Towards Network Awareness in LHC Computing

- **LHC Run1:** Discovery of a New Boson
- **LHC Run2:** Beyond the Standard Model
  
  Gateway to a New Era

Artur Barczyk / Caltech
Internet2 Technology Exchange
Indianapolis, October 27, 2014
The Model until recently

Tier 1

- 10 – 40 to 100 Gbps
- London
- Paris
- Taipei
- Chicago

Tier 2

- 10 to N X 10 Gbps
- CACR

Tier 3

- 10 to N X 10 Gbps
- Institute

Tier 4

- 1 to 10 Gbps
- Workstations

The Culture of Innovation and Partnership continues, fostering a new generation of Networks: LHCONE, ANSE
PhEDEx: 10+ Years of Managed Data Transfers in CMS

In Addition: “Location Independent Access” (AAA)

To 3 PB/Week LHC Data Taking is Not the Only Driver
Larger Data Flows are Ahead, During LHC Run 2
ATLAS Data Flow by Region: 2009-2014

Excellent Grid performance was crucial for fast discovery of the Higgs boson.

~50 Gbps Average, 112 Gbps Peak
171 Petabytes Transferred During 2012

After vs Before LHC Run1
A New Plateau
6X Growth in 2013 Vs 2009

Most US LHC Tier2 sites upgraded to 100G
Location Independent Access: Blurring the Boundaries Among Sites + Analysis vs Computing

- Once the archival functions are separated from the Tier-1 sites, the functional difference between Tier-1 and Tier-2 sites becomes small [and the analysis/computing-ops boundary blurs]
- Connections and functions of sites are defined by their capability, including the network!!

Scale tests ongoing: 20% of data across WAN: 200k jobs, 60k files, (100TB)/day
● Peak upload rate: 26.9 Gbps
● Average upload rate over 1h of manual transfer requests: 23.4 Gbps
● Average upload rate over 2h (1h manual + 1h automatic): 20.2 Gbps
● Peak rate to CNAF alone: 20 Gbps

Now: 20 Gbps Milestone for US CMS Tier2 Sites with 100G
Plan by SC14: ~50 Gbps of production traffic from Caltech Tier2
Downloading Terabyte Datasets to Tier3s and Tier4s (to the desktop/laptop) is a use case that will be explored by CMS
Transfer Caltech ➔ Europe elevates usage of Internet2 to > 40% occupancy on some segments

Now: Developing path selection and load balancing methods using the Internet2 Transfer Caltech ➔ Europe raises usage to > 40% occupancy on some segments.
Map of the Global LHCONE Virtual Routing and Forwarding (VRF) Infrastructure Supporting Tier1/2/3 Connectivity

LHCONE: A global infrastructure for the LHC Tier1 data center and Tier 2/3 analysis center connectivity

The Major Network R&E Players in Support of the LHC Program

W. Johnston, ESnet

12 August 2014

See http://lhcone.net for details.
ANSE: Advanced Network Services for Experiments: Manage LHC data flows

- CC NIE project by Caltech, Vanderbilt, U. Michigan, UT Arlington
  - working with U. Victoria and Princeton
- Includes both US CMS and US ATLAS
- Interface advanced network services with LHC data management systems
  - PanDA in (US) Atlas [De et al.]
  - PhEDEx in (US) CMS [Wildish et al.]
- Network-Application (Middleware) interaction:
  - Monitoring (perfSONAR, MonALISA)
  - Bandwidth allocation (OSCARS, AutoBAHN, etc)

Example: PanDA Workflow Management System

- Production managers
- PanDA server
- Data Management System (DQ2)
- Local Replica Catalog (LFC)
- Logging System
- NDGF
- ARC Interface (aCT)

End-user → Worker Nodes
- Pilot
- Condor-g
- EGEE/EGI
- OSG

Submit → Submit
- Pull
- Job
- Etc.
Integrating Network Awareness in ATLAS Distributed Computing

ANSE: Kaushik De

25M Jobs at > 100 Sites Now Completed Each Month

6X Growth in 3 Years (2010-13)

Production and Distributed Analysis

- **STEP1:** Import network information into PanDA
- **STEP2:** Use network information directly to optimize workflow for data transfer/access; at a higher level than individual transfers alone
  - Start with simple use cases leading to measurable improvements in workflow/user experience
1. Faster User Analysis
   - Analysis jobs normally go to sites with local data: sometimes leads to long wait times due to queuing
   - **Could use network information to assign work** to ‘nearby’ sites with idle CPUs and good connectivity

2. Cloud Selection
   - Tier2s are connected to Tier1 “Clouds”, manually by the ops team (may be attached to multiple Tier1s)
   - To be automated using network info: **Algorithm under test**

3. PD2P = PanDA Dynamic Data Placement: Asynchronous usage-based
   - Repeated use of data or Backlog in Processing → **Make add’l copies**
   - Rebrokerage of queues → **New data locations**
     - PD2P is perfect for network integration
       - Use network for site selection – to be tested soon
       - **Try SDN provisioning** since this usually involves large datasets; requires some dedicated network capacity

**USE CASES**

Kaushik De
PhEDEx (CMS): Physics Experiment Data Export

- Tier2 sites pull data based on managers’ pre-placement decisions
- Download agents communicate via a Database [“blackboard”]
- Original assumption: network is a scarce and fragile resource
- Now need to adapt to higher speed + more reliable networks

Desired Upgrades (Wildish):
- Agile/flexible source-site selection
- CMS-wide scheduling to avoid competition on shared links and end points
- Wider range of use cases: Possibly include downloads directly to desktops
- Dynamic Circuits to control bandwidth for determinism in data movement

Hundreds of Petabytes transferred since 2004

https://cmsweb.cern.ch/phedex/
ANSE: CMS Developments

- Implemented circuit interface in PhEDEx
- Developed a site circuit agent
  - receives creation requests from download agents
  - checks the database and the lookup server to see if circuits are actually allowed on the current link
  - Handles the creation of the circuits

- Testbed: using dynamic circuits between Geneva and Amsterdam
  - Over US LHCNet, using OSCARS
  - First results very promising
- Plans: include other DYNES sites; move to pre-production then production use

More on ANSE/PhEDEX in Vlad Lapadatescu’s presentation
ANSE Deployment in LHCONE

- The LHCONE Point-to-point Bandwidth-on-Demand (BoD) Service is currently being developed, starting as a demonstrator (aka “experiment”) system
  - Step 1: multi-domain BoD connectivity based on emerging standards (NSI)
  - Step 2: add end-sites and HEP end-systems

- ANSE integration with ATLAS and CMS middleware provides the first application to leverage this infrastructure directly in the mainstream workflow systems of the LHC experiments

Participating domains, exchange points and end-sites
Picture compiled from notes by W. Johnston
Summary

• LHC will restart, at higher energy, in spring 2015
• Changes in the computing models and infrastructures ongoing
  – Upgrades to 100G
  – Move towards more flexible transfer patterns
  – Move towards more dynamic data movement
  – Move towards remote data access

• We investigate active use of network as a resource for increased determinism in
  – Workflow management
  – Data transfer management

• Use network information to take informed decisions at application level
• Use bandwidth reservation for deterministic data movement

• More details in following presentations in this session
THANK YOU!

Artur.Barczyk@cern.ch