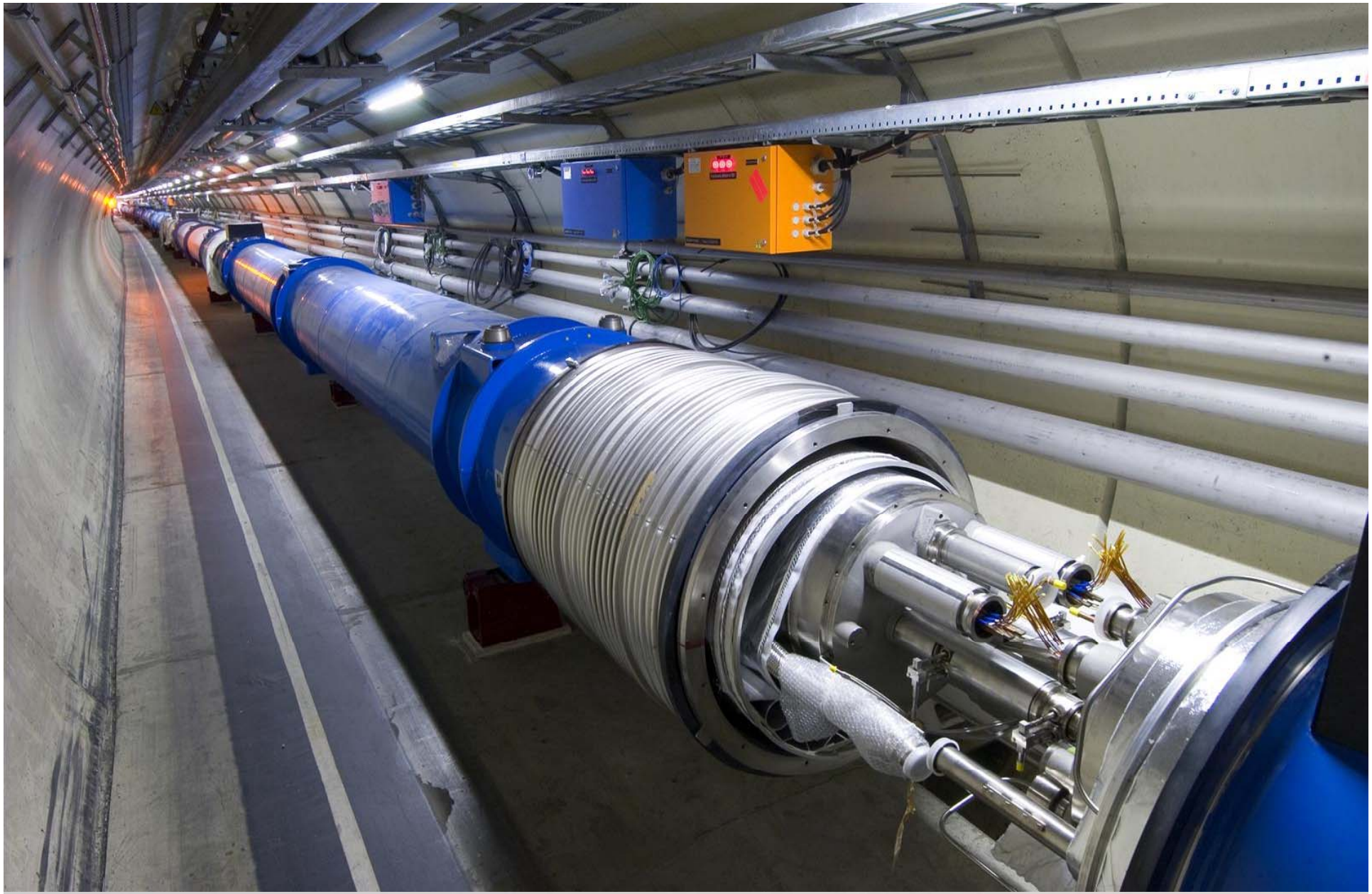
LHC加速器ATLAS実験
大規模転送処理演習の経験

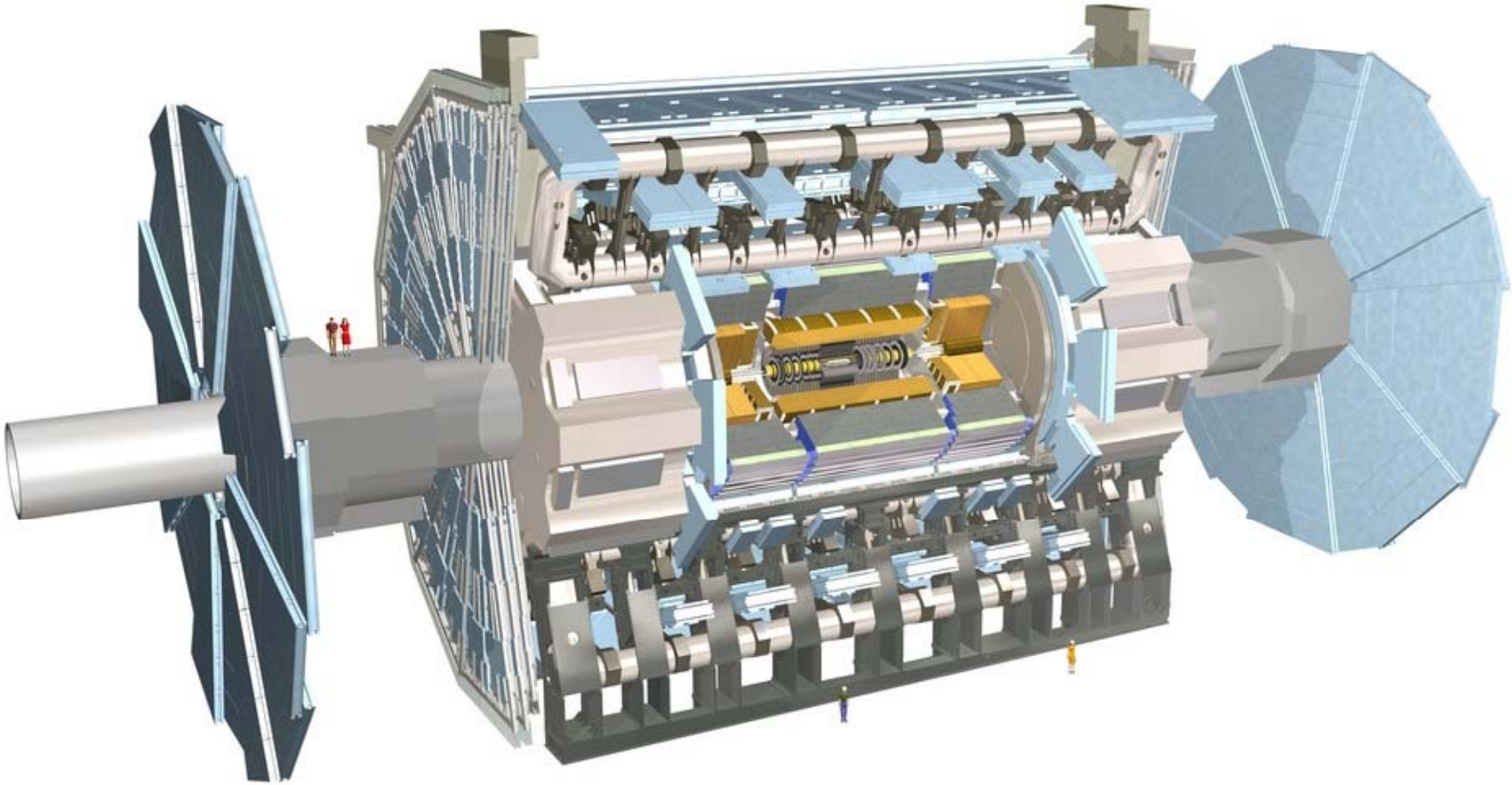
坂本　宏

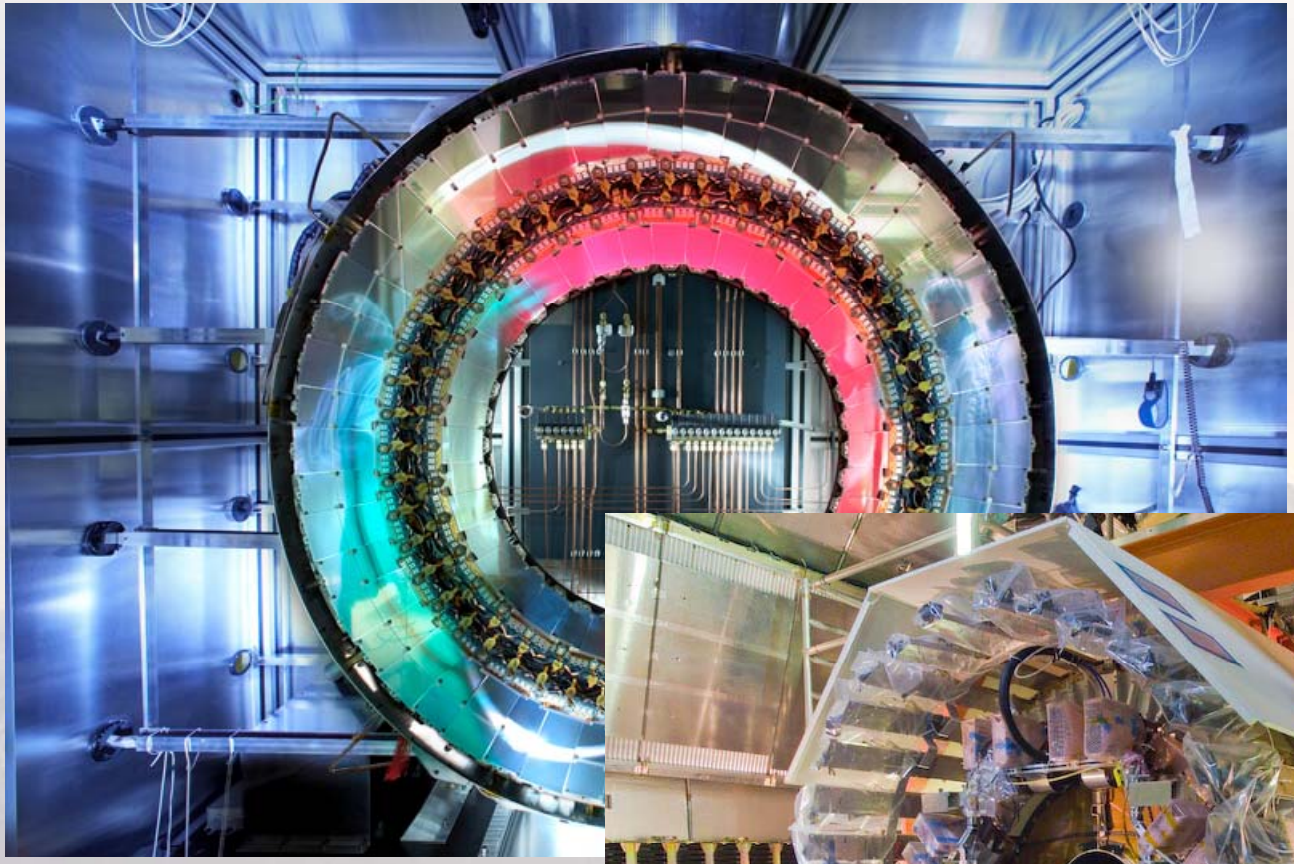
東京大学素粒子物理国際研究センター

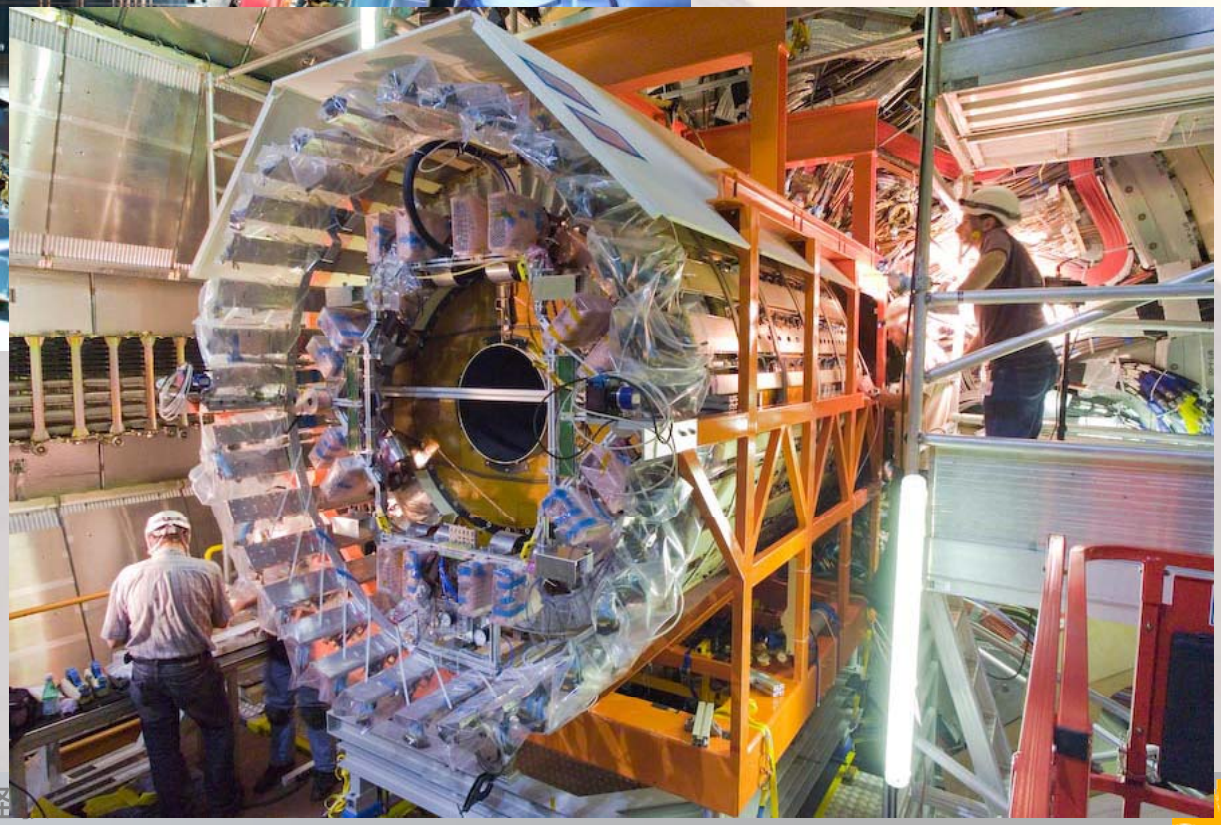
目次

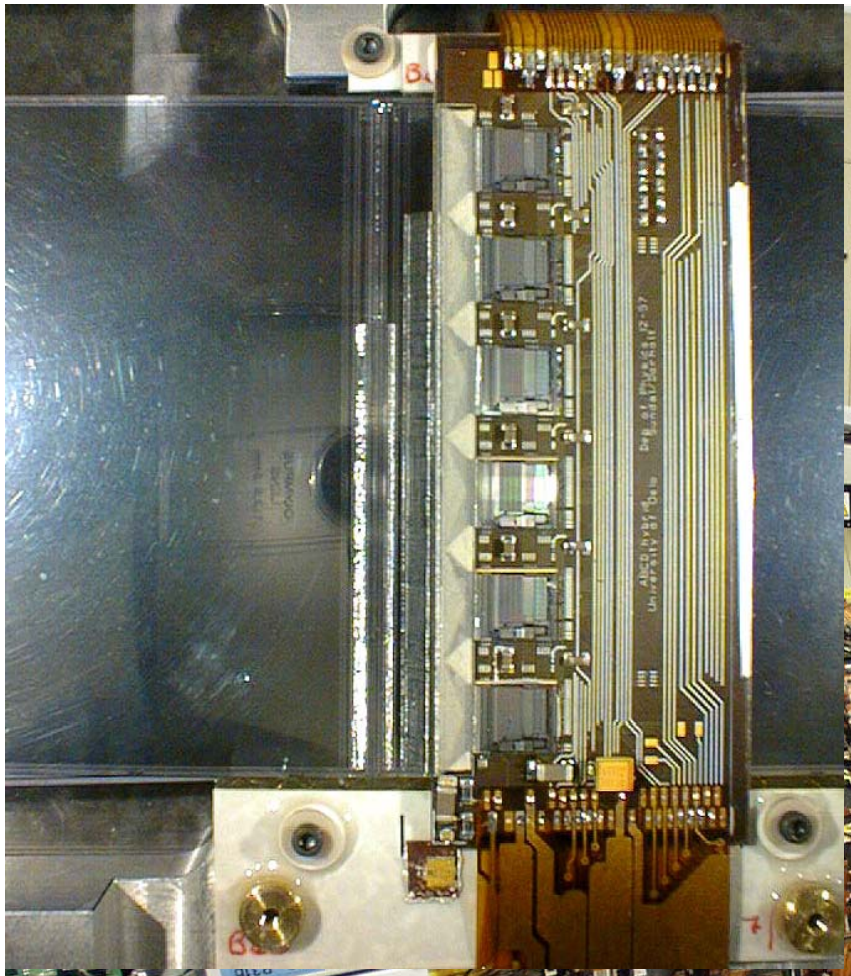
- LHC加速器ATLAS実験
- LHC Computing Grid (LCG)
- ATLAS実験でのLCGの利用
  - グリッドジョブ管理
  - 分散データ管理
- STEP09総合演習
  - ATLAS全体
  - 東京大学(TOKYO-LCG2)
- 今後の予定

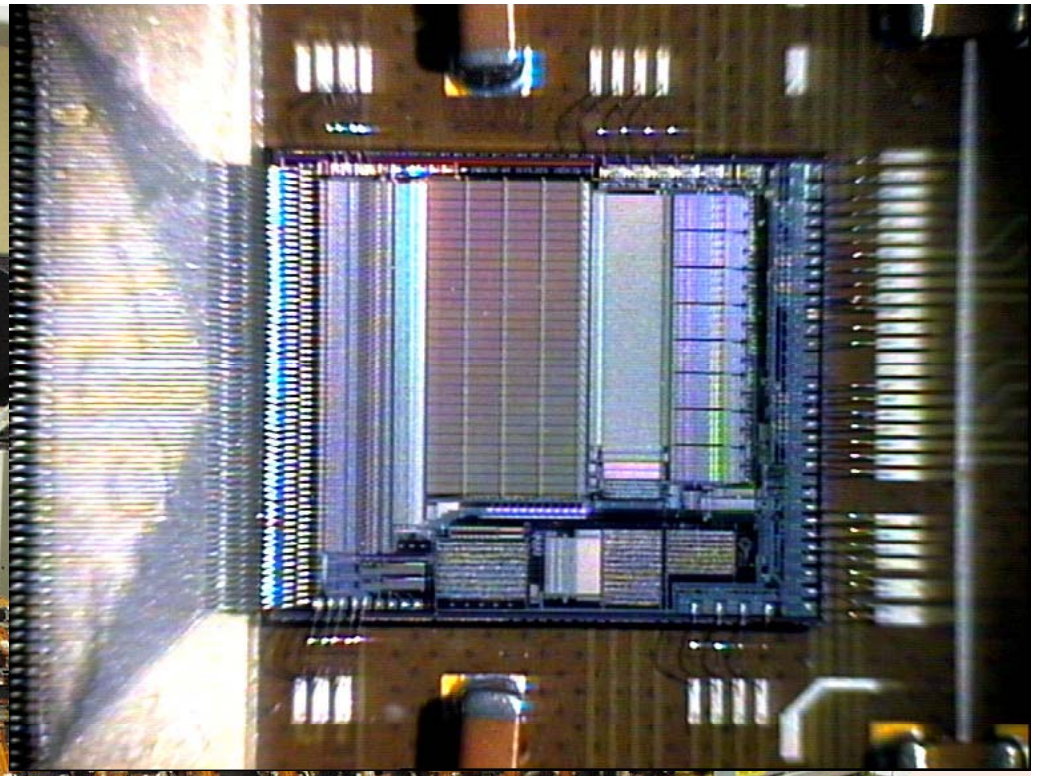


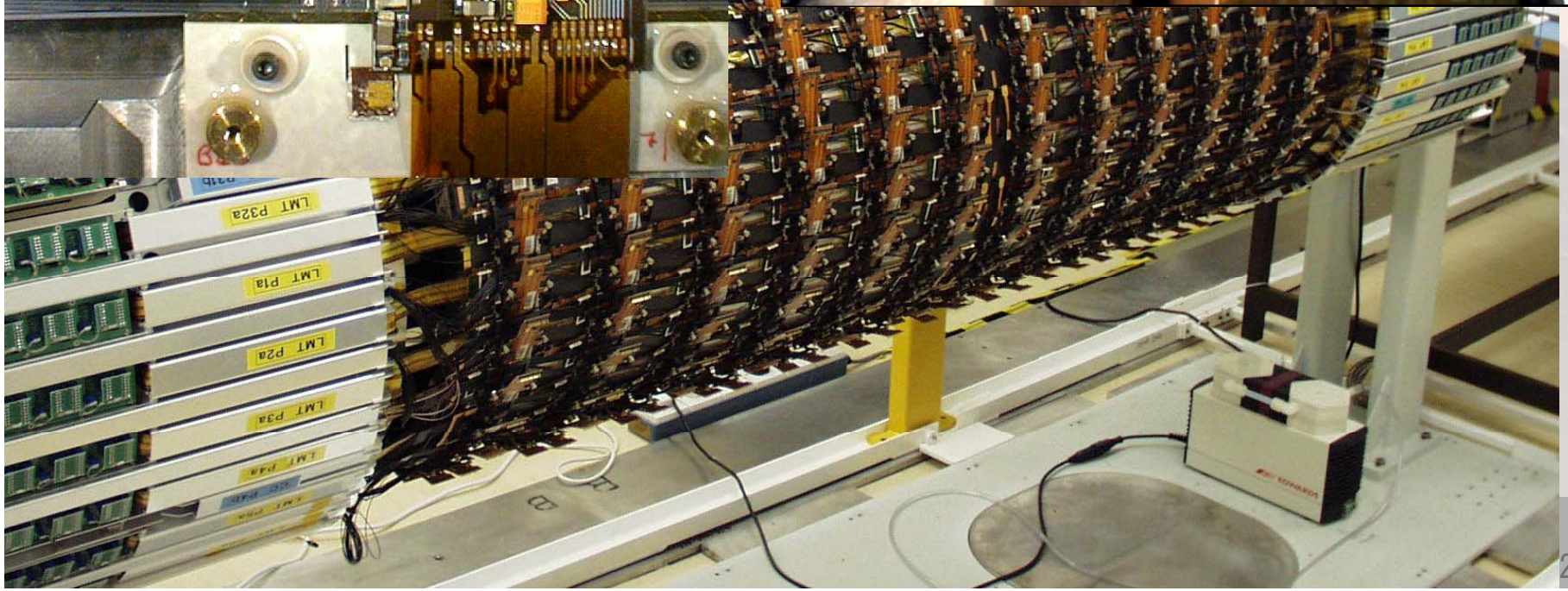


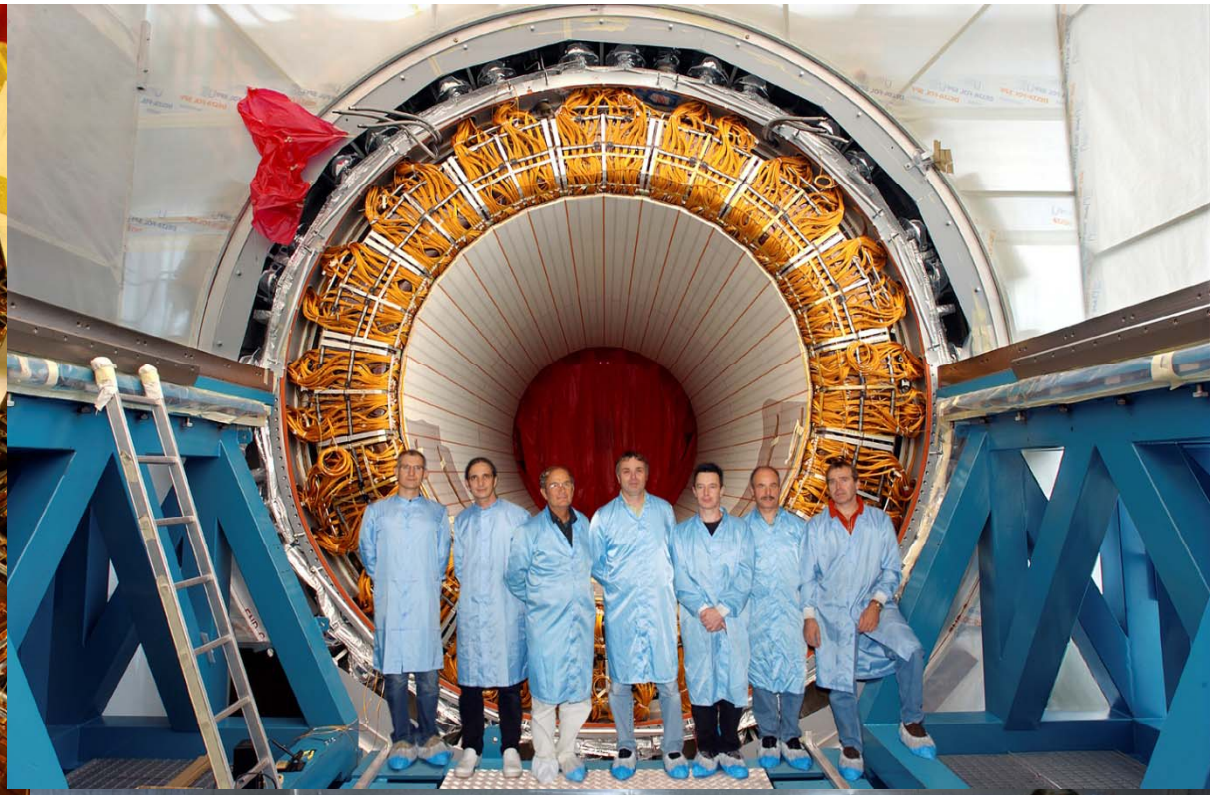


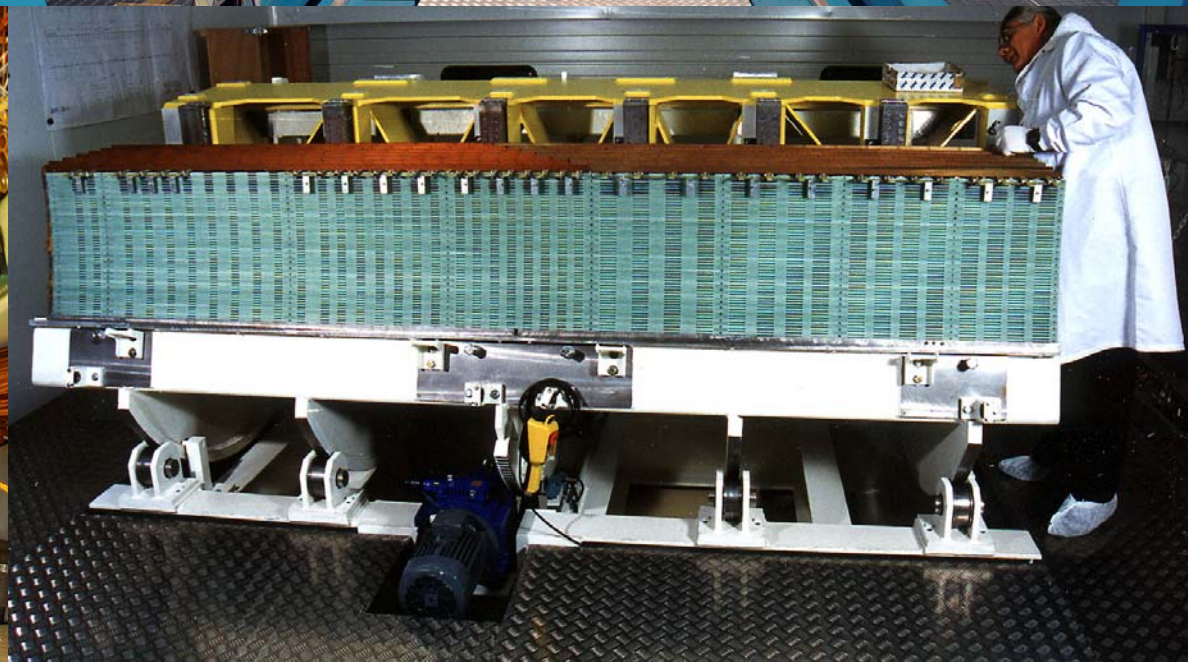


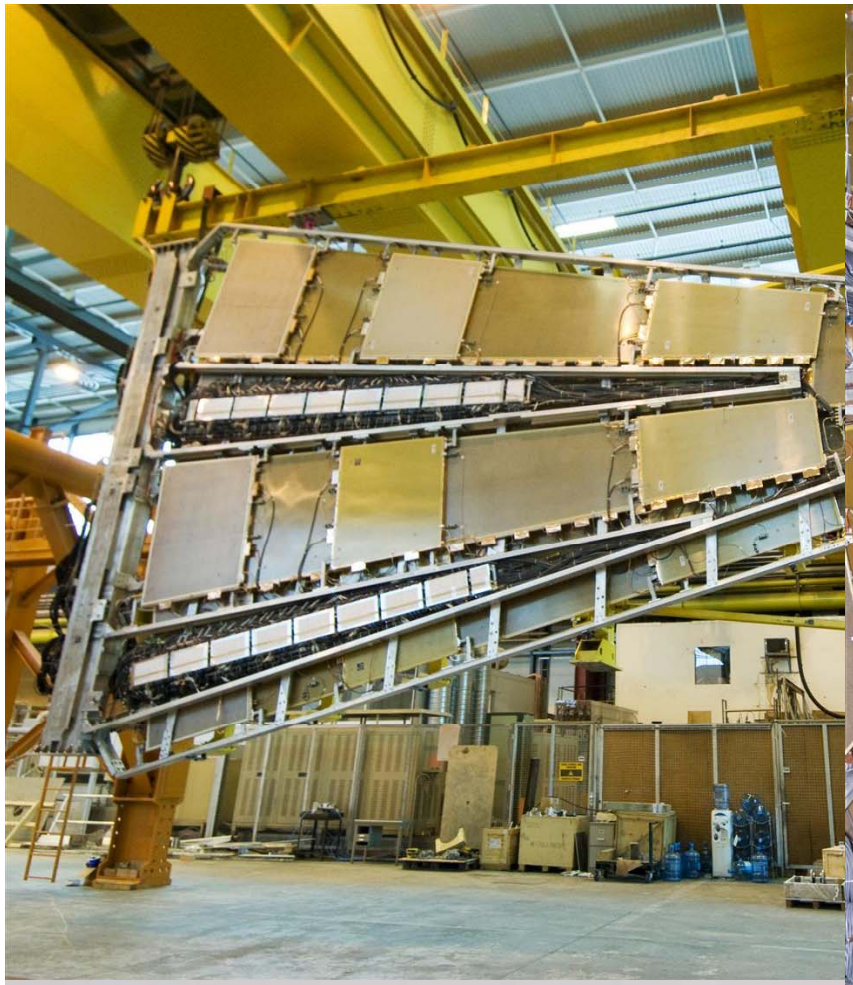




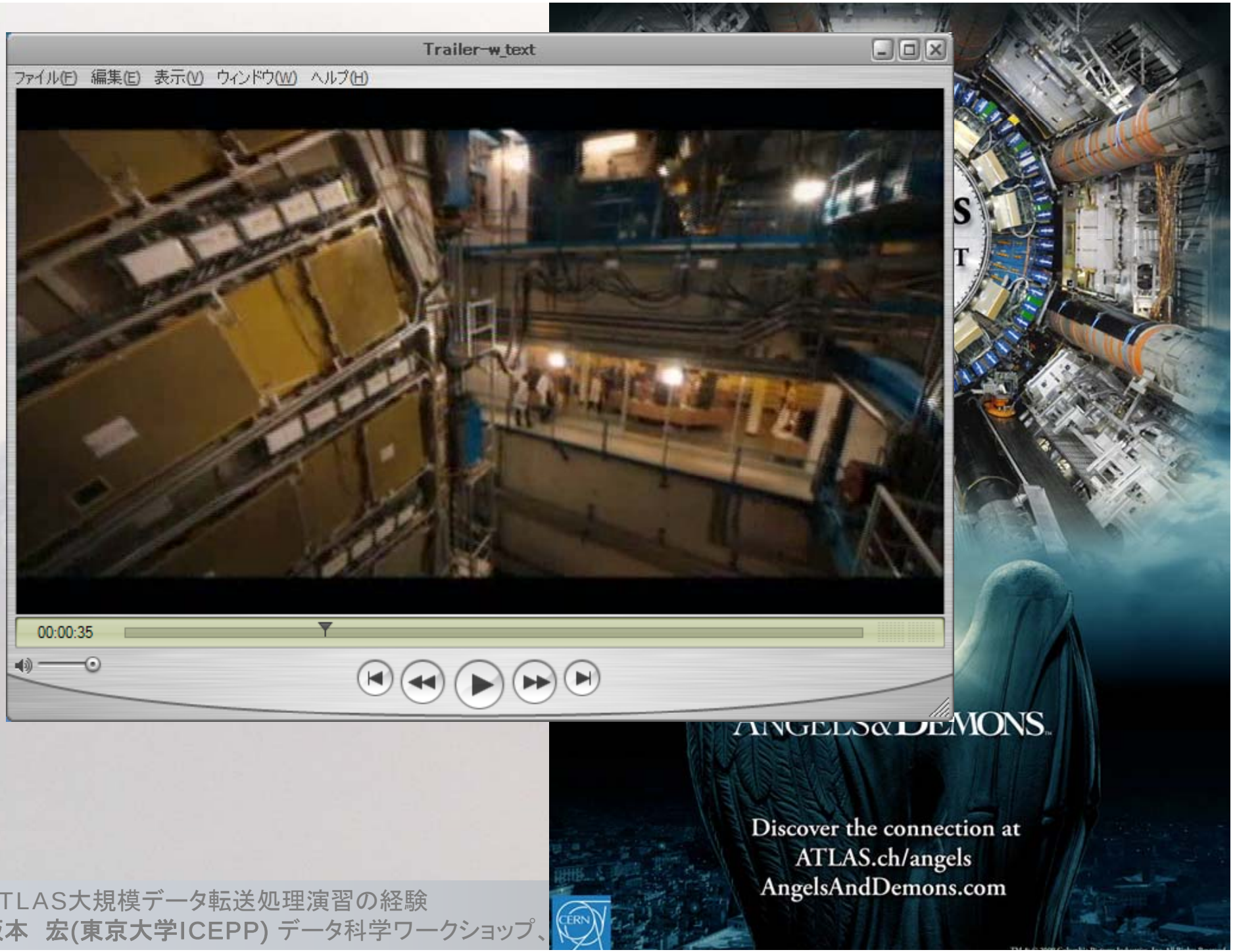

LMT P32a
LMT P1a
LMT P2a
LMT P3a







ATLAS大規模データ転送処理演習の経験
坂本　宏(東京大学ICEPP) データ科学ワークショッ



ATLAS大規模データ転送処理演習の経験

坂本　宏(東京大学ICEPP) データ科学ワークショップ、

Data from ATLAS

320MB/s Throughput ~ 6 seconds for 2GB file

3.2PB/2GB = 1.6M files

	Rate(Hz)	sec/year	Events/y	Size(MB)	Total(TB)
Raw Data	200	1.00E+07	2.00E+09	1.6	3200
ESD (Reconstruction out)	200	1.00E+07	2.00E+09	0.5	1000
General ESD	180	1.00E+07	1.80E+09	0.5	900
General AOD (Analysis)	180	1.00E+07	1.80E+09	0.1	180
General TAG (Event db)	180	1.00E+07	1.80E+09	0.001	2
Calibration					40
MC Raw			1.00E+08	2	200
ESD Sim			1.00E+08	0.5	50
AOD Sim			1.00E+08	0.1	10
TAG Sim			1.00E+08	0.001	0
Tuple				0.01	

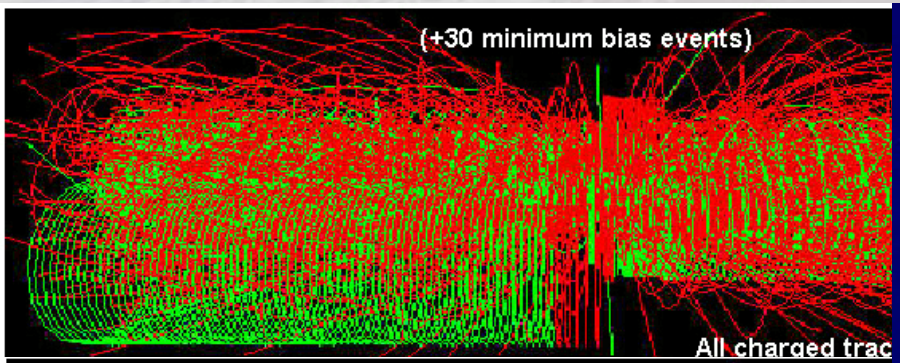
Processing Power for ATLAS

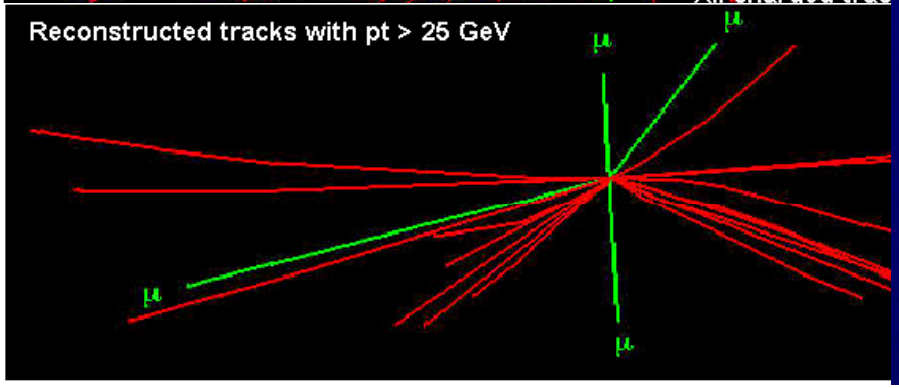
Reconstruction:

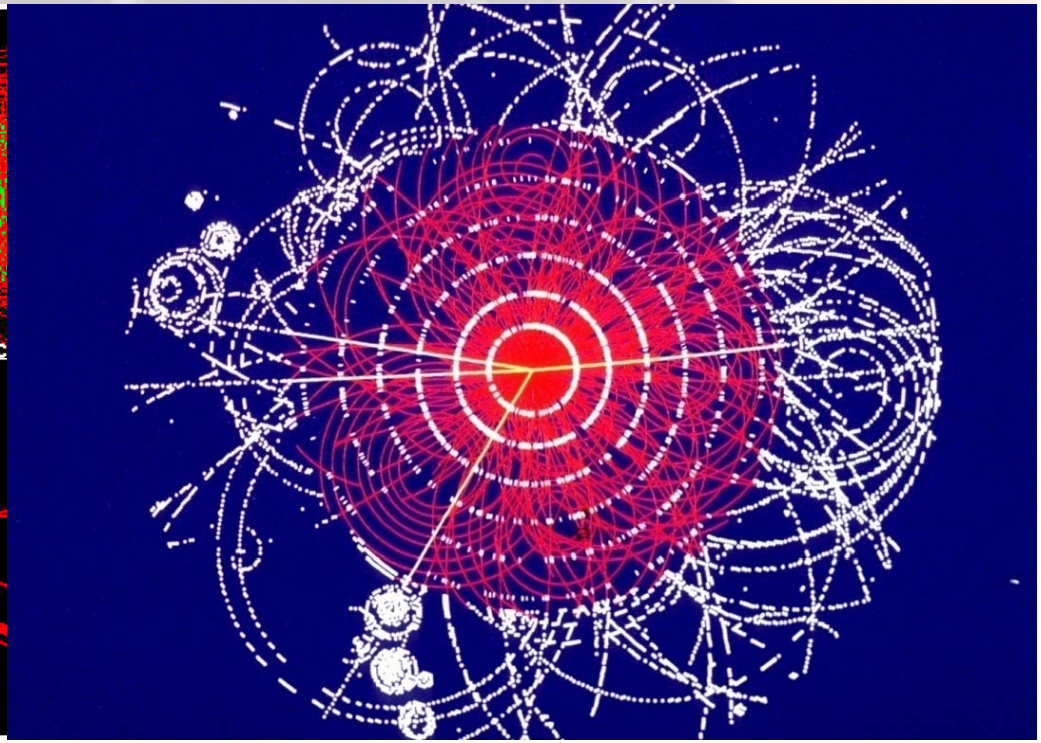
15kSI2ksec/event - 3000kSI2k (200Hz)

Simulation:

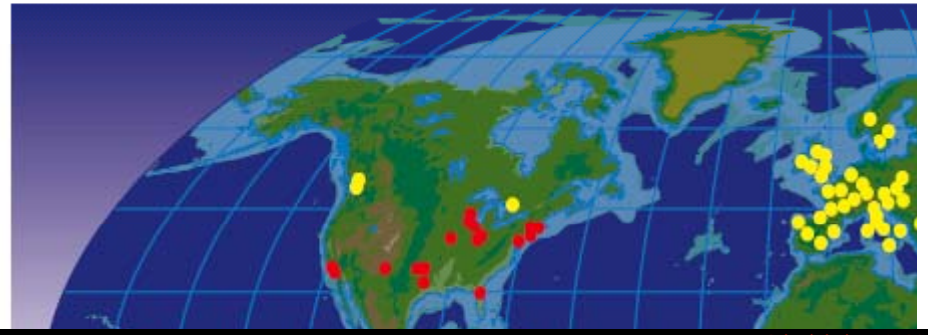
400kSI2ksec/event - 4000kSI2k (10Hz)

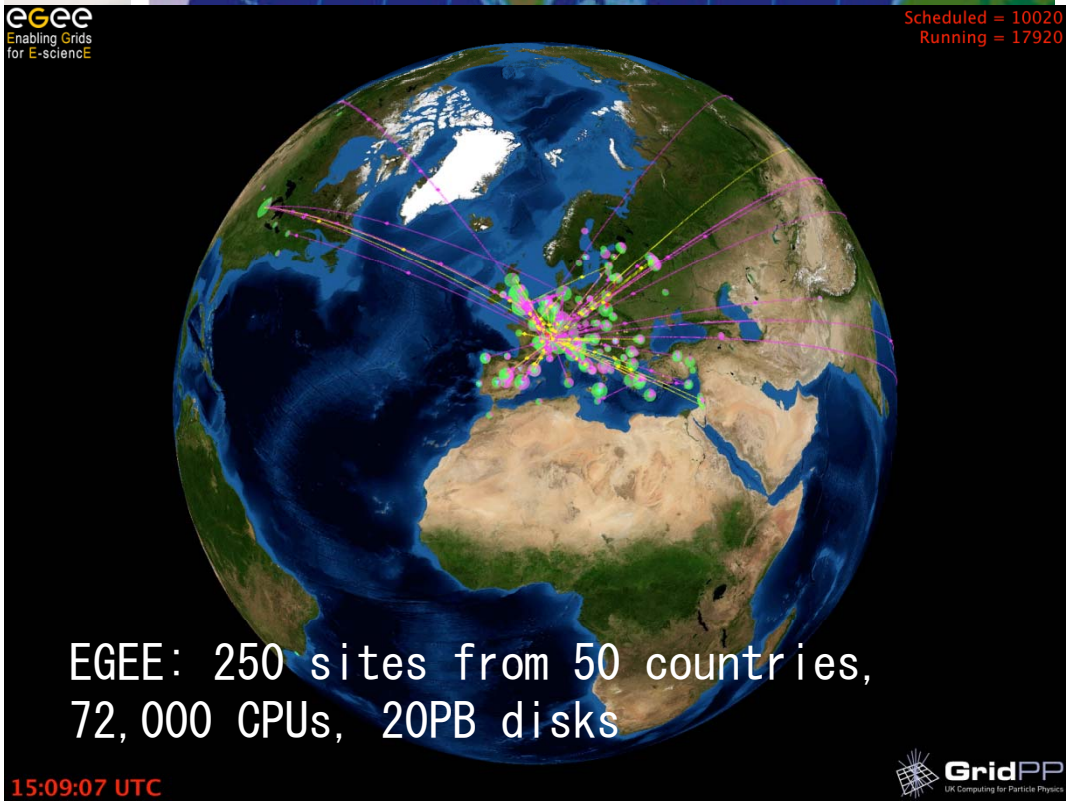


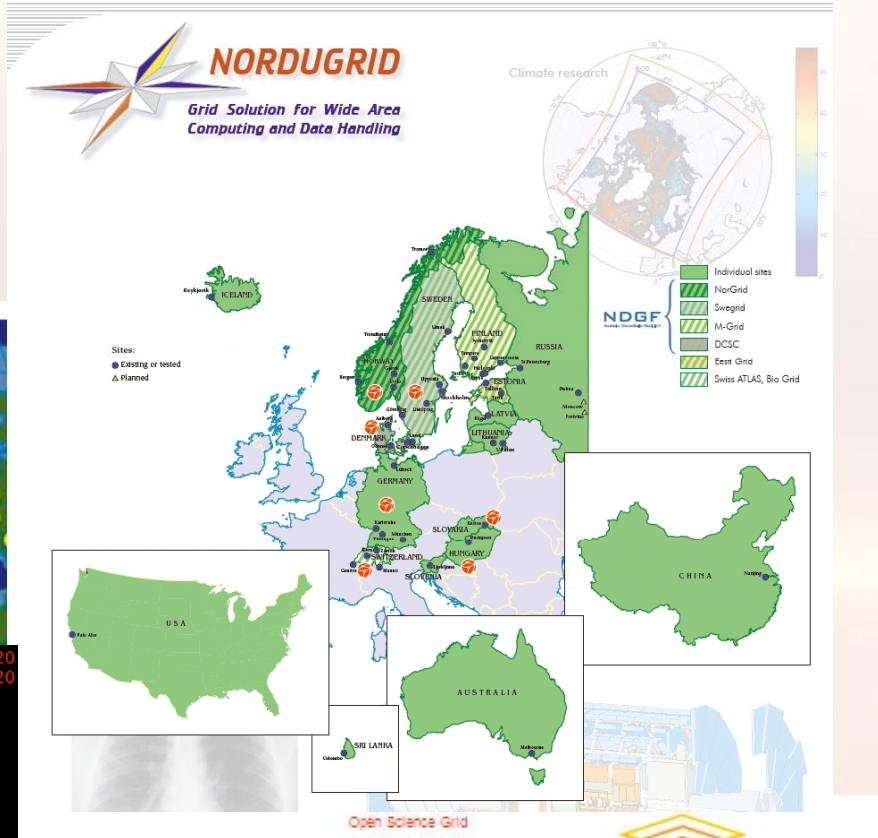


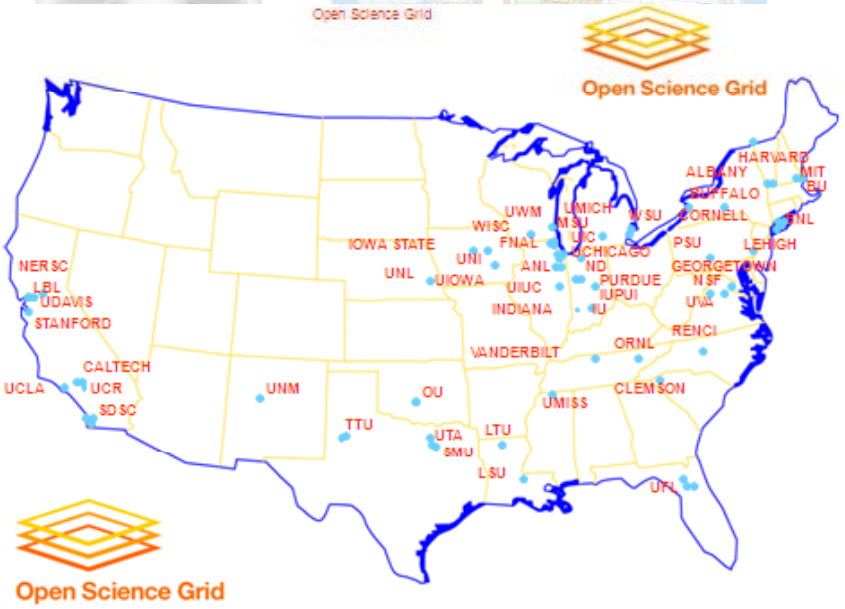


Worldwide LHC Computing Grid (WLCG)





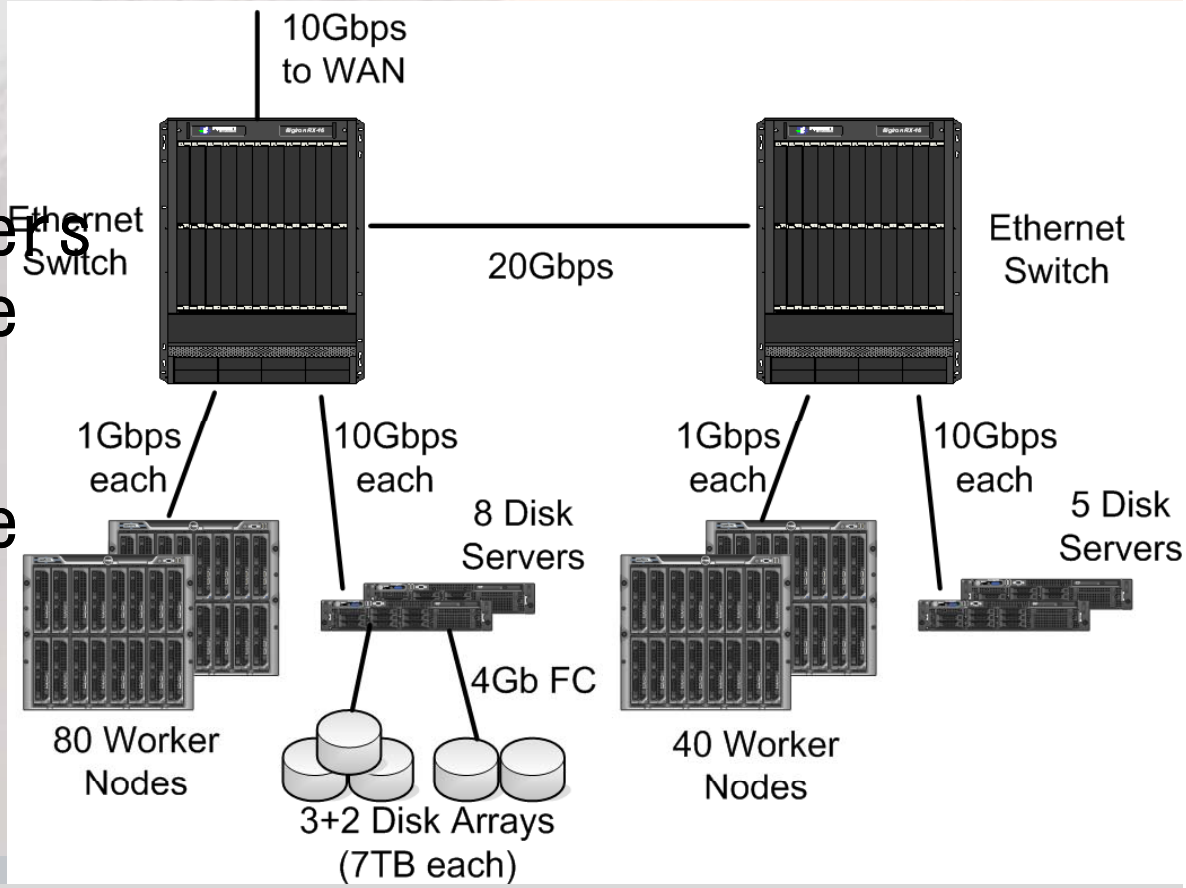




Grid Deployment in Asia Pacific Region

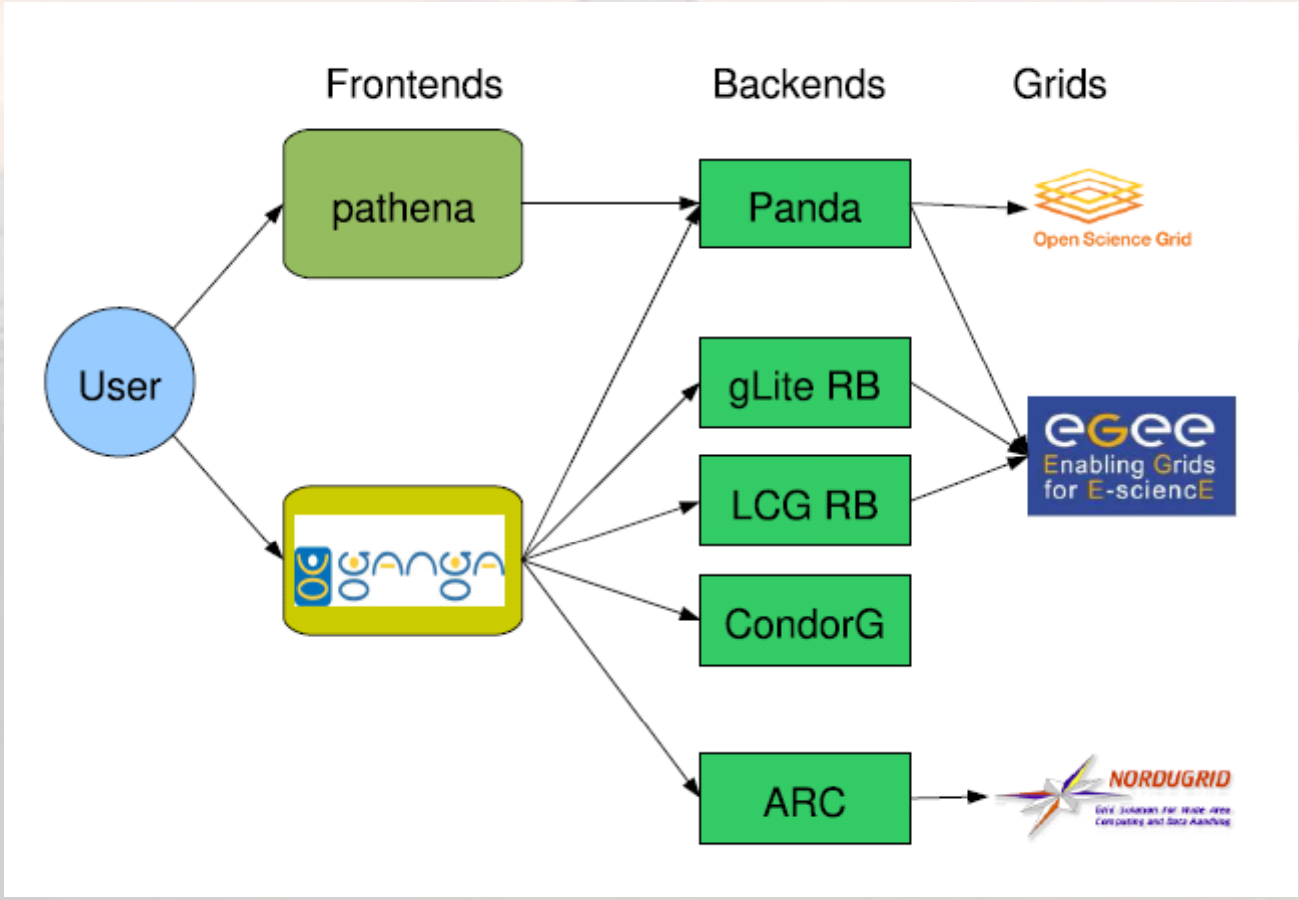
Site Configuration during STEP09

- Tier-2 dedicated to ATLAS
- SE:
  - DPM 1.7.0
  - 13 disk servers + 1 head node
- WN:
  - 4 cores/node
  - 120 nodes (480 cores)

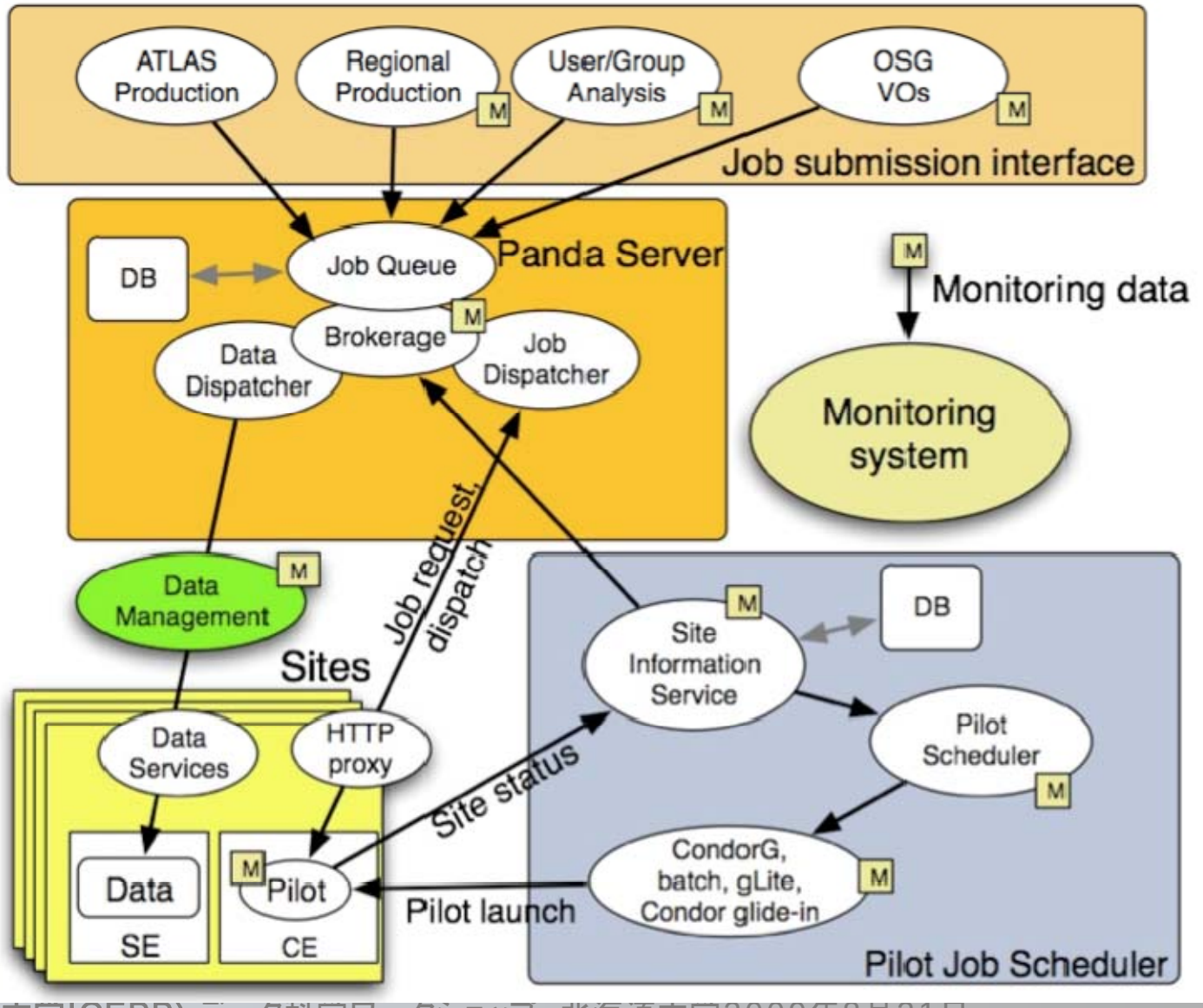


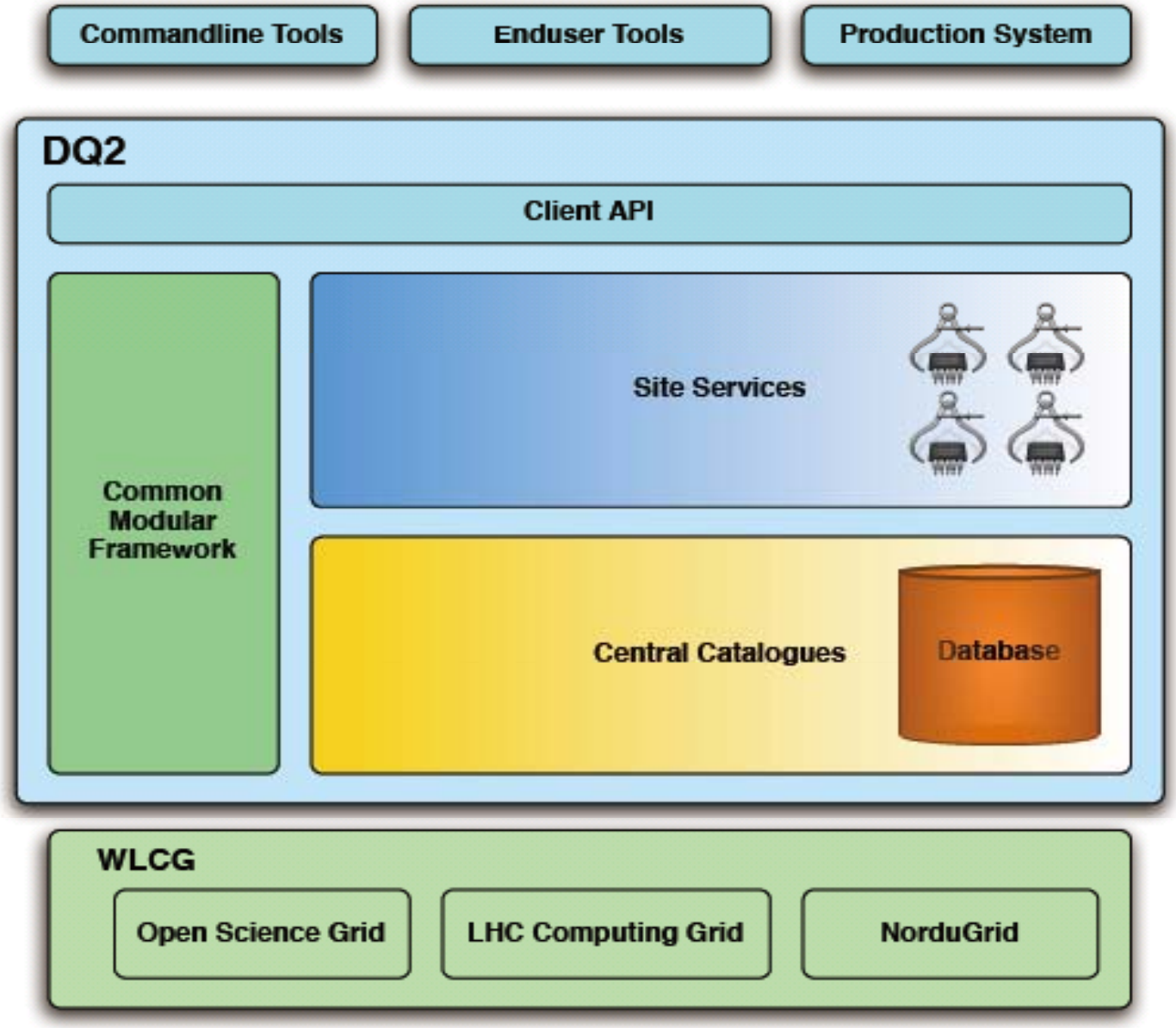
H. Matsunaga, STEP09 Post-mortem workshop Jul.

ATLAS Job Submission on Grids



PanDA System Schematic



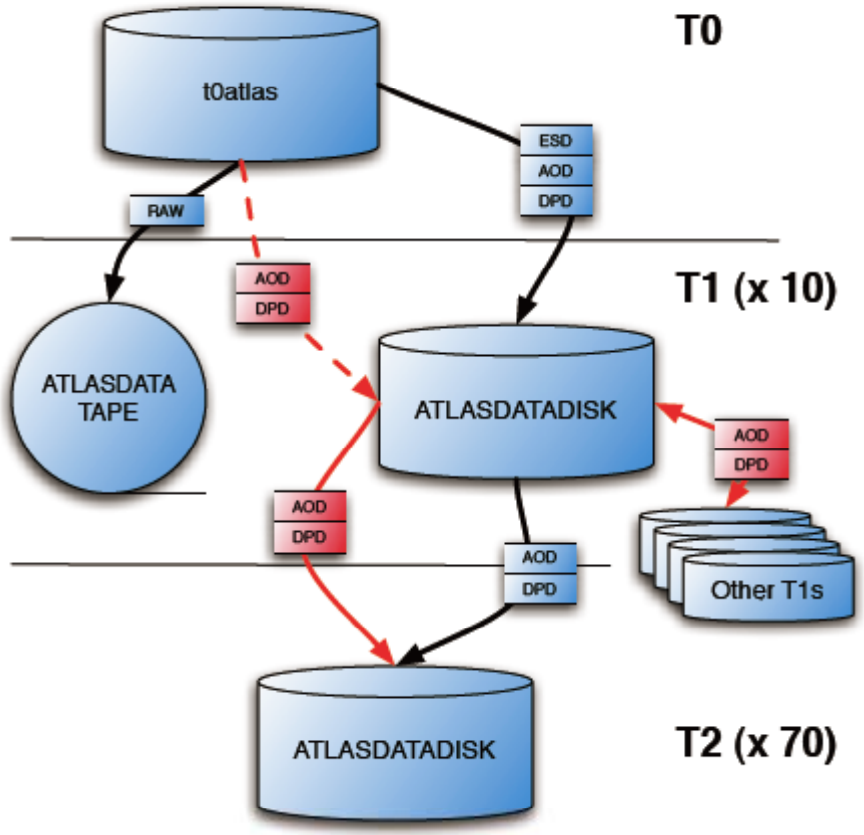

Commandline Tools
Enduser Tools
Production System
DQ2
Client API
Common Modular Framework
Site Services
Central Catalogues
Database
WLCG
Open Science Grid
LHC Computing Grid
NorduGrid

STEP09
Scale Test for the Experiment Program 09

- LHCの4実験が同時に演習
- 実際の実験時と同じ処理
  - 各段階のデータ(生→再構成→物理オブジェクト)配布
  - Tier1でのデータ再処理(テープ→ディスク)
  - モンテカルロプロダクション
  - ユーザ解析
  - ソフトウエアインフラ～データベース等

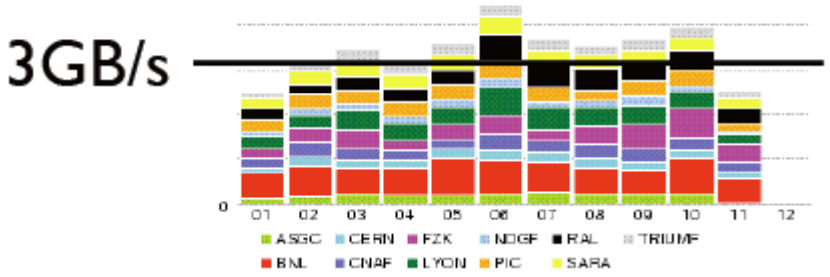
Data Distribution

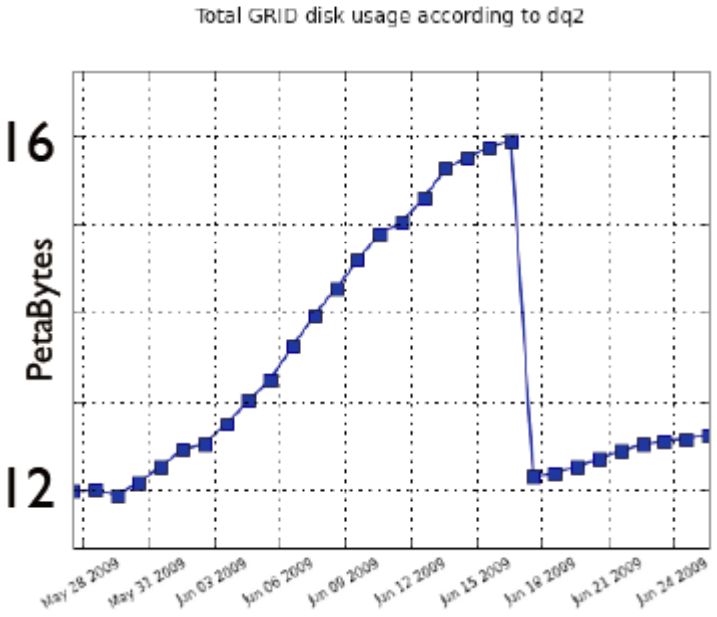
- Data taking and first reconstruction passes
  - RAW and ESD from CERN → distributed to T1 sites (1, 2 copies respectively, RAW to tape)
  - AOD and DPD from CERN distributed to all T1 sites (10 copies)
  - AOD and DPD from CERN distributed to T2 from their parent T1 (1 to 2.7 copies per cloud)
- Reprocessing at Tier-1s →
  - AOD and DPD distributed from all T1s to all other T1s
  - AOD and DPD from all T1s distributed to all T2s from their parent T1
    - In STEP this involved an extra T0→T1 step, but this is a minor perturbation ⇢



Data Distribution Results

- 4PB of data distributed
  - Large files! Large files!
- T1s all passed, some small problems in T1-T1 distribution identified and cured
- 54/67 T2 sites also made the metric (ranged from 100% to 5% share of data)
  - 13 fell short from slightly (99.7% complete) to catastrophically (25.9% complete)
  - Problems numerous: transfer service misconfiguration, SE instability, out of space, network bottlenecks





Cloud	Efficiency	Transfers Throughput
ASGC	99%	397 MB/s
BNL	84%	1128 MB/s
CERN	100%	334 MB/s
CNAF	98%	561 MB/s
FZK	85%	556 MB/s
LYON	96%	620 MB/s
NDGF	84%	137 MB/s
PIC	93%	429 MB/s
RAL	99%	838 MB/s
SARA	53%	262 MB/s
TRIUMF	100%	297 MB/s

Peaks of 5.5GB/s

Thursday, 9 July 2009 5

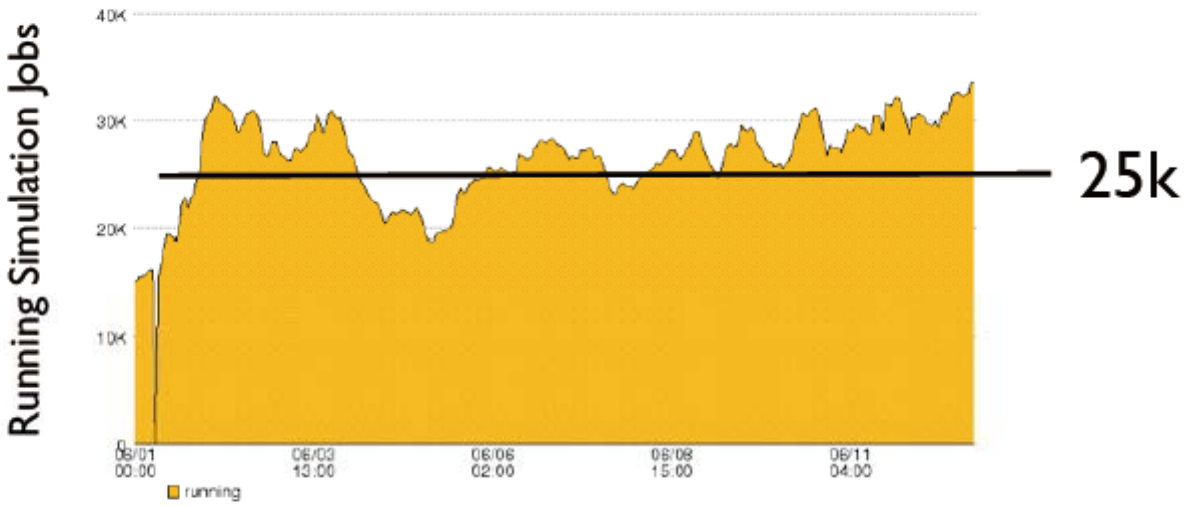
Reprocessing Results

T1	Base Target	Result	Comment
ASGC	10 000	4 782	Many batch system and basic setup problems
BNL + SLAC	50 000	99 276	Also ran high priority validation and other tasks
CNAF	10 000	29 997 ★	
FZK	20 000	17 954	Big tape system problems pre-STEP; no CMS
LYON	30 000	29 187	Very late start due to tape system upgrade, then good
NDGF	10 000	28 571 ★	
PIC	10 000	47 262 ★	
RAL	20 000	77 017 ★	
SARA	30 000	28 729	Tape system performance very patchy
TRIUMF	10 000	32 481 ★	Also ran high priority validation and other tasks

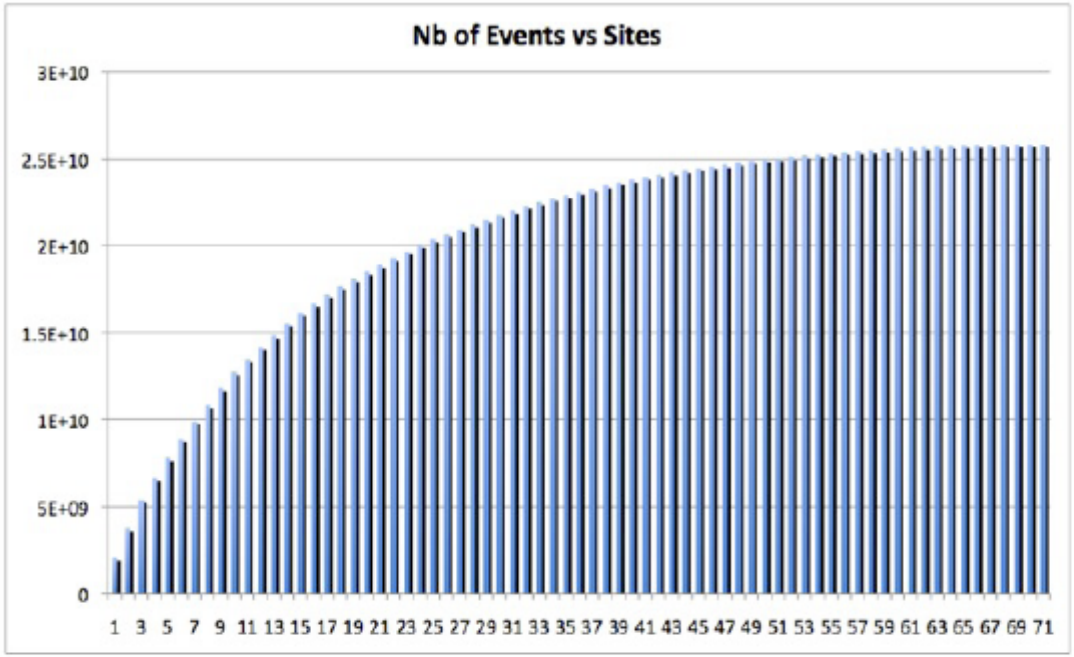
- Reprocessing from tape now validated in 6/10 T1s, with 3 more very close - improvements and retests planned
- Xavi will give details tomorrow

MonteCarlo Production

- Millions of hours of simulation production done
  - Production already well validated by increasingly large production runs
  - Operationally this is a solved problem
- N.B. Simulation filled all free resources to produce 12M events during STEP which matches ATLAS' mc09 requirements



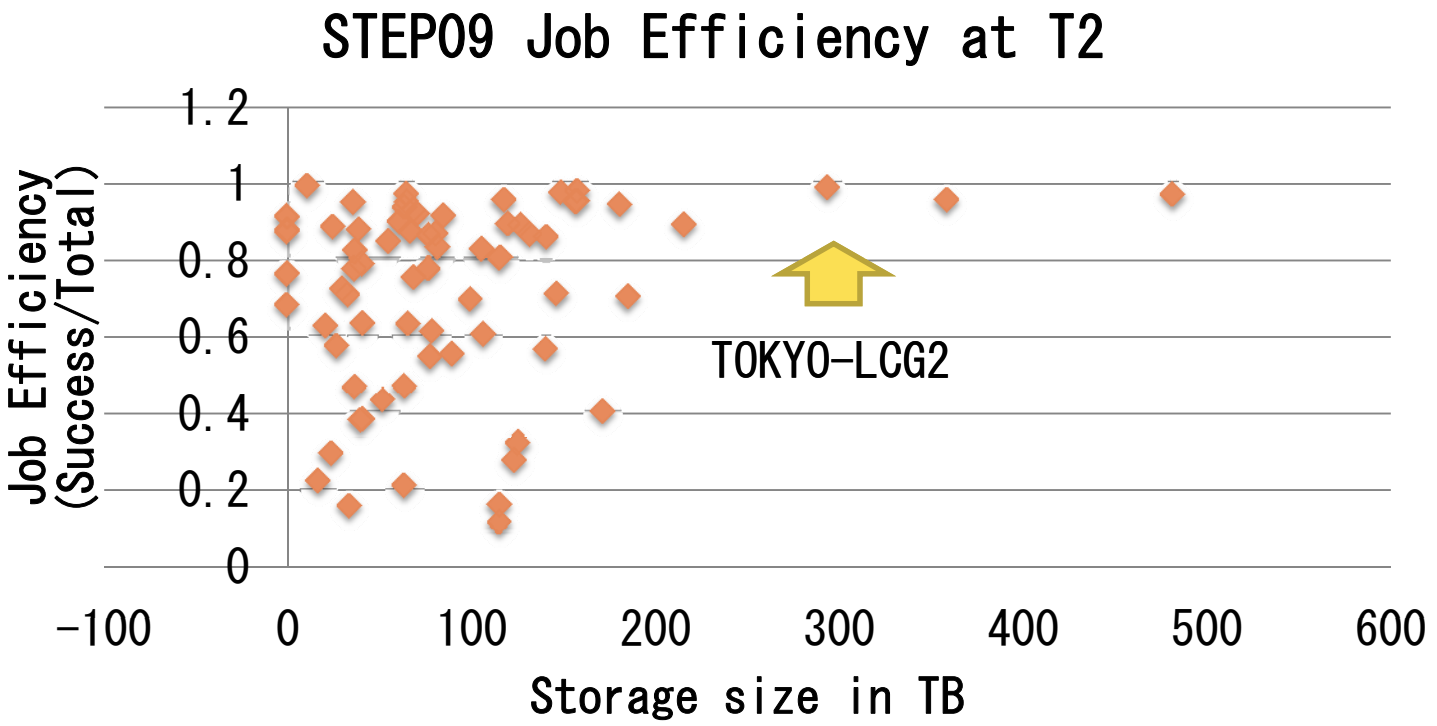
Cumulative Events per Site



- 50% of all events are processed in 11 sites
- 90% of all events are processed in 37 sites
- Of course, we know we have large and small sites, but...

Thursday, 9 July 2009 13

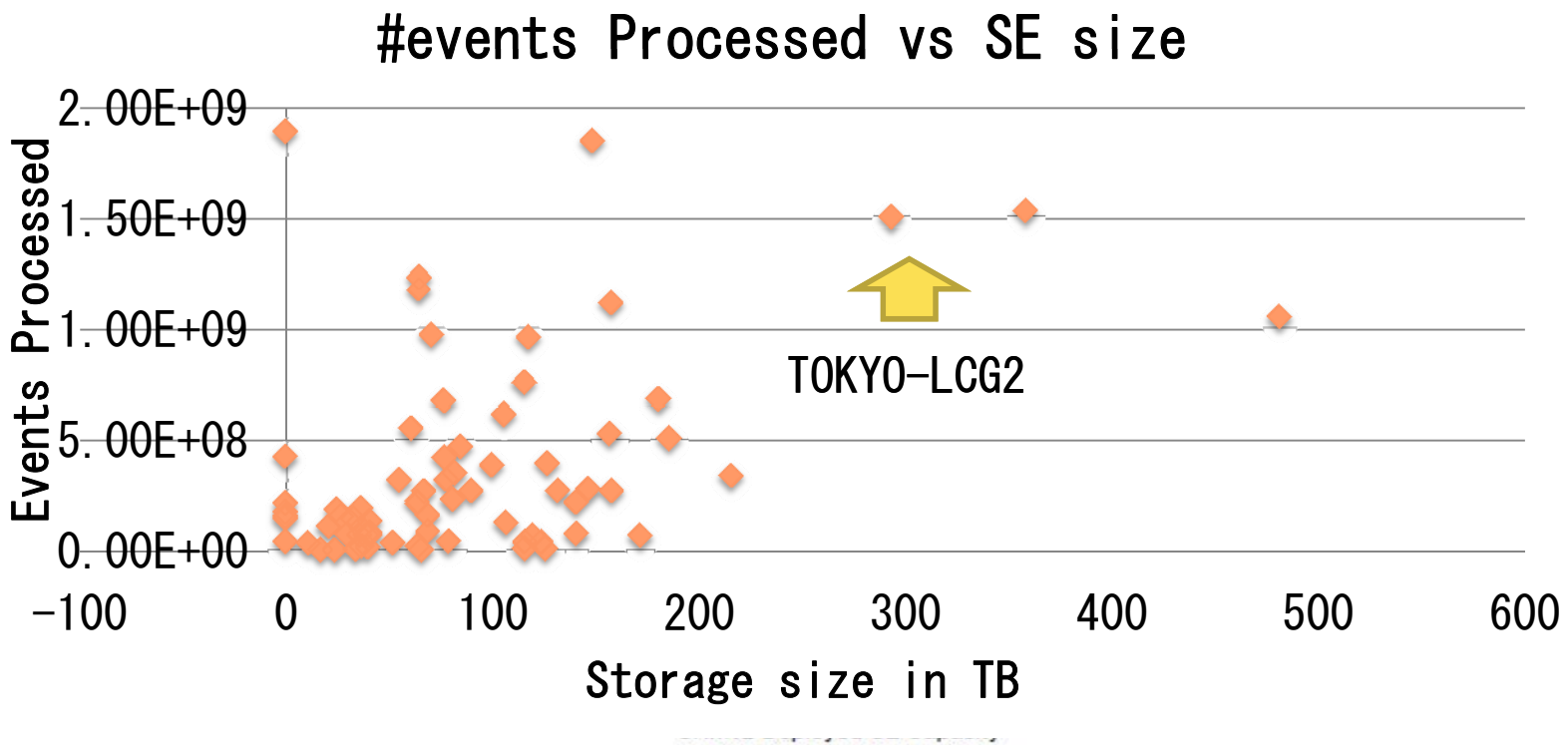
Visible T2 Site Resources I



- No strong correlation between job efficiency and storage resource size
  - Which is ok!
- N.B. Job efficiency is (success/total) not (cpu/wall)

Thursday, 9 July 2009 15

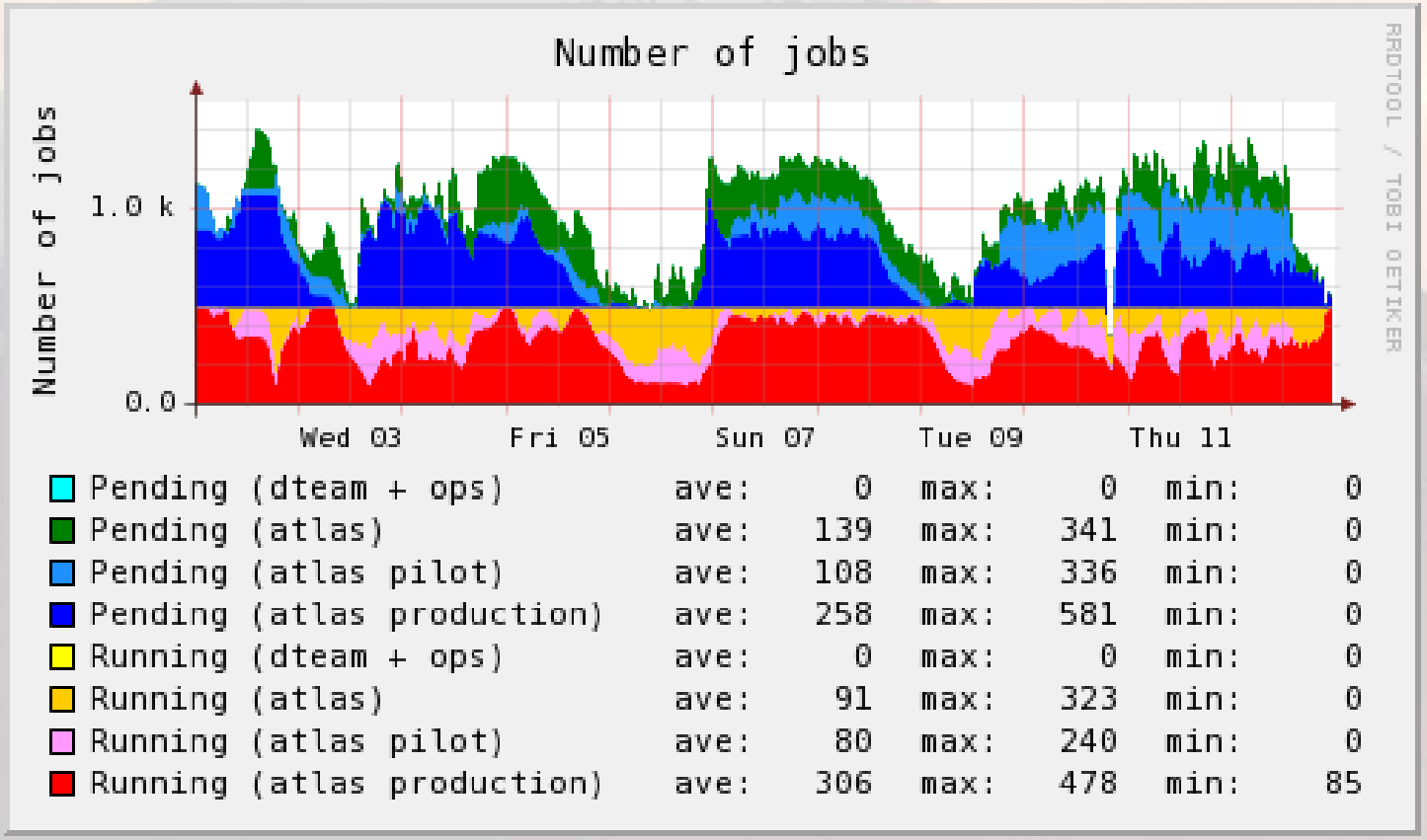
Visible T2 Site Resources II



- No strong correlation between storage size and number of events analysed
  - Which is bad!

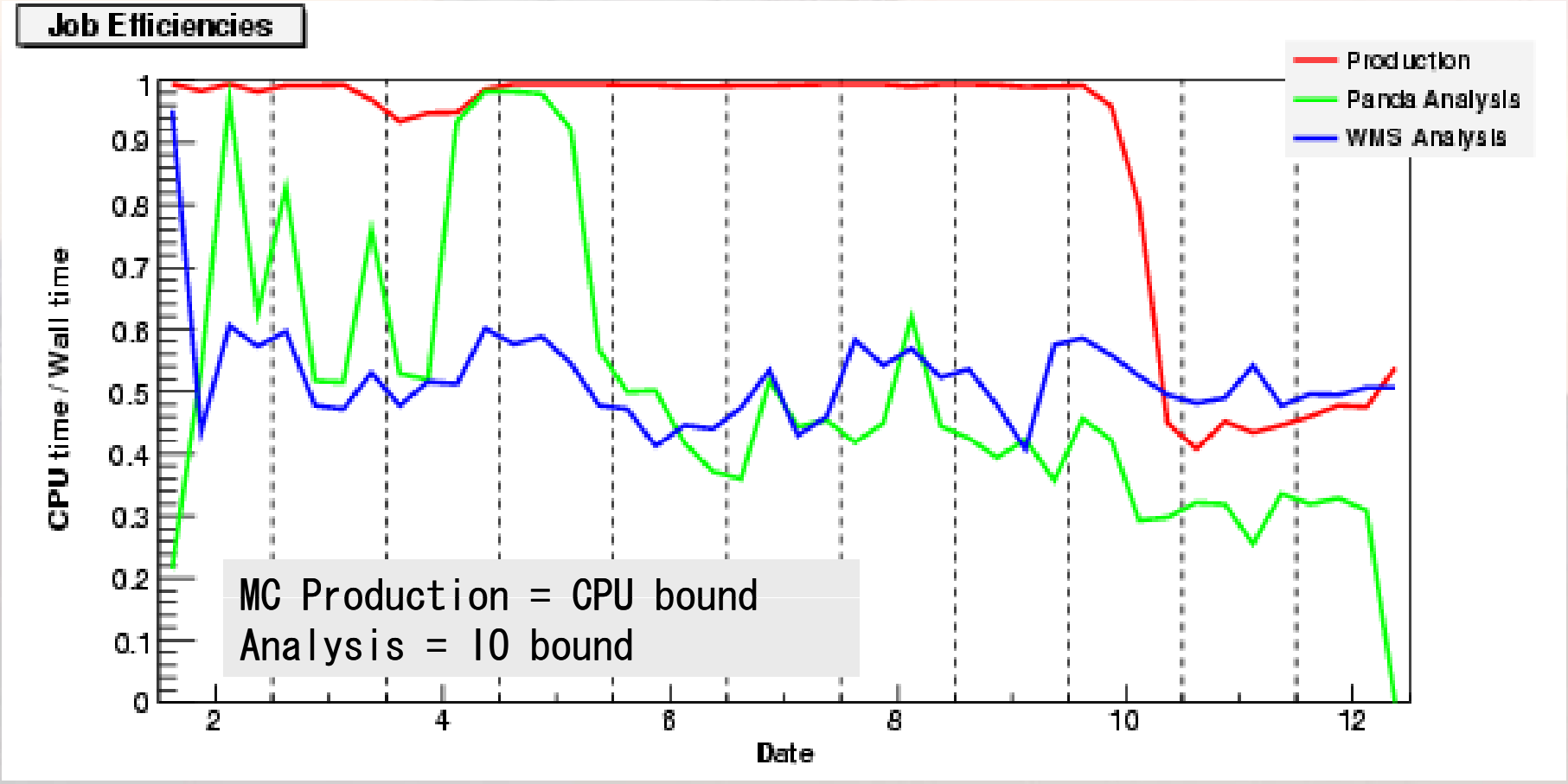
Thursday, 9 July 2009 16

Tokyo LCG2 in STEP09
Jobs

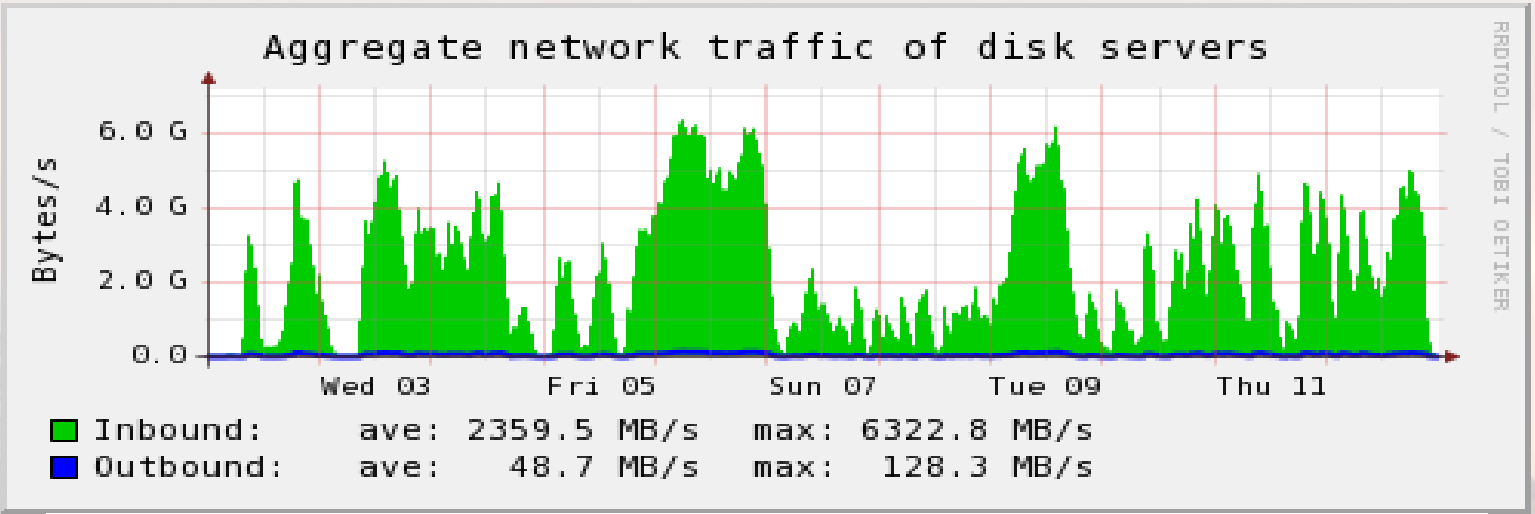


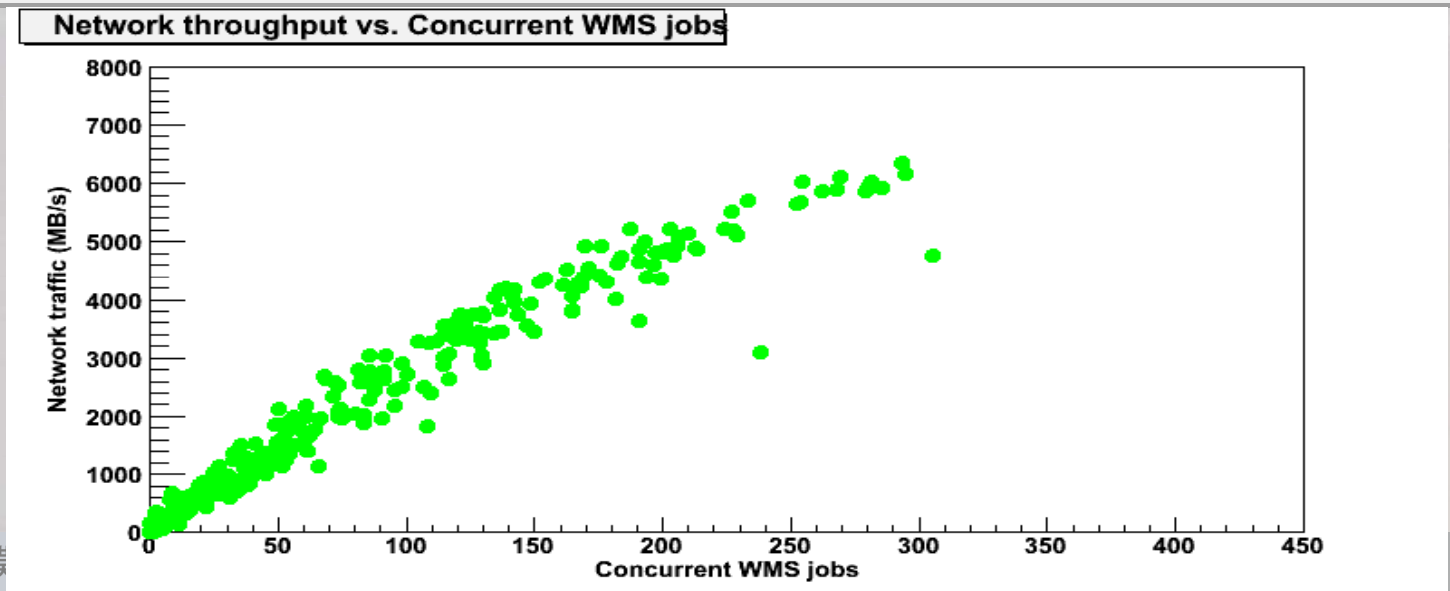
H. Matsunaga, STEP09 Post-mortem workshop Jul.

Job Efficiency = CPU time/Wall clock time

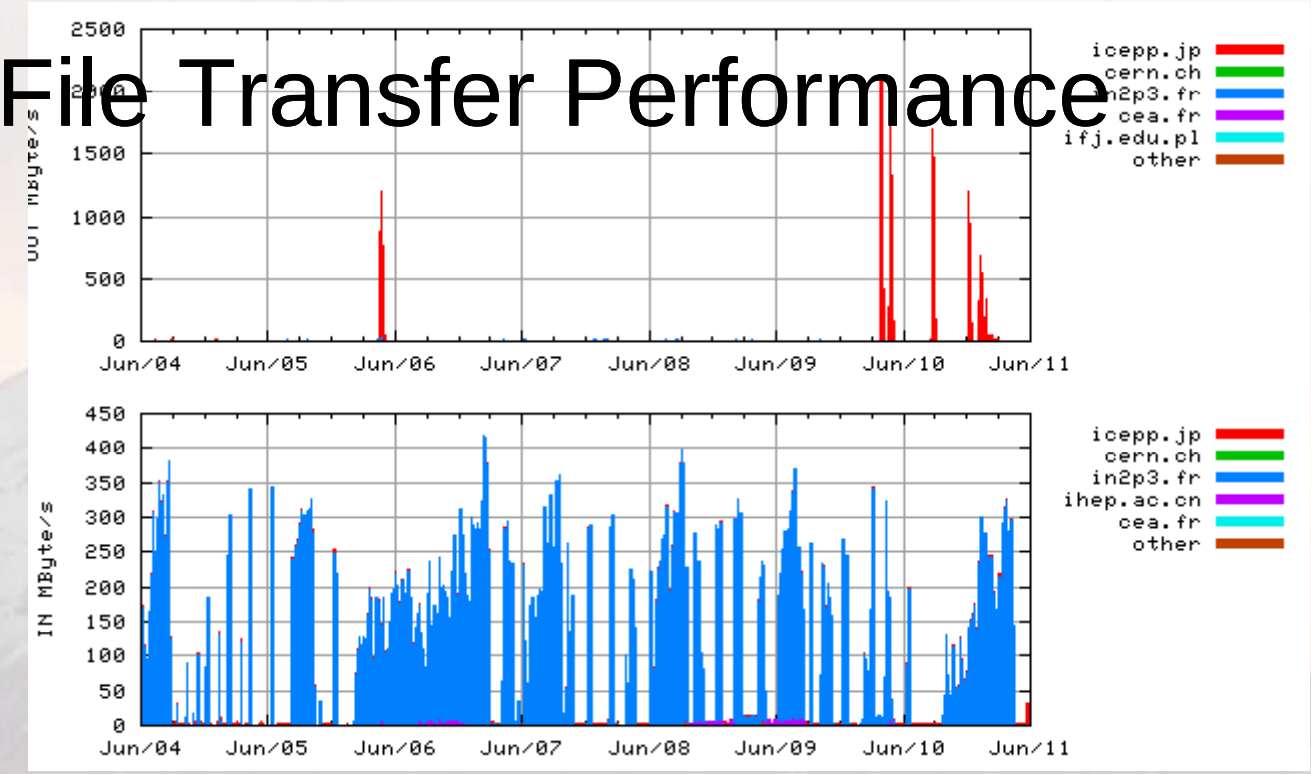


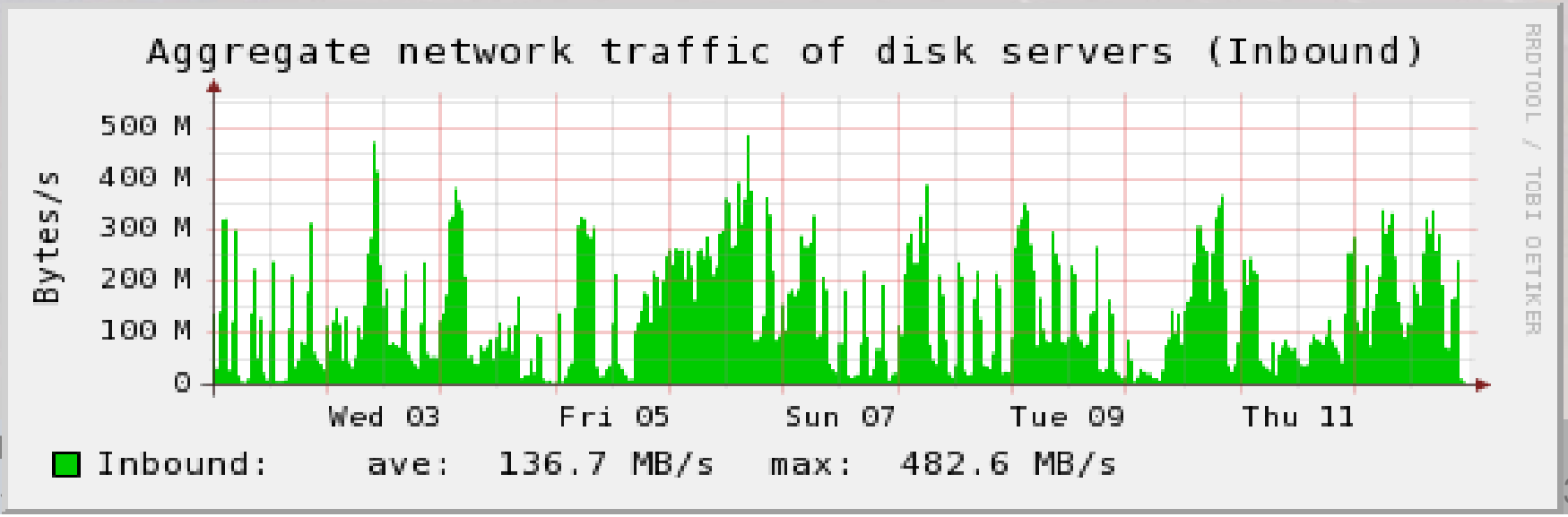
Local Area Network Throughput

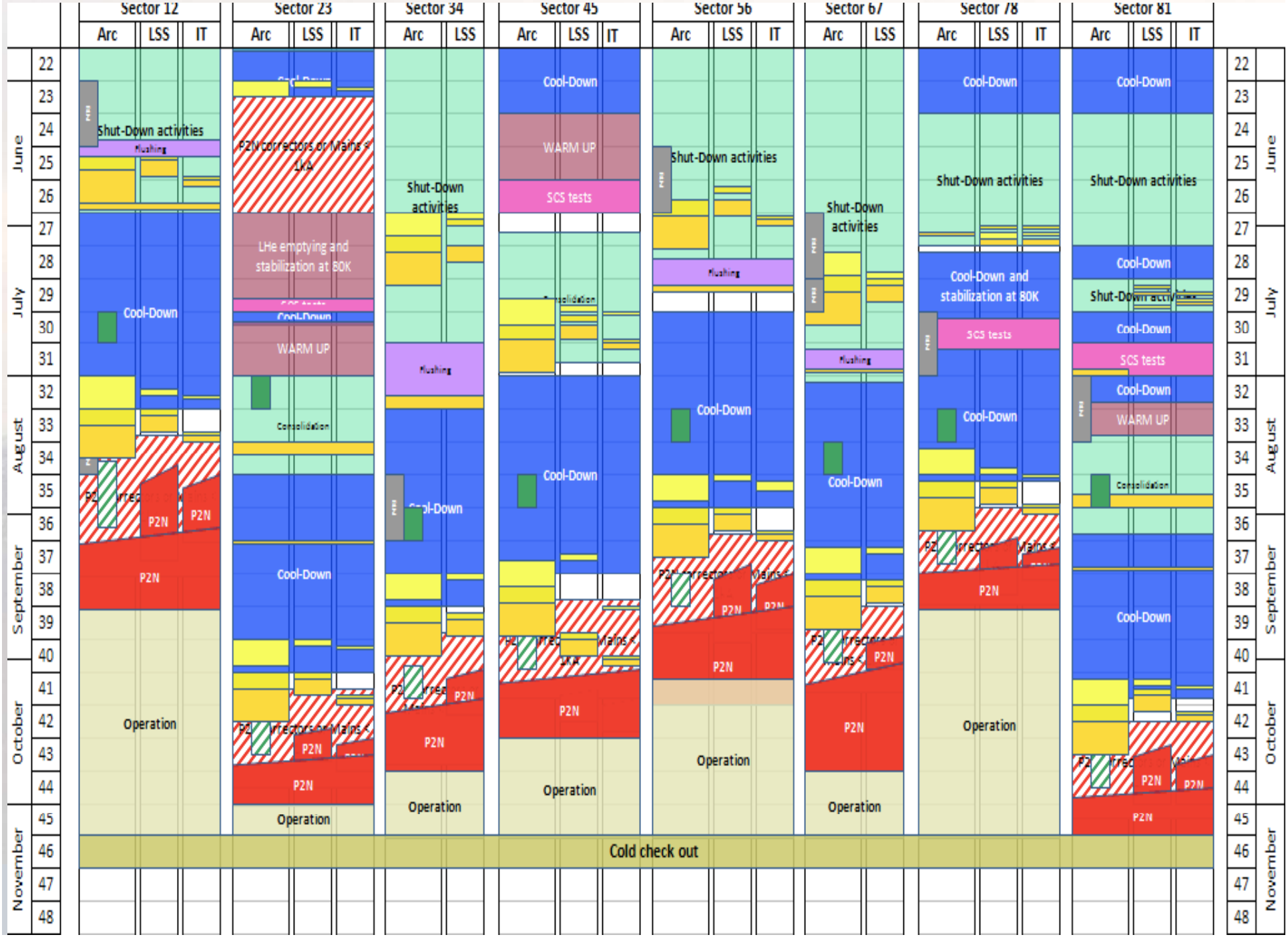




File Transfer Performance






Sector 12
Sector 23
Sector 34
Sector 45
Sector 56
Sector 67
Sector 78
Sector 81
Arc
LSS
IT
June
July
August
September
October
November
Shut-Down activities
Flushing
Cool-Down
WARM UP
SCS tests
LHe emptying and stabilization at 80K
Consolidation
Cool-Down and stabilization at 80K
P2N correctors or Mains < 1kA
P2N
Operation
Cold check out

Summary

- ATLAS detector is ready
- LHC computing gird is well tested
  - Successful STEP09 campaign
- LHC will restart operation soon
  - Accelerator will be turned on in November
  - First collision is expected in December
  - Full year run in 2010, at 3.5TeV (1/2 of target)