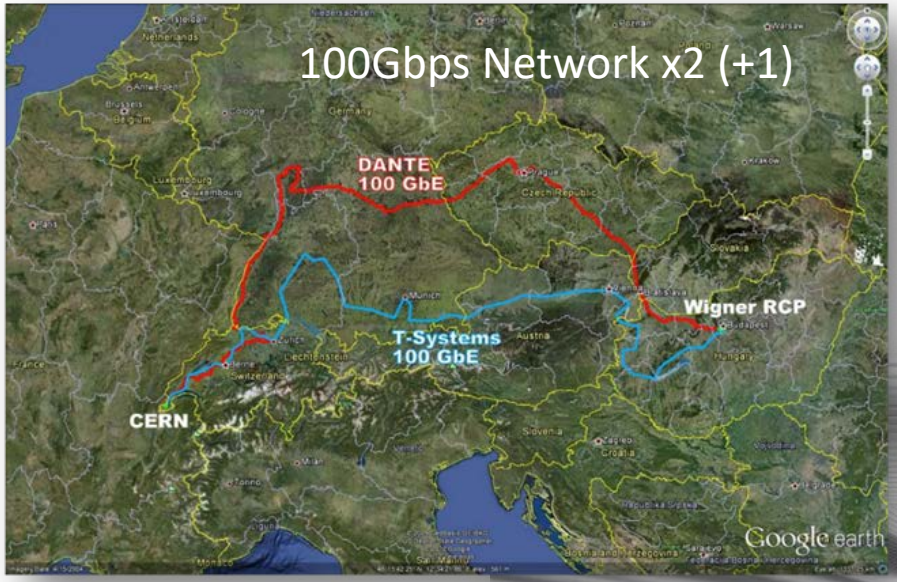
ATLASグリッド概要

中村智昭 (KEK-CRC)

CERN計算機



データ量の増大
- 加速器の高輝度化
- 検出器の高精度化
- データ収集・記録技術の向上

集中から分散へ
→グリッドコンピューティング

仮想化による最適化
→クラウド技術

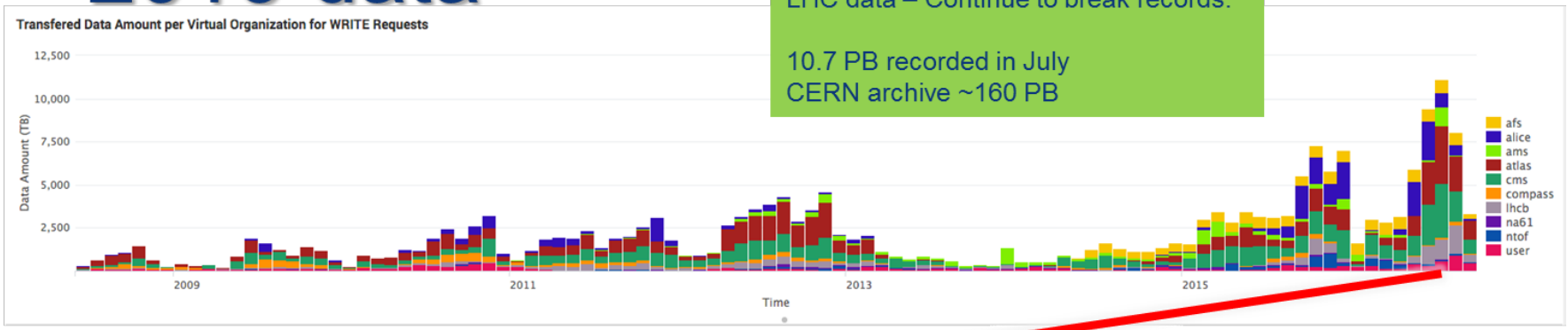
Meyrin
~15万 CPUコア
~165 PB disk
~200 PB tape

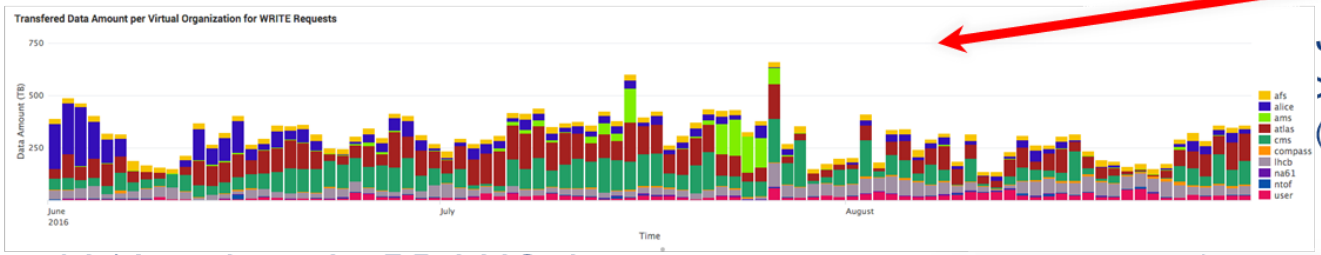
Wigner Data Center
~6万 CPUコア
~100 PB disk

2016 data

LHC data – Continue to break records:

10.7 PB recorded in July
CERN archive ~160 PB

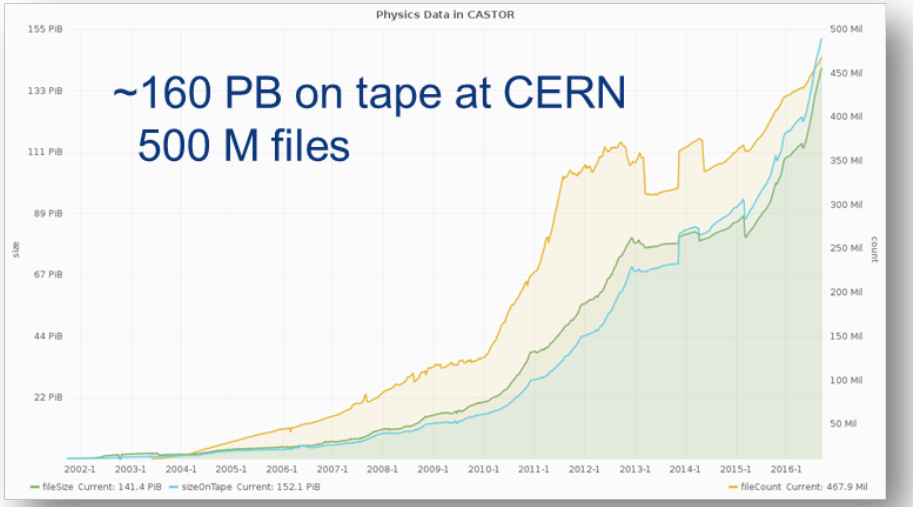




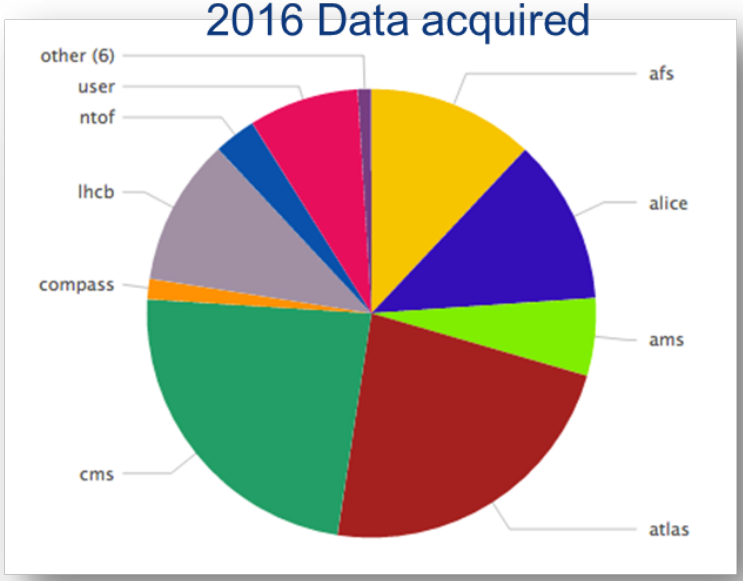
June-Aug 2016
>500 TB / day
(Run 1 peak for HI was 220 TB)

I. Bird

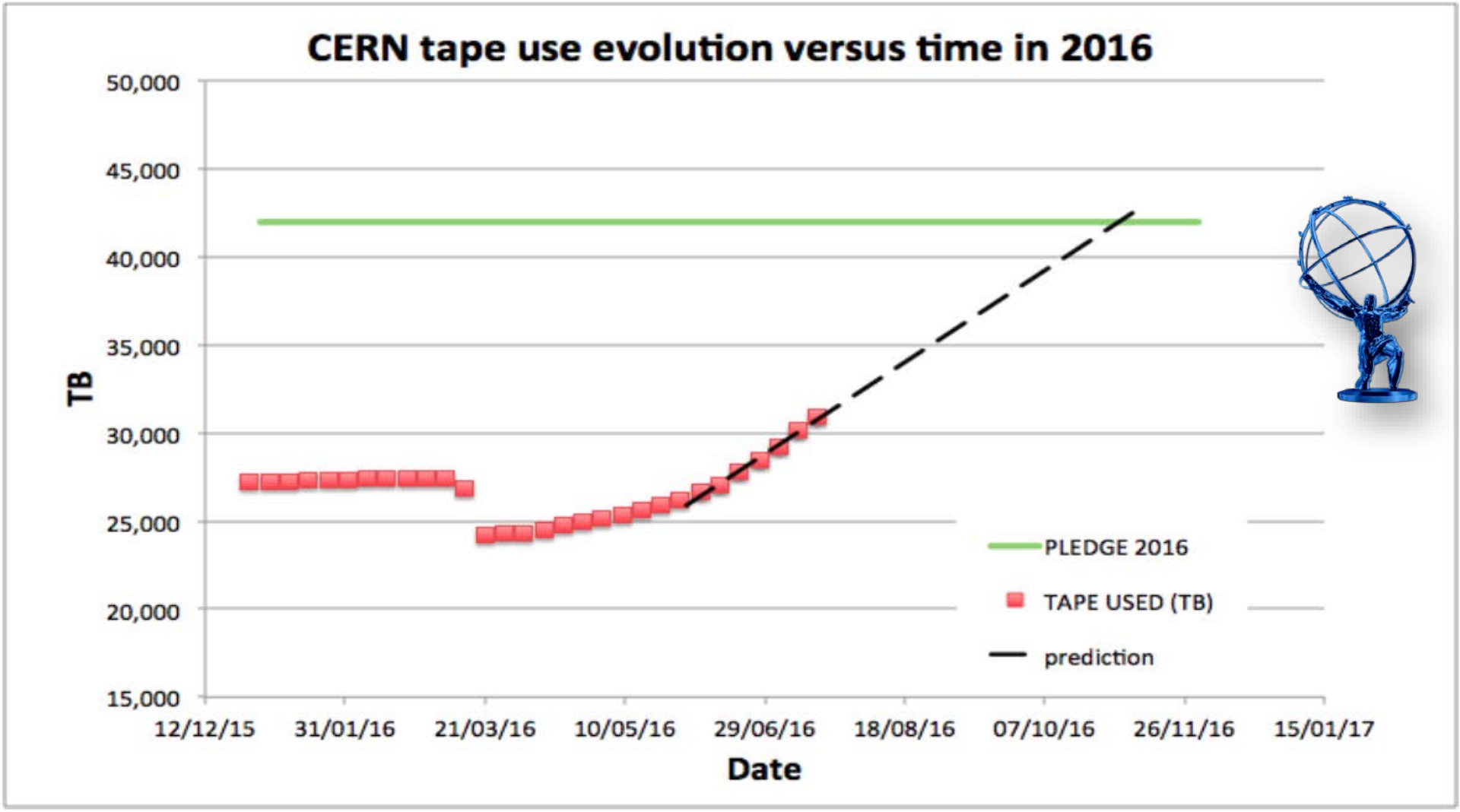
2016 to date: 35 PB LHC data:
ALICE 6, ATLAS 11.6, CMS 11.9, LHCb 5.4)



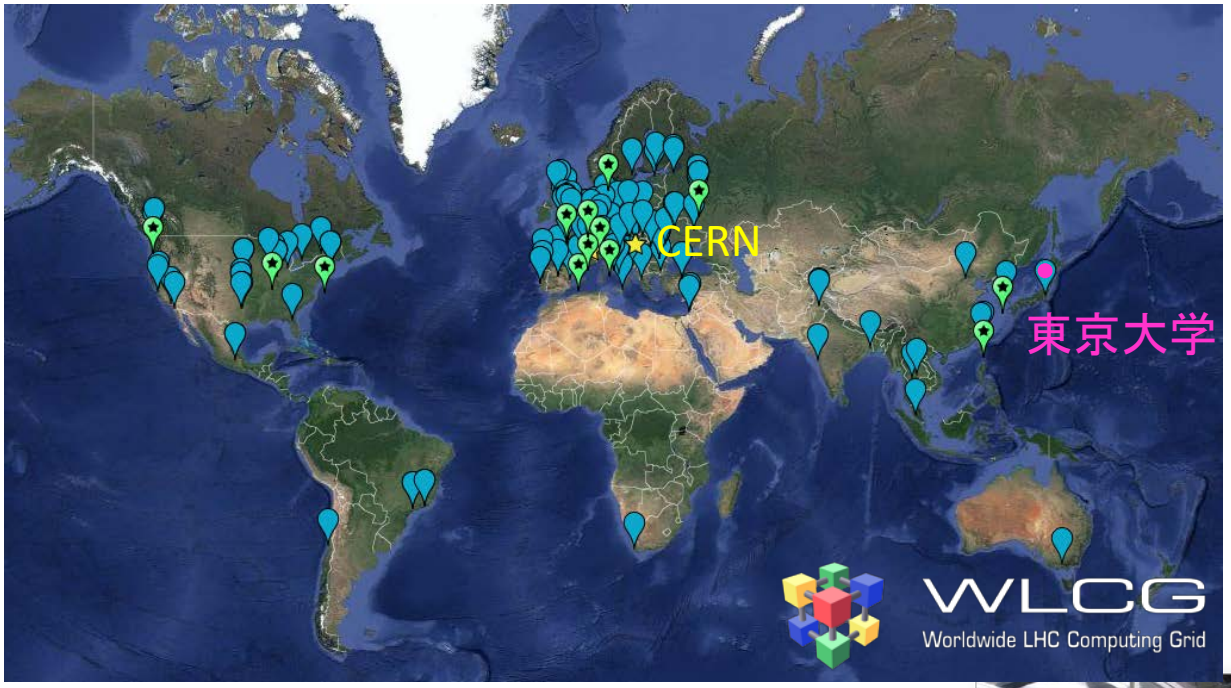
2016 Data acquired



ATLAS raw data（CERNのTapeアーカイブ）



Worldwide LHC Computing Grid



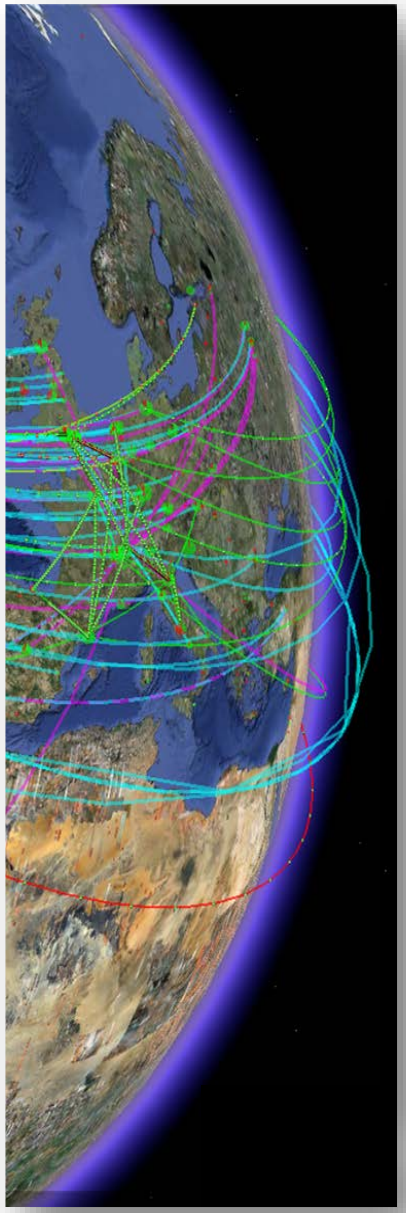
研究者は，自分のコンピューターからグリッドにアクセスすることにより，データや計算機の所在を意識することなく，即座に物理解析結果を得ることができる。

実験データの取得から物理解析結果を導くまで，パイプラインでの処理を可能とする仕組み。

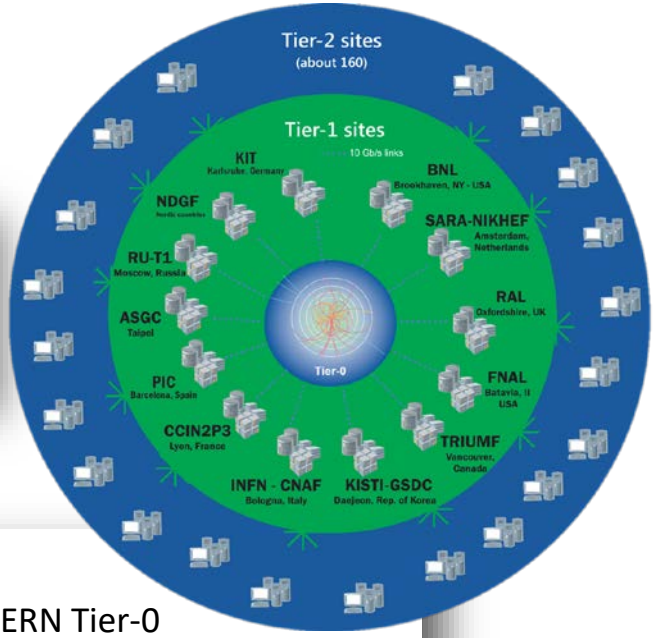
40か国に点在する170の拠点が高帯域学術ネットワーク（10Gbps～100Gbps）で接続されている。
全体で50万個のCPUと50PBのディスクストレージが利用でき，常時200万のデータ解析とシミュレーションが実行されている。（CERNの計算機の割合は全体の10%程度）

東京大学素粒子物理国際研究センターに配備するATLAS実験用のグリッド拠点

Worldwide LHC Computing Grid



WLCG
Worldwide LHC Computing Grid



ATLASグリッド

データアーカイブと再構成:	CERN Tier-0
分散データアーカイブ:	10サイトのTier-1
MCプロダクションとデータ解析:	100サイト以上のTier-2

計算機資源量 (WLCG pledge 2016)

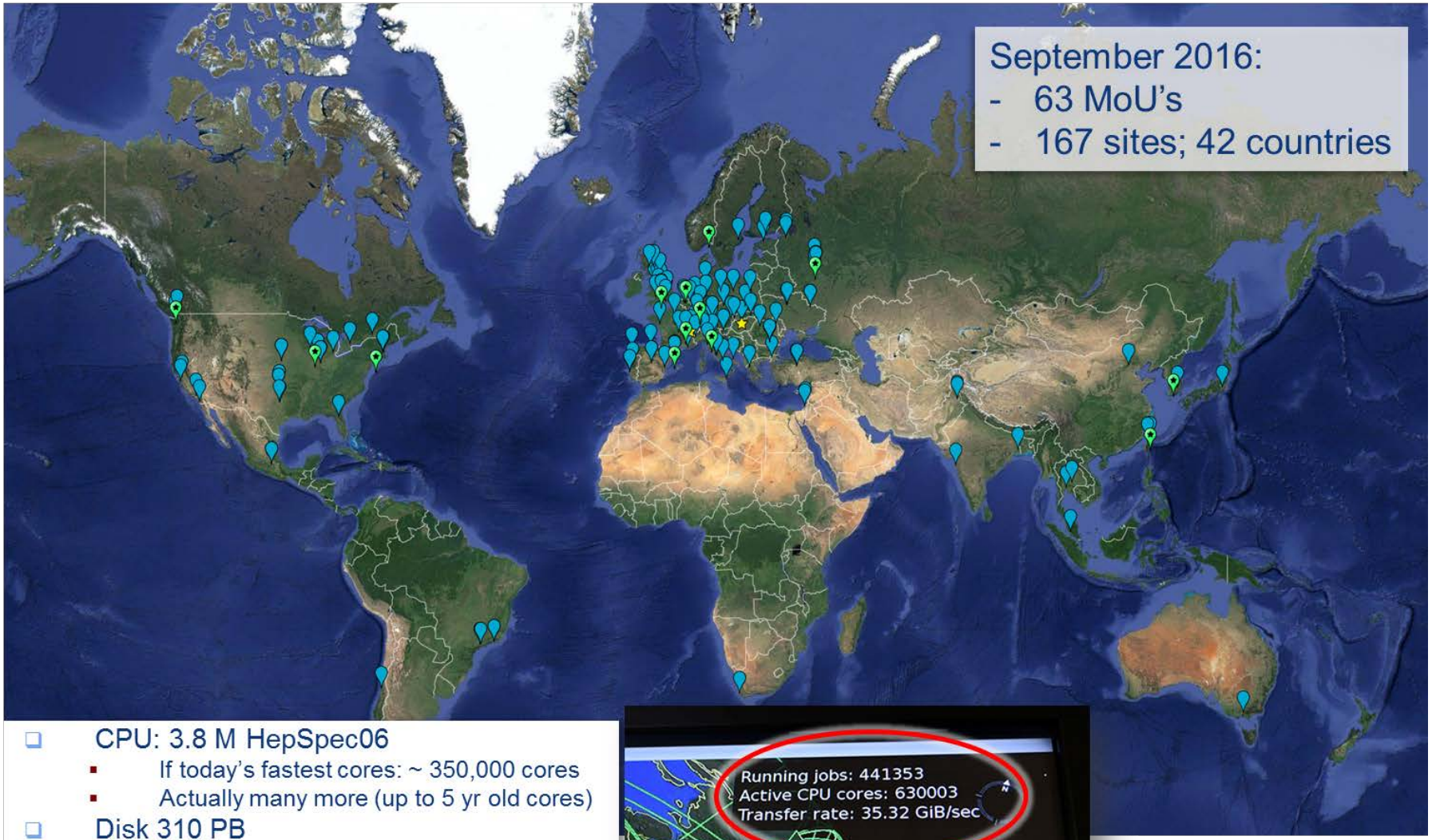
CPU:	~3.8 MHS06 (10~20HS06/core = ~40万コア)
Disk:	~310 PB
Tape:	~390 PB

オーバープレッジ

各サイトの構成

グリッドミドルウェアは大きく分けて3種類 (EGI, OSG, NorduGrid)
サイトごとに採用する設備・ハードウェアはバラバラ
国ごとに予算獲得方法も違う

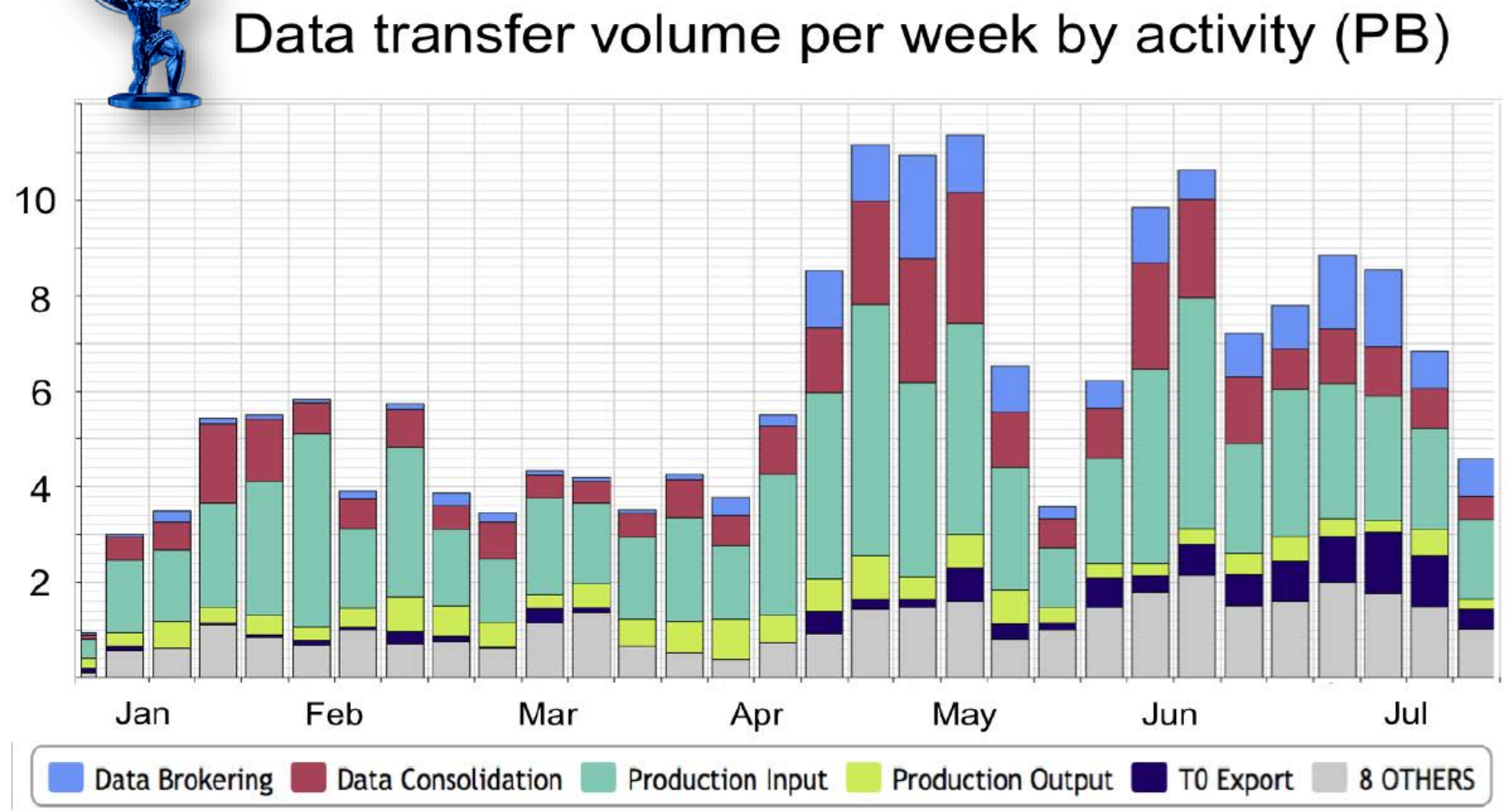
WLCG Collaboration



- CPU: 3.8 M HepSpec06
  - If today's fastest cores: ~ 350,000 cores
  - Actually many more (up to 5 yr old cores)
- Disk 310 PB
- Tape 390 PB

I. Bird

ATLASサイト間でのデータ転送



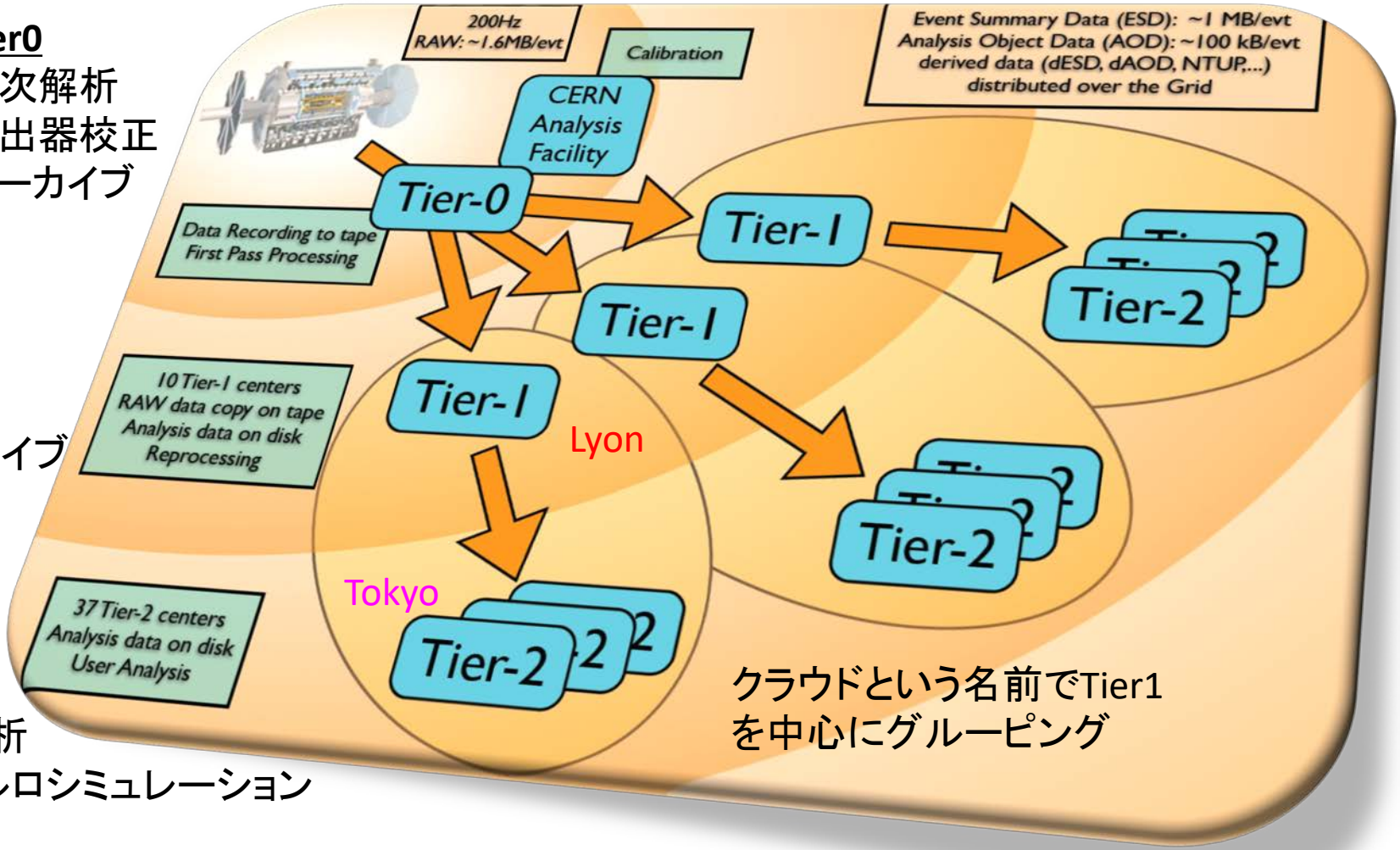
旧データ分配とワークフロー

モナークモデル

Tier0
一次解析
検出器校正
アーカイブ

Tier1
再解析
分散アーカイブ

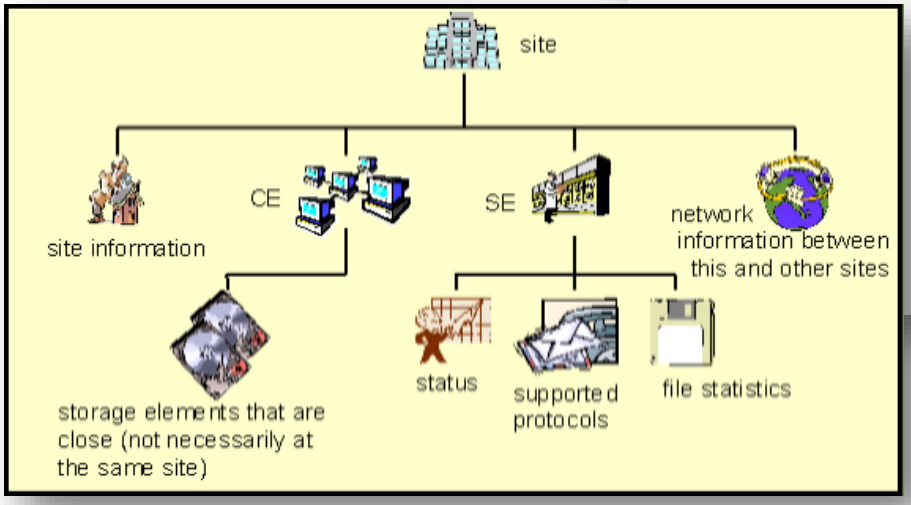
Tier2
ユーザ解析
モンテカルロシミュレーション

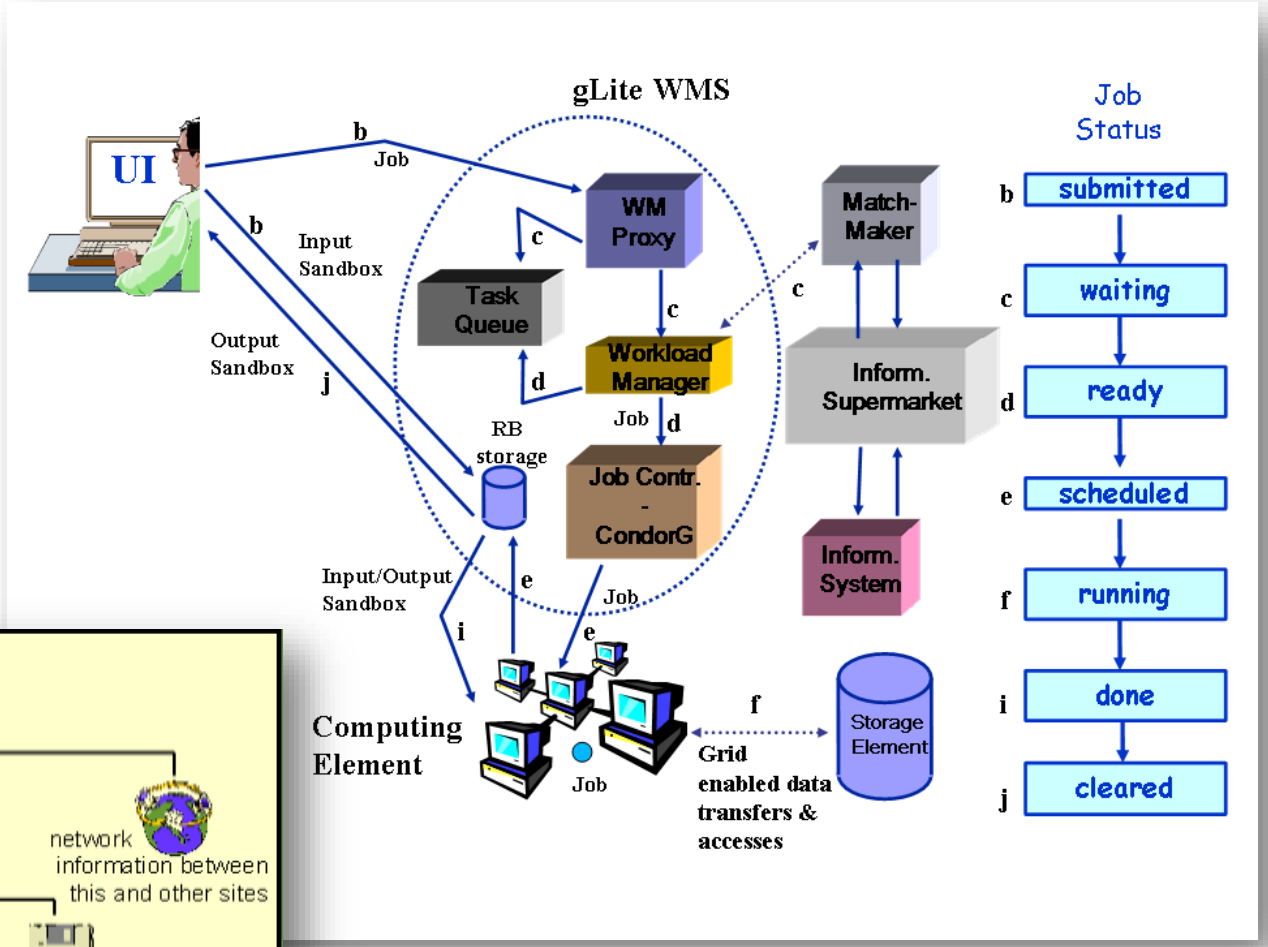


クラウドという名前でTier1を中心にグルーピング

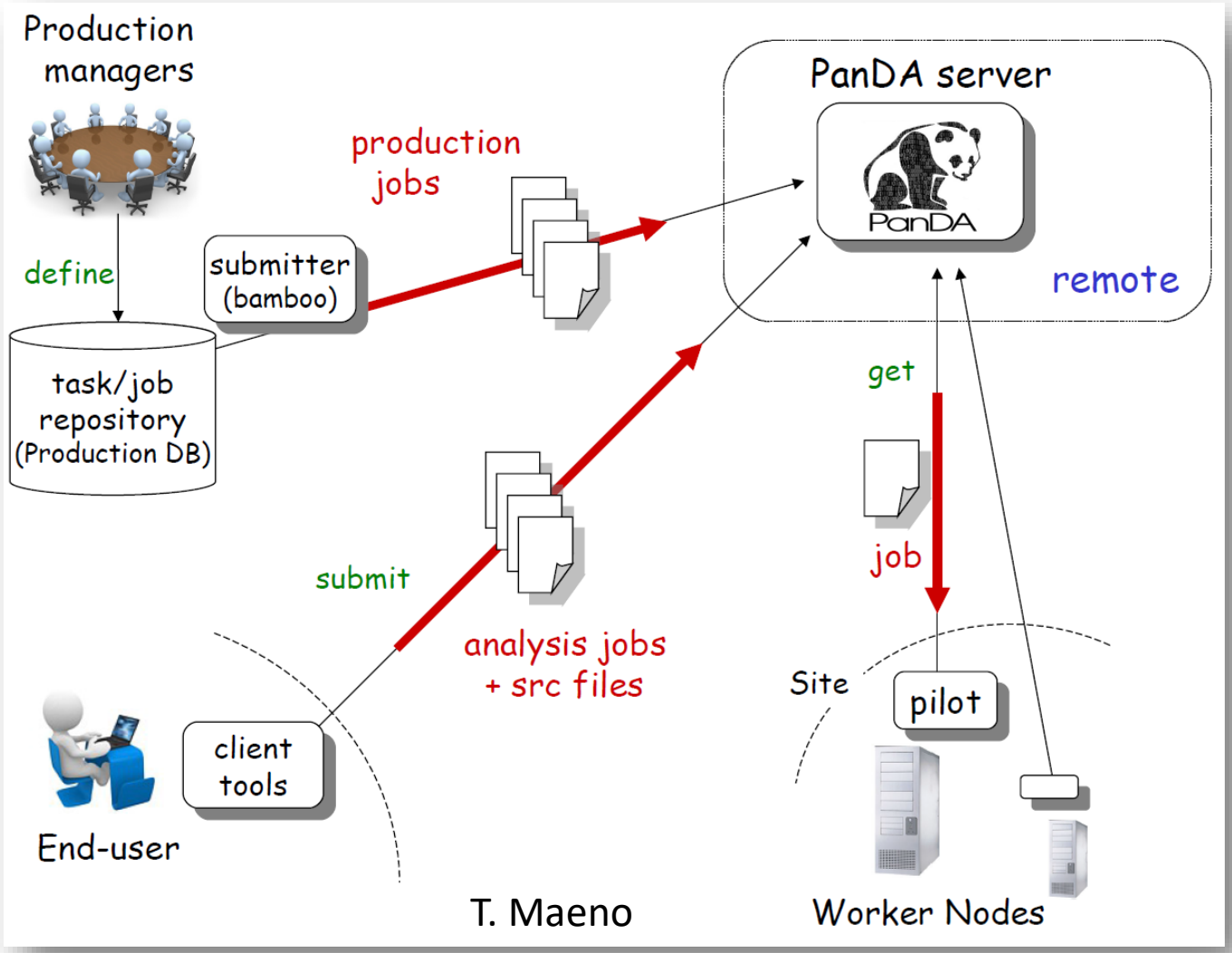
グリッド初期

グリッド証明書を使った認証以外にはローカルのバッチクラスターと同じような使い方





PanDA at LHC Run1



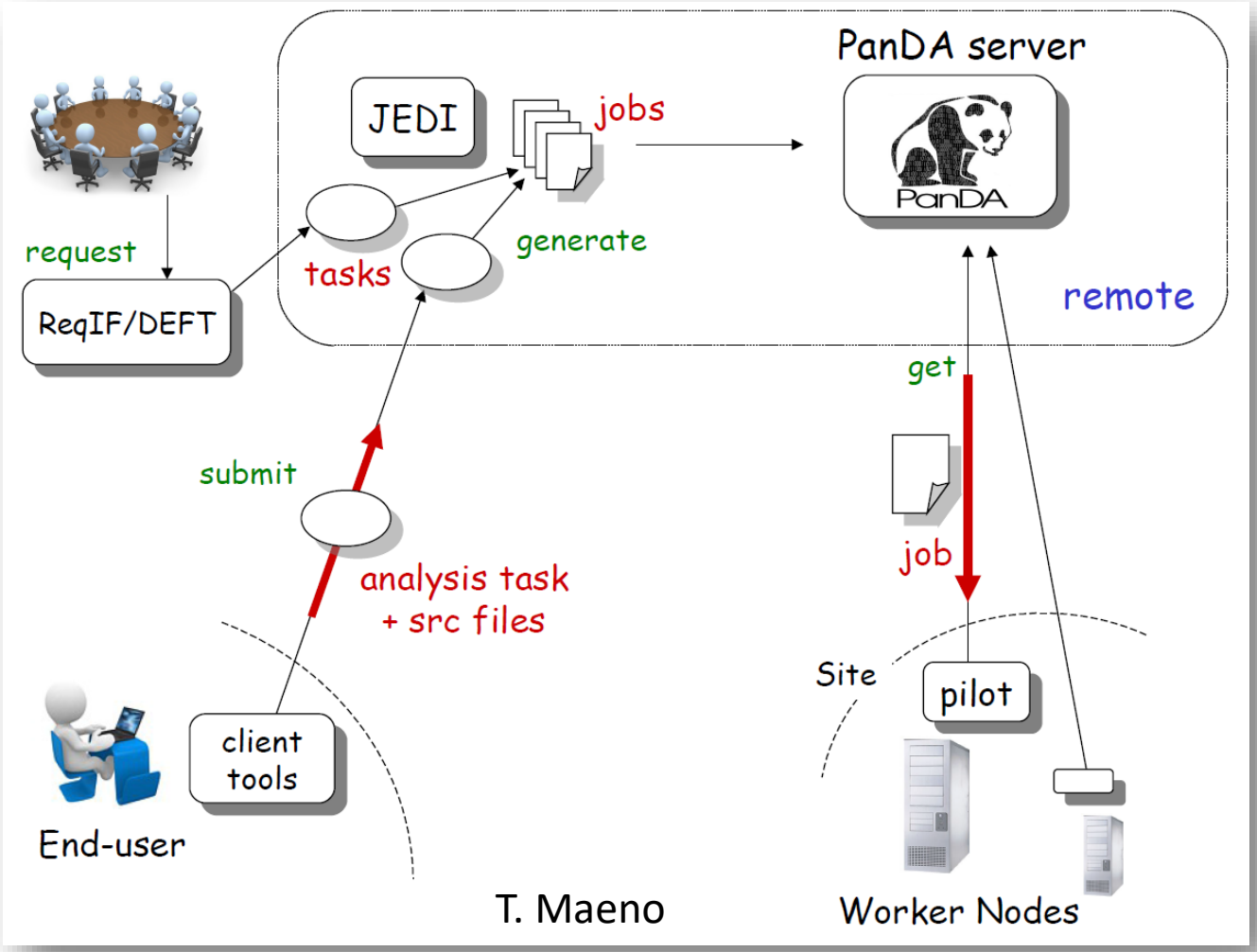
プロダクション:

マネージャがジョブの塊としてのタスクを定義してデータベースに登録

ユーザー解析:

入出力ファイルを意識して処理プログラムを投入

Improvement for Run2

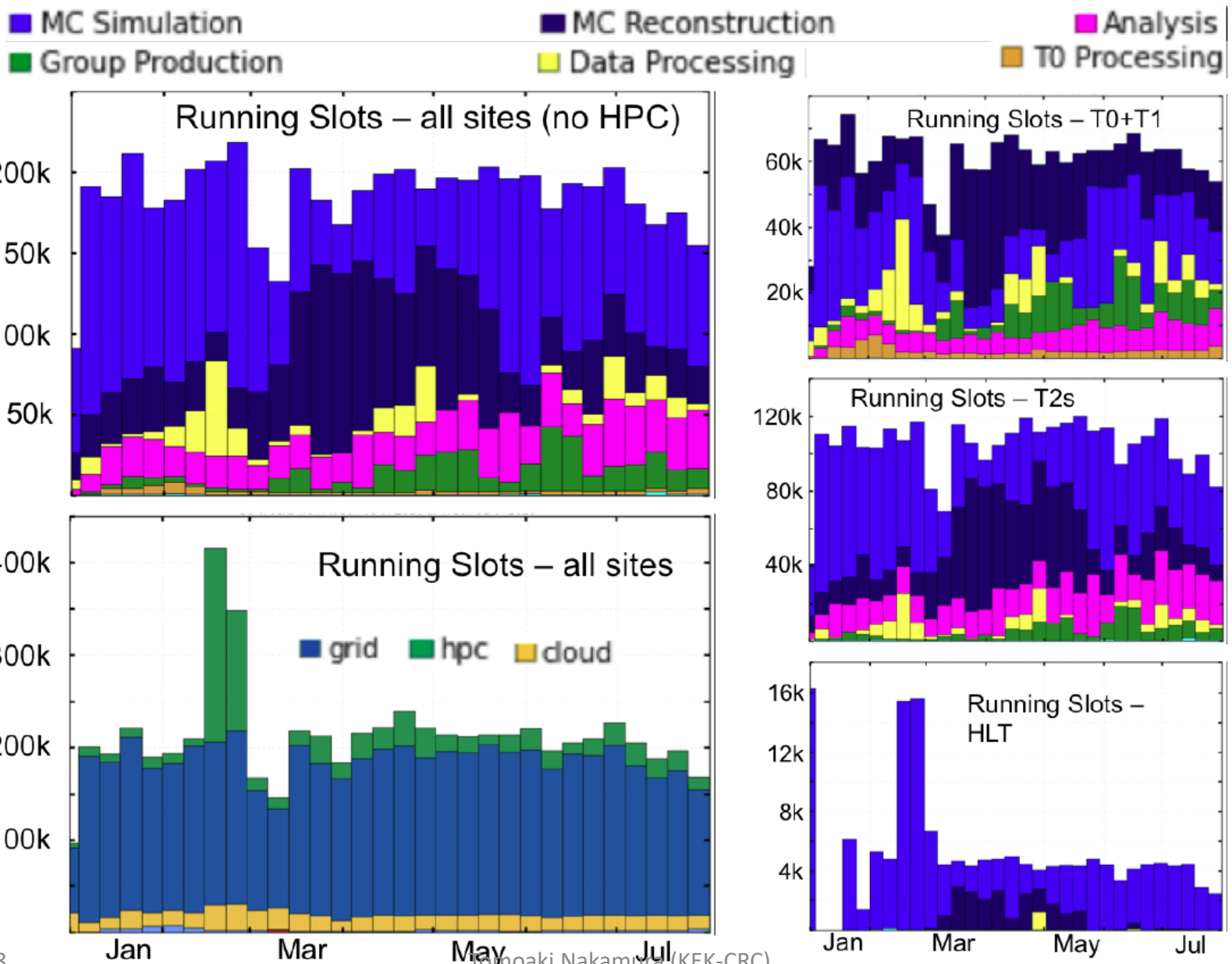


ユーザーは各サイトにおける最適なジョブパラメタを知らない
- ジョブ実行時間
- 入出力ファイルサイズ
- 使用可能なメモリサイズ
- パラレル数
- ネットワーク接続性

→ユーザーはタスクのみを投入
→最適ジョブを自動生成
→資源利用の効率化

大量の失敗ジョブ対策
→サンプルジョブを投入
→失敗率の高いタスクを却下
→20~30%の無駄を削減

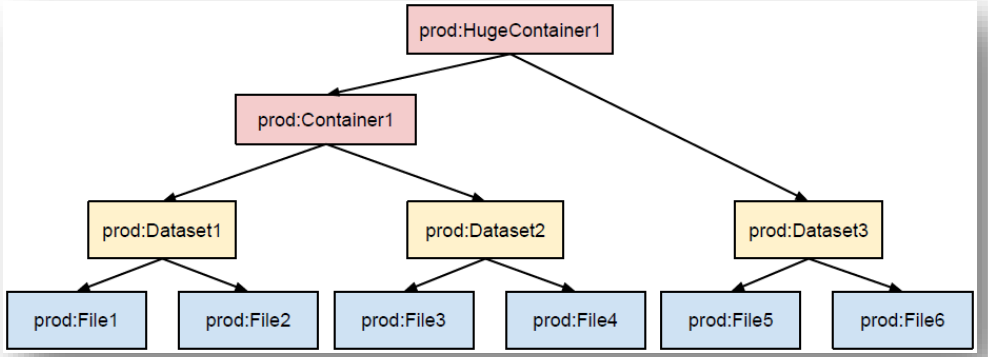
グリッドジョブ


MC Simulation
MC Reconstruction
Analysis
Group Production
Data Processing
T0 Processing
Running Slots – all sites (no HPC)
200k
150k
100k
50k
Running Slots – T0+T1
60k
40k
20k
Running Slots – T2s
120k
80k
40k
Running Slots – all sites
grid
hpc
cloud
400k
300k
200k
100k
Running Slots – HLT
16k
12k
8k
4k
Jan
Mar
May
Jul

Rucio (データマネージメント)

Rucio Storage Element

- スケールしないLFCから脱却
  - データベースから再設計
- ファイル，データセット，レプリカ情報の概念は保持
- グリッドワイドなアカウント管理を導入
  - ユーザー，グループ，プロジェクト
  - クォータ，パーミッション
- メタデータを強化
  - ファイルに含まれるイベント数
- 既存のオープンソースソフトウェとの互換性を持たせる
  - RESTfulなインターフェイス



http://rucio.cern.ch/client_tutorial.html

```
# rucio get-metadata
 data_test:data_test.00250001.calibration_DcmDummyProcessor.daq.RAW._lb0000._SFO-5._0001.data
campaign: None
updated_at: 2015-01-30 20:51:37
is_new: None
is_open: None
guid: e657bff1aea8e411b4450030489eba28
availability: AVAILABLE
deleted_at: None
panda_id: None
provenance: None
accessed_at: None
version: None
scope: data_test
hidden: False
md5: None
events: 2444
adler32: 06f2f6c2
...
```

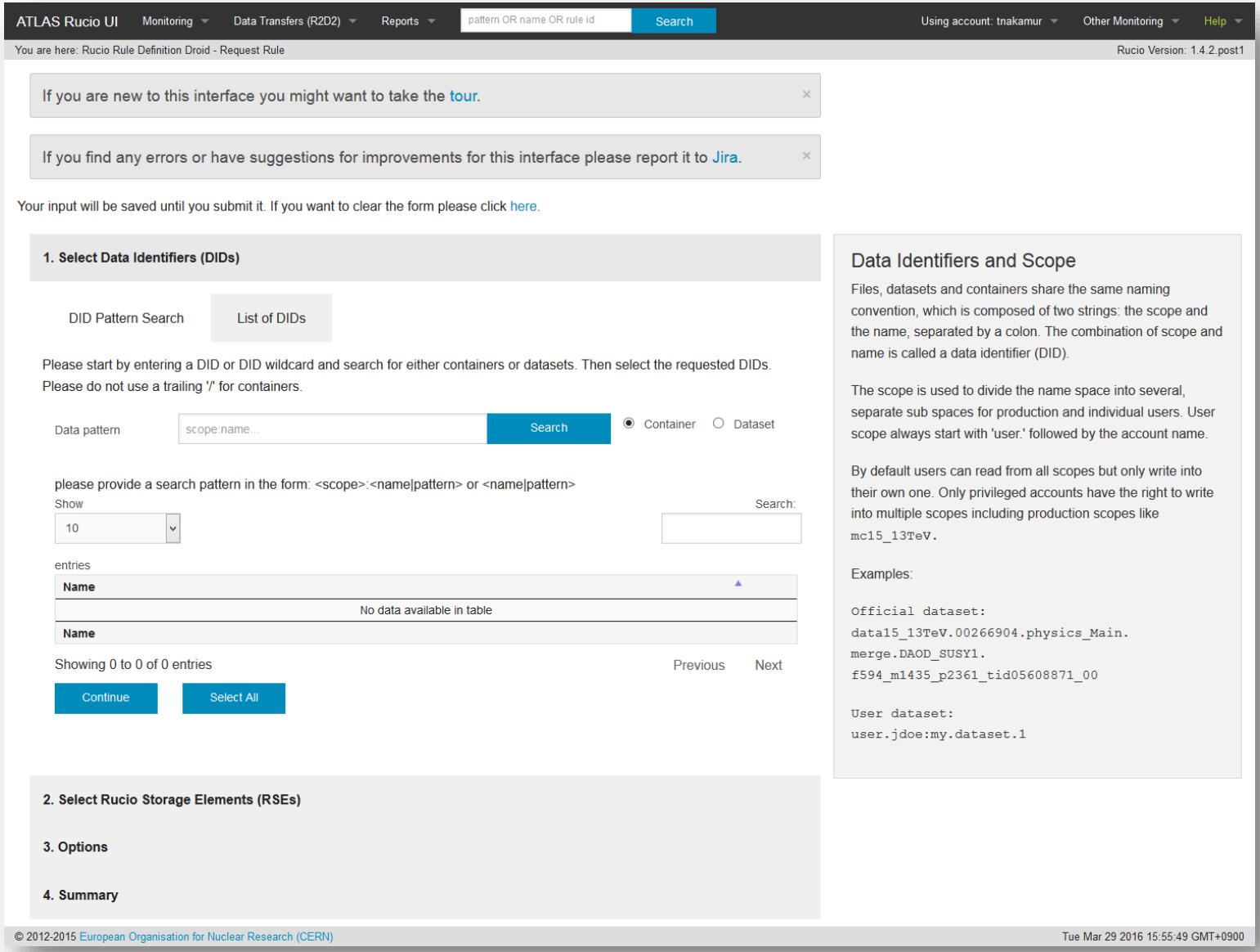
```
# rucio list-dids --recursive  user.serfon:user.serfon.test.1234.31052013.214
|     |- user.serfon:user.serfon.test.1234.31052013.212 [CONTAINER]
|     |     |- user.serfon:user.serfon.test.24092014.1 [DATASET]
|     |     |- user.serfon:user.serfon.test.25092014.1 [DATASET]
|     |     |- user.serfon:user.serfon.test.26092014.1 [DATASET]
|     |     |     |- user.serfon:file1.beaf170153b34b12b86b8a667848747d [FILE]
|     |     |     |- user.serfon:file2.beaf170153b34b12b86b8a667848747d [FILE]
|     |     |     |- user.serfon:file3.beaf170153b34b12b86b8a667848747d [FILE]
|     |- user.serfon:user.serfon.test.1234.31052013.215 [CONTAINER]
```

```
> curl -i -H "X-Rucio-Account: jdoe" --
cacert $X509_USER_PROXY --cert
$X509_USER_PROXY --capath /etc/grid-
security/certificates/ -X GET https:
//voatlasrucio-auth-prod.cern.
ch/auth/x509_proxy
```

```
[~]$ rucio download --rse TOKYO-LCG2_DATADISK dids
```

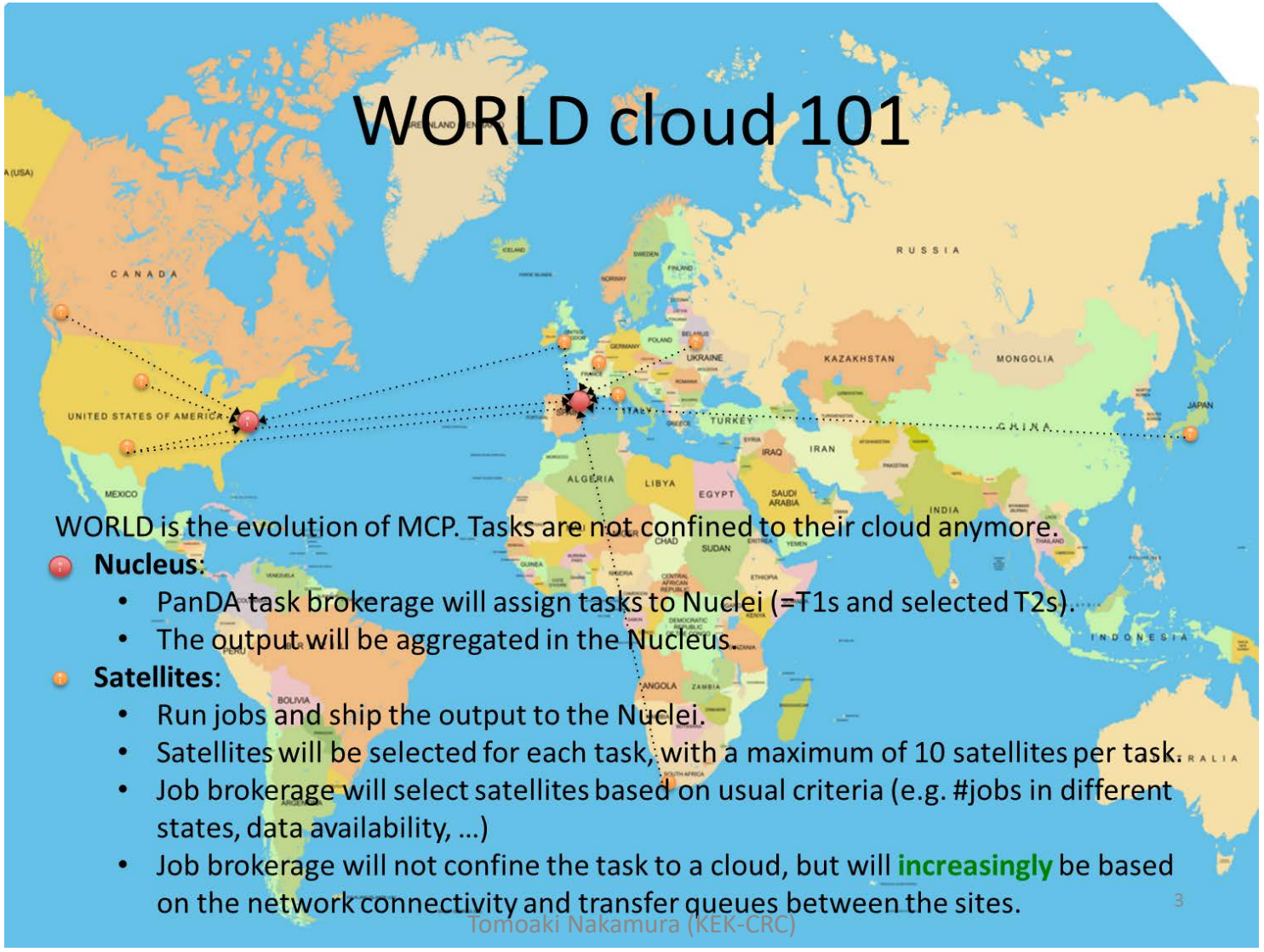
Rucio request interface

https://rucio-ui.cern.ch/r2d2/request

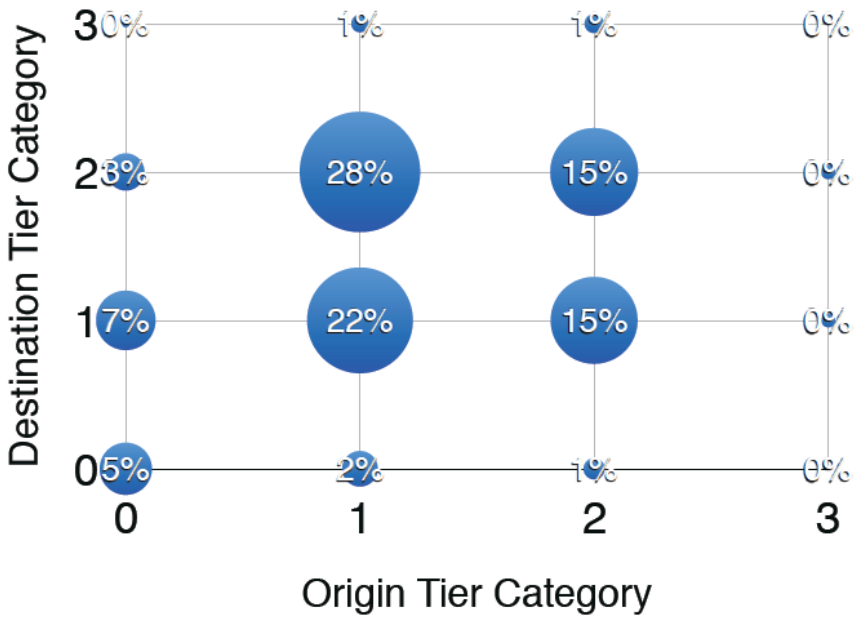


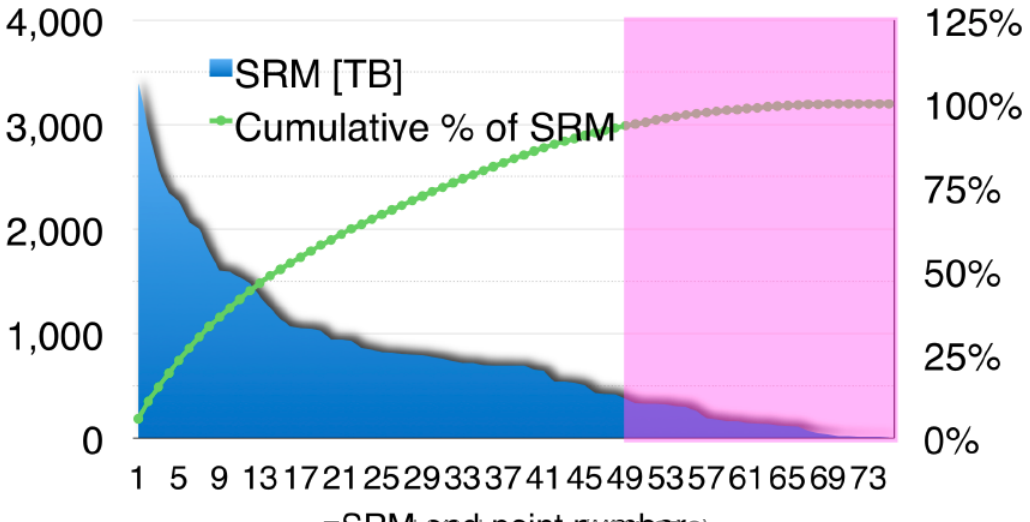
現在：Worldクラウド

- Availabilityが高いサイト
- リソース量が多いサイト
- ネットワーク接続性が良いサイト

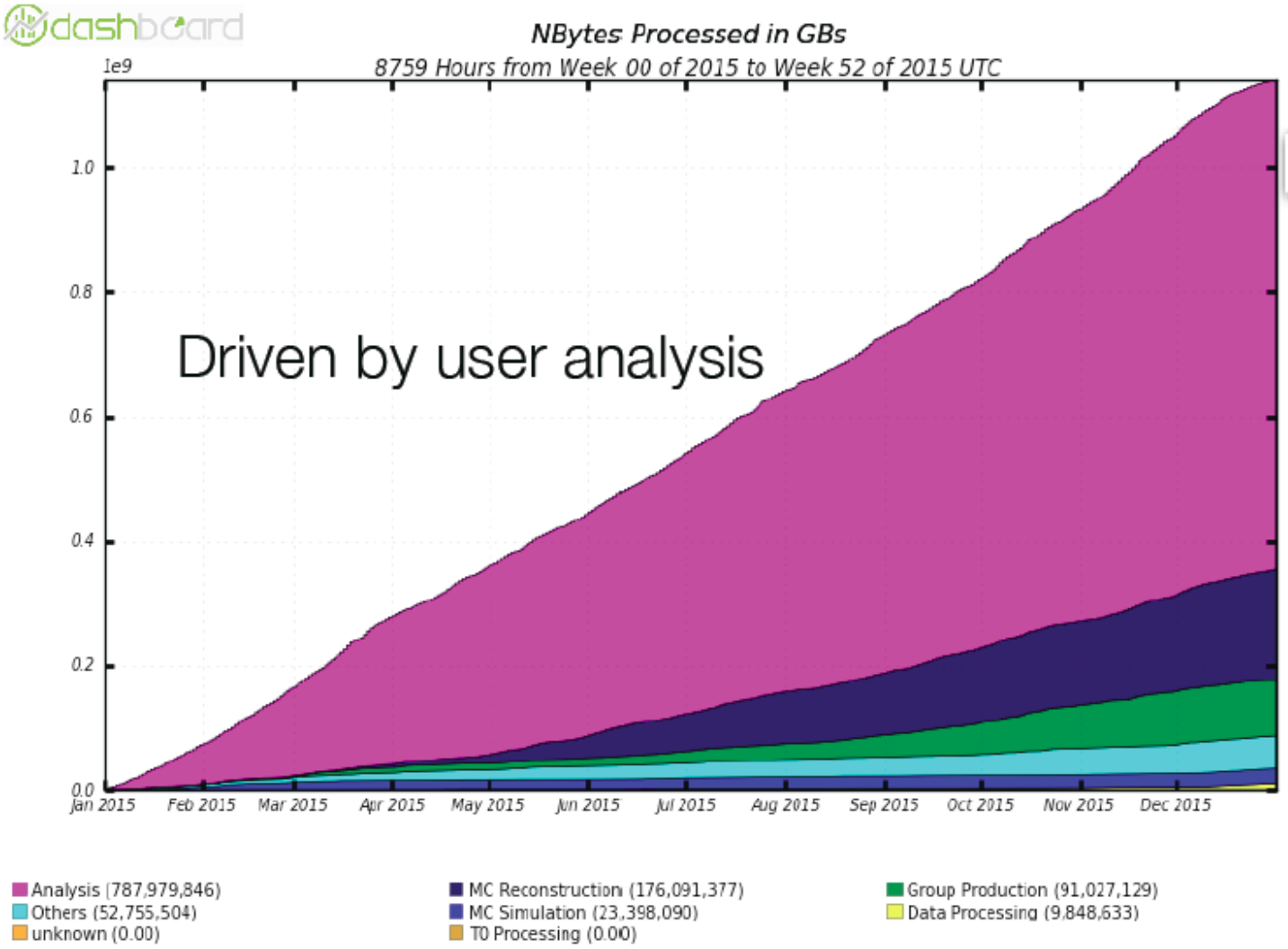


サイトの使われ方


Destination Tier Category
3
2
1
0
0%
1%
1%
0%
3%
28%
15%
0%
7%
22%
15%
0%
5%
2%
1%
0%
0
1
2
3
Origin Tier Category

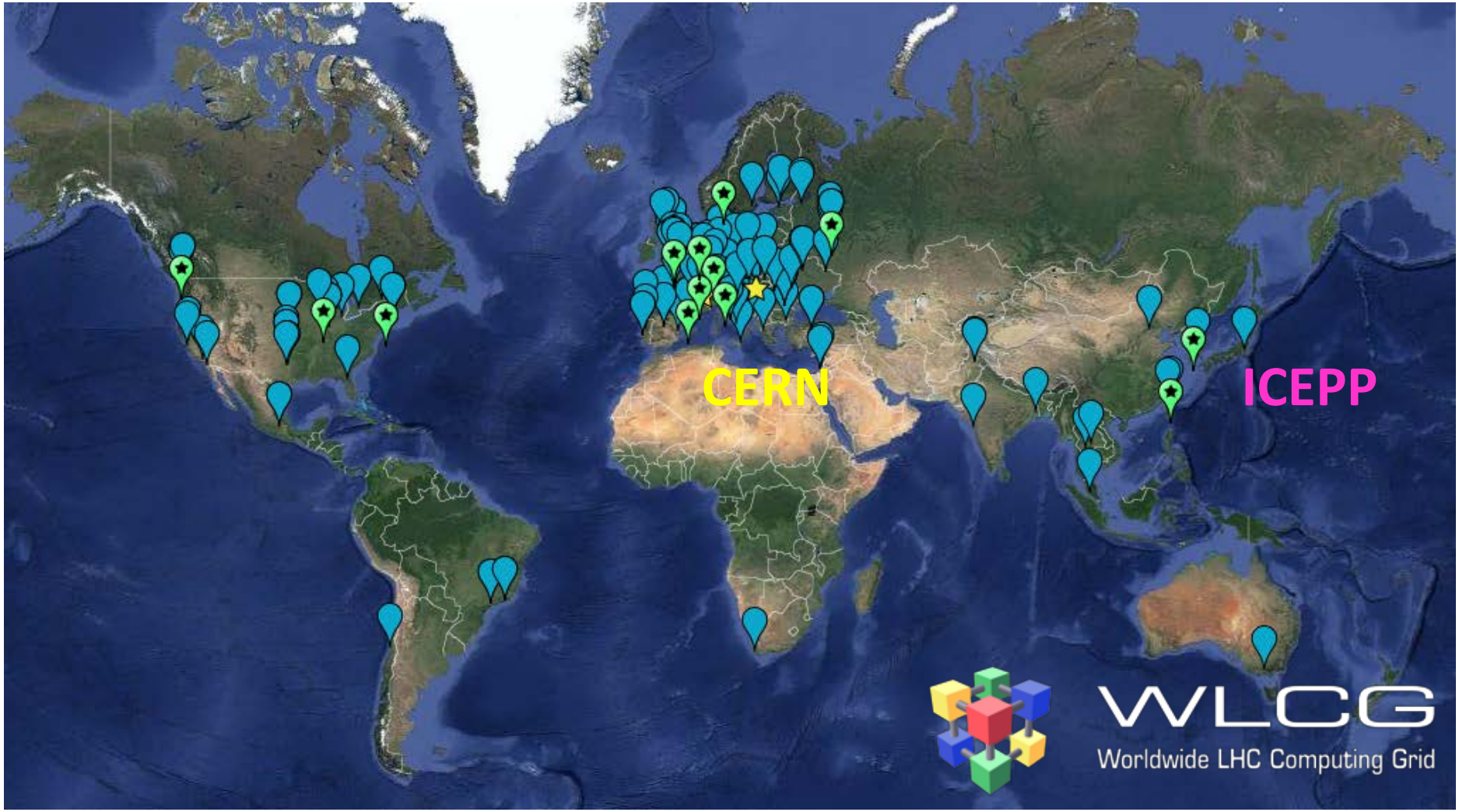

4,000
3,000
2,000
1,000
0
125%
100%
75%
50%
25%
0%
SRM [TB]
Cumulative % of SRM
1 5 9 13 17 21 25 29 33 37 41 45 49 53 57 61 65 69 73
SRM end point number

ATLASの処理したデータ量



1 Exabyte

ICEPPサイト (TOKYO-LCG2)



40カ国，170サイト

ICEPP地域解析センター（TOKYO-Tier2）

Worker nodes

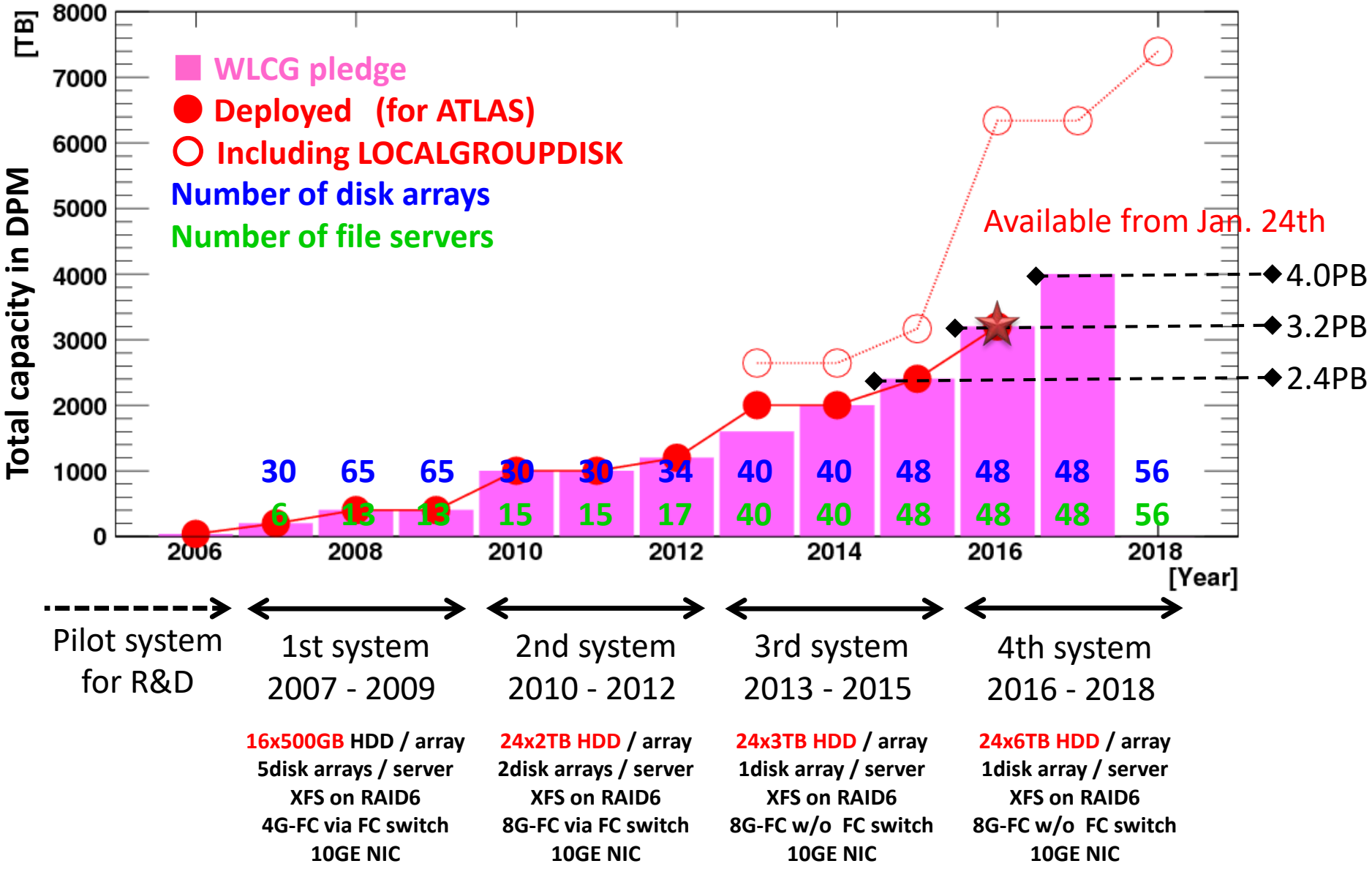
Disk arrays

システム構成)

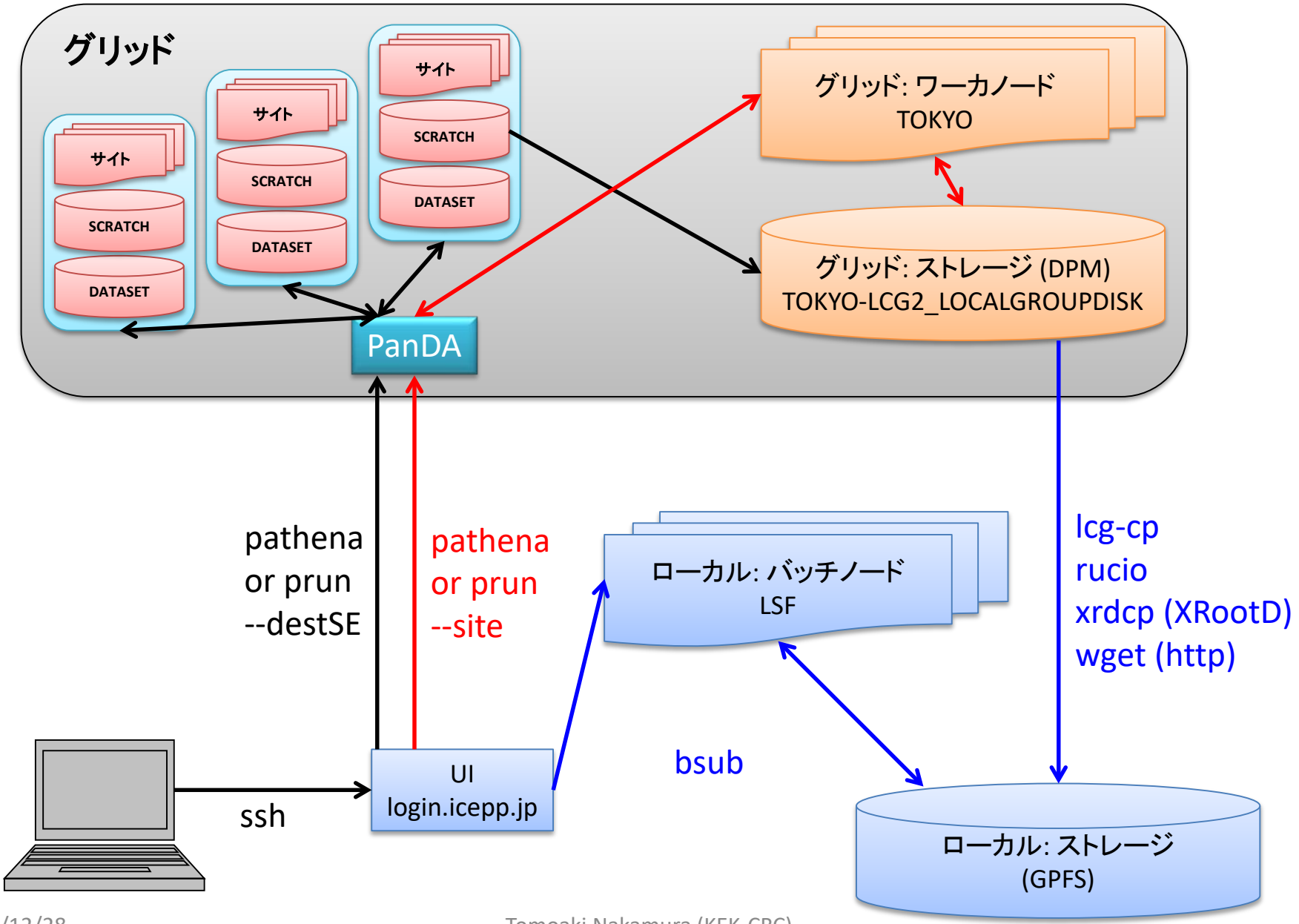
2016年1月末からTier2として本格再稼働

		3rd system (2013-2015)	4th system (2016-2018)
Computing node	Total	Node: **624 nodes, 9984 cores** (including service nodes) CPU: Intel Xeon E5-2680 (Sandy Bridge 2.7GHz, 8cores/CPU)	Node: **416 nodes, 9984 cores** (including service nodes) CPU: Intel Xeon E5-2680 v3 (Haswell 2.5GHz, 12cores/CPU)
	Tier2 pledge 2016 28 kHS06 pledge 2017 32 kHS06	Node: **160 nodes, 2560 cores** Memory: 32GB/node, 64GB/node NIC: 10Gbps/node Network BW: 80Gbps/16 nodes Disk: 600GB SAS x 2	Node: **256 nodes, 6144 cores** Memory: 64GB/node (2.66GB/job slots) NIC: 10Gbps/node Network BW: 80Gbps/16 nodes Disk: 1.2TB SAS x 2
Disk storage	Total	Capacity: **6732TB** (RAID6) Disk Array: 102 (3TB x 24) File Server: 102 nodes (1U) FC: 8Gbps/Disk, 8Gbps/FS	Capacity: **10560TB** (RAID6) + α Disk Array: 80 (6TB x 24) File Server: 80 nodes (1U) FC: 8Gbps/Disk, 8Gbps/FS
	Tier2	DPM: **3.168PB**	DPM: **6.336PB (+1.056PB)**
Network bandwidth	LAN	10GE ports in switch: 352 Switch inter link : 160Gbps	10GE ports in switch: 352 Switch inter link : 160Gbps
	WAN	ICEPP-UTNET: **10Gbps** SINET-USA: 10Gbps x 3 ICEPP-EU: 10Gbps (+10Gbps)	ICEPP-UTNET: **20Gbps (+20Gbps)** SINET-USA: 100Gbps + 10Gbps ICEPP-EU: 20Gbps (+20Gbps)

Disk容量（TOKYO-Tier2）

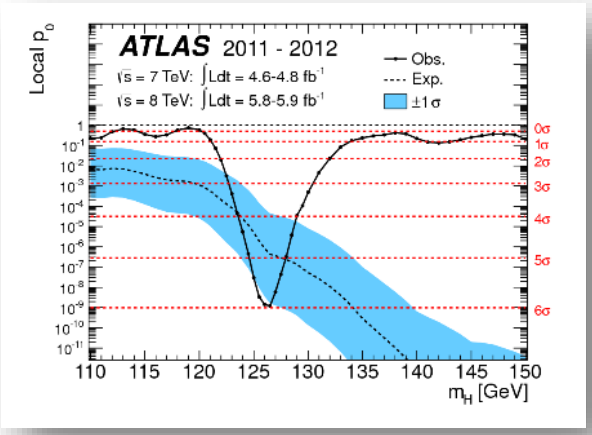


End user analysis workflow



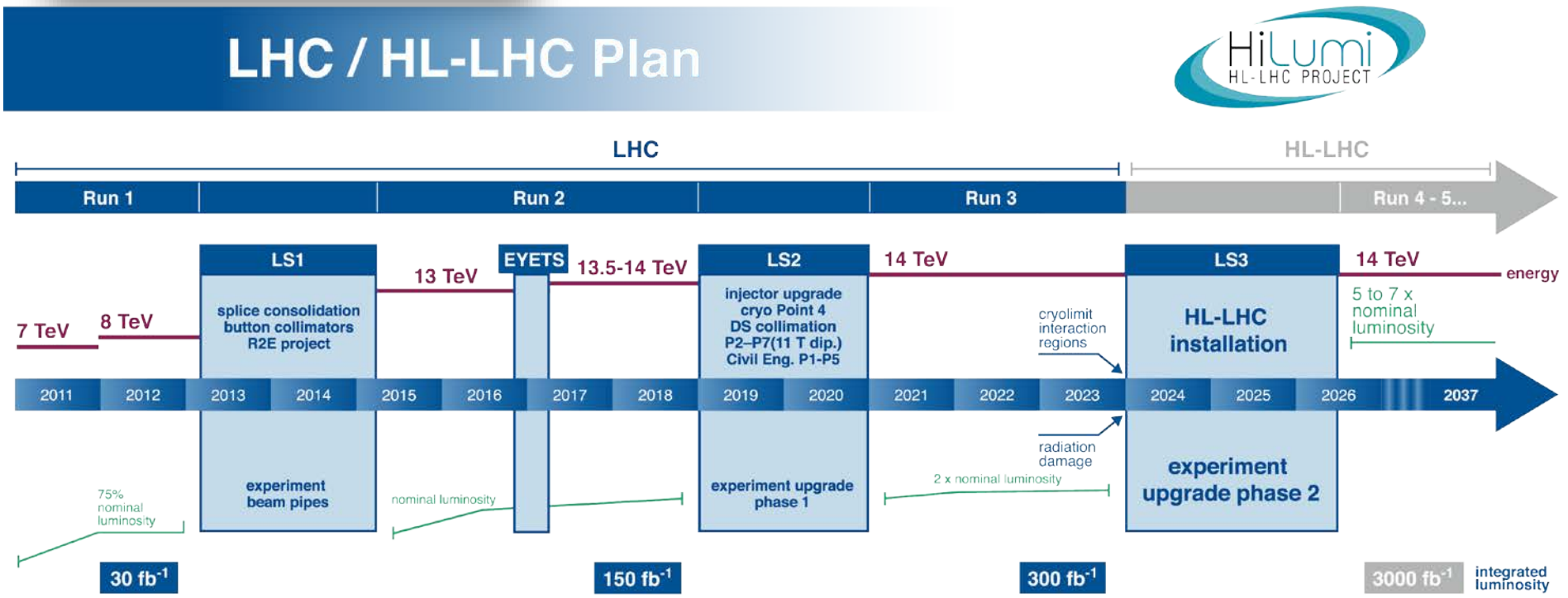
to HL-LHC

LHC-Run1で発見されたヒッグス粒子

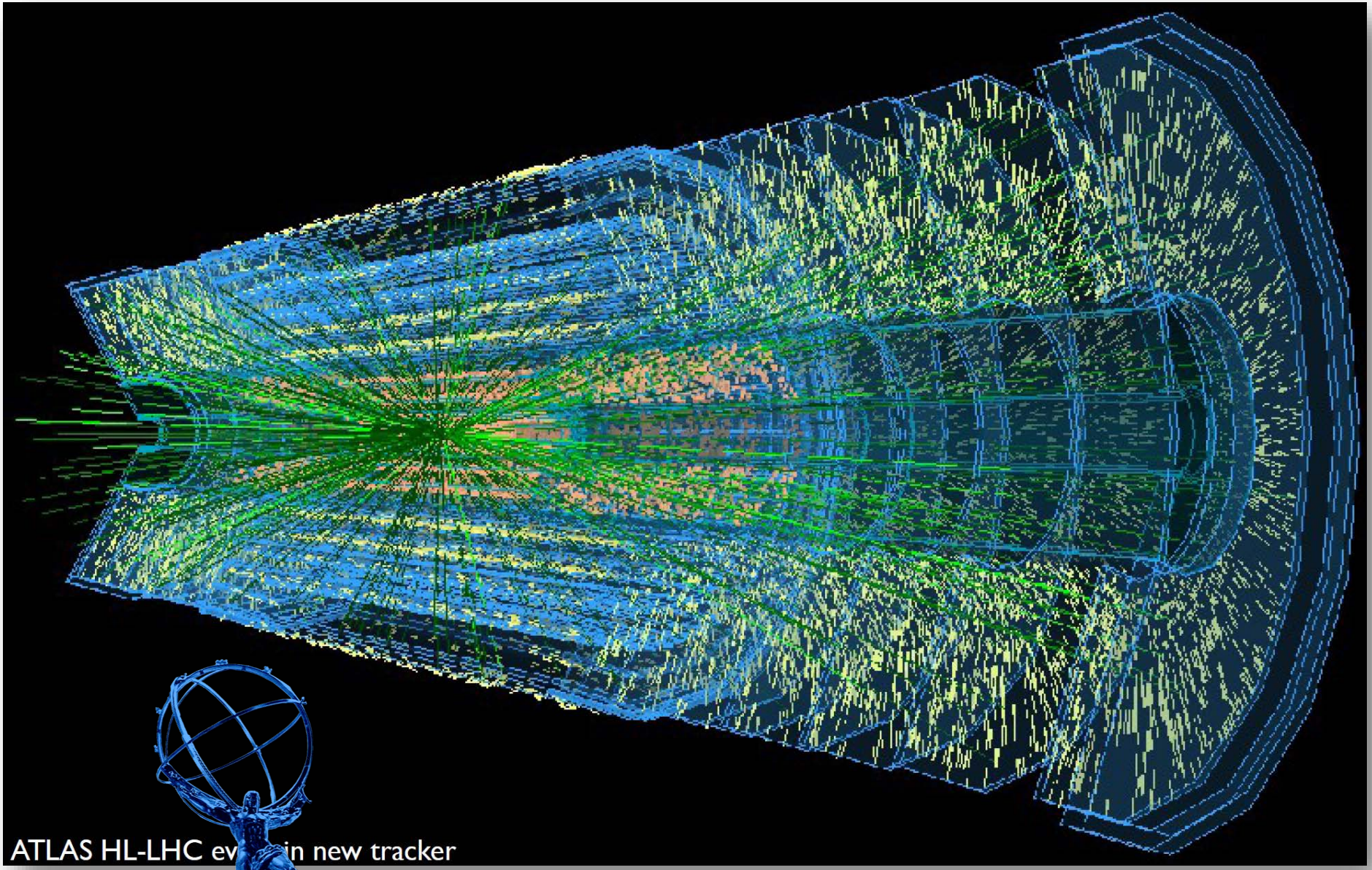


これから20年間でRun1の100倍の実験データを取得

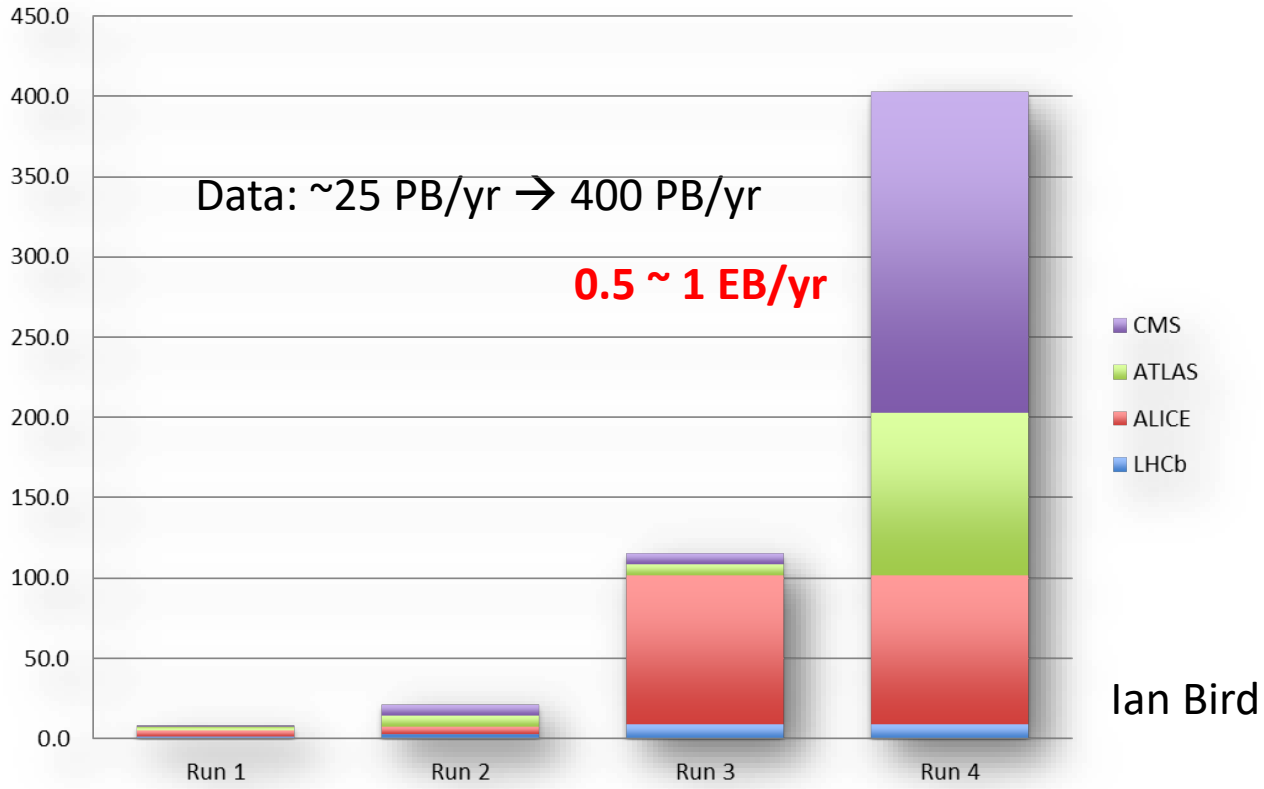
- ヒッグス粒子の詳細な性質を明らかにする
- 標準理論を超える物理（新粒子）を探索する



HL-LHC (150-200パイルアップ)



Data volume in future



Ian Bird

	2010-2012	2015-2018	2020-2022	2025-2035
Integrated Luminosity	25 fm^{-1}	50 fm^{-1}	300 fm^{-1}	3000 fm^{-1}
Collision Energy	7-8TeV	13-14TeV	14TeV	14TeV

ATLAS 400Hz (400MB/s)

ATLAS 1kHz (1-1.5GB/s)

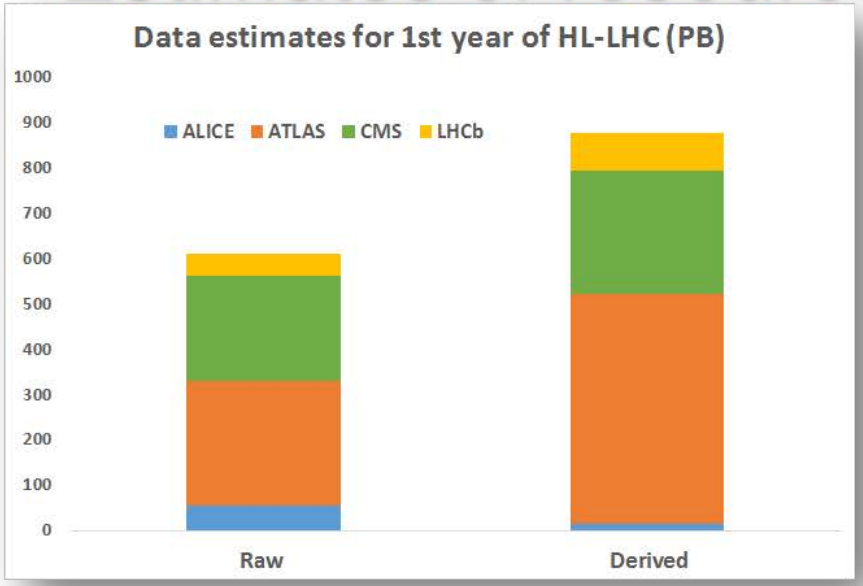
LHCb 20kHz (2GB/s)

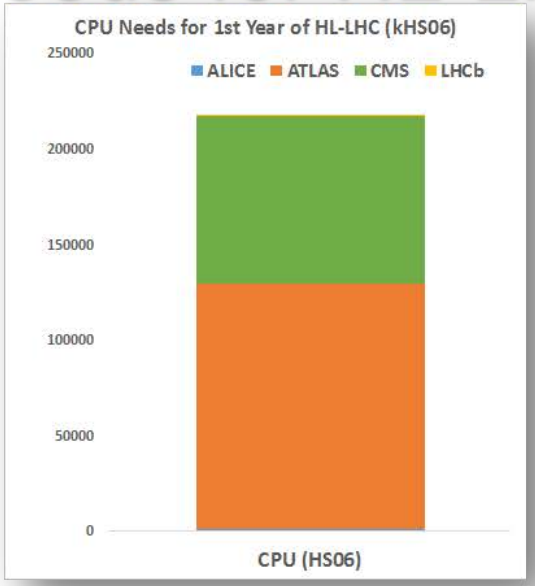
ALICE 50kHz (75GB/s)

ATLAS 5-10kHz (10-20GB/s)

CMS 10kHz (40GB/s)

Estimates of resource needs for HL-LHC





I. Bird

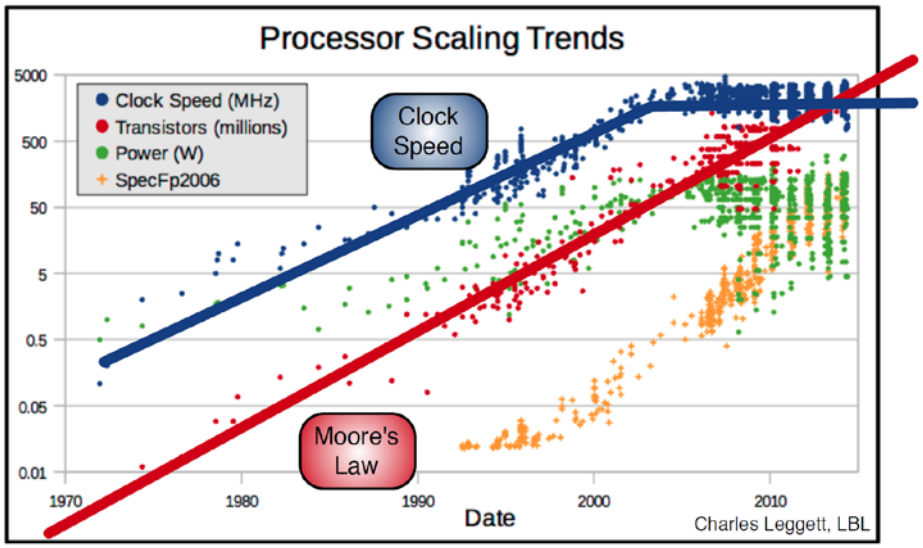
Data:
- Raw 2016: 50 PB → 2027: 600 PB
- Derived (1 copy): 2016: 80 PB → 2027: 900 PB

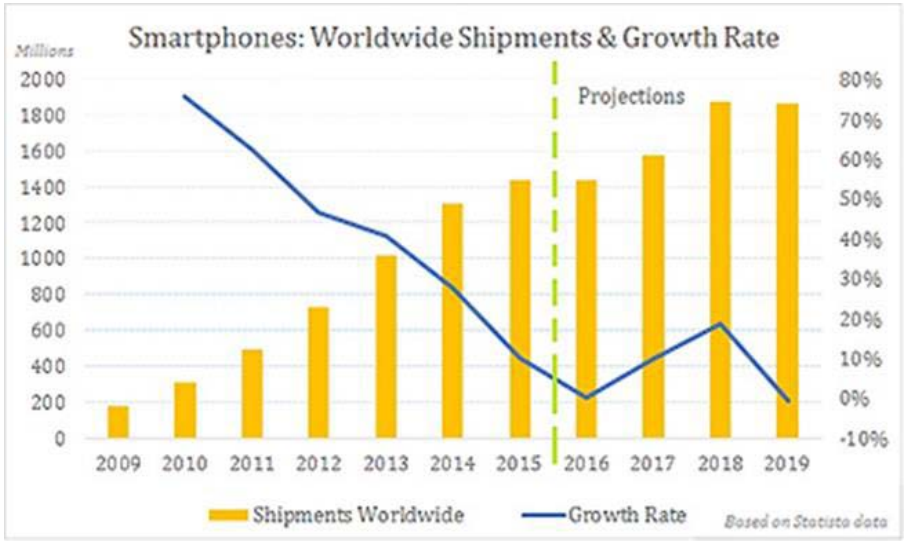
CPU:
- x60 from 2016

Technology at ~20%/year will bring x6-10 in 10-11 years

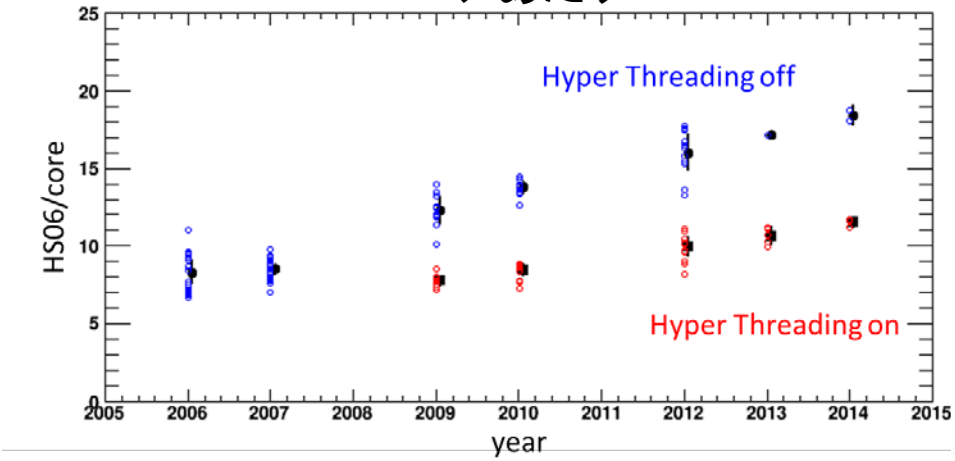
- Simple model based on today's computing models, but with expected HL-LHC operating parameters (pile-up, trigger rates, etc.)
- At least x10 above what is realistic to expect from technology with reasonably constant cost

Processor

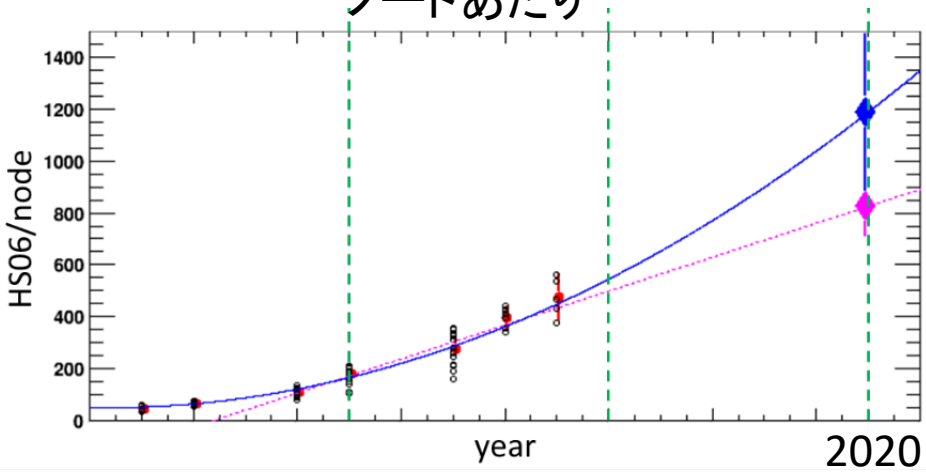




コアあたり

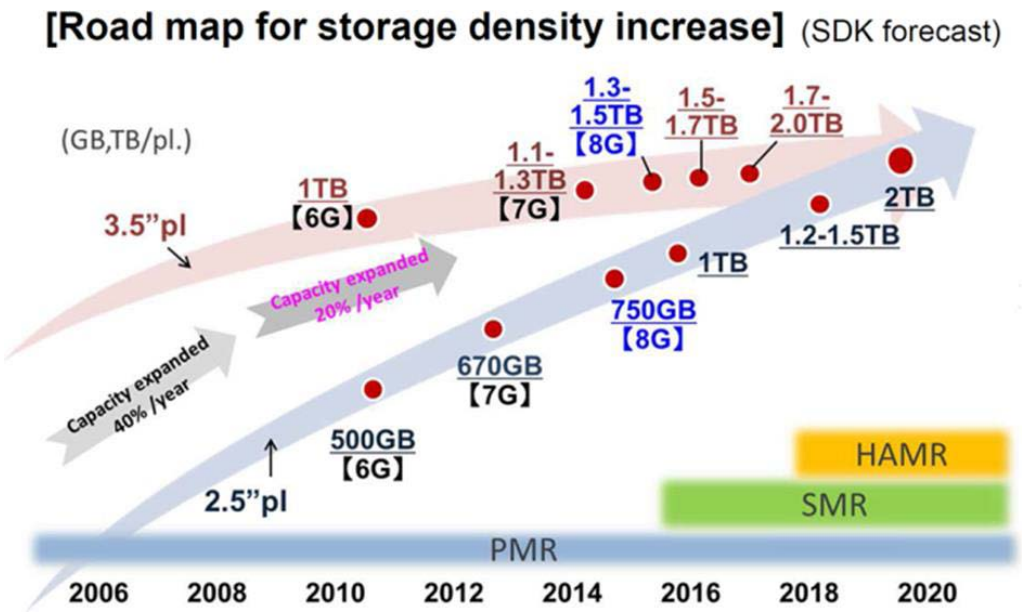


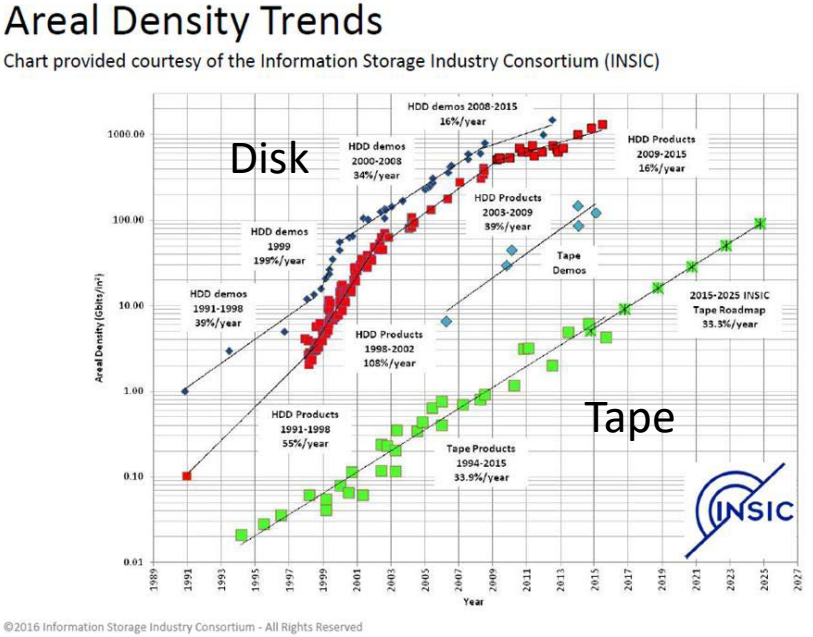
ノードあたり

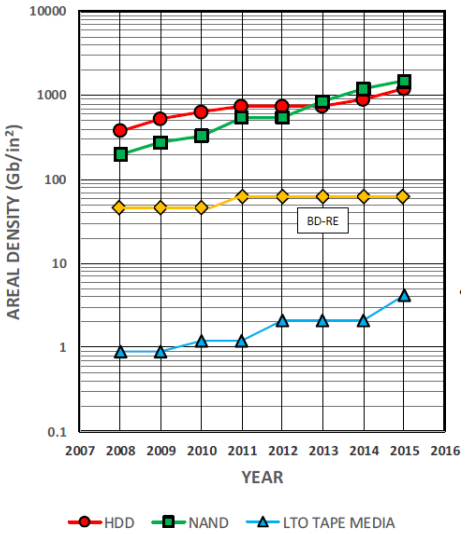


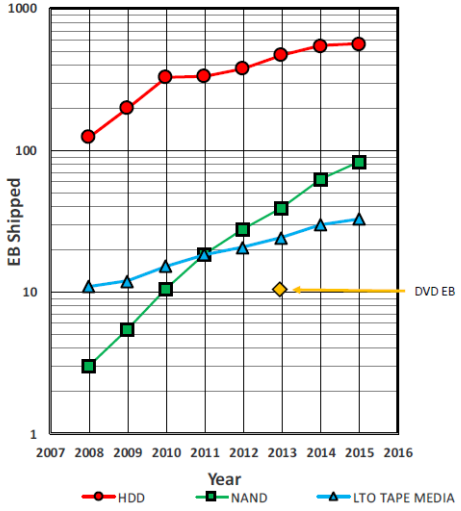
2% improve / year

Storage (HDD, SSD, Tape)







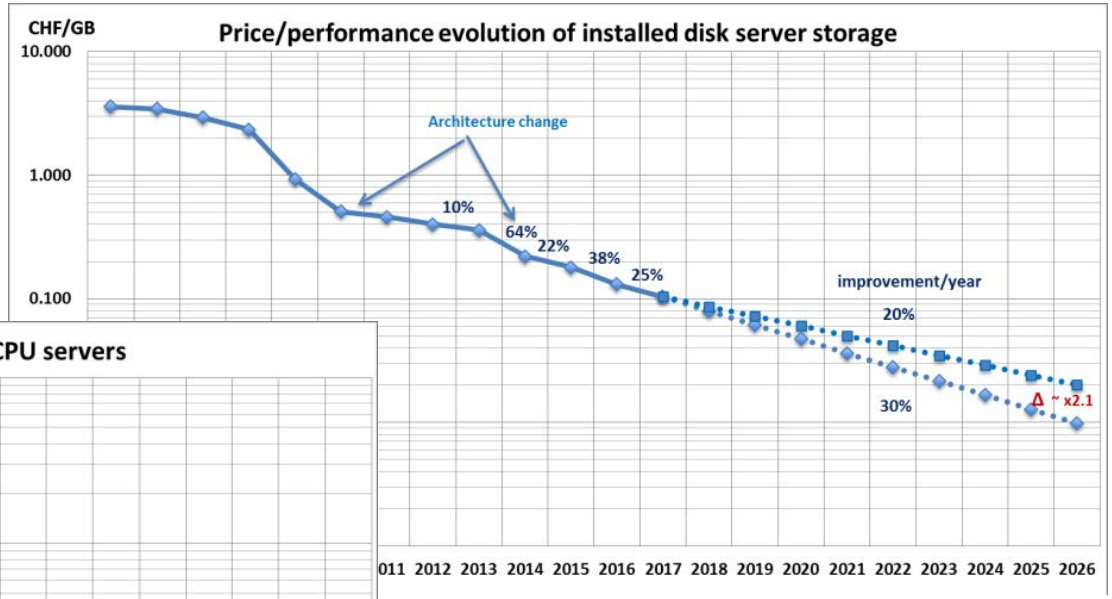


計算機技術の進歩によるコスト減

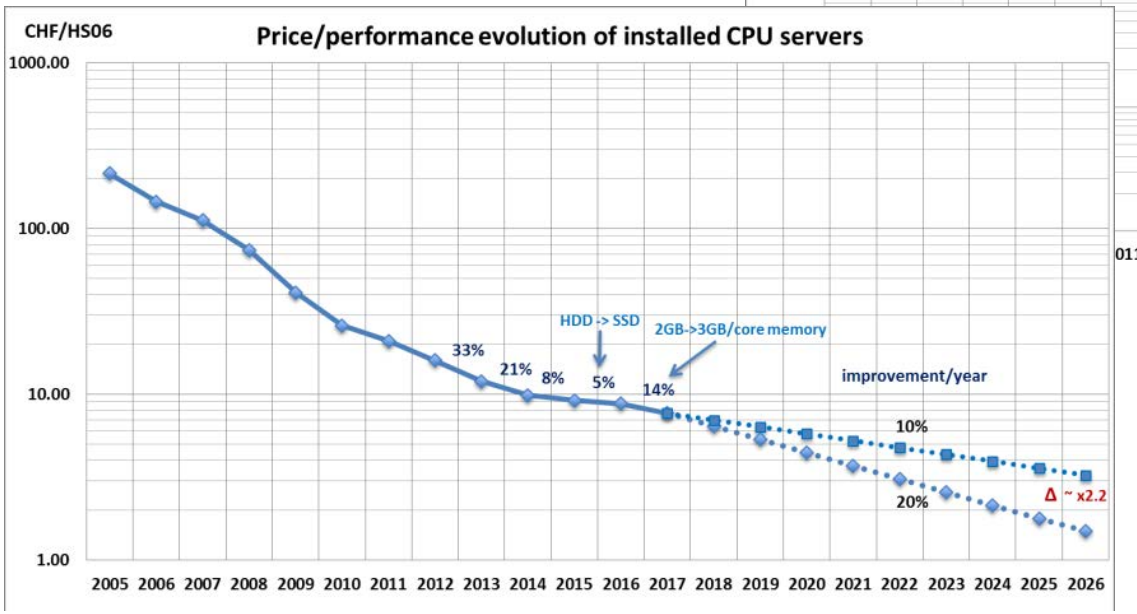
CERNの調達に基づく予測

H. Meinhard, B. Panzer-Steindel

Disk



CPU



他にも、
Memory, GPU, Network....

HL-LHC computing cost parameters