HEP Networking and Grids in Japan

on 9-11 October 2006 at Krakow ICFA-WS

Yukio.Karita@KEK.jp

Outline

- National R&E Networks in Japan
  - Domestic
  - International
- HEP Networking in Japan
- HEP Computing in Japan
  - Belle Data Analysis, Lattice-QCD, …
- Grids in Japan
- Grids at KEK

R&E Networks in Japan

- Production Networks
  - SuperSINET/SINET operated by NII, Ministry of Education and Science
    - Domestic/International
    - 10Gbps to MANLAN, 2.5Gbps to LA
  - APAN operated by APAN-Consortium
    - International
    - 10Gbps to LA, 2.5Gbps to HK, 2.5Gbps to KR, …
- Testbeds
  - JGN2 operated by NICT, Ministry of Communication
    - Domestic/International
    - 10Gbps to Starlight
  - IEEAF
    - International
    - 10Gbps to PacificWave

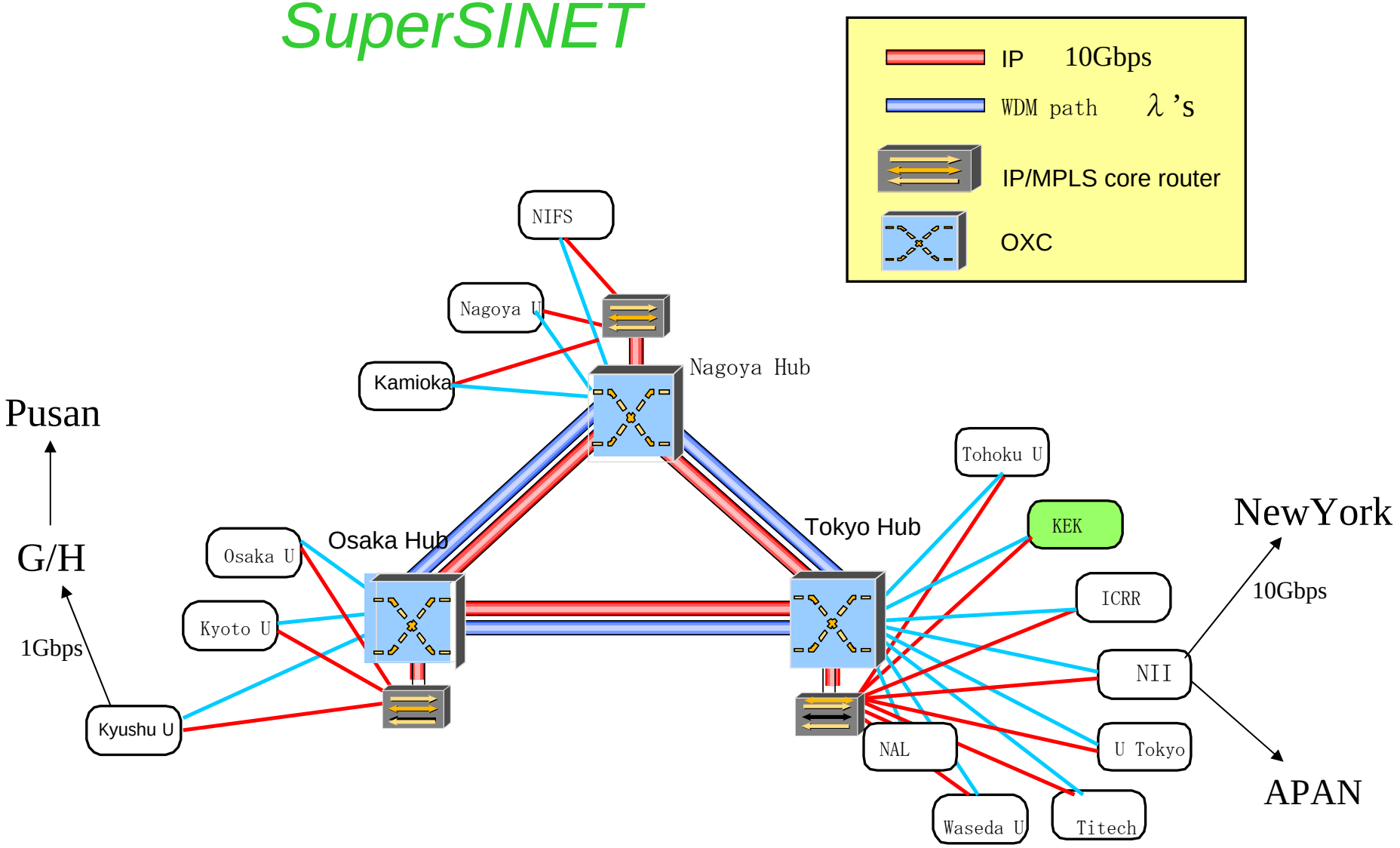
SuperSINET and JGN2

- National R&E Networks
- SuperSINET: Production network for sciences
  - Since Jan 2002
  - 10Gbps IP connection and p2p GigE's in Japan
  - 10Gbps IP to NY since Dec 2002
  - 2.5Gbps IP to LA since Apr 2005
- JGN2: Testbed for the research of the network
  - Since Apr 2004
  - 10Gbps L2 between Tsukuba and Tokyo (incl. T-LEX)
  - 10Gbps L2 to Starlight in Sep 2005

SuperSINET

- 10Gbps IP/MPLS Backbone
  - Star-topology OC192 connection from Hub
  - Non-shared 10Gbps
  - MPLS-VPN’s are configured on request
- GigE / 10GE Bridges for peer-connection
  - lambdas separate from the 10G IP/MPLS connection
  - Lightwave permanent path
  - L1 p2p service
  - Tagged-VLAN’s can run on a path
- Operation of Optical Cross Connect (OXC) for fiber / wavelength switching
- Operational from 4th January, 2002

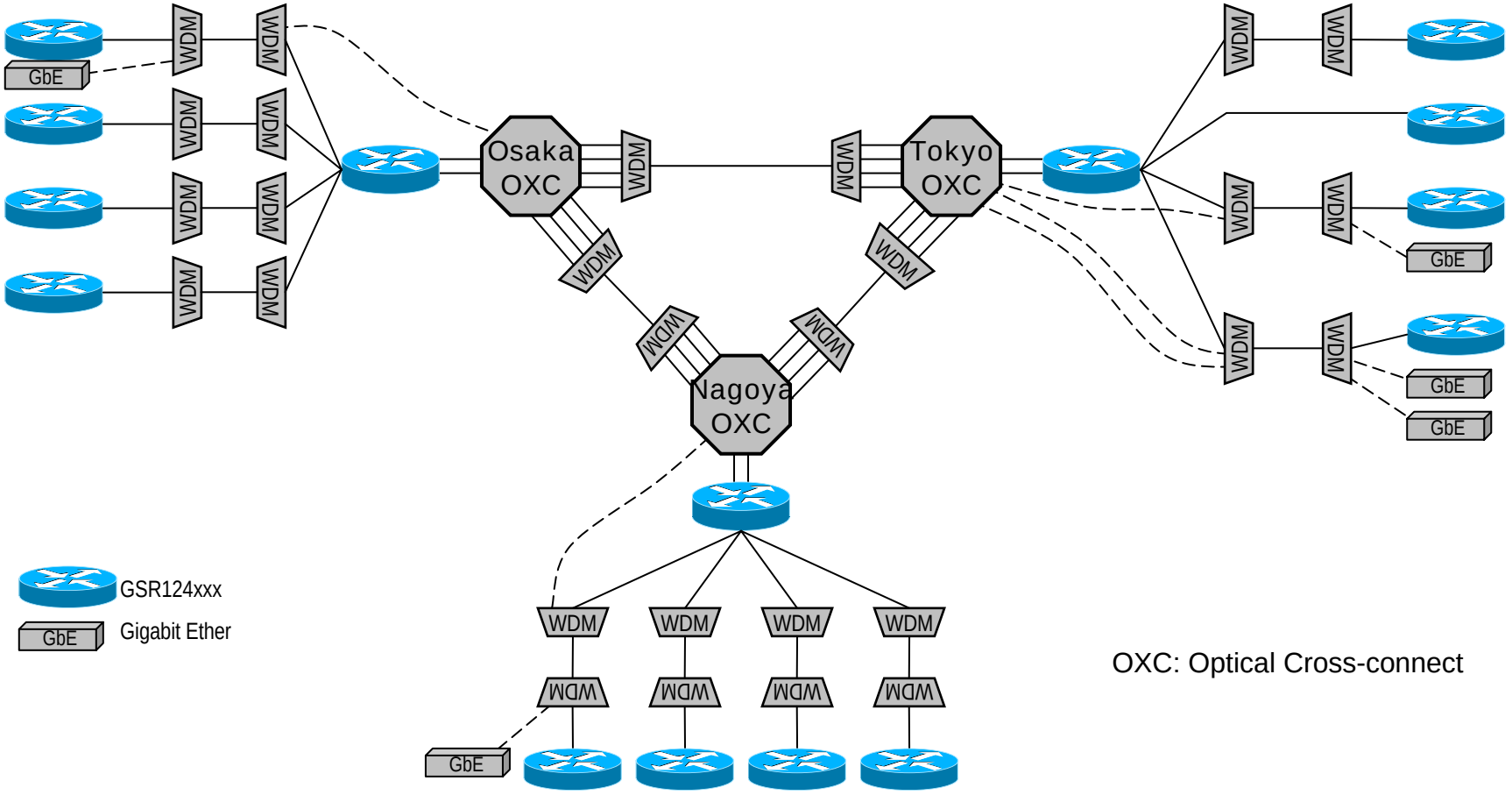
SuperSINET



Only some nodes are shown.

SuperSINET Network Configuration

SuperSINET is composed of multiple lambdas constructed with dark fibers and DWDM.



SuperSINET/SINET→SINET3

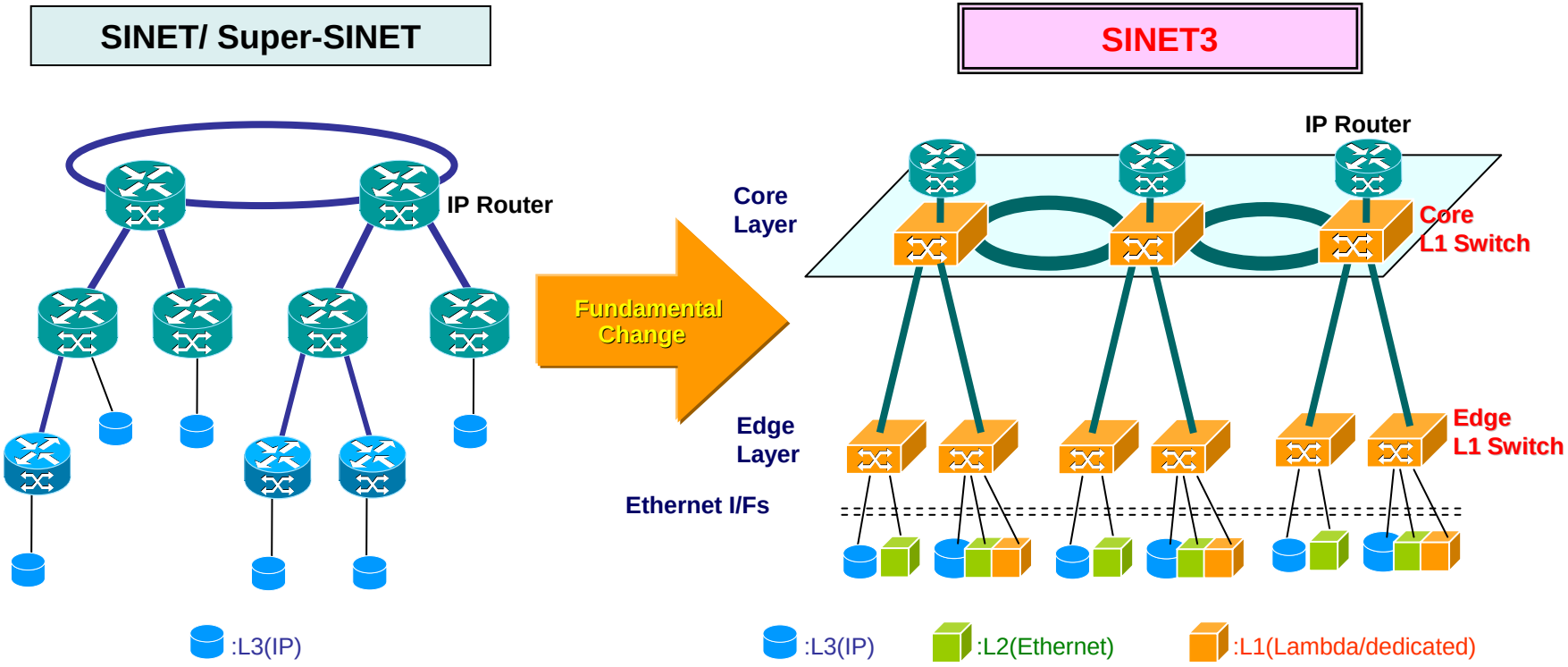
- SuperSINET (will end in Mar 2007)
  - 10G IP/MPLS
  - Dedicated λ's (GbE and OC48)
    - End-end bandwidth guaranteed

  eg. 10G IP/MPLS + 10xDedicated GbE for KEK
- SINET3 (will start in Apr 2007)
  - 10GbE or 2 x 10GbE or OC48 or GbE
    - In-band VLAN's
    - End-end bandwidth not guaranteed

  eg. 2x10GbE – 2xOC48 for KEK

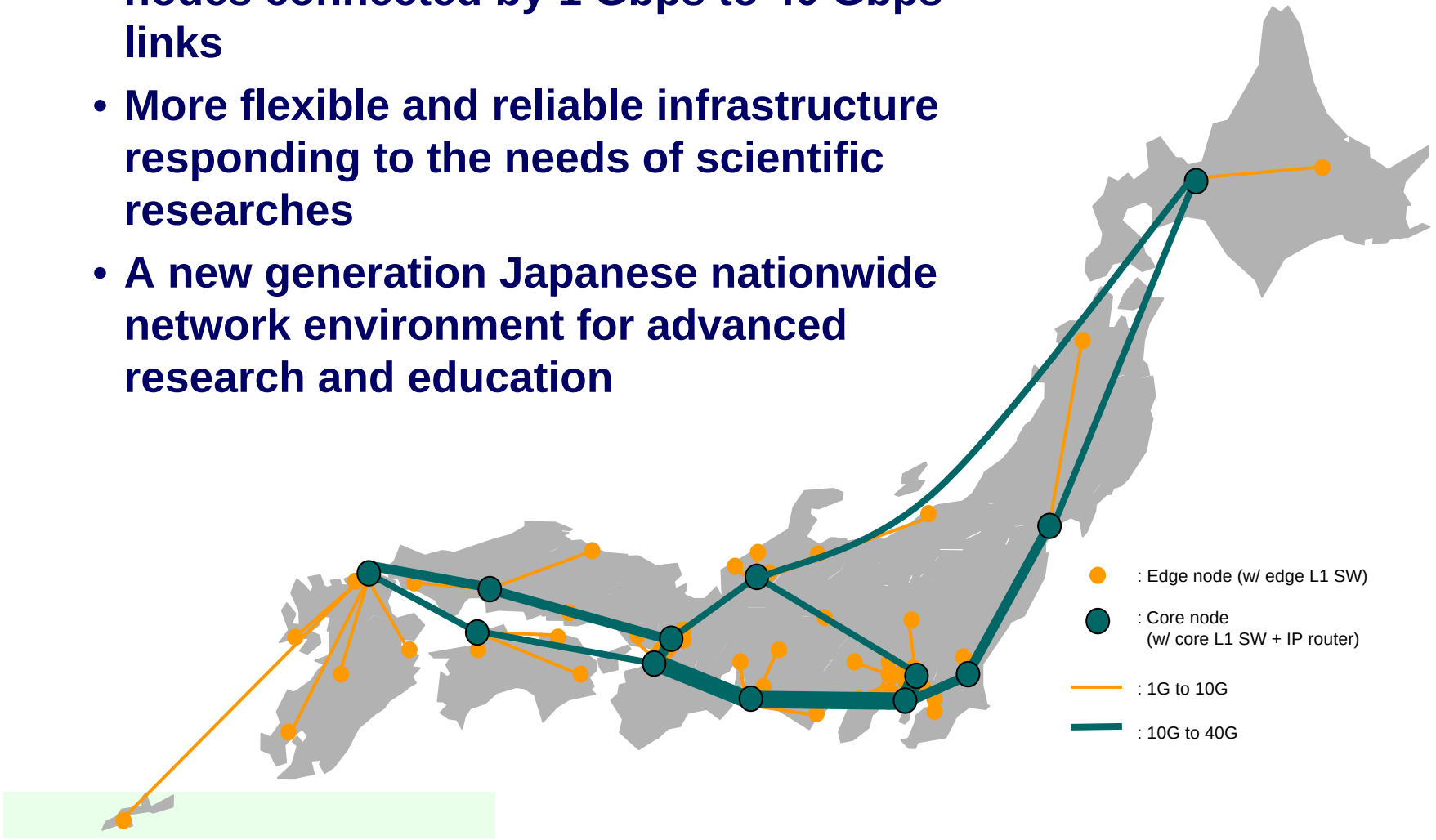
SINET3: Change in Network Structure

- **Two tier structure with edge and core layers.**
- **The edge layer consists of edge layer 1 switches with Ethernet interfaces to accommodate users' equipment.**
- **The core layer consists of core layer 1 switches and high-performance IP routers and constitutes a nationwide reliable backbone network.**



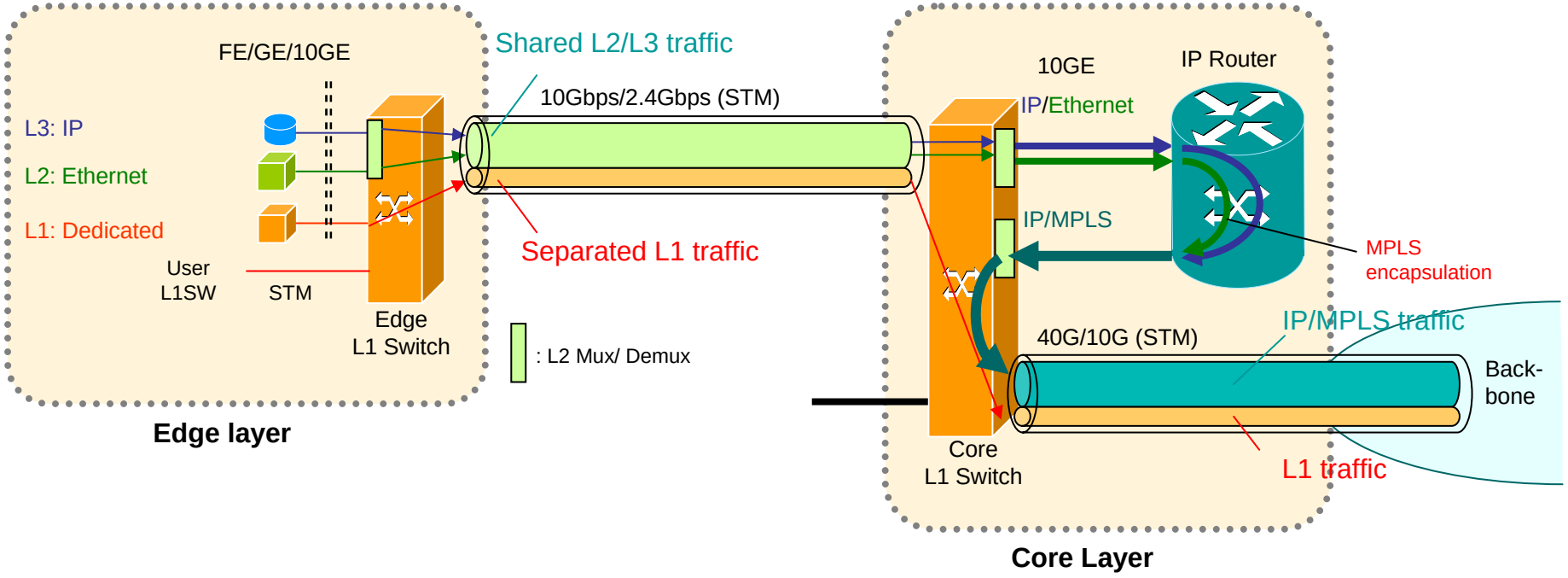
SINET3: Nationwide Multi-layer Network

- **More than 60 edge nodes and 12 core nodes connected by 1 Gbps to 40 Gbps links**
- **More flexible and reliable infrastructure responding to the needs of scientific researches**
- **A new generation Japanese nationwide network environment for advanced research and education**



SINET3 Traffic Accommodation for Layers 1 to 3

- **Edge L1 switch:**
  - Users' L1/L2/L3 traffic is accommodated and transferred to a 10Gbps(STM) line.
  - L1 traffic is assigned a dedicated bandwidth and separated from L2/L3 traffic.
  - L2/L3 traffic shares the remaining bandwidth by L2 multiplexing.
- **Core L1 switch:**
  - L1 path is switched internally.
  - L2/L3 traffic is forwarded to and received back from IP router.
- **IP Router:**
  - IP/MPLS traffic is forwarded. L2 traffic is encapsulated using MPLS.



SuperSINET→SINET3

- Major changes
  - DWDM device → L1 switch
  - Protected circuit → Non-protected circuit (L2)
  - Protected circuit (L1)'s are provided if requested and if approved, but having them lowers the bandwidth for the shared traffic.
  - So we decided to have all the HEP traffic on the shared bandwidth.
  - Having end-end L2 paths becomes very easy.
  - It is said that if the bandwidth usage becomes high the bandwidth will be upgraded, but how immediately can it be made?

HEP networking

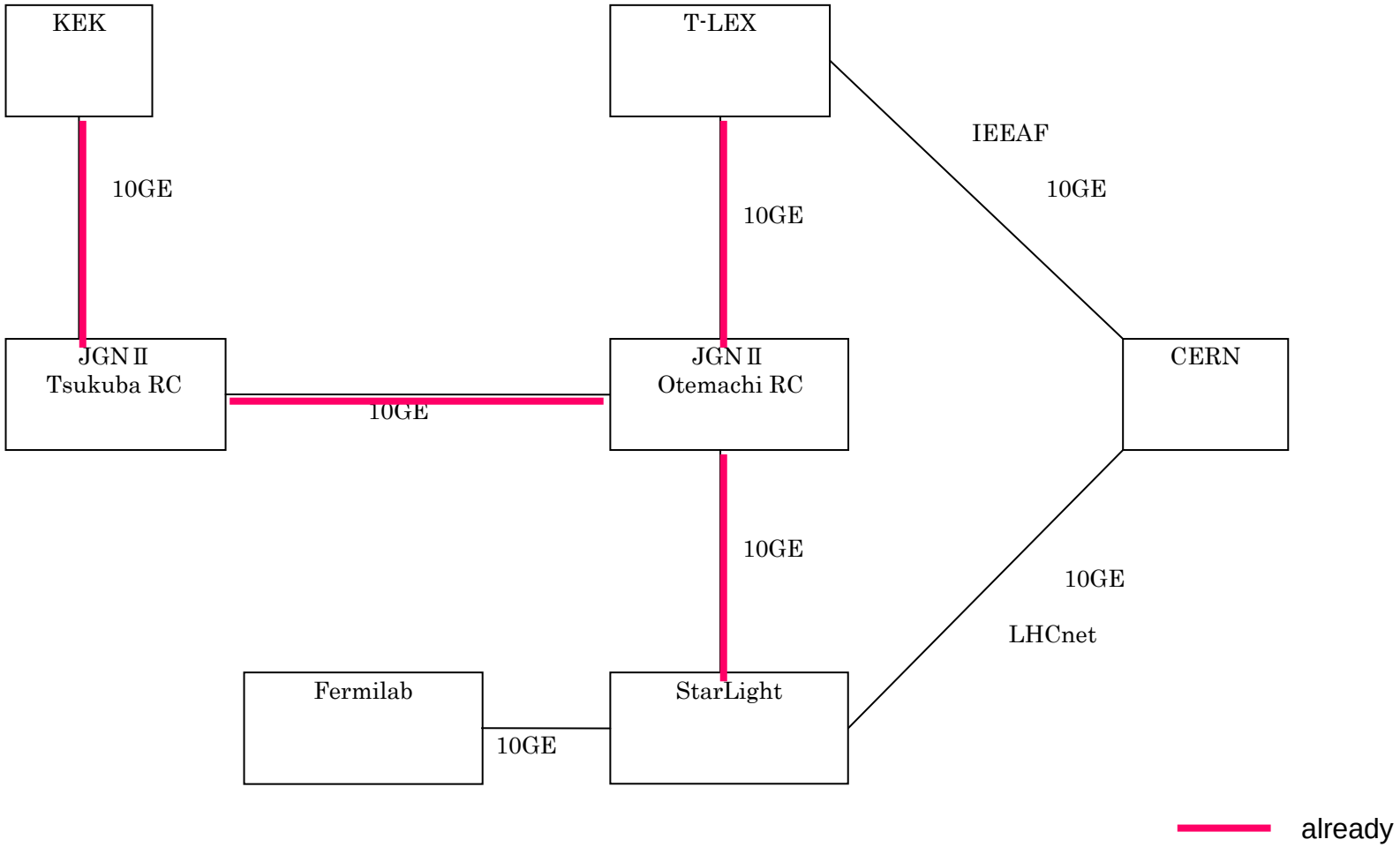
- Production Network
  - Domestic
    - SuperSINET/SINET → SINET3
    - Wide-area Ethernet (for some non-SINETuniv's)

    Creating VLANs for Grid VOs is very easy.
  - International
    - SINET
    - APAN (for connections in Asia-Pacific region)
    - Dedicated line (to BINP, will be terminated soon.)
- Testbed
  - JGN2

Our use of JGN2

- Testbed
  - Participate in the UltraLight.
  - Evaluate the advantages of “international broadband L2 connection”
    - in Grid, in security, in performance, …
    - KEK-CERN, KEK-FNAL, …
  - Evaluate GMPLS (or on-demand L1)
  - And propose some in the upgrade of the SINET3
- L2 paths for peering
  - Direct peering with APAN
  - Direct peering with ASCC

KEK connected to JGN2 and to Starlight with 10GE in September 2005.



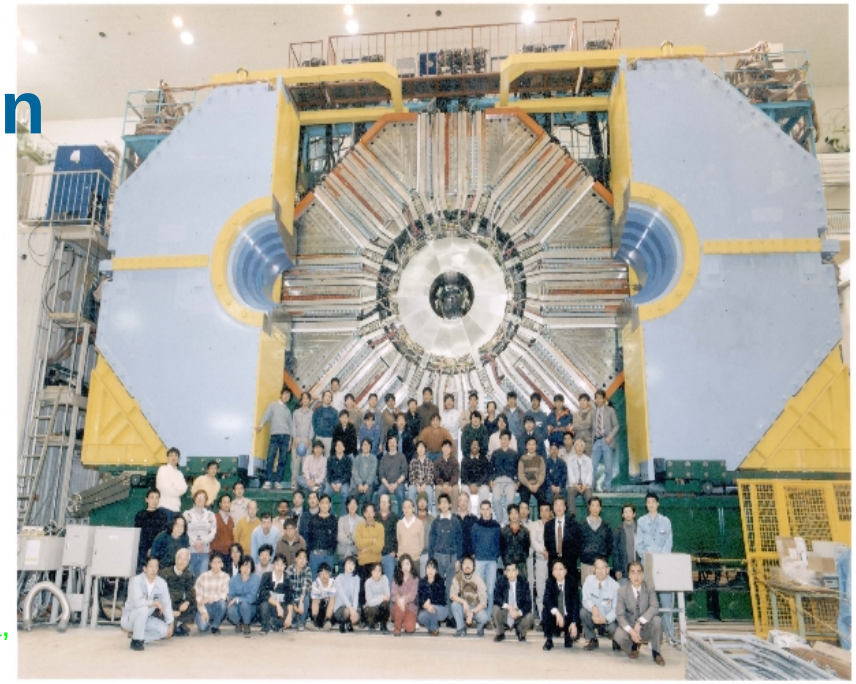
HEP computing in Japan

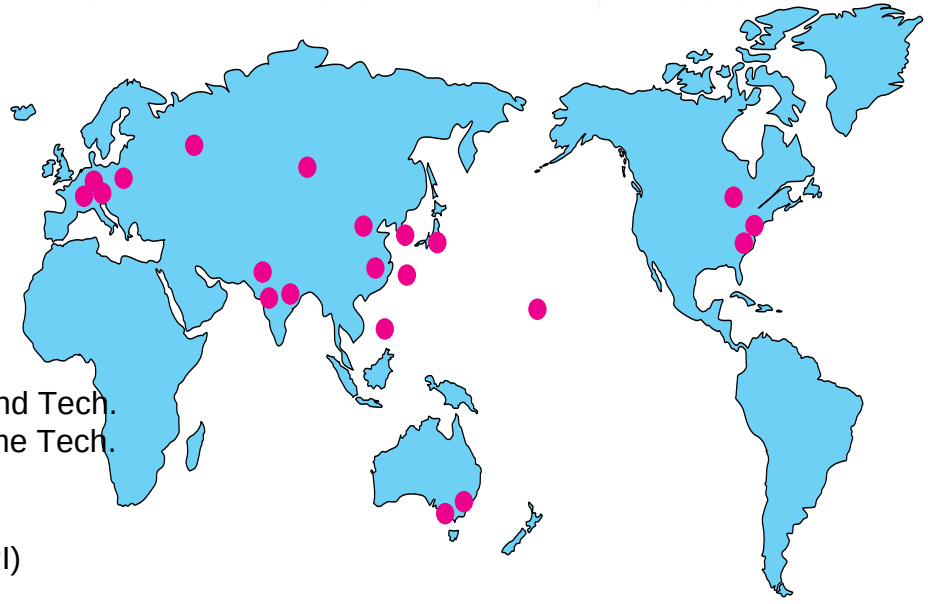
- In addition to the central computer system, KEK has two other computer systems dedicated to:
  - Belle Data Analysis
  - Lattice QCD
- Universities have some amount of computing resources.

The Belle Collaboration

>300 researchers from 55 institutes

- Aomori Univ.
- Budker Inst. of Nucl. Physics, RU
- Chiba Univ.
- Chuo Univ.
- Univ. of Cincinnati
- Univ. of Frankfurt
- **Gyeongsang Nat'l Univ., KR**
- Univ. of Hawaii
- Hiroshima Inst. of Tech.
- Hiroshima Coll. of Maritime Tech.
- Inst of Cosmic Ray Res., U of Tokyo
- IHEP, CN
- ITEP, RU
- Joint Crystal Collab. Group
- Kanagawa Univ.
- KEK
- **Korea Univ., KR**
- **Krakow** Inst of Nucl Physics
- Kyoto Univ.
- **Kyungpook Nat'l Univ (CHEP), KR**
- Univ. of Melbourne., AU
- Nagasaki Inst. of Applied Science
- Nagaya Univ.
- Nara Woman's Univ
- Nat'l Central Univ., TW
- Nat'l Kaoshiung Univ, TW
- Nat'l Lien-Ho Coll. of Tech., TW
- Nat'l Taiwan Univ., TW
- Nihon Dental Coll.
- Niigata Univ.
- Osaka Univ.
- Osaka City Univ.
- Panjab Univ., IN
- Peking Univ., CN
- Saga Univ.
- **Seoul Nat'l Univ., KR**
- Univ. of Sci. and Tech. of China, CN
- Sugiyama Woman's Coll.
- **Sungkyunkwan Univ., KR**
- Univ. of Sydney, AU
- Tata Inst., IN
- Toho Univ.
- Tohoku Univ.
- Tohoku-gakuin Univ.
- Univ. of Tokyo
- Tokyo Inst. of Tech.
- Tokyo Metropolitan Univ.
- Tokyo Univ. of Agriculture and Tech.
- Toyama Nat'l Coll. of Maritime Tech.
- Univ. of Tsukuba
- Utkal Univ., IN
- Virginia Polytechnic Inst (VPI)
- Yokkkaichi Univ.
- **Yonsei Univ., KR**





An on-going HEP experiment at KEK, Japan

New Computer System for Belle

- **Installed in March 23. 2006**
- **History of the Computer System for Belle**

Performance＼Year	1997- (4years)	2001- (5years)	2006- (6years)
Computing Server **（SPECint2000 rate）**	~100 **(WS)**	~1,250 **(WS+PC)**	~42,500 **(PC)**
Disk Capacity （TB）	~4	~9	1,000 (1PB)
Tape Library Capacity （TB）	160	620	3,500 (3.5PB)
Work Group server **（# of hosts ）**	3+(9)	11	80+16FS
User Workstation **（# of hosts）**	25WS +68X	23WS +100PC	128PC

Moore’s Law 1.5y=twice　4y=~6.3　5y=~10

New Computer System for Belle

1 Enclosure = 10 nodes/ 7U space
1 rack = 50 nodes
25 racks = 4 arrays

- **Computing Server (CS)**
  - **CS＋WG servers(80)**
    = 1208 nodes=2416CPU
    =45,662 SPEC CINT 2k rate
    =8.7THz
  - **DELL Power Edge 1855**
    Xeon3.6GHz x2,
    memory 1GB
  - **Linux**
    (CentOS/CS,REL/WGS)

New Computer System for Belle

- Storage System (SS)

  –disk-

  – 1,000TB, 42FileServ.
  – Nexan + ADTeX + SystemWks
  – SATAII **500G dr.**

    ×~2000

    (~1.8 failures/day?)
  – HSM = 370TB non HSM (no Bck)

    = 630TB

New Computer System for Belle

- **Storage System (SS)**

  -tape-

  – **HSM**
    - **3.5PB + 60drv + 13srv**
    - **SAIT 500GB/volume**
    - **30MB/s drive**
    - **Petaserv(SONY)**

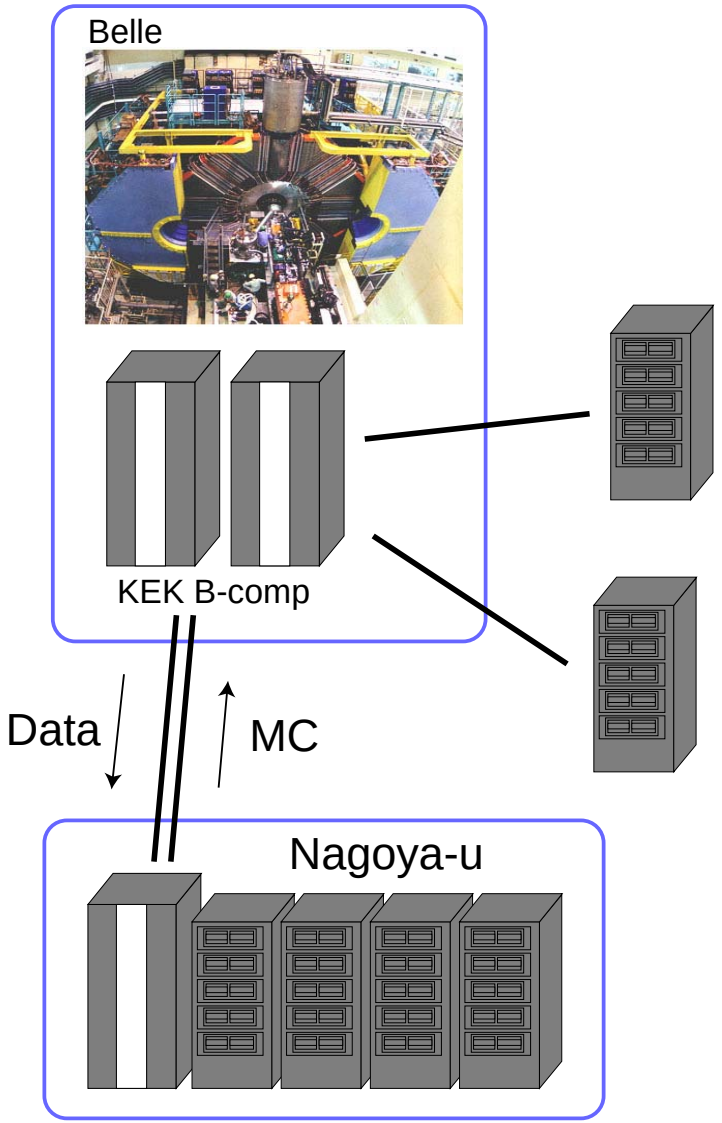
  – **WFS backup**
    - **90TB + 12drv + 3srv**
    - **LTO3 400GB/volume**
    - **NetVault**

New Computer System for Belle

- System usage
  - User Workstations (PC) are used for the network terminal.
  - Users login to Work Group Server（WGS; 4~5persons/host）.
  - 1208 Computing Servers divided into 3 LSF (Batch System) clusters.
  - WGS(80svr) shares WFS(16srv+80TB) as NFS user home directories.
  - Not-so-frequently-modified applications/Libraries are held in 50 NFS servers. 1140 CS shares the NFS server.
  - Exp. Data (Many and Big size)
    - is transferred between CS(1140) and Storage Servers（42)
    - by using a Belle self-made simple TCP/socket application.
    - Data is managed by cooperation with DB system.
  - Storage System ⇔Computing server transfer performance spec.
    - CS/WGS 1/3(540) ⇔ SS = 10GB/s
    - CS 2/3(1080) ⇔ SS/HSM = 0.5GB/s
    - SS/HSM ⇔ SS/nonHSM = 0.5GB/s

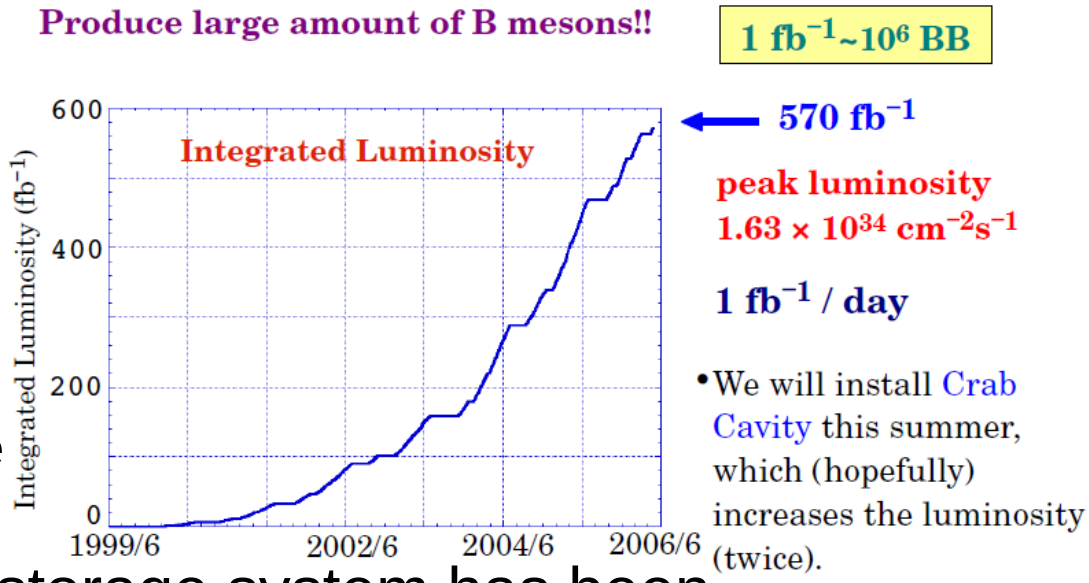
Universities in Belle
eg. Nagoya University

- Storage for Belle at KEK
  - Disk 1PetaBytes
  - Tape 3.5PetaBytes
- Storage and computers for Belle at Nagoya-U
  - Disk 530TeraBytes
  - Linux PC farm ~1200GHz
- Use of computers at Nagoya-U
  - Belle data analysis
  - Generation of Monte-Carlo samples
  - Simulation for detector development
- Data transfer between KEK and Nagoya-U
  - Use of L2 connction in SuperSINET→SINET3
  - Requiring more than 1Gbps



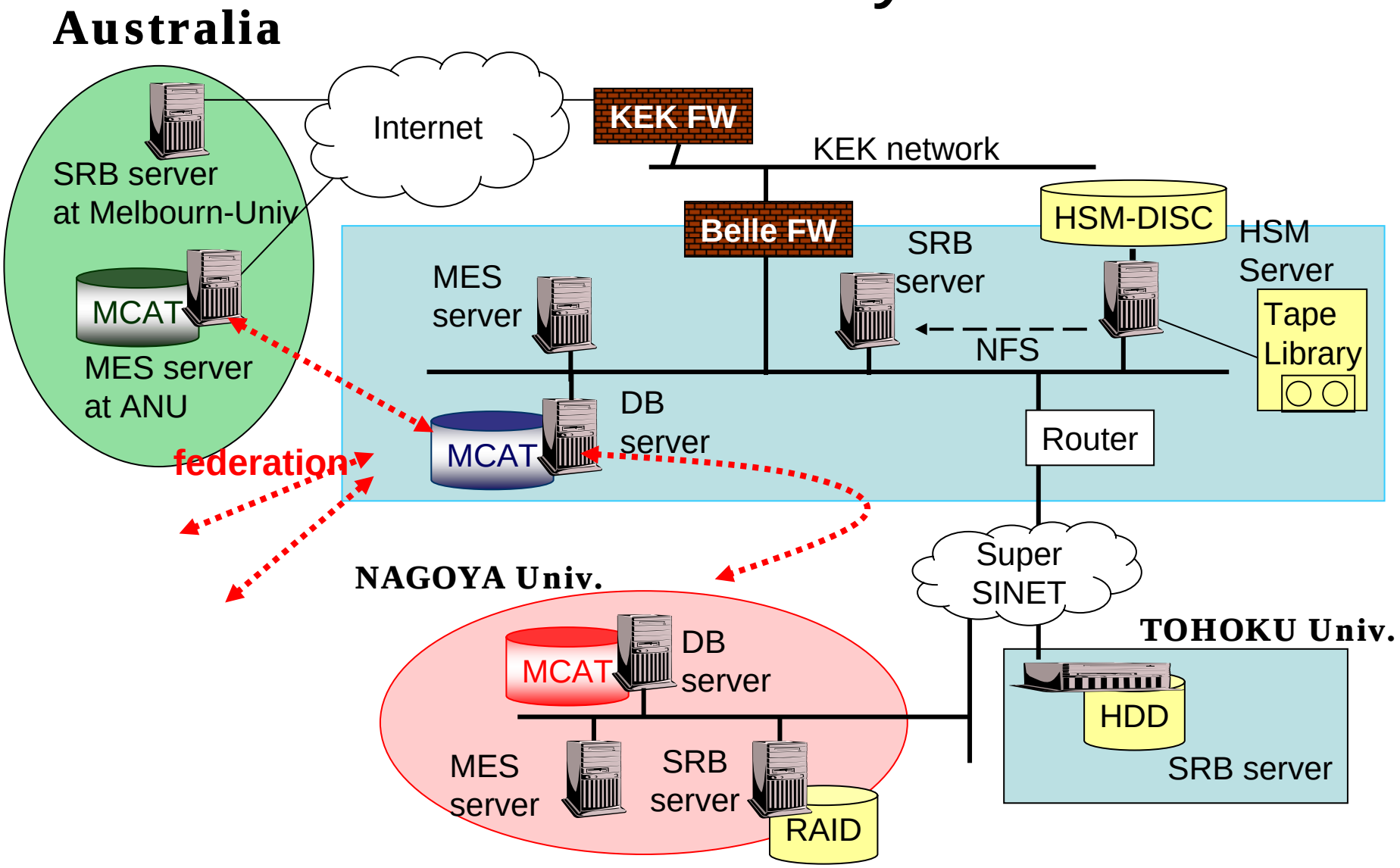
KEK B Factory : Belle Experiment

- Data accumulated so far
  - 1.5 PB
    including Simulation data
  - Recent data acquisition rate
    ~ 1.0TB/day



- SRB servers for real data storage system has been implemented in Aug. 2005.
  - Current active data sharing
    - among KEK, U. Melbourne (Australia), and Nagoya Univ.
    - Target storage space 120 TB
    - files registered to MCAT
      ~ 423 files as of Sep. 6

The Belle SRB system



GRID deployment at KEK

- Bare Globus
  - We started with this in 2001
  - No production use, only R&D
- SRB (2002 -)
  - GSI authentication or password
  - SRB-DSI became available
    - Works as SRM for the SRB world from LCG side
    - Performance test will be done
  - Performance tests among RAL, CC-IN2P3 and KEK is on going
- Gfarm
  - Collaboration with AIST
  - Lattice QCD GRID
- LCG (2004-
  - JP-KEK-CRC-01
    - For R&D
  - JP-KEK-CRC-02
    - For production
    - Interface to HPSS
  - Test bed (JP-KEK-CRC-00)
    - Staff training and tests
- We have our own GRID CA
  - In production since this January
  - Accredited by APGRID PMA
- NAREGI test bed
  - Under construction

Pre-production LCG site for Belle

(JP-KEK-CRC-01)

- Pre-production site was built up with LCG2.7 March, 2006
- Certification by APROC for registration to GOCDB has been done at the end of March
- New VO: Belle has been registered to the LCG/EGEE as a global VO.
- Initial collaboration sites expected:
  - Melbourne, ASGC, Krakow, Jozef Stefan Institute (Slovenia), IHEP Vienna
  - Nagoya U.

LCG Deployment plan at KEK

- New Computer Systems
  - Central Information System since Feb. 20. ’06
  - Belle Computer System since Mar. 23. ‘06
- 1st Phase
  - LCG and SRB for production usage are available on the Grid System in the new Central Information System.
    - Not for public usage, but for supporting projects
    - Under system maintenance in contract with IBM-Japan
    - WN: 36 nodes x 2 =72 CPU
    - Storage: Disk (2TB) + HPSS(~200TB)
    - Supported VO: Belle, APDG, Atlas_J
  - Service start in ~ May 2006
- 2nd Phase
  - Full support in the Belle production system

Belle GRID

- Starting slowly using SRB and LCG
  - LCG site: JP-KEK-CRC-02
- Data distribution service using SRB-DSI
  - Belle already have a few PBs data in total including 100s TB DST and MC
    - Bulk file register helps us: Sregister
    - we do not move any of them
  - Benefits both for native SRB users and LCG users
- VO is supported by KEK
  - Nagoya, Melbourne, Academia Sinica, Krakow and etc

New Supercomputer System at KEK

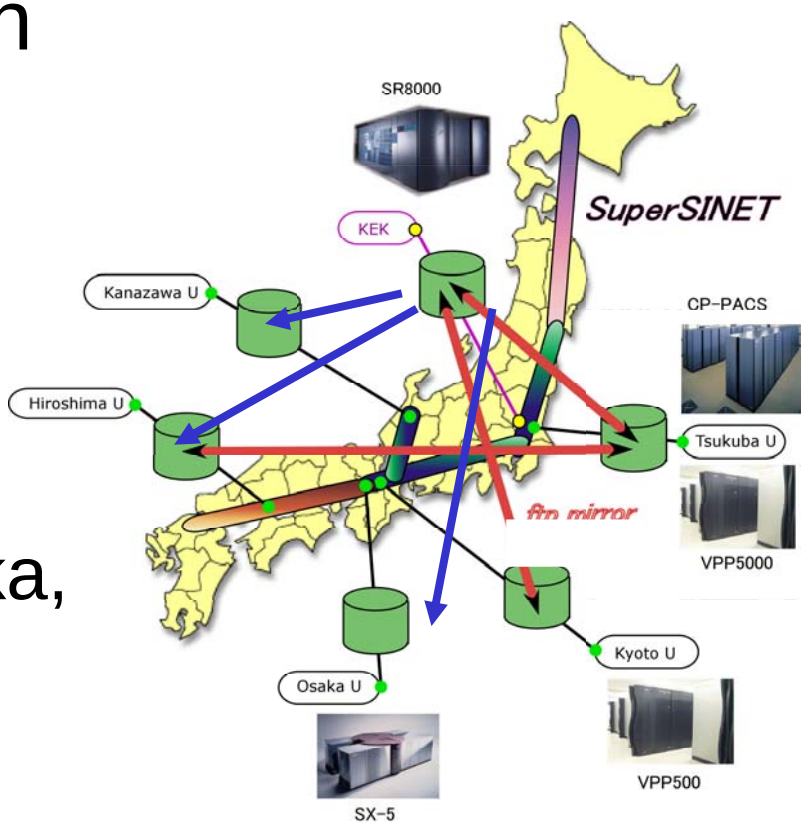
From 2006 March 1st

- Large Scale Simulation for particle and nuclear physics research and accelerator-related scientific studies
  - Hitachi SR11000 K1 System,
    - 16 nodes
    - 2.15 Tflops (Theoretical peak)
    - Large memory capacity 32GB/64 GB/node
  - IBM BlueGene Solution,
    - 10 racks
    - 57.3 Tflops (Theoretical peak)
    - Massive parallel system for Lattice QCD simulation

About 50 times faster than former Supercomputer system (Hitachi SR8000 100 node system)

HEPnet-J/sc

- HEPnet-J/sc
  - A VPN constructed in SuperSINET
  - Connecting major lattice QCD sites in Japan
    - KEK, Tsukuba, Osaka, Kyoto, Kanazawa, Hiroshima
  - File mirroring



Lattice QCD Data Grid

In lattice QCD, gauge field configuration is essential data

- Generated by Monte Carlo algorithms
  - Full QCD requires large computational resources
- Once generated,

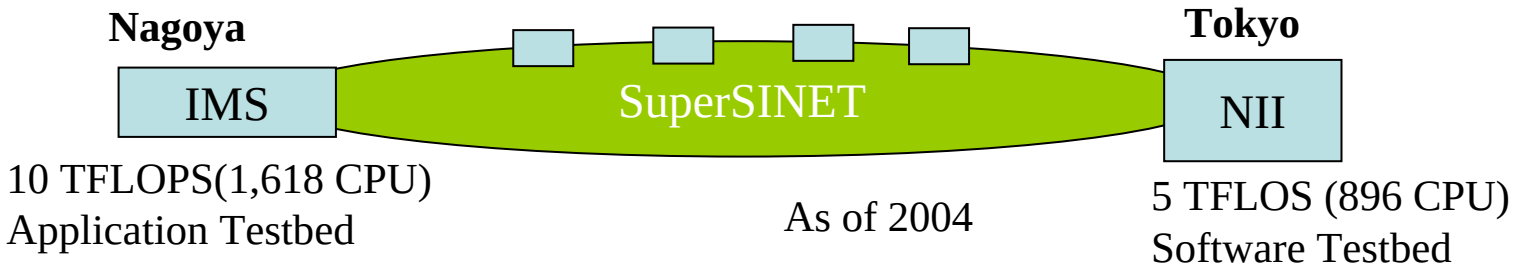
  various correlation functions can be measured
  - Hadron spectra, decay constants, matrix elements, etc.
  - Exotic hadrons, interaction between hadrons
- Data size
  - Degree of freedom: SU(3)$\times$site(x,y,z,t)$\times$direction(4),
  - statistics O(1000)
- Various actions, sizes, parameters
  - Extrapolations to continuum, small quark mass, large volume limits
  - Comparison for consistency check
- Sharing gauge configuration is now world movement

ILDG and JLDG

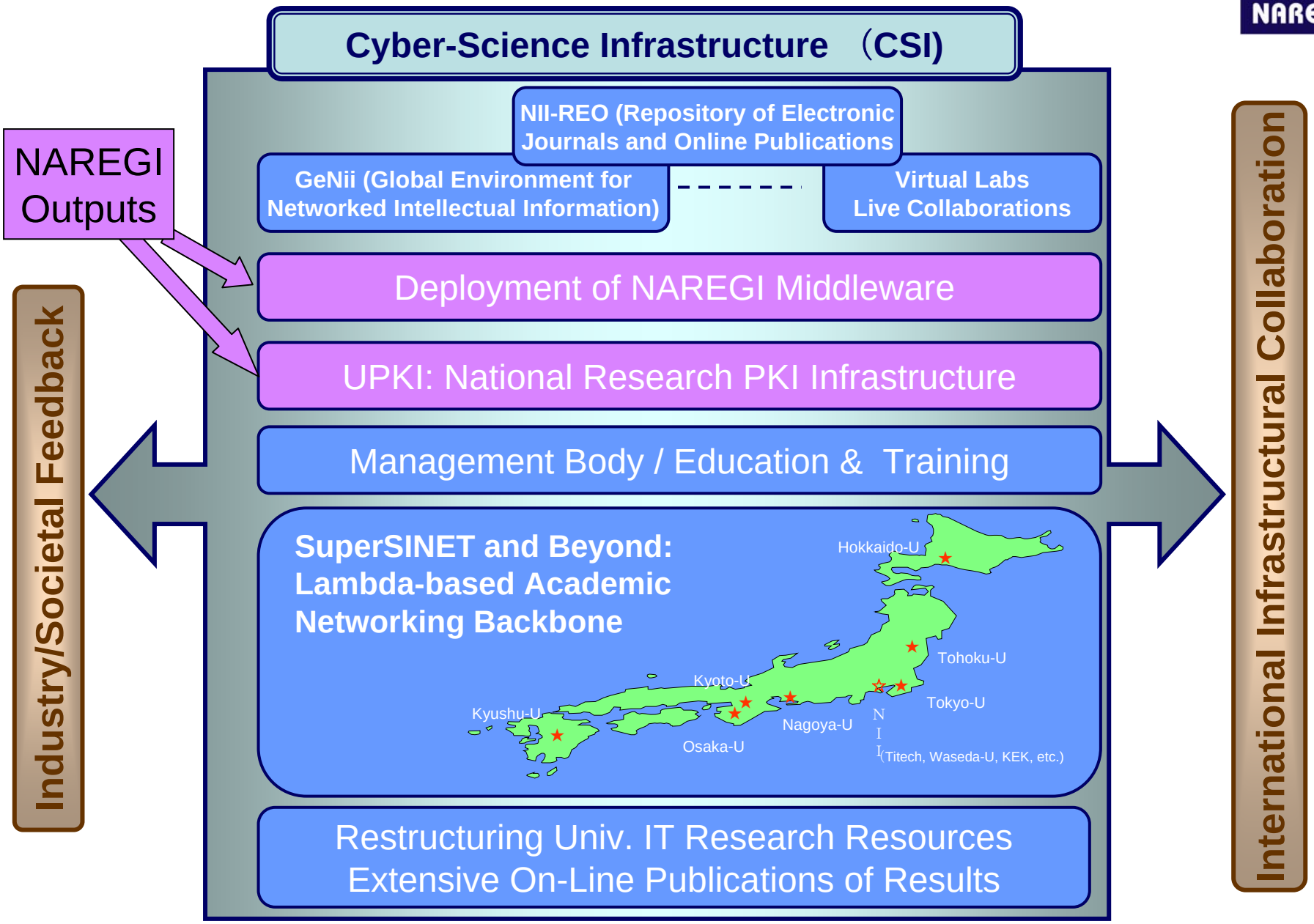
- ILDG: International Lattice DataGrid
  - International organization for data sharing
  - Developing mark-up language, middleware
  - Officially started in June 2006
  - Several sites are already providing data:
    - LQA(Lattice QCD Archive)@Tsukuba Univ.
    - Gauge Connection (NERSC, USA)
- JLDG: Japan Lattice DataGrid
  - National community to share lattice data on HEPnet-J/sc
  - Provides data to ILDG
  - Developping file system and middleware (interface to ILDG)

National Research Grid Initiatives (NAREGI)

- Apr. 2003 MEXT funded NAREGI 5 years Project
- Development of Grid infrastructure and an application for promotion of national economy
- Target application is nano science and technology for new material design
- Players
  - Computing & networking: NII, AIST, TITEC
  - Material scientists :IMS, U. Tokyo, Tohoku U., Kyushu U., KEK, ..
  - Companies: Fujitsu, Hitachi, NEC
- Distributed facility: Computing Grid up to 100 TFLOPS in total
- Extended to 2010 as a part of National Peta-scale Computing Project



National Institute of Informatics



NAREGI

- What we expect for NAREGI
  - Better quality
  - Easier deployment
  - Better support in the native language
- What we need but still looks not in NAREGI
  - File/replica catalogue and data GRID related functionalities
    - Need more assessments
- Comes a little bit late
  - Earlier is better for us
    - We need something working today!
- Require commercial version of PBS for β 1

gLite and NAREGI interoperability

- NAREGI has much interests on interoperability because they came late and they decided to establish in their side
- First meeting at CERN
  - March 2006
  - NAREGI, gLite, and KEK
- Second meeting at GGF Tokyo

KEK plan

- VO hosted at KEK using LCG
  - Belle, APDG, ILC
- Ask NAREGI to implement LFC on their middleware
  - We assume job submission between them will be realized soon
  - Share the same file/replica catalogue space between LCG/gLite and NAREGI
    - Move data between them using GridFTP
  - Try something by ourselves
    - Brute force porting of LFC on NAREGI
- NAREGI<->SRB<->gLite will be tried also

- Assessments will be done for
  - Command level compatibility (syntax) between NAREGI and gLite
  - Job description languages
  - Software in experiments, especially ATLAS
    - How depends on LCG/gLite?

Domestic support

- KEK will try to support domestic institutions centrally
  - We seek funding and technical schemes
    - Proposal to funding agencies has been done to establish HEPNET-J VO for the test
      - Installation and operation
        - » Send technicians and engineers from KEK temporally for the installation
        - » Operation and monitor centrally

Future strategy

- ILC, International Linear Collider, will be a target
  - interoperability among gLite, OSG and NAREGI will be required
- ILC Japan people want to start to work with French collaborators as soon as possible using LCG
    - What we do relating AIL?
      - DESY already hosts some VO's for ILC
- SRB and ROD (Resource on Demand)s in the future

APDG

- Asia Pacific Data GRID
- Collaboration among Academia Sinica(TW), Center for HEP-Korea, University of Melbourne and KEK
- Regular meetings, workshops and conferences
- KEK is seeking tighter collaboration with ASGC(Academia Sinca Grid Computing Centre),
  - GOC in Asia

KEK-ICEPP Relation

- ICEPP (International Center for Elementary Particle Physics), U of Tokyo is the tier-2 center of ATLAS in Japan
    - No tier-1 in Japan
- KEK and ICEPP are collaborating each other on GRID related issue
- Universities in ATLAS will get a technical support from KEK
- CA services are provided by KEK

Strategy on GRID

- Deployment at KEK for major groups
  - BELLE
  - ILC
- University support
  - education and training
  - Deployment at smaller centers
    - HEPNET-J VO

Thank you!