The Management of Heterogeneous Resources in Belle II

Pacific Northwest NATIONAL LABORATORY

Proudly Operated by Battelle *Since 1965*

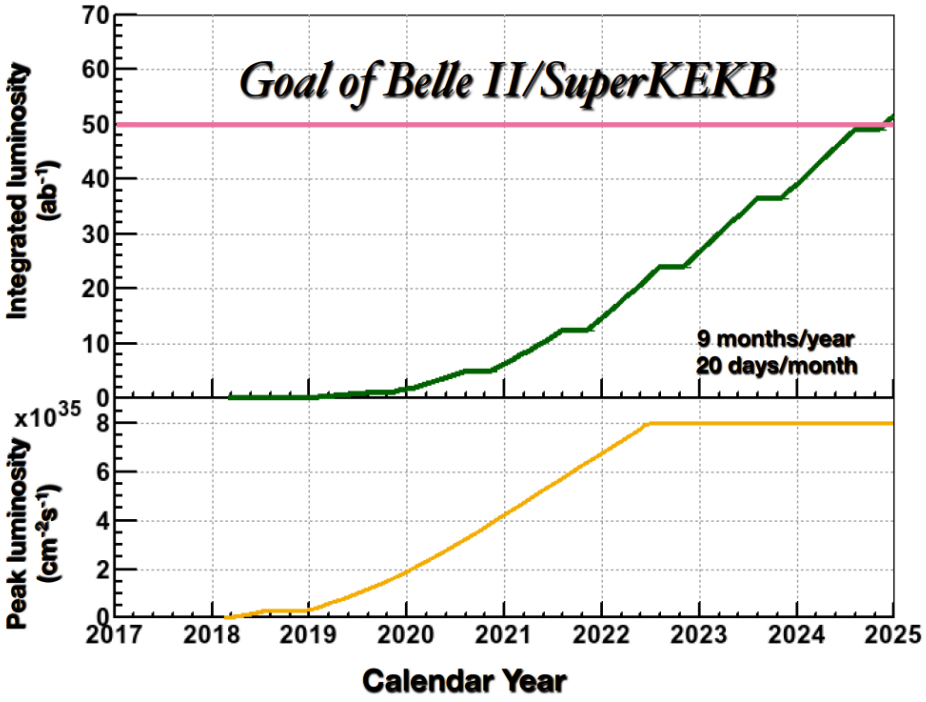
Malachi Schram, Vikas Bansal, Antonio Ledesma
Pacific Northwest National Laboratory
August 22nd, 2017

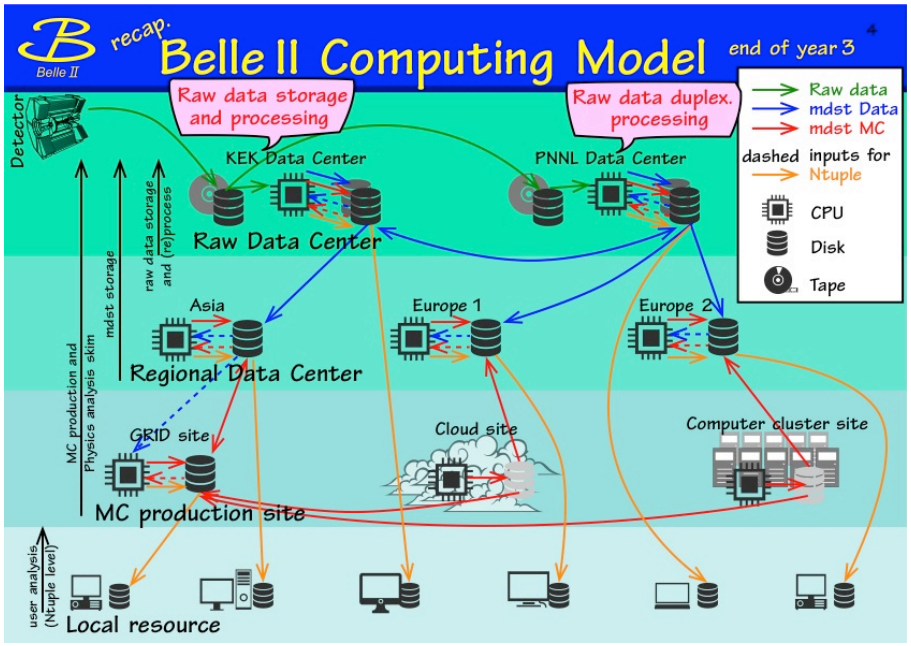
Belle II Overview

- **Goal**: discover new particles and phenomena beyond the Standard Model of particle physics
- Collaboration among **725+ physicists** from 104 institutes in 24 countries
- **50x** the data volume, **40x** rate of collisions relative to previous Belle experiment
- **PNNL led U.S.** (DOE) contribution to Belle II detector construction – now complete
- **Largest** ever U.S. science investment in Japan
  - More Ph.D. physicists (50+) and more institutions (14) than any other country
- SuperKEKB: single beam circulation was done **successfully** (phase 1 in 2016)
- Cosmic-ray **data taking is on-going**
- **Physics** run will start in **2018**

Belle II Computing Requirements

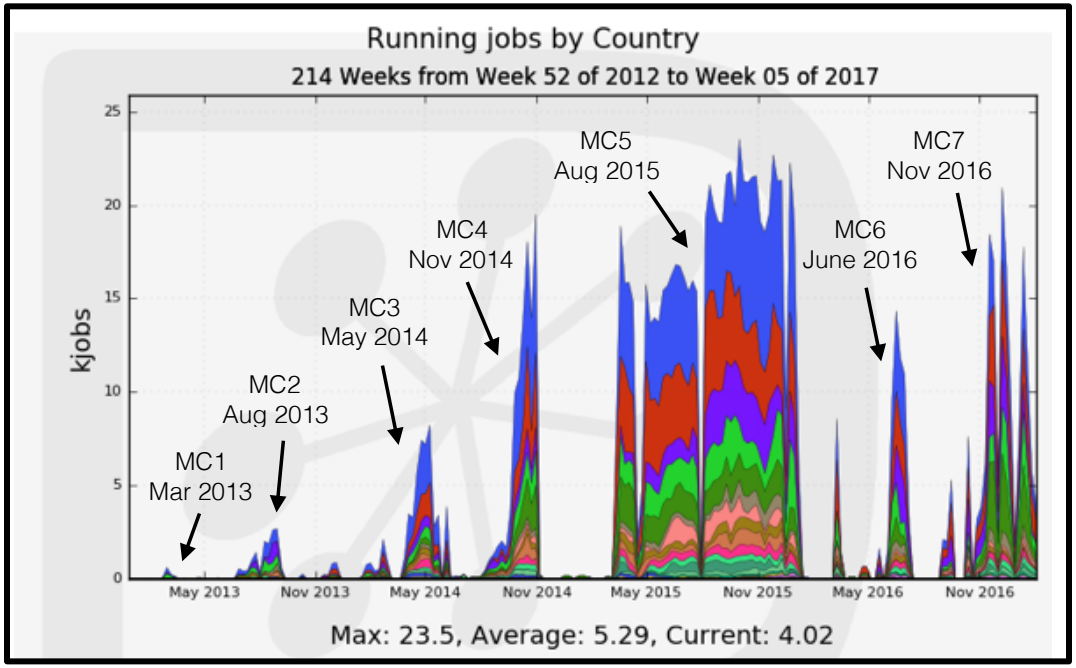
- Expected data rates from the Belle II experiment are high
  - Event rate of 6 kHz, corresponds to 11 PetaBytes per year starting in 2022
  - 100 PetaBytes raw data volume by 2024, total data volume ~190 PB
- Processed data samples will be distributed worldwide (Asia, North America, Europe)
- The Belle II Computing Steering Group resource estimations and feedback from Belle Program Advisory Committee and DOE reviews
- Resource requirements depend upon luminosity, event size, replicas, MC samples, etc.



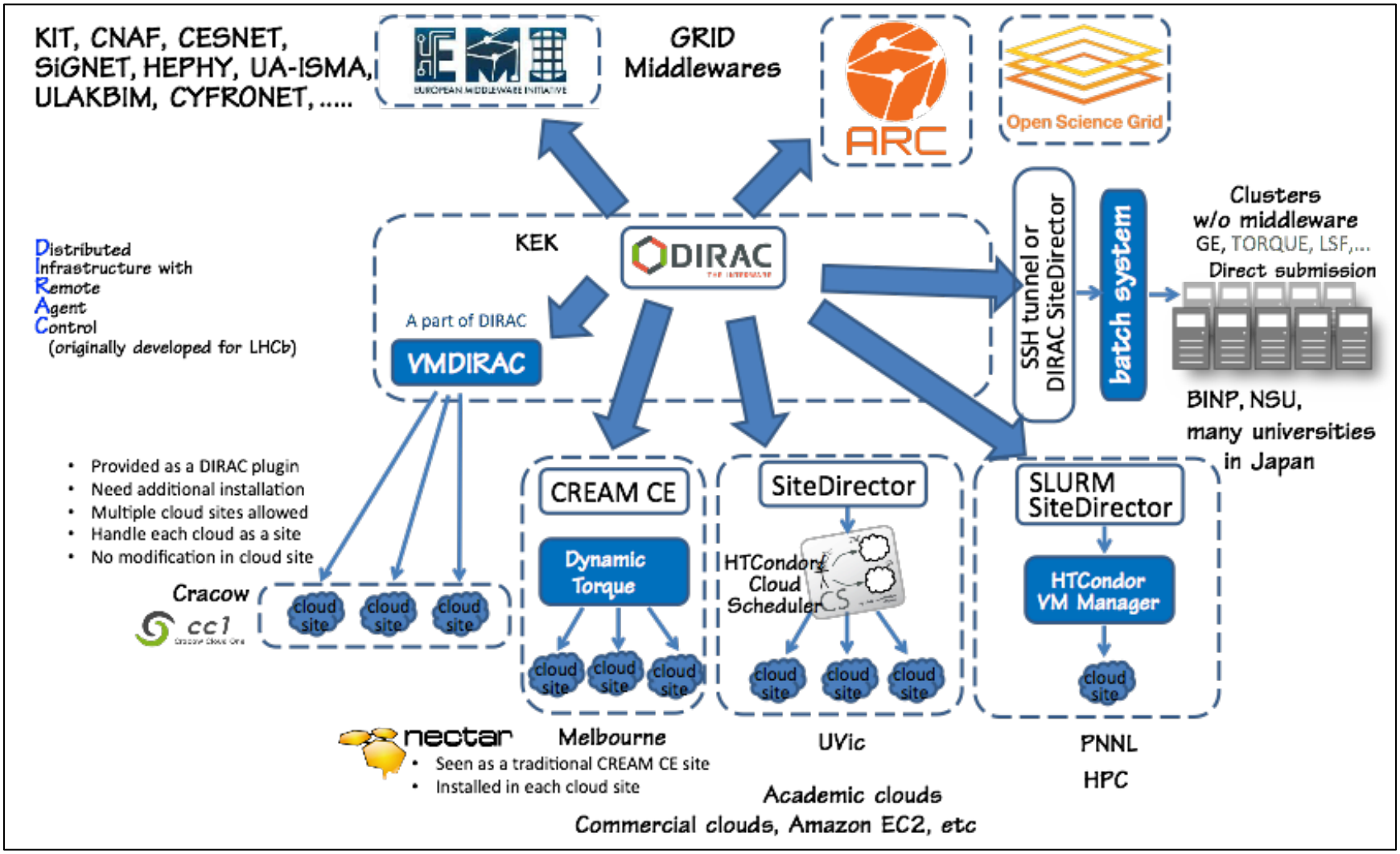


Belle II Monte Carlo Production

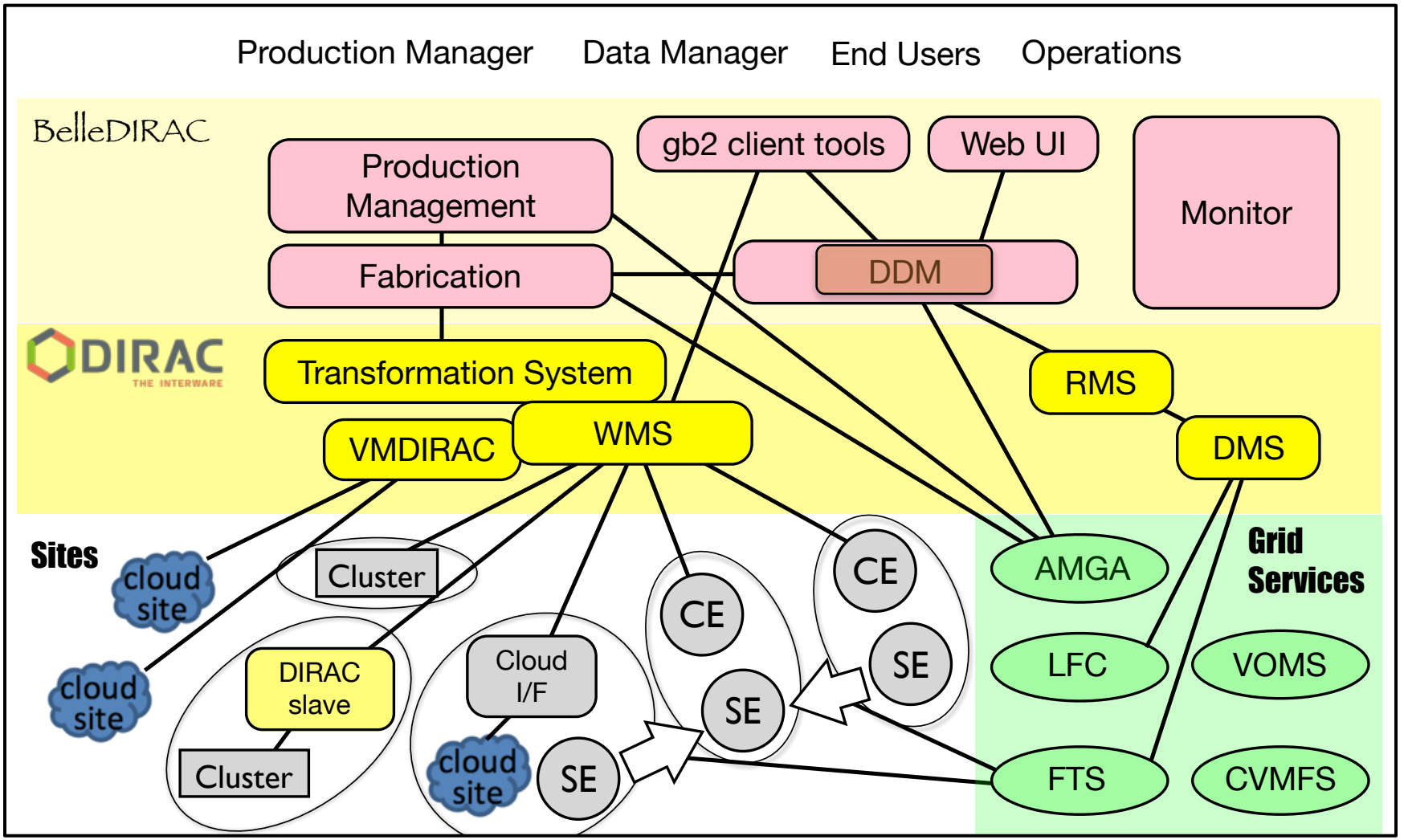
- **Accomplishments:**
  - MC production are critical exercises to test the latest software, scalability, and provide physics samples for analysis.
  - Monte Carlo #7 ran from Nov. 1st 2016 to Feb. 7th 2017
    - Approximately 27 billion events simulated
    - Reached 25k concurrent jobs (>200 kHEPSpec)
  - Monte Carlo #8 started on Feb 16th and is done … MC9 is ramping up
    - New Belle II computing record ~300kHEPSpec as measured by DIRAC pilot jobs



DIRAC (Distributed Infrastructure with Remote Agent Control) INTERWARE



Belle II Distributed Computing



U.S. Belle II Computing Requirements

- Resource requirements depend upon luminosity profile & event size (data volume), (re)processing, MC generation, and user analysis
- MC production accounts for nearly 50% of US Belle II computing resources

- Using of Leadership Class Facilities (LCFs) for MC production
- Investigating LCFs for scalable detector and physics studies
  - Rare physics channels such as B→D*taunu and B→K(*)tautau
  - iTOP "ring image" PID performance studies using Deep Learning
  - Detailed systematic studies when using Deep Learning models
  - Exhaustive hyper parameter scans

	CY17	CY18	CY19	CY20	CY21
CPU [kHEPSpecs]	20.11	27.56	58.90	69.71	82.97
Storage [PB]	0.31	0.81	5.04	6.50	9.28
Networking In/Out [Gbps]	0.30/0.30	0.49/0.36	1.06/0.26	1.56/0.31	1.89/0.83

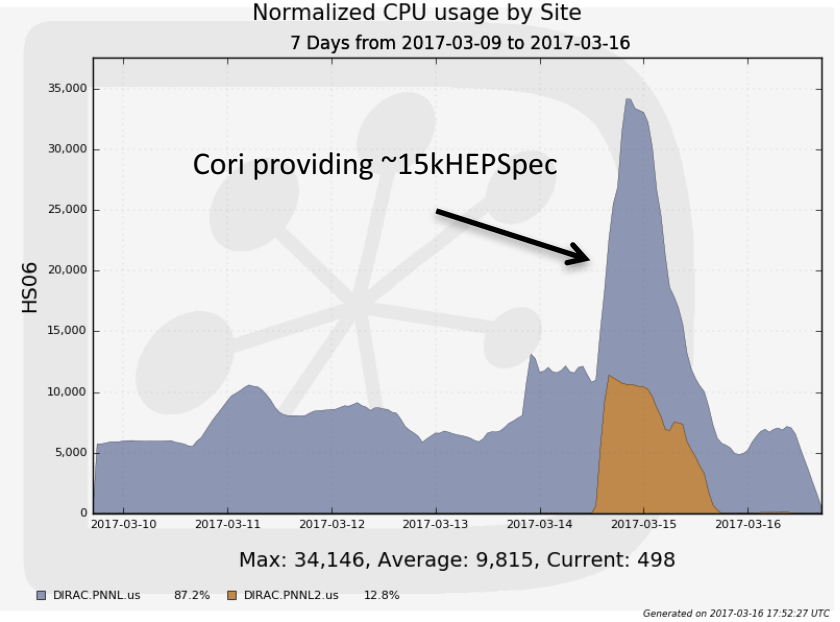
	CY17	CY18	CY19	CY20	CY21
CPU @ PNNL [kHEPSpecs]	0.41	6.35	40.52	29.01	41.58
CPU @ LCF [kHEPSpecs]	19.70	21.21	18.38	40.70	41.39

PNNL Computing Facilities & Infrastructure

- **Description:**
  - PNNL leverage virtualization technologies using Kubernetes and OpenStack to provide a fault tolerance and flexible infrastructure
  - HEP Computing is funded by a mix of DOE, PNNL HEP project money, and US-Japan funding
- **Major Components:**
  - Gridftp servers w/ bestman2 SRM
  - OpenStack
  - Enterprise Linux-compatible distros
  - Docker
  - DIRAC
  - Condor
  - Kubernetes
  - Ceph object store
  - Lustre shared filesystem
- **Services provided via private cloud:**
  - DIRAC server instances
  - Compute nodes
  - FTS3 to manage file transfers
  - GUMS, VOMS servers
  - CVMFS Stratum 0 & 1
  - Squids
  - Custom services on relational databases (e.g. Belle2db)

Belle II Computing on NERSC

- NERSC has two clusters with grid access:
  - Edison
  - Cori
- NERSC Allocation for 2017
  - Used to develop/test/validate the use of HPC for Belle II
- DIRAC GlobusComputingElement is being modified to provide extra options to enable the use Edison and Cori
- MC9 phase 3 just started and will allow us to monitor the availability of the NERSC resources



Project Title	US Belle II HPC Workflow	**Funded by DOE Office of Science?**	Y
Organization	Pacific Northwest National Laboratory (PNNL)	**DOE Manager**	Kevin Flood, Glen Crawford
DOE Office & Program	HEP - HEP Other Research	**MPP Hours requested in ERCAP**	10,000,000
Science Category	High Energy Physics	**SRUs requested in ERCAP**	1,000
Project class	DOE Base Funding		

Grid Components for NERSC

DIRAC

Workload Management Agent

SiteDirector
- Using modified *GlobusComputingElement*

Resource Definition

OSG.CORI.us
- Defined as *GlobusComputingElement*
- *Mapped to PNNL StorageElement*

OSG.EDISON.us
- Defined as *GlobusComputingElement*
- *Mapped to PNNL StorageElement*

/cvmfs/belle.cern.ch
-Sync repo to docker

Docker
pnnlhep/osg-compute

Belle II HPC Docker
pnnlhep/b2-AB-CD-DE

NERSC

Pull and register into Edison/Cori shifter

Grid submit with docker image and volume host/docker mount points for input/ouput/repo

LCFs Status and Plans

For NERSC:

- ✓ Request allocation; 10M core hours for CY17
- ✓ Setup for NERSC Cori and Edison setup are done
- ✓ Docker with software release 00-09-01 used for MC9 is done
- Develop MPI wrapper to submit “big” jobs and test on NERSC
- Create multiple DIRAC SiteDirector to submit different size jobs to NERSC

For ORLCF:

- ✓ Request allocation; 1M core hours for CY17
- Develop new DIRAC “SiteDirector”
- Develop backlog job optimization for scheduling

Belle II Distributed Data Management Computing Software

DIRAC is the main software to orchestrate Belle II computing resources

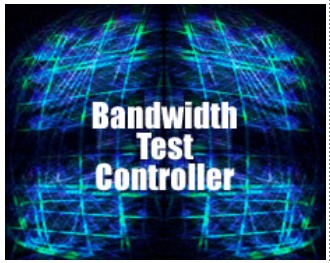
DIRAC

WAN Data Challenge Setup

File Transfer Service (FTS3) to schedule transfers between sites

FTS3

BWCTL



BWCTL for initial network test. Setup on perfSONAR boxes

Iperf3 , traceroute

MaDDash

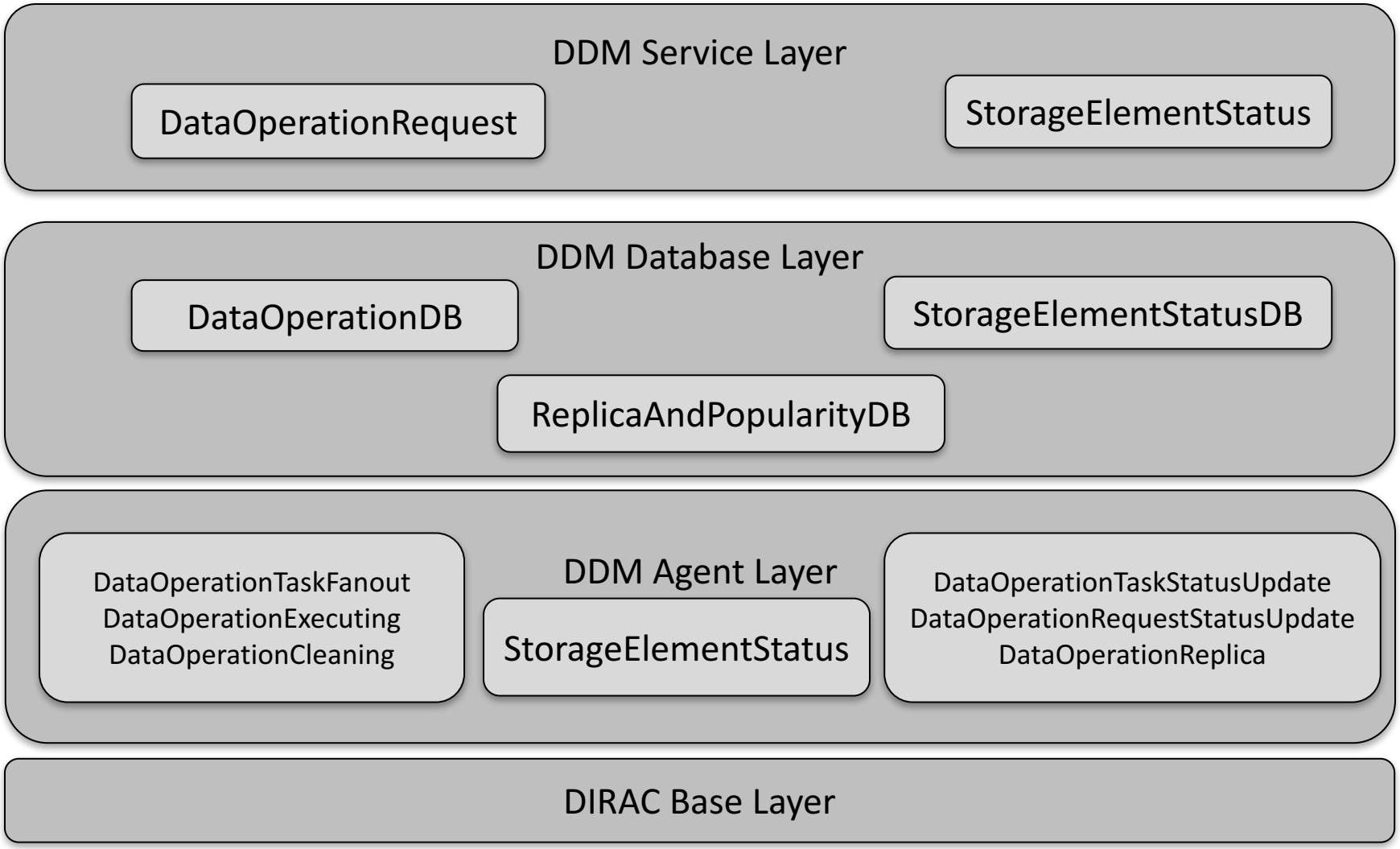
perfSONAR

MaDDash provides network monitoring.
Belle II sites are part of MaDDash mesh

GRIDFTP

GridFTP servers render file Transfers between sites

Belle II Distributed Data Management Overview



Belle II Networking

- **Description:**
  - Coordinate with the relevant National Research and Education Network (NREN) providers to organize and evaluate the WAN network status and requirements for the Belle II collaboration
  - Coordinate the development effort to integrate the networking information into the Belle II distributed computing framework
  - Establish a full design and specification of the data transfer and processing/reprocessing workflow between KEK and PNNL

- **Accomplishments:**
  - Full integration into LHCONE
  - Network upgrade by NRENs and major sites: 2x10Gbps dedicated optical path from Seattle, 1x10Gbps dedicated optical path backup path through Boise (tested)
  - Detailed network diagram for each major site (WAN, DTNs, storage backend)
  - Several Network Data Challenges (NDC) before/after SINET upgrade:
    KEK to PNNL transfer rates based on NDC results is 16Gbps (max 20Gbps)
  - NDC results demonstrates that Belle II networking requirements are satisfied

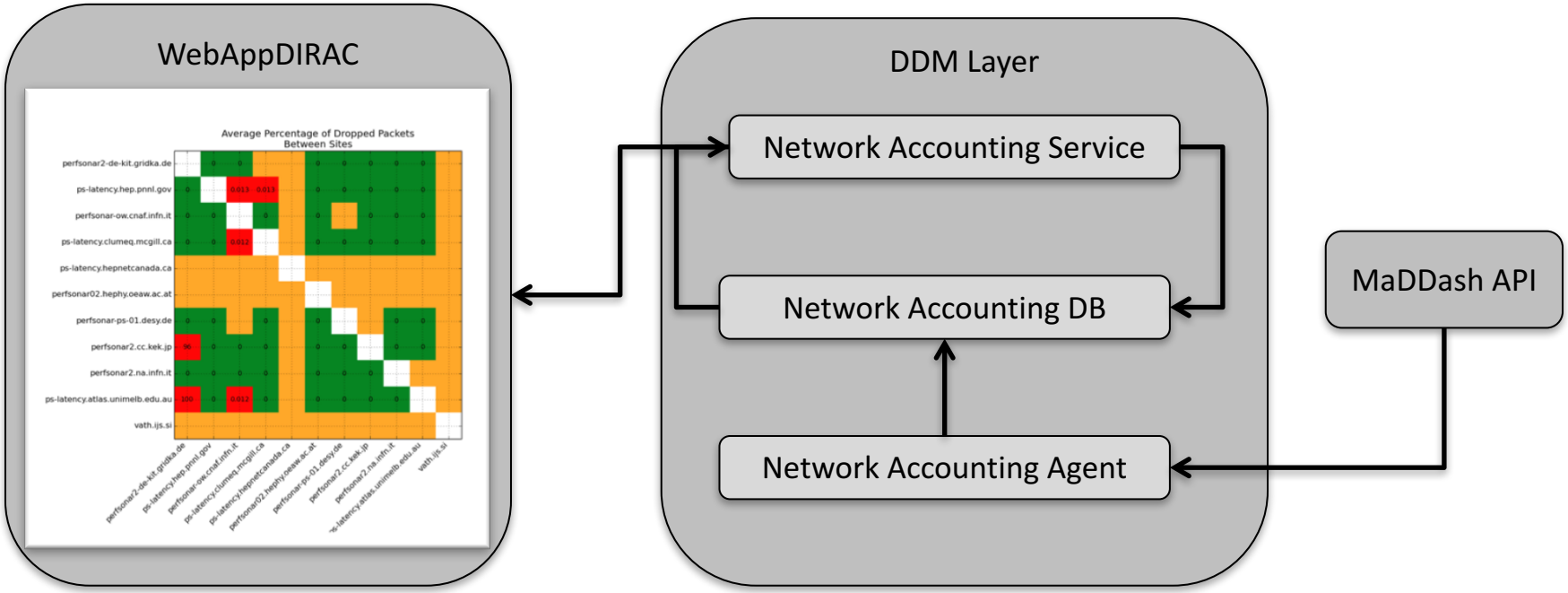
Latest Belle II Network Data Challenge Results

- A lot of moving part that requires frequent re-evaluation
  - Site reconfiguration: Network, Storage, Data Transfer Nodes, etc.
  - A stable perfSONAR mesh will simplify the network validation process
- New NDC is planned this fall

Source→ Destination↓	KEK (Gbps)	PNNL (Gbps)	DESY (Gbps)	KIT (Gbps)	CNAF (Gbps)	NAPOLI (Gbps)	SiGNET (Gbps)
KEK		6.2	11.0	5.0	9.2	15.0	3.0
PNNL	16.0		10.0	6.0	14.0	10.0	-
DESY	6.0	6.6		8.0	8.0	8.0	3.0
KIT	5.6	4.8	8.0		8.0	6.0	3.0
CNAF	18.0	14.0	10.0	6.0		8.0	3.0
NAPOLI	16.0	6.0	3.0	3.0	3.0		3.0
SiGNET	1.6	0.6	5.0	5.0	5.0	2.0	

Integrating Networking in the DDMS

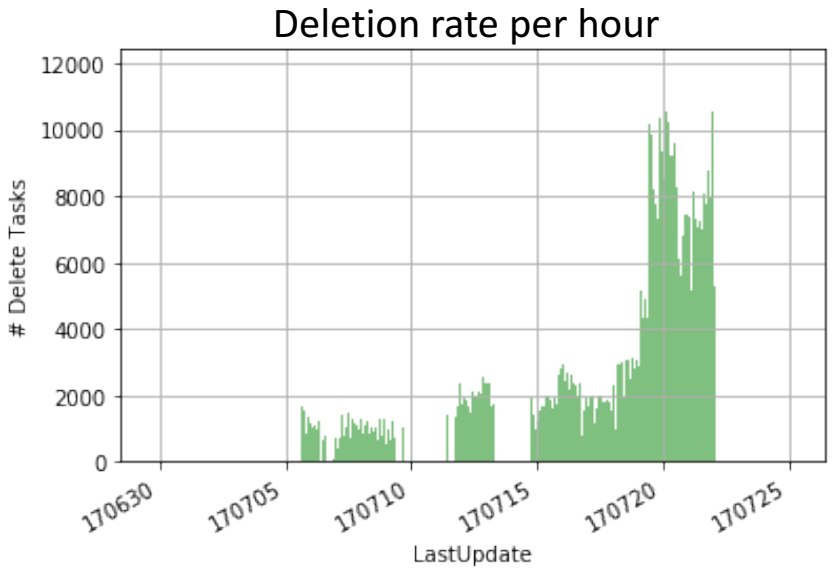
- perfSONAR servers set up at Belle II sites for network monitoring
- Belle II MaDDash mesh is online and provides bandwidth and latency information
- Network information is being part of distributed data management (DDM) inside DIRAC
- Automate notification to sites with network problems and storage status update



DDMS Status and Plans

Status:

- ✓ File replication rates reached ~20k files per hour
- ✓ File deletion rates reached ~10k files per hour



Short Term Plans:

- Datablock-level replica and popularity catalog implementation
- Performance tuning with multithread agents
- Streamline BelleDIRAC access to FTS3 server
- Job data access lock
- Integrate networking information