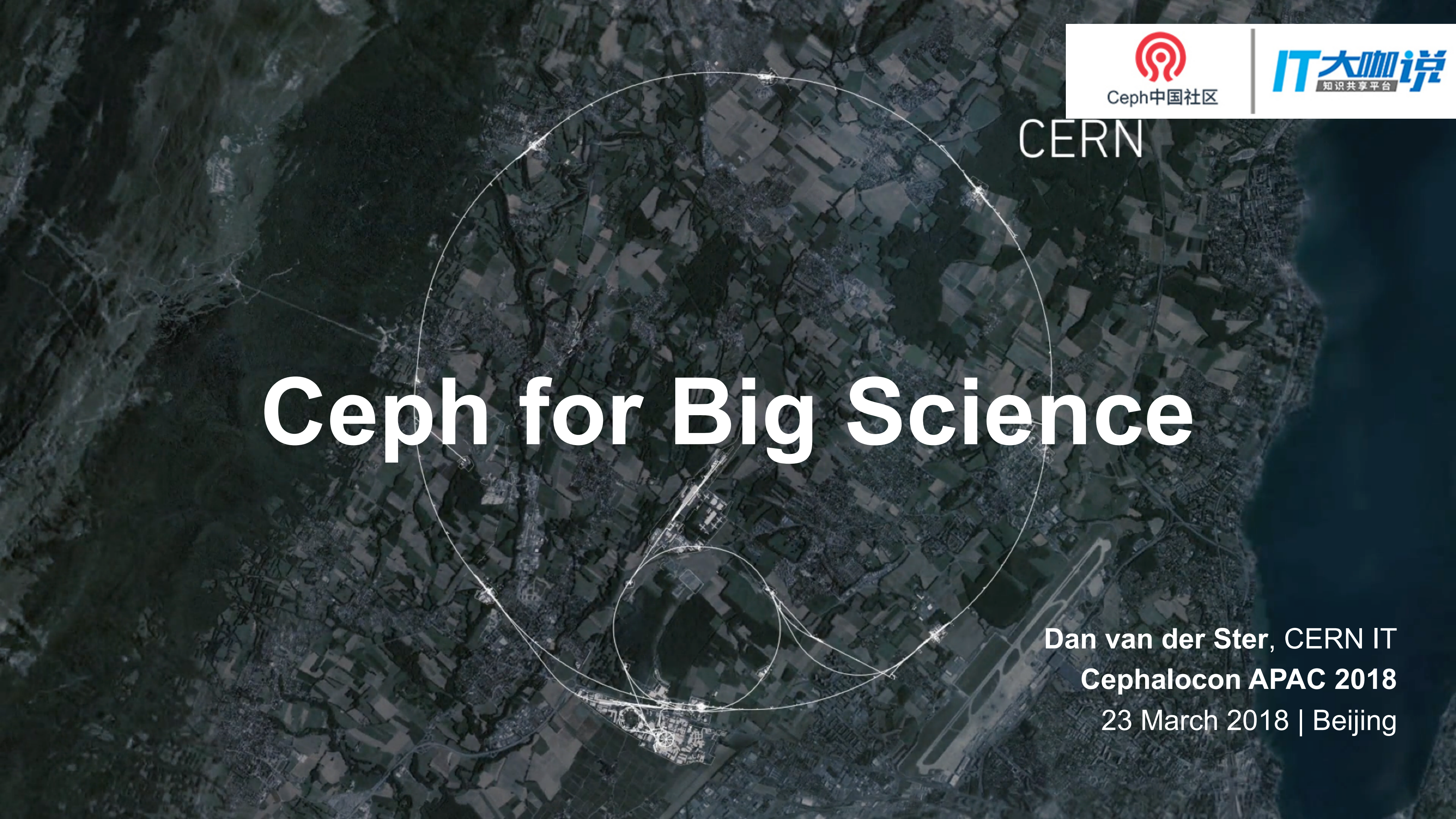
CERN

Ceph中国社区 | IT大咖说 知识共享平台



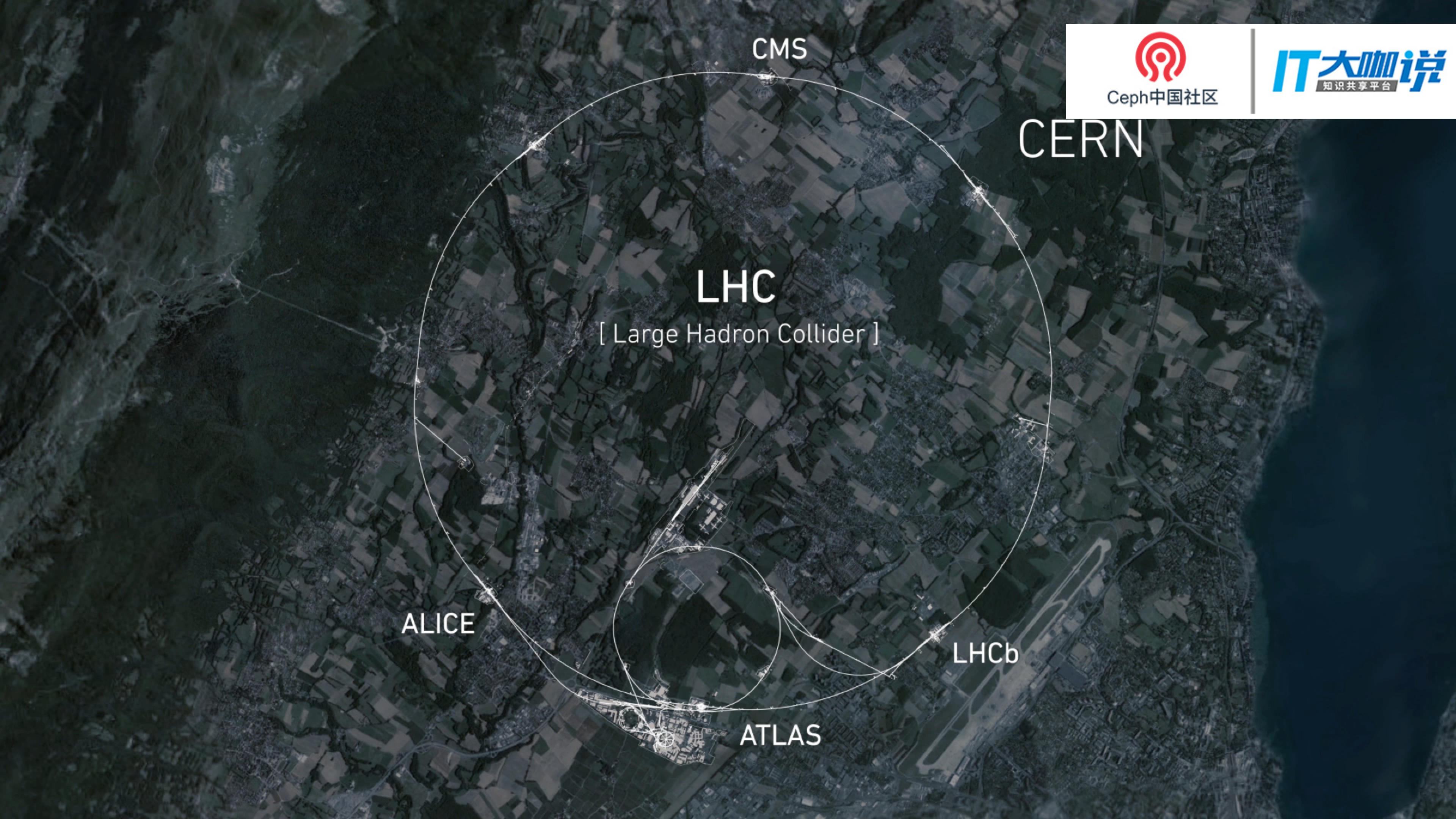
CERN

Ceph for Big Science

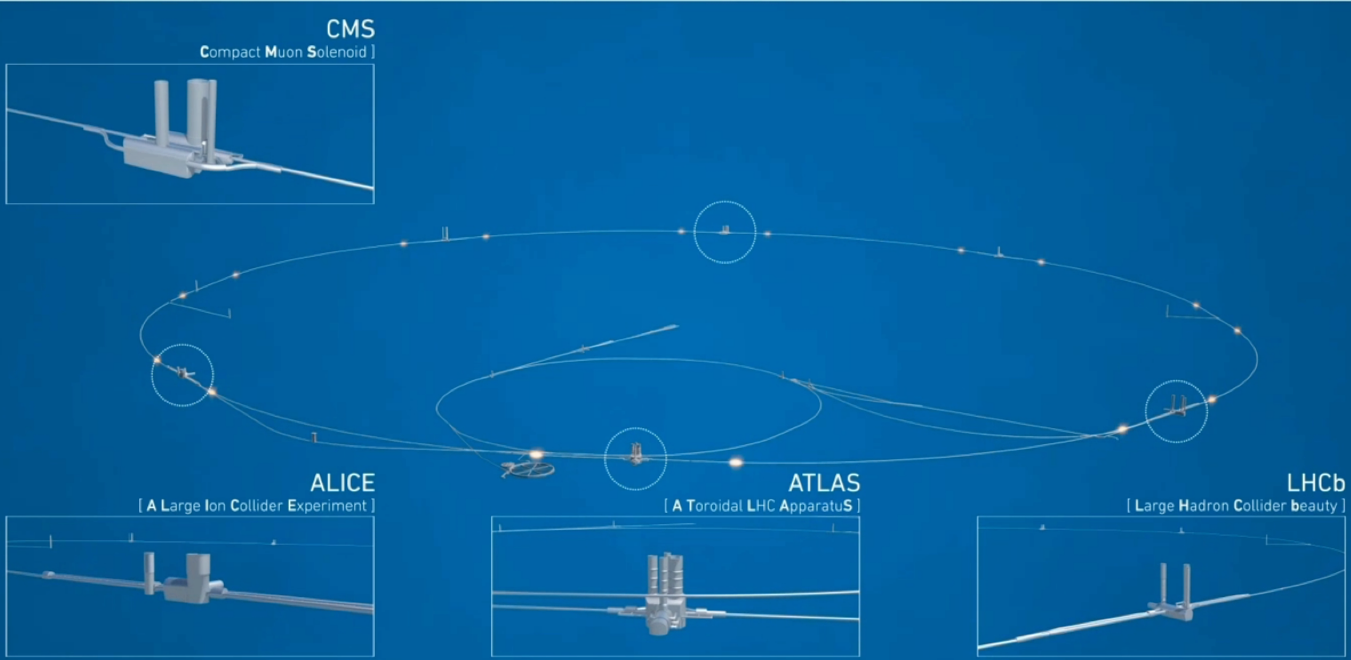
Dan van der Ster, CERN IT

Cephalocon APAC 2018

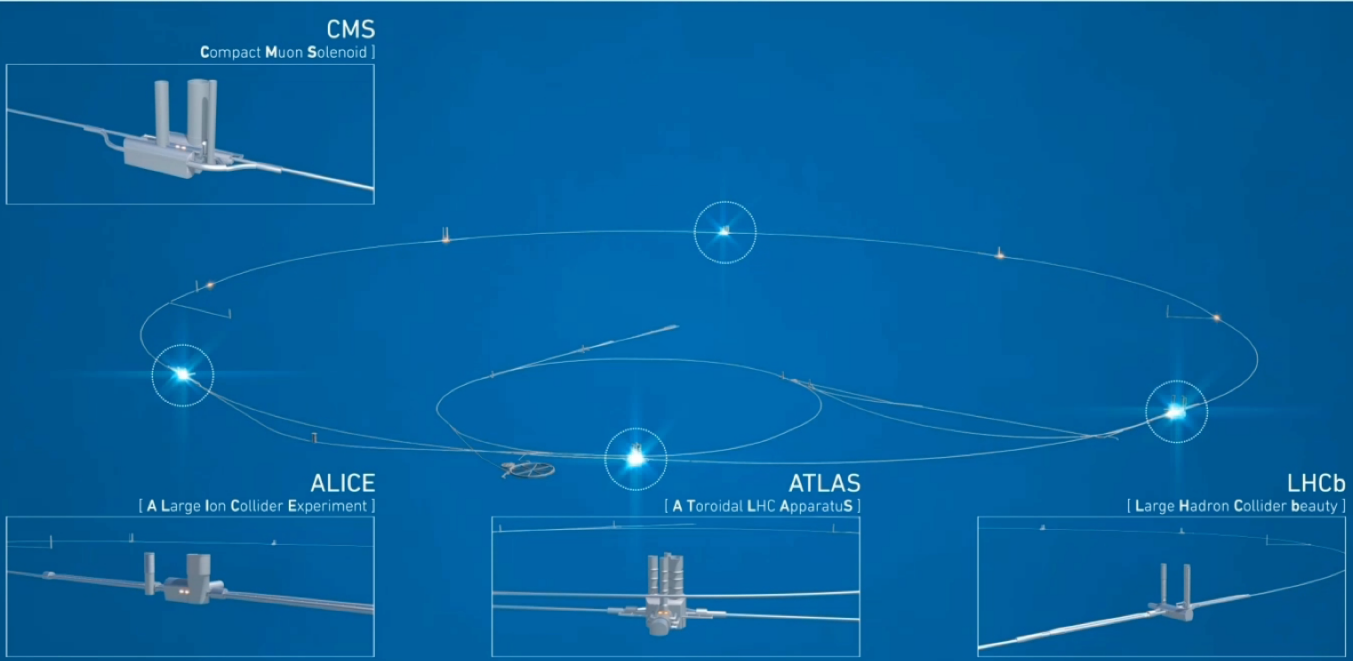
23 March 2018 | Beijing


CMS
CERN
LHC
[Large Hadron Collider]
ALICE
LHCb
ATLAS

27 km circumference

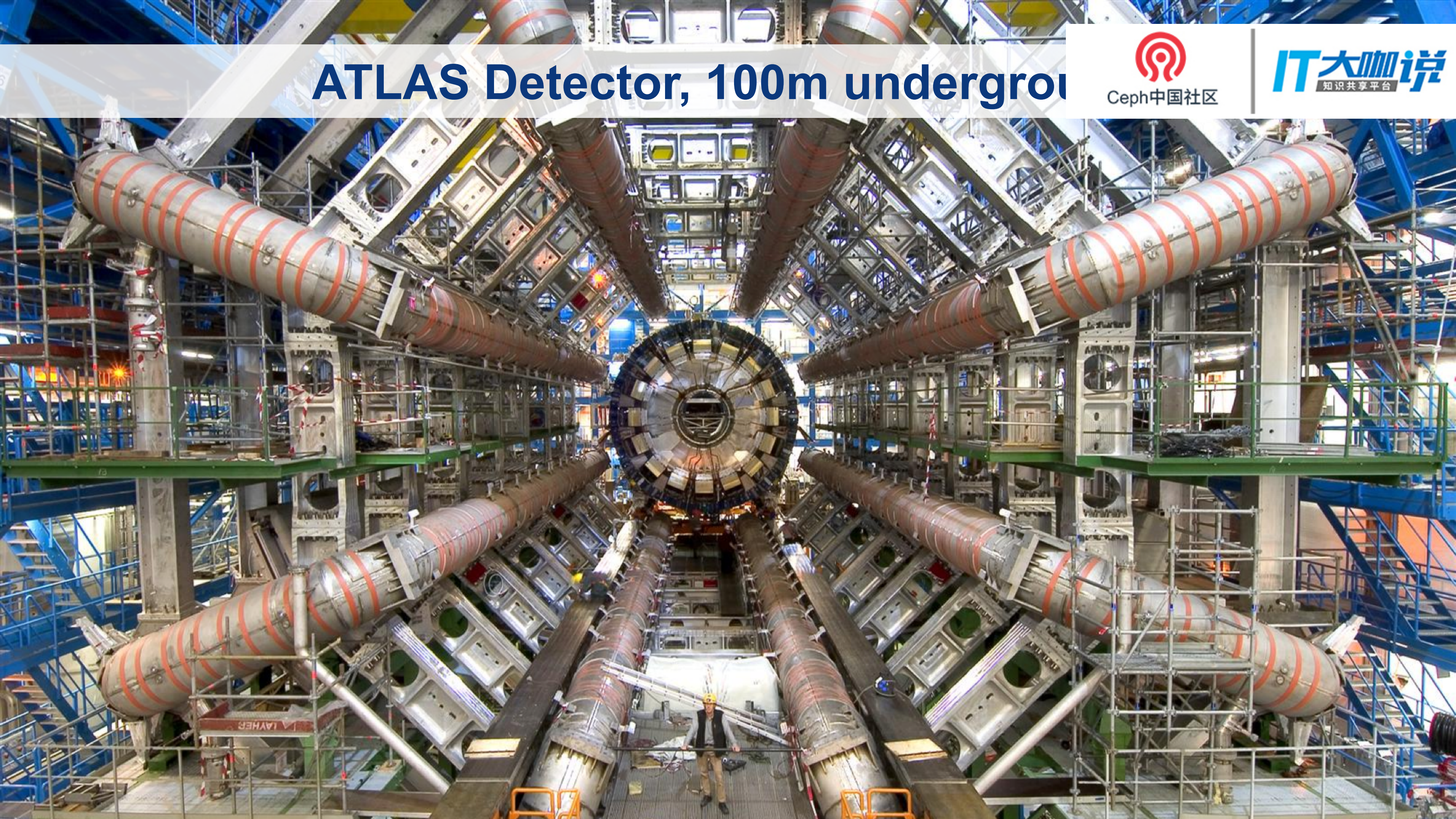

CMS
Compact Muon Solenoid]
ALICE
[A Large Ion Collider Experiment]
ATLAS
[A Toroidal LHC ApparatuS]
LHCb
[Large Hadron Collider beauty]

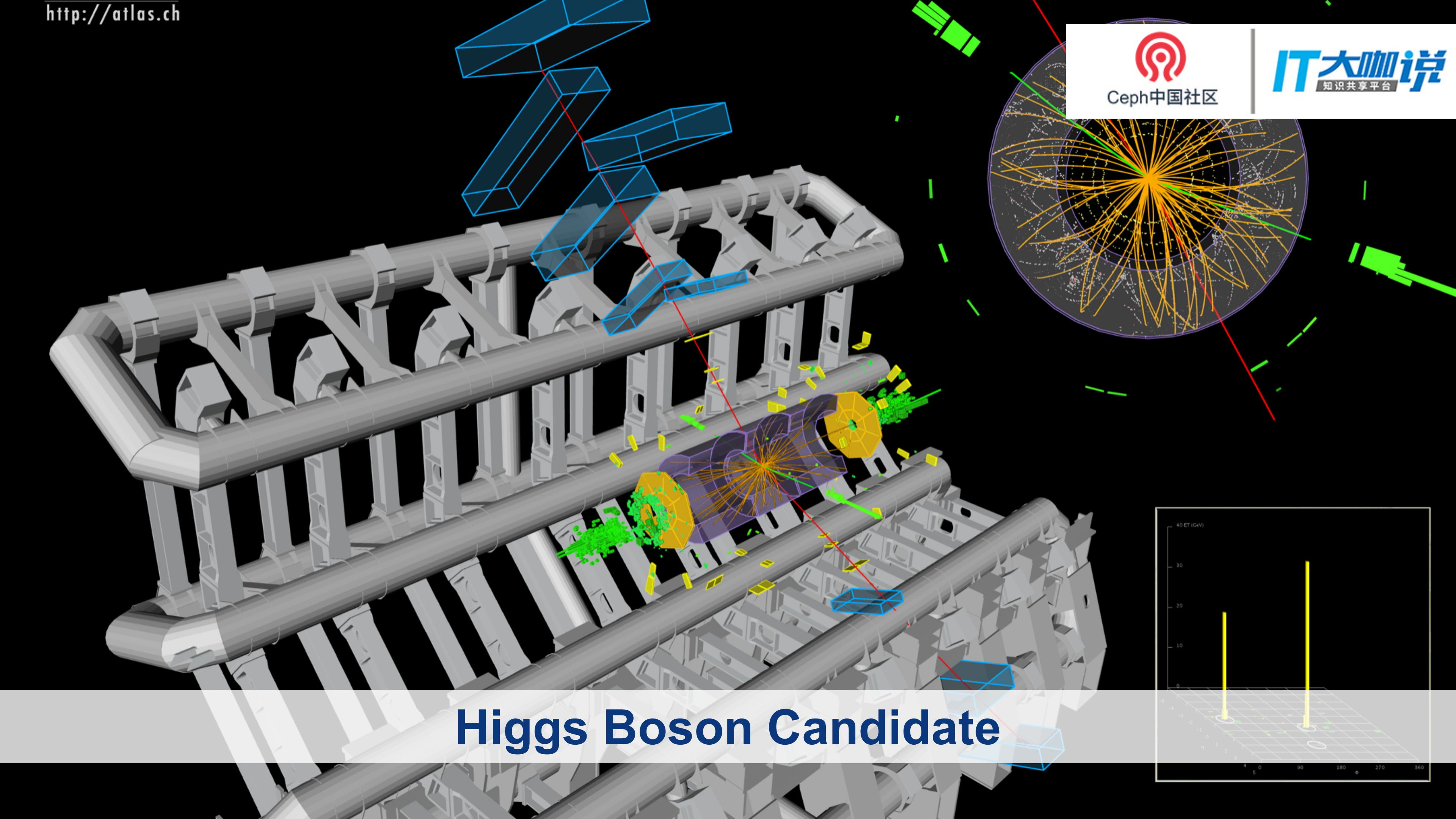
~30MHz interactions filtered to ~1kHz recorded collisions


CMS
[Compact Muon Solenoid]
ALICE
[A Large Ion Collider Experiment]
ATLAS
[A Toroidal LHC ApparatuS]
LHCb
[Large Hadron Collider beauty]

~30MHz interactions filtered to ~1kHz recorded collisions

ATLAS Detector, 100m undergrou



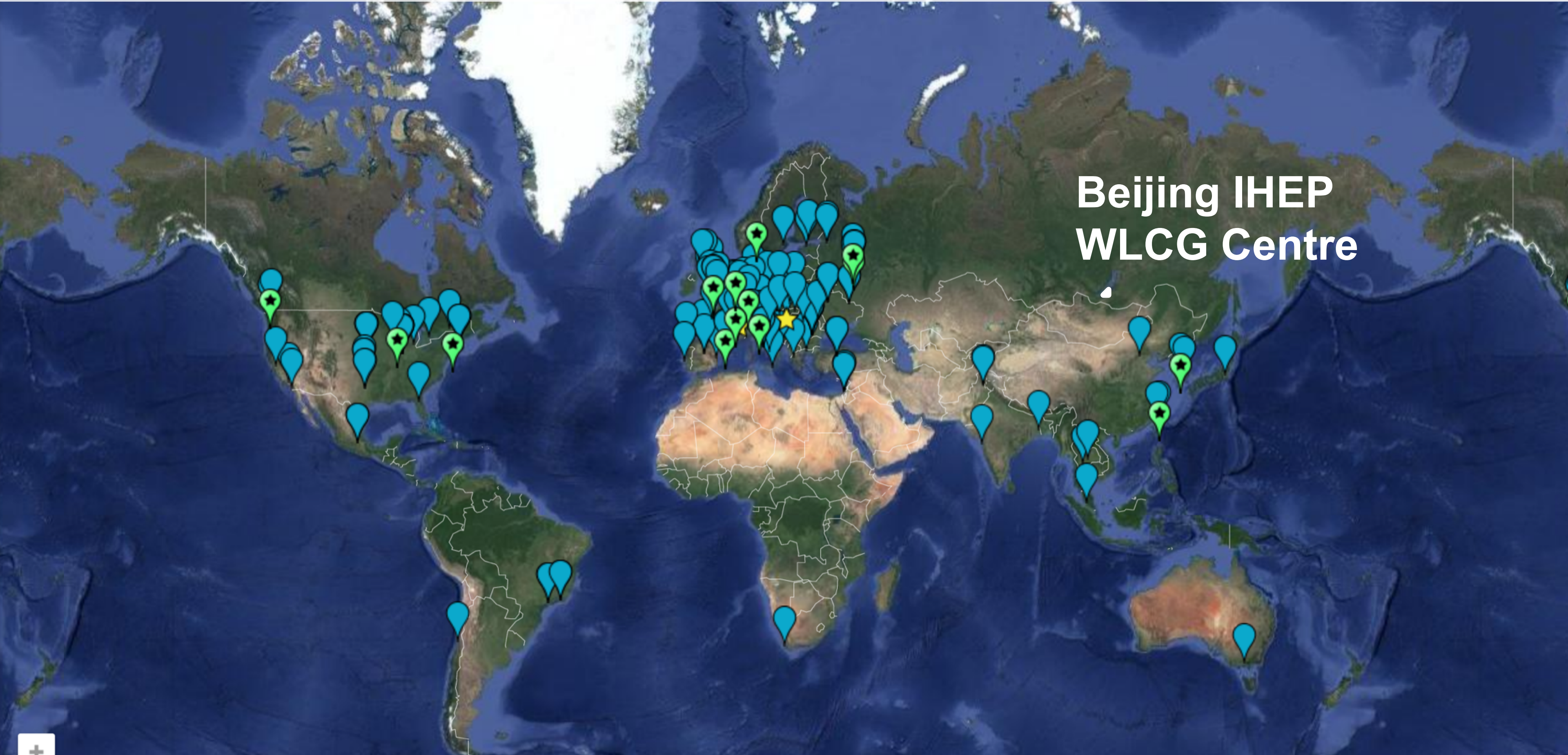


Higgs Boson Candidate

300 petabytes storage, 230 000 CPU

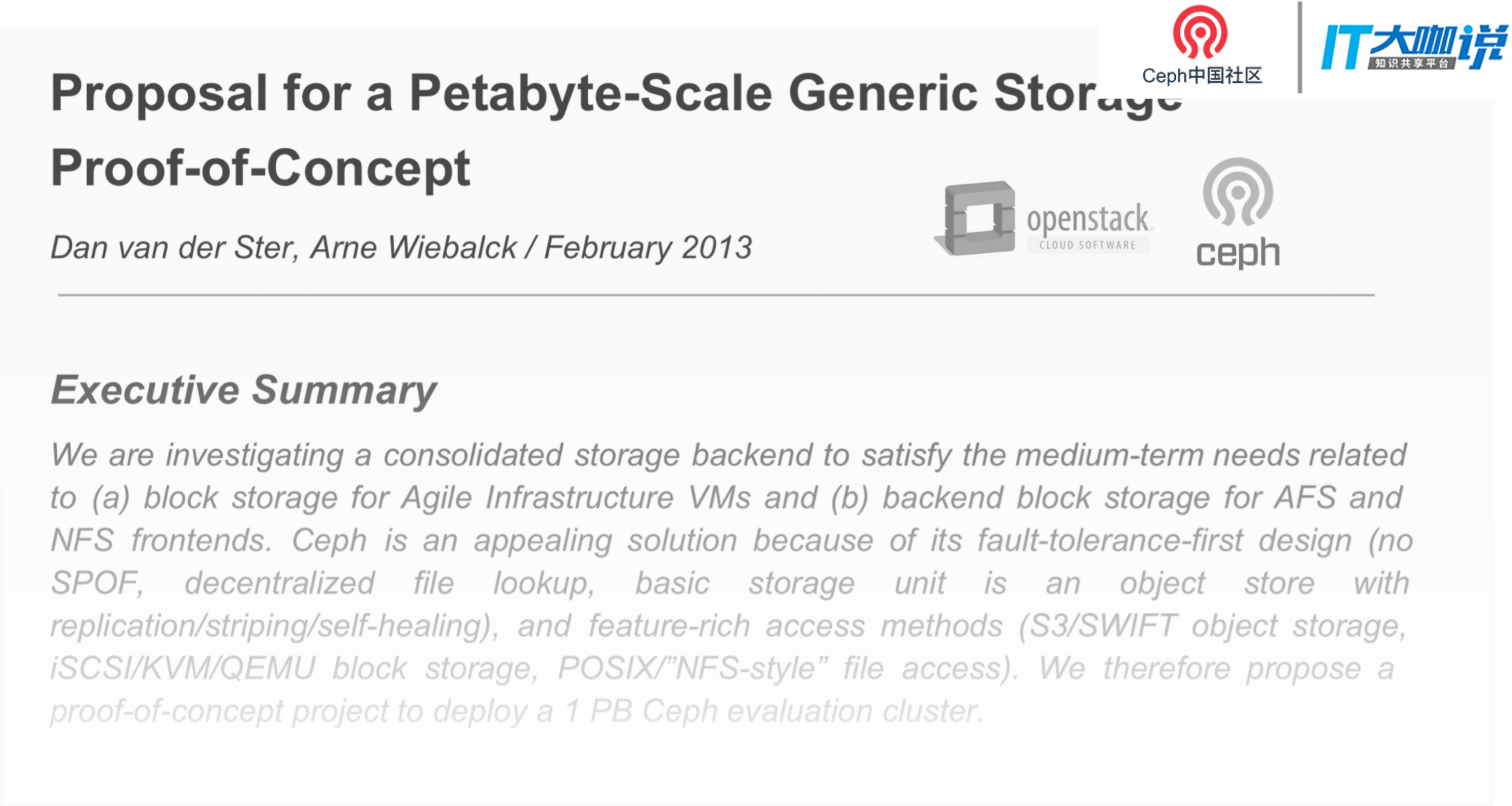
Worldwide LHC Compute Grid

Worldwide LHC Compute Grid



Beijing IHEP
WLCG Centre

Ceph at CERN: Yesterday & Today



Proposal for a Petabyte-Scale Generic Storage Proof-of-Concept

Dan van der Ster, Arne Wiebalck / February 2013

openstack CLOUD SOFTWARE

ceph

Executive Summary

We are investigating a consolidated storage backend to satisfy the medium-term needs related to (a) block storage for Agile Infrastructure VMs and (b) backend block storage for AFS and NFS frontends. Ceph is an appealing solution because of its fault-tolerance-first design (no SPOF, decentralized file lookup, basic storage unit is an object store with replication/striping/self-healing), and feature-rich access methods (S3/SWIFT object storage, iSCSI/KVM/QEMU block storage, POSIX/"NFS-style" file access). We therefore propose a proof-of-concept project to deploy a 1 PB Ceph evaluation cluster.

First production cluster built mid to late 2013 for OpenStack Cinder block storage.

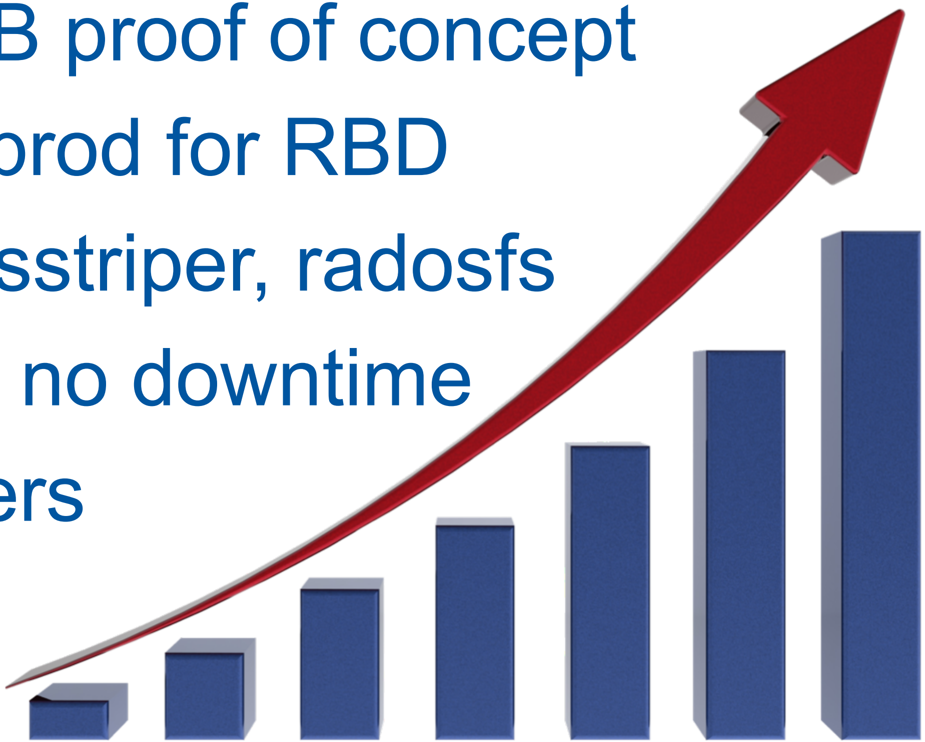
3 PB, 48x24x3TB drives, 200 journaling SSDs

Ceph *dumpling* v0.67 on Scientific Linux 6

We were very cautious: 4 replicas! (now 3)

History

- March 2013: 300TB proof of concept
- Dec 2013: 3PB in prod for RBD
- 2014-15: EC, radosstriper, radosfs
- 2016: 3PB to 6PB, no downtime
- 2017: 8 prod clusters

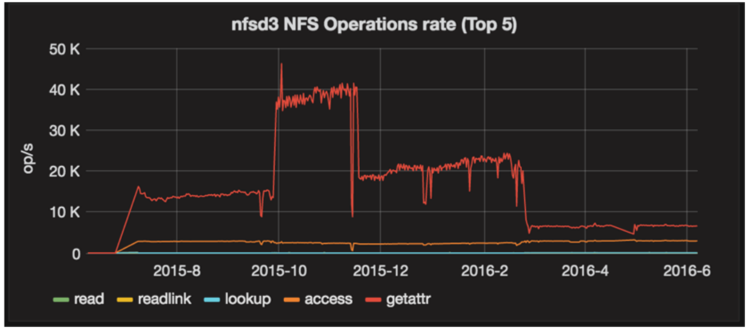


CERN Ceph Clusters		Size	version
OpenStack Cinder/Glance	*Production*	5.5PB	jewel
	Satellite data centre (1000km away)	0.4PB	luminous
CephFS (HPC+Manila)	*Production*	0.8PB	luminous
	Manila testing cluster	0.4PB	luminous
	Hyperconverged HPC	0.4PB	luminous
CASTOR/XRootD	*Production*	4.2PB	luminous
	CERN Tape Archive	0.8PB	luminous
S3+SWIFT	*Production*	0.9PB	luminous

CephFS

CephFS: Filer Evolution

- Virtual NFS filers are stable and perform well:
  - nfsd, ZFS, zrep, OpenStack VMs, Cinder/RBD
  - We have ~60TB on ~30 servers
- High performance, but not scalable:
  - Quota management tedious
  - Labour-intensive to create new filers
  - Can't scale performance horizontally



puppet labs

GitLab

collaborate with TWiki

Platform Computing an IBM Company

BOINC

OPENSHIFT

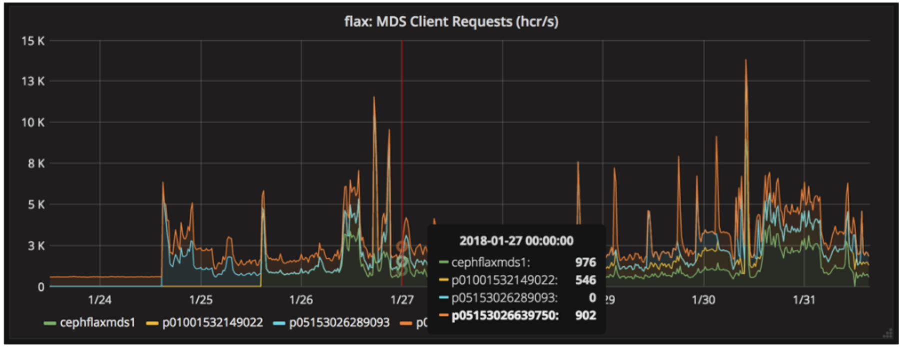
elastic

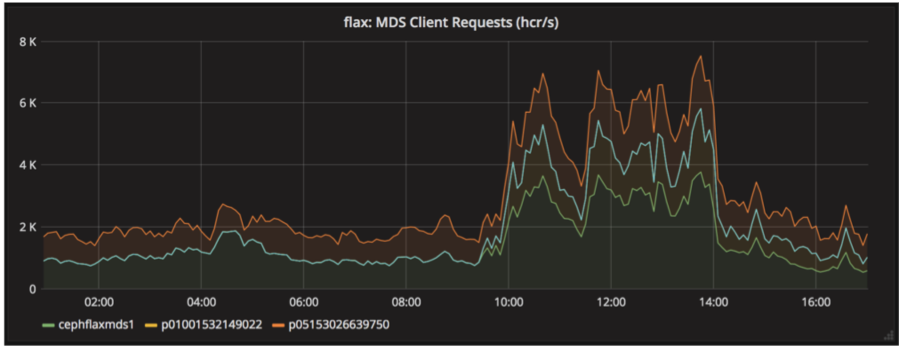
CephFS: Filer Evolution

- OpenStack Manila (with CephFS) has most of the needed features:
  - Multi-tenant with security isolation + quotas
  - Easy self-service share provisioning
  - Scalable performance (add more MDSs or OSDs as needed)
- Successful testing with preproduction users since mid-2017.
  - Single MDS was seen as a bottleneck. Luminous has stable multi-MDS.
- Manila + CephFS now in production:
  - One user already asked for 2000 shares
  - Also using for Kubernetes: we are working on a new CSI CephFS plugin
    - Really need kernel quota support!

Multi-MDS in Production

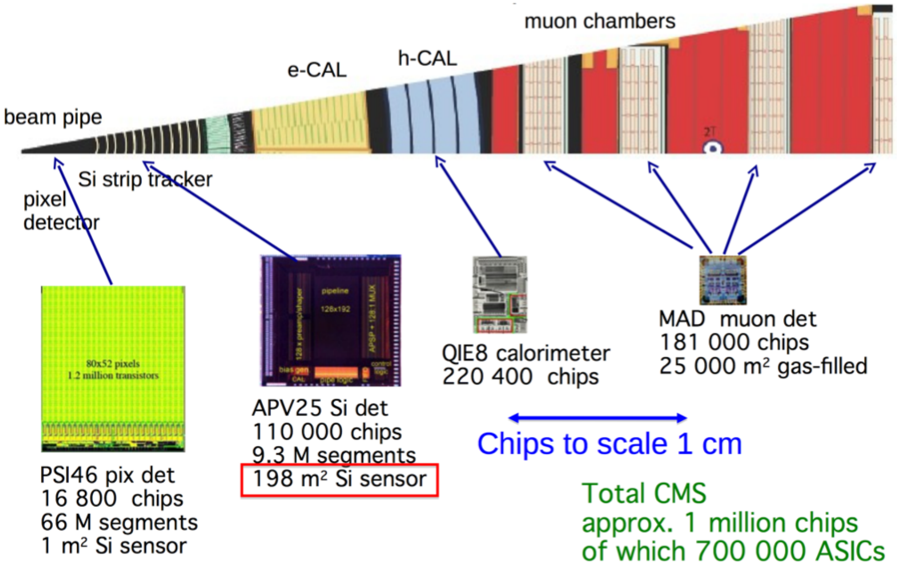
- ~20 tenants on our pre-prod environment for several months
  - 2 active MDSs since luminous
- Enabled multi-MDS on our production cluster on Jan 24
- Currently have 3 active MDSs
  - default balancer and pinning





HPC on CephFS?

- CERN is mostly a high *throughput* computing lab:
  - File-based parallelism
- Several smaller HPC use-cases exist within our lab:
  - Beams, plasma, CFD, QCD, ASICs
  - Need full POSIX, consistency, parallel IO



PSI46 pix det
16 800 chips
66 M segments
1 m² Si sensor

APV25 Si det
110 000 chips
9.3 M segments
198 m² Si sensor

QIE8 calorimeter
220 400 chips

MAD muon det
181 000 chips
25 000 m² gas-filled

Chips to scale 1 cm

Total CMS
approx. 1 million chips
of which 700 000 ASICs

“Software Defined HPC”

- CERN’s approach is to build HPC clusters with commodity parts: *“Software Defined HPC”*
  - Compute side is solved with HTCondor & SLURM
  - Typical HPC storage is not very attractive (missing expertise + budget)
- 200-300 HPC nodes accessing ~1PB CephFS since mid-2016:
  - Manila + HPC use-cases on the same clusters. HPC is just another user.
  - Quite stable but not super high performance

IO-500

- Storage benchmark announced by John Bent on ceph-users ML (from SuperComputing 2017)
- *« goal is to improve parallel file systems by ensuring that sites publish results of both "hero" and "anti-hero" runs and by sharing the tuning and configuration »*
- We have just started testing on our CephFS clusters:
  - `IOR` throughput tests, `mdtest` + `find` metadata test
  - Easy/hard mode for shared/unique file tests

IO-500 First Look…No tuning

Test	Result
ior_easy_write	2.595 GB/s
ior_hard_write	0.761 GB/s
ior_easy_read	4.951 GB/s
ior_hard_read	0.944 GB/s

Test	Result
mdtest_easy_write	1.774 kiops
mdtest_hard_write	1.512 kiops
find	50.00 kiops
mdtest_easy_stat	8.722 kiops
mdtest_hard_stat	7.076 kiops
mdtest_easy_delete	0.747 kiops
mdtest_hard_read	2.521 kiops
mdtest_hard_delete	1.351 kiops

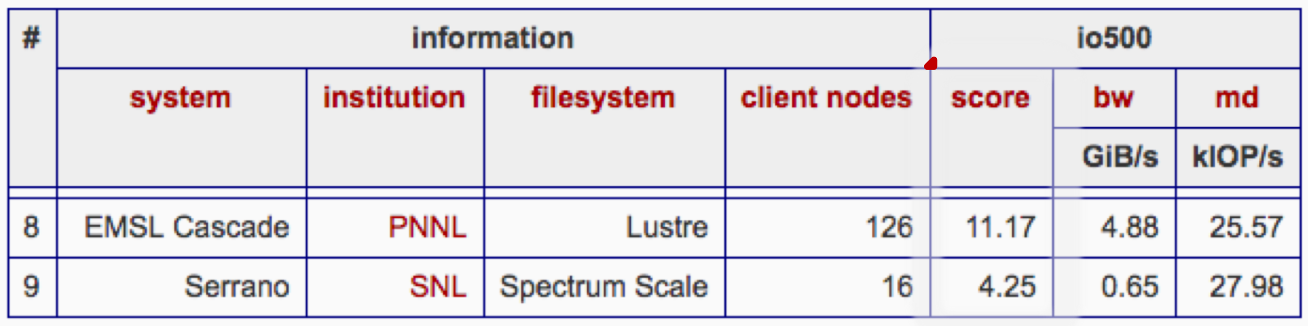
```
Luminous v12.2.4 -- Tested March 2018

411 OSDs: 800TB SSDs, 2 per server
    OSDs running on same HW as clients

2 active MDSs running on VMs
```

```
[SCORE] Bandwidth 1.74 GB/s : IOPS 3.47 kiops : TOTAL 2.46
```

IO-500 First Look…No tuni

Test	Result
ior_easy_write	2.595 GB/s
ior_hard_write	0.761 GB/s

Test	Result
mdtest_easy_write	1.774 kiops
mdtest_hard_write	1.512 kiops



#	information				io500		
	system	institution	filesystem	client nodes	score	bw	md
						GiB/s	kIOP/s
8	EMSL Cascade	PNNL	Lustre	126	11.17	4.88	25.57
9	Serrano	SNL	Spectrum Scale	16	4.25	0.65	27.98

Luminous v

411 OSDs:
OSDs r

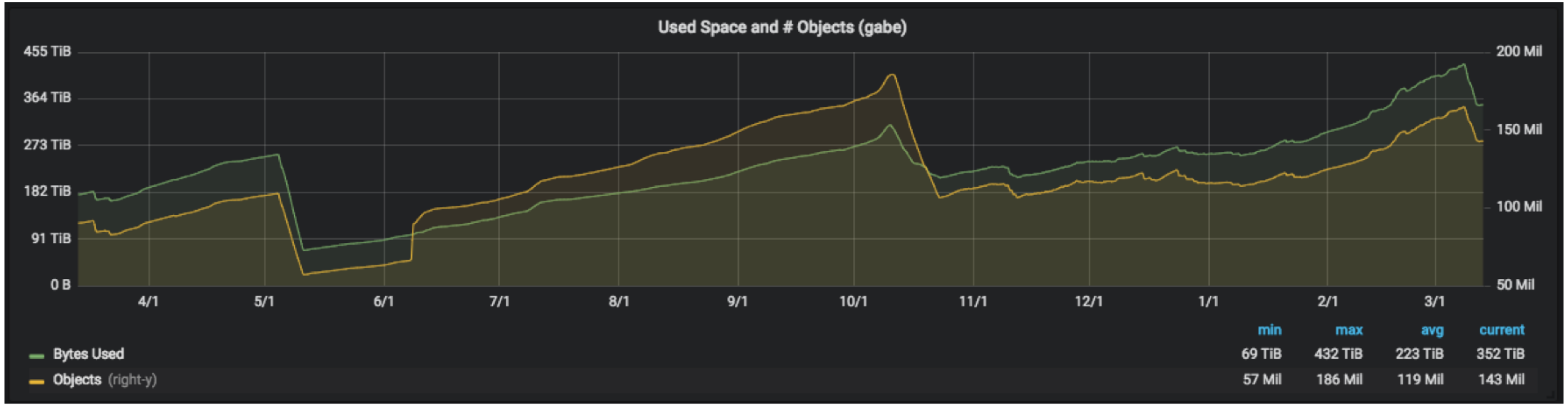
2 active MDSs running on VMs

```
[SCORE] Bandwidth 1.74 GB/s : IOPS 3.47 kiops : TOTAL 2.46
```

RGW

S3 @ CERN

- Ceph luminous cluster with VM gateways. Single region.
  - 4+2 erasure coding. Physics data for small objects, volunteer computing, some backups.
  - Pre-signed URLs and object expiration working well.
- HAProxy is very useful:
  - High-availability & mapping *special* buckets to dedicated gateways

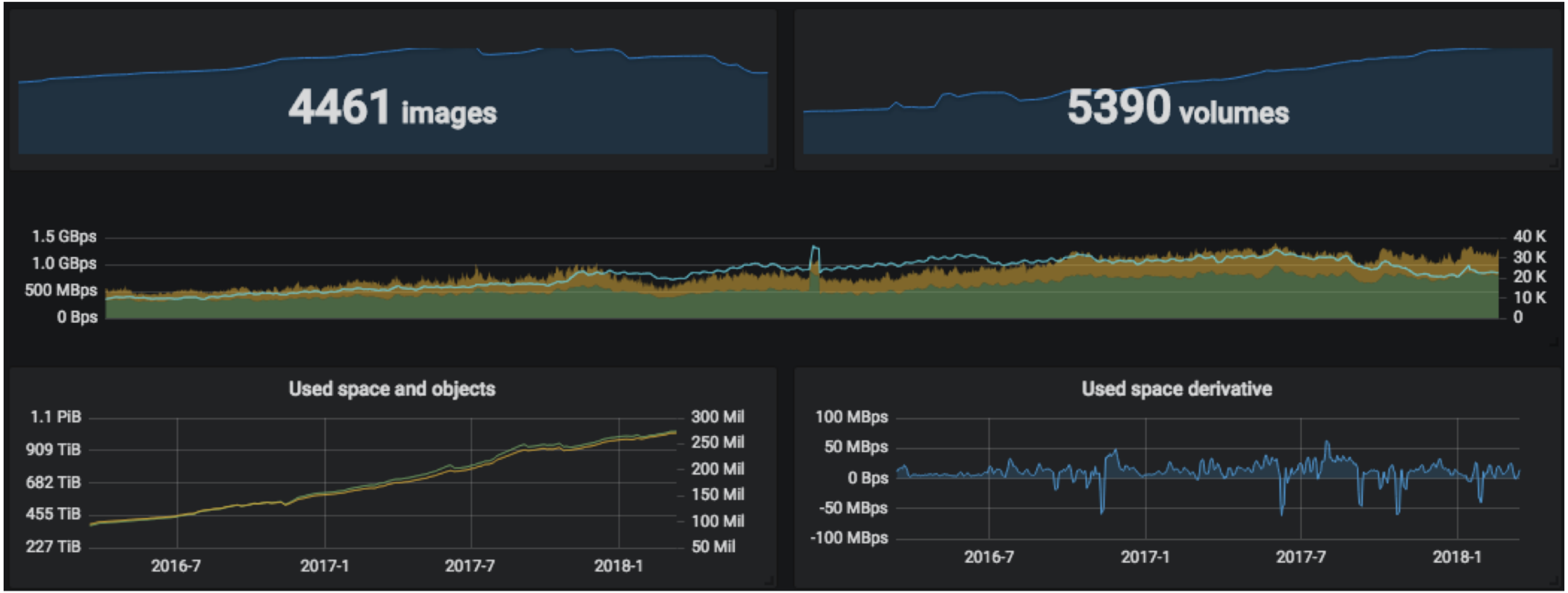


Used Space and # Objects (gabe)

	min	max	avg	current
Bytes Used	69 TiB	432 TiB	223 TiB	352 TiB
Objects (right-y)	57 Mil	186 Mil	119 Mil	143 Mil

RBD

RBD: Ceph + OpenStack



Cinder Volume Types

Volume Type	Size (TB)	Count
standard	871	4,758
io1	440	608
cp1	97	118
cpio1	269	107
wig-cp1	26	19
wig-cpio1	106	13
io-test10k	20	1
Totals:	**1,811**	**5,624**

RBD @ CERN

- OpenStack Cinder + Glance use-cases continue to be highly reliable:
  - QoS via IOPS/BW throttles is essential.
  - *Spectre/Meltdown reboots updated all clients to luminous!*
- Ongoing work:
  - Recently finished an expanded *rbd trash* feature
  - Just starting work on a persistent cache for librbd
    - **CERN Openlab collaboration with Rackspace!**
  - Writing a backup driver for glance (RBD to S3)

Hyperconverged Ceph/Clo

- Experimenting with co-located ceph-osd on HVs and HPC:
  - New cluster with 384 SSDs on HPC nodes
- Minor issues related to server isolation:
  - cgroups or NUMA pinning are options but not yet used.
- Issues are related to our operations culture:
  - We (Ceph team) don’t own the servers – need to co-operate with the cloud/HPC teams.
  - E.g. When is it ok to reboot a node? how to drain a node? Software upgrade procedures.

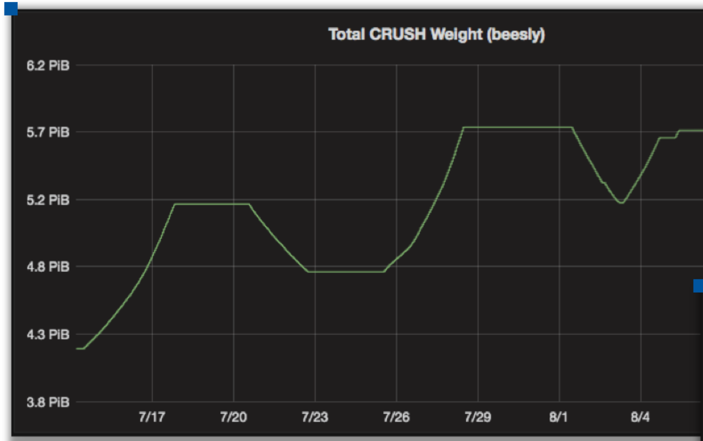
User Feedback:
From Jewel to Luminous

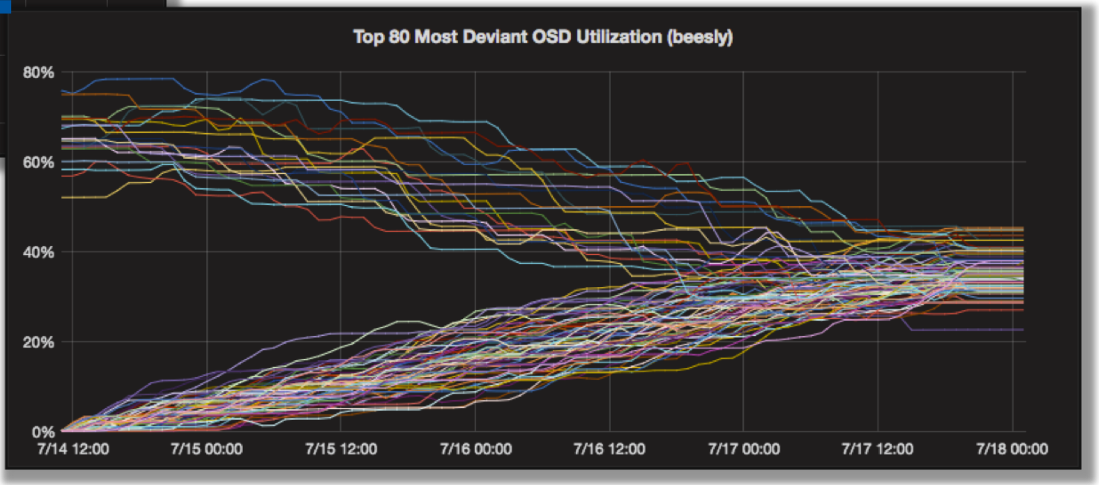
Jewel to Luminous Upgrad

- In general upgrades went well with no big problems.
- New/replaced OSDs are BlueStore (ceph-volume lvm)
  - Preparing a FileStore conversion script for our infrastructure
- ceph-mgr balancer is very interesting:
  - Actively testing the crush-compat mode
  - Python module can be patched in place for quick fixes

How to replace many OSDs?

Fully replaced 3PB of block storage with 6PB new hardware over several weeks, transparent to users.





GitHub cernceph/ceph-scripts

Current Challenges

- RBD / OpenStack Cinder:
  - Ops: how to identify active volumes?
    - “`rbd top`”
  - Performance: µs latencies and kHz IOPS.
    - Need persistent SSD caches.
  - On the wire encryption, client-side volume encryption
  - OpenStack: volume type / availability zone coupling for hyper-converged clusters

Current Challenges

- CephFS HPC:
  - HPC: parallel MPI I/O and single-MDS metadata perf (IO-500!)
  - Copying data across /cephfs: need "`rsync --ceph`"
- CephFS general use-case:
  - Scaling to 10,000 (or 100,000!) clients:
    - client throttles, tools to block/disconnect noisy users/clients.
      - Need “`ceph client top`”
    - native Kerberos (without NFS gateway), group accounting and quotas
    - HA CIFS and NFS gateways for non-Linux clients
  - How to backup a 10 billion file CephFS?
    - e.g. how about binary diff between snaps, similar to ZFS send/receive?

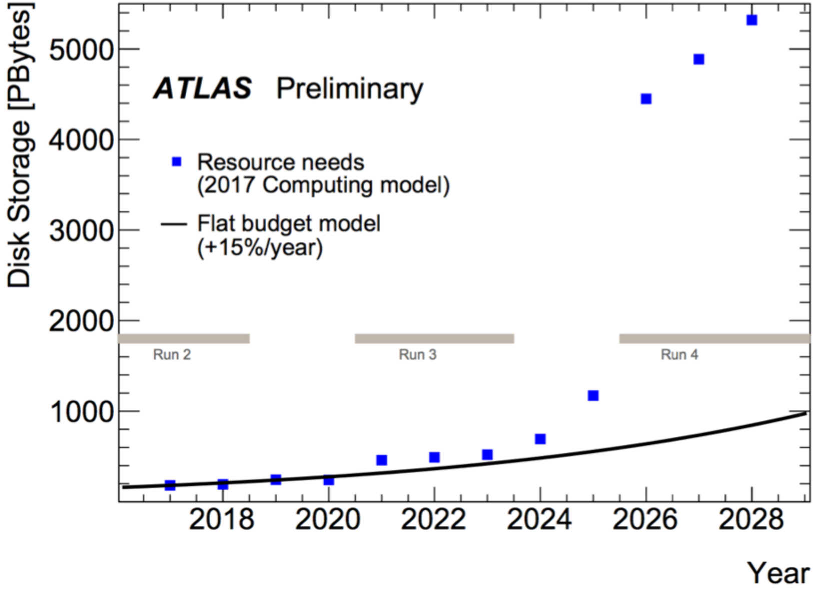
Current Challenges

- RADOS:
  - How to phase in new features on old clusters
    - e.g. we have 3PB of RBD data with *hammer* tunables
  - Pool-level object backup (convert from replicated to EC, copy to non-Ceph)
    - `rados export` the diff between two pool snaphots?
- Areas we cannot use Ceph yet:
  - Storage for large enterprise databases *(are we close?)*
  - Large scale batch processing
  - Single filesystems spanning multiple sites
  - HSM use-cases (CephFS with tape backend?, Glacier for S3?)

Future…

HEP Computing for the 202

- Run-2 (2015-18):
  ~50-80PB/year
- Run-3 (2020-23):
  ~150PB/year
- Run-4: ~600PB/year?!

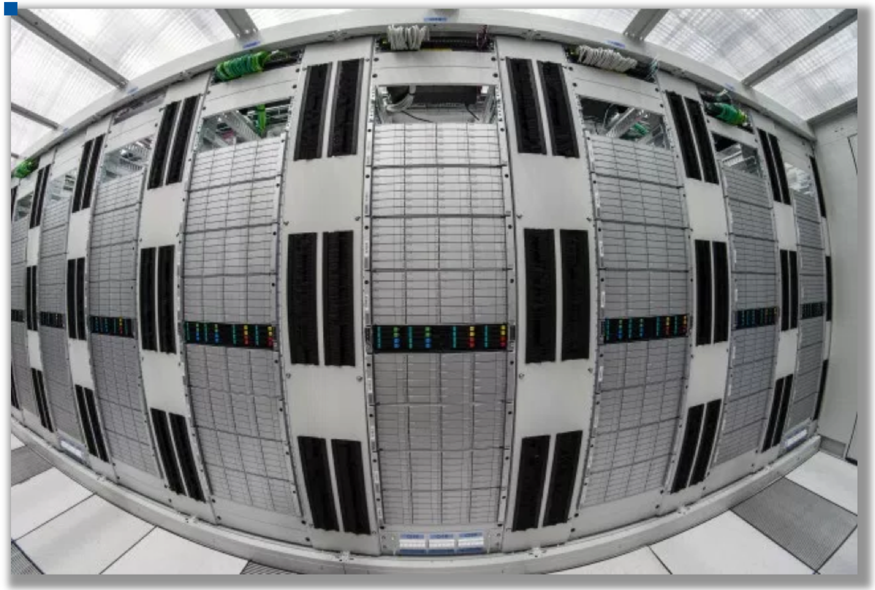


"Data Lakes" – globally distributed, flexible placement, ubiquitous access

https://arxiv.org/abs/1712.06982

Ceph Bigbang Scale Testing

- *Bigbang* scale tests mutually benefit CERN & Ceph project
- *Bigbang I:* 30PB, 7200 OSDs, Ceph hammer. Several *osdmap* limitations
- *Bigbang II:* Similar size, Ceph jewel. Scalability limited by OSD/MON messaging. Motivated *ceph-mgr*
- *Bigbang III:* 65PB, 10800 OSDs



https://ceph.com/community/new-luminous-scalability/

Thanks…

Thanks to my CERN Collea

- Ceph team at CERN
  - Hervé Rousseau, Teo Mouratidis, Roberto Valverde, Paul Musset, Julien Collet
  - Massimo Lamanna / Alberto Pace (Storage Group Leadership)
  - Andreas-Joachim Peters (Intel EC)
  - Sebastien Ponce (radosstriper)
- OpenStack & Containers teams at CERN
  - Tim Bell, Jan van Eldik, Arne Wiebalck (also co-initiator of Ceph at CERN), Belmiro Moreira, Ricardo Rocha, Jose Castro Leon
- HPC team at CERN
  - Nils Hoimyr, Carolina Lindqvist, Pablo Llopis

Thanks!

谢谢！

home.cern