Belle II distributed computing

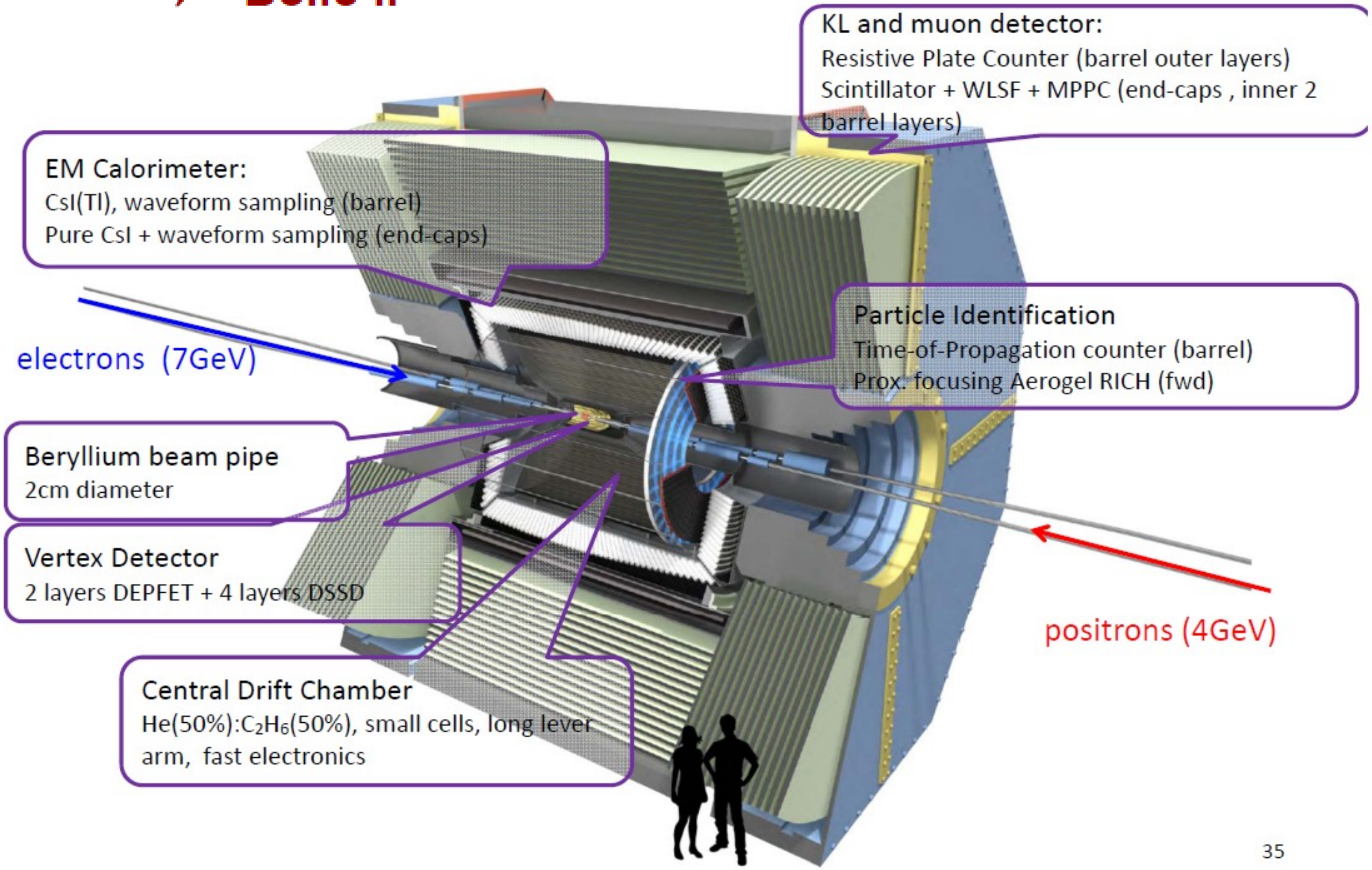
Pavel Krokovny
Novosibirsk State University and
Budker Institute of Nuclear Physics

Outline:
- Belle II experiment
- Computing model
- Resource estimation
- Production system
- Computing in BINP / NSU

ACAT 2014
– bridging disciplines
1-5 September 2014
Prague, Czech Republic
Europe/Prague timezone

The Super B factory project

	KEKB		SuperKEKB	
Luminosity:	2.1×10^{34}	→	8×10^{35}	(x 40)
Total Data:	1 ab^{-1}	→	50 ab^{-1}	(x 50)
Detector:	Belle	→	Belle II	



KL and muon detector:
Resistive Plate Counter (barrel outer layers)
Scintillator + WLSF + MPPC (end-caps , inner 2 barrel layers)

EM Calorimeter:
CsI(Tl), waveform sampling (barrel)
Pure CsI + waveform sampling (end-caps)

Particle Identification
Time-of-Propagation counter (barrel)
Prox. focusing Aerogel RICH (fwd)

Beryllium beam pipe
2cm diameter

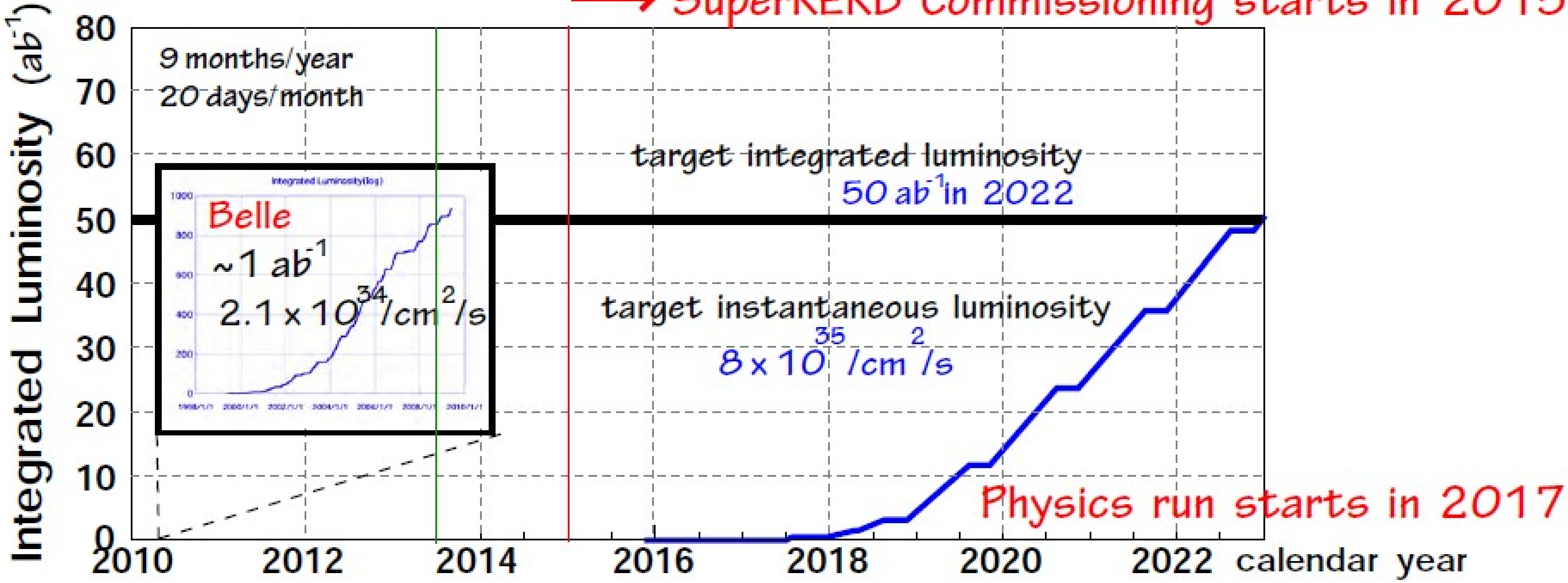
Vertex Detector
2 layers DEPFET + 4 layers DSSD

Central Drift Chamber
$He(50\%):C_2H_6(50\%)$, small cells, long lever arm, fast electronics

Belle II Collaboration:
- 26 countries,
- 97 institutes,
- ~600 scientists

Luminosity prospect

→ SuperKEKB Commissioning starts in 2015



Physics run starts in 2017

Experiment	Event size [kB]	Rate @ Storage [event/sec]	Rate @ Storage [MB/sec]
Belle II	300	6,000	1,800
ALICE (Pb-Pb)	50,000	100	4,000
ALICE (p-p)	2,000	100	200
ATLAS	1,500	600	700
CMS	1,500	150	225 (<~1000)
LHCb	55	4,500	250

(@ max. luminosity)

(LHC experiments: as seen in 2011/2012 runs)

Software system

- A “framework” system with dynamic module loading, parallel processing, Python steering, and ROOT I/O
- Full detector simulation with Geant4
- C++ 11 and gcc 4.7
- Supporting major Linux distributions: SL, Fedora, Ubuntu, etc

- Code management: Subversion
- Formatting tool: astyle
- Building: scons and buildbot system
- Documentation: Doxygen, Twiki
- Issue tracking: Redmine

Computing model

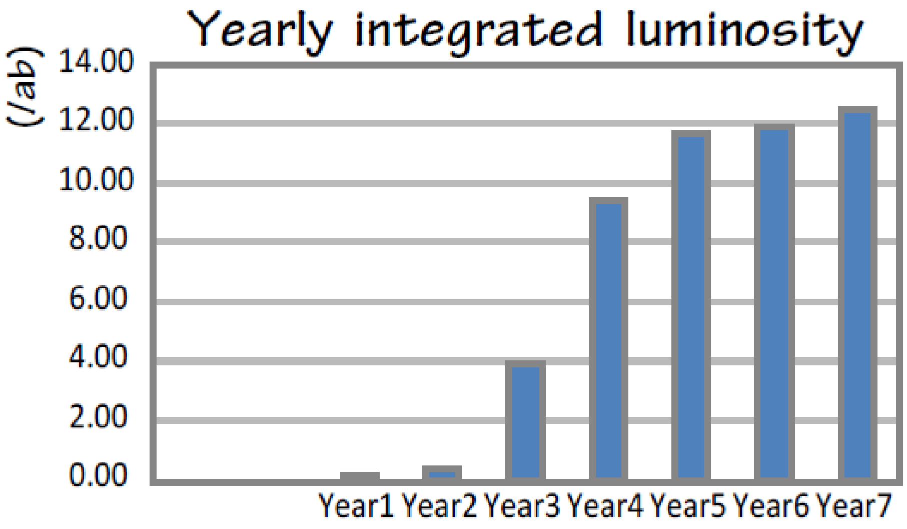
The BELLE II Computing model has to accomplish, in a distributed environment, the following main tasks:

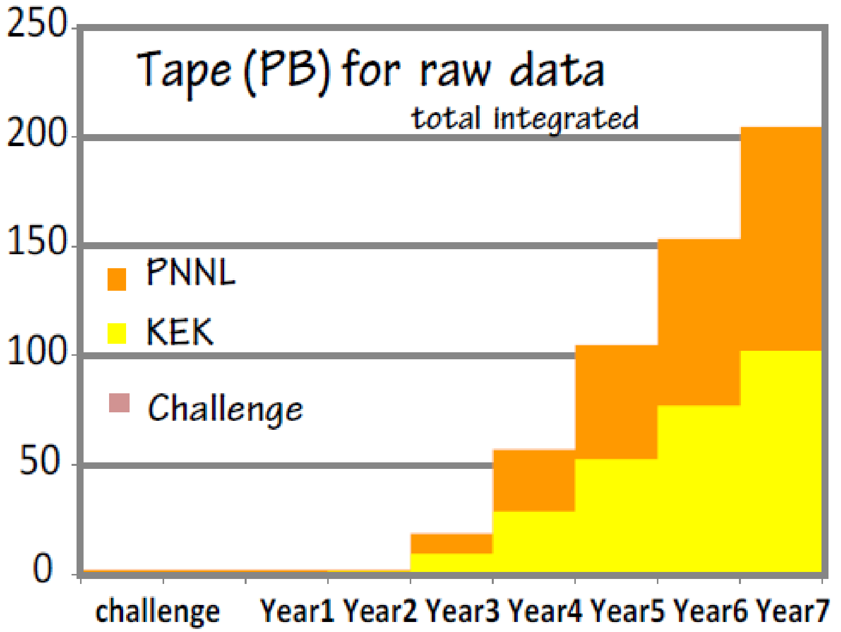
- RAW data processing
- Monte Carlo Production
- Physics analysis
- Data Storage and Data Archiving

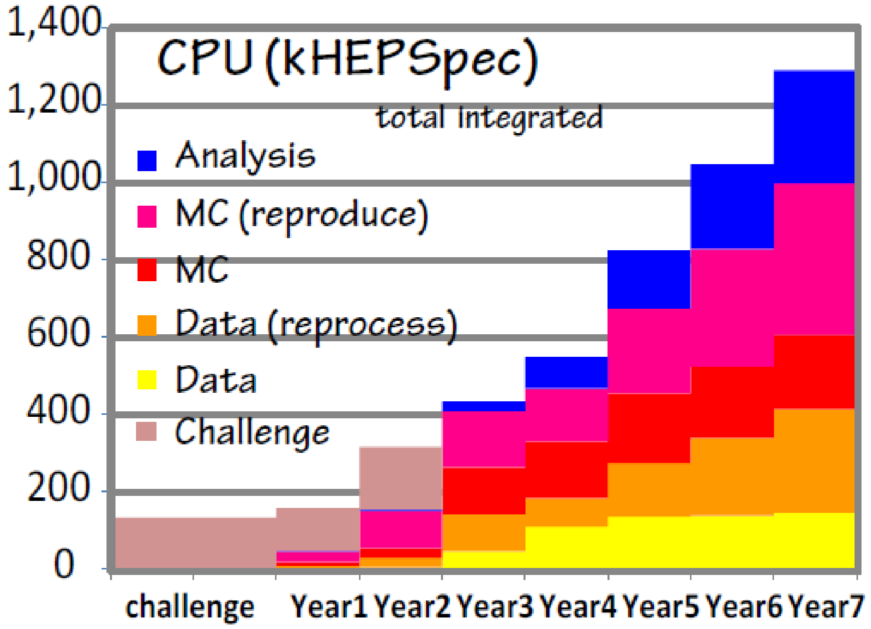
Resource Estimation:

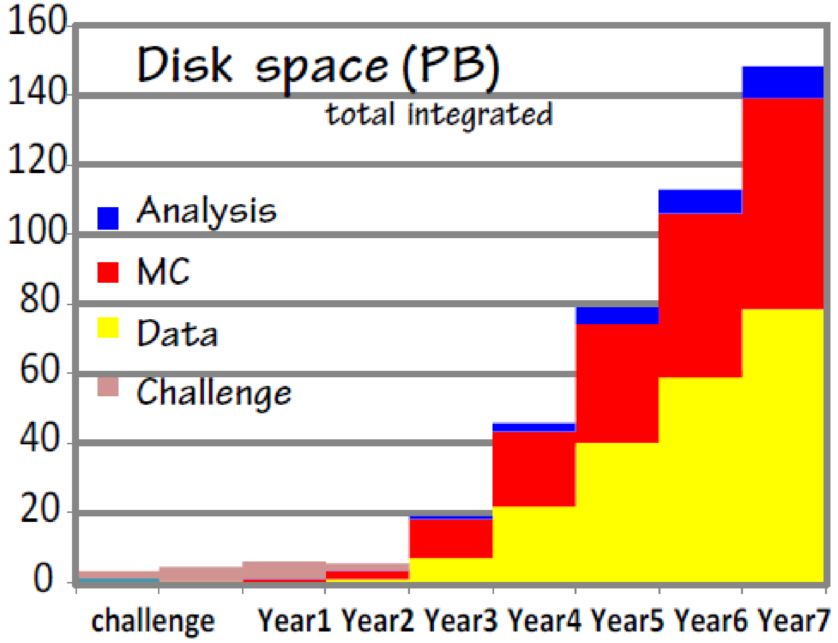
- Event size for RAW (mDST) data is 300 (40) Kb
- RAW data (MC) processing: 45 (90) HepSPEC*s / event

Required hardware resources

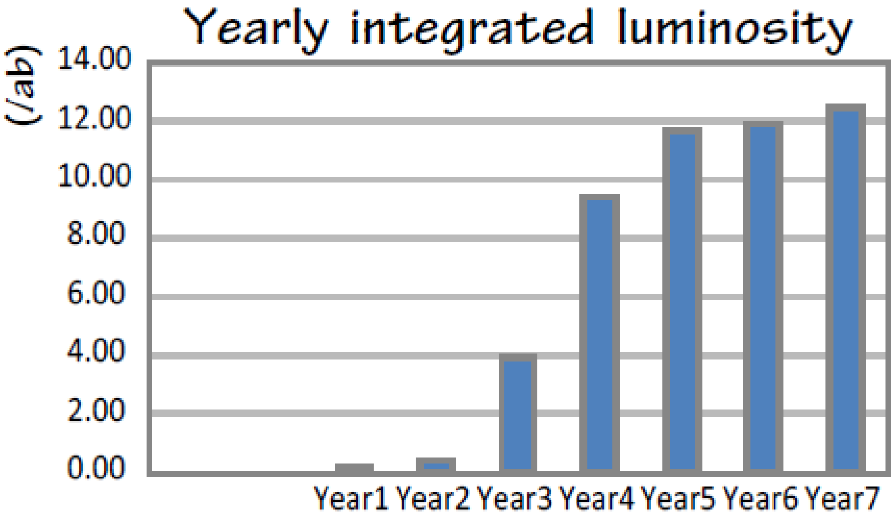


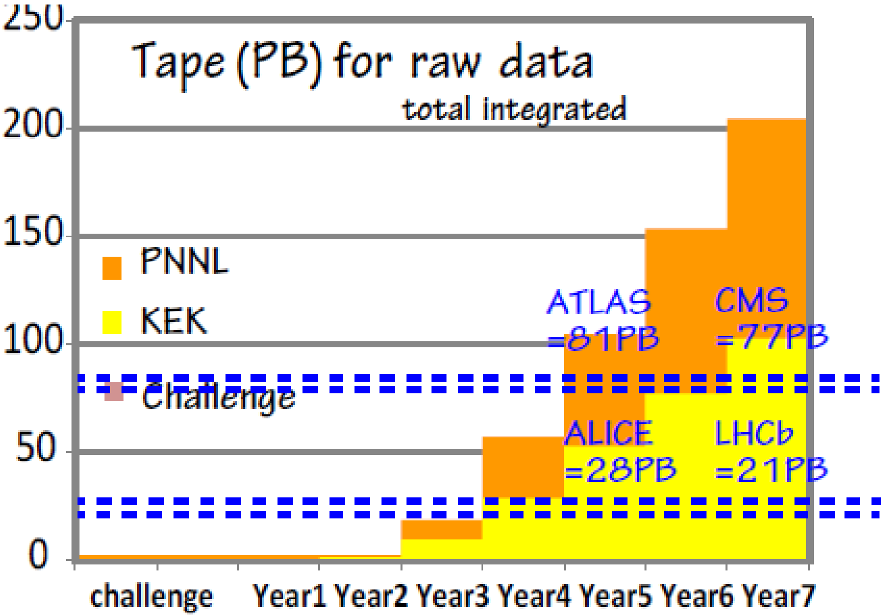


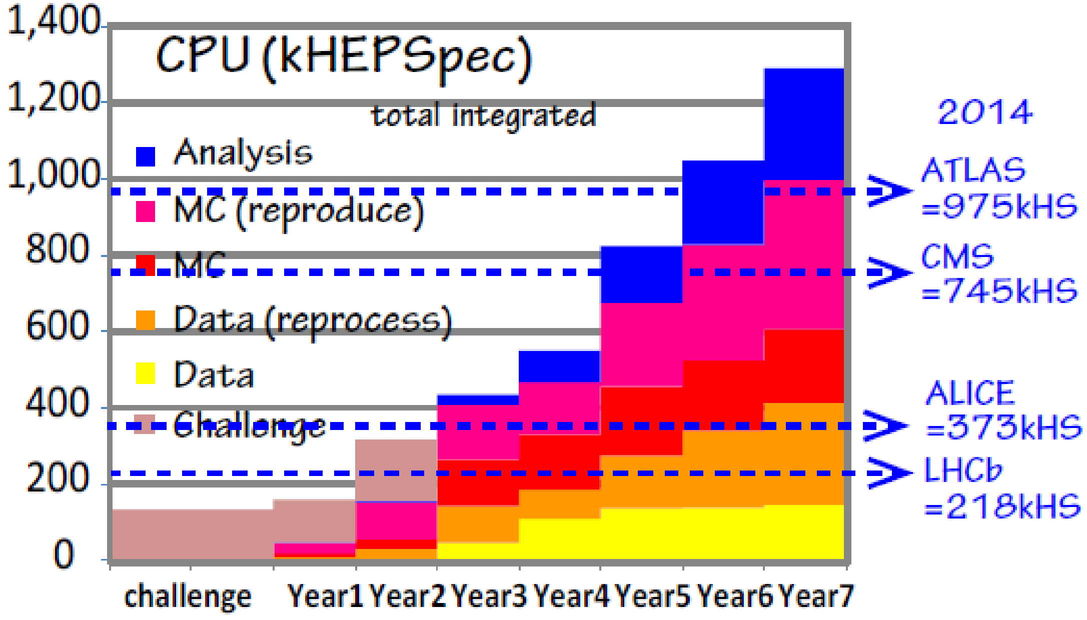


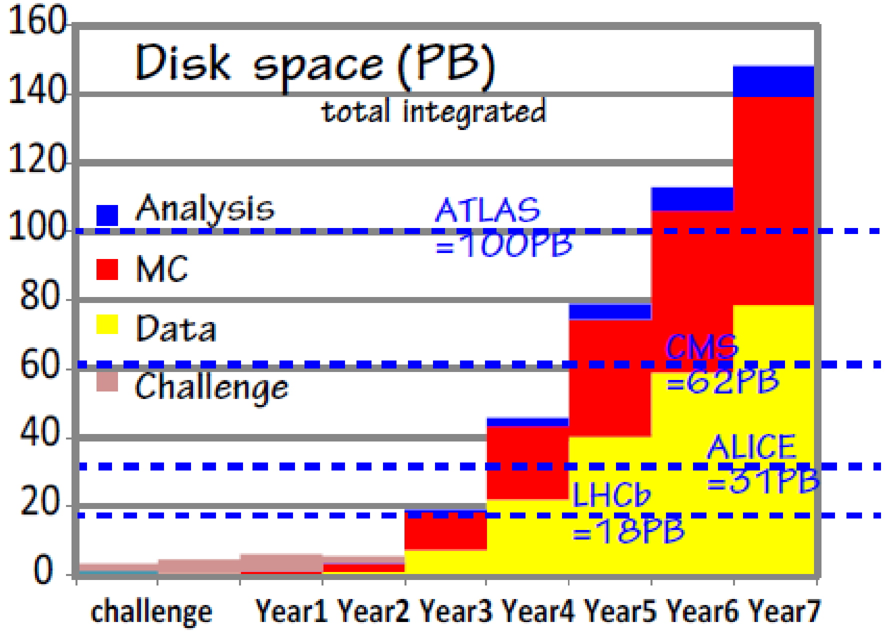


Comparison of hardware resources



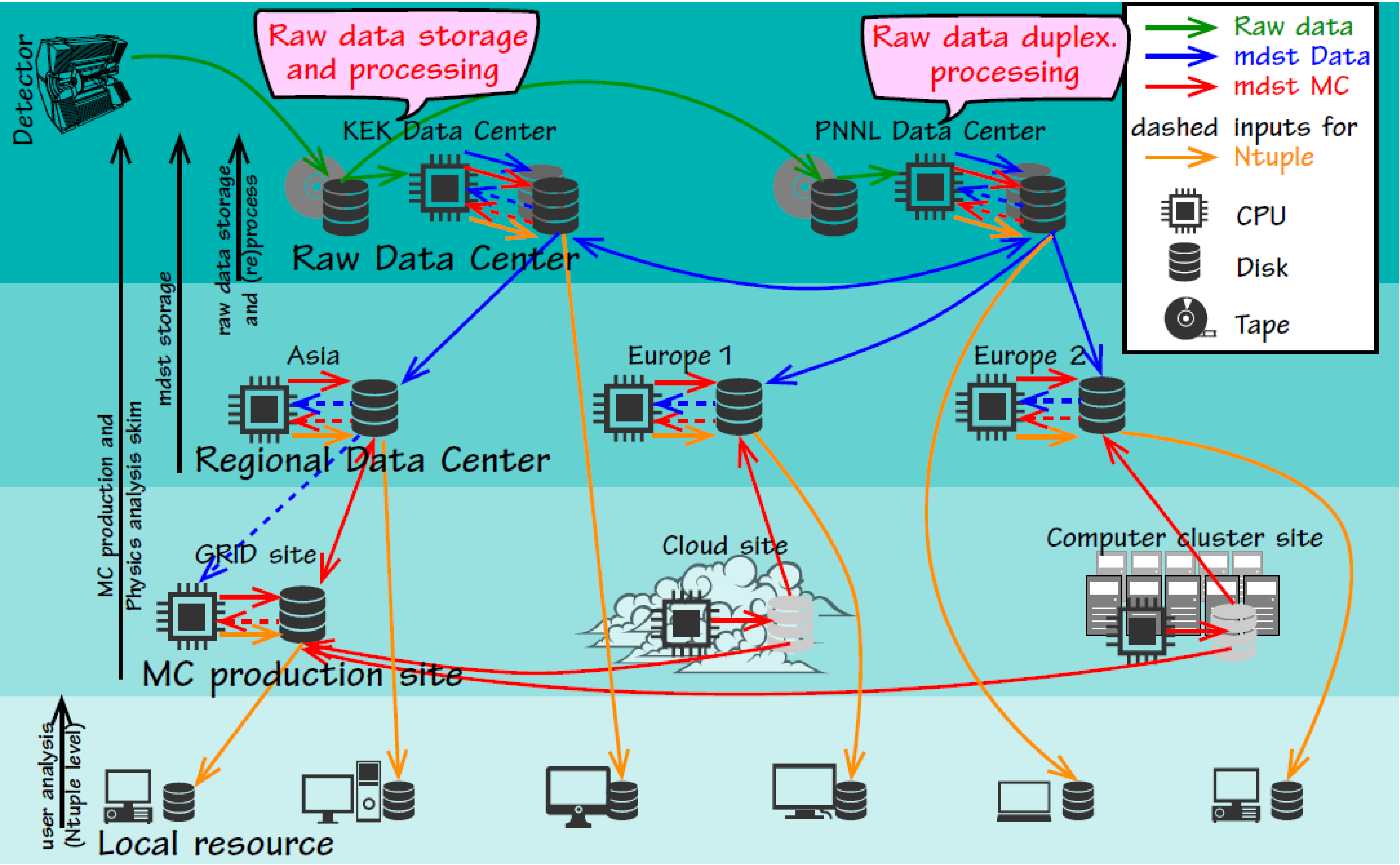




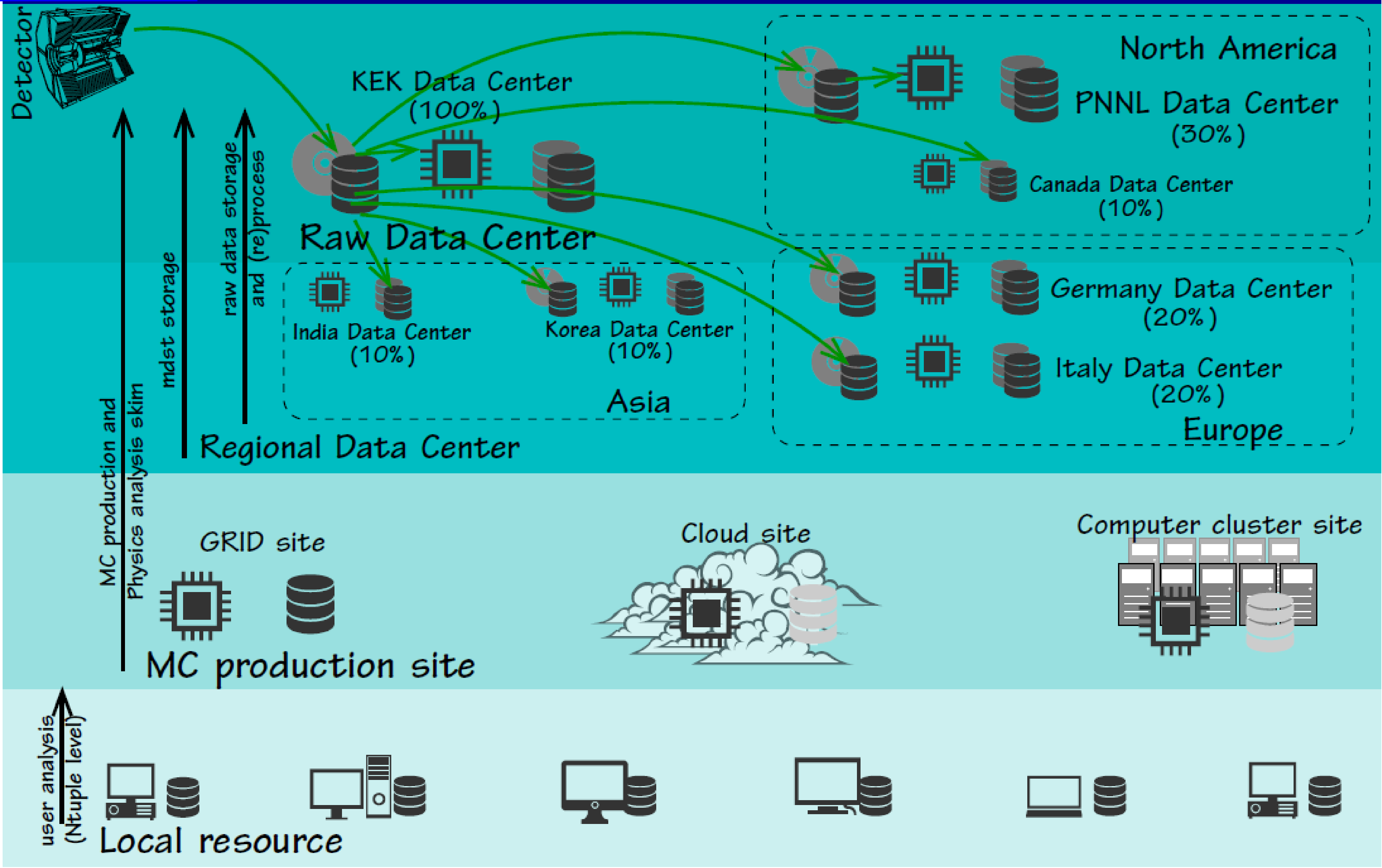


Pledge summary of LHC experiments : http://wlcg-rebus.cern.ch/apps/pledges/summary/

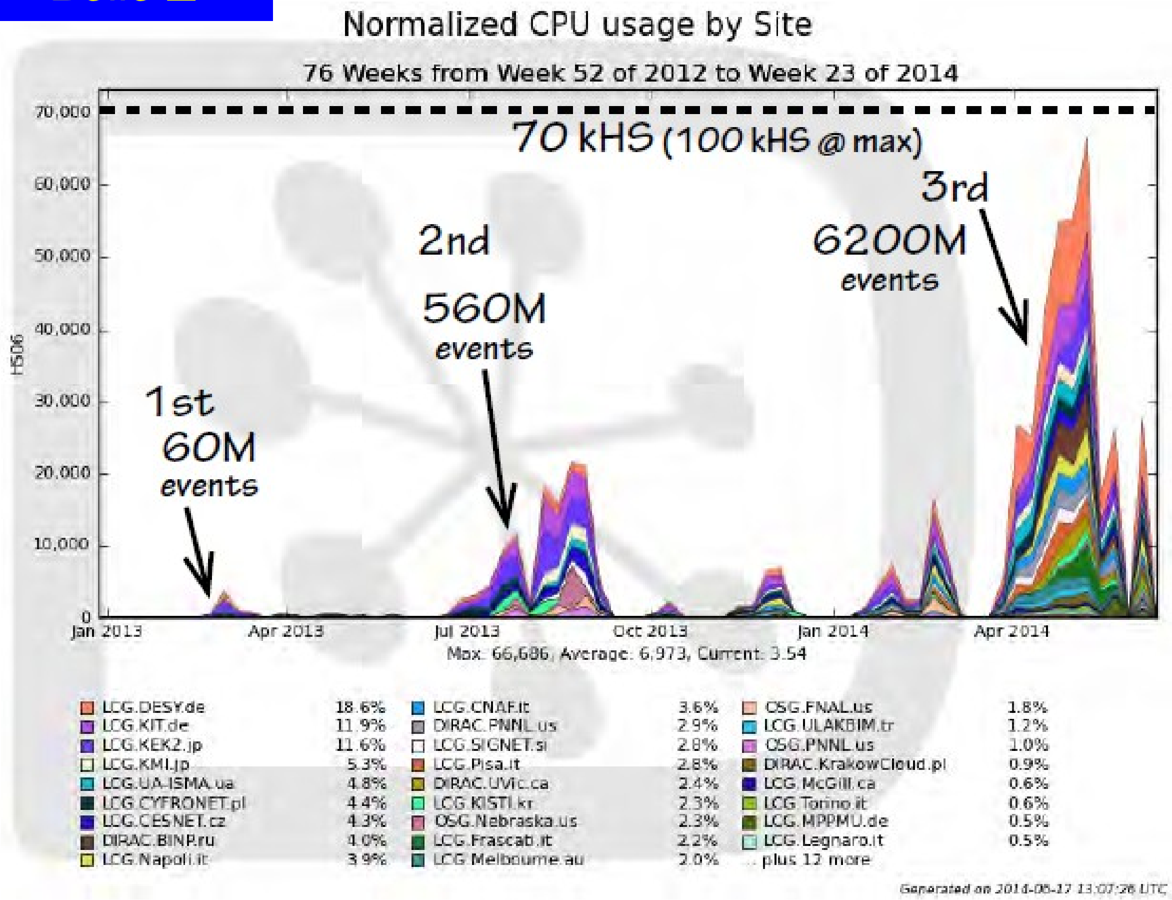
Initial computing model (first 3 years)



Raw data storage (2nd stage)



Current status



70 kHS (100 kHS @ max)

1st 60M events

2nd 560M events

3rd 6200M events

15 countries/regions

27 sites (+ 2 non-Belle II sites)

HEPHY (Vienna) and MPPMU (Munich) joined recently

GRID, Cloud, local cluster is available

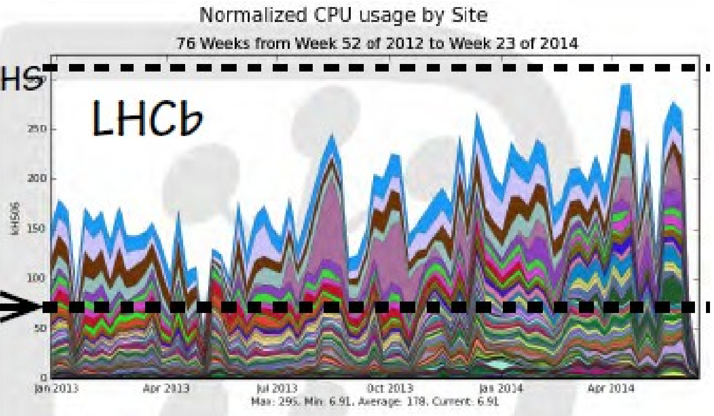
First official release of MC samples

BB generic decay/continuum

tau pair

(corresponding to 100fb^{-1} w/ and w/o BG)

Trans-pacific / trans-atlantic network data tranfer challenge



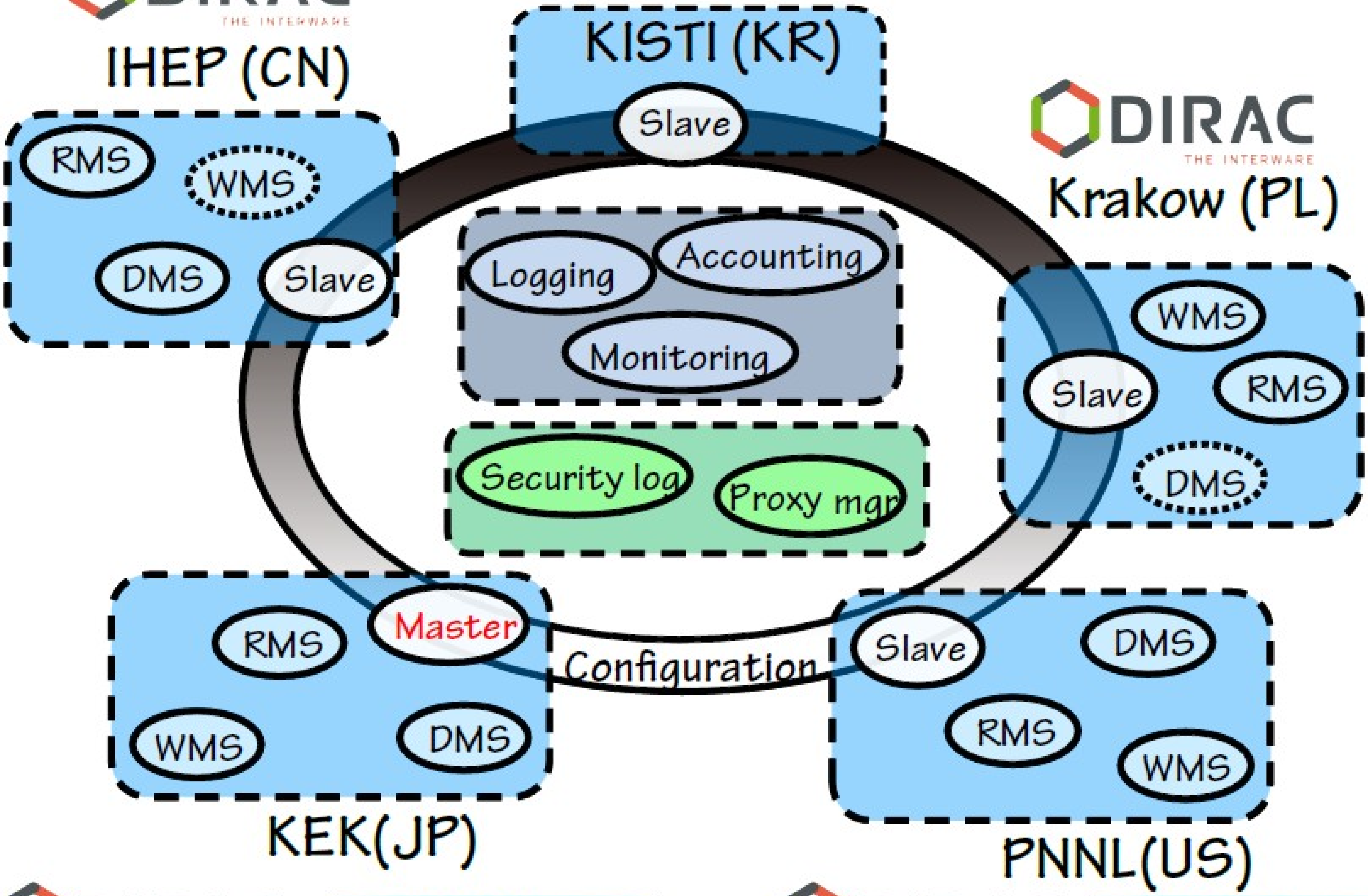
300 kHS

LHCb

120 sites

Belle II now
70 kHS

Production system (core services)



AMGA ARDA METADATA CATALOGUE PROJECT

KISTI (KR)
- Slave

DIRAC THE INTERWARE

IHEP (CN)
- RMS
- WMS
- DMS
- Slave

DIRAC THE INTERWARE

Krakow (PL)
- WMS
- Slave
- RMS
- DMS

- Logging
- Accounting
- Monitoring

- Security log
- Proxy mgr

KEK(JP)
- RMS
- Master
- WMS
- DMS

Configuration

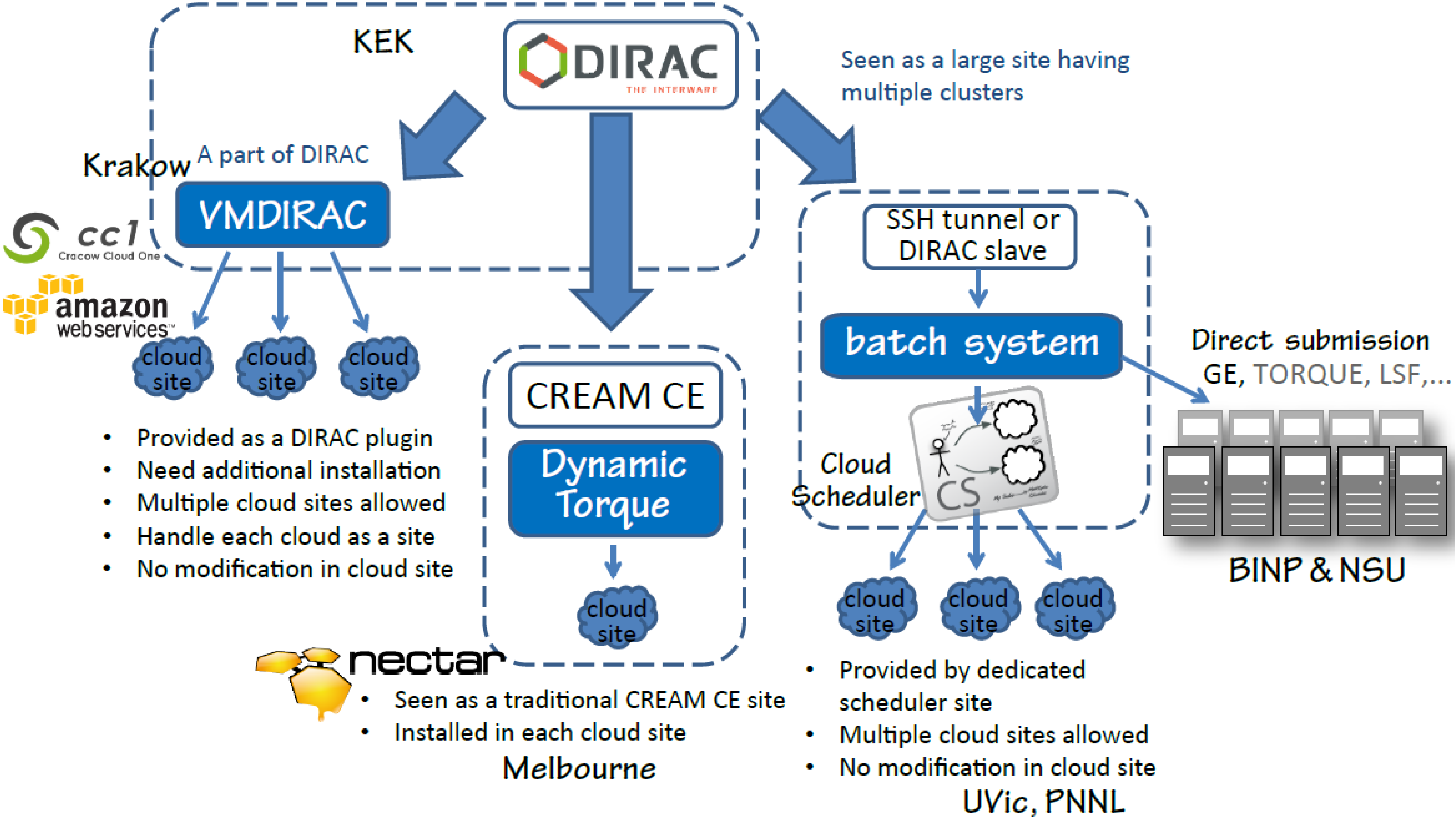
PNNL(US)
- Slave
- DMS
- RMS
- WMS

DIRAC THE INTERWARE AMGA ARDA METADATA CATALOGUE PROJECT

DIRAC THE INTERWARE AMGA ARDA METADATA CATALOGUE PROJECT

CernVM File system cvmfs is used for software installation

Production system (cloud usage)



- Provided as a DIRAC plugin
- Need additional installation
- Multiple cloud sites allowed
- Handle each cloud as a site
- No modification in cloud site

- Seen as a traditional CREAM CE site
- Installed in each cloud site

Melbourne

- Provided by dedicated scheduler site
- Multiple cloud sites allowed
- No modification in cloud site

UVic, PNNL

Computing in BINP & NSU

Experiments
- BINP: CMD-3, SND, KEDR
- International collaborations: Belle II, ATLAS, LHCb

Computing resources
- BINP: 512 CPU cores, 6 Tflops
- Novosibirsk State University: 2432 cores, 29 Tflops
- Supercomputer Center: 30 + 85 Tflops (CPU + GPU)

We want:
- keep the specific computing environment and user’s experience
- be like a normal SC user

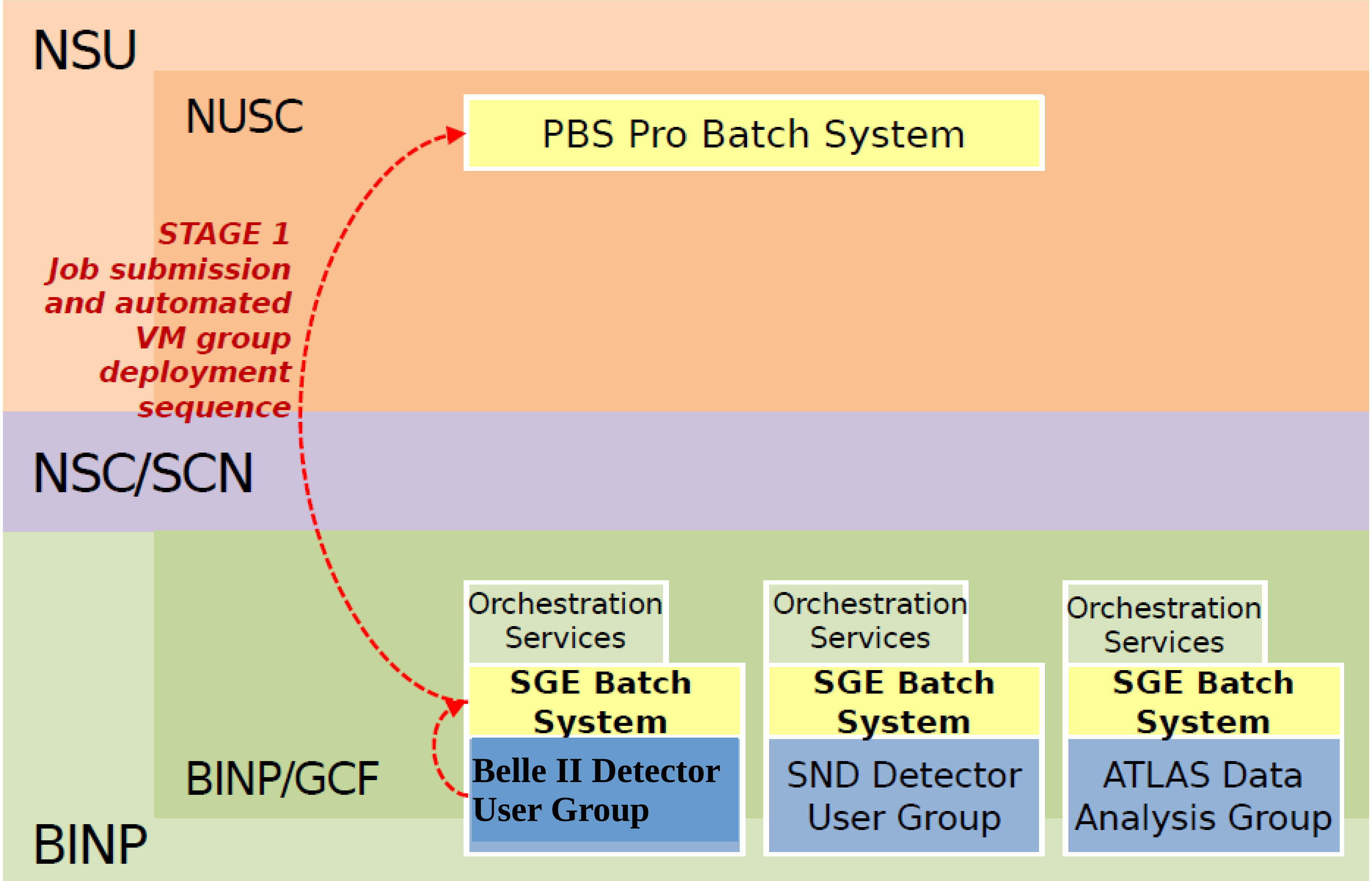
The solution is:
- run HEP tasks inside virtual machines,
- run VMs inside supercomputer’s batch system jobs.

Key features of the virtualized infrastructure

- Virtualization by KVM
  - included in modern Linux distributions,
  - quite stable,
  - does not require modified Linux kernel.

- VM disk images are located on SC’s file system and accessible via InfiniBand.
- Local snapshots of VM images are used on physical nodes thus leaving master images unchanged.
- Input/output data are located at BINP and accessed by VMs via NFS.
- VMs are just regular batch tasks at a supercomputer.
- VMs are started automatically on user’s demands.

Batch System Integration Mechanisms



NSU

NUSC

PBS Pro Batch System

STAGE 1
Job submission
and automated
VM group
deployment
sequence

NSC/SCN

BINP/GCF

Orchestration Services

SGE Batch System

Belle II Detector User Group

Orchestration Services

SGE Batch System

SND Detector User Group

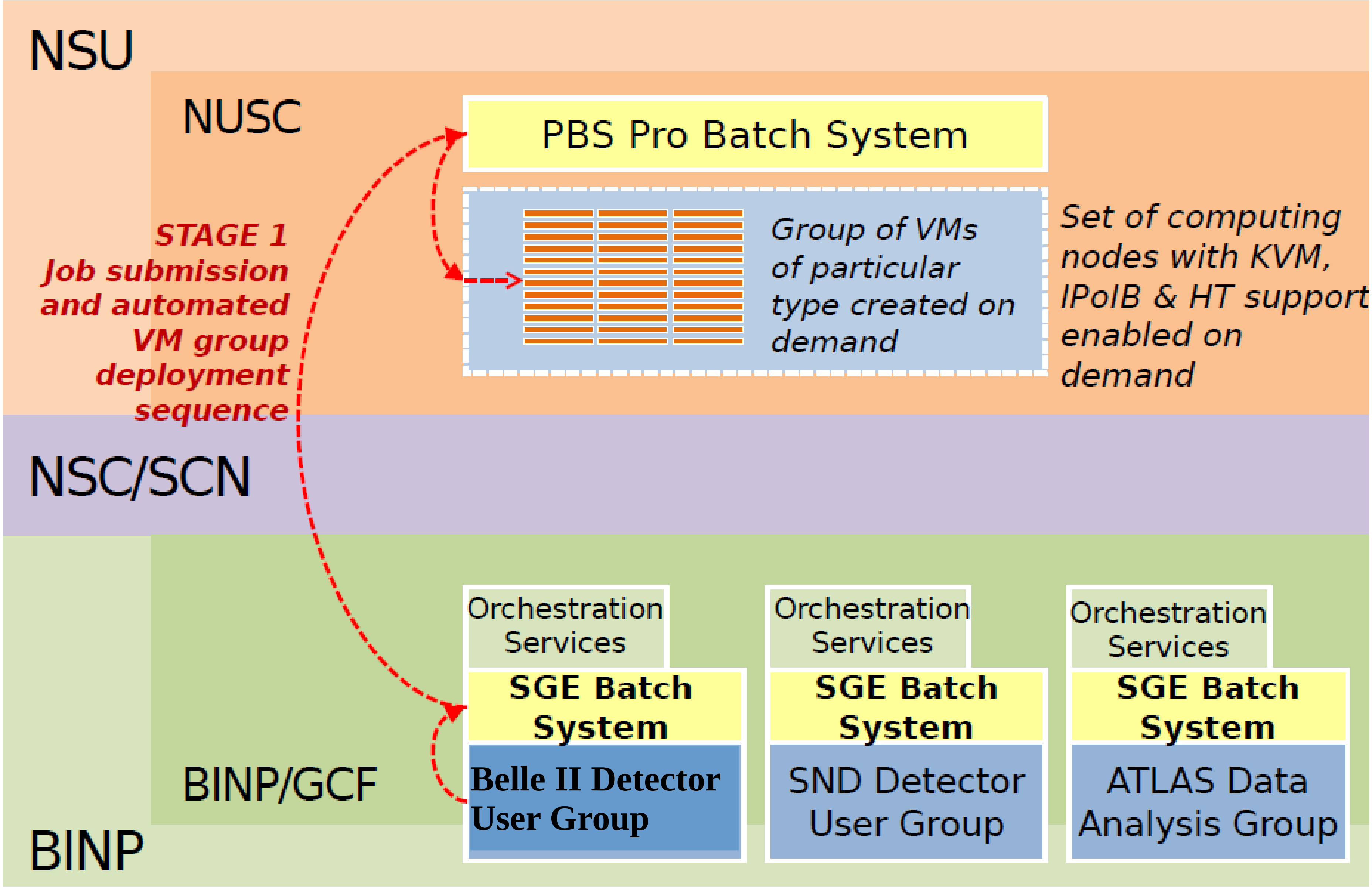
Orchestration Services

SGE Batch System

ATLAS Data Analysis Group

BINP

Batch System Integration Mechanisms



NSU

NUSC

PBS Pro Batch System

Group of VMs of particular type created on demand

Set of computing nodes with KVM, IPoIB & HT support enabled on demand

STAGE 1

Job submission and automated VM group deployment sequence

NSC/SCN

BINP

BINP/GCF

Orchestration Services

SGE Batch System

Belle II Detector User Group

Orchestration Services

SGE Batch System

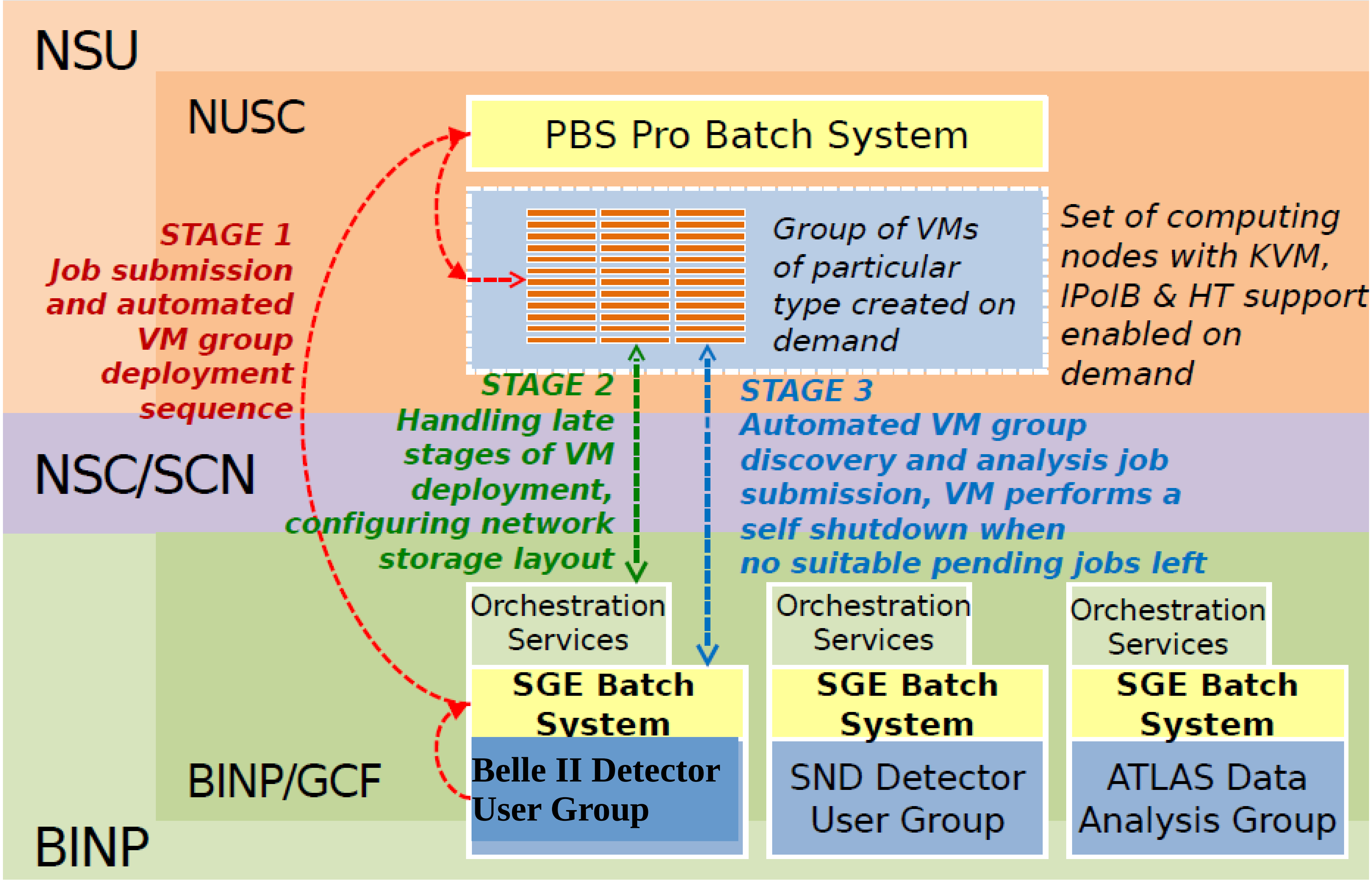
SND Detector User Group

Orchestration Services

SGE Batch System

ATLAS Data Analysis Group

Batch System Integration Mechanisms



NSU

NUSC

PBS Pro Batch System

STAGE 1
Job submission and automated VM group deployment sequence

Group of VMs of particular type created on demand

Set of computing nodes with KVM, IPoIB & HT support enabled on demand

NSC/SCN

STAGE 2
Handling late stages of VM deployment, configuring network storage layout

STAGE 3
Automated VM group discovery and analysis job submission, VM performs a self shutdown when no suitable pending jobs left

Orchestration Services

SGE Batch System

Belle II Detector User Group

Orchestration Services

SGE Batch System

SND Detector User Group

Orchestration Services

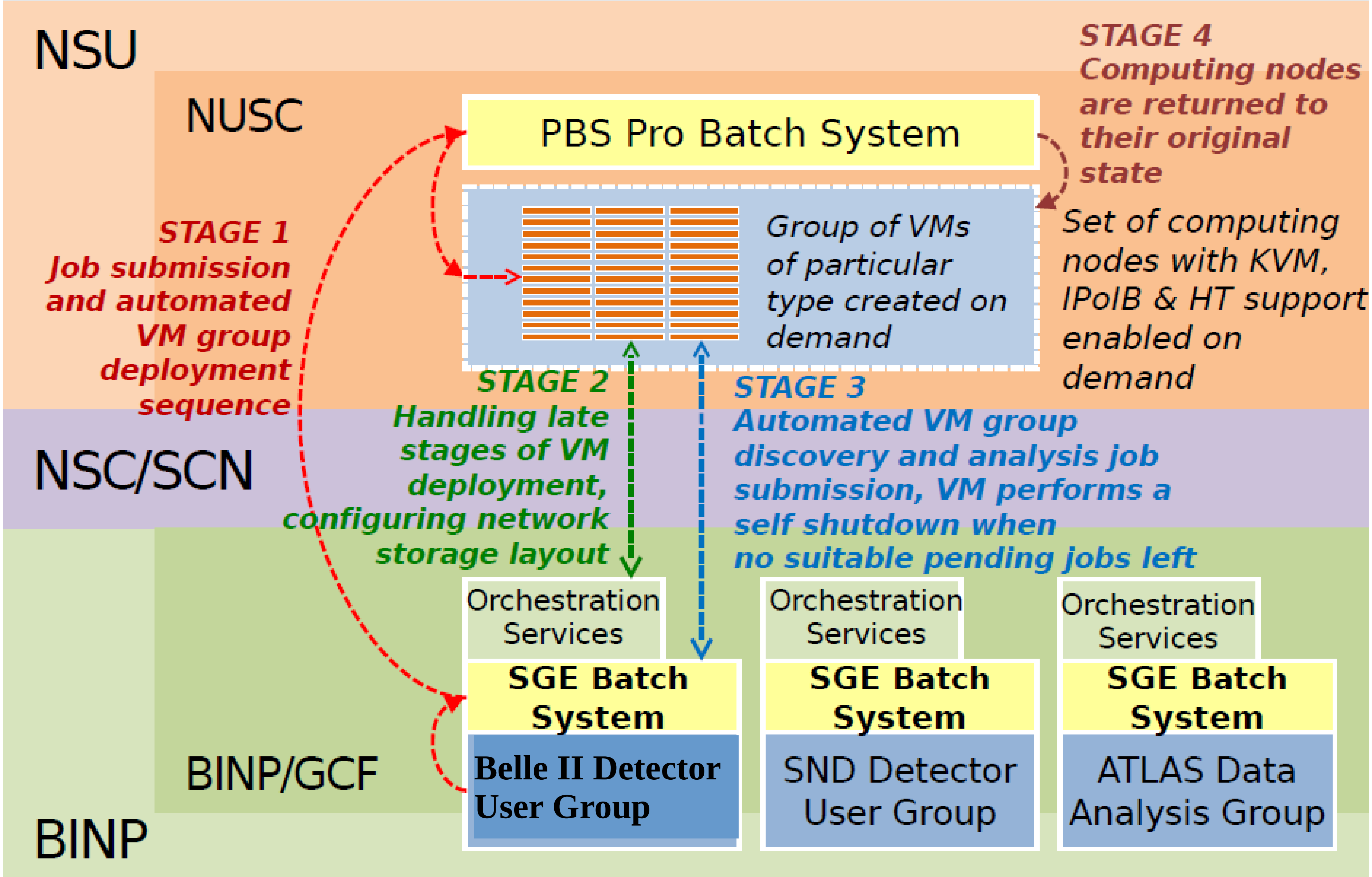
SGE Batch System

ATLAS Data Analysis Group

BINP/GCF

BINP

Batch System Integration Mechanisms



NSU

NUSC

PBS Pro Batch System

STAGE 4
Computing nodes are returned to their original state

Set of computing nodes with KVM, IPoIB & HT support enabled on demand

STAGE 1
Job submission and automated VM group deployment sequence

Group of VMs of particular type created on demand

STAGE 2
Handling late stages of VM deployment, configuring network storage layout

STAGE 3
Automated VM group discovery and analysis job submission, VM performs a self shutdown when no suitable pending jobs left

NSC/SCN

Orchestration Services

SGE Batch System

Belle II Detector User Group

Orchestration Services

SGE Batch System

SND Detector User Group

Orchestration Services

SGE Batch System

ATLAS Data Analysis Group

BINP/GCF

BINP

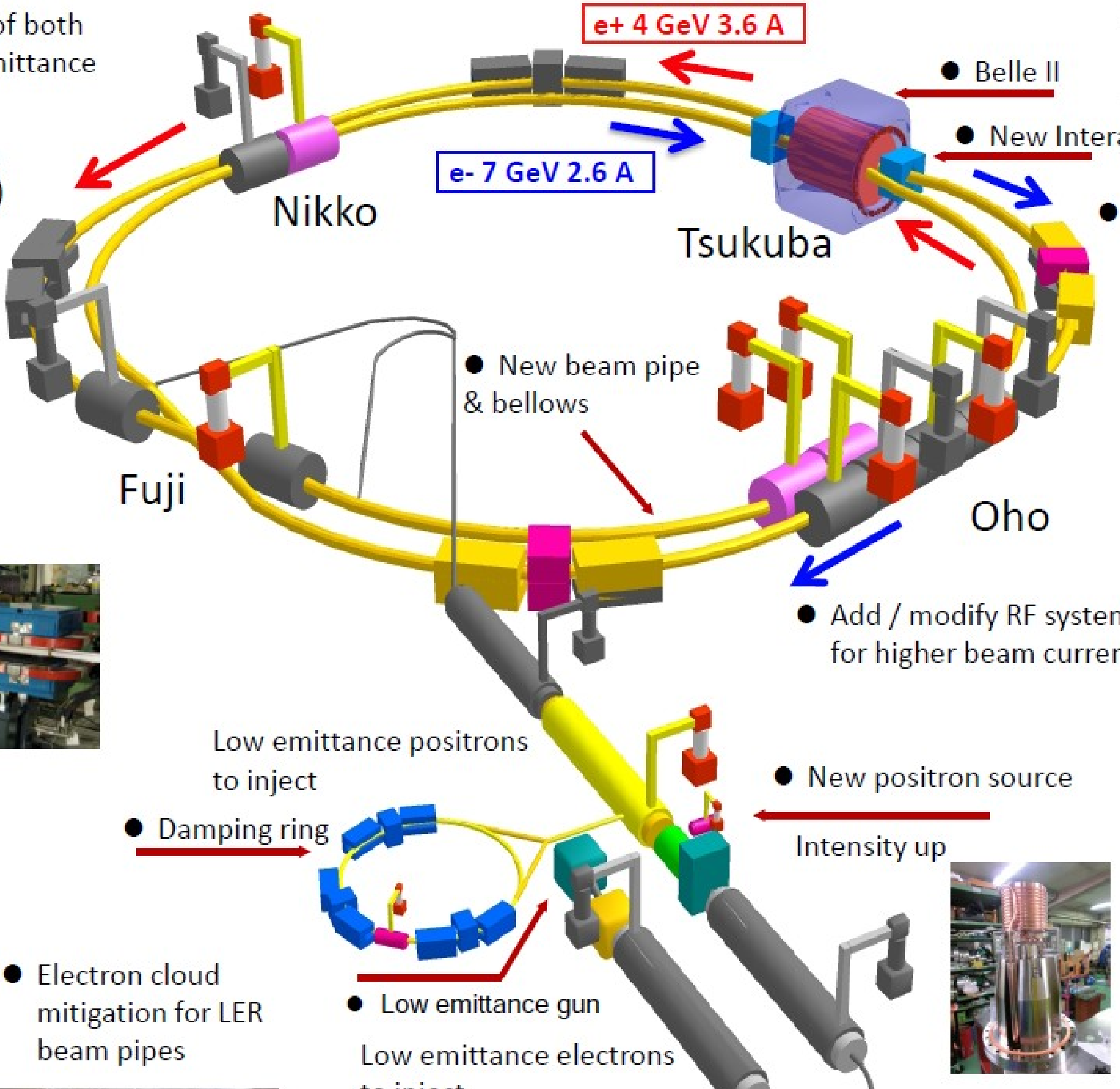
Summary

- Belle II experiment has successful experience in distributed computing using GRID, Cloud and local clusters.
- Current CPU power is about 70 kHS (25% of LHCb) and used for MC production.
- BINP successfully uses Virtual Machines technique to run up to 1K jobs simultaneously for 6 experiments on 3 supercomputer sites.

Backup

SuperKEKB accelerator

- Redesign the lattices of both rings to reduce the emittance
- Replace short dipoles with longer ones (LER)



- Belle II
- New Interaction region
- New beam pipe & bellows
- Add / modify RF systems for higher beam current

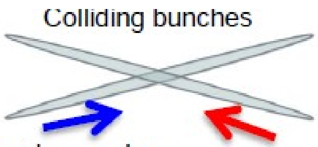
Low emittance positrons to inject

- Damping ring
- New positron source

Intensity up

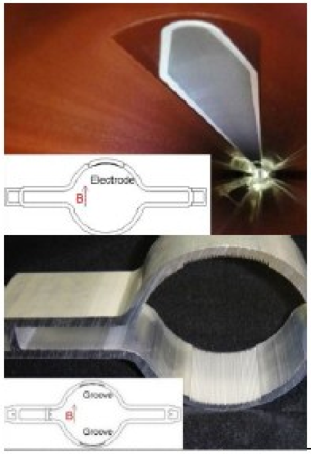
- Low emittance gun

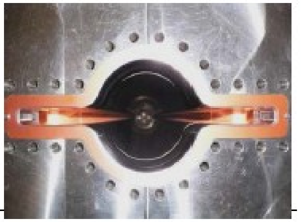
Low emittance electrons to inject



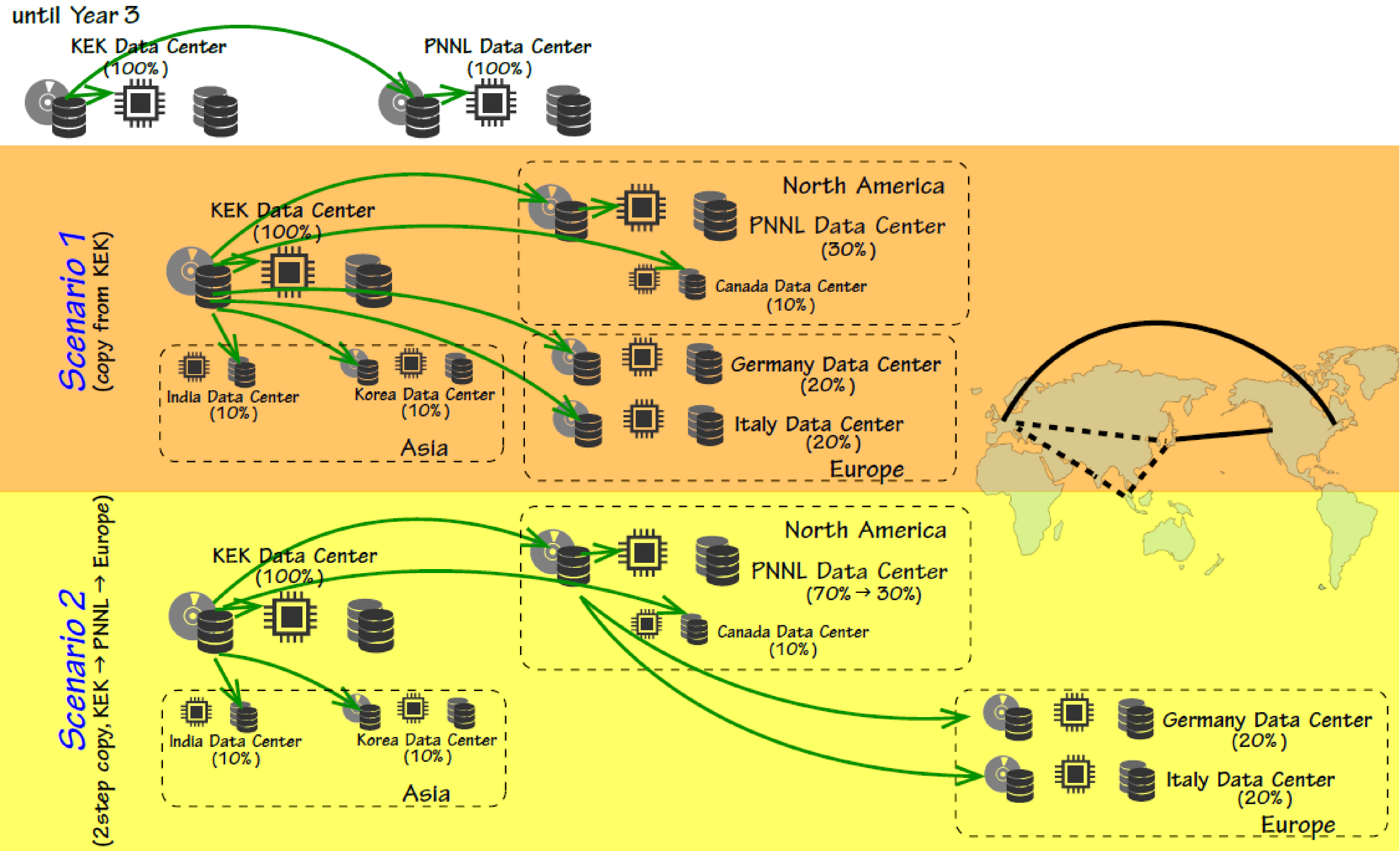
- New superconducting final focusing quads near the IP

- Electron cloud mitigation for LER beam pipes





Raw data distribution



until Year 3

KEK Data Center
(100%)

PNNL Data Center
(100%)

Scenario 1
(copy from KEK)

KEK Data Center
(100%)

North America

PNNL Data Center
(30%)

Canada Data Center
(10%)

India Data Center
(10%)

Korea Data Center
(10%)

Asia

Germany Data Center
(20%)

Italy Data Center
(20%)

Europe

Scenario 2
(2step copy, KEK → PNNL → Europe)

KEK Data Center
(100%)

North America

PNNL Data Center
(70% → 30%)

Canada Data Center
(10%)

India Data Center
(10%)

Korea Data Center
(10%)

Asia

Germany Data Center
(20%)

Italy Data Center
(20%)

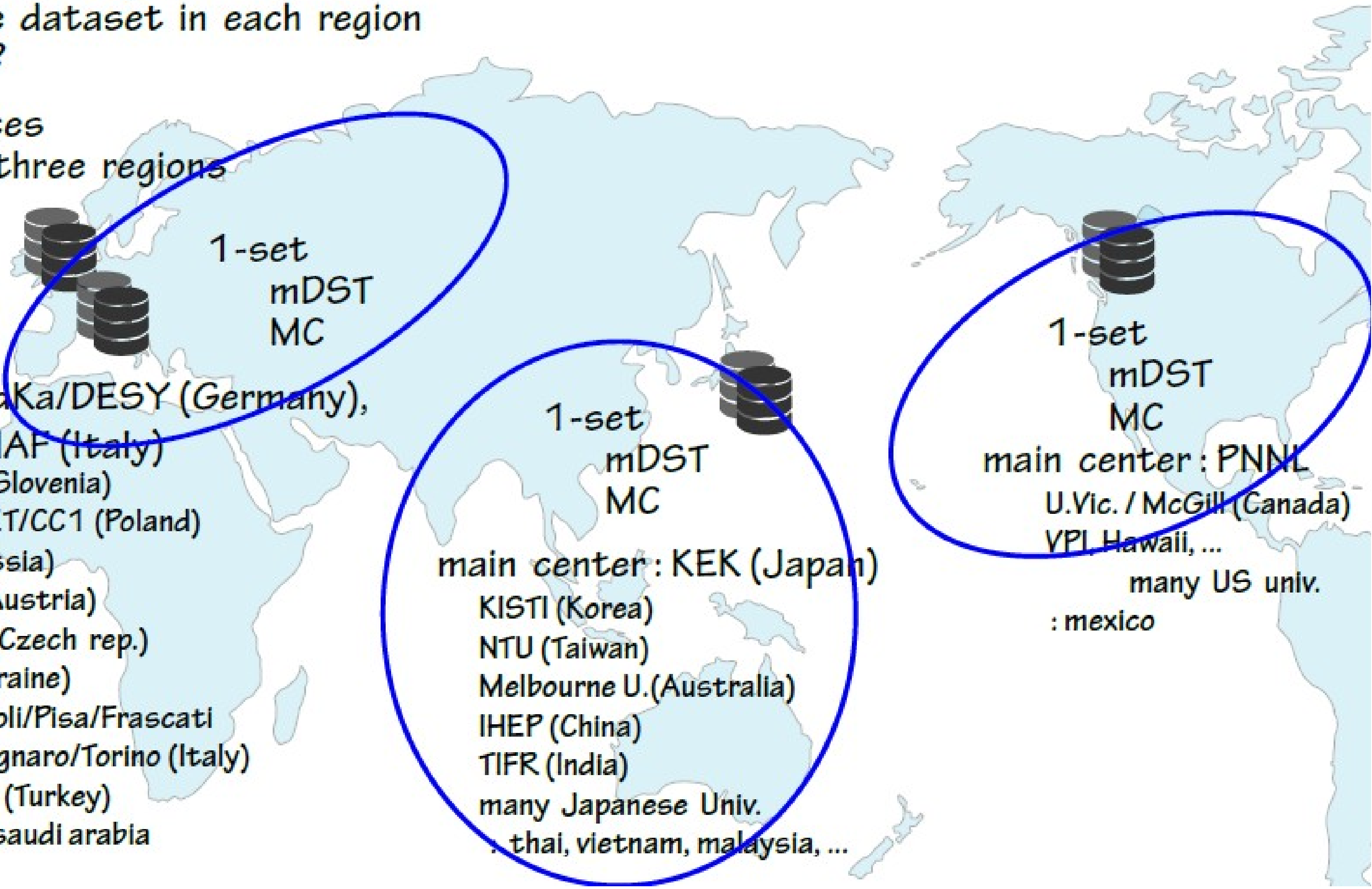
Europe

mDST & MC data distribution

mDST (data) is copied in Asia, Europe, and USA

For the MC data seems to be natural to be
    the similar structure
        better network ? in each region
        completeness of the dataset in each region
        easier maintenance ?

        unbalance of resources
        data copy between three regions



1-set
mDST
MC

main center : GridKa/DESY (Germany),
CNAF (Italy)
- SiGNET (Slovenia)
- CYFRONET/CC1 (Poland)
- BINP (Russia)
- HEPHY (Austria)
- CESNET (Czech rep.)
- ISMA (Ukraine)
- INFN Napoli/Pisa/Frascati /Legnaro/Torino (Italy)
- ULAKBIM (Turkey)
- : spain, saudi arabia

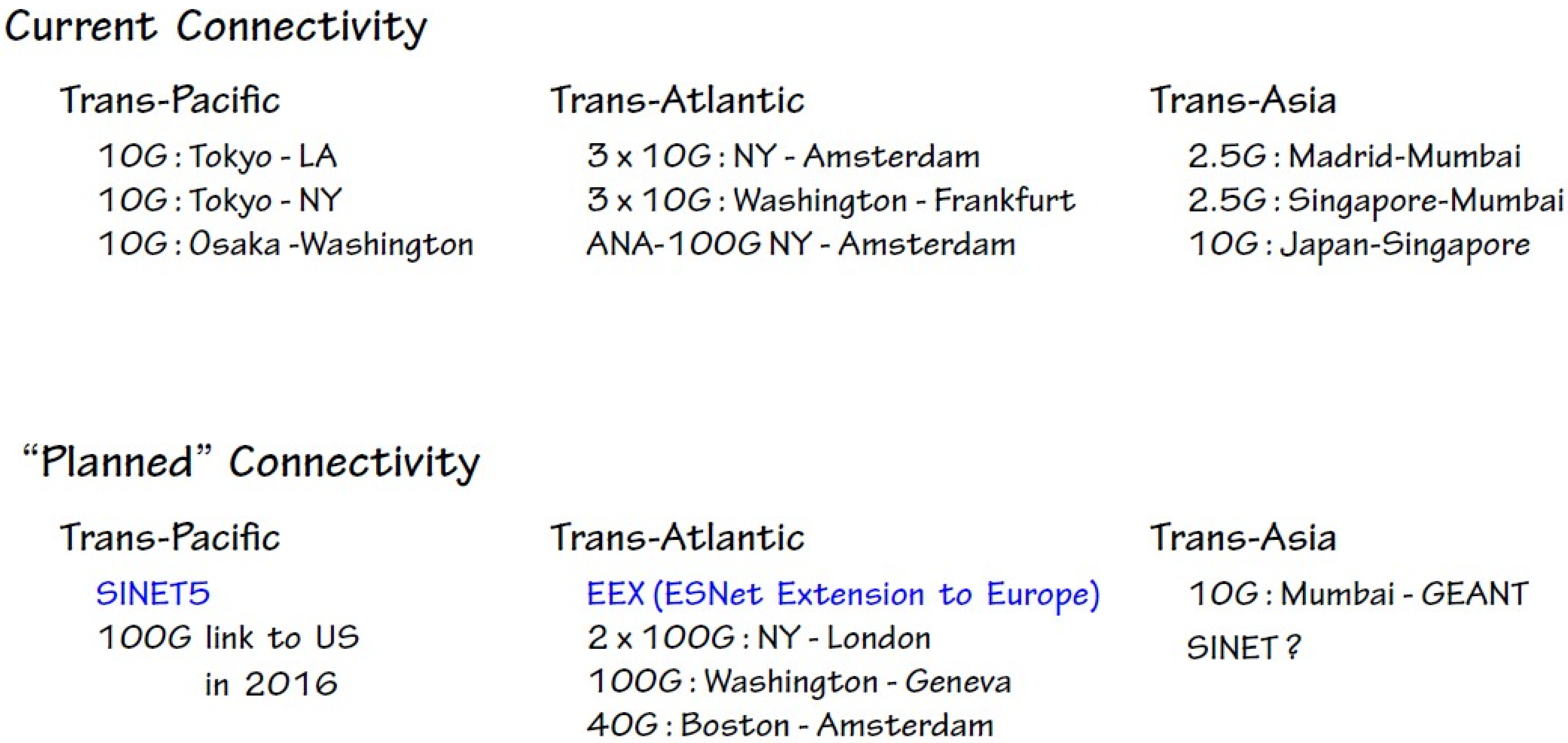
1-set
mDST
MC

main center : KEK (Japan)
- KISTI (Korea)
- NTU (Taiwan)
- Melbourne U.(Australia)
- IHEP (China)
- TIFR (India)
- many Japanese Univ.
- : thai, vietnam, malaysia, ...

1-set
mDST
MC

main center : PNNL
- U.Vic. / McGill (Canada)
- VPI, Hawaii, ...
- many US univ.
- : mexico

Network



Current Connectivity

Trans-Pacific

10G : Tokyo - LA
10G : Tokyo - NY
10G : Osaka -Washington

Trans-Atlantic

3 x 10G : NY - Amsterdam
3 x 10G : Washington - Frankfurt
ANA-100G NY - Amsterdam

Trans-Asia

2.5G : Madrid-Mumbai
2.5G : Singapore-Mumbai
10G : Japan-Singapore

“Planned” Connectivity

Trans-Pacific

SINET5
100G link to US
in 2016

Trans-Atlantic

EEX (ESNet Extension to Europe)
2 x 100G : NY - London
100G : Washington - Geneva
40G : Boston - Amsterdam

Trans-Asia

10G : Mumbai - GEANT

SINET ?

Deployment of the virtualized computing infrastructure

Deployment stages:

- Virtualizing of an experimental group's computing environment.
- Tests of virtual machines locally on BINP resources.
- Transferring the VMs to a supercomputer and running them under remote batch system's control.
- Integration of local and remote batch systems.

Finally we have dynamical virtualized computing cluster. Physicists use computing resources in a conventional way.