LHCONE Point-to-Point Service Workshop:
a brief summary

Daniele Bonacorsi
[University of Bologna, Italy - deputy CMS Computing coordinator]

*[CAVEAT: this is my summary as an attendant, and **not** an executive summary on the workshop.*
The latter is being prepared by the workshop chairs, and will be circulated later on.]

General

The workshop took place on December 13th-14th, 2012:

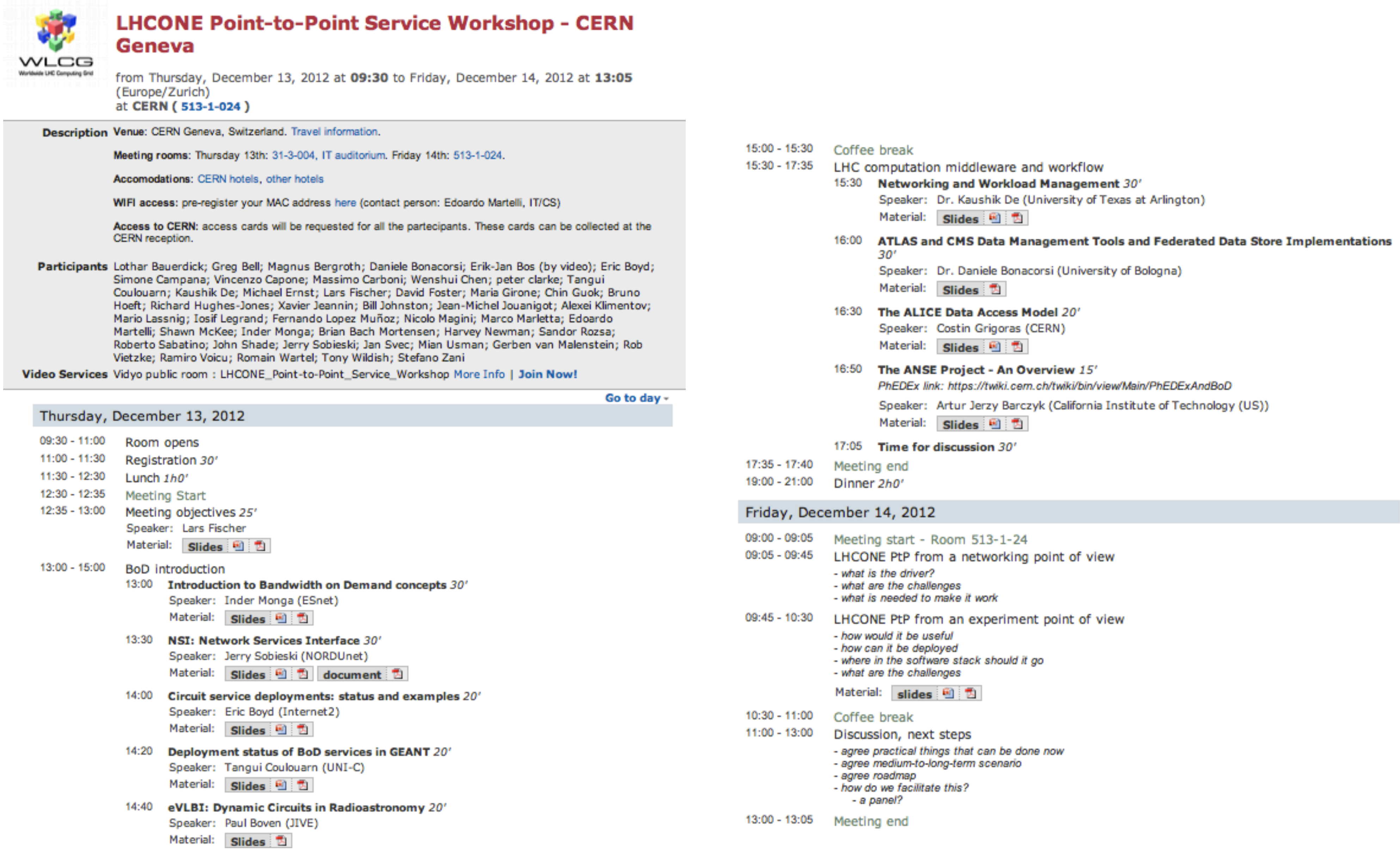
- ✦ http://indico.cern.ch/conferenceOtherViews.py?view=lcg&confId=215393
- ✦ ~50 participants
  - representatives from Network community and NRENs, LHC experiments, CERN-IT, and more

In my opinion, it was an efficiently chaired workshop with a wisely-mixed content:

- ✦ An effective balance of talks vs discussion slots:
  - several talks by Network experts + few (summary) talks by experiments
  - *plenty* of discussion time, especially on the second day
- ✦ A good mixture of “Network(s)” vs “Experiment(s)” flavors
  - Workload Management and Data Management/Access aspects in experiments
  - BoD from concepts to existing solutions, projects, deployment ideas, etc

Agenda

http://indico.cern.ch/conferenceOtherViews.py?view=lcg&confId=215393



LHCONE Point-to-Point Service Workshop - CERN Geneva

from Thursday, December 13, 2012 at **09:30** to Friday, December 14, 2012 at **13:05**
(Europe/Zurich)
at **CERN (513-1-024)**

Description **Venue**: CERN Geneva, Switzerland. Travel information.

Meeting rooms: Thursday 13th: 31-3-004, IT auditorium. Friday 14th: 513-1-024.

Accomodations: CERN hotels, other hotels

WIFI access: pre-register your MAC address here (contact person: Edoardo Martelli, IT/CS)

Access to CERN: access cards will be requested for all the partecipants. These cards can be collected at the CERN reception.

Participants Lothar Bauerdick; Greg Bell; Magnus Bergroth; Daniele Bonacorsi; Erik-Jan Bos (by video); Eric Boyd; Simone Campana; Vincenzo Capone; Massimo Carboni; Wenshui Chen; peter clarke; Tangui Coulouarn; Kaushik De; Michael Ernst; Lars Fischer; David Foster; Maria Girone; Chin Guok; Bruno Hoeft; Richard Hughes-Jones; Xavier Jeannin; Bill Johnston; Jean-Michel Jouanigot; Alexei Klimentov; Mario Lassnig; Iosif Legrand; Fernando Lopez Muñoz; Nicolo Magini; Marco Marletta; Edoardo Martelli; Shawn McKee; Inder Monga; Brian Bach Mortensen; Harvey Newman; Sandor Rozsa; Roberto Sabatino; John Shade; Jerry Sobieski; Jan Svec; Mian Usman; Gerben van Malenstein; Rob Vietzke; Ramiro Voicu; Romain Wartel; Tony Wildish; Stefano Zani

Video Services Vidyo public room : LHCONE_Point-to-Point_Service_Workshop More Info | **Join Now!**

Go to day

Thursday, December 13, 2012

09:30 - 11:00 Room opens

11:00 - 11:30 Registration *30'*

11:30 - 12:30 Lunch *1h0'*

12:30 - 12:35 Meeting Start

12:35 - 13:00 Meeting objectives *25'*
Speaker: Lars Fischer
Material: **Slides**

13:00 - 15:00 BoD introduction

13:00 **Introduction to Bandwidth on Demand concepts** *30'*
Speaker: Inder Monga (ESnet)
Material: **Slides**

13:30 **NSI: Network Services Interface** *30'*
Speaker: Jerry Sobieski (NORDUnet)
Material: **Slides** **document**

14:00 **Circuit service deployments: status and examples** *20'*
Speaker: Eric Boyd (Internet2)
Material: **Slides**

14:20 **Deployment status of BoD services in GEANT** *20'*
Speaker: Tangui Coulouarn (UNI-C)
Material: **Slides**

14:40 **eVLBI: Dynamic Circuits in Radioastronomy** *20'*
Speaker: Paul Boven (JIVE)
Material: **Slides**

15:00 - 15:30 Coffee break

15:30 - 17:35 LHC computation middleware and workflow

15:30 **Networking and Workload Management** *30'*
Speaker: Dr. Kaushik De (University of Texas at Arlington)
Material: **Slides**

16:00 **ATLAS and CMS Data Management Tools and Federated Data Store Implementations** *30'*
Speaker: Dr. Daniele Bonacorsi (University of Bologna)
Material: **Slides**

16:30 **The ALICE Data Access Model** *20'*
Speaker: Costin Grigoras (CERN)
Material: **Slides**

16:50 **The ANSE Project - An Overview** *15'*
PhEDEx link: https://twiki.cern.ch/twiki/bin/view/Main/PhEDExAndBoD
Speaker: Artur Jerzy Barczyk (California Institute of Technology (US))
Material: **Slides**

17:05 **Time for discussion** *30'*

17:35 - 17:40 Meeting end

19:00 - 21:00 Dinner *2h0'*

Friday, December 14, 2012

09:00 - 09:05 Meeting start - Room 513-1-24

09:05 - 09:45 LHCONE PtP from a networking point of view
- what is the driver?
- what are the challenges
- what is needed to make it work

09:45 - 10:30 LHCONE PtP from an experiment point of view
- how would it be useful
- how can it be deployed
- where in the software stack should it go
- what are the challenges
Material: **slides**

10:30 - 11:00 Coffee break

11:00 - 13:00 Discussion, next steps
- agree practical things that can be done now
- agree medium-to-long-term scenario
- agree roadmap
- how do we facilitate this?
- a panel?

13:00 - 13:05 Meeting end

What’s common in DM/WM ?

E.g. ATLAS and CMS DM/WM concepts and implementations are similar in many ways

- ... and getting more and more similar in Computing Models evolutions
  - Both in DM and in WM
  - Both in the main concepts and in the operational choices

- A non-exhaustive list:
  - towards a “mesh” model of data transfers
  - evaluation of FTS3 as a file-level transfer service (as done for FTS-1/2)
  - transition from DQ2 to Rucio in ATLAS (e.g. catalogue evolutions)
  - local data access complemented with WAN access
  - deployment of data federations (i.e. AAA in CMS, FAX in ATLAS)
  - disk vs tape separation
  - work on “Common Solutions” (PanDA and CRAB3) for the ATLAS/CMS WM sector
  - more “dynamic” data placement tactics (subscriptions and deletions based on popularity data)
  - ...

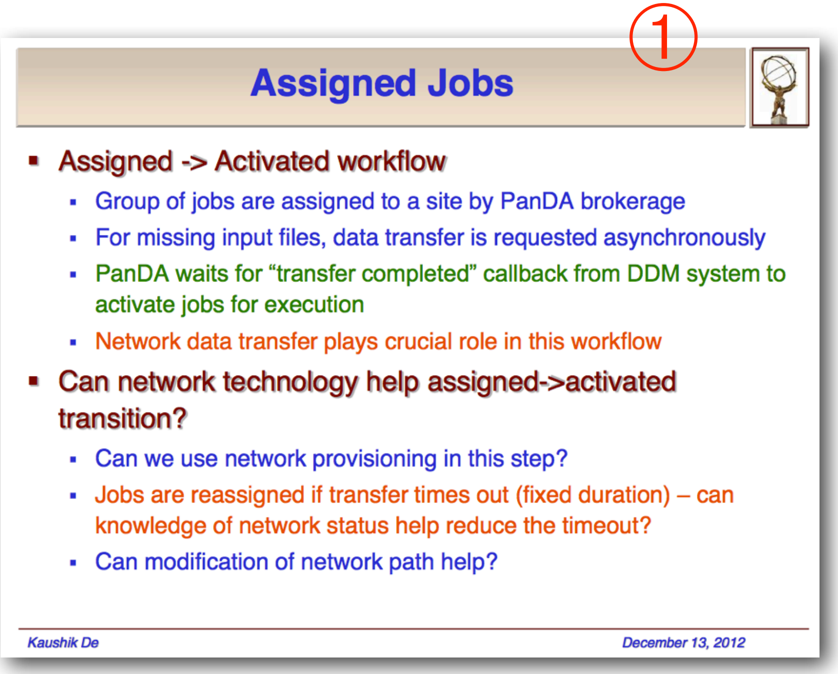
Thoughts on ATLAS/CMS WM

http://indico.cern.ch/getFile.py/access?contribId=5&sessionId=1&resId=0&materialId=slides&confId=215393

Kaushik presented examples based on the PanDA experience, but general in many ways

- ✦ May network provisioning help in quickly completing incomplete input datasets? ①
- ✦ May network status info help to decide to wait for transfer jobs to complete, or rerun? ②
- ✦ Should we consider network as a resource in brokerage? ③

①

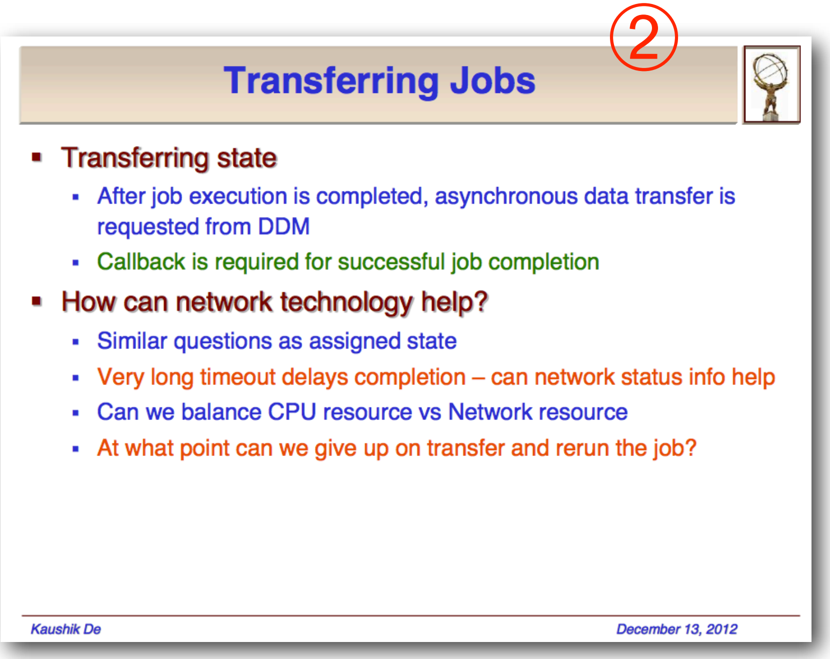


Assigned Jobs

- Assigned -> Activated workflow
  - Group of jobs are assigned to a site by PanDA brokerage
  - For missing input files, data transfer is requested asynchronously
  - PanDA waits for “transfer completed” callback from DDM system to activate jobs for execution
  - Network data transfer plays crucial role in this workflow
- Can network technology help assigned->activated transition?
  - Can we use network provisioning in this step?
  - Jobs are reassigned if transfer times out (fixed duration) – can knowledge of network status help reduce the timeout?
  - Can modification of network path help?

Kaushik De December 13, 2012

②

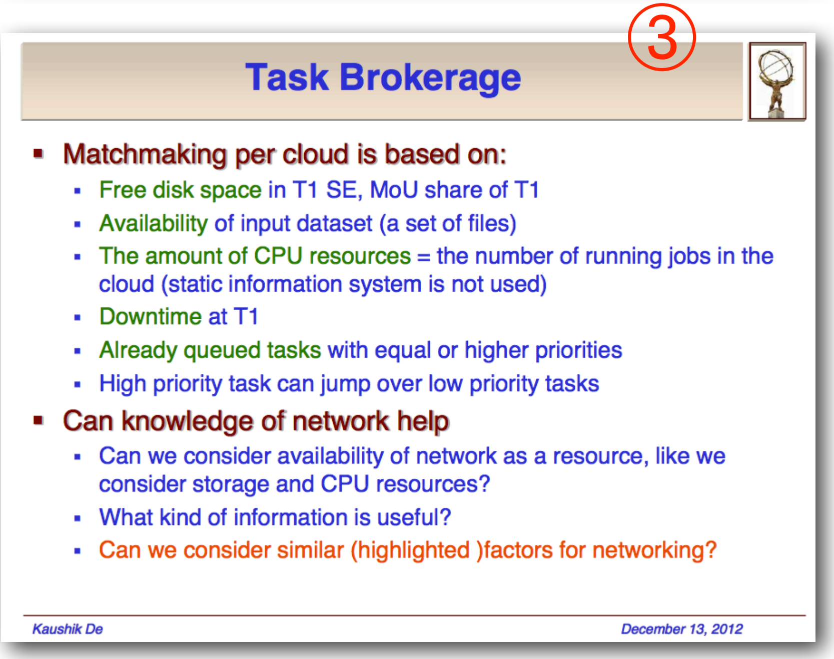


Transferring Jobs

- Transferring state
  - After job execution is completed, asynchronous data transfer is requested from DDM
  - Callback is required for successful job completion
- How can network technology help?
  - Similar questions as assigned state
  - Very long timeout delays completion – can network status info help
  - Can we balance CPU resource vs Network resource
  - At what point can we give up on transfer and rerun the job?

Kaushik De December 13, 2012

③



Task Brokerage

- Matchmaking per cloud is based on:
  - Free disk space in T1 SE, MoU share of T1
  - Availability of input dataset (a set of files)
  - The amount of CPU resources = the number of running jobs in the cloud (static information system is not used)
  - Downtime at T1
  - Already queued tasks with equal or higher priorities
  - High priority task can jump over low priority tasks
- Can knowledge of network help
  - Can we consider availability of network as a resource, like we consider storage and CPU resources?
  - What kind of information is useful?
  - Can we consider similar (highlighted)factors for networking?

Kaushik De December 13, 2012

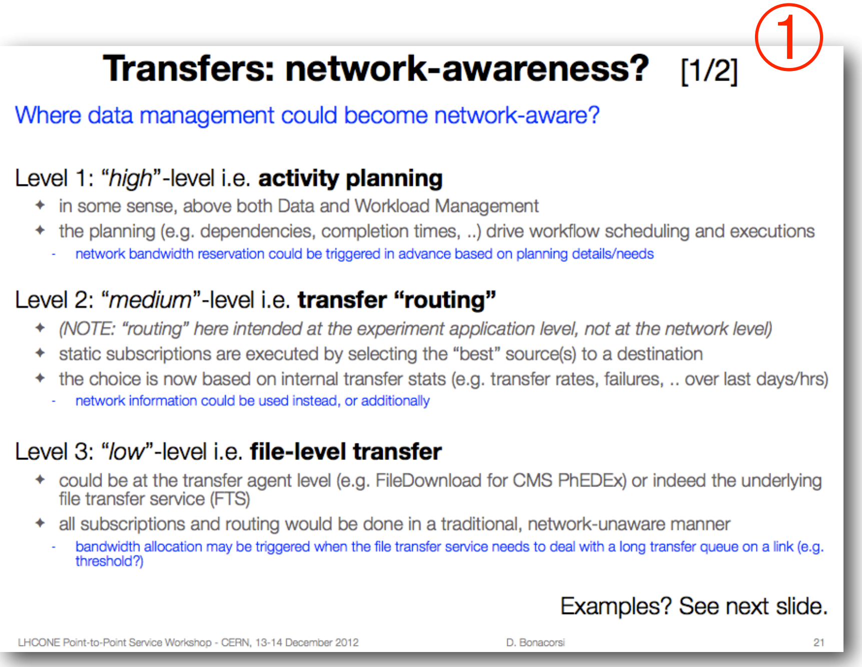
Thoughts on ATLAS/CMS DM

http://indico.cern.ch/getFile.py/access?contribId=6&sessionId=1&resId=0&materialId=slides&confId=215393

I presented a comparison of ATLAS/CMS DM systems, also in their evolutions

Plus some thoughts as of whether network-awareness may help in DM and "at which level"

- "Where" in the DM components some network-awareness could be plugged in? ①
- Possible paths to explore, for each of the "levels" above ②
- Network-awareness in data federations? ③



①

Transfers: network-awareness? [1/2]

Where data management could become network-aware?

Level 1: "*high*"-level i.e. **activity planning**

- in some sense, above both Data and Workload Management
- the planning (e.g. dependencies, completion times, ..) drive workflow scheduling and executions
  - network bandwidth reservation could be triggered in advance based on planning details/needs

Level 2: "*medium*"-level i.e. **transfer "routing"**

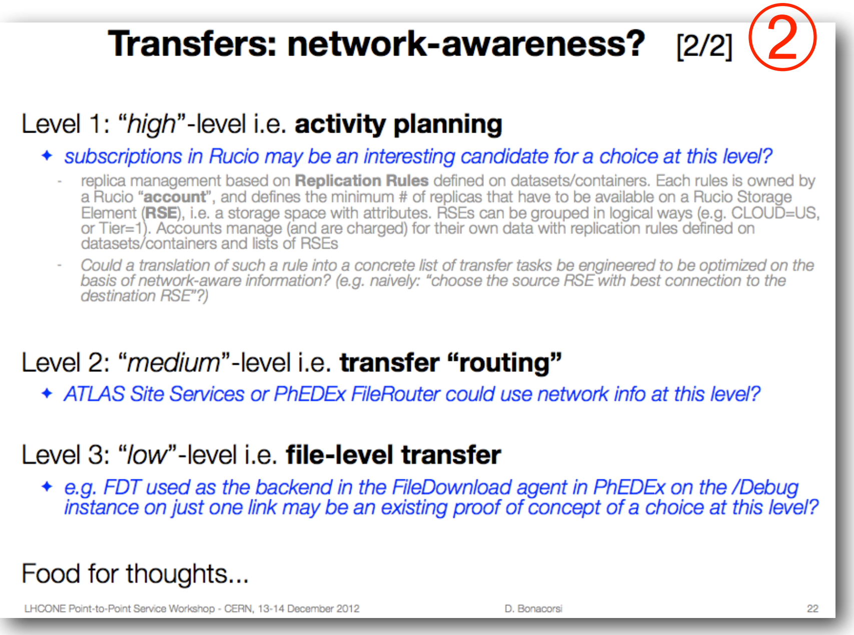
- *(NOTE: "routing" here intended at the experiment application level, not at the network level)*
- static subscriptions are executed by selecting the "best" source(s) to a destination
- the choice is now based on internal transfer stats (e.g. transfer rates, failures, .. over last days/hrs)
  - network information could be used instead, or additionally

Level 3: "*low*"-level i.e. **file-level transfer**

- could be at the transfer agent level (e.g. FileDownload for CMS PhEDEx) or indeed the underlying file transfer service (FTS)
- all subscriptions and routing would be done in a traditional, network-unaware manner
  - bandwidth allocation may be triggered when the file transfer service needs to deal with a long transfer queue on a link (e.g. threshold?)

Examples? See next slide.

LHCONE Point-to-Point Service Workshop - CERN, 13-14 December 2012 D. Bonacorsi 21



Transfers: network-awareness? [2/2] ②

Level 1: "*high*"-level i.e. **activity planning**

- *subscriptions in Rucio may be an interesting candidate for a choice at this level?*
  - replica management based on **Replication Rules** defined on datasets/containers. Each rules is owned by a Rucio "**account**", and defines the minimum # of replicas that have to be available on a Rucio Storage Element (**RSE**), i.e. a storage space with attributes. RSEs can be grouped in logical ways (e.g. CLOUD=US, or Tier=1). Accounts manage (and are charged) for their own data with replication rules defined on datasets/containers and lists of RSEs
  - *Could a translation of such a rule into a concrete list of transfer tasks be engineered to be optimized on the basis of network-aware information? (e.g. naively: "choose the source RSE with best connection to the destination RSE"?)*

Level 2: "*medium*"-level i.e. **transfer "routing"**

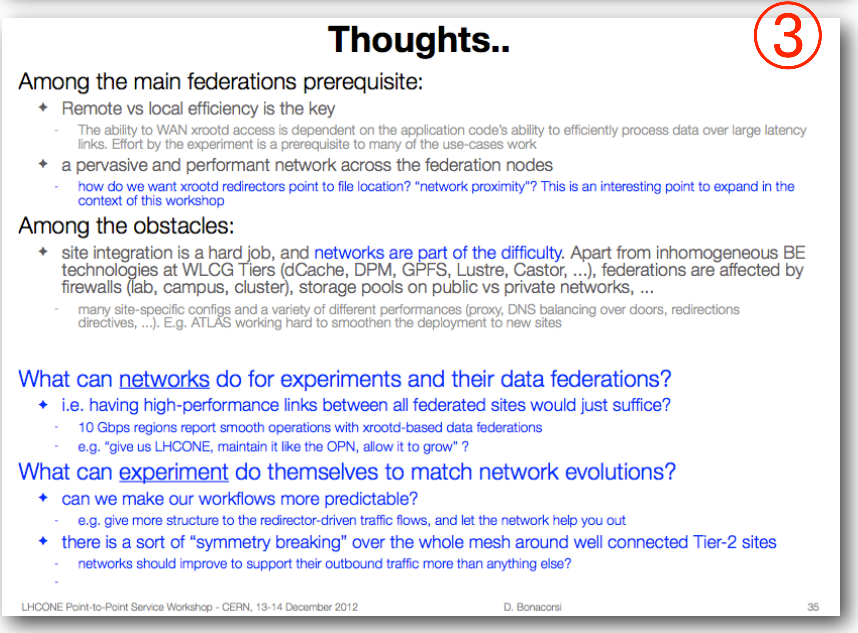
- *ATLAS Site Services or PhEDEx FileRouter could use network info at this level?*

Level 3: "*low*"-level i.e. **file-level transfer**

- *e.g. FDT used as the backend in the FileDownload agent in PhEDEx on the /Debug instance on just one link may be an existing proof of concept of a choice at this level?*

Food for thoughts...

LHCONE Point-to-Point Service Workshop - CERN, 13-14 December 2012 D. Bonacorsi 22



Thoughts.. ③

Among the main federations prerequisite:

- Remote vs local efficiency is the key
  - The ability to WAN xrootd access is dependent on the application code's ability to efficiently process data over large latency links. Effort by the experiment is a prerequisite to many of the use-cases work
- a pervasive and performant network across the federation nodes
  - how do we want xrootd redirectors point to file location? "network proximity"? This is an interesting point to expand in the context of this workshop

Among the obstacles:

- site integration is a hard job, and networks are part of the difficulty. Apart from inhomogeneous BE technologies at WLCG Tiers (dCache, DPM, GPFS, Lustre, Castor, ...), federations are affected by firewalls (lab, campus, cluster), storage pools on public vs private networks, ...
  - many site-specific configs and a variety of different performances (proxy, DNS balancing over doors, redirections directives, ...). E.g. ATLAS working hard to smoothen the deployment to new sites

What can networks do for experiments and their data federations?

- i.e. having high-performance links between all federated sites would just suffice?
  - 10 Gbps regions report smooth operations with xrootd-based data federations
  - e.g. "give us LHCONE, maintain it like the OPN, allow it to grow" ?

What can experiment do themselves to match network evolutions?

- can we make our workflows more predictable?
  - e.g. give more structure to the redirector-driven traffic flows, and let the network help you out
- there is a sort of "symmetry breaking" over the whole mesh around well connected Tier-2 sites
  - networks should improve to support their outbound traffic more than anything else?
  -

LHCONE Point-to-Point Service Workshop - CERN, 13-14 December 2012 D. Bonacorsi 35

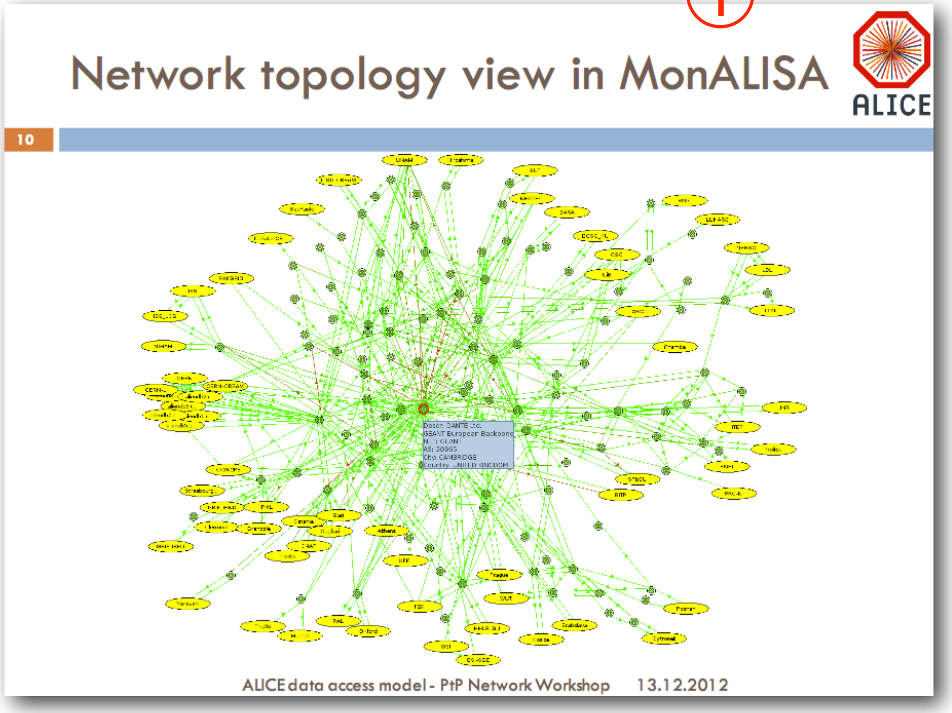
Thoughts on ALICE DM/WM

http://indico.cern.ch/getFile.py/access?contribId=7&sessionId=1&resId=0&materialId=slides&confId=215393

Costin presented an overview of ALICE DM/WM sectors and data access model

- An ambitious model with no dedicated SEs/sites/links for particular tasks, which worked because network capacity exceeded expectations ①
- Now the problems are more at an operational level, need more meaningful/realistic monitoring data, from site fabric to the backbone (e.g. replica discovery may profit from reliable network info) ②

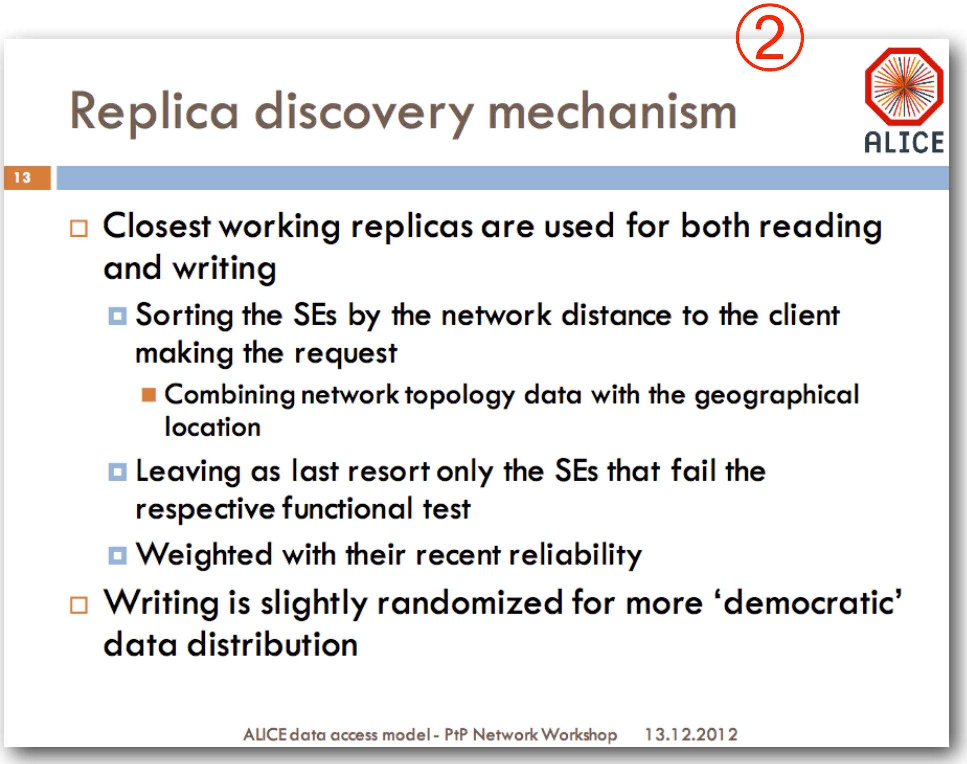
①



Network topology view in MonALISA

ALICE data access model - PtP Network Workshop 13.12.2012

②



Replica discovery mechanism

- Closest working replicas are used for both reading and writing
  - Sorting the SEs by the network distance to the client making the request
    - Combining network topology data with the geographical location
  - Leaving as last resort only the SEs that fail the respective functional test
  - Weighted with their recent reliability
- Writing is slightly randomized for more 'democratic' data distribution

ALICE data access model - PtP Network Workshop 13.12.2012

Network input

It was massive, in talks and discussions, and too much for this brief summary

- ✦ find more on slides and in the official “executive” summary that will come out soon

Let me just quote **ANSE** presented by Artur

- ✦ **A**dvanced **N**etwork **S**ervices for (LHC) **E**xperiments (ANSE), NSF funded (2 yrs)
- ✦ Goal: deterministic, optimized workflow
  - Use network resource allocation along with storage and CPU resource allocation in planning data and job placement
  - Improve overall throughput and task times to completion
- ✦ Integrate advanced network-aware tools in the mainstream production workflows of ATLAS and CMS
  - use tools and deployed installations where they exist
  - extend functionality of the tools to match experiments’ needs
  - identify and develop tools and interfaces where they are missing
- ✦ More at:
  - http://indico.cern.ch/getFile.py/access?contribId=8&sessionId=1&resId=1&materialId=slides&confId=215393
- ✦ In my opinion, quite close to the approach that experiments would get the highest benefit from

Summary

It seems promising to explore network-aware approaches in DM and WM sectors of LHC experiments

Carefulness is a must, though

- ✦ not obvious that e.g. ATLAS and CMS would interact to PtP/PtMP scheduling in the same way. Be pragmatic and see if we are actually in that position or if it requires significant work, more than the gain we would obtain in the experiment workflows

But, LS1 is an *excellent* opportunity window

- ✦ if not now, when?

Experiments are interested in trying to evaluate new approaches; choice driven by:

- ✦ level of pragmatic and productive interactions among Experiment and Network communities
- ✦ quality, cost, features of the interfaces offered by the Network experts
- ✦ the ATLAS/CMS manpower available to perform meaningful tests is limited

This workshop was an excellent starting point, and happened at the right time

- ✦ many thanks to Michael and Lars to organize it!
- ✦ an executive summary by the organizers will be available soon
- ✦ work-in-progress on a white paper, and some first collaborations established
- ✦ the WLCG Network group chaired by Michael is the perfect body to follow up

Credits: to all people who participated to the workshop discussions.