Estimations of the speedup of large data set analysis with geographically distributed computing facilities

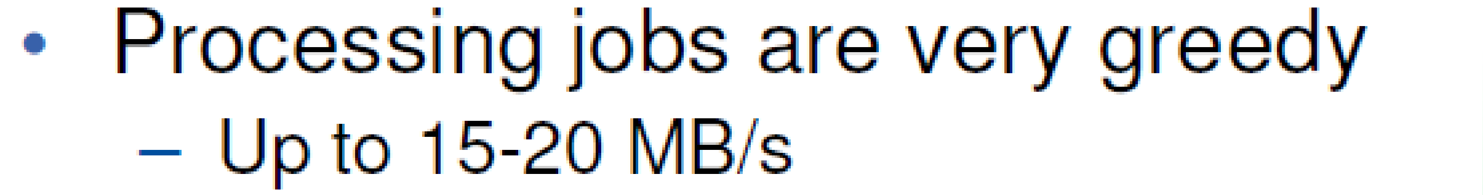
Andrey Y Shevel

Quick overview

- Importance of the matter
- Large analysis systems for HEP (examples)
- Scale of a number of used software packages
- Scale of required human resources
- Small physics analysis group (dozen+ physists)
- Estimation of the speedup and Consideration

The data channel bandwidth is hot topic (e.g. ACAT-2010 http://indico.cern.ch/conferenceDisplay.py?confId=59397)

- Michael Zerola et al «Building efficient Data Planner for Peta-scale Science»
- Jerome Lauret, Axel Nauman «Computing Technology for Physics Research»
- Fabrizio Furano «Data Access in the HEP community»



- Processing jobs are very greedy
  - – Up to 15-20 MB/s

The Event Data Model (EDM)

(ATLAS computing model)

- RAW:
    - "ByteStream” format, ~1.6 MB/event (~16 PB/year)
- ESD (Event Summary Data):
    - Full reconstraction, ~ 1MB/event (~1 PB/year)
- AOD (Analysis Object Data):
    - nominal size 100 kB/event (currently roughly double that) (> 0.1 PB/year)
- DPD (Derived Physics Data):
    - nominally 10 kB/event on average
        - Large variations depending on physics channels
- TAG:
    - nominal size 1 MB/event initially.

The Operational Model

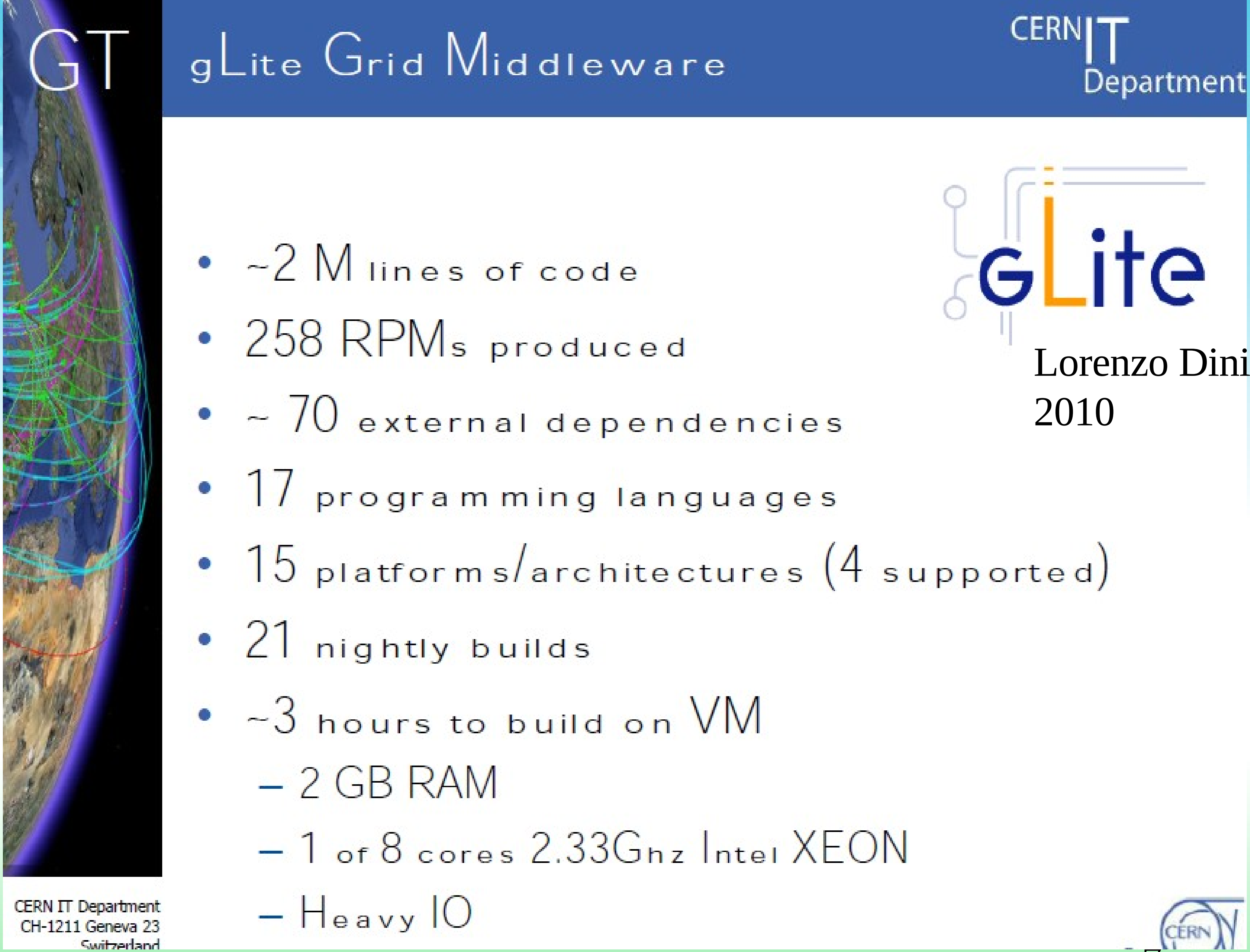
(ATLAS computing model)

- Tier-0 (CERN):
  - Copy RAW data to CERN Castor for archival & Tier-1s for storage and reprocessing
  - Run first-pass calibration/alignment
  - Run first-pass reconstruction (within 48 hrs)
  - Distribute reconstruction output (ESDs, AODs, DPDs & TAGS) to Tier-1s
- Tier-1 (x10):
  - Store and take care of a fraction of RAW data (forever)
  - Run “slow” calibration/alignment procedures
  - Rerun reconstruction with better calib/align and/or algorithms
  - Distribute reconstruction output to Tier-2s
  - Keep current versions of ESDs and AODs on disk for analysis
  - Run large-scale event selection and analysis jobs for physics and detector groups
  - Looks like some user access will be granted, but limited and NO ACCESS TO TAPE or LONG TERM STORAGE

Operational model - 2

(ATLAS computing model)

- Tier-2 (x~35):
  - Run analysis jobs (mainly AOD and DPD)
  - Run simulation (and calibration/alignment when/where appropriate)
  - Keep current versions of AODs and samples of other data types on disk for analysis
- **Tier-3**:
  - Provide access to Grid resources and local storage for end-user data
  - Contribute CPU cycles for simulation and analysis if/when possible



GT

gLite Grid Middleware

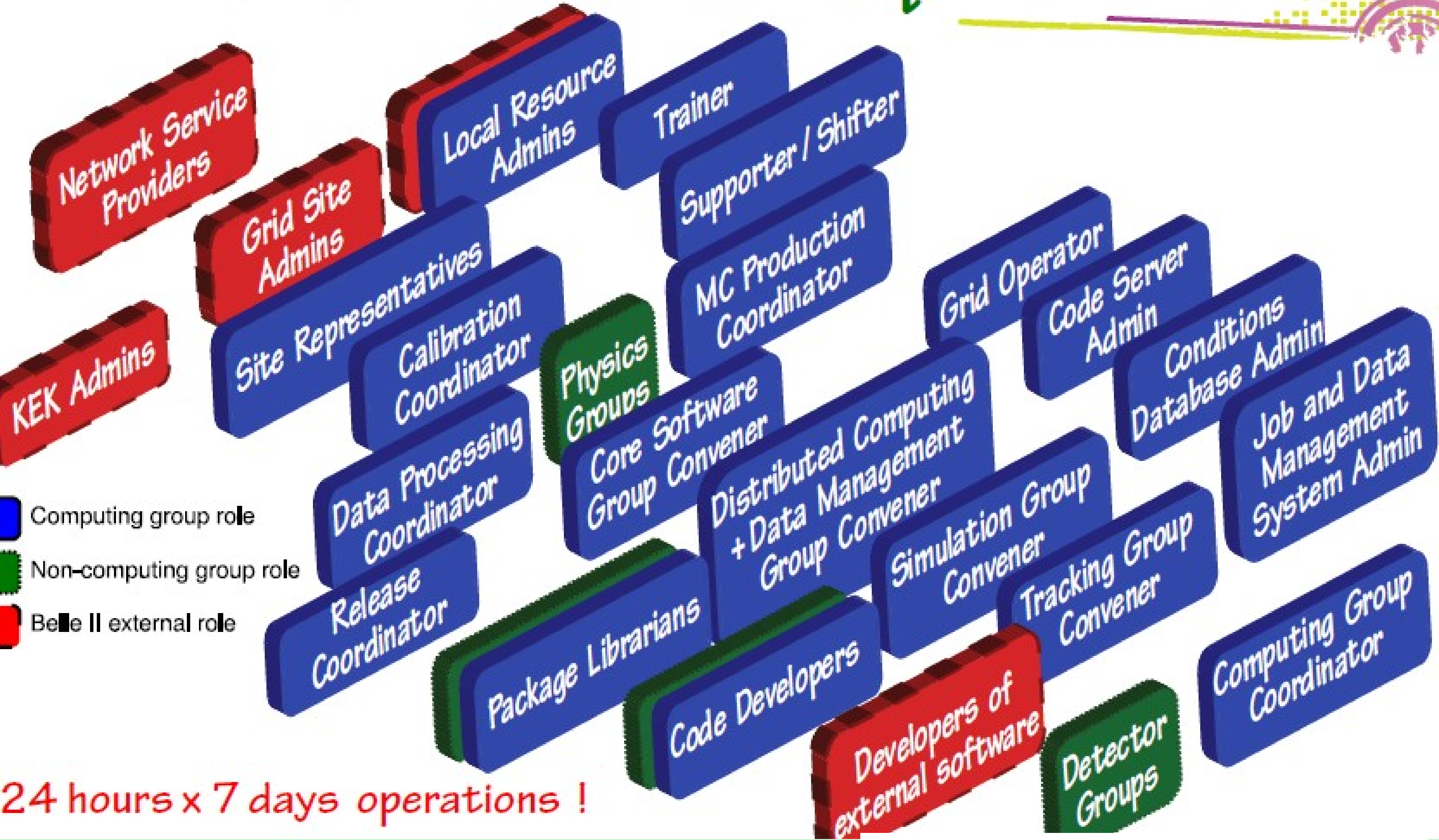
CERN IT Department

- ~2 M lines of code
- 258 RPMs produced
- ~ 70 external dependencies
- 17 programming languages
- 15 platforms/architectures (4 supported)
- 21 nightly builds
- ~3 hours to build on VM
  - 2 GB RAM
  - 1 of 8 cores 2.33Ghz Intel XEON
  - Heavy IO

CERN IT Department
CH-1211 Geneva 23
Switzerland

Lorenzo Dini
2010

Human resource requirement



Network Service Providers

Grid Site Admins

Local Resource Admins

Trainer

Supporter / Shifter

KEK Admins

Site Representatives

Calibration Coordinator

Physics Groups

MC Production Coordinator

Grid Operator

Code Server Admin

Conditions Database Admin

Job and Data Management System Admin

Data Processing Coordinator

Core Software Group Convener

Distributed Computing + Data Management Group Convener

Simulation Group Convener

Tracking Group Convener

Release Coordinator

Package Librarians

Code Developers

Developers of external software

Detector Groups

Computing Group Coordinator

🟦 Computing group role

🟩 Non-computing group role

🟥 Belle II external role

24 hours x 7 days operations !

Small physics group scenario

- To select some fraction of data on Tier-1
    - If possible it is better to analyse the data on the same cluster (let say «local» cluster)
  - However if the analysis ability is limited or not possible on «local» cluster user could try to do analysis on «local» + another cluster (or more clusters) [let say «remote»];
    - It is assumed that before start first analysis job on remote cluster you need to move required data to the cluster(s). The volume of the data does matter.

It is good to plan ...

- Data volumes:
  - *How much data will be required by one job*
  - *What is total volume of the data for one analysis job run*
- Data transfer to remote cluster:
  - What is real bandwidth available to you
    - is the bandwidth stable over time?
- In many cases such values might be taken into account almost automatically by replication system, but ...

Important conditions

- Let us assume that remote cluster has following features:
  - The remote cluster is more or less stable over time required for your analysis;
  - The data trasfer speed is more or less stable over time required to transfer of your data.
  - Remote managers are responsible and friendly persons who help you to get analysis done.

Further asumptions

- Let us introduce parameters:
  - T_{al} — the average time to analyse the portion of the data on local cluster
    - Portion is any part of data, for example, event or file with events
  - T_{ar} — the average time to analyse the portion of the data on remote cluster
  - T_{dt} — the time for data transfer of one portion of the data
  - T_{oo} — other time overheads
  - P_l — local cluster performance available to you = $1/T_{al}$
  - P_g — total performance (local cluster + remote cluster)

Trivial calculations

- $P_{ar}=1/(T_{dt}+T_{oo}+T_{ar})$
- $P_{g}=1/T_{al} + 1/(T_{dt}+T_{oo}+T_{ar})$
- Speedup = $P_{g}/P_{al} = [1/T_{al} + 1/(T_{dt}+T_{oo}+T_{ar})] / (1/T_{al})$
- Condition $P_{g}/P_{al} >= N$ (it is condition for spedup)
- Let N = 2 then $T_{al} >= 2 * (T_{dt}+T_{oo}+T_{ar})$
- In ideal situation $T_{oo} = T_{ar} =\sim 0$
- It gives us $T_{al} >= N * T_{dt}$ (data tranfer time is one of the key issues)

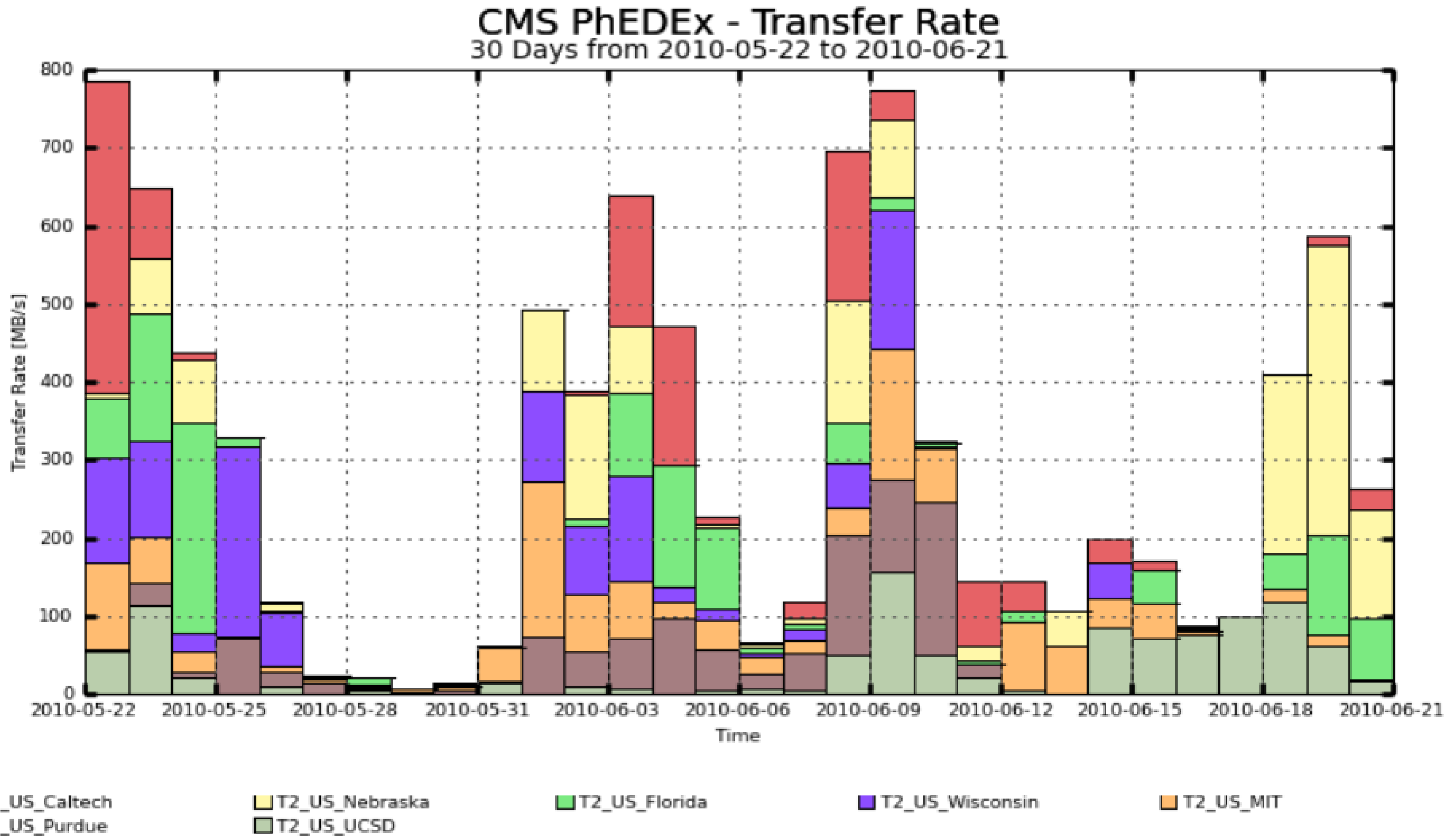
Example-US for bandwidth estimation

Graph Transfer Rate **by** Destination **filter source** **destination** T2_US

hide MSS nodes

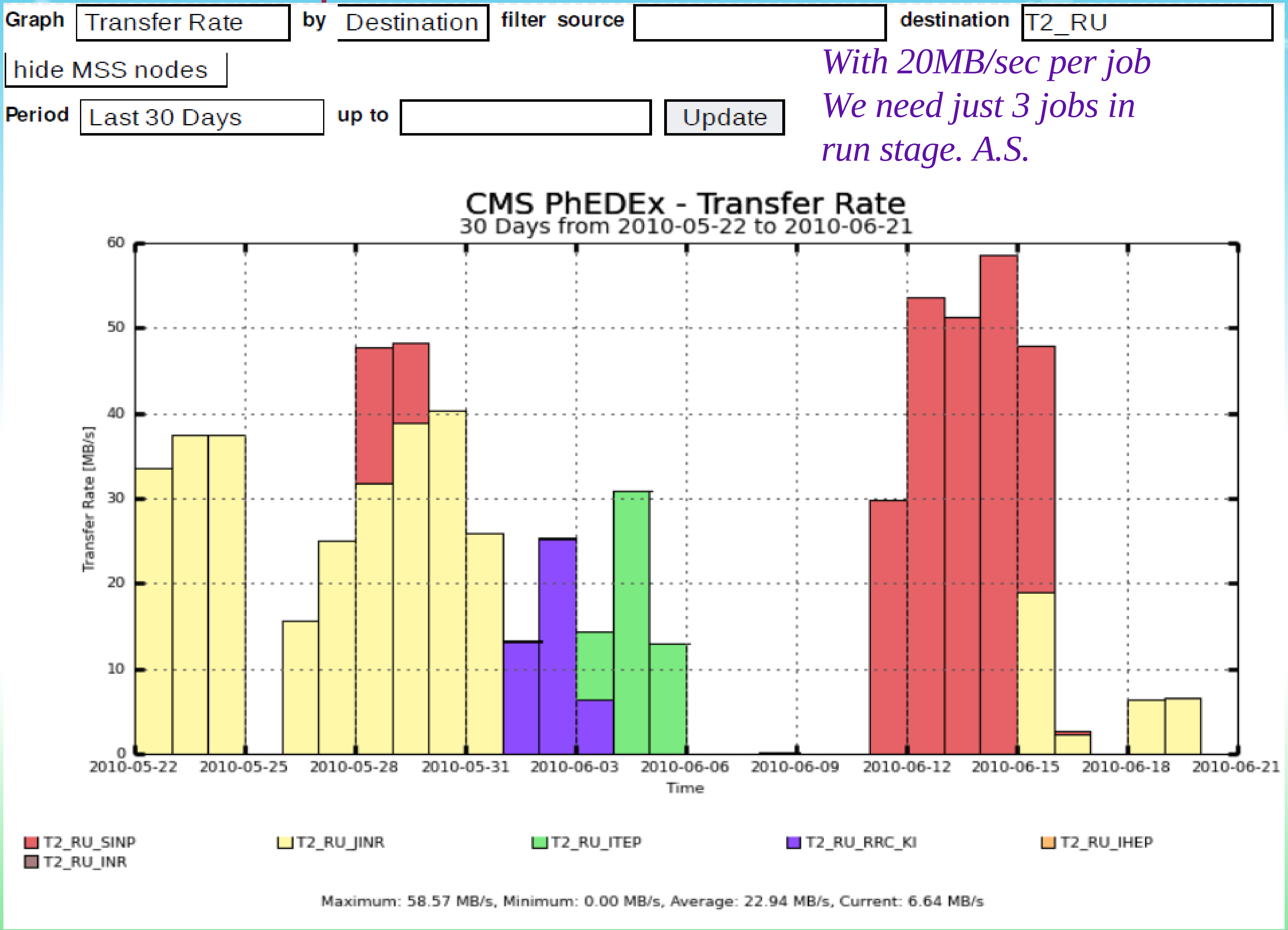
Period Last 30 Days **up to** Update

Taking into account 20MB/sec per job we need just 40 jobs in run stage. A.S.



Maximum: 786.54 MB/s, Minimum: 7.47 MB/s, Average: 295.18 MB/s, Current: 262.81 MB/s

Example-RU for bandwidth estimation



Graph: Transfer Rate by: Destination filter source: destination: T2_RU

hide MSS nodes

Period: Last 30 Days up to: Update

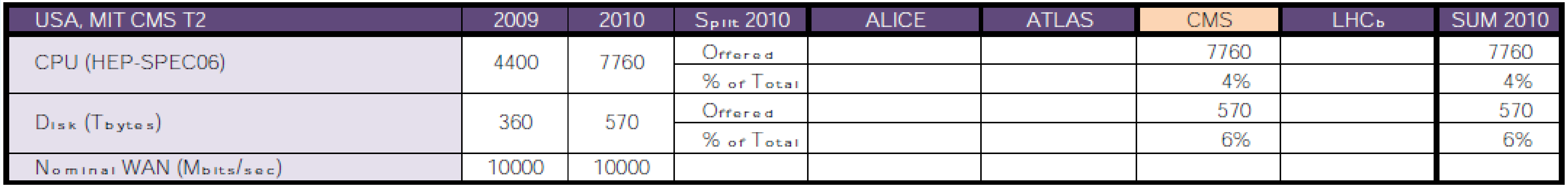
Maximum: 58.57 MB/s, Minimum: 0.00 MB/s, Average: 22.94 MB/s, Current: 6.64 MB/s

With 20MB/sec per job We need just 3 jobs in run stage. A.S.

Nominal bandwidth (US; RU)

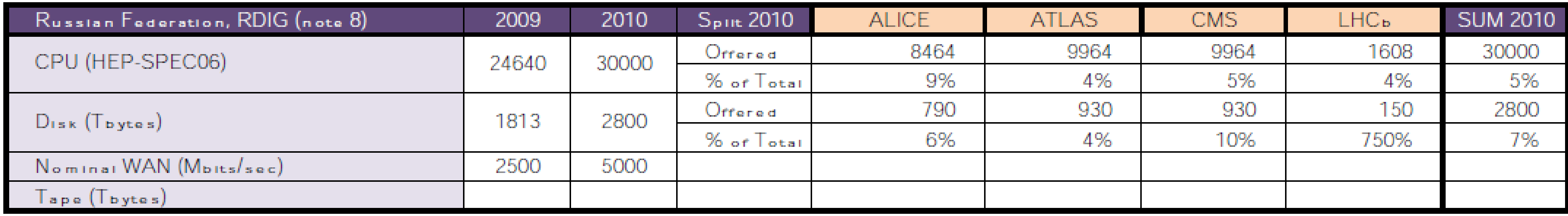
(http://lcg.web.cern.ch/LCG/Resources/WLCGResources-2009-2010_12APR10.pdf)

US



USA, MIT CMS T2	2009	2010	Split 2010	ALICE	ATLAS	CMS	LHCb	SUM 2010
CPU (HEP-SPEC06)	4400	7760	Offered			7760		7760
			% of Total			4%		4%
Disk (Tbytes)	360	570	Offered			570		570
			% of Total			6%		6%
Nominal WAN (Mbits/sec)	10000	10000						

RU

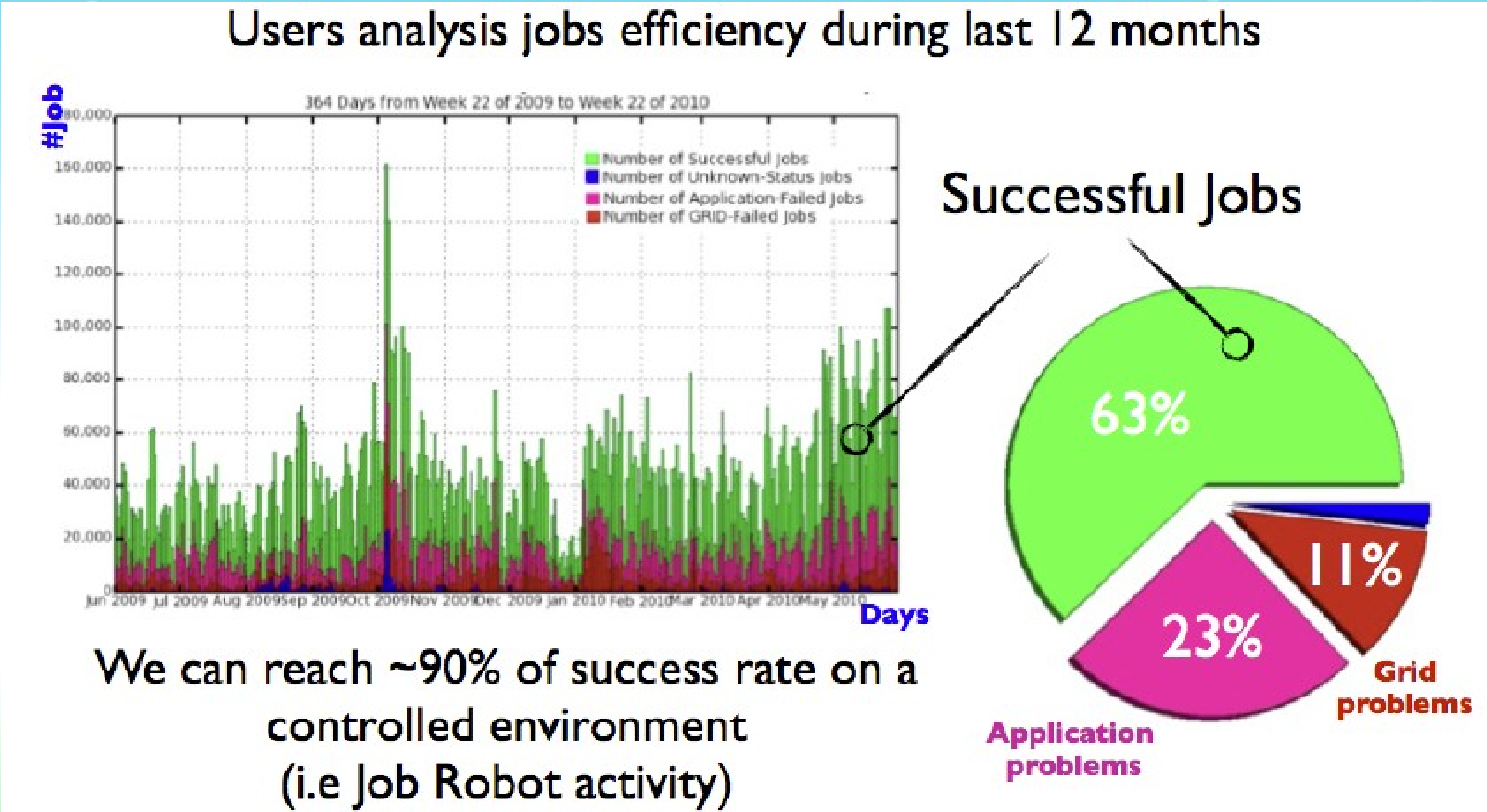


Russian Federation, RDIG (note 8)	2009	2010	Split 2010	ALICE	ATLAS	CMS	LHCb	SUM 2010
CPU (HEP-SPEC06)	24640	30000	Offered	8464	9964	9964	1608	30000
			% of Total	9%	4%	5%	4%	5%
Disk (Tbytes)	1813	2800	Offered	790	930	930	150	2800
			% of Total	6%	4%	10%	750%	7%
Nominal WAN (Mbits/sec)	2500	5000						
Tape (Tbytes)								

Considerations on the effective number of jobs

- Total volume of data for analysis (might vary)
  - 1 TB — 20 TB
- If we have around 20 MB/sec per job and we have centralized storage with access bandwidth 1 GB/sec *(BTW, it means 10**3 — 2*10**4 seconds just for data reading)*
- That means max number of jobs might be around
  - 10**3 MB/sec / 20 MB/sec = 50 jobs in run stage
- Intermediate conclusion: **data transfer bandwidth is important but more important the point of balance between bandwidth, computing power, interests, etc**

Case of remote cluster: important remark



Users analysis jobs efficiency during last 12 months

We can reach ~90% of success rate on a controlled environment (i.e Job Robot activity)

Kind of conclusion: appropriate cluster size for small physics laboratory

- About *dozen+* physicists who involved into real data analysis (run jobs, got new analysis results)
- **The small cluster** (~12-24 modern nodes) is more or less feasible solution for small physics laboratory
    - *Easy to reconfigure to fit the concete needs*
    - *Easy to maintain*
    - *Not expensive*
    - *Good as the gateway to external large computing facility (Tier 1 or so)*
    - *100-200 TB of disk space will be near perfect solution for analysis.*
- Some technology support for **Tier-3** (ability to prepare quite small slices of Dbs, data selection, etc)
- **Tier-3** will increase the performance of whole computing Grid.

Spare slide: how many cores per node is effective?

- If we have same asumption as before and have bandwidth for eth0 equal to 1 Gbit, that woud mean 1 Gbit ~ 100 MB/sec / 20 MB/sec ~ 5 jobs.
- In other words 8 cores is more than enough for our conditions with one network interface 1 Gbit .

Thank you !   Questions?