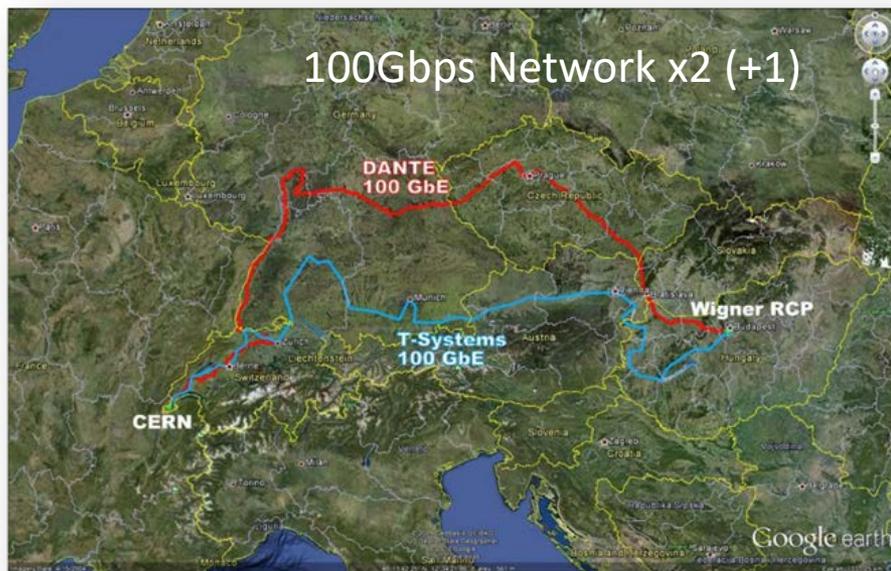

ATLASグリッド概要



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

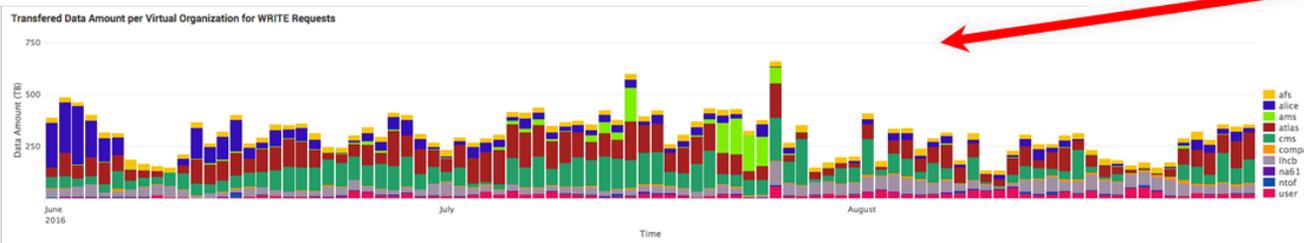
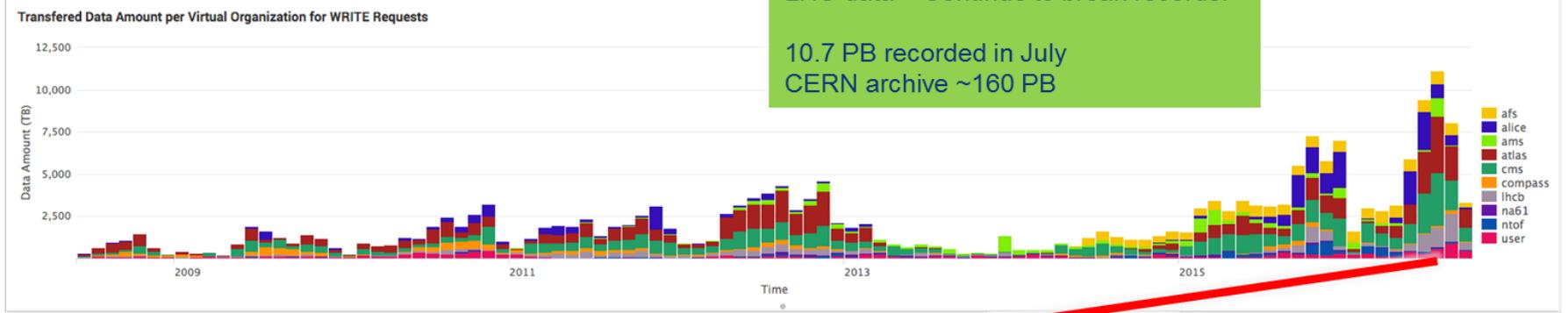
~100 PB disk



2016 data

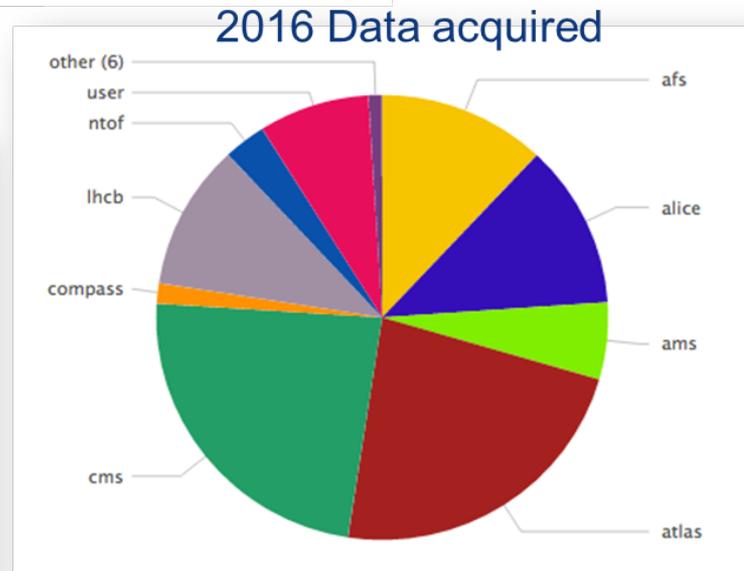
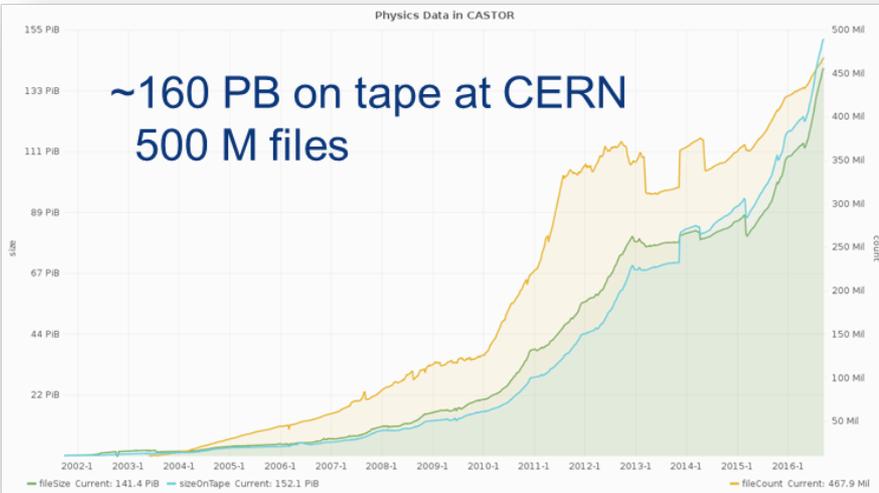
LHC data – Continue to break records:

10.7 PB recorded in July
CERN archive ~160 PB



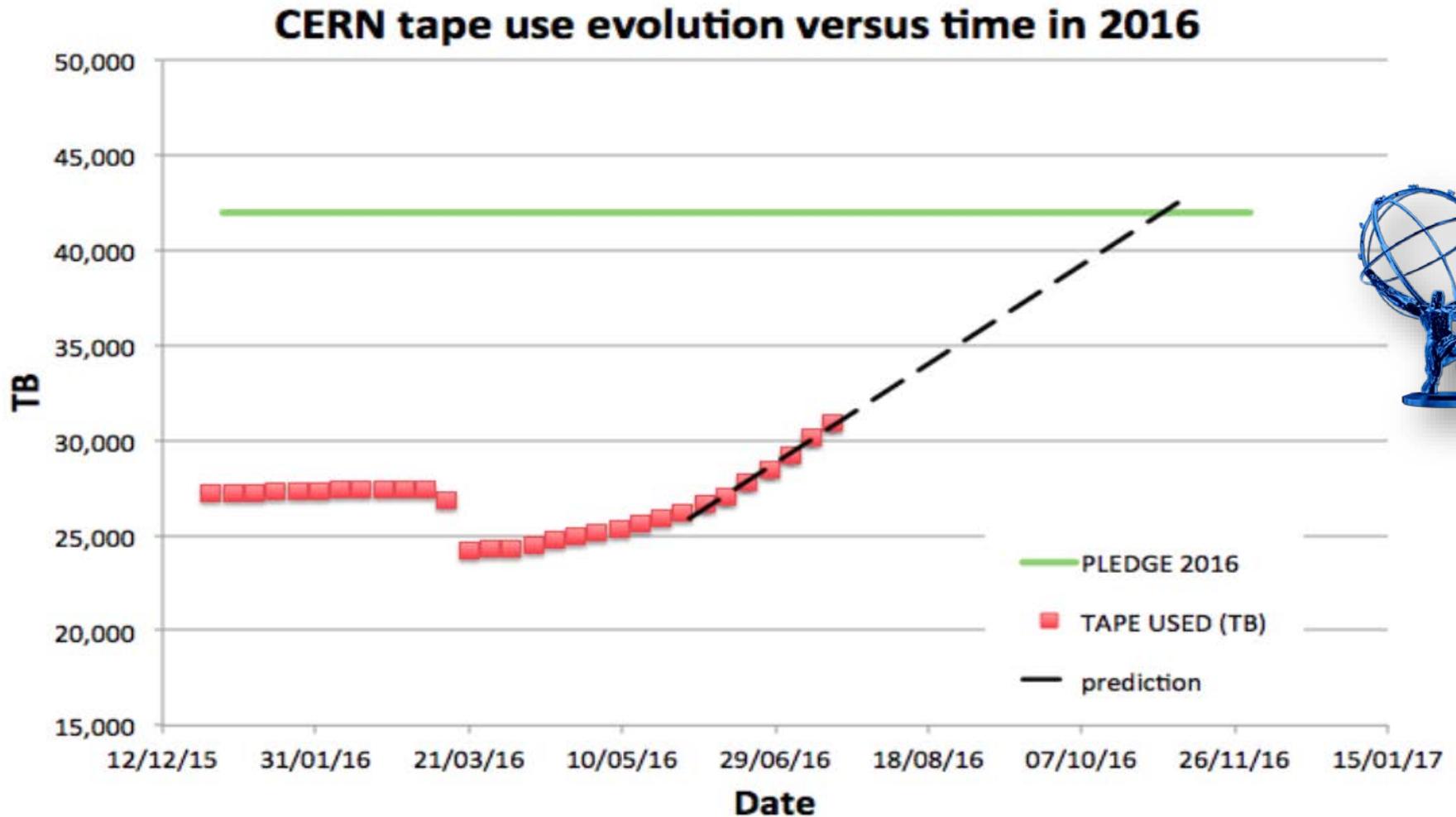
June-Aug 2016
>500 TB / day
(Run 1 peak for HI was 220 TB)

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

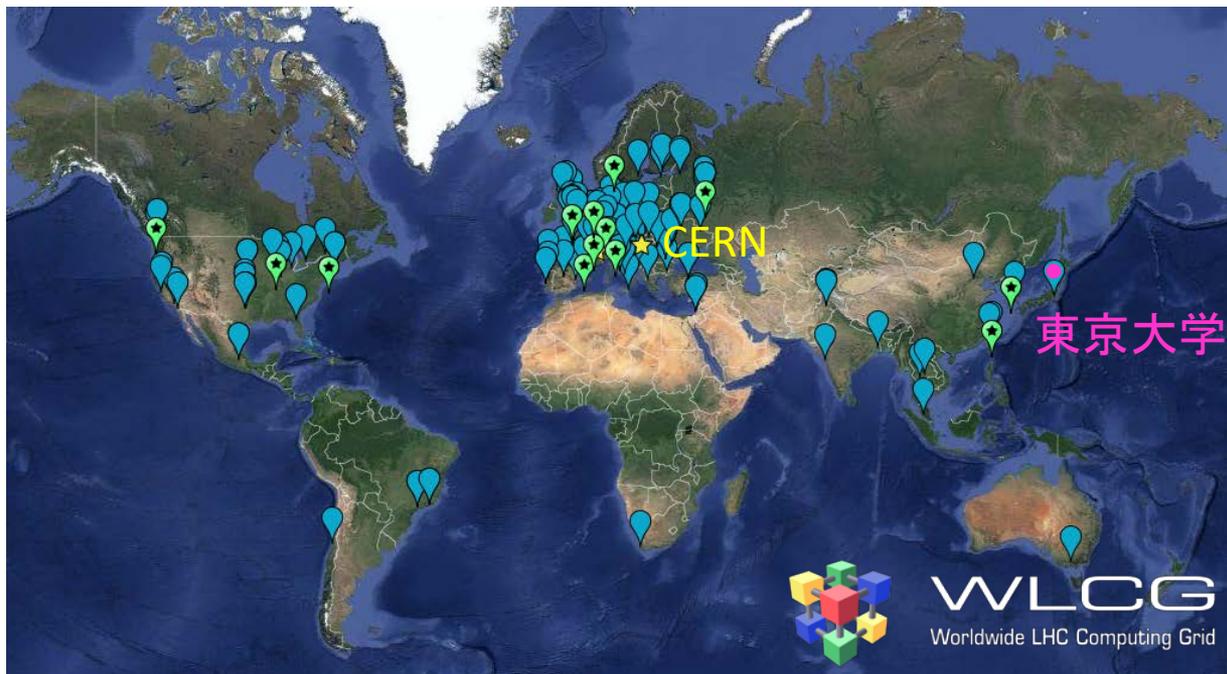


I. Bird

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)



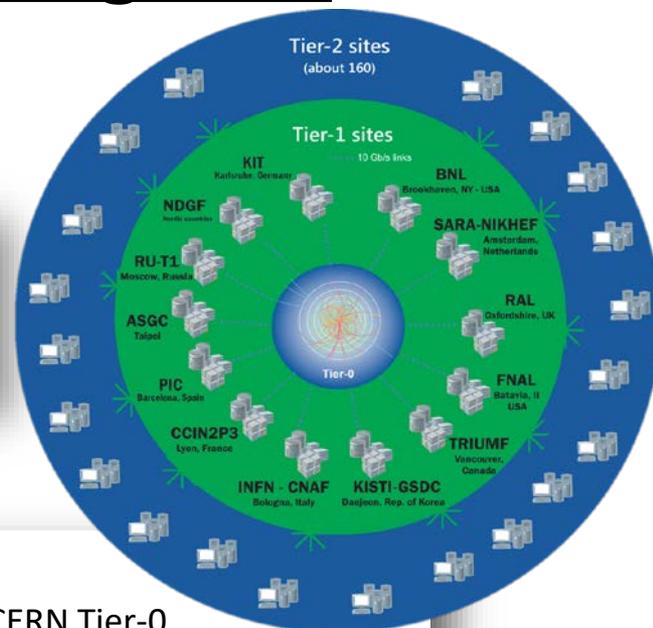
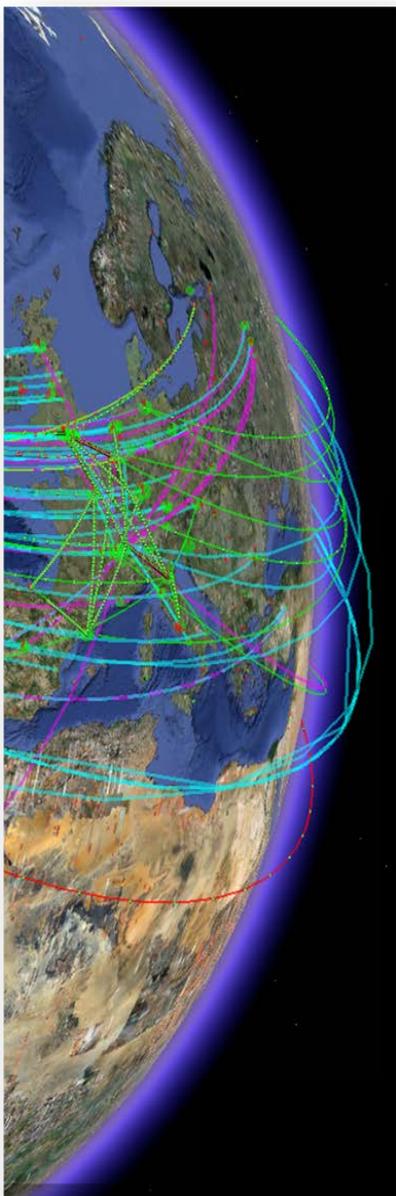
計算ノード



ディスクストレージ

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



ATLASグリッド

データアーカイブと再構成:
分散データアーカイブ:
MCプロダクションとデータ解析:

CERN Tier-0
10サイトのTier-1
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

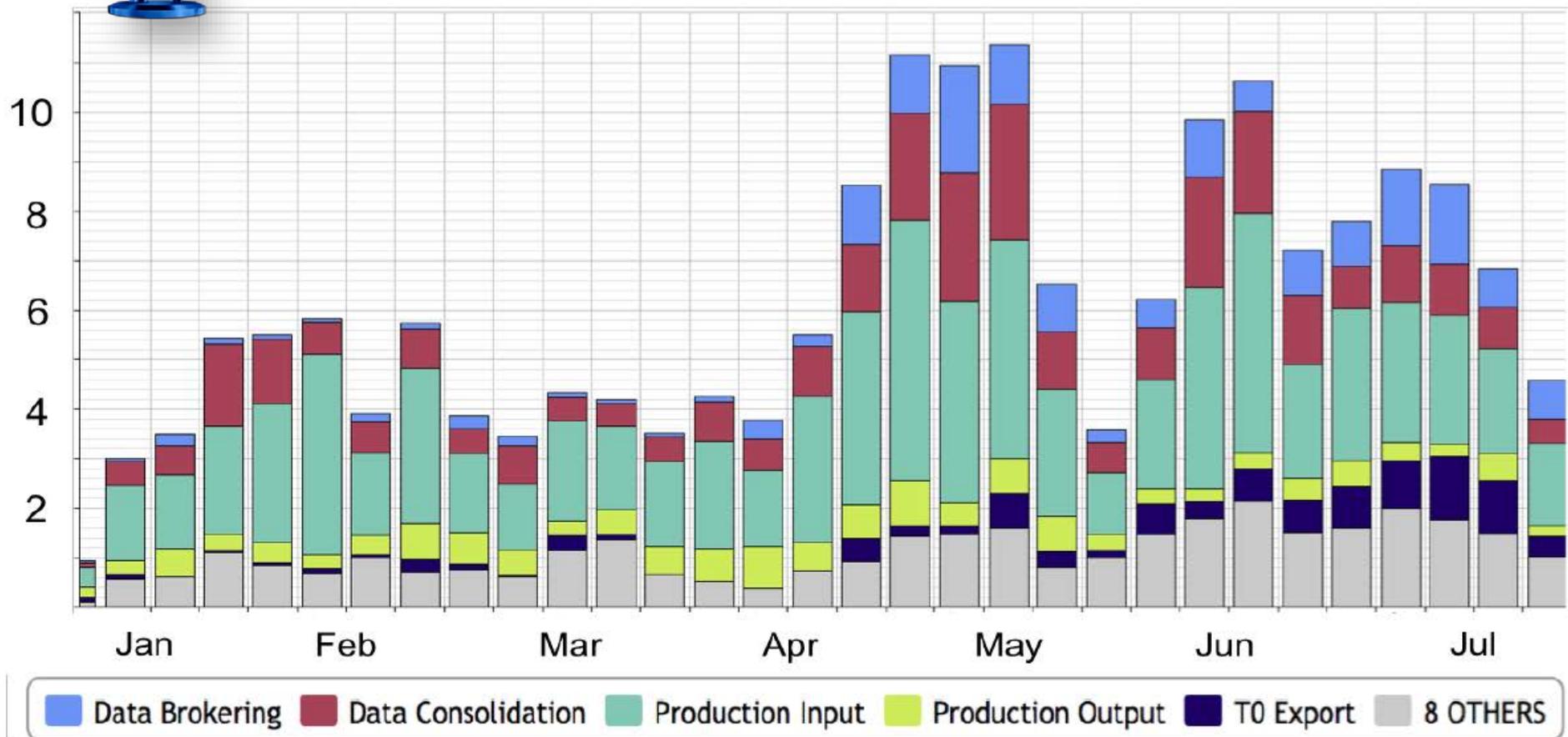
WLCG Collaboration



ATLASサイト間でのデータ転送



Data transfer volume per week by activity (PB)



旧データ分配とワークフロー

モナークモデル

Tier0

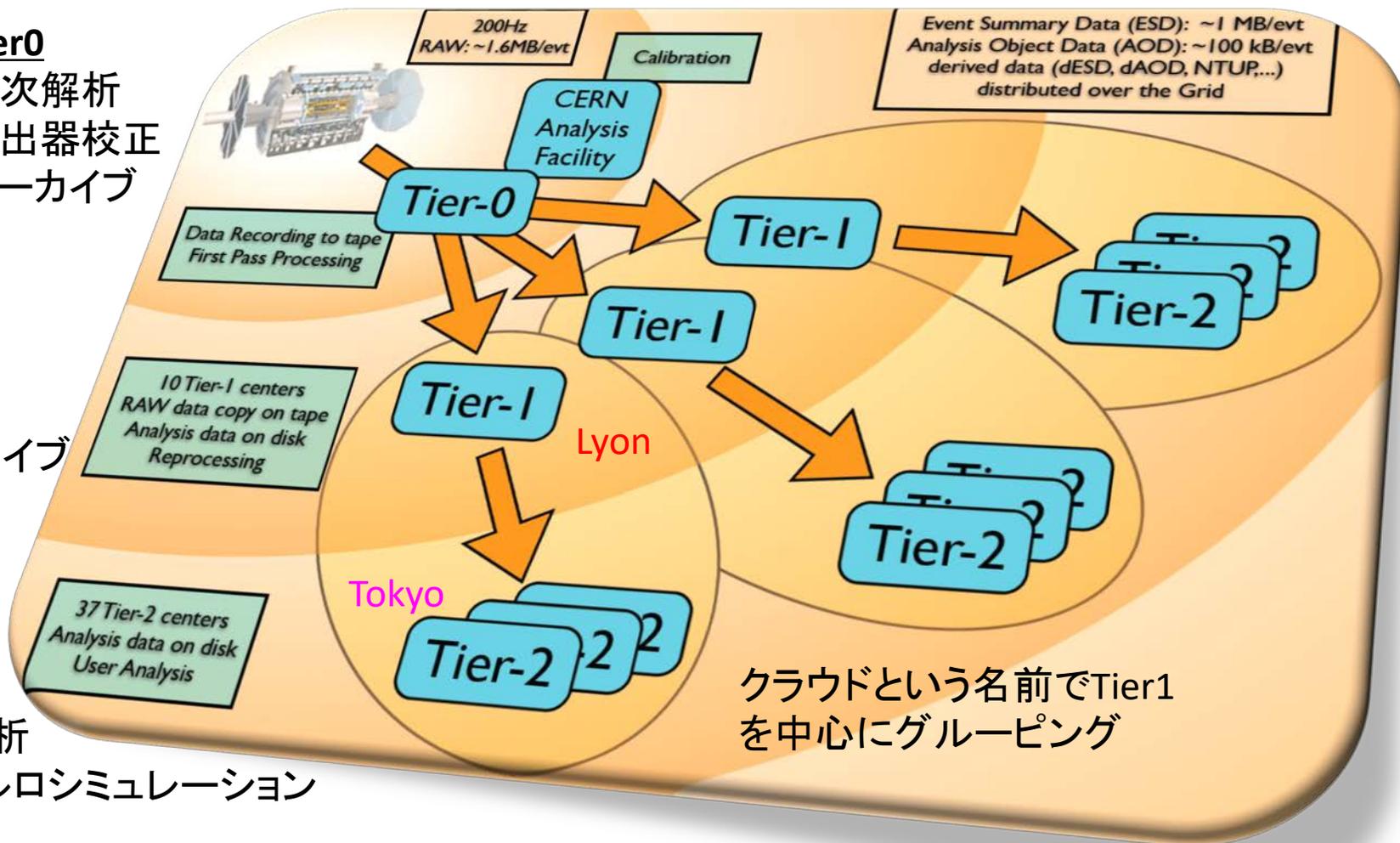
一次解析
検出器校正
アーカイブ

Tier1

再解析
分散アーカイブ

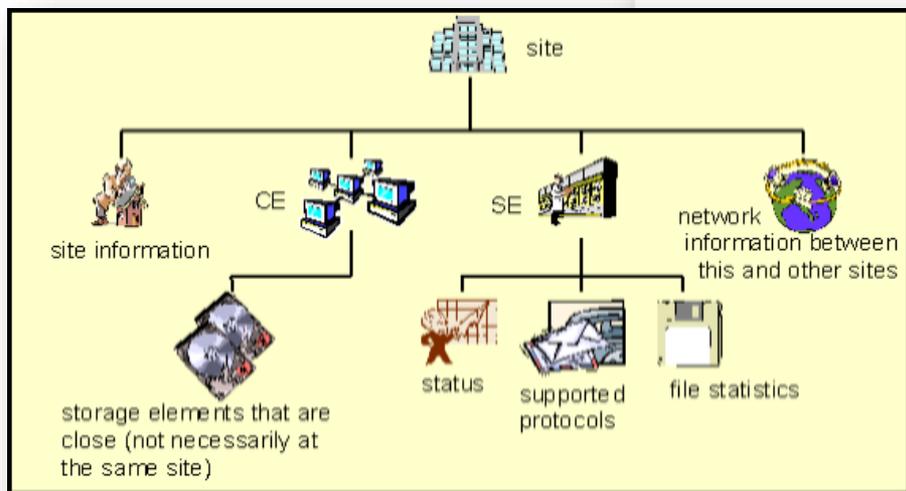
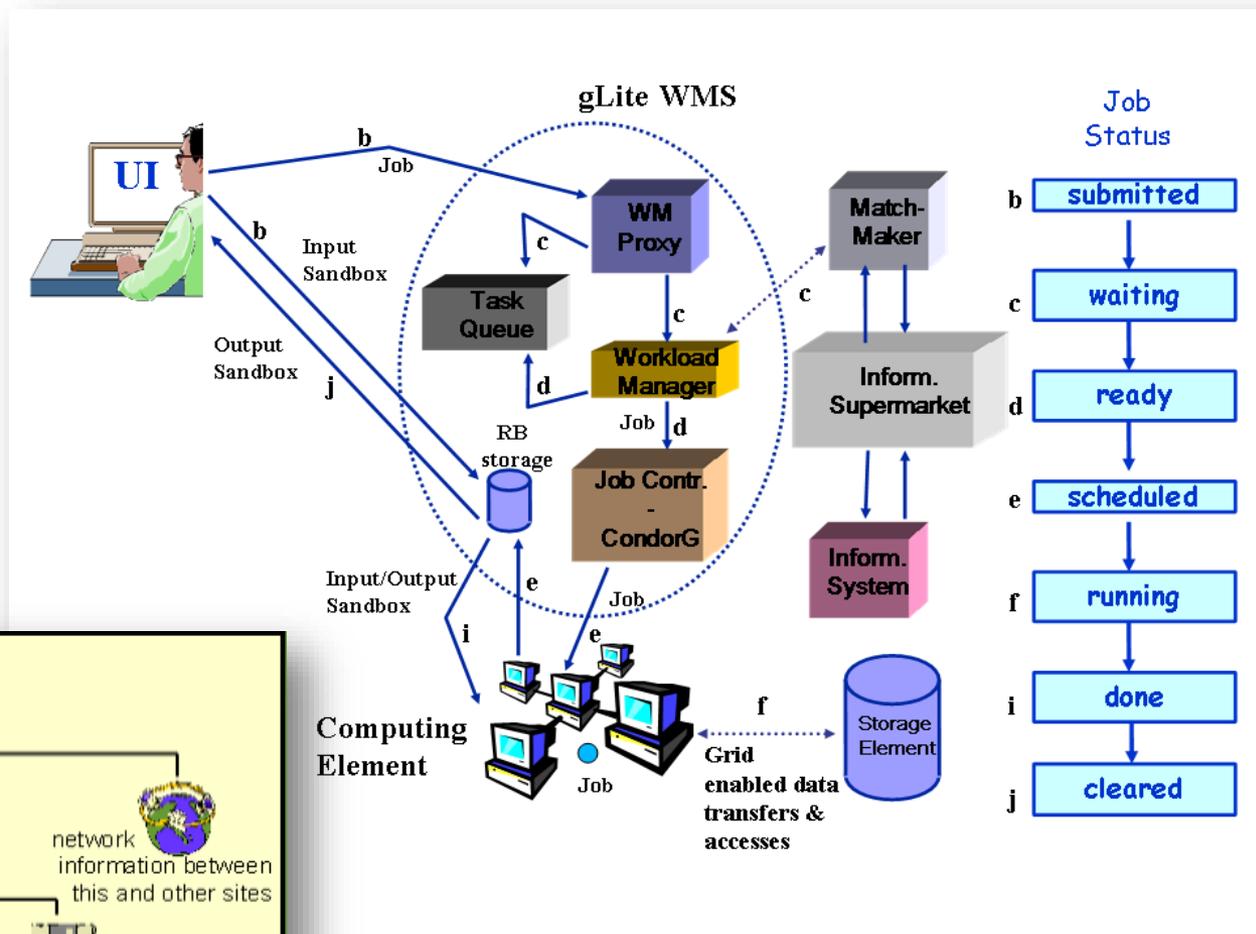
Tier2

ユーザ解析
モンテカルロシミュレーション

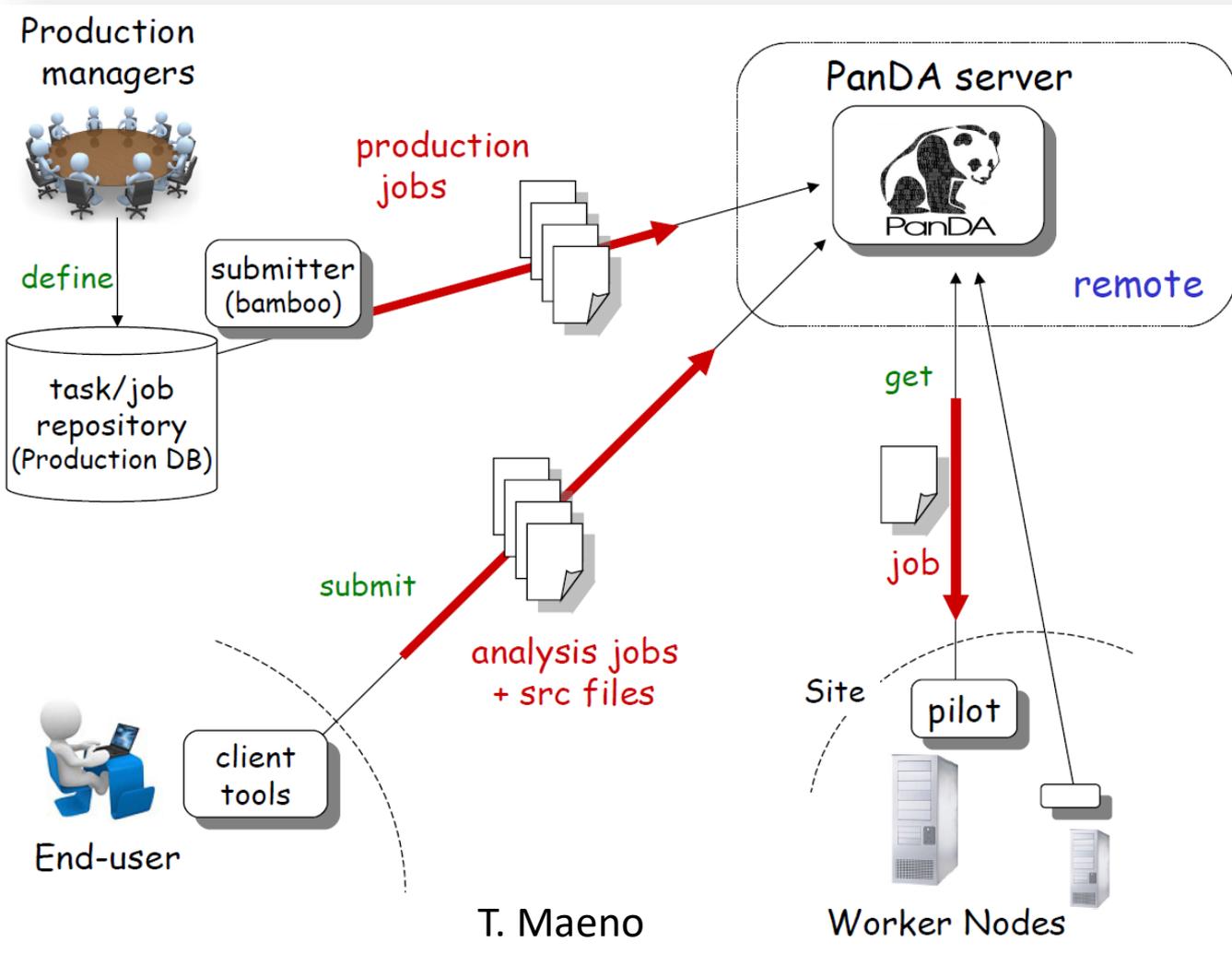


グリッド初期

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

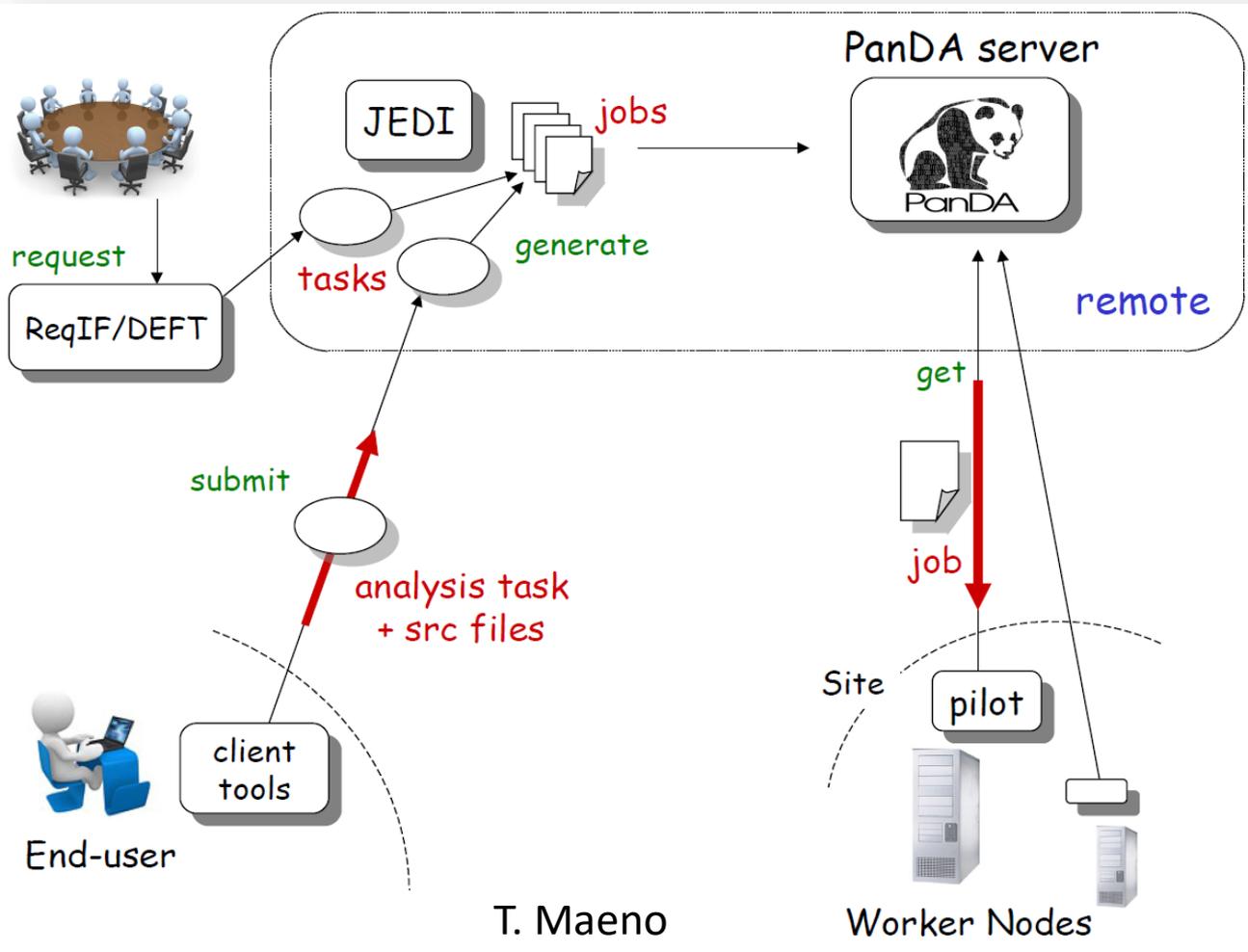
ユーザー解析:

入出力ファイルを意識して処理プログラムを投入

T. Maeno

Worker Nodes

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

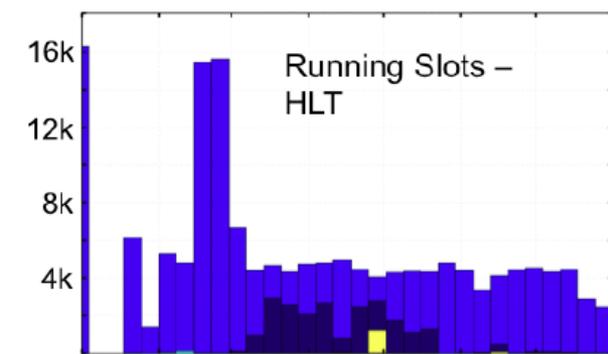
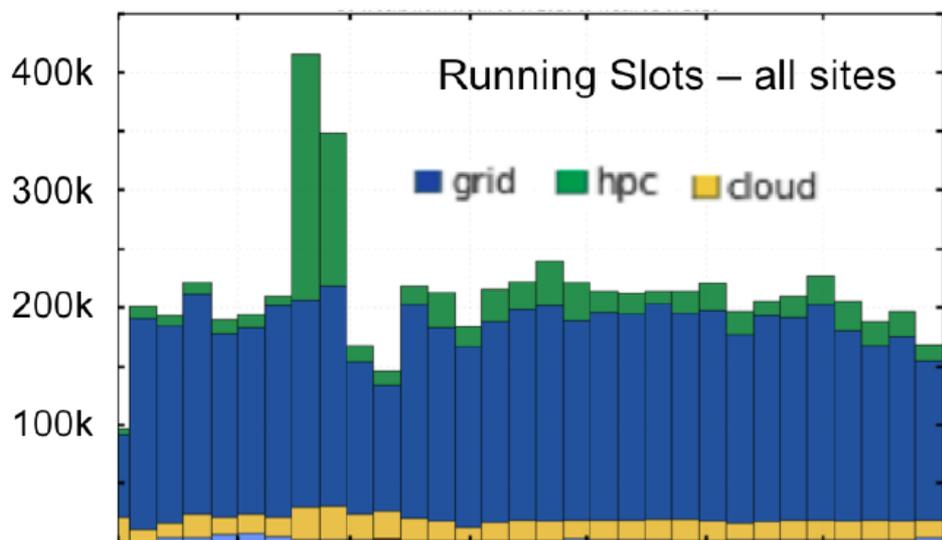
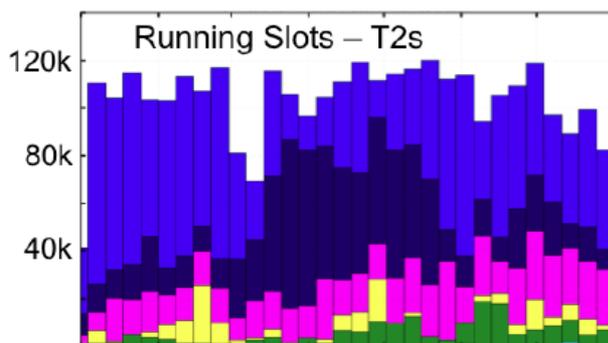
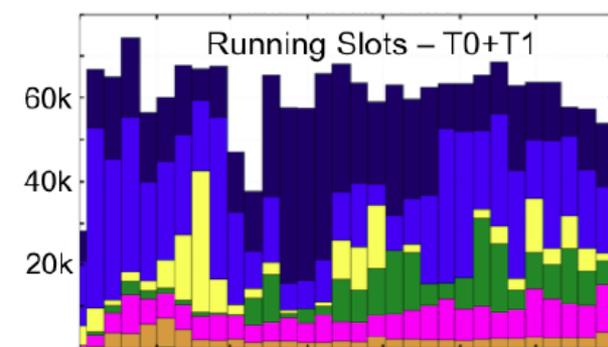
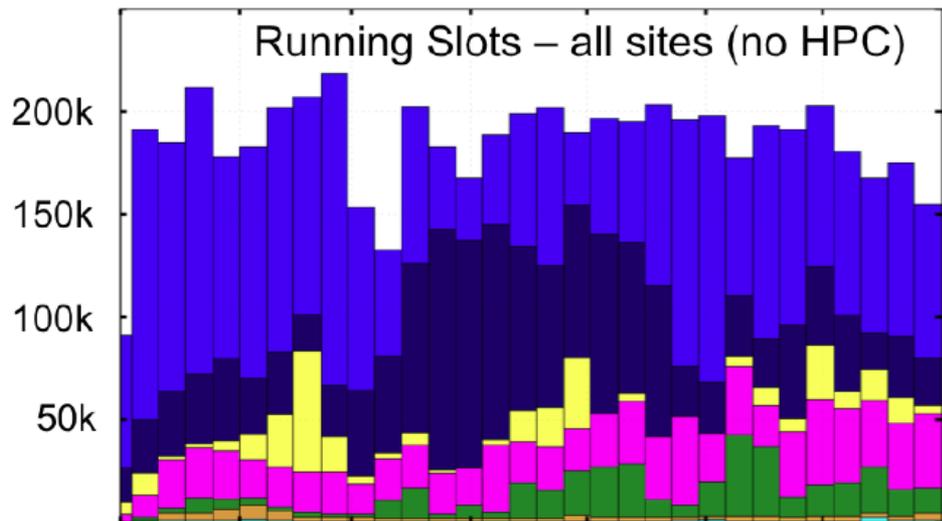
- ジョブ実行時間
- 入出力ファイルサイズ
- 使用可能なメモリサイズ
- パラレル数
- ネットワーク接続性

- 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



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

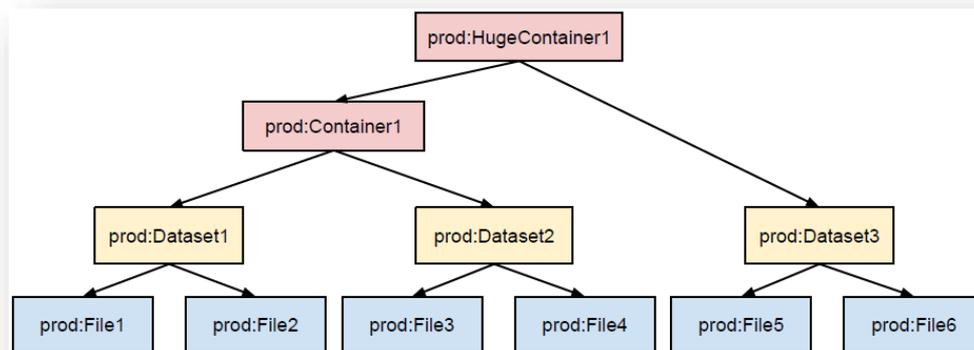
クォータ, パーミッション

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ア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._SF0-5._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 download --rse TOKYO-LCG2_DATADISK dids
```

```
# 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 request interface

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

ATLAS Rucio UI Monitoring Data Transfers (R2D2) Reports pattern OR name OR rule id Search Using account: tnakamur Other Monitoring Help

You are here: Rucio Rule Definition Droid - Request Rule Rucio Version: 1.4.2.post1

If you are new to this interface you might want to take the [tour](#).

If you find any errors or have suggestions for improvements for this interface please report it to [Jira](#).

Your input will be saved until you submit it. If you want to clear the form please click [here](#).

1. Select Data Identifiers (DIDs)

DID Pattern Search List of DIDs

Please start by entering a DID or DID wildcard and search for either containers or datasets. Then select the requested DIDs. Please do not use a trailing '/' for containers.

Data pattern Search Container Dataset

please provide a search pattern in the form: <scope>:<name|pattern> or <name|pattern>

Show Search:

Name
No data available in table
Name

Showing 0 to 0 of 0 entries Previous Next

[Continue](#) [Select All](#)

2. Select Rucio Storage Elements (RSEs)

3. Options

4. Summary

Data Identifiers and Scope

Files, datasets and containers share the same naming convention, which is composed of two strings: the scope and the name, separated by a colon. The combination of scope and name is called a data identifier (DID).

The scope is used to divide the name space into several, separate sub spaces for production and individual users. User scope always start with 'user.' followed by the account name.

By default users can read from all scopes but only write into their own one. Only privileged accounts have the right to write into multiple scopes including production scopes like mc15_13TeV.

Examples:

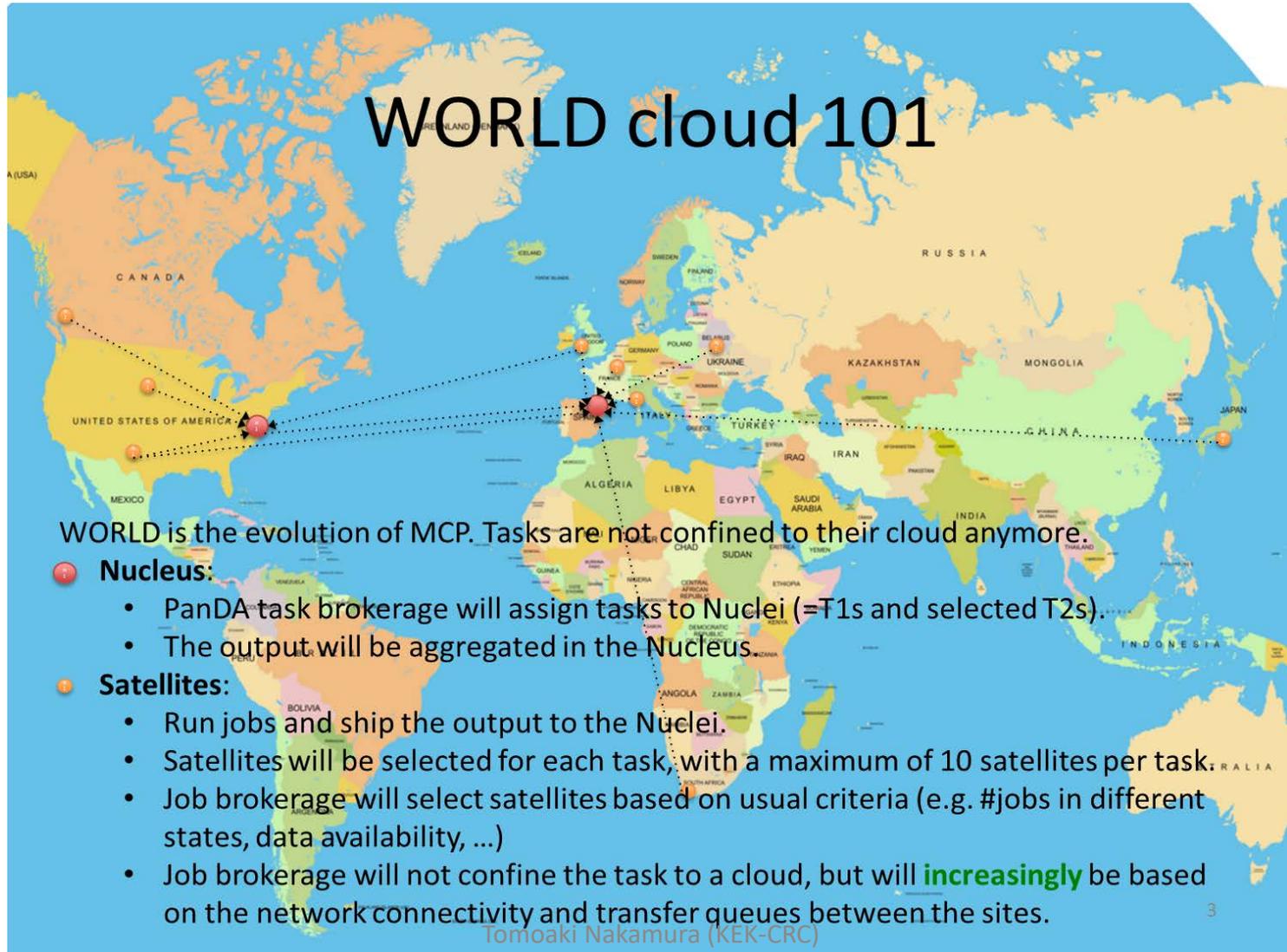
Official dataset:
data15_13TeV.00266904.physics_Main.
merge.DAOD_SUSY1.
f594_m1435_p2361_tid05608871_00

User dataset:
user.jdoe:my.dataset.1

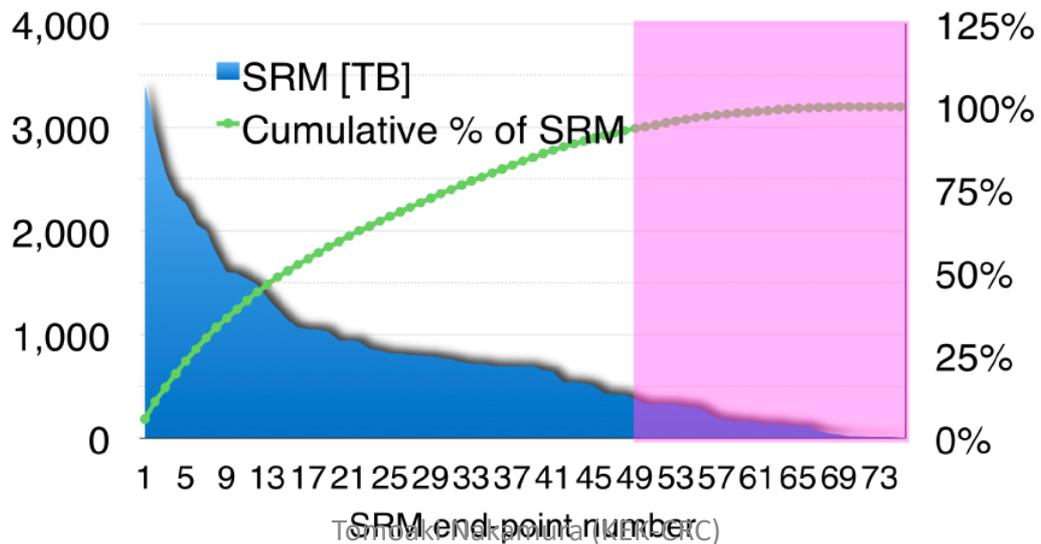
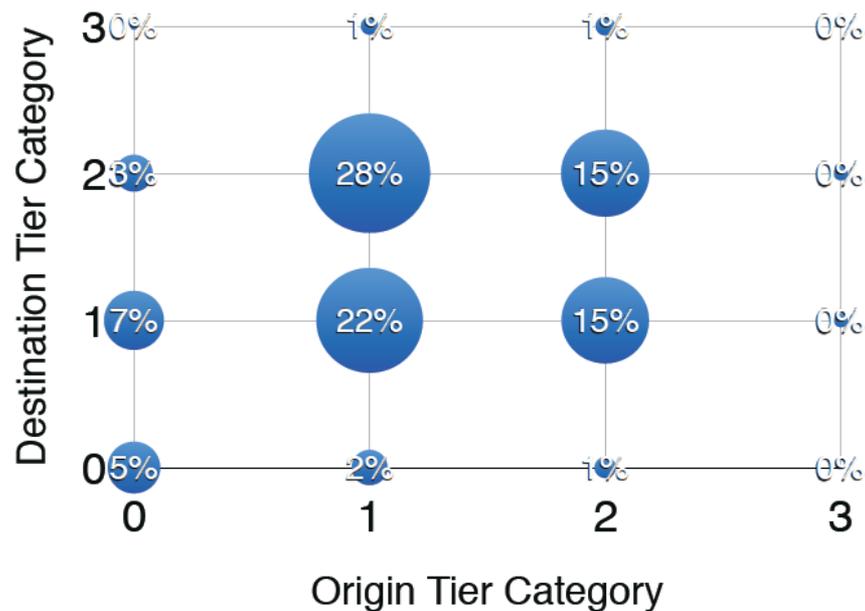
© 2012-2015 European Organisation for Nuclear Research (CERN) Tue Mar 29 2016 15:55:49 GMT+0900

現在 : Worldクラウド

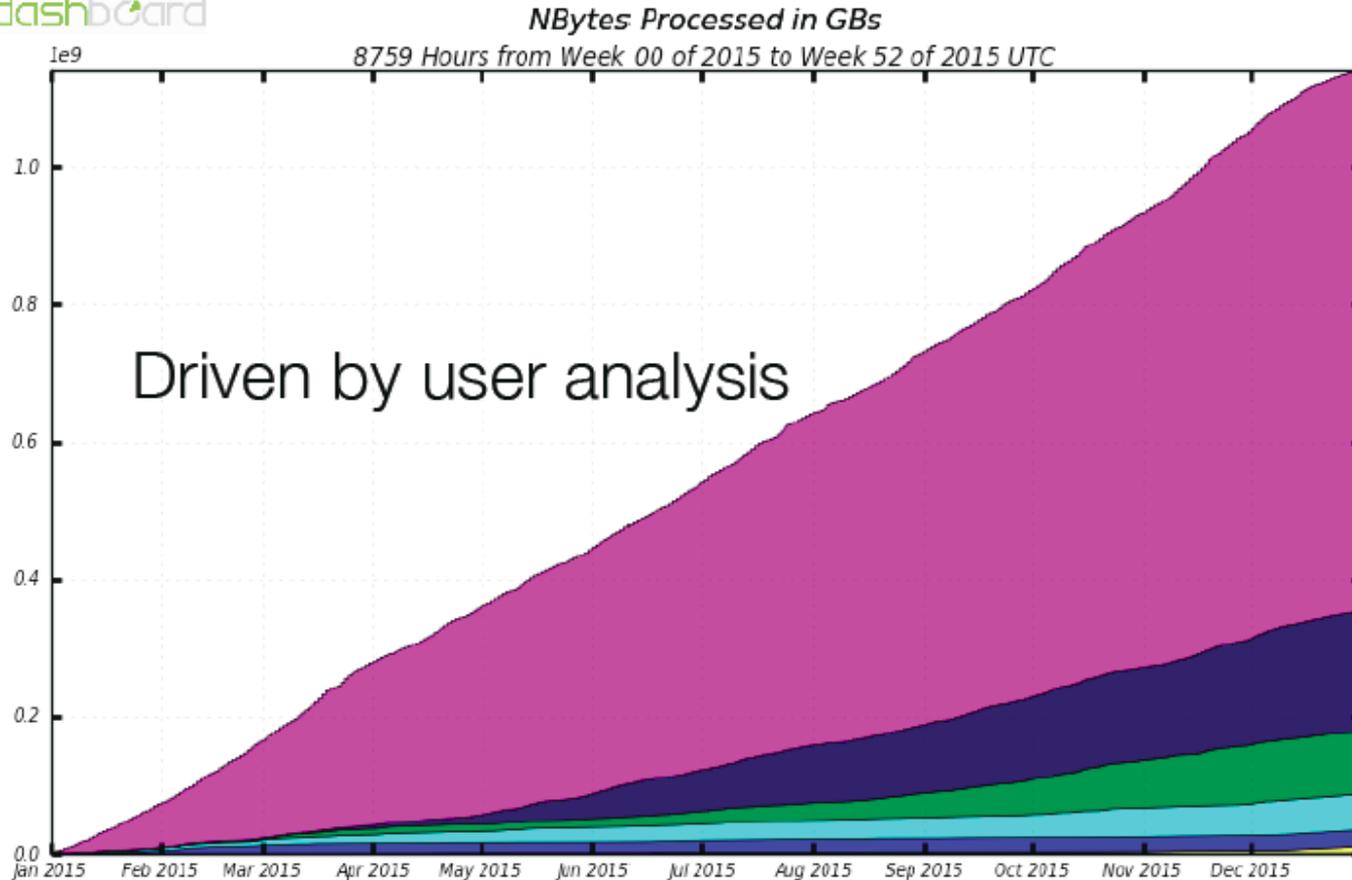
- Availabilityが高いサイト
- リソース量が多いサイト
- ネットワーク接続性が良いサイト



サイトの使われ方



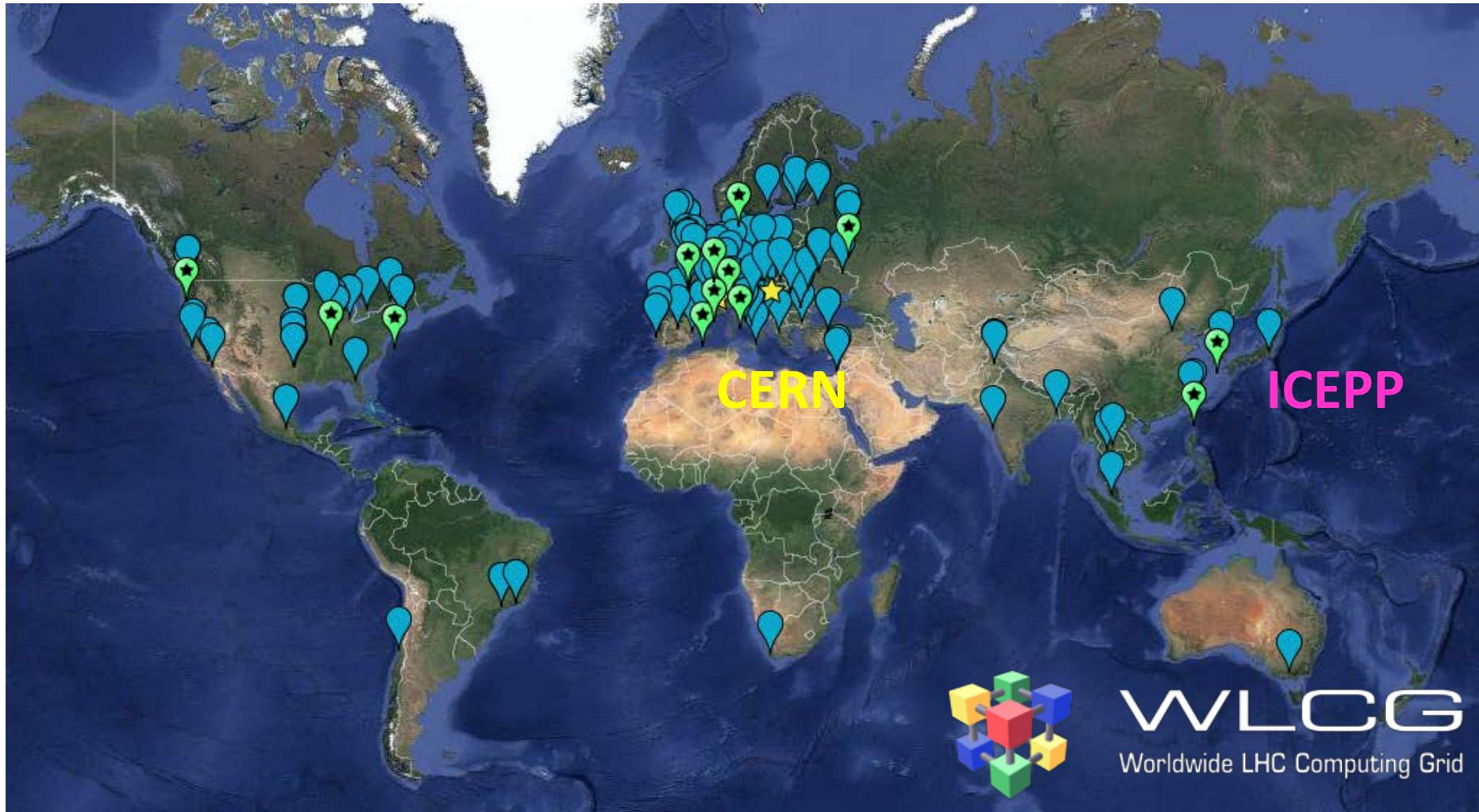
ATLASの処理したデータ量



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



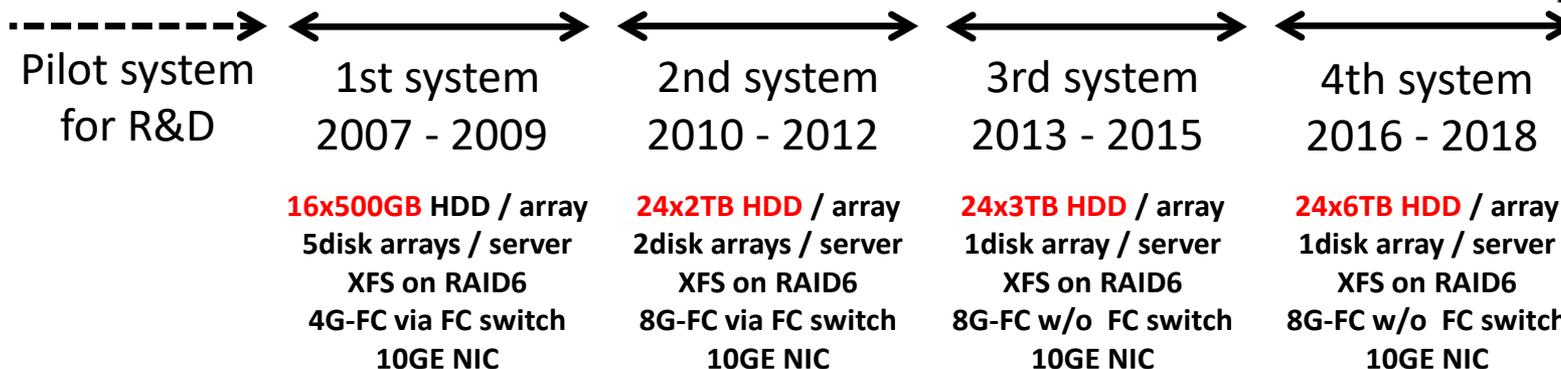
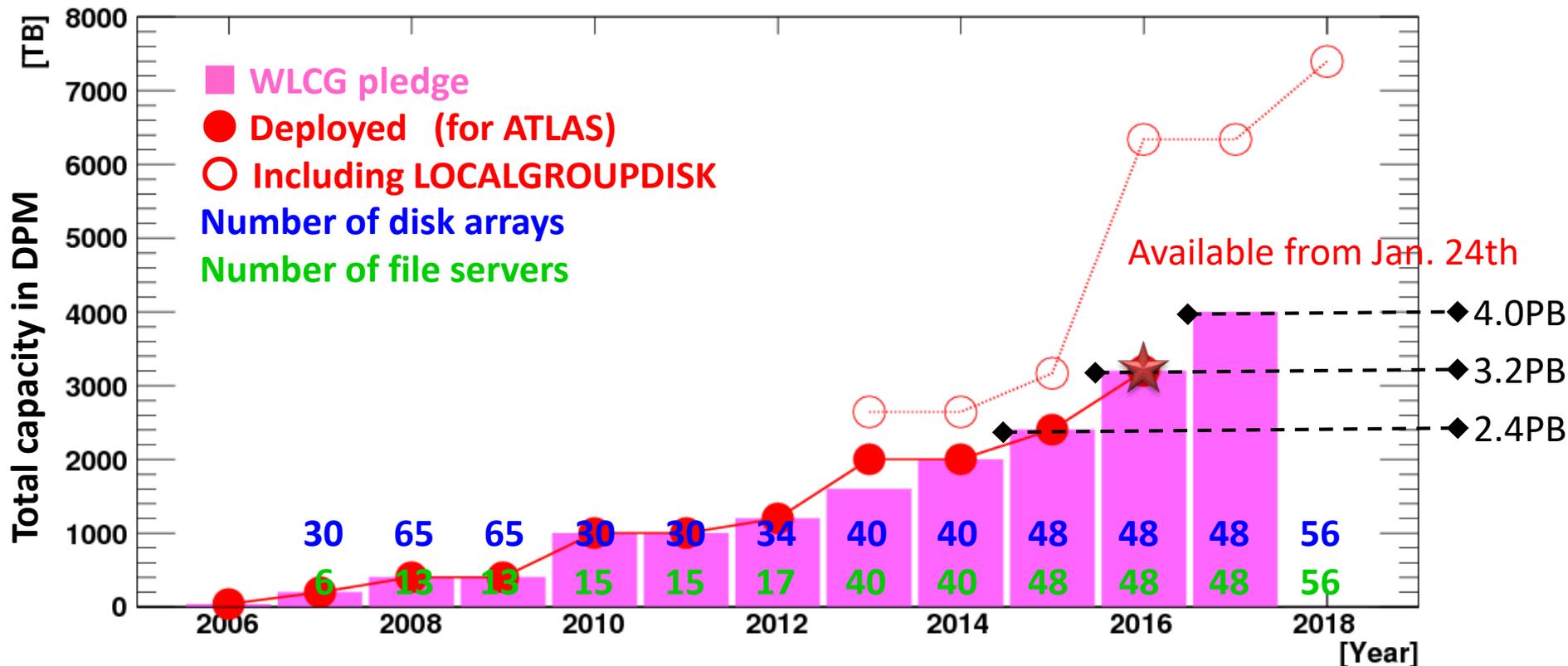
Disk arrays

システム構成)

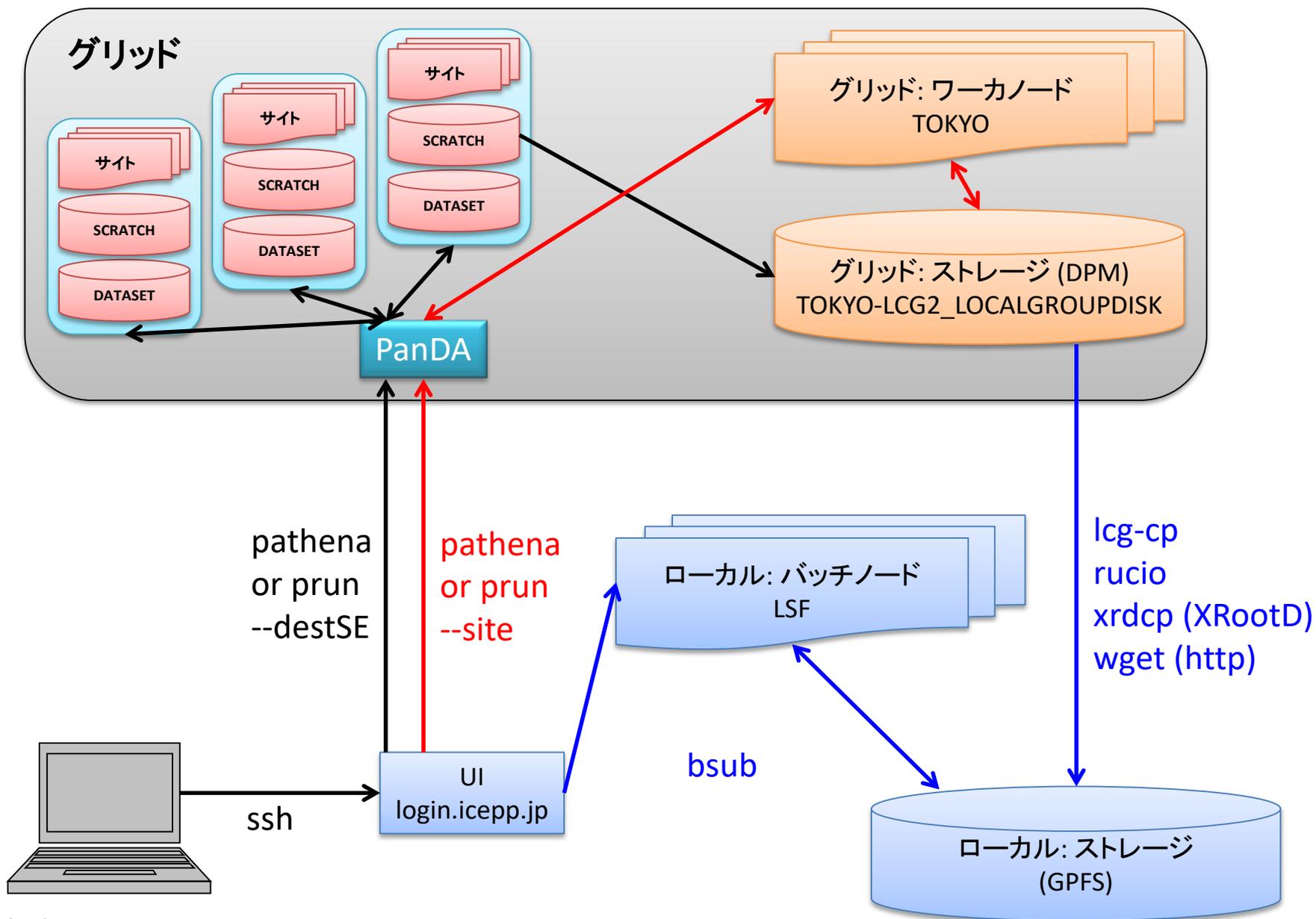
2016年1月末からTier2として本格稼働

		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)

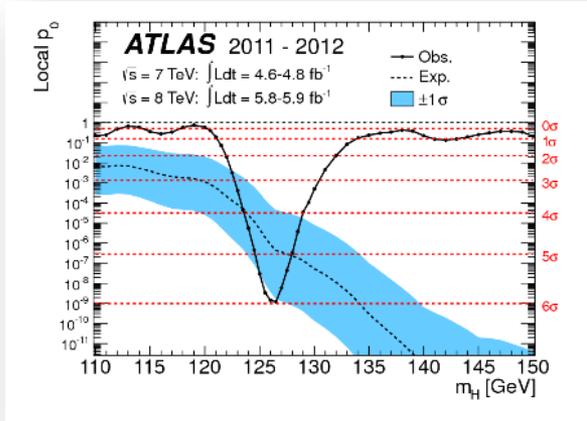


End user analysis workflow



to HL-LHC

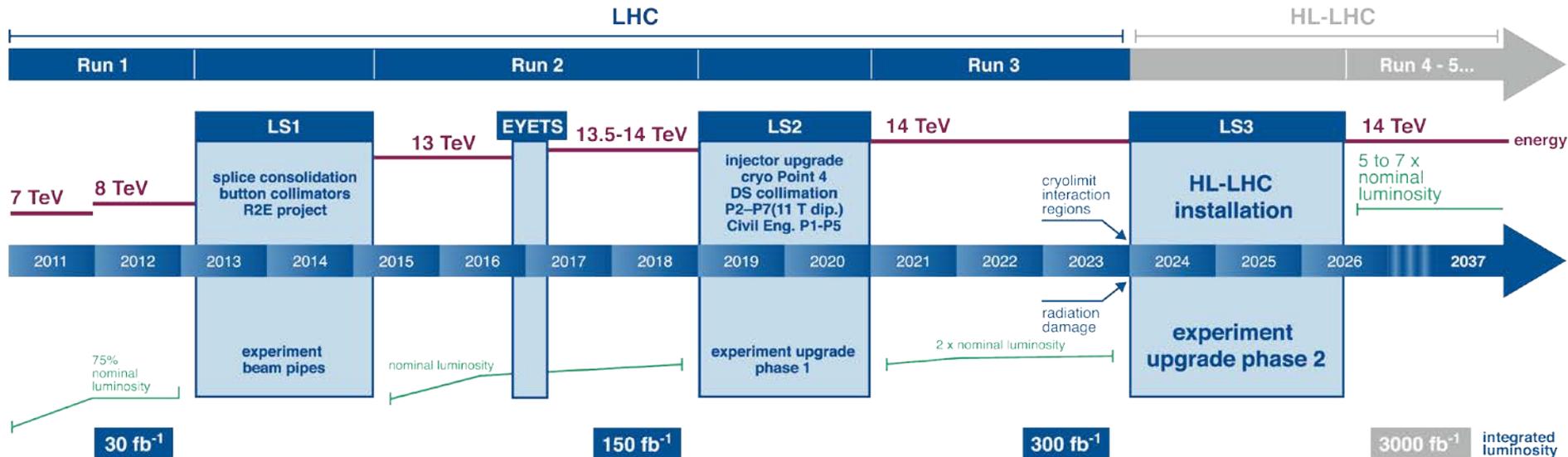
LHC-Run1で発見されたヒッグス粒子



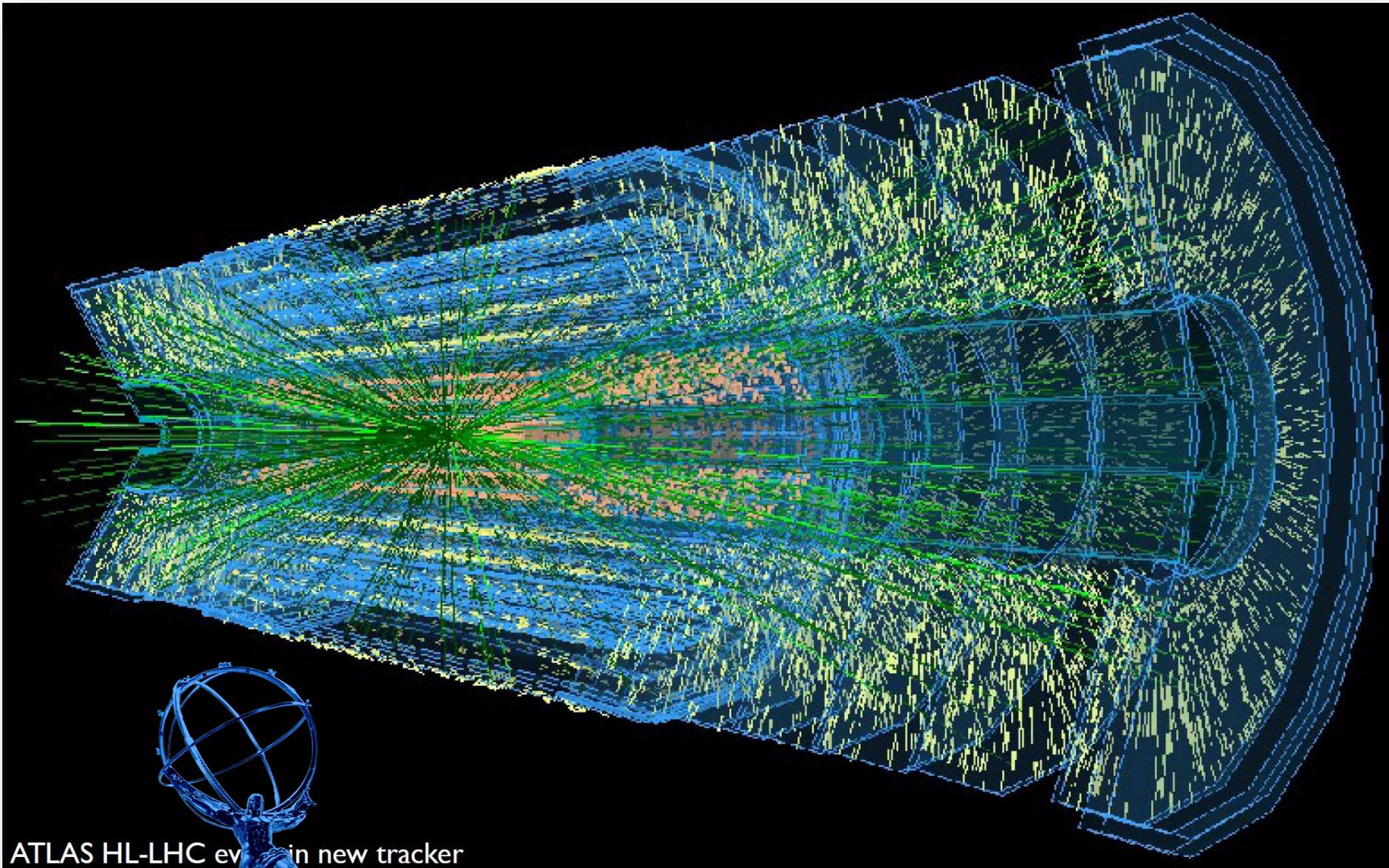
これから20年間でRun1の100倍の実験データを取得

- 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

LHC / HL-LHC Plan

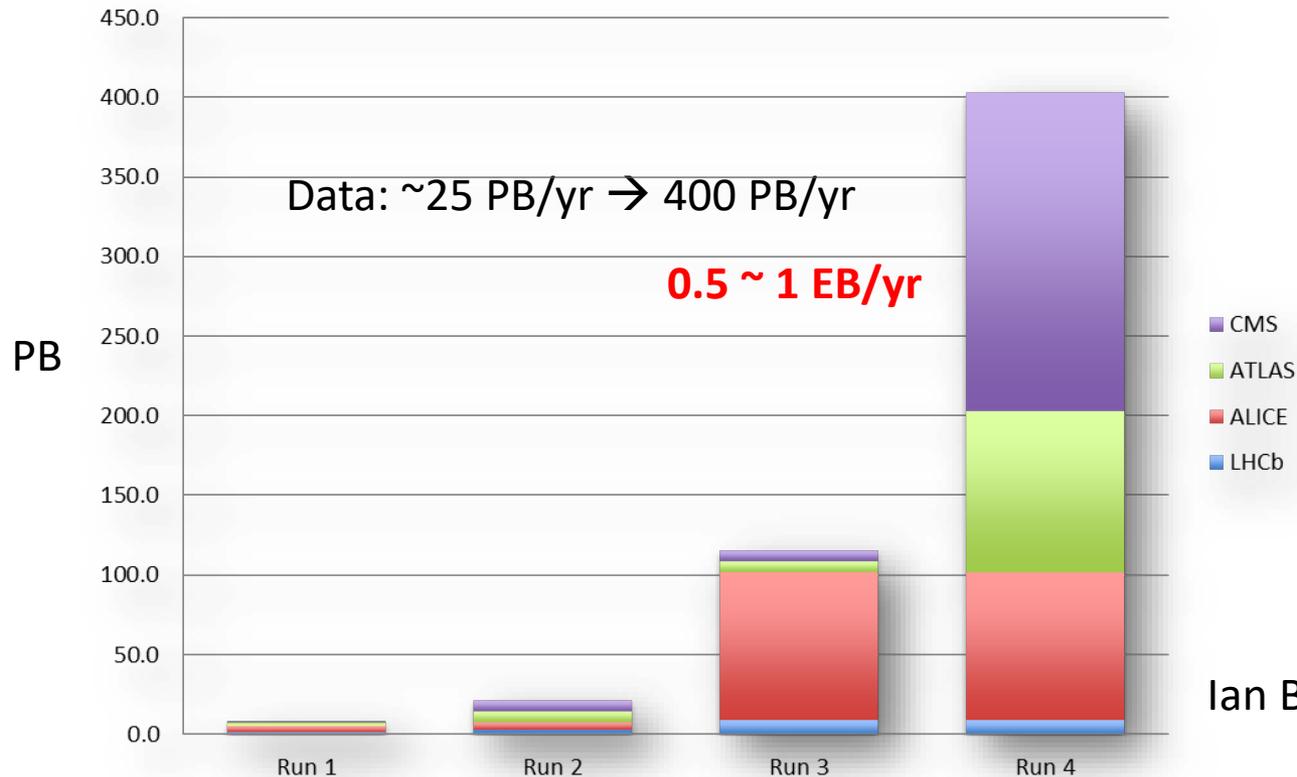


HL-LHC (150-200ピイルアップ)



ATLAS HL-LHC event in new tracker

Data volume in future



Ian Bird

	Run 1	Run 2	Run 3	Run 4
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)

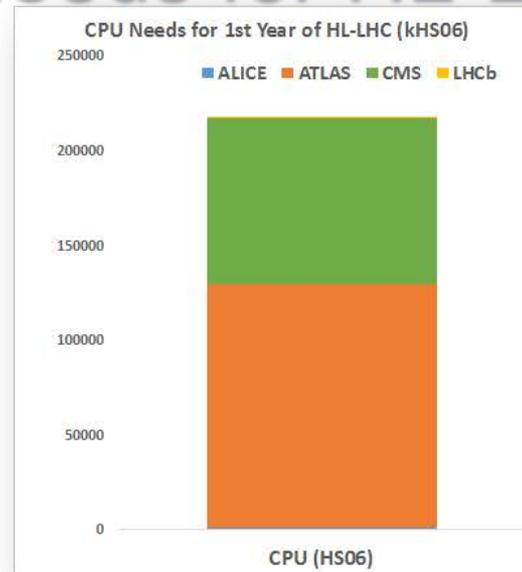
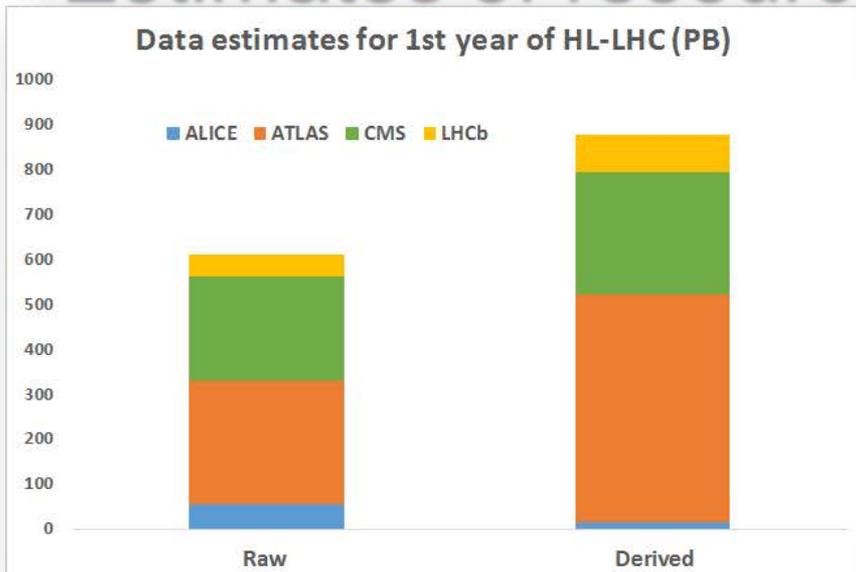
LHCb 20kHz (2GB/s)

ALICE 50kHz (75GB/s)

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

CMS 10kHz (40GB/s)

Estimates of resource needs for HL-LHC



Data:

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

CPU:

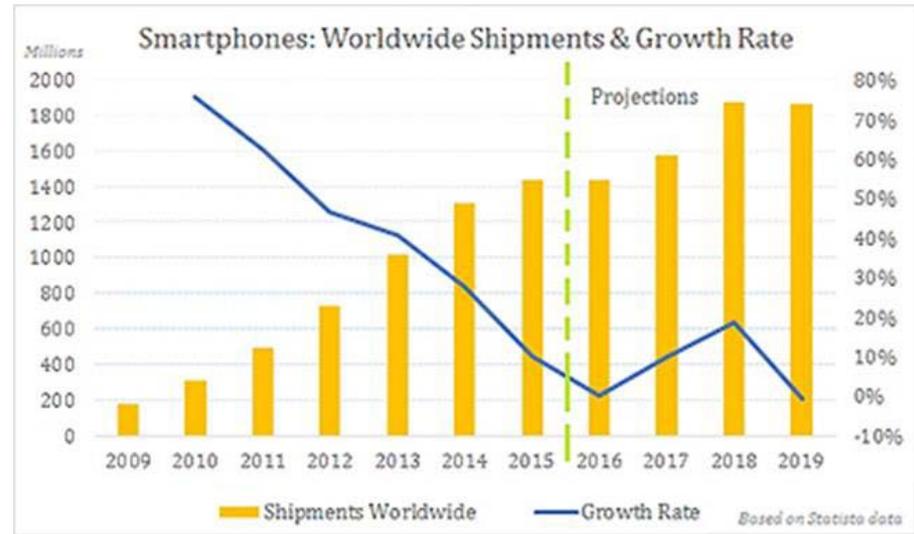
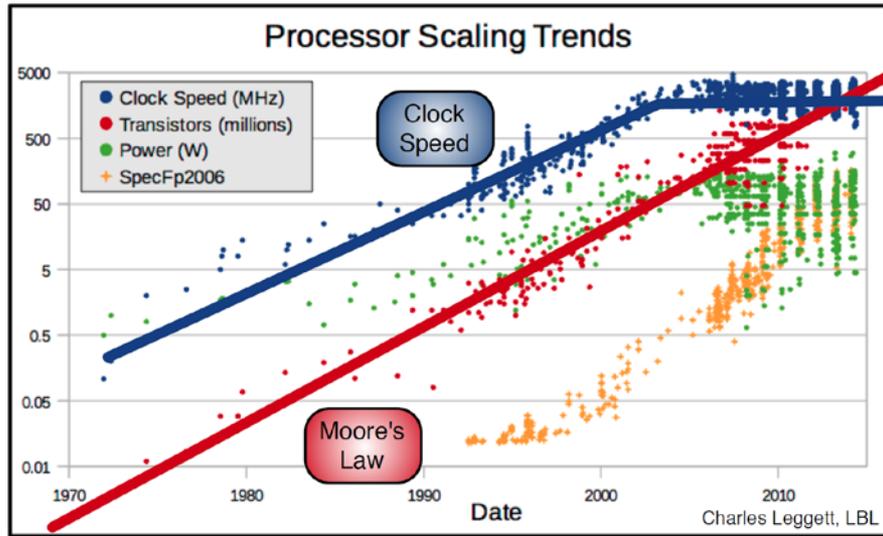
- x60 from 2016

I. Bird

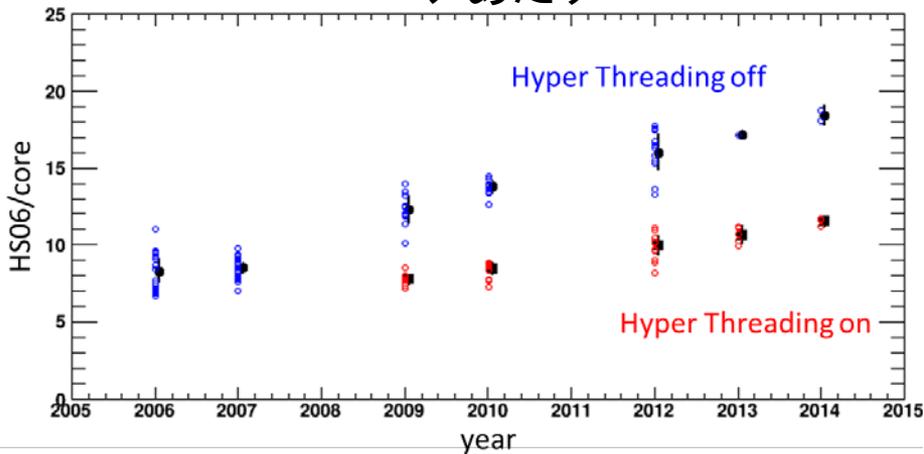
Technology at ~20%/year will bring x6-10 in 10-11 years

- ❑ 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

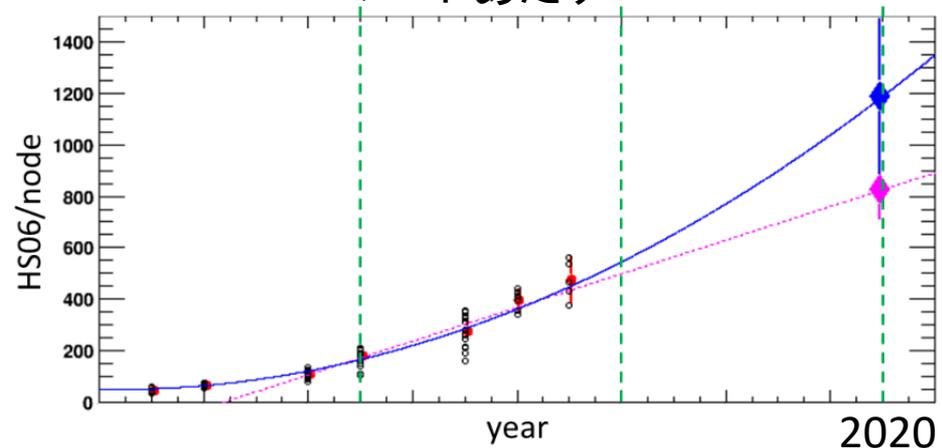


コアあたり



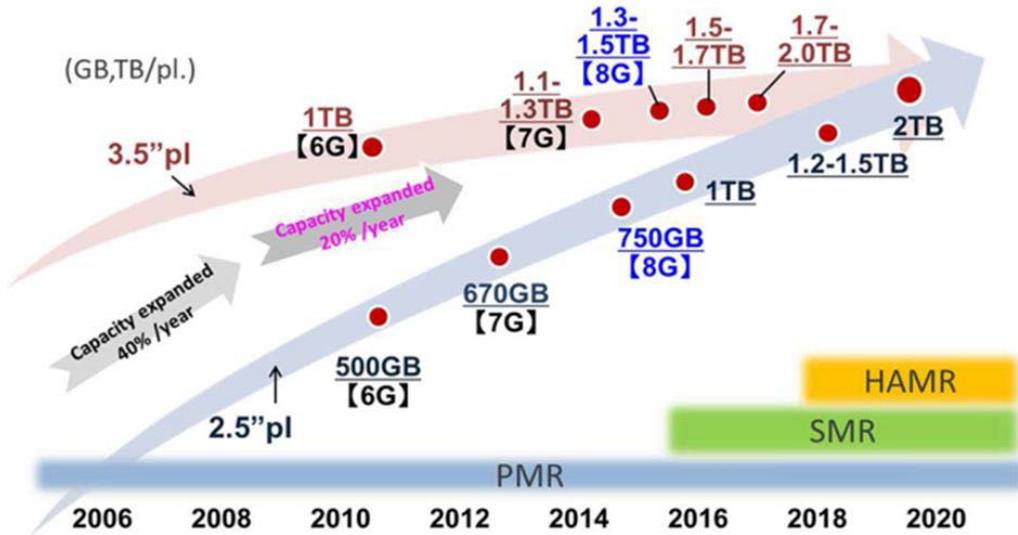
2% improve / year

ノードあたり



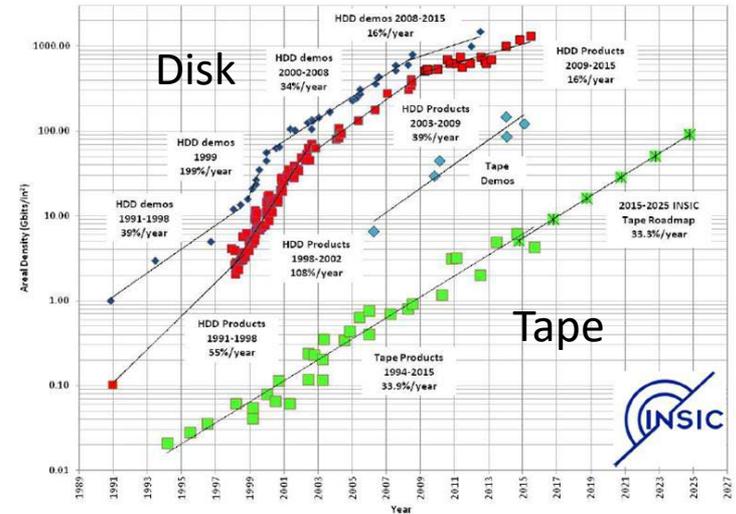
Storage (HDD, SSD, Tape)

[Road map for storage density increase] (SDK forecast)

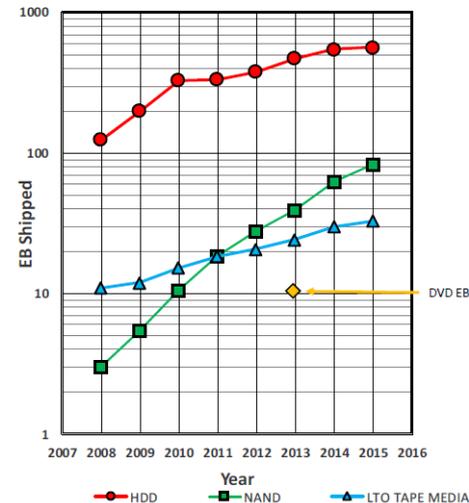
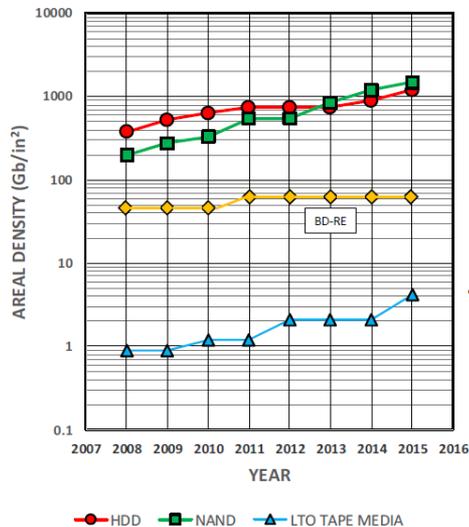


Areal Density Trends

Chart provided courtesy of the Information Storage Industry Consortium (INSIC)



©2016 Information Storage Industry Consortium - All Rights Reserved



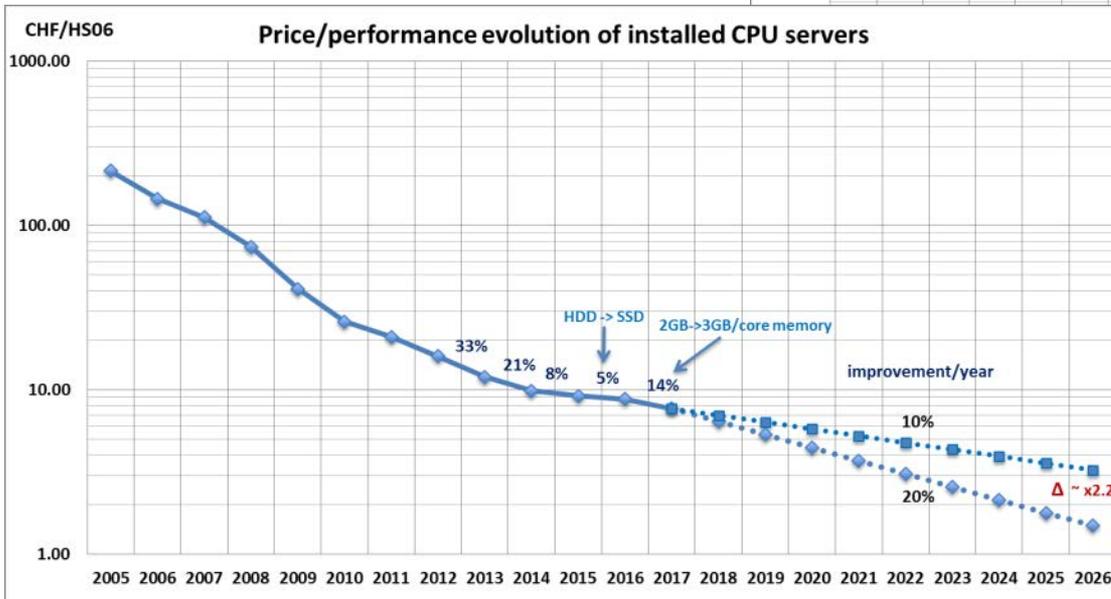
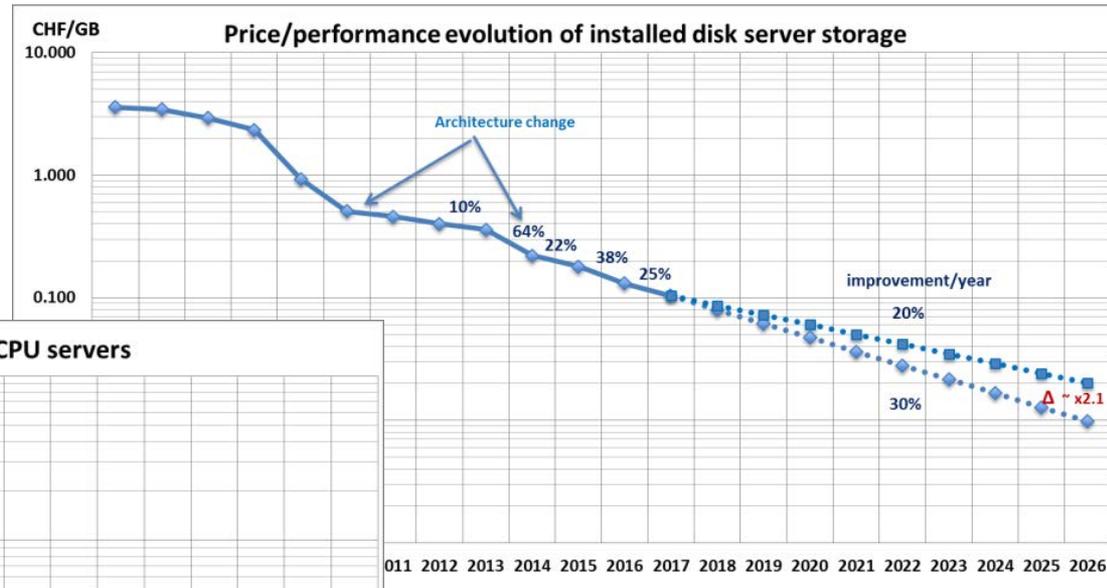
計算機技術の進歩によるコスト減

Disk

CERNの調達に基づく予測

H. Meinhard, B. Panzer-Steindel

CPU



他にも、
Memory, GPU, Network....

HL-LHC computing cost parameters

