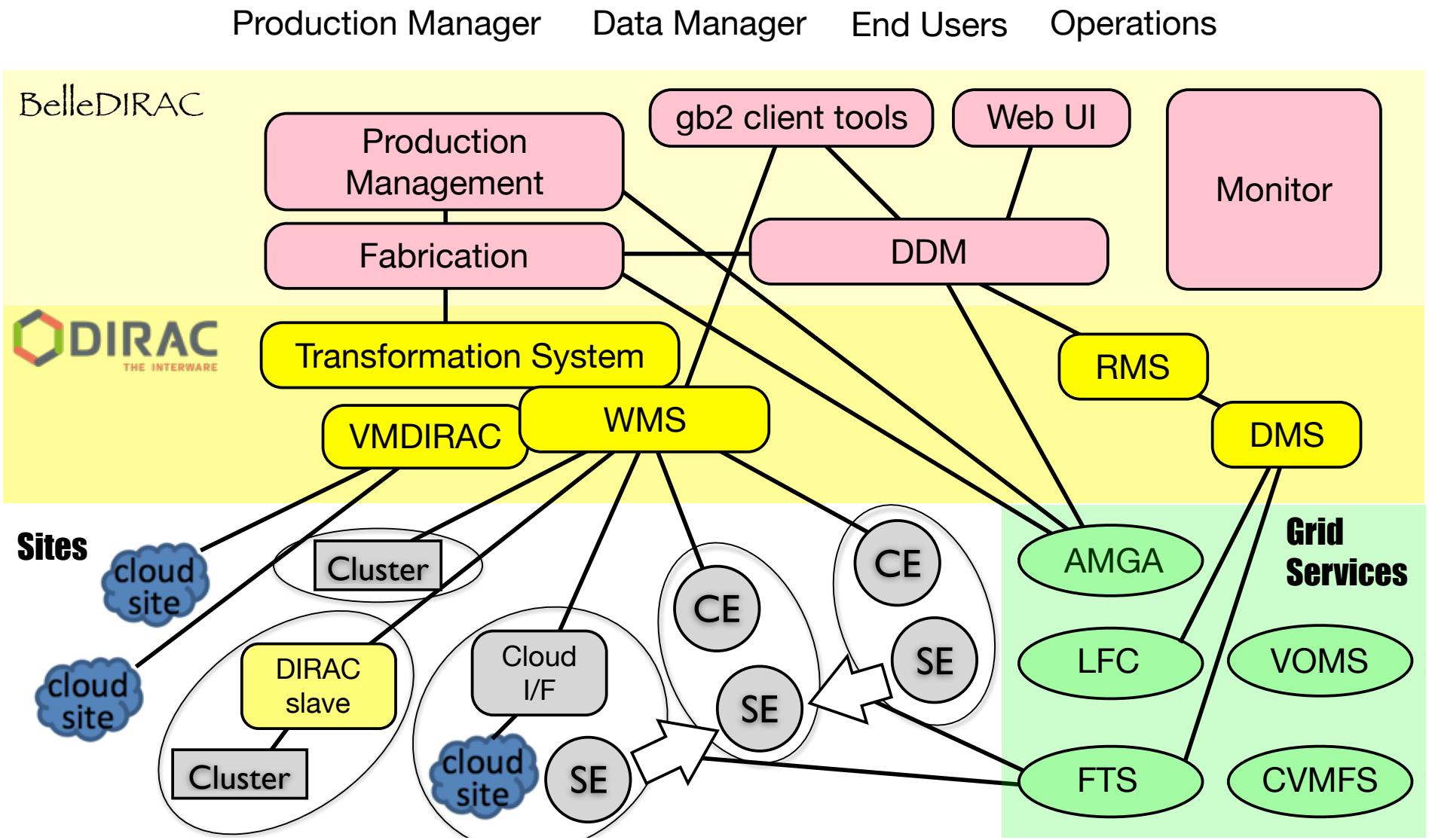
DIRAC (Distributed Infrastructure with Remote Agent Control) **INTERWARE**

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

KIT, CNAF, CESNET, SIGNET, HEPPHY, ULAKBIM …



Belle II Distributed Computing Layers



Core BelleDIRAC Extensions

"Production Manager" (human)

- to define a production : sw release, event type, number of events to produce, …
- and put the parameters into Production Management System

Production Management System

- to create Fabrication instance(s) : simulation, reprocess, skim, merge, …
- and chain them as needed

Fabrication System

- to create and look after jobs, replace failed jobs, verify output of successful jobs
  - simulation jobs can run everywhere, merge jobs can run at the sites with "major" SEs.
- to feed output files to Distributed Data Management System

Distributed Data Management System

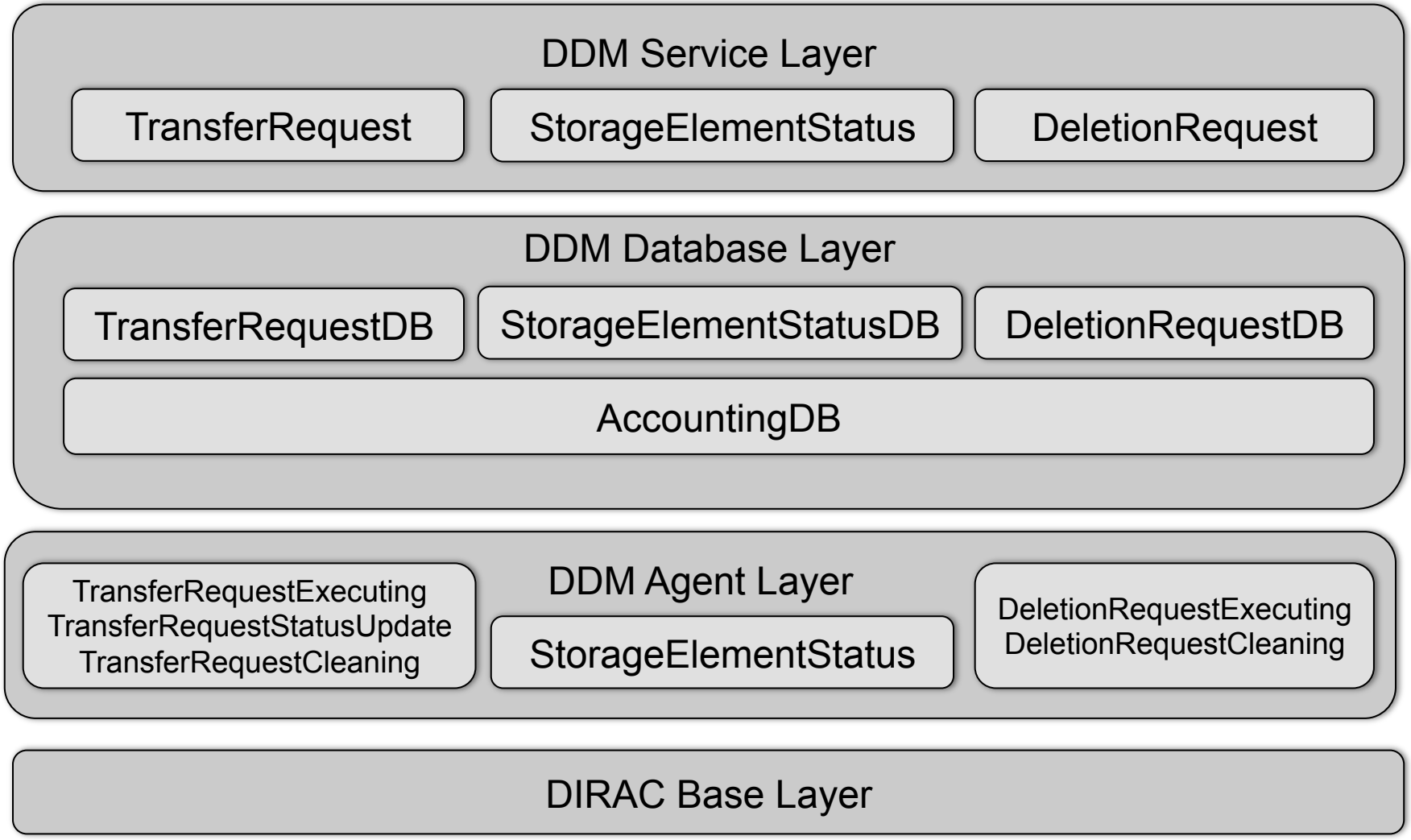
- to gather output files to "major" SEs in blocks (important for merge)
- to distribute "final" products over the grid (for further processing / analysis)

Distributed Data Management System

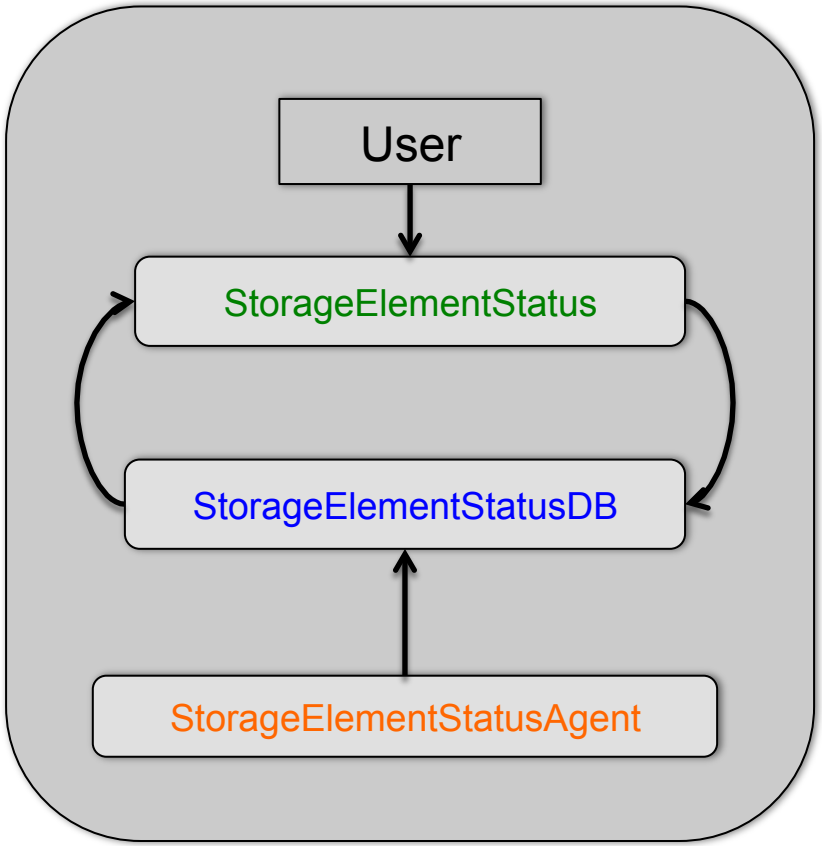
Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

- **Storage element accounting components**
  - Provide health of each registered Storage Element (SE)
  - Overall storage availability at each SE
  - Group and user level accounting
  - Network Monitoring
- **Data transfer components**
  - Provides tools to transfer data between sites
  - Provide tools for users to retrieve their samples
- **Bulk data deletion components**
  - Provides tools to do large scale data deletion
- **Data Integrity**
  - Insures that the physical data is pristine and consistent with the File-Catalogs

Distributed Data Management System



StorageElementStatus Workflow



Purpose:

- Provide near real time SRM storage elements information

Current implementations:

① StorageElementStatus provide API layer for internal and external users.

② StorageElementStatusAgentDB provide persistified information

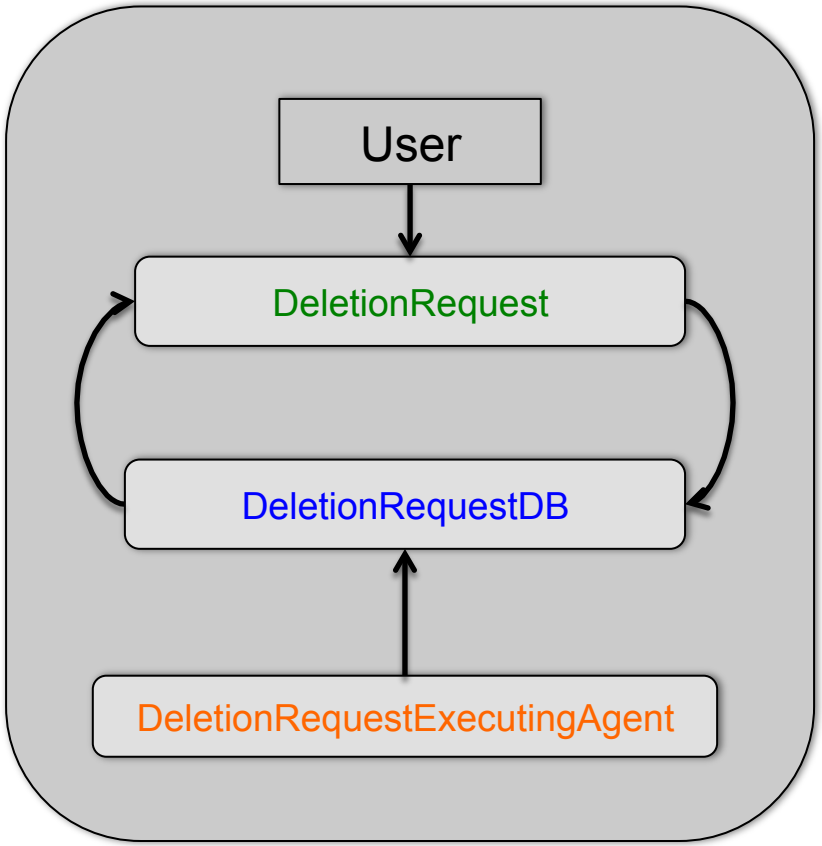
③ StorageElementStatusAgent :

- Available space
- Access rights (w/r) at file and director level
- ADLER32 Checksum

Extra features:

- Include full ACL

DeletionRequest Workflow



Purpose:

- Provide bulk deletion request

Current implementations:

① DeletionRequest provide API layer for internal and external users.

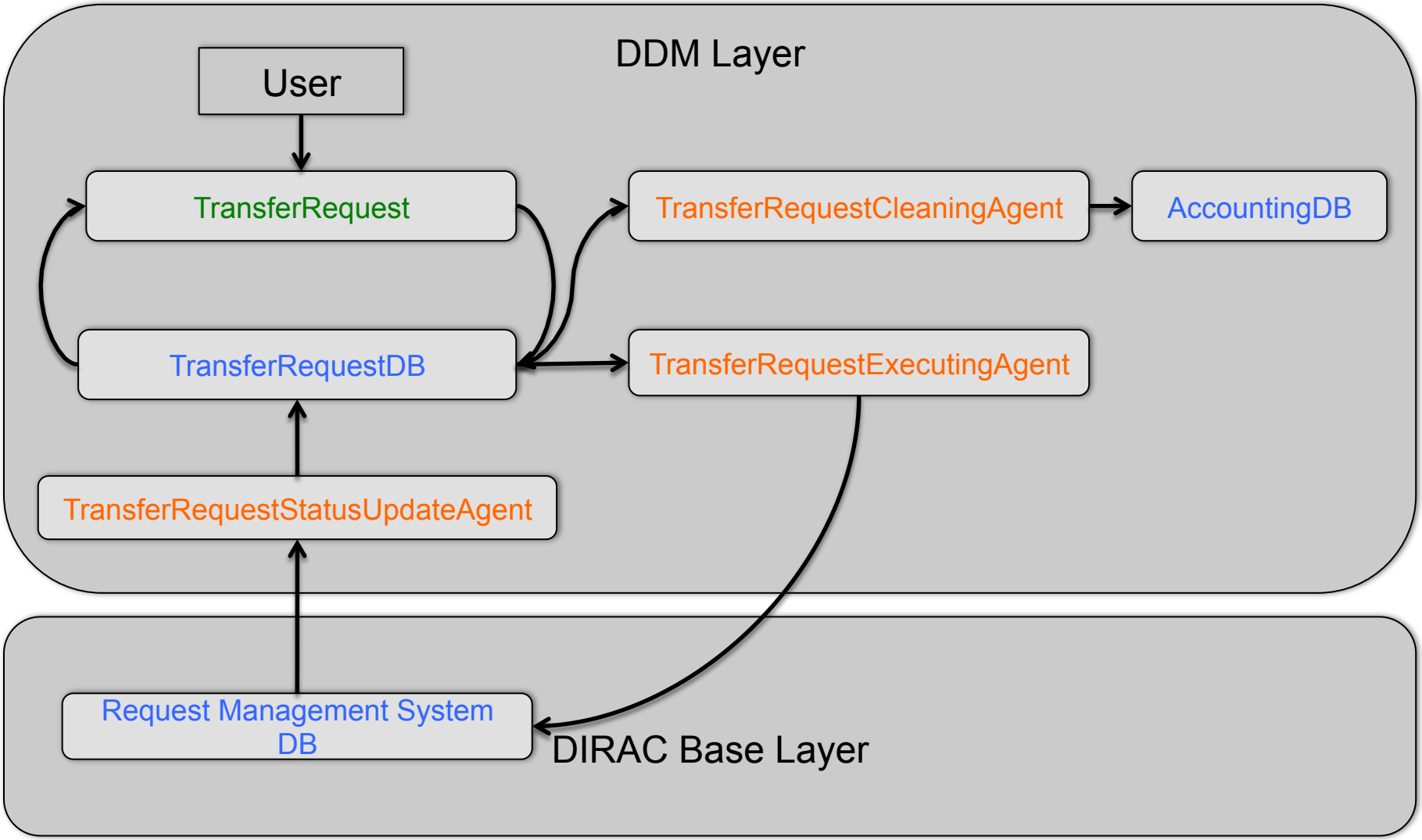
② DeletionRequestAgentDB provide persistified information of requests

③ DeletionRequestExecutingAgent:

Extra features:

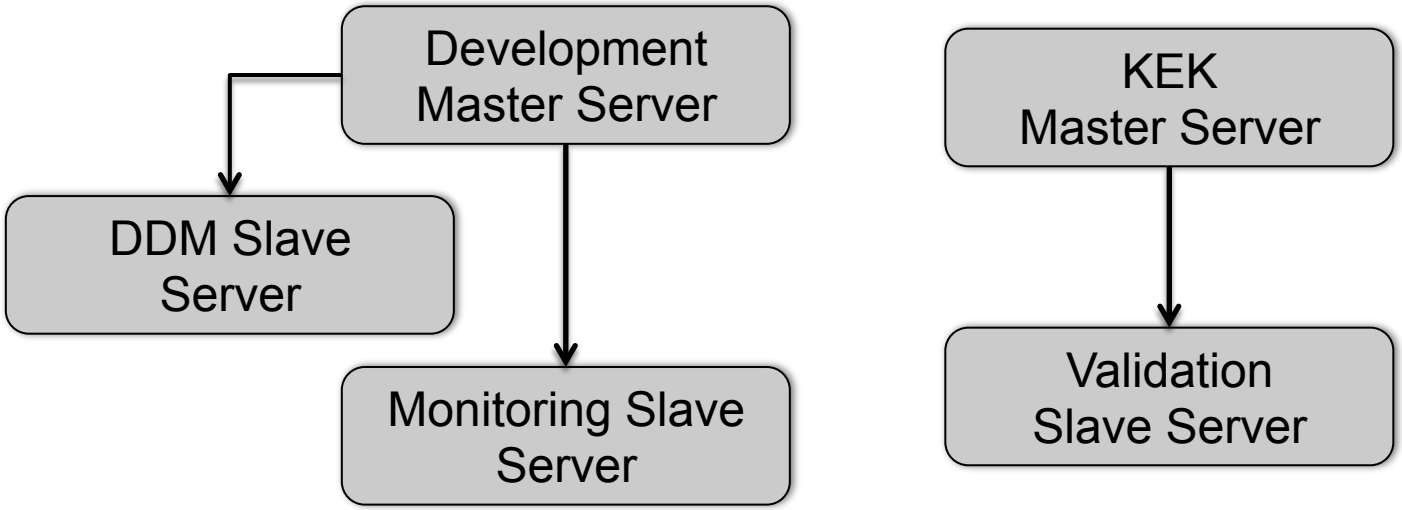
- Priority feature

Transfer Request Workflow



Development Life Cycle

- DDM group work on 3 week development cycle (Agile Sprint)
- Dedicated development DIRAC master server used to configure test setups
  - Runs all systems
  - Common global resources access for development slave servers
- Dedicated DDM slave server for dynamic development
  - PNNL and Melbourne
- Dedicated validation server
  - Runs all systems
  - Uses common global resources for validating development

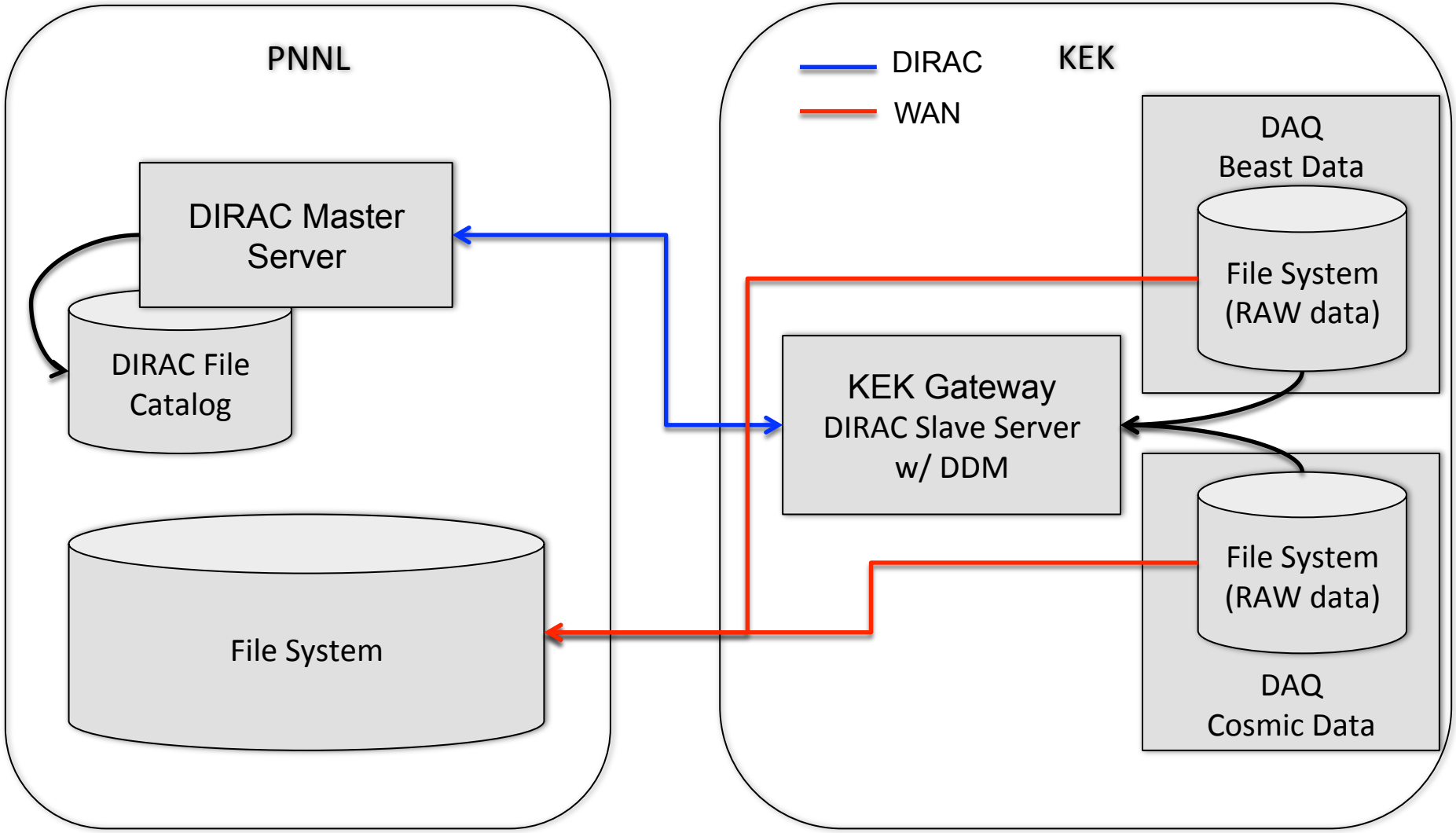


Cosmic Ray Test and DDM

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

- **Purpose**
  - Transfer raw CRT data from KEK to PNNL
  - Socialization exercise between DAQ and Distributed Computing group
- **Phases 1**
  - Provide access to while KEKCC is down
  - Restricted network between KEK and PNNL
  - No access to KEKCC grid computing
- **Phases 2**
  - Detailed study of network and storage performance
  - Use new KEKCC storage
  - Use LHCONE network

Comic Ray Test: Phase 1



Road to full production system

MC3 (2014 May): no Production System

- 2.5 M jobs: **manual submission** by the **shifters** (with a script)

MC4 (2014 Nov): with "prototype" Fabrication System

- ~5 M jobs: **automatic submission** by the system, controlled by **a single system manager**
- Output files uploaded directly from WNs to "major SEs"

MC5 (2015 Aug — 2016 Feb): with improved "prototype" Fabrication System

- ~10 M jobs: automatic submission by the system, controlled by a single system manager
- Output files **transferred by FTS** from production sites to "major SEs"
- **Monitoring and automatic issue detection** for shift tasks and expert investigation

Cosmic ray data replication (2016 Jun ~):

- An agent to register raw data files as they get available on a storage, and make transfer requests
- **Automatic replication of raw data**

Road to full production system (2)

MC6 (2016 Jun): with Fabrication System and DDM

- O(1) M jobs: automatic submission by the system, controlled by a single system manager
- Output file transfers by FTS **controlled by DDM**
- Monitoring and automatic issue detection for shift tasks and expert investigation

MC7 (2016 Nov ~ 5 months): with the Production Management System

- **The software should be well tested by the software group, validated by a production manager**
- Automatic job submission by the system, **controlled by the system**
- Automatic file transfers by the system, **controlled by the system**
- Monitoring of **production progress**
- **Automatised operations** (eg. ggus tickets, downtime, storage health)
- **Tune and improve the system** during the production and afterwards
- **The system will be ready for workflows** other than MC production (raw dta processing, reprocessing, skim, …)

Road to full production system (3)

Scale Tests (to be repeated as needed in between "production" activities)

- To see if we have any limitation in job or file transfer throughputs beyond the current level
- And address the issues found

Cosmic ray data processing (2017)

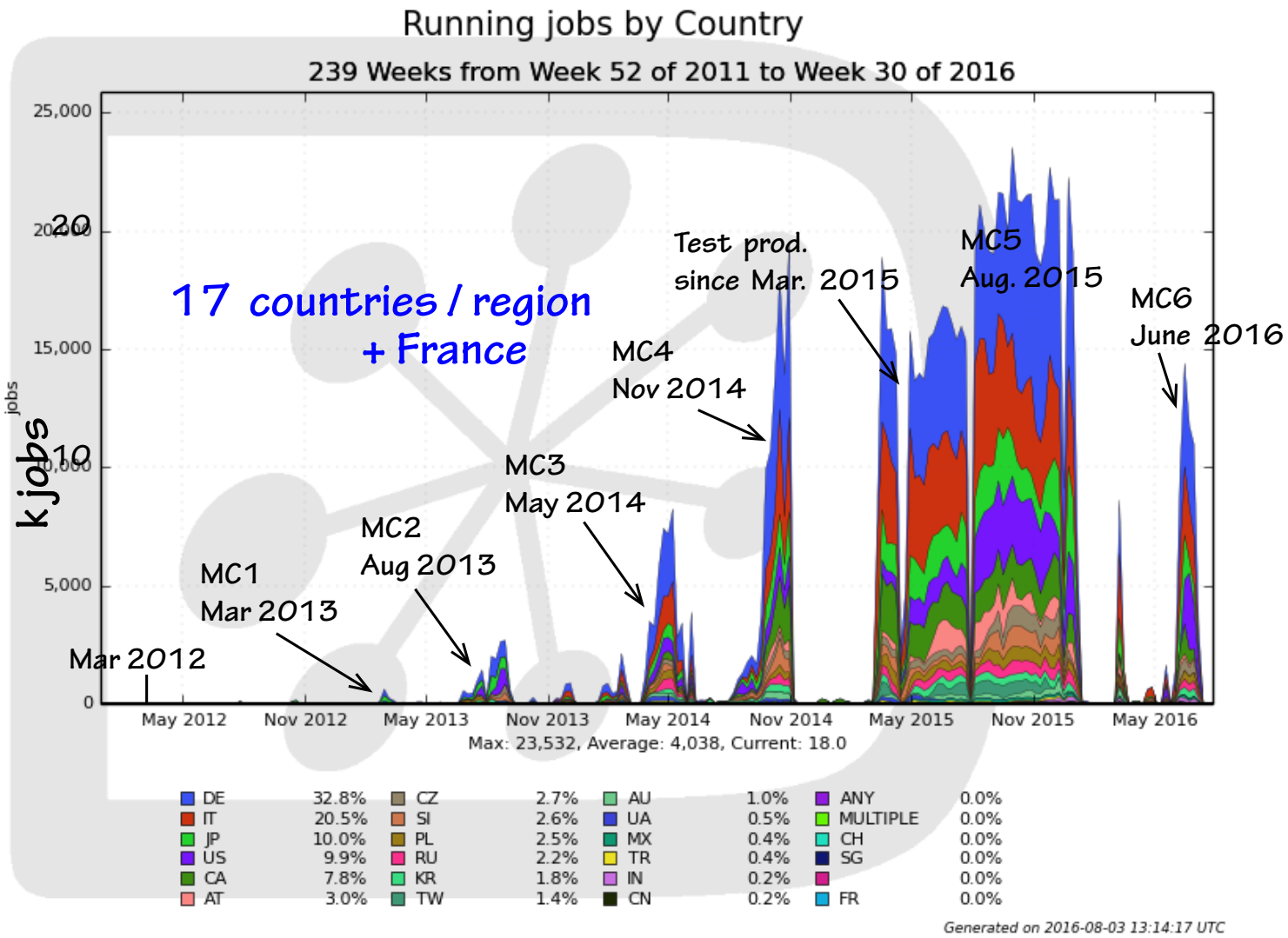
- **Raw data format and reconstruction software need to be ready** and well tested by this time
- **The production system should be ready,** it has been developed not only for MC production, but with other use cases (incl. raw data processing)
- Still, this will be the **first real use case to try raw data processing workflow**

System dress rehearsal (2017): before Phase-2 runs

- To try out the **full chain workflow** from raw data to skim
- Find issues and address them before phase-2 runs

System full dress rehearsal (2018): before Physics runs

Monte Carlo Campaign History



Belle II Networking Requirements

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

- In 2014, Belle II was added to the LHCONE network
- International network requirements are based on (*production only*):
  - RAW data
  - mDST from data and Monte Carlo - after data taking and reprocessing
- KEK estimated outgoing (incoming) traffic of 20Gbps (5Gbps) in 2024.
- PNNL estimated maximal incoming (outgoing) traffic 8Gbps (4Gbps)
- European estimated maximal incoming (outgoing) traffic 10Gbps (4Gbps)
- India, Canada and Slovenia will require ~2Gbps traffic
- Other site (Australia) will be interested by a traffic less than 1Gbps

Latest Network Data Challenges

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

- DC 1 (2015):
  - KEK ←→PNNL
  - KEK → EU
- DC 2 (2015/2016): LHCONE sites
  - PNNL ←→ EU
  - EU←→ EU
- DC 3 (2016): SINET5 (phase 1)
  - KEK ←→PNNL
  - KEK → EU
- Software stack used:
  - perfSONAR suite
  - SRM storage elements + gridftp servers
  - FTS3 with manual channel configuration

Network Data Challenge Results

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

- Results of key sites based on the latest data challenges

Source →	KEK	PNNL	DESY	KIT	CNAF	NAPOLI	SIGNET
KEK	N/A	6.3 Gbps	8 Gbps	7 Gbps	7 Gbps	6.6 Gbps	2.5 Gbps
PNNL	3.9 Gbps	N/A	6.6 Gbps	6 Gbps	9 Gbps	8 Gbps	-
DESY	3 Gbps	6.6 Gbps	N/A	8 Gbps	8 Gbps	8 Gbps*	3 Gbps
KIT	3.2 Gbps	4 Gbps*	8 Gbps	N/A	8 Gbps	6 Gbps	3 Gbps
CNAF	3.8 Gbps	8 Gbps	10 Gbps	6 Gbps	N/A	8 Gbps*	3 Gbps
NAPOLI	3 Gbps	3 Gbps*	3 Gbps*	3 Gbps	3 Gbps*	N/A	3 Gbps*
SIGNET	0.8 Gbps	0.6* Gbps	5 Gbps	2 Gbps	5 Gbps	2 Gbps*	N/A

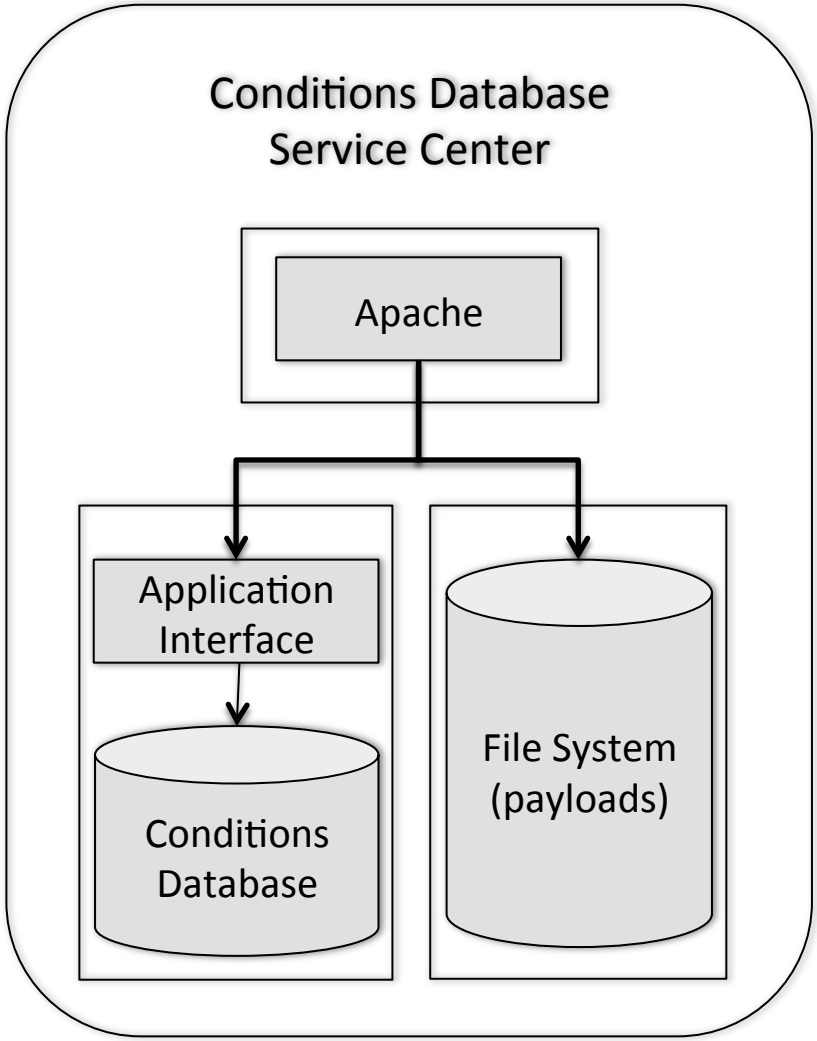
Networking Status and Plans

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

- **Current Status**
  - Network data challenge verify that:
    - KEK→PNNL requirements are satisfied until 2019
    - KEK outgoing traffic appears to be limited to ~3-4Gbps
    - PNNL→EU requirements are satisfied
    - EU ←→ EU requirements are satisfied
  - Next data challenge will test new KEKCC and full LHCONE setup (Fall 2016)

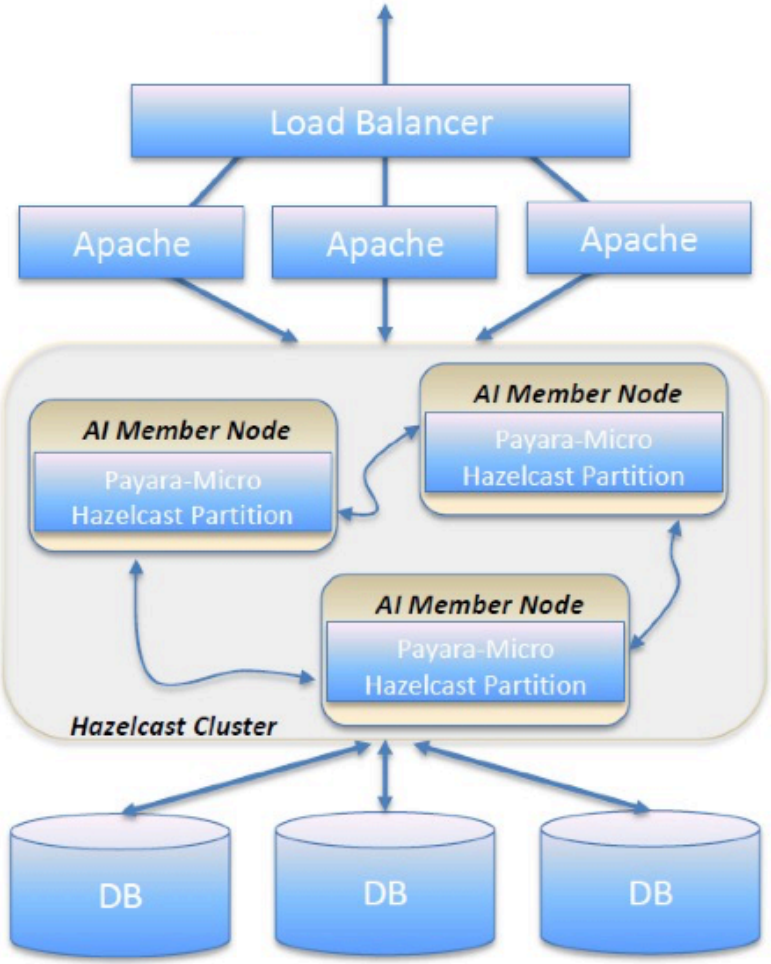
- **Moving Forward (FY16-17):**
  - Integrate networking transfer tools into the Distributed Data Management System
  - Integrate KEK in LHCONE and perform new Data Challenge
  - Update network requirements to include estimation on non-production traffic

Conditions Database Overview



- Provide read/write access to calibration, alignment, geometry, environmental parameters
- Adopted REST API
  - ***Apache*** server
  - Application Interface (***GlassFish***)
  - ***PostgreSQL*** database
  - ***Lustre*** file system
- A global tag is a unique identifier to define a collection of payloads with defined IOVs
- Belle II will use global tags to define processing runs, etc

Scaling the Conditions Database



- Prototype developed using ***OpenStack***
- Use ***Docker*** to develop modular components which allows for quick scaling as needed
- Use ***Swagger*** for automated API documentation
- User ***Payara*** and ***Hazelcast cluster*** for flexible and scalable in-memory data grid solution
- Testing ***OpenStack Trove*** provide redundancy and high availability

Extra Slides

- **Resources**
  - PNNL Tier-1 site
  - Several services are in place at PNNL: CVMFS, VOMS, etc.
  - Investigating increasing the US contribution: Universities, OSG, HPC
  - Conditions Databases in place
  - CRT and BEAST DIRAC setup are available … yes you can analysis TOP data!

- **Computing**
  - Prototype production system is available
  - Working on scalability test of each components and overall system
  - Three separate systems: Develop/Validate/Production
  - Developing automated unit testing

If you're interested join us!!!