The Tokyo Regional Analysis Center Site Report

Michiru Kaneda

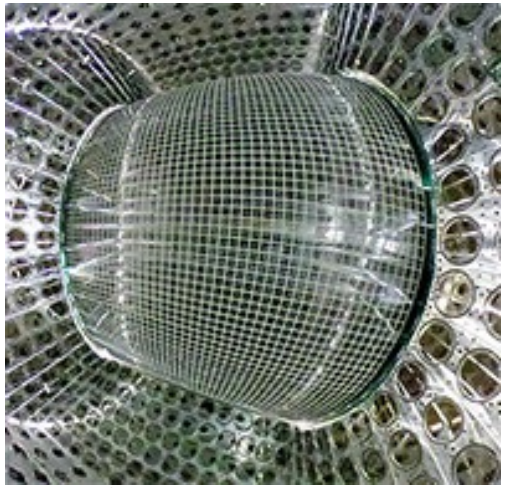
The International Center for Elementary Particle Physics (ICEPP), The University of Tokyo

24/Oct/2019, The 5th Asia Tier Center Forum, Mumbai, India

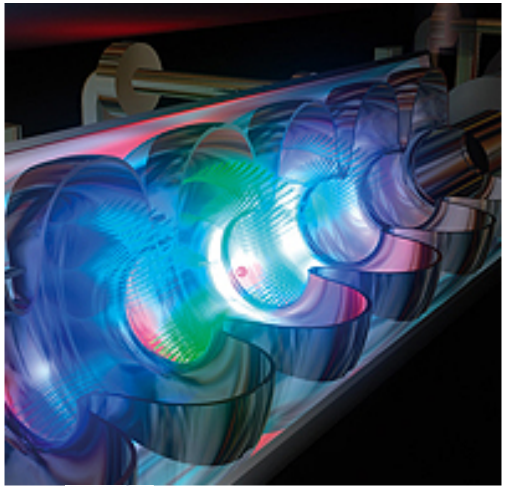
International Center for Elementary Particle Physics

ICEPP
The University of Tokyo

ATLAS



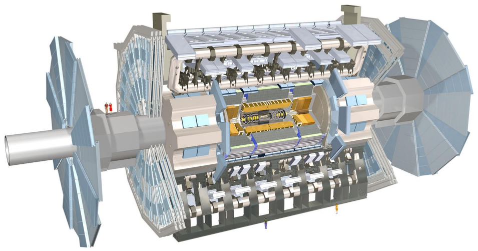
MEG



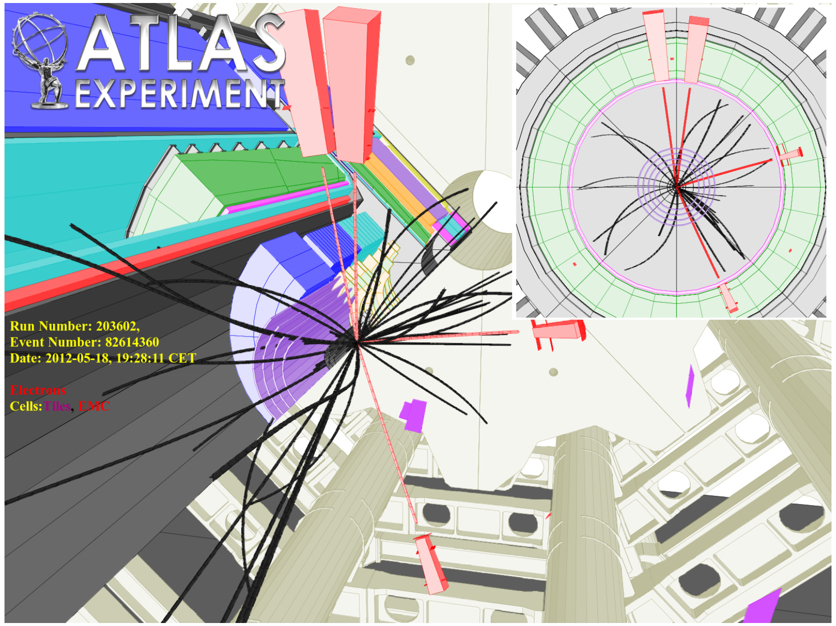
ILC

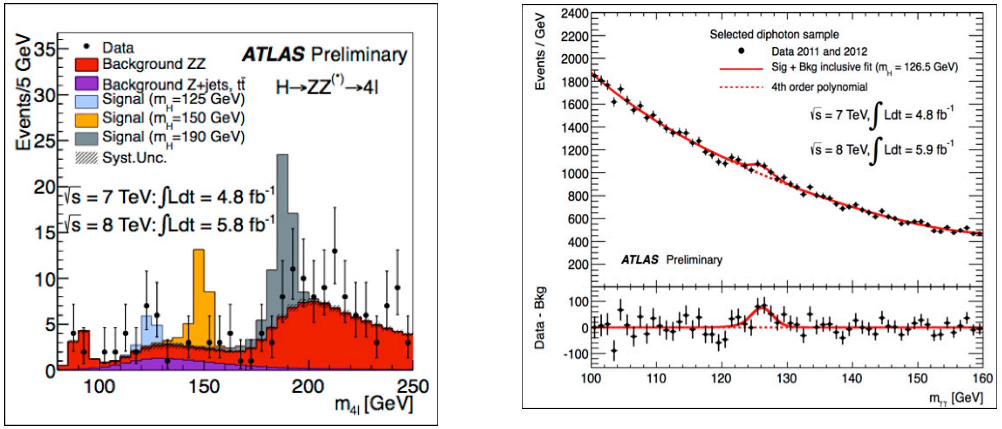
- The Tokyo regional analysis center at ICEPP:
  →Computing center for the ATLAS experiment
  →WLCG Tier2 site (only site in the ATLAS Japan)

The ATLAS Experiment



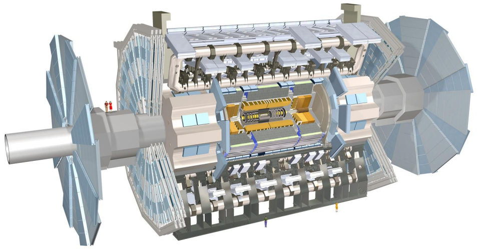
ALICE
CMS
ATLAS
LHCb





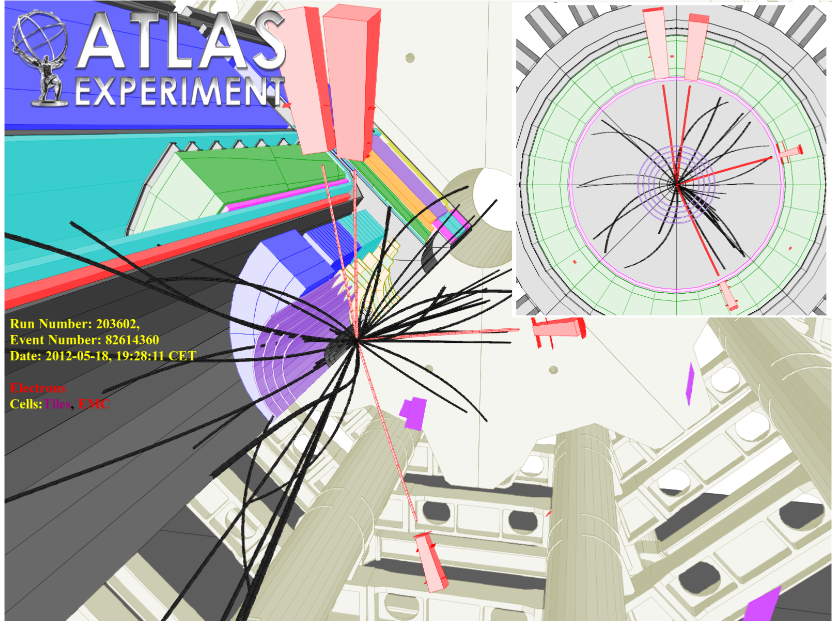
The Higgs Boson Discovery in 2012

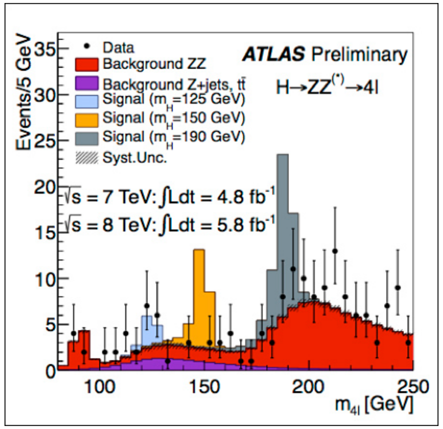
The ATLAS Experiment

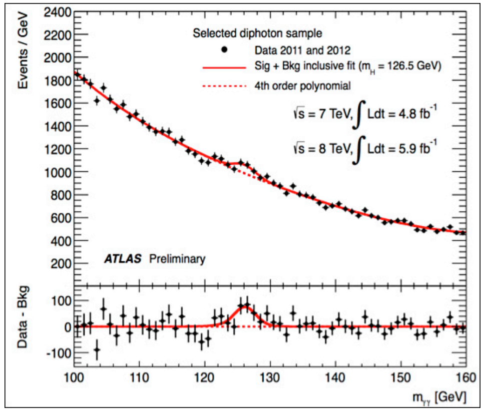


ALICE
CMS
ATLAS
LHCb

Raw data: ~1GB/s, ~10PB/year, Current total data size (including MC): >200PB

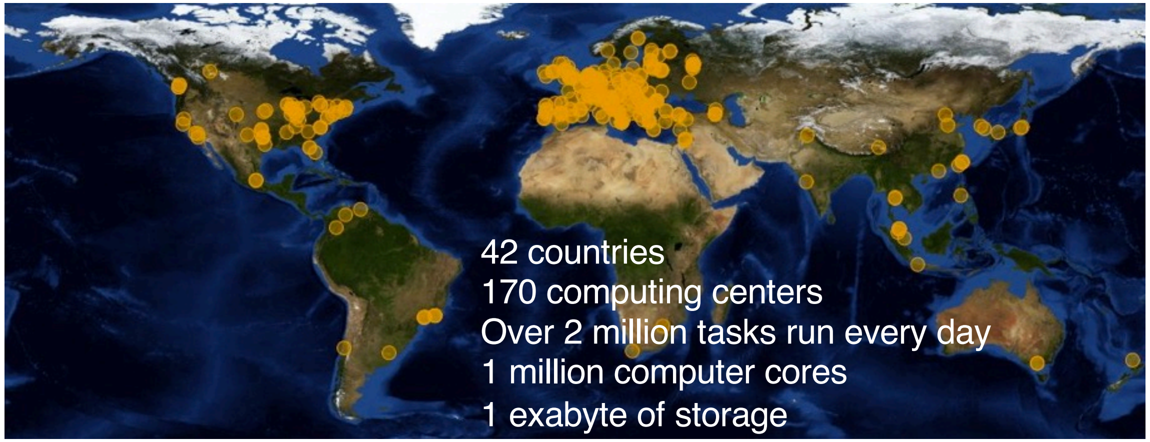




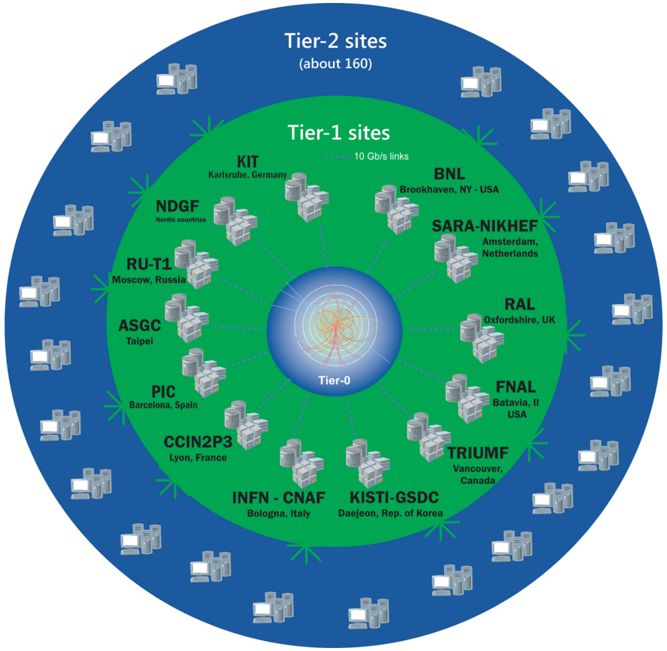


The Higgs Boson Discovery in 2012

Worldwide LHC Computing Grid (WLCG)

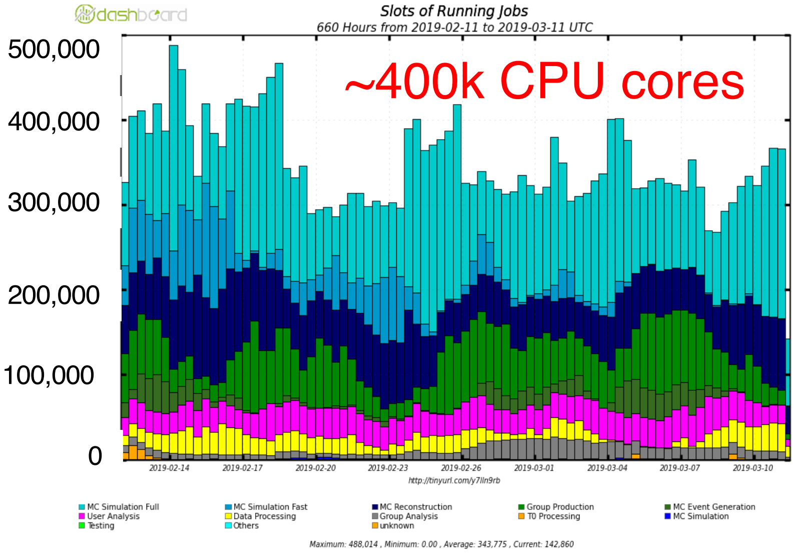


42 countries
170 computing centers
Over 2 million tasks run every day
1 million computer cores
1 exabyte of storage



- A global computing collaboration for LHC
  → Tier0 is CERN

- The Tokyo regional analysis center is one of Tier2 for ATLAS

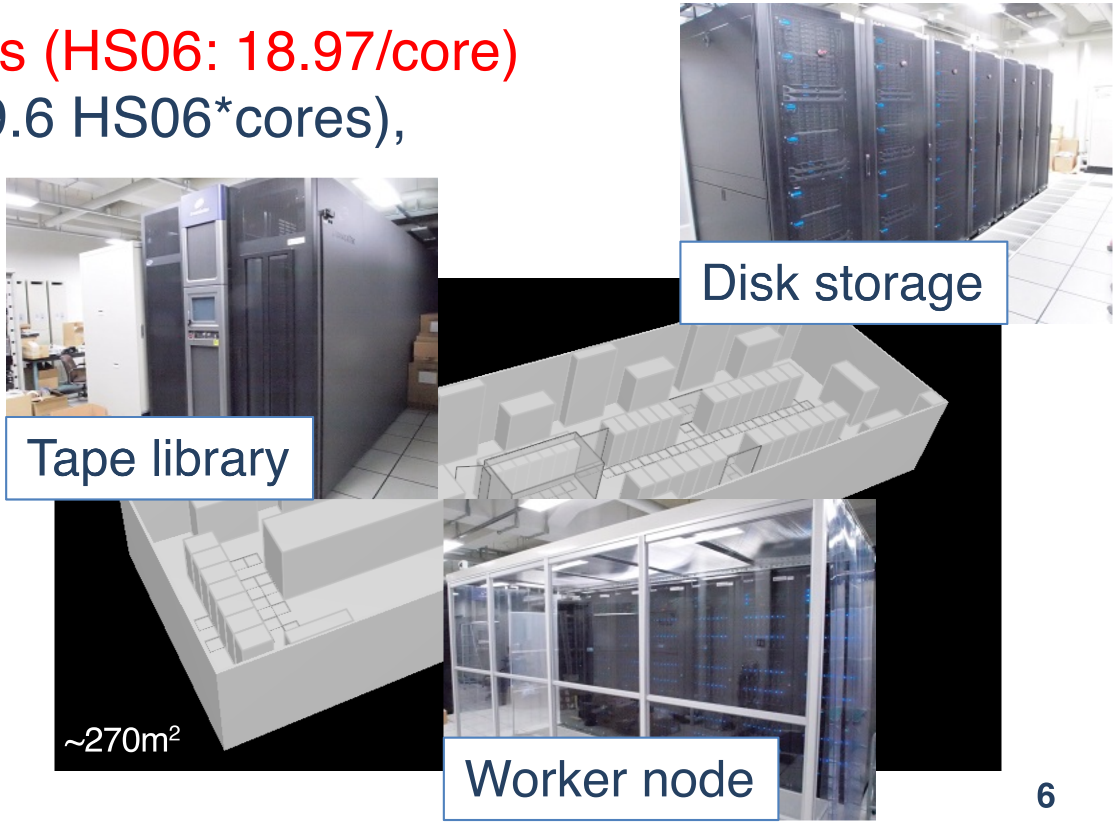


~400k CPU cores

Number of cores used by ATLAS

The Tokyo regional analysis center

- The computing center at ICEPP, the University of Tokyo
- Supports ATLAS VO as one of the WLCG Tier2 sites
  →Provides local resources to the ATLAS Japan group, too
- All hardware devices are supplied by the three years rental
  → All hardware devices are renewed in three years
- Current system (Starting from Jan/2019):
  →Worker node: 10,752cores (HS06: 18.97/core) (7,680 for WLCG, 145689.6 HS06*cores), 3.0GB/core
  →File server: 15,840TB, (10,560TB for WLCG)



The Tokyo regional analysis center

- The computing center at ICEPP, the University of Tokyo
- Supports ATLAS VO as one of the WLCG Tier2 sites

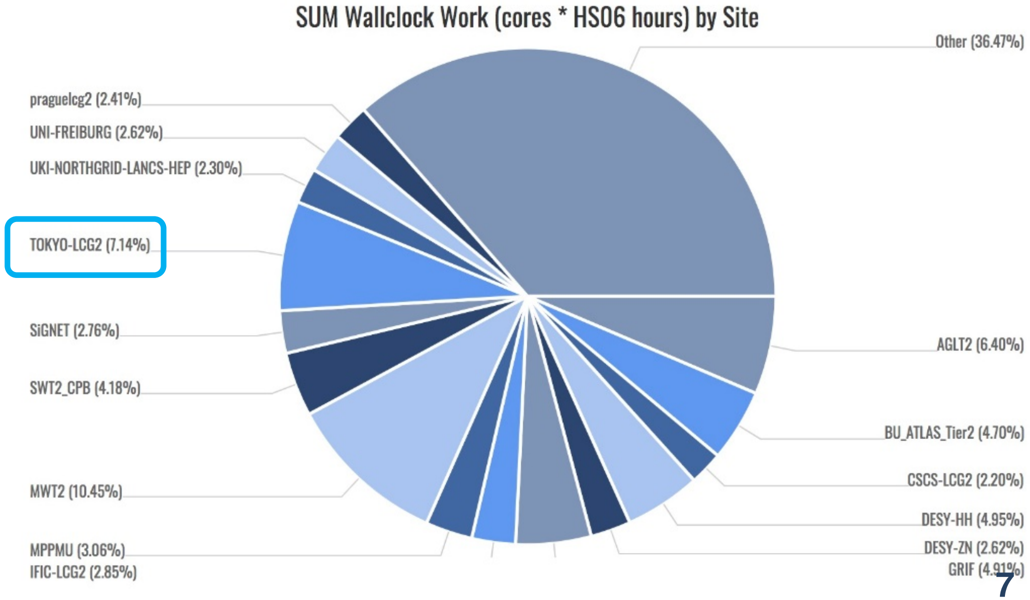
  →Provides local resources to the ATLAS Japan group, too
- All hardware devices are supplied by the three years rental

  → All hardware devices are renewed in three years
- Current system (Starting from Jan/2019):

  →Worker node: 10,752cores (HS06: 18.97/core) (7,680 for WLCG, 145689.6 HS06*cores), 3.0GB/core

  →File server: 15,840TB, (10,560TB for WLCG)

Tier 2 Grid Accounting



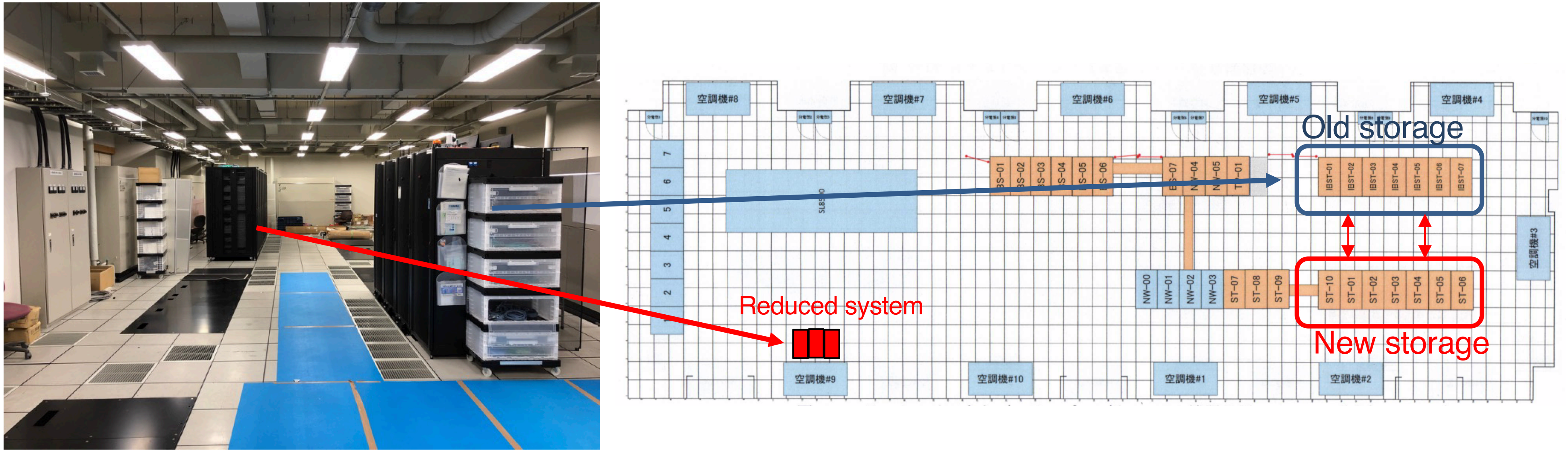
TOKYO-LCG2 provides 7% of Tier 2

Reference number:
ATLAS authors: ~ 3000
ATLAS Japan authors: ~150 (5%)

System migration

- Devices are renewed in Dec/2018

During the migration



- Installation took 10 working days

  →The reduced system worked to minimize the downtime (only 16h)

  → 768 CPU cores

- Data migration took 60 hours for 5.8PB data

  →Connected new-old storages by fiber channel cables

  →Copied data by using cp/rsync at each disk array

  → ~500MB/sec, 97 disk arrays

System migration

- Devices are renewed in Dec/2018

During the migration

After the migration

- Installation took 10 working days

  →The reduced system worked to minimize the downtime (only 16h)

  → 768 CPU cores

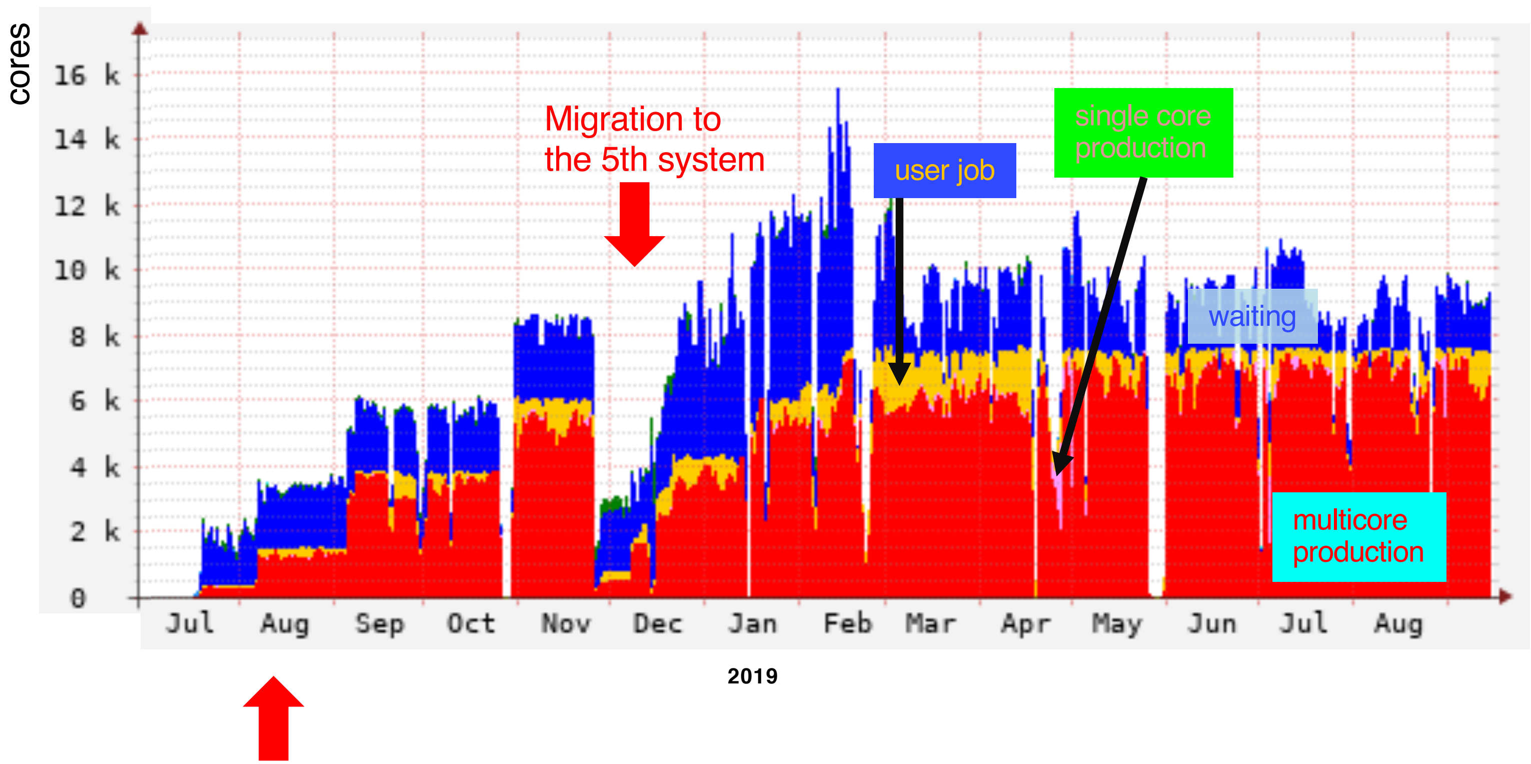
- Data migration took 60 hours for 5.8PB data

  →Connected new-old storages by fiber channel cables

  →Copied data by using cp/rsync at each disk array

  → ~500MB/sec, 97 disk arrays

System migration



Migration to the 5th system

user job

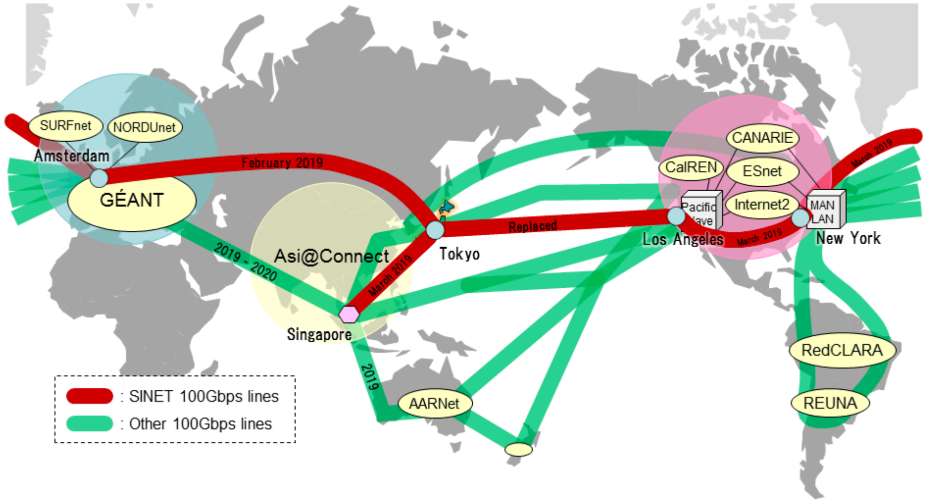
single core production

waiting

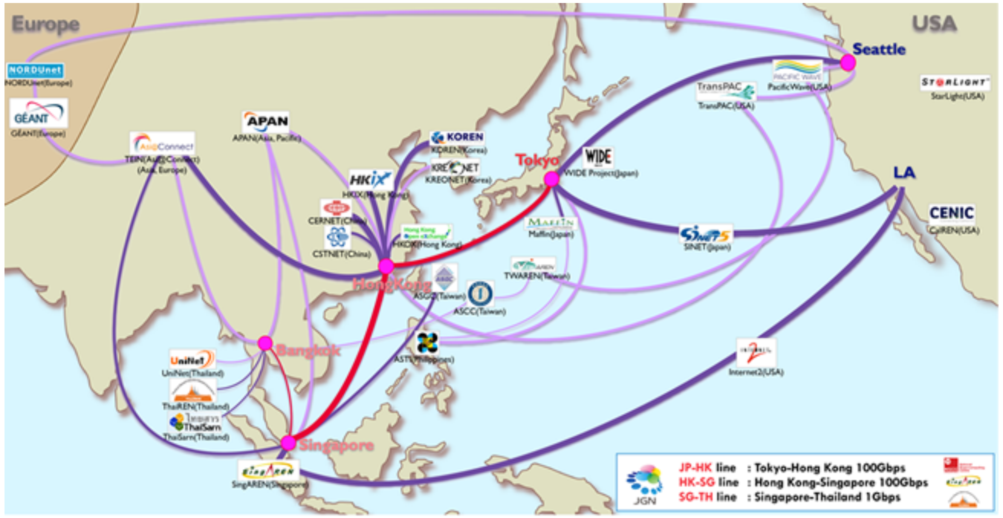
multicore production

OS Migration (From SL6 to CentOS7)

Network



https://www.sinet.ad.jp/en/news_en/2019-03-01news-2



https://testbed.nict.go.jp/jgn/english/networks/index.html

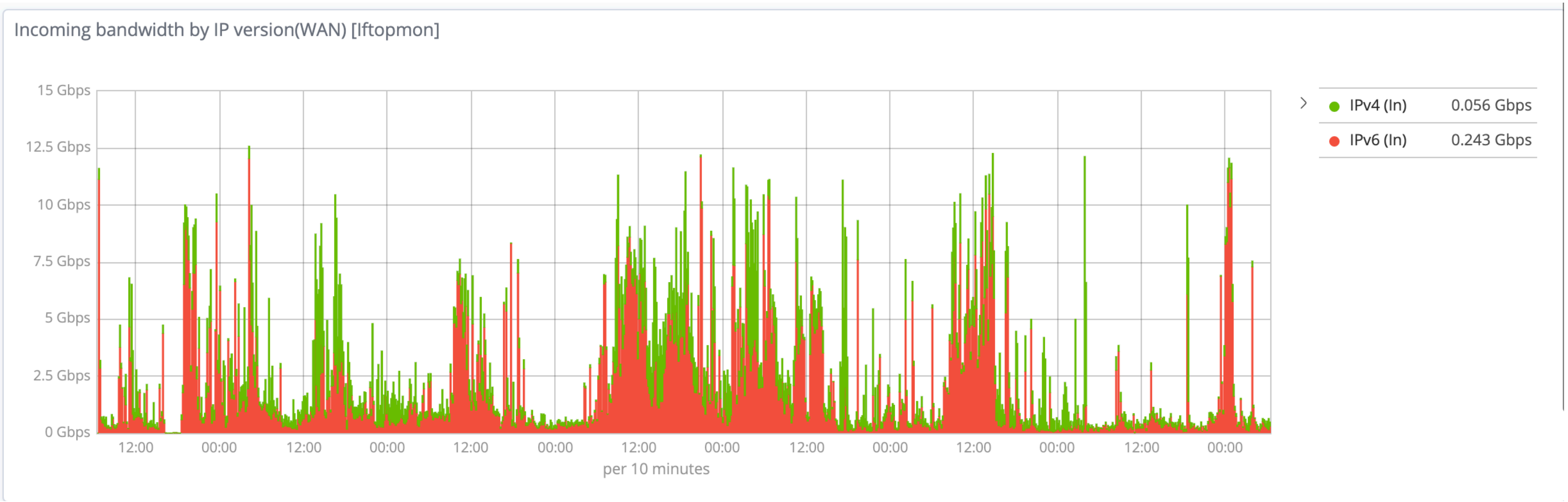
- SINET and JGN (NREN of Japan) made 100Gbps international network connections
- JGN has connection of Tokyo-Hong Kong
- SINET upgraded 100Gbps connections
  →Tokyo-Amsterdam-New York-Los Angeles global ring
  →Connection to Singapore
- Connection of our center to SINET is currently 20Gbps
  →Will be 40Gbps in this weekend

IPv6/IPv4 Dual Stack

- WLCG requires IPv6

  →It becomes difficult to get new IPv4 addresses
- IPv6/IPv4 dual-stack was deployed for the storage element
- Connections to major sites of EU/US by LHCONE are also adapted to IPv6

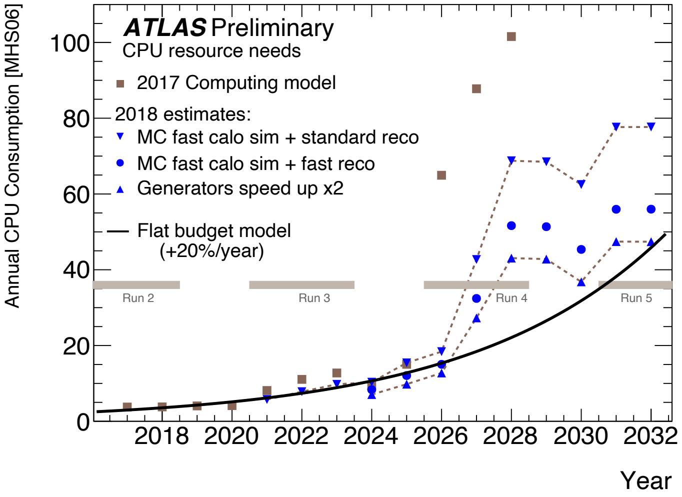
Most of connections are IPv6



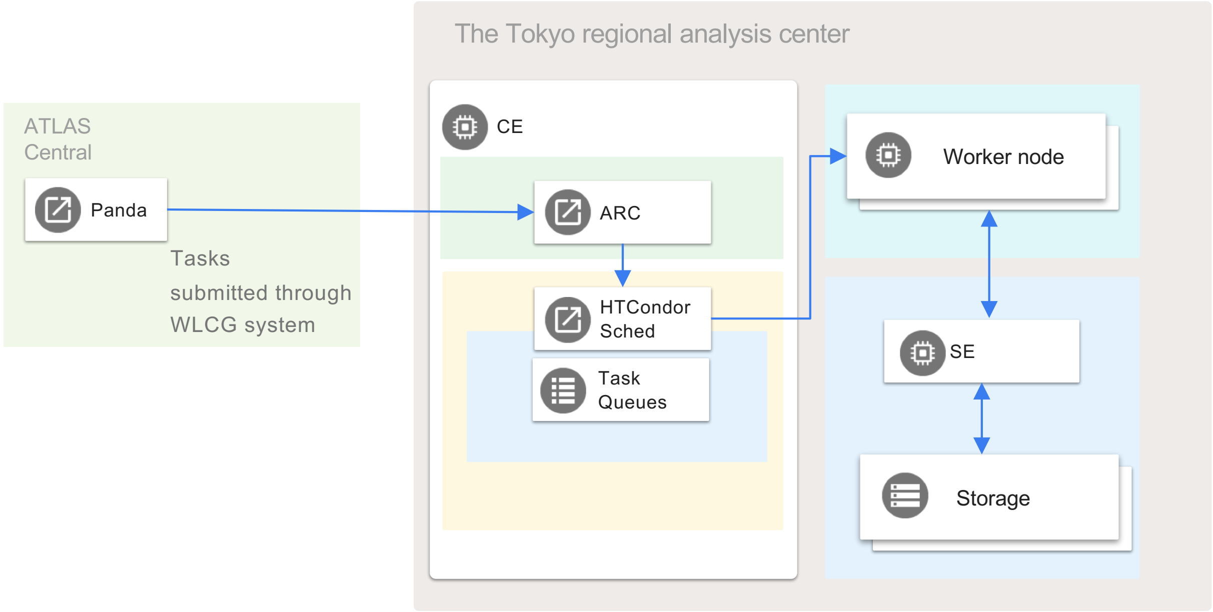
Future Computing Resources

- WLCG have provided enormous computing resources and made it possible to give a lot of results by the LHC experiments
  →But we will need more resources for the future experiments

- CERN plans High-Luminosity LHC in 2026
  →The peak luminosity: x 5
  →The current system cannot provide enough resources with expected budgets
  →More improvements or new ideas are necessary
  →Software update
  →New devices: GPGPU, FPGA, (QC)
  →New grid structure: Data Cloud
  →External resources: HPC, Commercial cloud

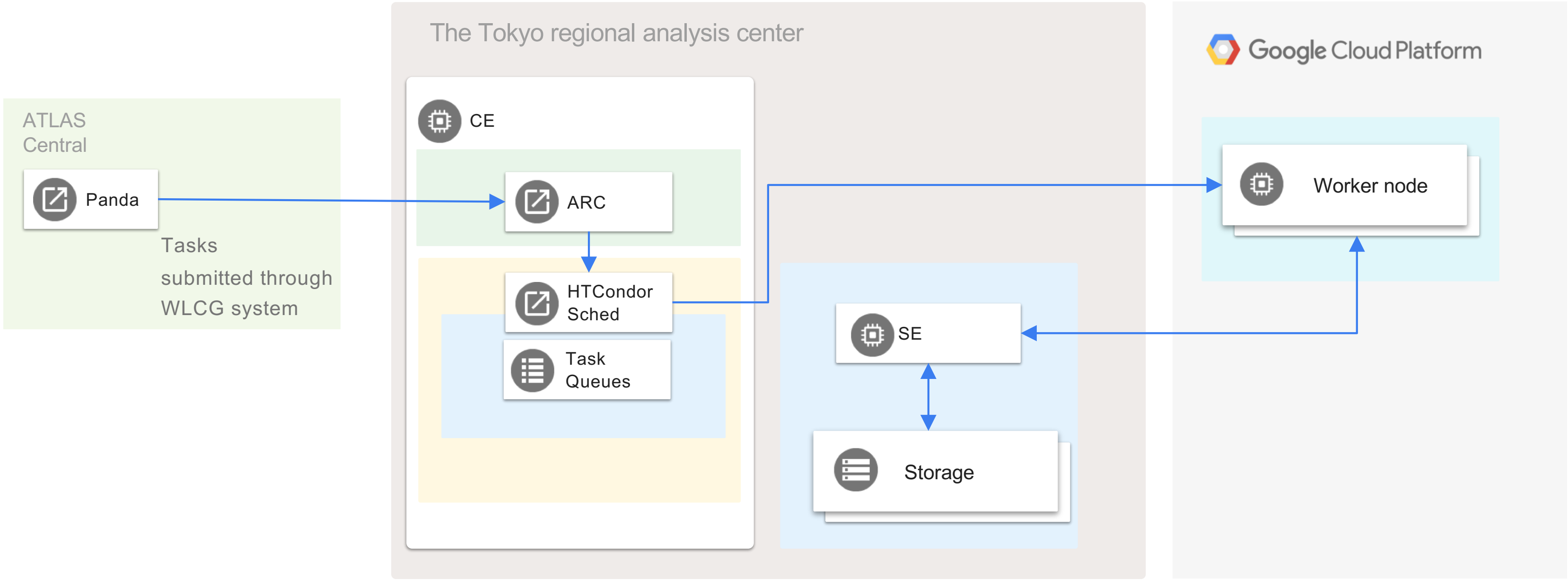


Our Local System


The Tokyo regional analysis center
ATLAS Central
Panda
Tasks submitted through WLCG system
CE
ARC
HTCondor Sched
Task Queues
Worker node
SE
Storage

- Panda: ATLAS job management system, using WLCG framework
- ARC-CE: Grid front-end
- HTCondor: Job scheduler

Hybrid System with Google Cloud Platform



- Cost of storage is high
  →Additional cost to extract data
- Only worker nodes (and some supporting servers) were deployed on cloud, and other services are in on-premises
  →Hybrid system

Google Cloud Platform Condor Pool Manager

- Google Cloud Platform Condor Pool Manager (GCPM)

  → https://github.com/mickaneda/gcpm

  → Can be installed by pip:

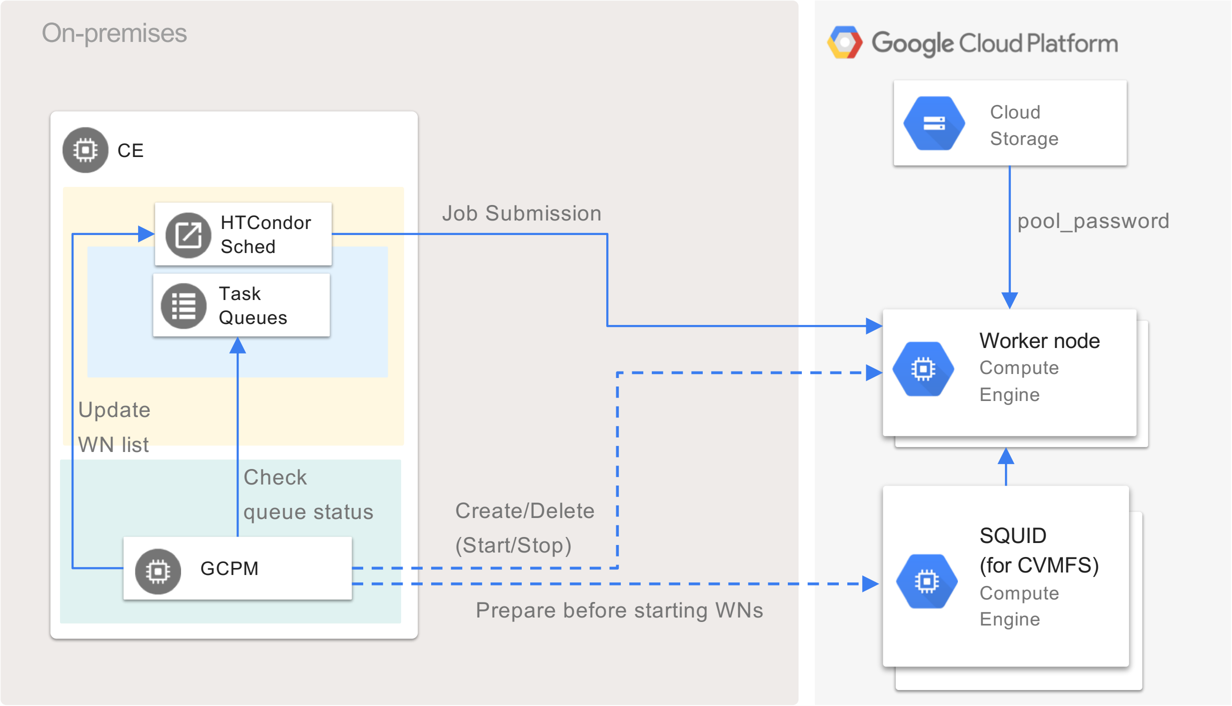
  → ***$ pip install gcpm***

- Manage GCP resources and HTCondor's worker node list

  → On-demand instance preparation

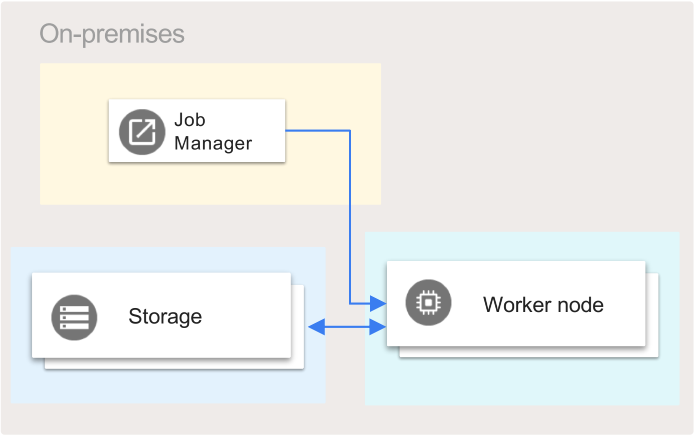
- Can be used for any of HTCondor systems

  → Useful for high-peak needs of CPUs, GPGPU instances, many cores instances, or high-memory instances which are needed once in a while



Cost Estimation

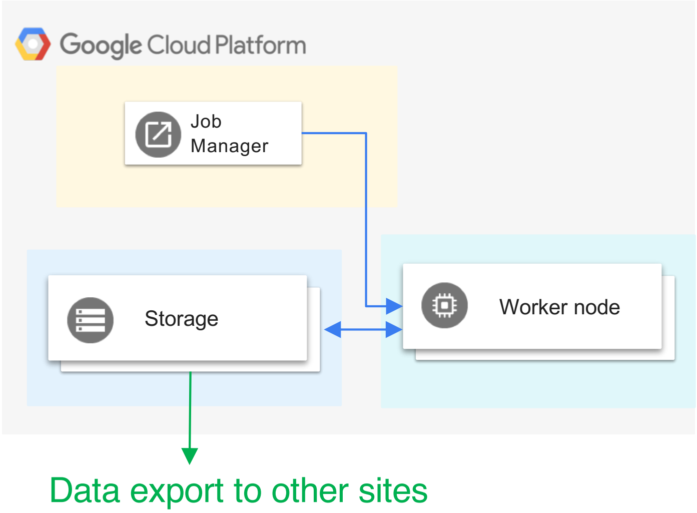
Full on-premises system



- Estimated with Dell machines
- 10k cores, 3GB/core memory, 35GB/core disk: $5M
- 16PB storage: $1M
- Power cost: $20k/month

→ For 3 years usage: ~$200k/month (+Facility/Infrastructure cost, Hardware Maintenance cost, etc…)

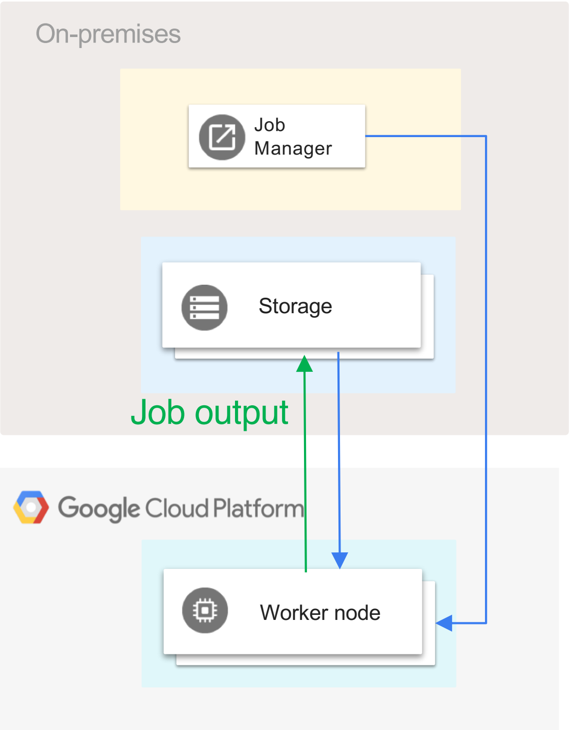
Full cloud system



Resource	Cost/month
vCPU x20k	$130k
3GB x20k	$52k
Local Disk 35GBx20k	$28k
Storage 8PB	$184k
Network Storage to Outside 600 TB	$86k

Total cost: $480k/month

Hybrid System



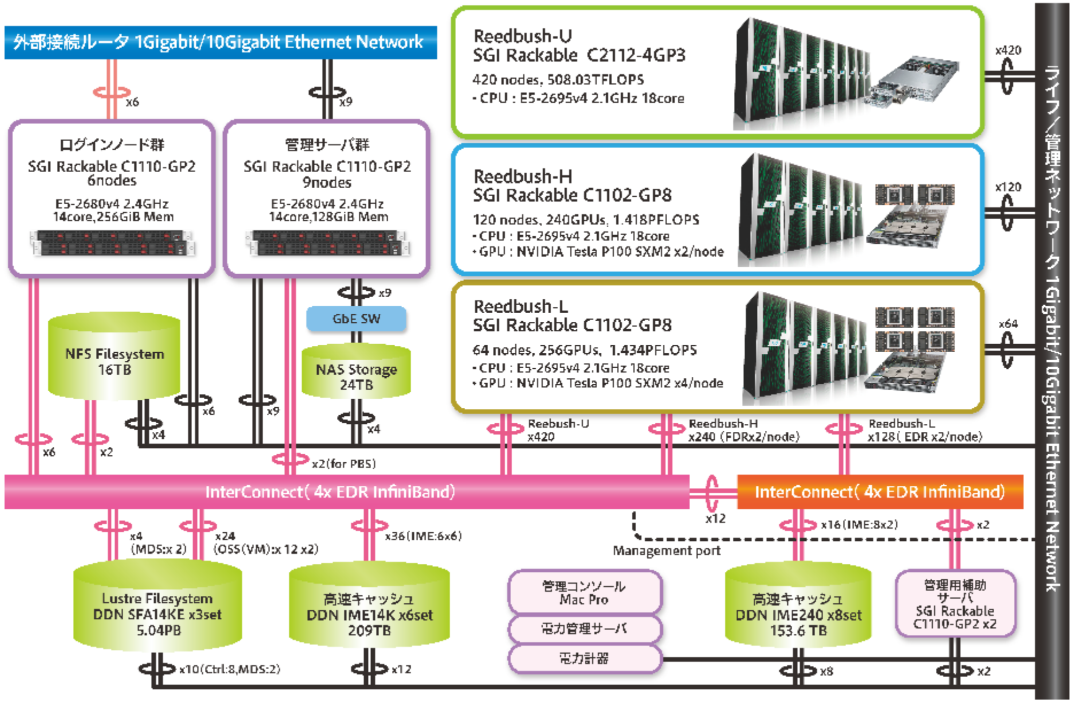
Resource	Cost/month
vCPU x20k	$130k
3GB x20k	$52k
Local Disk 35GBx20k	$28k
Network GCP WN to ICEPP Storage 300 TB	$43k

Total cost: $243k/month
+ on-premises costs
(storage $30k/month + others)

For GCP, use 20k to have comparable spec, as HT on
Use Preemptible Instance

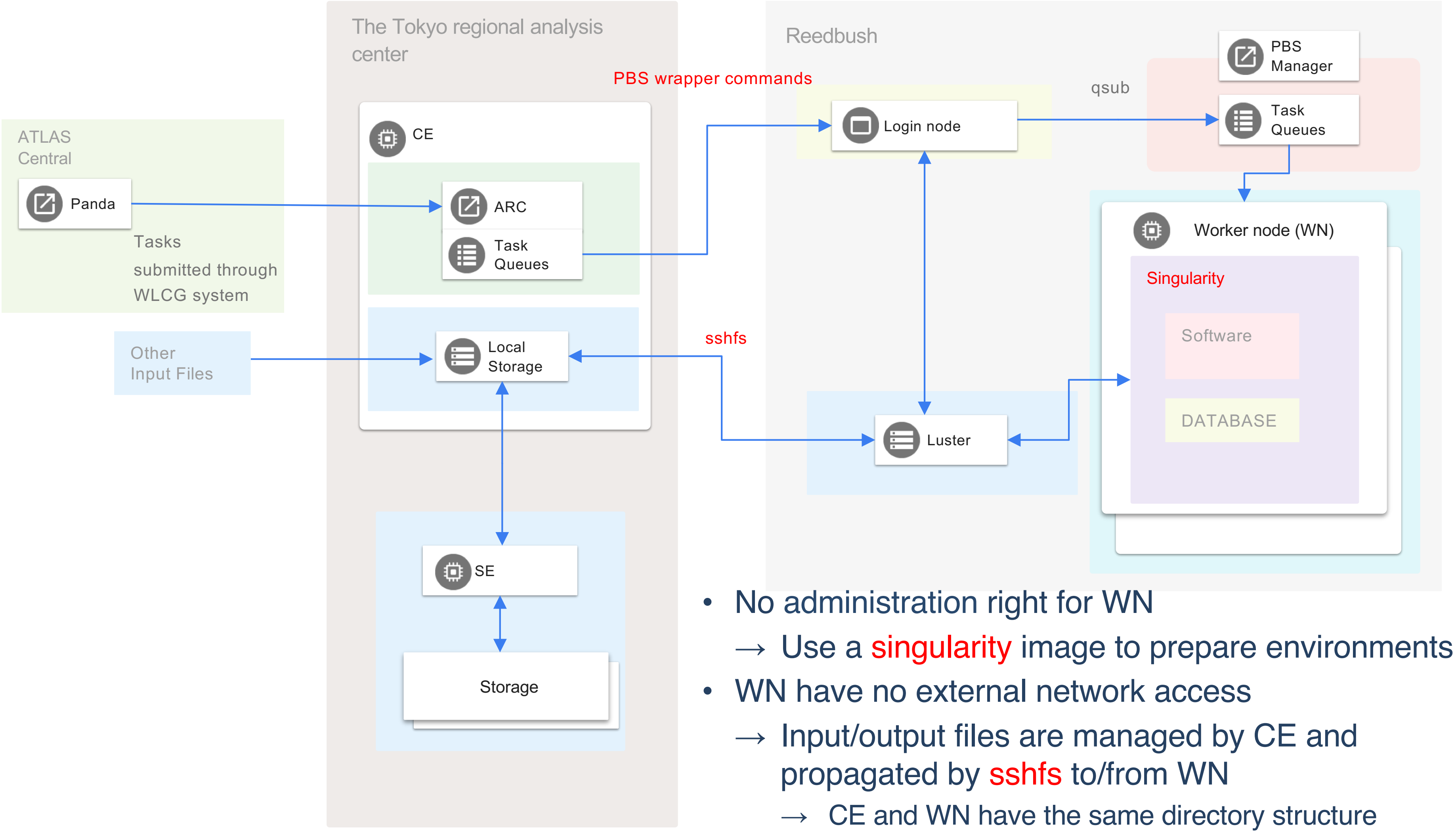
Reedbush: HPC@The Univ. Tokyo

- Supercomputer system @Information Technology Center, The University of Tokyo
  →CPU:Intel Xeon (2CPUs/node (36cores/node))
  →GPU: NVIDIA Tesla P100
- CPU only nodes and GPU nodes
- OS: Read Hat Enterprise Linux 7



- PBS for the job management
- Lustre file system
- No external network access from each WN

System with HPC

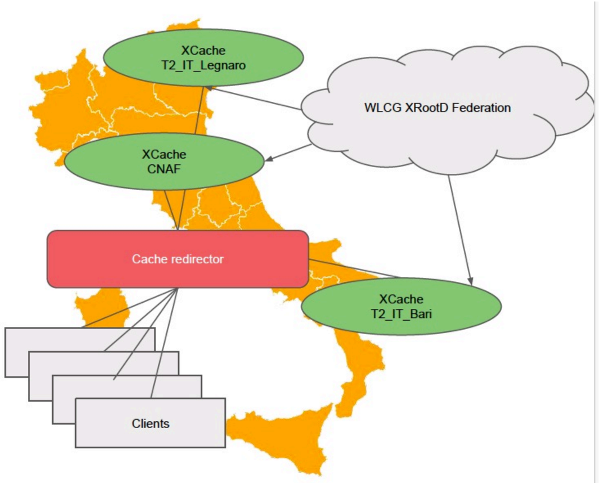


- No administration right for WN
  → Use a singularity image to prepare environments
- WN have no external network access
  → Input/output files are managed by CE and propagated by sshfs to/from WN
    → CE and WN have the same directory structure
- Reedbush uses PBS for the job management
  → Available only on the login node
  → To manage jobs from CE, PBS wrapper commands are used
    → qsub:

```
ssh user@reedbush “cd $work_dir && qsub job.sh”
```

Collaboration in Asia

- Some European countries started to construct “data lake” structures



https://indico.cern.ch/event/769507/

- Italy caching layer prototype for CMS
  →Using Xcache
  →Some storage-less Tier2s

- Data lake of Asia?
  →One of the collaboration ways of Asia
    →But each of us supports different VOs…
  →Network connection between Tokyo to Asia has been improved
    →SINTE will make more connections if we have valuable usage
      →Tokyo – Korea, Tokyo –Taiwan, etc…

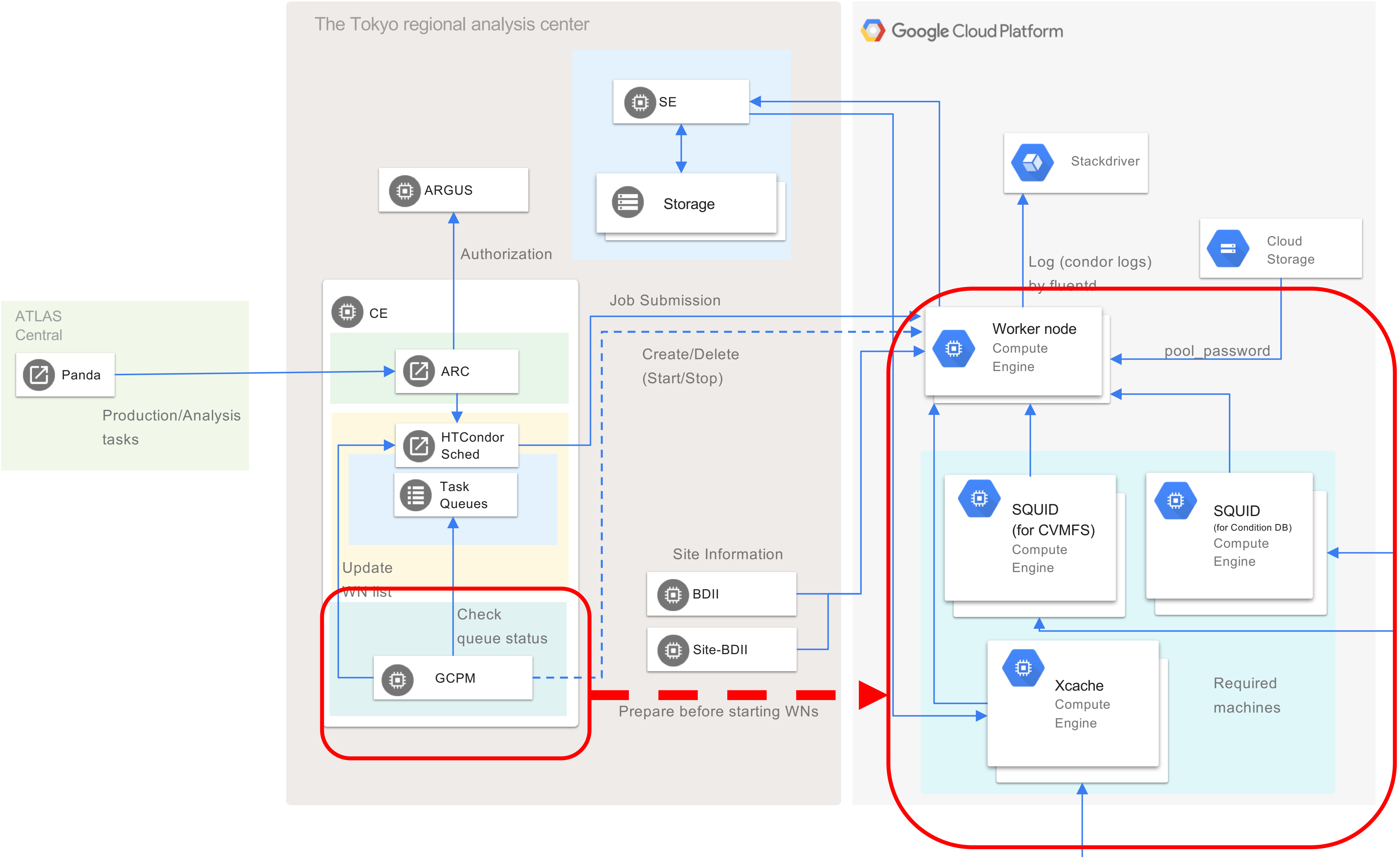
Summary

- The Tokyo regional analysis center:
  →Tier2 of WLCG for ATLAS
  →Renewed to 5th system in Dec/2018
    →Successfully migrated
    →New system has 10k CPU cores and 16PB disk
- SINET established global 100Gbps network
- Some R&D for the future extensions are on going
  →Cloud resources, HPCs

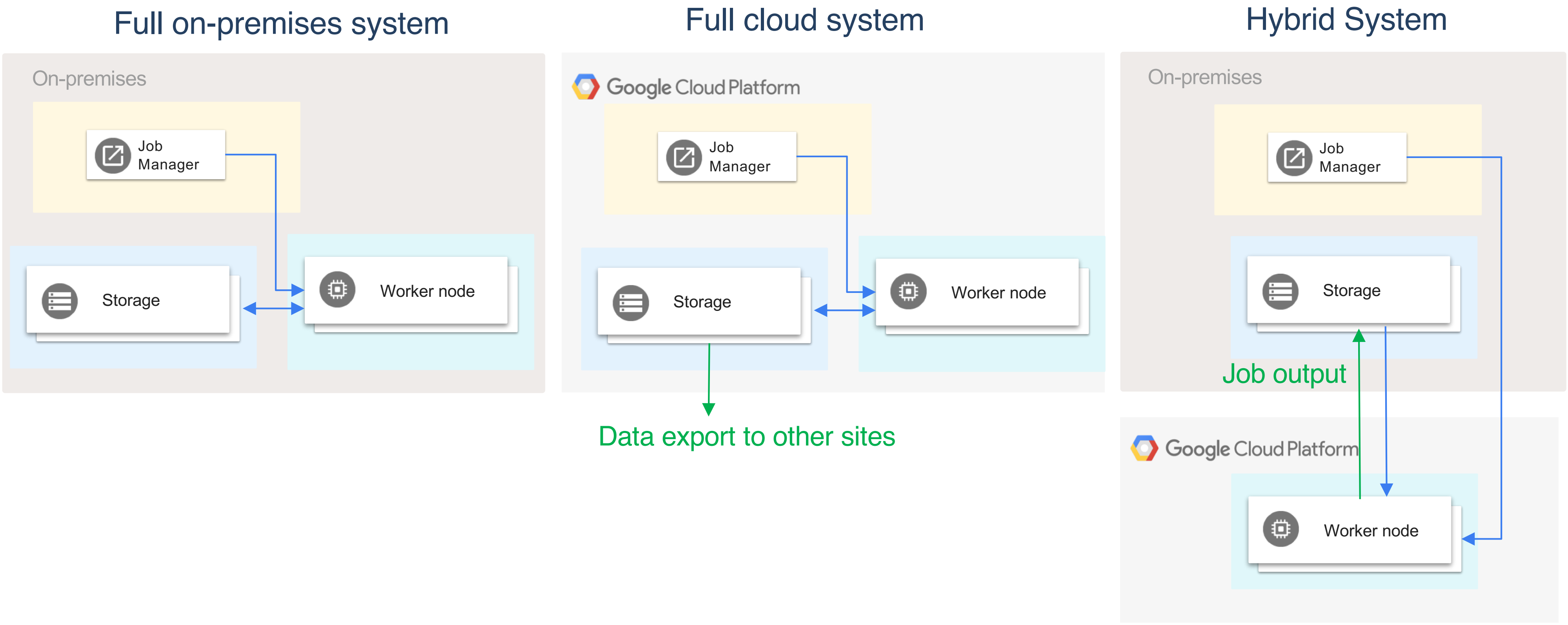
- How can we make a collaboration in Asia?
  →Data lake could be one of the way

Backup

System for R&D



Cost Estimation



Full on-premises system

Full cloud system

Hybrid System

- Estimated with Dell machines
- 10k cores, 3GB/core memory, 35GB/core disk: $5M
- 16PB storage: $1M
- Power cost: $20k/month

  → For 3 years usage: ~$200k/month (+Facility/Infrastructure cost, Hardware Maintenance cost, etc…)

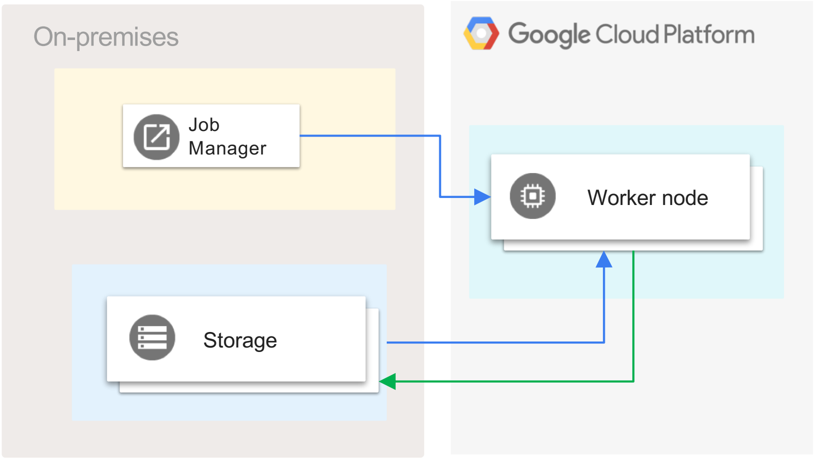
- For GCP, use 20k to have comparable spec

  → Use Preemptible Instance (Hyperthreading On, half)

- 8PB storage which is used at ICEPP for now
- Cost to export data from GCP

https://cloud.google.com/compute/pricing
https://cloud.google.com/storage/pricing

1 Day Real Cost



Hybrid system: 1k cores, 2.4GB/core memory

→ Cost for month (x30), with 20k cores (x20): ~$240k + on-premises costs

1 Day Real Cost (13/Feb)

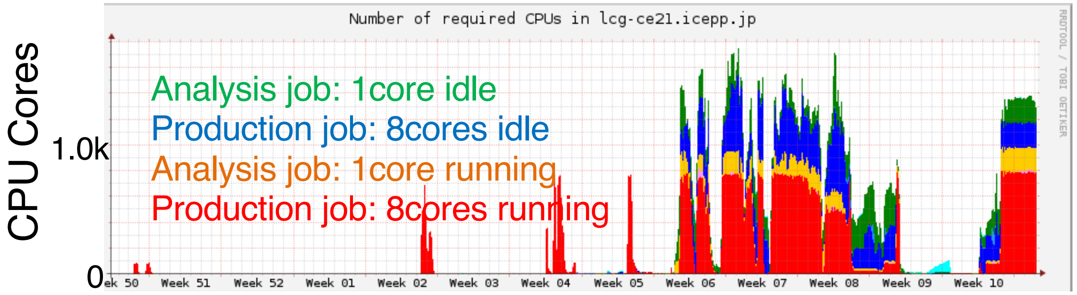
	Usage	Cost/day	x30x20
vCPU (vCPU*hours)	20046	$177	$106k
Memory (GB*hours)	47581	$56	$34k
Disk (GB*hours)	644898	$50	$30k
Network (GB)	559	$78	$47k
Other services		$30	$18k
Total		$391	$236k

vCPU: 1vCPU instances max 200, 8 vCPUs instances max 100
Memory: 2.4 GB/vCPU
Disk: 50GB for 1vCPU instance, 150 GB for 8 vCPUs instance

Cost Estimation

Resource	Cost/month
vCPU x20k	$130k
3GB x20k	$42k
Local Disk 35GBx20k	$28k
Network GCP WN to ICEPP Storage 300 TB	$43k
Total	$243k

ATLAS jobs on GCP and Reedbush

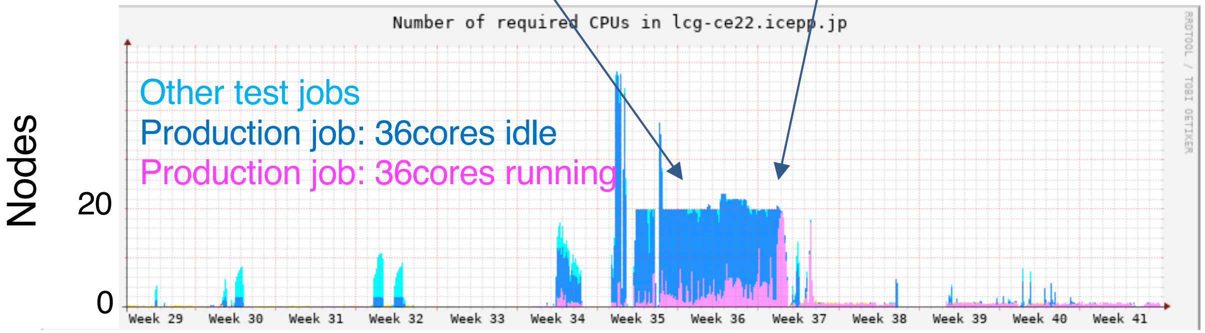


HTCondor status monitor for GCP

Max CPU Cores = 1k

Tested with the small queue
(Only a few nodes are available)

Tested with the large queue



PBS status monitor for Reedbush

Max nodes = 20 (=720 CPU cores)

Performance Comparison

System	Hyper Threading	Core(vCPU)	Memory	CPU	HEPSPEC/ core	ATLAS simulation 1000events (hours)	Walltime*cores/Events
ICEPP local system	Off	32	96GiB	Intel(R) Xeon(R) Gold 6130 CPU @ 2.10GHz	18.97	(8core job) 5.19	0.042
Google Cloud Platform	On	8	24GiB	Intel(R) Xeon(R) Gold 6138 CPU @ 2.00GHz	12.62	(8core job) 9.27	0.074
Reedbush	Off	36	256GB	Intel(R) Xeon(R) CPU E5-2695 v4 @ 2.10GHz	16.78	(36 core job) 1.1	0.040

HEPSPEC (06): Benchmark for HEP

- The ATLAS production jobs can run with multi-processing mode

  → Normally 8 cores are used at WLCG sites

  → Will be multi-threading
- All GCP’s instances are set as hyper-threading on

  → ~half performance of other systems
- Reedbush nodes have 36 cores

  → Each job occupies all cores in the node: Run 36 processes mode

Cost Comparison

System	Cost for 10k cores/Month
On-premises	$200k
Reedbush	$40k
Google Cloud Platform	$250k

- On-premises:
  - → Total server cost of 10k CPU cores, 16PB storage (Dell)/3 years
  - → Additional cost: infrastructure, maintenance
- Reedbush:
  - → Non-university groups also can apply to use the system (price: x1.2)
  - → Only limited number of resources
    - → Currently max number of nodes is ~ 20 (~700cores)
  - → Additional cost: on-premises storage and other service components
- GCP:
  - → Hyper Threading On: Need double number of CPU cores (calculated by assuming 20k cores)
  - → Reduced cost by using preemptible instances
  - → Including network cost
  - → Additional cost: on-premises storage and other service components