Direct exploitation of a top500 supercomputer in the analysis of CMS data

*IFCA CSIC – Universidad de Cantabria  & **Universidad de Oviedo
SPAIN
THIS IS THE STORY OF A CHALLENGE...

- A general purpose, powerful (Top500 list), supercomputer in production in the University of Cantabria (SPAIN)
- A group of CMS researchers using Tier-2 CMS grid-based resources at IFCA center, but with peak demanding DATA processing needs to urgently prepare contributions to new papers (like Higgs WW, top cross section)
- An strategy: EXPLOIT RESOURCES, MAXIMIZE IMPACT, MINIMIZE EFFORT

...AND A SOLUTION
The Scene & the Actors

• **IFCA, Institute of Physics of Cantabria, Universidad de Cantabria-CSIC**
  - Basic Research center in Santander, SPAIN
  - HEP, Astrophysics, Statistical Physics & Computing
  - IFCA Data Center: Tier-2 center for CMS + additional NGI & FEDCLOUD resources
    - Several Clusters (>3600 cores)
    - HPC data storage (>2 petabytes)
    - 10Gb backbone and dark fiber to NReN
  - Management of the Spanish NGI (GRID expertise)
  - Hosting the UC node of the Spanish Supercomputing Network (RES)

• **ALTAMIRA SUPERCOMPUTER**
  - A new supercomputer acquired by the University of Cantabria in 2012
  - Designed with the support of IBM and BSC (Barcelona Supercomputing Center)
  - Supported by a high level (but limited in number) technical team

• **CMS researchers at University of Oviedo and at IFCA**
  - With responsibilities in the top and Higgs into WW channels
ALTAMIRA supercomputer

- TOP 500 in June 2012
  - #358
  - #2 in Spain
  - 240 “nodes” (IBM idataplex dx360m4)
  - ~4000 Intel cores: 330 Gflops/node
  - Top Efficiency: #36 worldwide

- Last generation FDR Infiniband (Mellanox)
  - Latency between nodes <1 microsec.
  - 40 Gbps

- Best solution for message passing in parallel jobs (in June 2012)
  - Excellent performance reported by CERN expert in Lattice QCD (summer 2012)
ALTAMIRA technical details

- **Hardware**
  - IBM idataplex cluster, 240 nodes dx360m4
    - 2x SandyBridge E5-2670 2.6GHz/1600 20MB
    - 64GB RAM, 16x4G DDR3-1600 DIMMs (4GB/core)
    - 500GB 7200 rpm SATA II local HDD
    - 332.8Gflop/node
  - HPC Infiniband FDR10 (40 Gbps)
    - FDR10 IB HCA Mellanox
    - 36 ports switches, leafs+ core layer
    - **FAT TREE non-blocking**
    - Advance Management
  - **Plus**
    - 7 IBM dx360m3 GPUs TESLA
    - 11 IBM ps702 Power7

- **Open Software**
  - xCat (Management)
  - Linux (Scientific Linux & Centos)
  - SLURM (queue management)
  - MPI (mvapich2, openmpi-x86_64)
  - Compilers (gcc, INTEL)
INTEGRATING ACCESS AND RESOURCES

- **ALTAMIRA** is managed as supercomputer
  - “Local” accounts (user/gr + passw.)
  - Queue system (SLURM)
  - GPFS limited home & scratch space
  - Network layer 2 over Infiniband

- **IFCA Clusters** are managed “a la GRID”
  - “Grid” access with local users access via LDAP to User Interface
  - Storage Element via STORM, underlying GPFS over 10Gb Ethernet

**INTEGRATION**

- COMMON IDENTITY FOR LOCAL USERS (ldap based) (handle carefully user/grid)
- SHARE GPFS VOLUMES OVER BOTH SIDES (including common home)
  - Over Ethernet for GRID resources, Over Infiniband for ALTAMIRA resources
  - EXPLOIT SIMILARITY IN JDL (scripts need very minor modifications)
  - Data Transfer Tools (vsftp, gridftp, xroot) + CMS software (via CVFMS)

- **EXPLOIT PERFORMANCE IN ALTAMIRA:**
  - LARGE NUMBER OF NODES + **INFINIBAND, 40Gb to each node, FOR DATA PROCESSING**
  - RUN o(100) JOBS IN PARALLEL WITHOUT IMPACT ON DATA ACCESS
ALTAMIRA INTEGRATION: Hardware View

DISK STORAGE
2.1 PB

DDN
874TB SATA RAID6

8Gbps FC

4x IBM 3650M3

10Gbps ETH

CMS DATA GPFS

GROUPS AREA GPFS

IBM DCS3700
400TB SATA RAID5

8Gbps FC

4x IBM 3650M3

4x4x1Gbps ETH

4Gbps FC

4Gbps IB

IBM DCS3700
400TB SATA RAID5

8Gbps FC

4x IBM 3650M3

4x10Gbps ETH

4x10Gbps IB

IBM DS2524
3TB SAS RAID5

4Gbps FC

2x IBM 3650

2x10Gbps ETH

IB BACKBONE

ETH BACKBONE

GRID WN

112x IBM HS21 BLADE

6x HP DL585G7

36x FUJI BX920S2 BLADE

34x HP BL460c Gen8 BLADE

IFCA DATACENTER

376 Physical CPU’s
2160 Cores
3136 Slots

SERVICES

50 Virtual Services
Over iSCSI

HA-LVM

IBM DS4700
1.6 TB FOR RAID5

IBM TOTAL STORAGE

LT03
IBM TOTAL STORAGE

LT05

326 Physical CPU’s
2588 Cores
5120 GPU Cores

BACKUP

2Gbps FC

8Gbps FC

200TB SAS RAID5

2x10Gbps ETH

1x IBM 3650

2x10Gbps ETH

2x1Gbps ETH

163x IBM x360 M4

ALTAMIRA

1. Cabrillo

3. Instituto de Física de Cantabria

ALTAMIRA INTEGRATION

Hardware View
ALTAMIRA INTEGRATION: Network View

ETHERNET CORE

- Force10 S2410
  - 20x10Gbps CX4
  - 4x10Gbps SFP+

- 4xG8052
  - 48x1GB + 4x10Gb SFP

- G8052
  - 48x1GB + 4x10Gb SFP

- Brocade RX8
  - 48 x 1Gbps
  - 40x10Gbps SFP+

- Force10 C300
  - 24x1Gbps
  - 8x10Gbps SFP+
  - 1x10Gb ER

- RedIris

INFINIBAND BACKBONE

- Altamira Nodes
- 10x Mellanox SX6036
- 5x Mellanox SX6036
- Fat Tree
  - 180x40Gbps IB
  - 163x40 Gbps FDR

- Altamira GPFS Serv Login
- 163x1Gbps ETH
- 5x1Gbps ETH
- 5x4Gbps FDR
ALTAMIRA INTEGRATION: Global View
INTEGRATION: GPFS details

• Connect all ALTAMIRA nodes to Ethernet IP network
  – Cabling the nodes with 1Gb Ethernet connection to switches
  – Deploy an optic 10Gbps + 10Gbps Trunk between IFCA Brocade Ethernet backbone and ALTAMIRA Ethernet switches.

• Setup the new 4 ALTAMIRA GPFS storage servers
  – Direct FC connection to DCS3700 storage cabin (1 Petabyte)
  – 4 Gb interfaces to access also IP storage network
  – IB FDR interfaces to serve all ALTAMIRA nodes
  – Create the 2 new ALTAMIRA file systems

• Add all ALTAMIRA nodes to IFCA GPFS cluster
  – At this point Altamira is able to access to different IFCA file systems
    • cms data
    • user home area
    • other IFCA projects
INTEGRATION: GPFS details

- Setting up the IB RDMA storage Network (GPFS > 3.4 is needed)

```
[root@node01 ~]# ibstat
CA 'mlx4_0'
  CA type: MT4099
  Number of ports: 1
  Firmware version: 2.10.700
  Hardware version: 0
  Node GUID: 0x0002c9030030e820
  System image GUID: 0x0002c9030030e823
  Port 1:
    State: Active
    Physical state: LinkUp
    Rate: 40 (FDR10)
    Base lid: 95
    LMC: 0
    SM lid: 46
    Capability mask: 0x02514868
    Port GUID: 0x0002c9030030e821
    Link layer: InfiniBand
```

```
[root@node01 ~]# ibstat
Infiniband device 'mlx4_0' port 1 status:
  base lid: 0x5f
  sm lid: 0x2e
  state: 4: ACTIVE
  phys state: 5: LinkUp
  rate: 40 Gb/sec (4X FDR10)
  link_layer: InfiniBand
```

Keep an eye on the device name: mlx4_0
INTEGRATION: GPFS details

– Tell GPFS to active verbsRdma for altamira nodes

```bash
# mmchconfig verbsRdma=enable --N “node1,node2,...,nodeN”
```

– Tell GPFS to active verbsPorts for altamira nodes

```bash
# mmchconfig verbsRdma=“mlx4_0” --N “node1,node2,...,nodeN”
```

– Restart GPFS on Altamira nodes

Loading modules from /lib/modules/2.6.32-358.14.1.el6.x86_64/extra

<table>
<thead>
<tr>
<th>Module</th>
<th>Size</th>
<th>Used by</th>
</tr>
</thead>
<tbody>
<tr>
<td>mmfs26</td>
<td>1762439</td>
<td>0</td>
</tr>
<tr>
<td>mmfslinux</td>
<td>310536</td>
<td>1 mmfs26</td>
</tr>
<tr>
<td>tracedev</td>
<td>29456</td>
<td>2 mmfs26, mmfslinux</td>
</tr>
</tbody>
</table>

Wed Sep 25 13:11:26.505 2013: GPFS: 6027-310 mmfsd initializing. {Version: 3.5.0.10  Built: May 7 2013 17:30:30} ...


INTEGRATION: GPFS details

— Testing GPFS connection over IB

[root@node01 ~]# mmfsadm test verbs conn

NSD Client Connections:
<table>
<thead>
<tr>
<th>destination</th>
<th>status</th>
<th>curr RW</th>
<th>peak RW</th>
<th>file RDs</th>
<th>file WRs</th>
<th>file RD KB</th>
<th>file WR KB</th>
<th>file recv KB</th>
<th>file send KB</th>
<th>file rcv KB</th>
<th>file snd KB</th>
<th>idx cookie</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

NSD Server Connections:
<table>
<thead>
<tr>
<th>destination</th>
<th>status</th>
<th>curr rdma</th>
<th>wait rdma</th>
<th>rdma RDs</th>
<th>rdma WRs</th>
<th>rdma RDs KB</th>
<th>rdma WRs KB</th>
<th>rdma recv KB</th>
<th>rdma send KB</th>
<th>idx cookie</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;c0n148&gt;</td>
<td>IBV_QPS_RTS</td>
<td>0</td>
<td>0</td>
<td>936415</td>
<td>909909</td>
<td>35095190</td>
<td>33079486</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>&lt;c0n195&gt;</td>
<td>IBV_QPS_RTS</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>9189003</td>
<td>238654045</td>
<td>313060231</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>&lt;c0n198&gt;</td>
<td>IBV_QPS_RTS</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>78</td>
<td>44</td>
<td>3427</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>&lt;c0n41&gt;</td>
<td>IBV_QPS_RTS</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>85</td>
<td>6825</td>
<td>3668</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

— RDMA supports zero-copy networking by enabling the network adapter to transfer data directly to or from application memory, eliminating the need to copy data between application memory and the data buffers in the operating system.
### Testing the Integration Performance

#### (No IB) Ethernet GPFS Server
- IBM x3650 M3 8 Cores E5520 2.27GHz, 16GB RAM
- 2 x 8Gbps FC SAN access
- 10 Gbps Ethernet

#### Infiniband GPFS Server
- IBM x3650 M3 4 Cores E5520 2.27GHz, 16GB RAM
- 2 x 8Gbps FC SAN access
- IB FDR10 40 Gbps

#### (No IB) Ethernet GPFS Client
- HP 585GL 48 Cores AMD 6176 SE 2.30GHz, 225GB RAM
- 10 Gbps Ethernet

#### InfiniBand GPFS Client
- IBM dx360 M4 16 cores E5 2670 2.60GHz, 64 GB RAM
- IB FDR10 40 Gbps

---

```
./gpfsperf "create/read/write" seq "fs" –n 32g –r 1m –fsync –th "n"
```

<table>
<thead>
<tr>
<th></th>
<th>CREATE</th>
<th>READ</th>
<th>WRITE</th>
<th>READ 8TH</th>
<th>READ 16TH</th>
<th>READ 32TH</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Eth Server</strong></td>
<td>1200MB/s</td>
<td>920MB/s</td>
<td>1200MB/s</td>
<td>925MB/s</td>
<td>930MB/s</td>
<td>925MB/s</td>
</tr>
<tr>
<td><strong>Eth Client</strong></td>
<td>460MB/s</td>
<td>300MB/s</td>
<td>455MB/s</td>
<td>308MB/s</td>
<td>315MB/s</td>
<td>305MB/s</td>
</tr>
<tr>
<td><strong>IB Server</strong></td>
<td>1700MB/s</td>
<td>1500MB/s</td>
<td>1850MB/s</td>
<td>1170MB/s</td>
<td>1420MB/s</td>
<td>1370MB/s</td>
</tr>
<tr>
<td><strong>IB Client</strong></td>
<td>1600MB/s</td>
<td>2290MB/s</td>
<td>1600MB/s</td>
<td>1132MB/s</td>
<td>1135MB/s</td>
<td>1135MB/s</td>
</tr>
</tbody>
</table>
ALTAMIRA & VO DATA (CMS)

- Now all VO Data is accessible from any ALTAMIRA node

```
[root@node1 ~]# df -h
Filesystem    Size  Used  Avail  Use% Mounted on
/dev/sda4     428G  2.1G  404G   1% /
tmpfs         32G   0    32G   0% /dev/shm
/dev/sda2     248M  63M  173M  27% /boot
/dev/sda1     50M  252K  50M   1% /boot/efi
/dev/projects 262T 175T  88T  67% /gpfs/csic_projects
/dev/gpfs.cms  874T  582T 292T  67% /gpfs/gaes
/dev/gpfs.users 5.5T  2.0T  3.6T  36% /gpfs/csic_users
/dev/res_projects 88T  29T  59T  34% /gpfs/res_projects
/dev/res_scratch 88T  2.5T  85T   3% /gpfs/res_scratch
```

- Install `cvmfs` and assure all nodes can access IFCA squid servers
SATISFYING LOCAL USER ANALYSIS PEAK DEMANDS

- ALTAMIRA jobs are limited to 72 hours, typically using 32-512 cores
- Short jobs (below 6 h) are prioritized to optimize filling
- Instantaneous “large & efficient” capacity using embarrassing parallel jobs: “wrap” multiple jobs, demand large number of cores!
- CMS use case:
  - Analysis jobs launched get access to the Tier-2 file system, including official CMS software and data
  - Intensive data processing jobs: CMS EDM event selection and filtering (aka skimming), and ROOT Tree production
  - Multiple batch submissions each wrapping ~ 150 jobs
  - Carefully balance between total number of jobs, complexity of scripts and control, and saturation of data transfer capacity

Typical reduction of waiting time, compared to analysis in Tier-2:

an order of magnitude
SATISFYING LOCAL USER PEAK DEMANDS

• Real example (April 2013): Skimming and ROOT Tree production over 2012 data samples for a real SUSY search analysis
  – 257,000 CPU hours
  – 17 TB Data Input (3 loops)
    • 13TB at /gpfs/res_projects/csic/
    • 5TB at /gpfs/csic_projects/cms/
  – 2.5TB Data Output
    • Moved to /gpfs/csic_projects/cms/
      – Accessible to CMS Tier3
      – Accessible to SRM

• Estimated time for processing in Tier-2: two months

• Finished in ALTAMIRA in less than one week
  – High Data Throughput (R/W)
  – No stageout fails (Sataured SRM)
  – Multiple batch submissions (150 jobs)
CONCLUSIONS

• ALTAMIRE supercomputer: an ideal system for LARGE DATA PROCESSING
• INTEGRATION with TIER-2 RESOURCES: LDAP, GPFS
• GPFS over INFINIBAND assures very good data transfer to any node
• VO software (CMS) was installed with low managerial effort (through CVMFS).
• CMS researchers were required to introduce only a very small modifications on their job submission scripts
• Typical reduction of waiting time, compared to our Tier-2: an order of magnitude faster!
  (from months to weeks, or from weeks to days)
• More than 500K hours used during 2013 in HWW, top and SUSY analysis (results already published in papers)
• AND THE KEY ADVANTAGE:
  Extra & efficient power available for analysis at peak periods