Pacific Northwest NATIONAL LABORATORY

Proudly Operated by Battelle Since 1965

Belle II Computing Overview

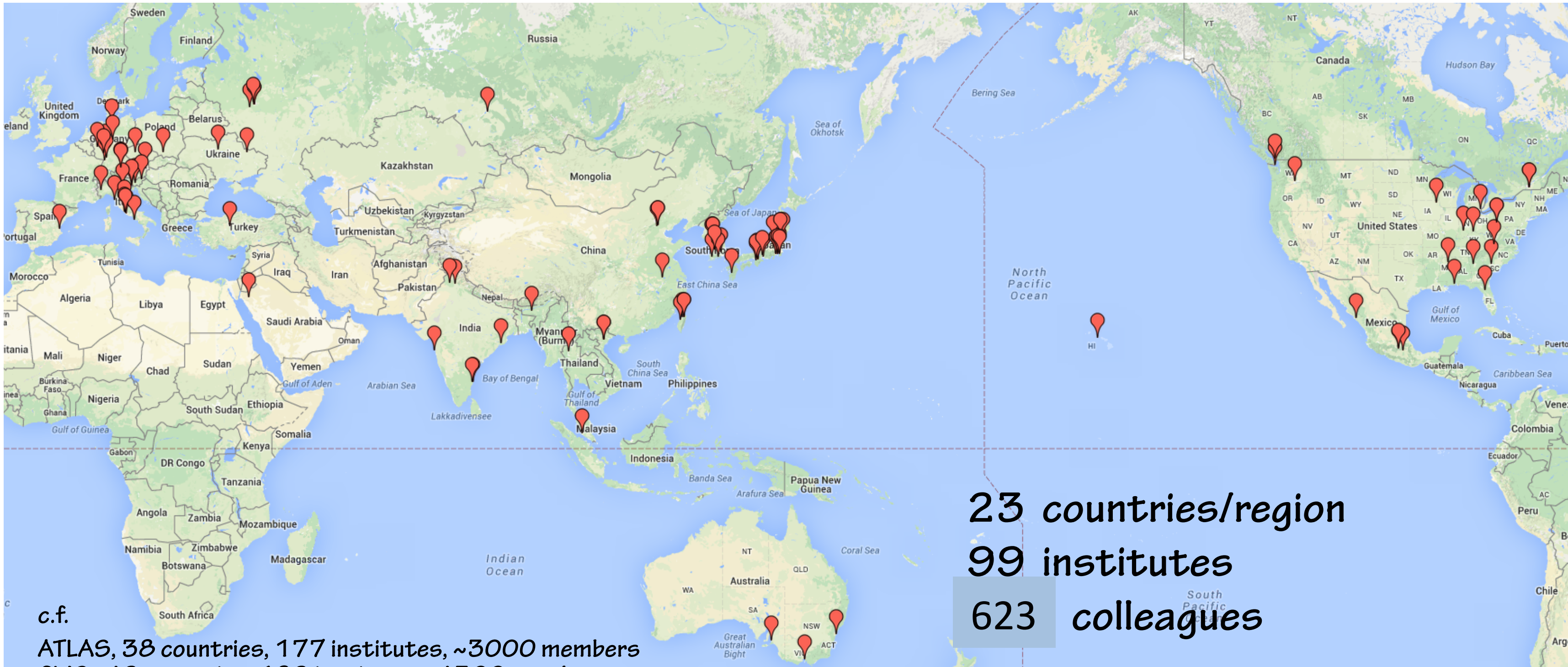
DAVID ASNER, DAVID COWLEY, KEVIN FOX, MALACHI SCHRAM, LYNN WOOD

Pacific Northwest National Laboratory

2015 Open Science Grid All-Hands Meeting

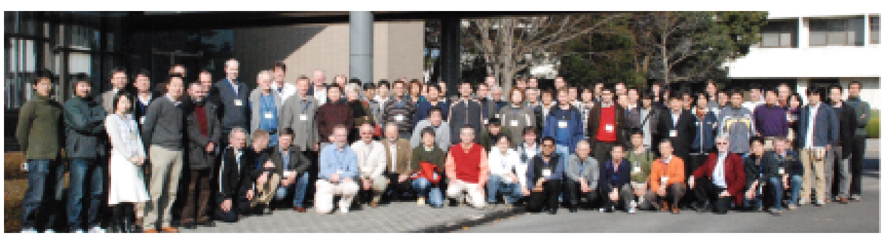
Belle II Collaboration

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965



23 countries/region
99 institutes
623 colleagues

c.f.
ATLAS, 38 countries, 177 institutes, ~3000 members
CMS : 42 countries, 182 institutes, 4300 members
ALICE : 36 countries, 131 institutes, 1200 members
LHCb : 16 countries, 67 institues, 1060 members



Asia : ~45%	N. America : ~16%	Europe : ~39%
Japan : 141	US : 72	Germany : 88
Korea : 36	Canada : 17	Italy : 61
Taiwan : 26	Mexico : 8	Russia : 38
India : 26		Slovenia : 17
China : 19		Austria : 14
Australia : 21		Poland : 11
		Czech rep. : 7

others : < 5 colleagues / country

Belle II Computing Model

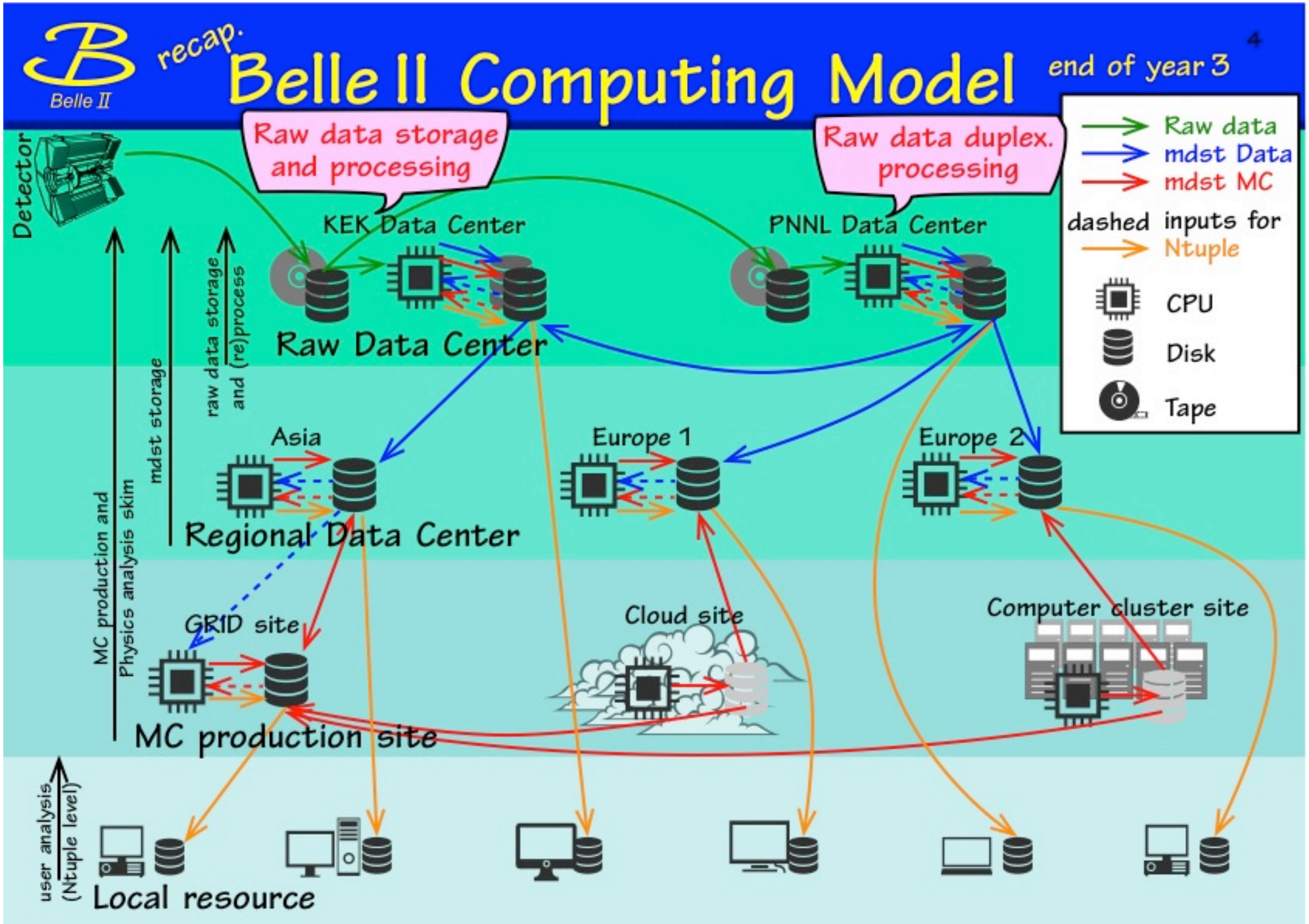
- The Belle II Computing model will manage in a geographically distributed environment the following main tasks:
  - RAW data processing.
  - Monte Carlo Production
  - Physics analysis
  - Data Storage and Data Archiving

- On going activities
  - Resource Estimation
  - Defining strategy for analysis and data distribution
  - Developing automated data replications and failover
  - Evaluating technologies

Site Categorization

- The Belle II Computing Sites are categorized as follow:
  - Raw data Center: Stores the RAW Data and data processing and/ or data reprocessing
  - Regional Data Center: Large data center that stores mDST and participates at the Monte Carlo production
  - MC Production site: Data Center that produces and stores Monte Carlo simulations, that included:
    - Grid Site
    - Cloud Site
    - Computing Cluster Site

Belle II Distributed Computing Model



US Belle II Computing Operations

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

- Provide all required Belle II services for proper site operations
- Lead the Belle II networking and data transfer effort
  - Coordinate network connectivity between sites (LHCONE, etc.)
  - Lead network data challenges
  - Network monitoring
- Develop and deploy components of the Data Management System
  - Develop Belle II Storage Accounting System (B2SAS)
  - Develop automated data replications and failover system
  - Develop and optimize FTS3/DIRAC interface
- Develop and deploy the Distributed Database System
- Participate in Monte Carlo and processing/re-processing efforts

US/PNNL Role: 15% of Belle II Compute/Data

- "*Each Belle II member institute provides… the computing resources to produce, store, and make available for analysis a fraction of the total simulated dataset… The fraction corresponds to the fraction of PhDs in the international Belle II collaboration*"

- PNNL is ideally suited to provide this capability:
  - Office of Science Laboratory
  - Core capability in advanced computer science, visualization, and data
  - EMSL and ARM user facilities
  - Wide Area Network Connectivity via ESNet
  - Ability to draw on resources across organizational boundaries/disciplines

- Belle II Resources come from multiple entities
  - PNNL: Olympus nodes, Data Transfer Node, /pic filesystem space, 10 Gigabit WAN connectivity
  - US-Japan: Servers, Storage, Network switches
  - DOE: Labor, Hardware

PNNL Belle II Site Planning

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

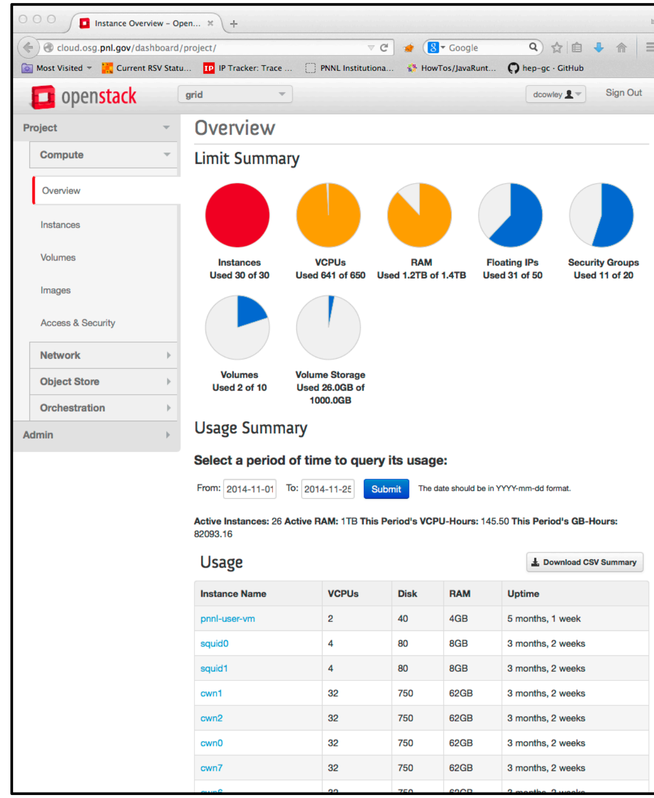
- Underlying infrastructure

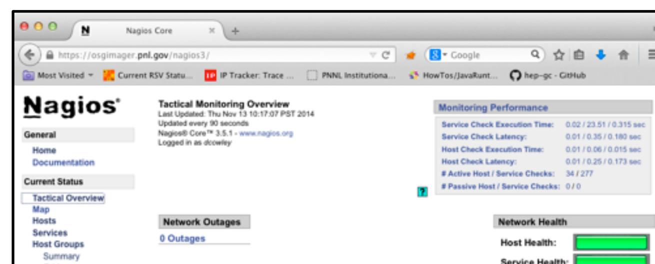
- OSG Software Stack
  - Authorization Service
  - Compute Element
  - Worker Nodes
  - Storage Elements

- Other Software
  - DIRAC servers
  - CVMFS stratum-0 and clients
  - Squid servers
  - Conditions database services
  - FTS3 servers
  - perfSONAR servers and mesh

Established System Support

- Experienced computational staff matrixed to the project:
  - High Performance Computing Operations
  - Wide Area Networking
  - Database Design and Development
- Dedicated 10 Gbps LHCONE
- Shared 10 Gbps via Public Internet
- Set up Belle II data infrastructure
  - 175 TB storage capacity
  - Mixture of Lustre filesystem and CEPH distributed object store
- Implemented system administration infrastructure with monitoring and trouble ticket systems for Belle II



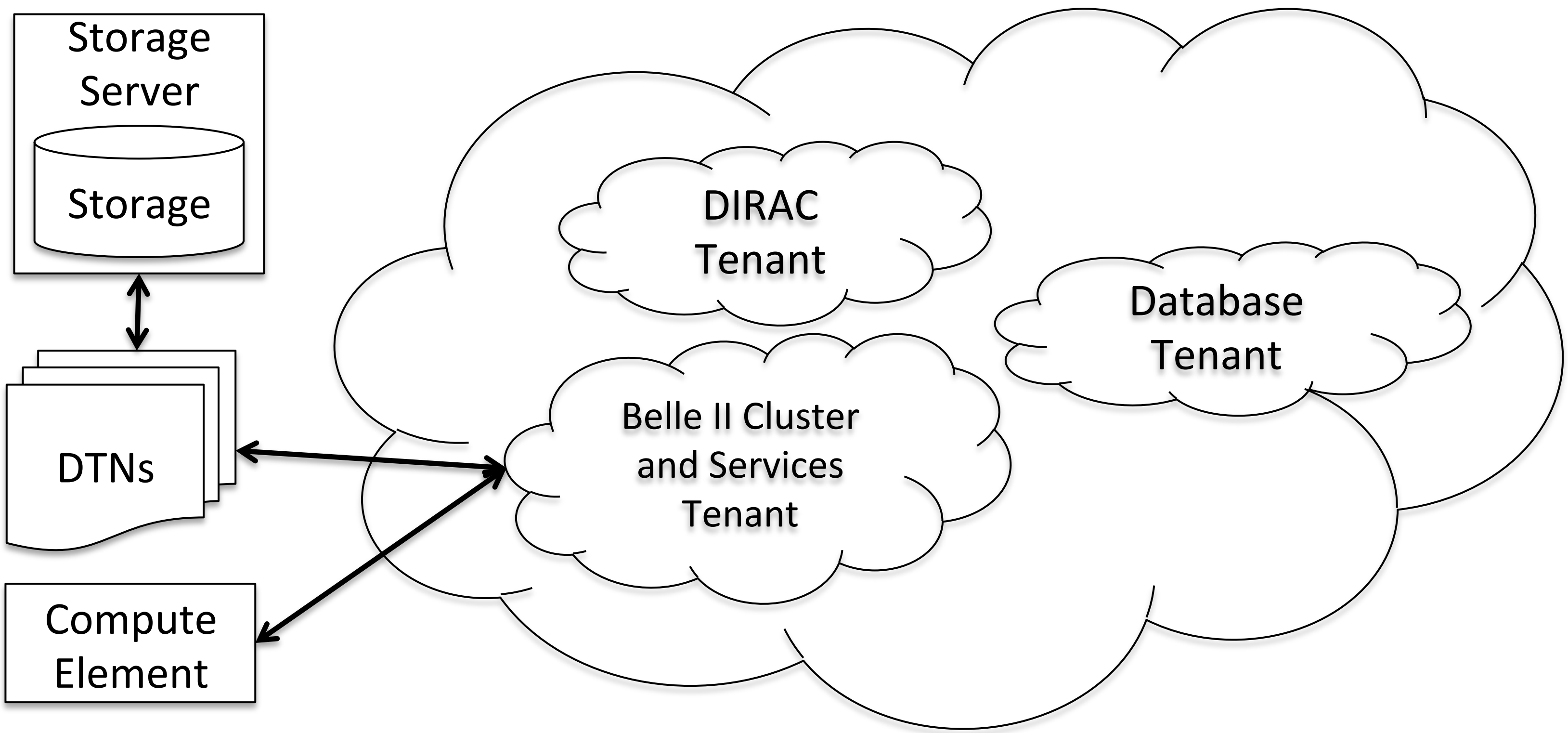


Established Software Stack

- Standardized on Scientific Linux 6 plus OpenStack private cloud technology for compute services
  - Enables rapid prototyping, development, testing, deployment of services
  - Increases flexibility, reliability of PNNL Belle II resources
  - Provides better utilization of compute resources by provisioning virtual machines
  - Job throughput (as reported by DIRAC) is *twice* that of the Olympus (PIC) system, using identical hardware
- Demonstrated that HEP jobs could productively be run in backfill windows on EMSL Cascade supercomputer

Creating a Private PNNL HEP Cloud

- Cloud tenants enable separation of production and development environments
- Current activities in development and testing on cloud tenants:
  - Grid services: CE, SE, Squid, CVMFS clients
  - Database: REST API, BASF2 Conditional Databases Services, CVMFS stratum-0
  - BelleDIRAC: Metadata catalog & distributed data system



Compute Element

- Originally using OSG CE stack (push-method)
  - Interaction between DIRAC/OSG-CE is inefficient for job scheduling
  - Limited batch system support (Condor, PBS, LSF, and SGE are presently supported)
  - Using CondorHT for dedicated HEP cluster queue (grid & local use)
    - Previous updates removed site specific configurations
- Migrate to DIRAC SiteDirector (pull-method)
  - Quick solution to increased job scheduling efficiency (nearly 100%)
  - Developed for SLURM batch system support for PNNL HPC resources
  - Easier to get cyber security approval
- We are re-visiting OSG CE stack to provide opportunistic compute cycles to OSG community

Worker Nodes

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

- Using OSG WN stack
- Worker nodes are launched on our private OpenStack cloud
- Using heat templates to configure/launch node in cloud
  - Base heat template and/or VM images from OSG?
- Using CVMFS client to access experiment repos (Belle II, ILC, and other HEP experiments)
  - Proxy through a local load balanced Squid farm
  - Access remote and local CVMFS stratum-0 and stratum-1

Storage Element

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by **Battelle** *Since 1965*

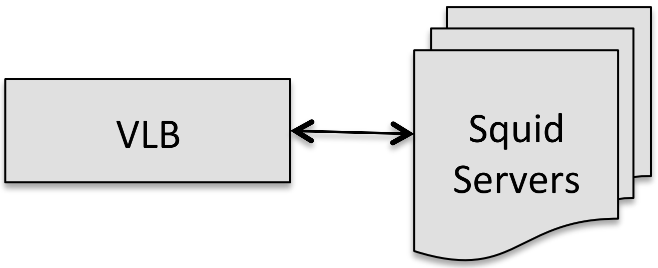
- Using Lustre backend file system

- Initially using Bestman2, however, problems:
  - Randomly unresponsive
  - Adler-32 support not working properly
  - Updates overrides existing configurations files
  - Unclear how to scale system: stateless vs. stateful information
  - No response from developers

- Investigating Storm SE
  - Clear stateless/stateful separation for scalability
  - Support Adler-32 out of the box
  - Active development and quick response

Other Services

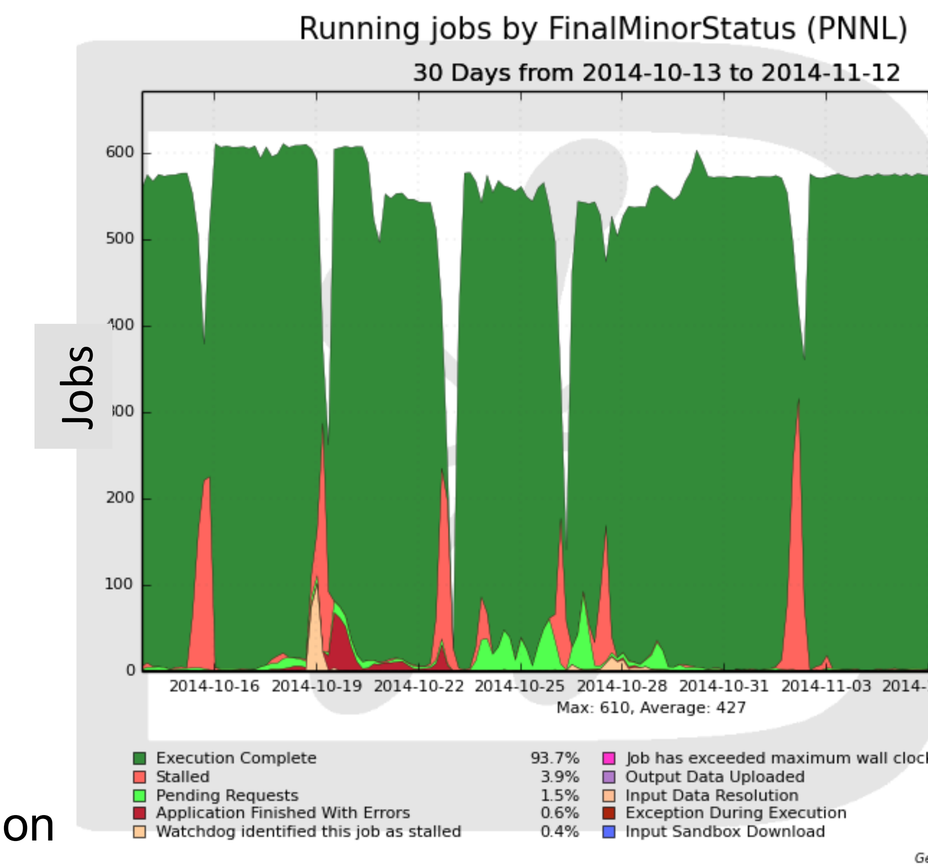
- Squid farm behind a Virtual Load Balancer
  - Common external hostname
  - Several modes to load balance



- CVMFS Stratum-0 at PNNL
  - Belle II conditional database payload repository
  - Other HEP experiments software (superCDMS, MiniCLEAN, etc.)
- File Transfer Service 3 (FTS3)
  - Used for mass data transfers
  - Clear stateful/stateless separation for scalability
- PerfSONAR server and mesh
  - Used for network monitoring
  - perfSONAR servers are online
  - Dedicated Belle II perfSONAR mesh is ready for testing

Belle II MC4 Monte Carlo Campaign Successful

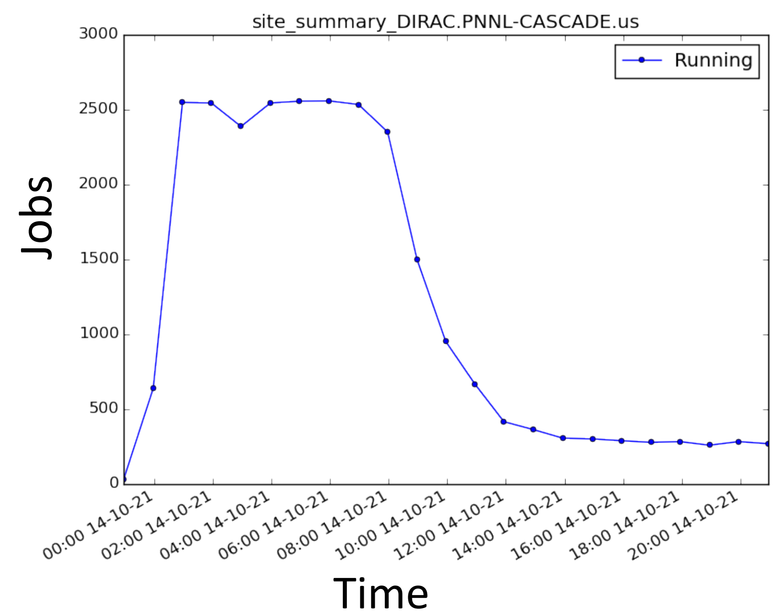
- September 22 – November 7 2014
- Typically ~576 jobs running
- PNNL Belle II Grid Site contribution:
  - 24 compute nodes
  - 90,102 jobs
  - 22,128.3 Wall Clock Days
- PNNL Cascade Supercomputer contribution:
  - 63,132 jobs
  - 5,804 Wall Clock Days
- Issue: periodic job stalling events.
  - Cause: aggressive port scans by PNNL's own cyber security organization
  - Fix: get them to turn down intensity



Enabling Opportunistic HPC Computing for HEP

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by **Battelle** *Since 1965*

- Ideal resource for stressing the distributed system while delaying procurements of dedicated resources
- PNNL has several large clusters (e.g. Cascade, etc.)
- We setup a gateway node on Cascade to provide gateway services for HEP computing (ATLAS/Belle II)
  - NAT through gateway node allows Cascade nodes
    to talk to selected HEP sites
  - Appropriate firewall changes were made
- OSG client software was integrated into Cascade node image
- Developed Belle II DIRAC SLURM Site Director agent (submitted to DIRAC developers)
- New queue was setup on Cascade to allow multiple single-core jobs to share a compute node
- Approximately 10% of the available CPU on Cascade was used at one time (>20kHEPSpec)



Proof of Concept: HEP jobs on DOE HPC

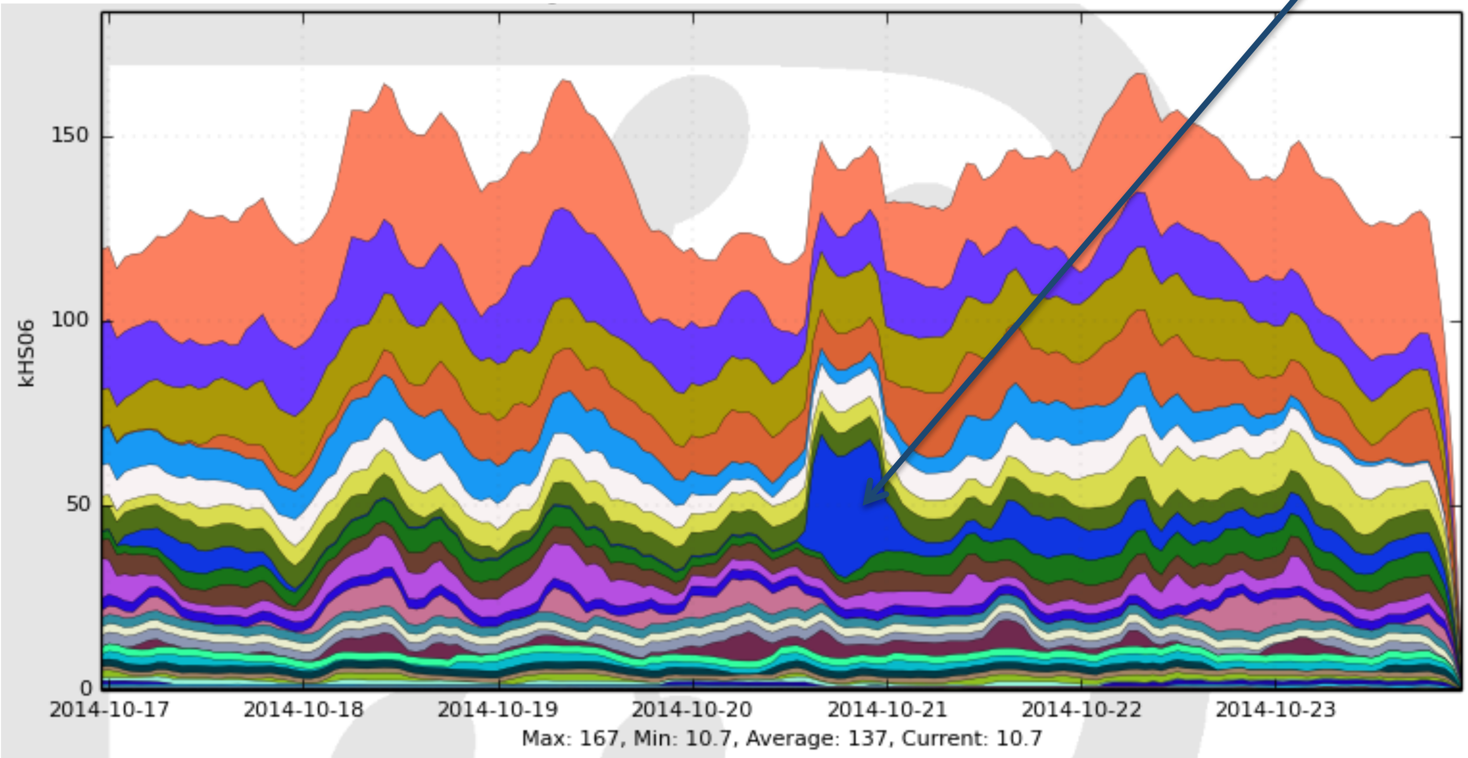
- Goal: Demonstrate that HEP jobs can run in “backfill” windows on parallel supercomputers
- 20,000 node hours allocated on PNNL/EMSL Cascade supercomputer
- DIRAC Site Director, Squid, and CVMFS set up on an ancillary Cascade node
- Cascade queuing policy adjusted to:
  - Create “Shared” queue that allowed more than one job on a node
  - Not clean the entire compute node when a job completed

HEP on HPC Lessons learned

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

- HEP workload can backfill very effectively between large parallel jobs!
- HEP jobs need to be tailored to fit well on parallel HPC systems
- Opportunities:
  - Run multicore jobs
  - “Right-size” job length for the target system

PNNL Cascade

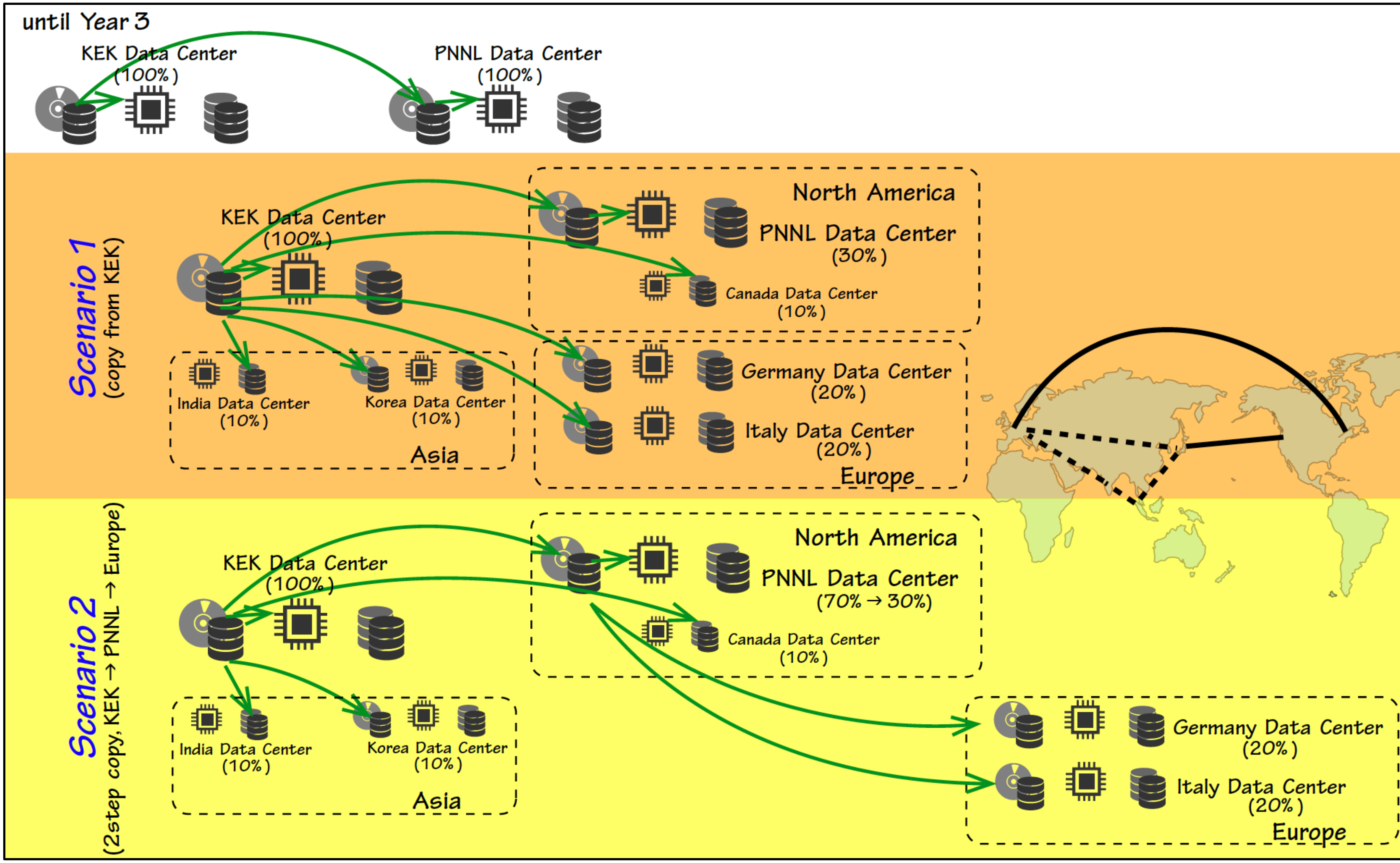


Networking & Data Transfer System

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

- Latest network requirements for Belle II estimates a total of 35Gbps in/out at designed luminosity
  - Does not include user analysis
- FTS3 Server:
  - Previously used for Data Challenges and performance studies
  - Requires continuous studies of the FTS3 transfers to tune the channels as needed (FTS3 optimization not perfect yet)
- DIRAC Integration:
  - Prototype Belle II FTS3 DIRAC Agent was developed to automate the Data Challenges and to gain experience with DIRAC/FTS3
  - Working on doing DIRAC transformation using FTS3
  - Monitoring and tuning distributed computing data transfers will be vital to stable operations
    - Using WLCG FTS3 Dashboard, however, we should move to DIRAC monitoring
    - We will deploy Belle II perfSONAR mesh this spring to provide automated network monitoring

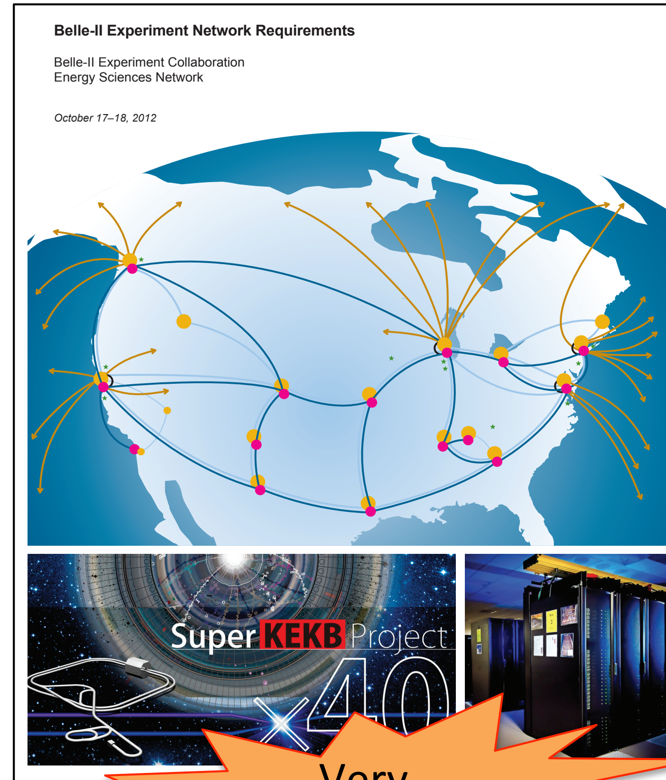
Belle II Raw Data Distribution



Goals Achieved based on Pacific Network and Computing Requirements

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

- The purpose was to address the wide-area networking requirements for science in general and of the Belle II experiment in particular.
- Report can be found at:
  http://www.es.net/assets/pubs_presos/Belle-II-Experiment-Network-Requirements-Workshop-v18-final.pdf
- Various goals were defined, for example:
- Setup KEK-PNNL VC — Done
- Preliminary data challenge goals are:



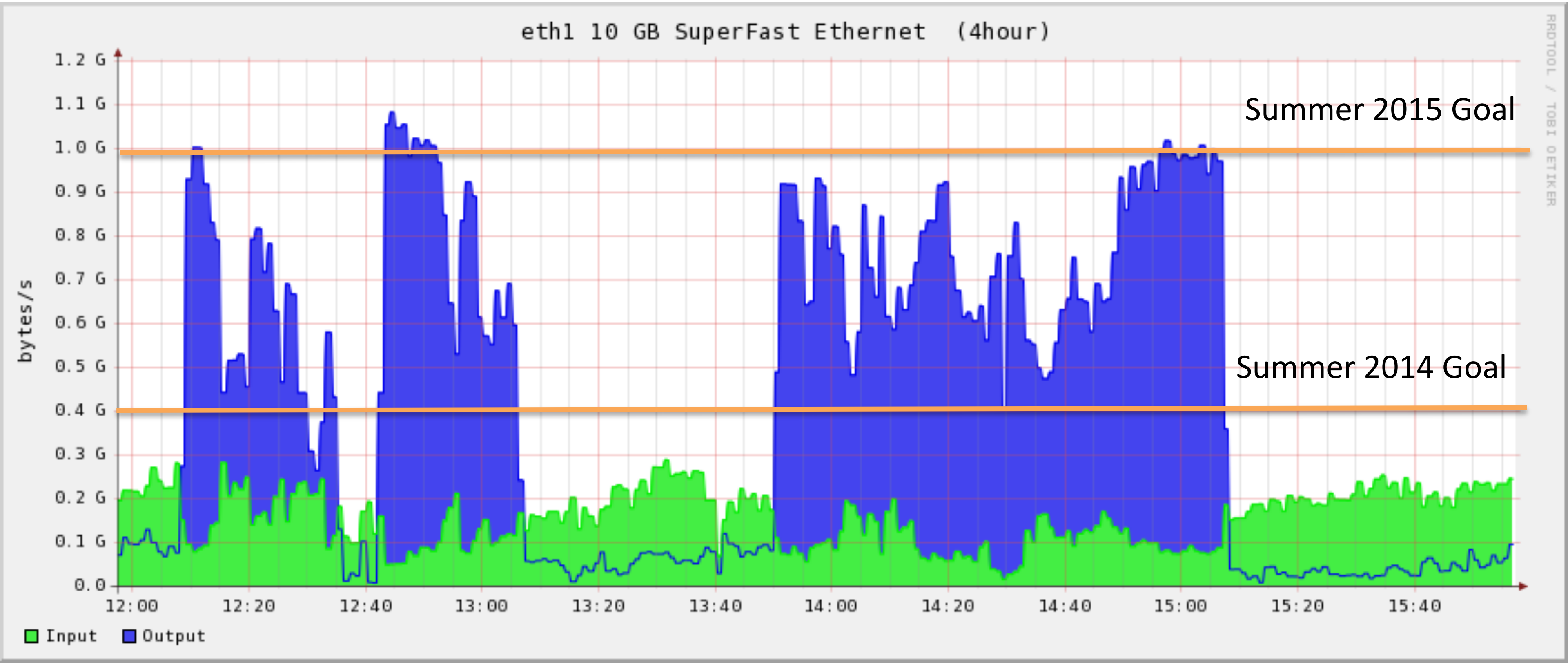
Achieved — Achieved — Very Challenging

Date	Summer 2013	Summer 2014	Summer 2015
Rate	100MB/sec	400MB/sec	1000MB/sec
Duration	24 hours	48 hours	72 hours

Higher Data Transfer Rates Achieved using File Transfer Service 3 (FTS3)

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

- Tuning FTS3 allowed us to reach ~8.8 Gbps transfer rates from PNNL to KEK
- Load balanced cloud implementation is currently being tested



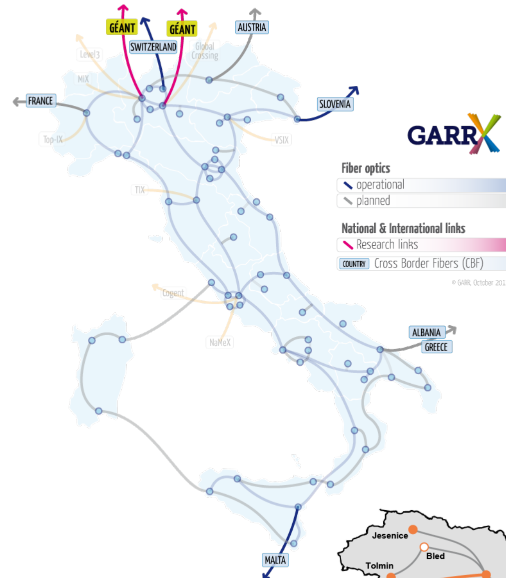
Goals Achieved based on European Networking Workshop

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

- Network requirements and milestones need to be reviewed

- The purpose of this workshop was to begin preparation for addressing the wide-area networking requirements for science in general and of the Belle II experiment in Europe

- Initial milestones:



Achieved | Achieved* | Achieved*

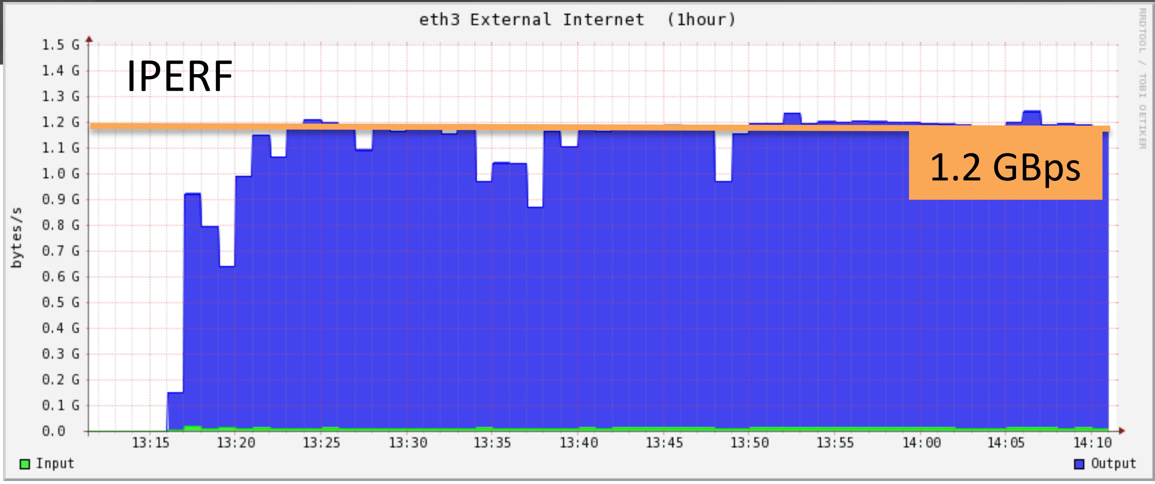
Date	Winter 2013	Summer 2014	Summer 2015
Rate	100MB/sec	200MB/sec	400MB/sec
Duration	24 hours	48 hours	72 hours

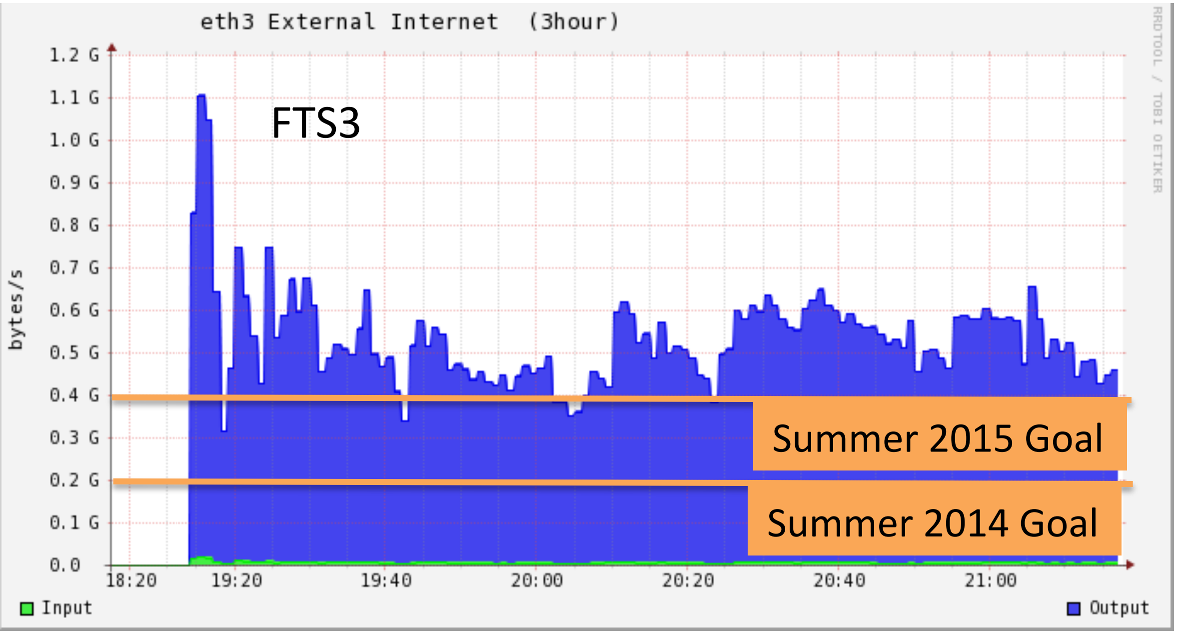
*Using ANA-100 link

Trans-Atlantic Goals Achieved using Belle II ANA-100 Setup

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

- First few days we conducted tests using *IPERF* for true network testing
- Required several parallel transfers to reach network saturation





- Transitioned to FTS3 server
- Reached network saturation but rates fell very quickly
- Large amount of dropped packets

Including Belle II into LHCONE

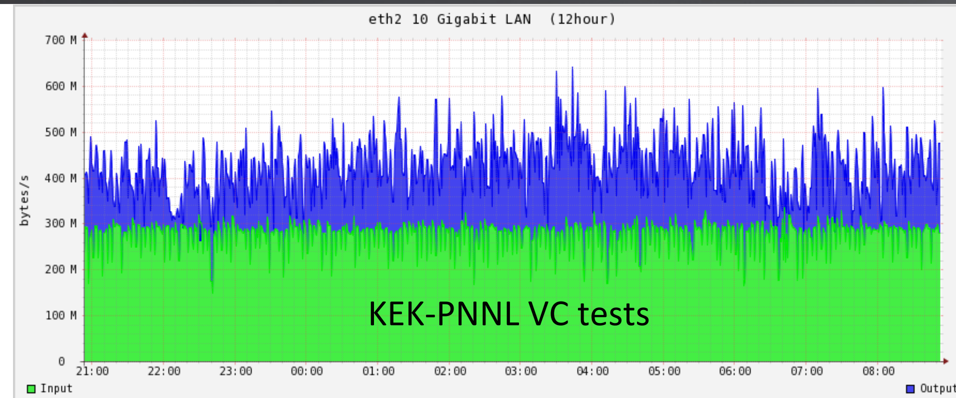
Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

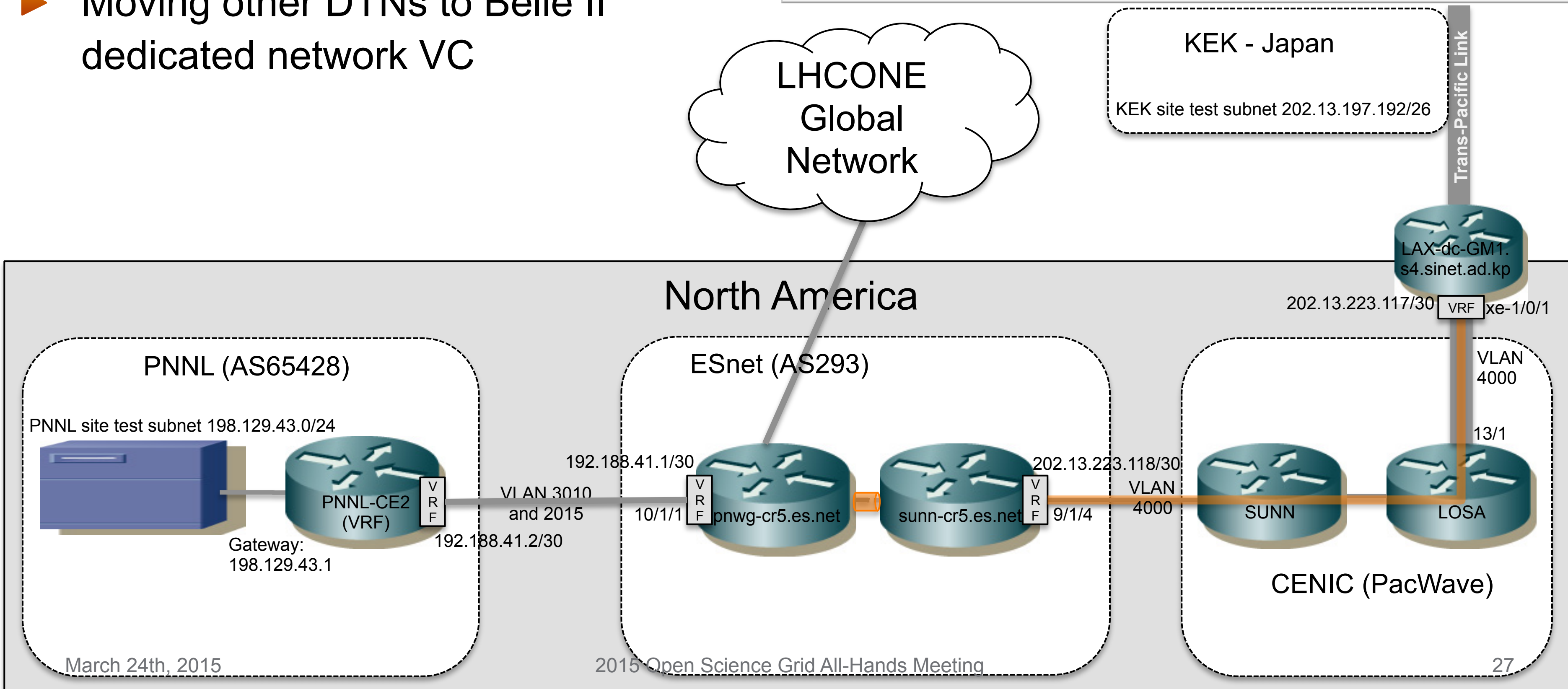
- The network requirements for Belle II are similar to the LHC experiments.
- Most Regional Data Centers are already part of LHCONE
- Belle II presented the computing/networking requirements at LHCONE Meetings
  - LHCONE Asia-Pacific Workshop - Nantou, Taiwan, 13 August 2014 (presented by Takanori Hara)
  - LHCOPN-LHCONE Meeting - Ann Arbor, Michigan, Sept. 15-16th 2014 (presented by Malachi Schram)
- LHCONE agreed to add Belle II at the Ann Arbor meeting
- New LHCONE Acceptable Use Policy explicitly includes Belle II: https://twiki.cern.ch/twiki/bin/view/LHCONE/LhcOneAup

Connecting PNNL, KEK, and LHCONE

(Setup to stay in place thru June 2016)

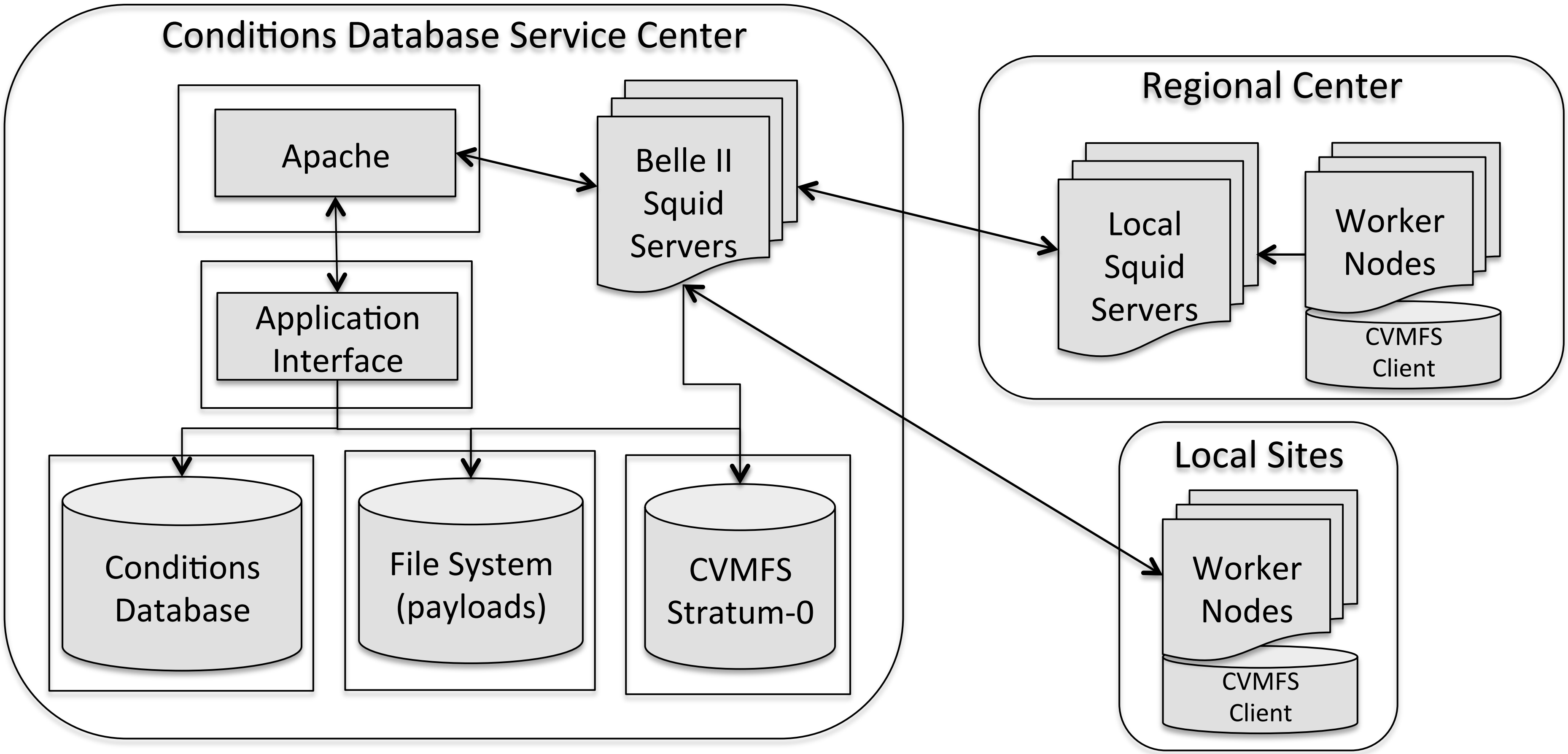
- KEK-PNNL VC transfer tests are ongoing
- Debugging LHCONE connection
- One dedicated Data Transfer Node
- Moving other DTNs to Belle II dedicated network VC





PNNL Distributed Database Testbed

- PNNL was explicitly requested by Belle II spokesperson and the computing and software coordinators to lead the Belle II database effort
- Provided a plan and the infrastructure to test the distributed database effort
- Belle II effort is being developed within a dedicated cloud tenant at PNNL



Summary

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

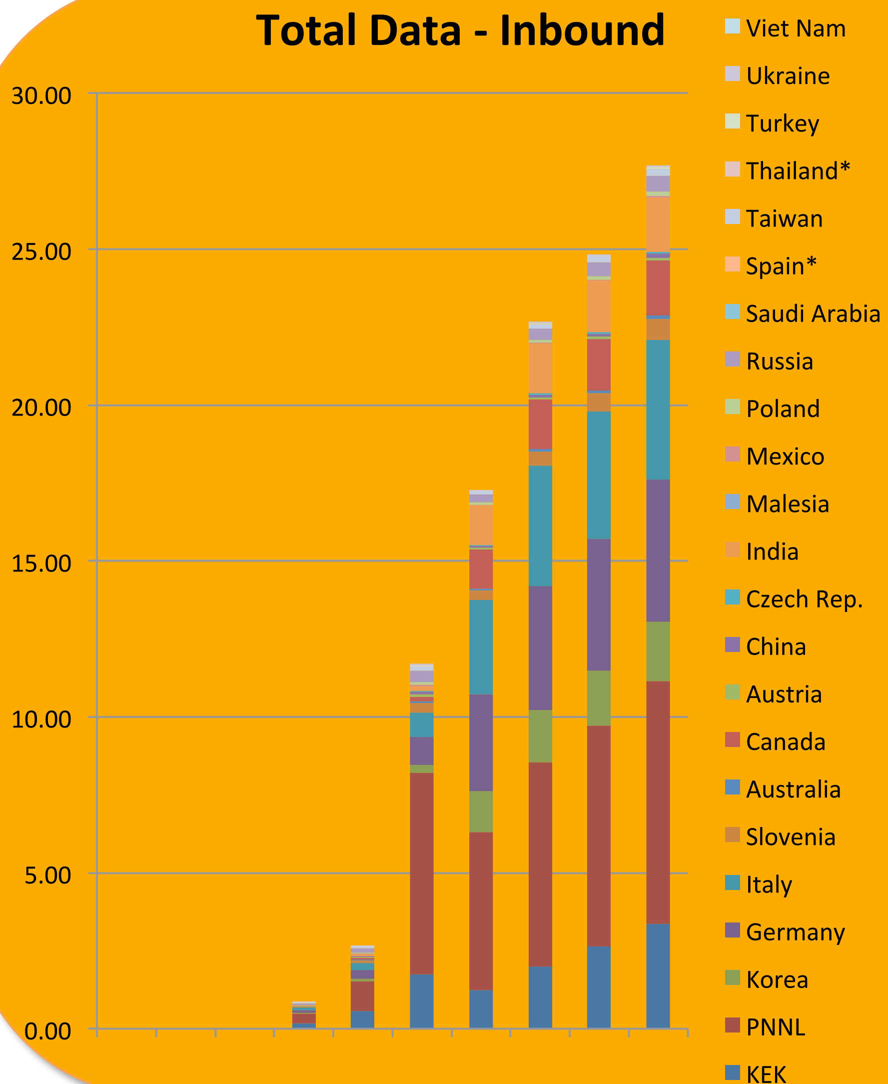
- Belle II Computing requirements are very similar to those of the LHC experiments
- PNNL is the led US laboratory and the primary off site computing center providing numerous grid services
- PNNL computing infrastructure is based on Cloud technology
  - PNNL has deployed several OSG software stack
  - Experienced problematic upgrades that overwrite site configurations
  - Will OSG provide Cloud templates or VM for grid services in the future?
- Example comments from Belle II University member:

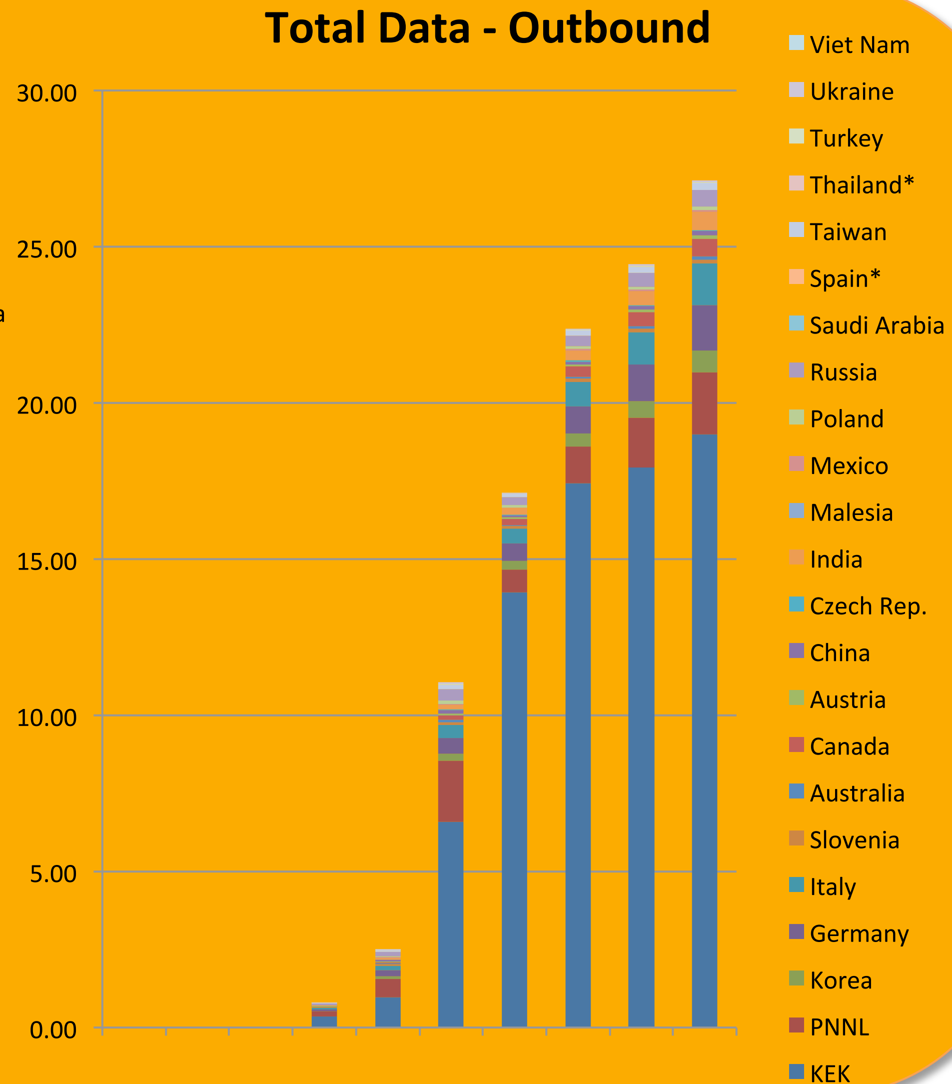
  "What we found was that setting up a new CE or SE from scratch is actually quite difficult; at some point we needed a dedicated person either on-site or off-site to help us, at least until the set-up was complete."

Backup Slides

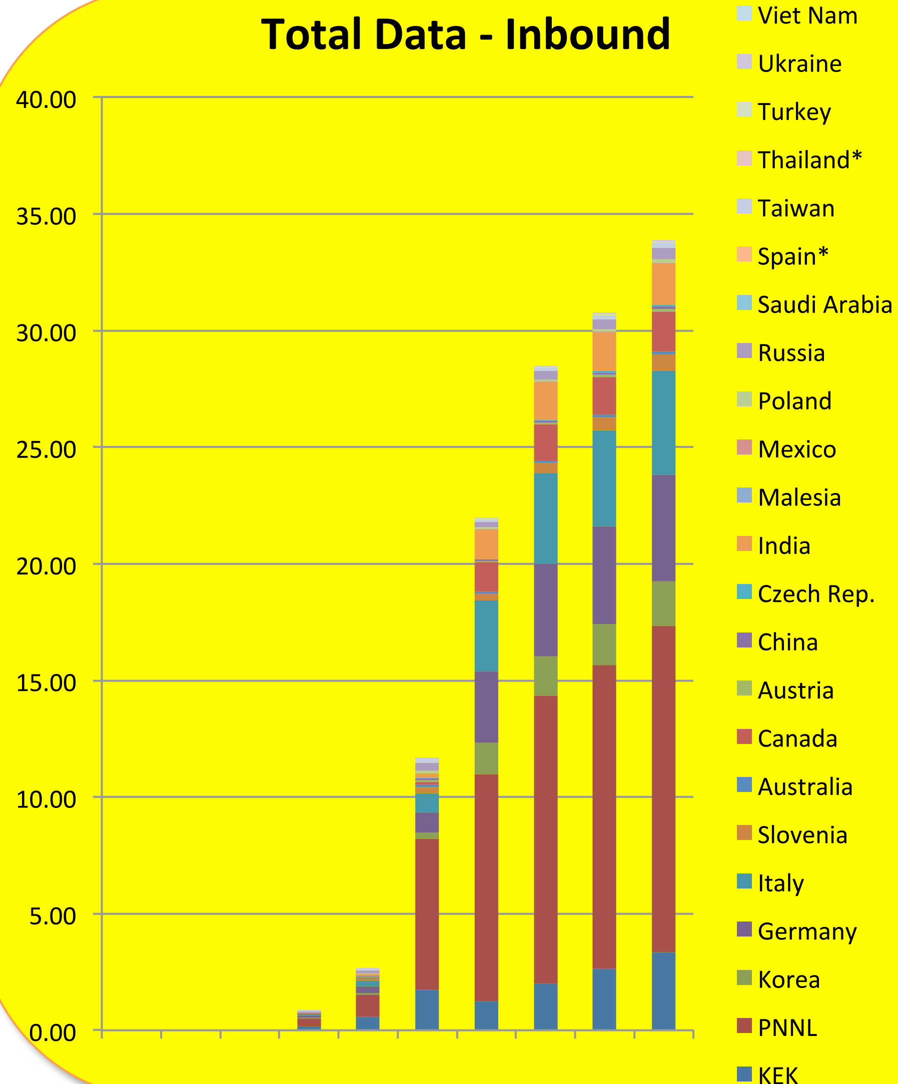
Belle II Total Data Network Requirements Scenario 1

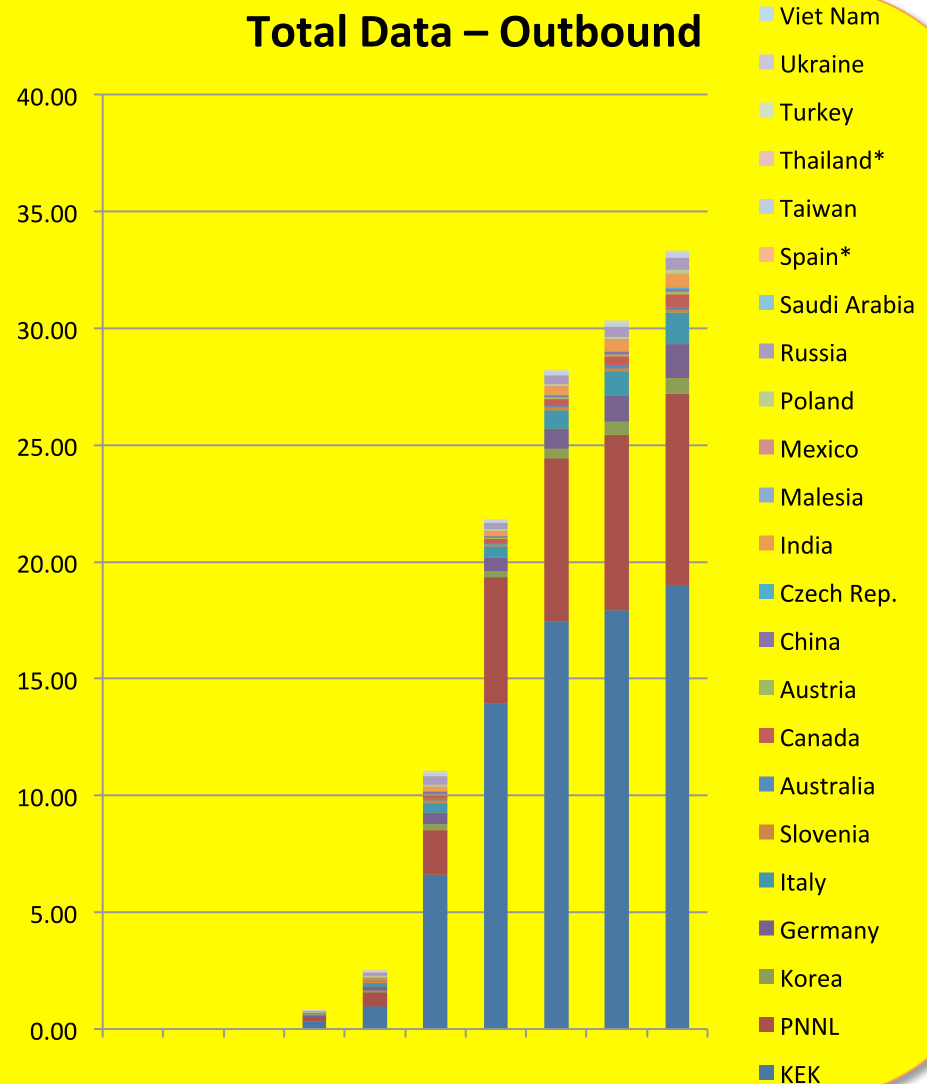
Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965





Belle II Total Data Network Requirements Scenario 2

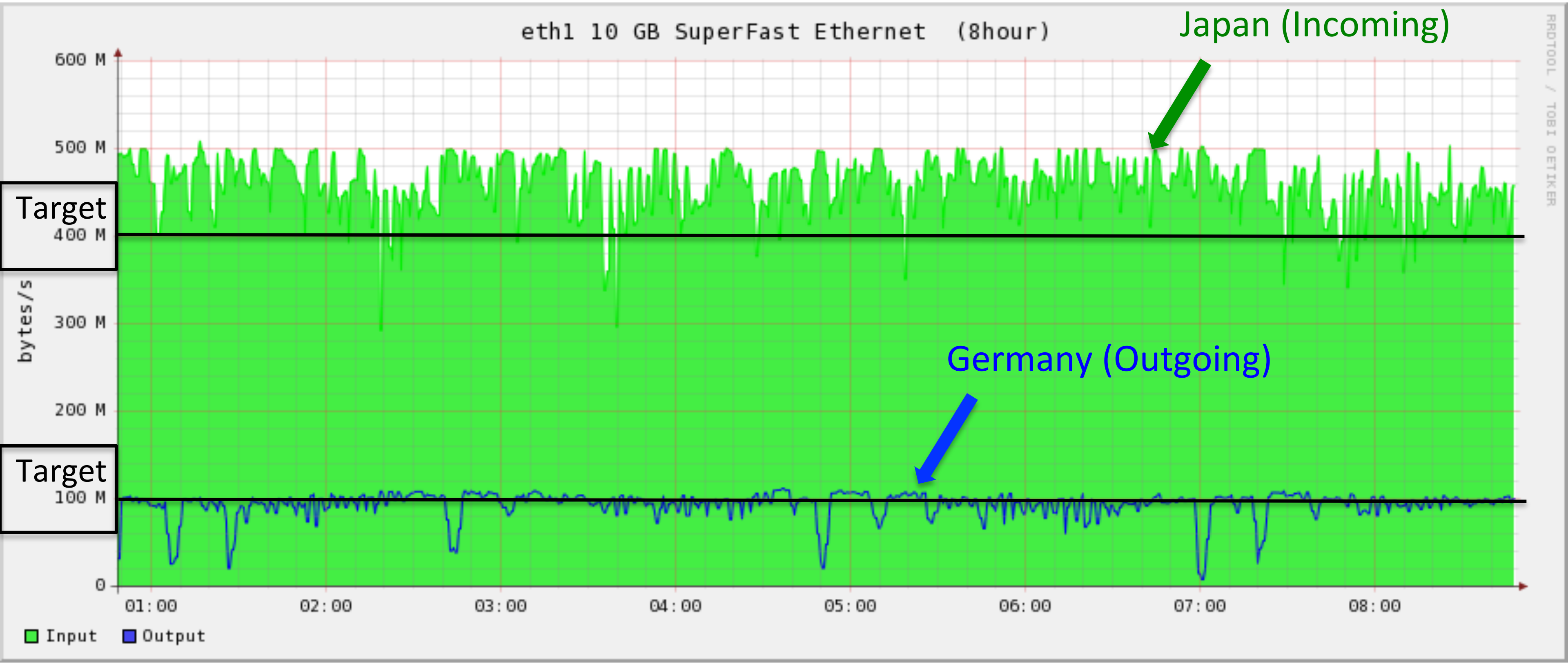




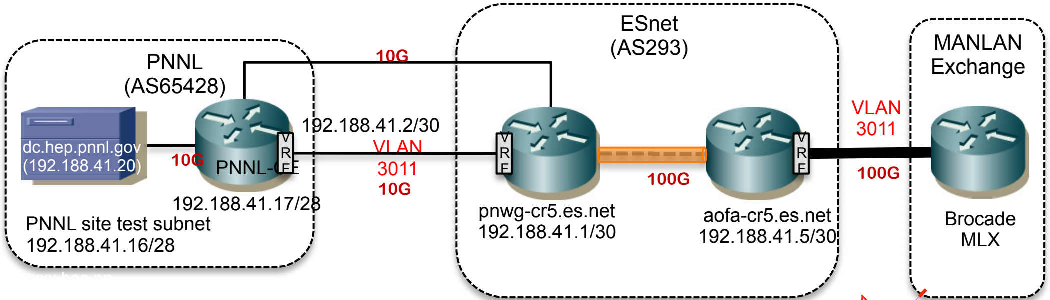
Summer 2014 Networking Goals Achieved

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by **Battelle** *Since 1965*

- Transfer rate from KEK to PNNL during 48hr stability test met summer 2014 goal (>400MBps)
- Transfer rate from PNNL to Germany during 24 hr stability test was >100MBps

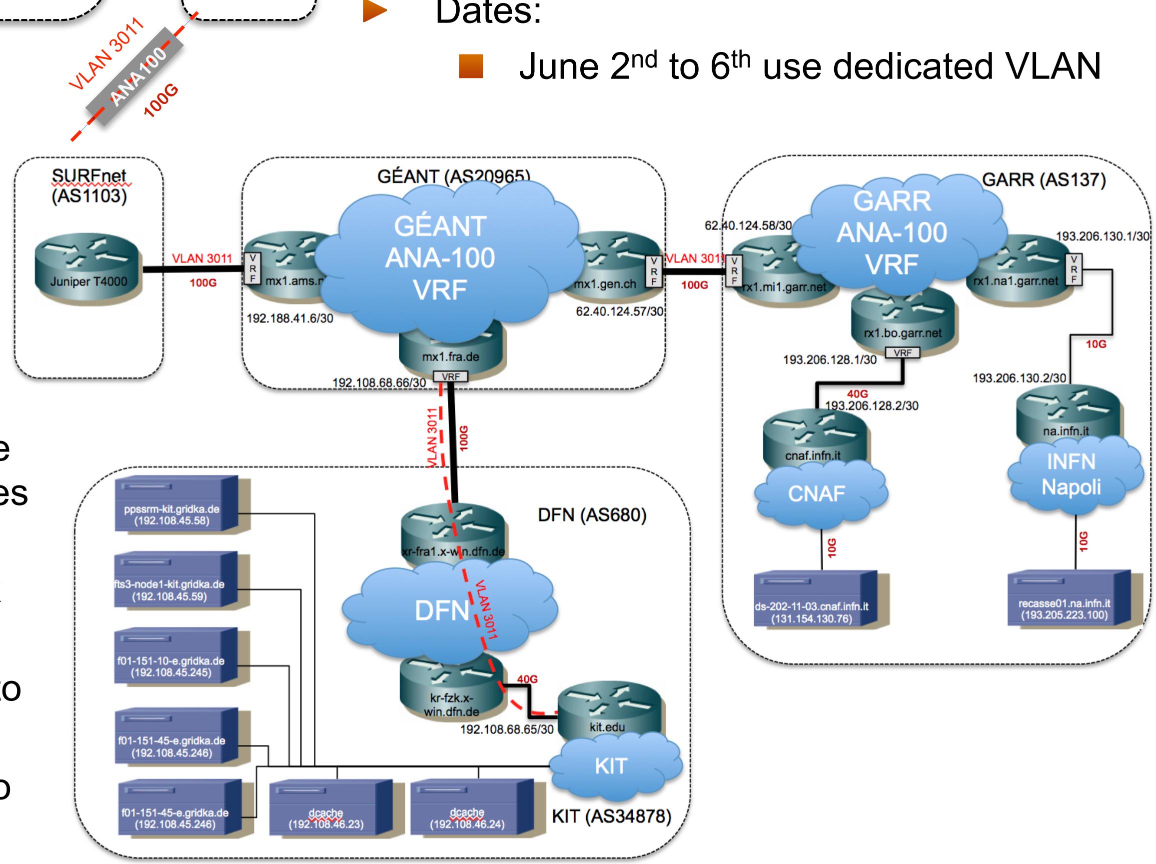


Setting Up Belle II ANA-100



- Goal
  - Test the ANA-100 Trans-Atlantic link
  - Test/tune/profile the performance of current Belle II data transfer tools
- Dates:
  - June 2nd to 6th use dedicated VLAN

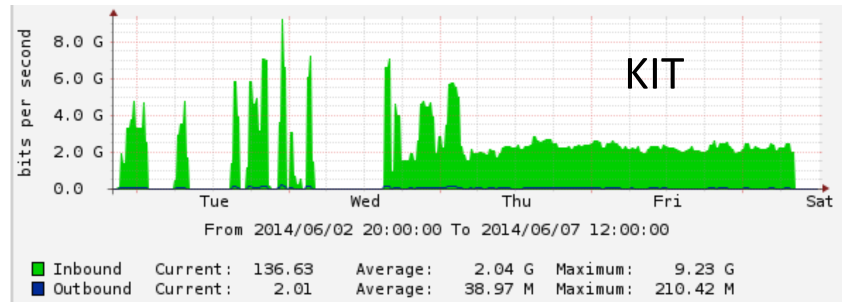
- Network Setup:
  - Network providers (Geant, ESnet, GARR, DFN) setup the VLAN
  - Local network providers and sites coordinated final configurations
  - Sites must configure hardware interface to match destinations
- Testing Tools:
  - *Traceroute* was used to confirm the routing to each Data Transfer Nodes (DTN)
  - *Iperf* was used to do initial network transfer rate tests
  - *gridftp* and/or *srm-copy* was used to test site
  - *FTS3 server* at GridKa was used to schedule data transfers

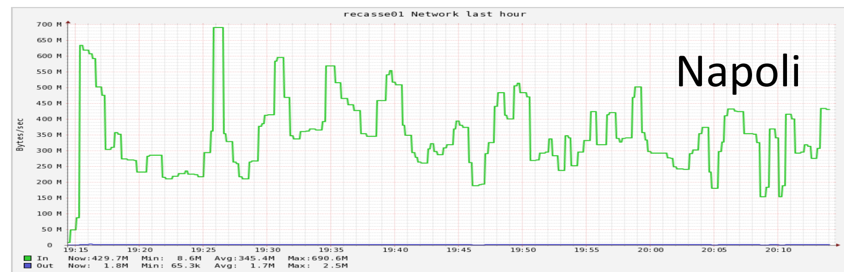


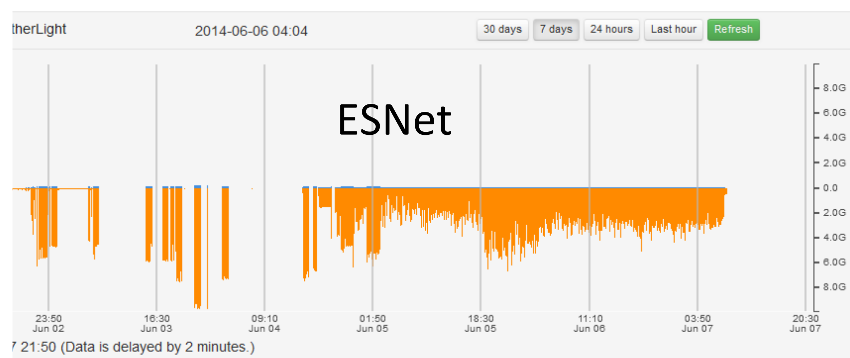
Belle II ANA-100: Lessons Learned

- Challenges encountered:
  - The main issue was the configuration of the local network apparatus
  - Having all the servers at each site using/ checking the proper network route
  - Hardware limitation (router, storage, etc.)
  - Not having dedicated setups (shared with ATLAS, etc.)

- Modification to sites to accommodate the increased rates:
  - Modification of TCP windows was performed at PNNL and Italy
  - Routing hardware interface
  - Configure/tune network interrupts for multicore
  - Modification of the FTS3 optimization







Sites and Data used for File Transfers

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

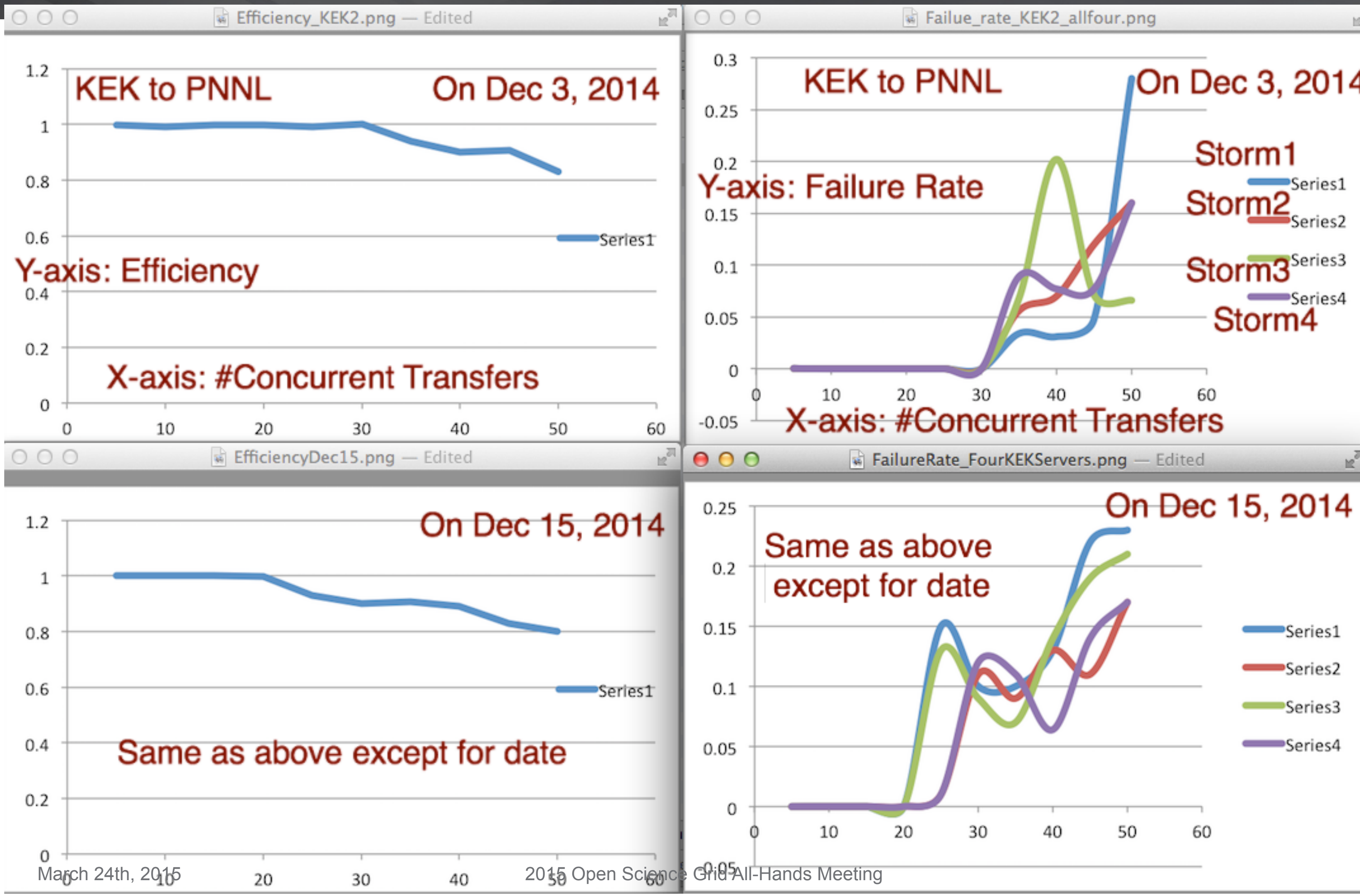
Sites:

- KEK2 - kek2-se01.cc.kek.jp
- PNNL - se.hep.pnnl.gov
- KIT - gridka-dcache.fzk.de
- DESY - dcache-se-desy.desy.de
- UVIC - charon01.westgrid.ca

Setup:

- All Regional Data Centers except KEK are now using LHCONE
- ~100 files with each file ~10 GB => This implies ~1 TB of data
- Each SE will temporarily require 2 x 1 TB of disk space
- Each SE has:
  - <SE-Path>/belle/DC2014/in
  - <SE-Path>/belle/DC2014/out
  - All previous files were deleted from belle/DC2014/ using srmrm command.
- 5x4 FTS3 transfers are initiated for 5 sites with each job assigned 20 files
- Channels are configured using fts-config-set + json

Transfer Efficiency & Failure Rate: From KEK to PNNL



WLCG FTS3 Belle II Dashboard

0 %

DESTINATIONS

	TOTAL-	charon01.westgri	dcache-se-desy.d	gridka-dcache.fzk	kek2-se01.cc.kek	se.hep.pnnl.gov+
TOTAL-	98 % 1 GB/s 8498 130	100 % 182 MB/s 1374 2	98 % 309 MB/s 2322 52	99 % 207 MB/s 1554 9	97 % 238 MB/s 1787 51	99 % 195 MB/s 1461 16
charon01.westgrid.ca+	97 % 19 MB/s 140 4		97 % 13 MB/s 95 3	100 % 546 kB/s 4 0	100 % 4 MB/s 27 0	93 % 2 MB/s 14 1
dcache-se-desy.desy.de+	99 % 303 MB/s 2288 29	100 % 51 MB/s 384 0		99 % 79 MB/s 597 6	97 % 93 MB/s 694 19	99 % 81 MB/s 613 4
gridka-dcache.fzk.de+	91 % 113 MB/s 852 83	95 % 5 MB/s 36 2	92 % 64 MB/s 486 45		87 % 23 MB/s 173 25	93 % 21 MB/s 157 11
kek2-se01.cc.kek.jp+	100 % 368 MB/s 2732 2	100 % 61 MB/s 455 0	100 % 145 MB/s 1080 0	100 % 70 MB/s 520 2		100 % 91 MB/s 677 0
se.hep.pnnl.gov+	100 % 328 MB/s 2486 12	100 % 66 MB/s 499 0	99 % 87 MB/s 661 4	100 % 57 MB/s 433 1	99 % 118 MB/s 893 7	