Track 6 Summary

Facilities, Infrastructure, Network

Phil DeMar, Josep Flix, Peter Hristov, Helge Meinhard, Eric Yen

CHEP2015 沖縄本島

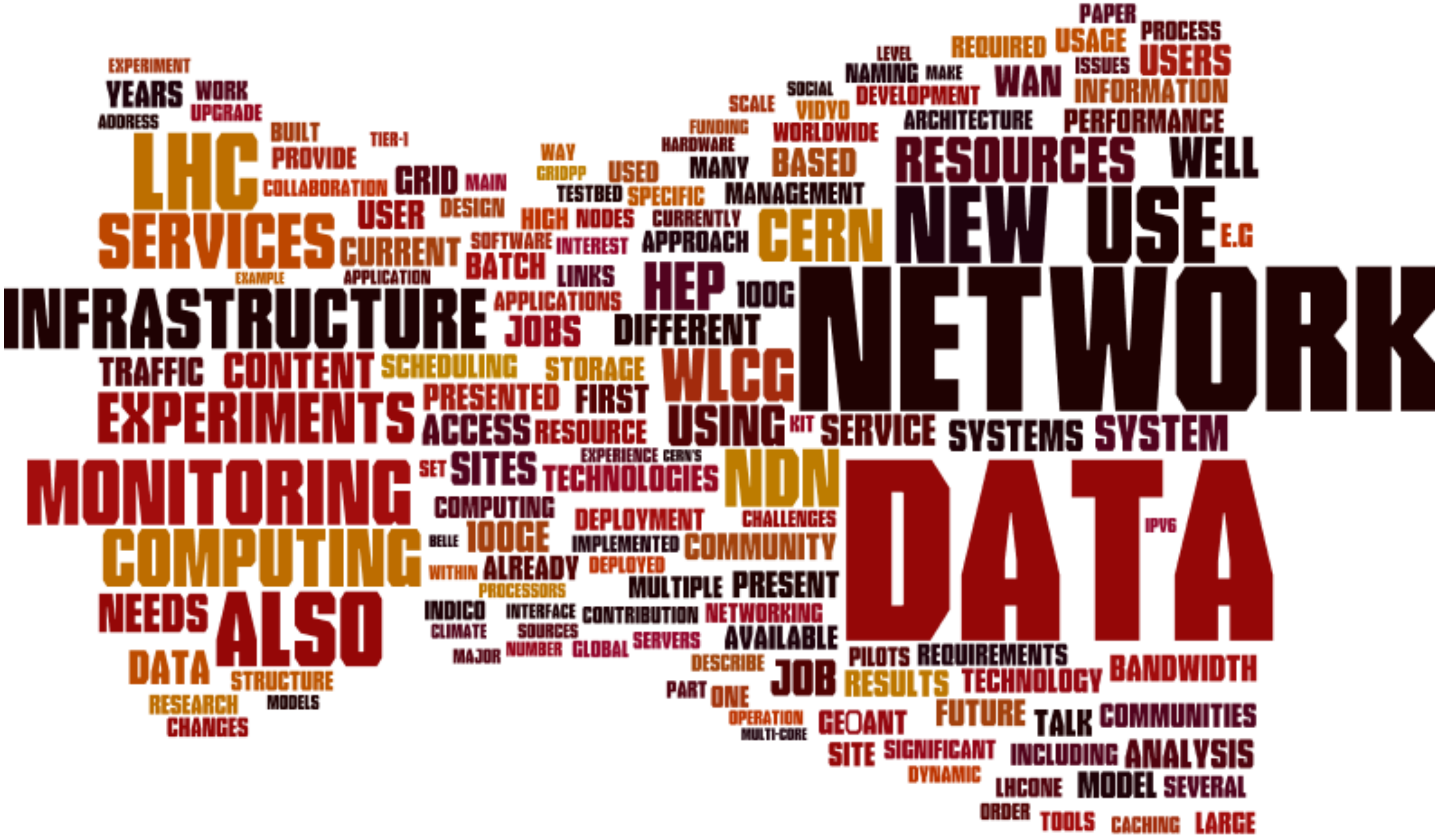
Overview Track 6 [Facilities, Infrastructure, Network]

- **54 contributions** presented to CHEP
  - 28 accepted as oral
  - 26 accepted as posters
- 4 sessions for Orals
  - Very good talks and with good attendance!
- 2 poster sessions
  - Very nice posters and coverage
  - Poster presenters very helpful

How did we prepare this summary?

- This summary is (should be) made in 25':
  - Disclaimer:
    - Difficult to credit all contributions with 1 slide/presentation or 1 slide/talk
  - We summarized the most important points for each of the sessions
    - It might be incomplete in some cases, sorry for this!
      - Look into the talks or chat with the authors if you are interested in specific issues
    - I am extremely grateful to all my convener colleagues for helping in this preparation
  - The poster contents are not covered in this talk
    - There is no time for it... Sorry! However, the posters were of excellent quality
    - I recommend you to get a look (or re-look) into the posters (uploaded to indico)

- ORALS -


PAPER
PROCESS
REQUIRED
USAGE
LEVEL
ISSUES
USERS
EXPERIMENT
NAMING
MAKE
WAN
INFORMATION
YEARS
WORK
SOCIAL
DEVELOPMENT
UPGRADE
SCALE
VIDYO
ADDRESS
ARCHITECTURE
PERFORMANCE
BUILT
FUNDING
WORLDWIDE
TIER-1
HARDWARE
LHC
PROVIDE
WAY
RESOURCES
WELL
MANY
BASED
GRIDPP
USED
COLLABORATION
GRID
MAIN
TESTBED
SPECIFIC
MANAGEMENT
USER
DESIGN
NEW
USE
HIGH
NODES
CURRENTLY
CERN
SERVICES
SOFTWARE
E.G
CURRENT
INTEREST
APPROACH
BATCH
EXAMPLE
APPLICATION
LINKS
HEP
100G
APPLICATIONS
INFRASTRUCTURE
NETWORK
JOBS
DIFFERENT
TRAFFIC
CONTENT
SCHEDULING
STORAGE
WLCG
PRESENTED
FIRST
EXPERIMENTS
USING
KIT
ACCESS
RESOURCE
SERVICE
SYSTEMS
SYSTEM
EXPERIENCE
CERN'S
SET
SITES
TECHNOLOGIES
NDN
MONITORING
COMPUTING
DATA
DEPLOYMENT
CHALLENGES
IPV6
BELLE
100GE
IMPLEMENTED
COMMUNITY
COMPUTING
WITHIN
ALREADY
DEPLOYED
PROCESSORS
MULTIPLE
PRESENT
NEEDS
ALSO
INDICO
INTERFACE
CONTRIBUTION
NETWORKING
CLIMATE
SOURCES
AVAILABLE
MAJOR
NUMBER
GLOBAL
SERVERS
PILOTS
REQUIREMENTS
DATA
DESCRIBE
STRUCTURE
JOB
BANDWIDTH
PART
RESULTS
TECHNOLOGY
ONE
RESEARCH
MODELS
OPERATION
FUTURE
CHANGES
GEANT
TALK
COMMUNITIES
MULTI-CORE
SITE
SIGNIFICANT
INCLUDING
ANALYSIS
DYNAMIC
LHCONE
MODEL
SEVERAL
ORDER
TOOLS
CACHING
LARGE

Session 1: Mon 14:00 h – 16:00 h [C210] → Convener: J. Flix

- Analysis of CERN computing infrastructure and monitoring data (Christian Nieke / CERN)
- Monitoring evolution at CERN (Pedro Andrade / CERN)
- Challenge and future of job distribution at a multi-VO Grid site (Birgit Lewendel / DESY-HH)
- Procuring Open-Compute hardware (Olof Barring / CERN)
- Building a Tier-3 based on ARM-64 SoC for WLCG (Peter Elmer / Princeton)
- Integrating network and transfer metrics to optimise transfer efficiencies and experiment workflows (Shawn McKee / Michigan)
- 100 G deployment at DE-KIT (Bruno H Hoeft / KIT)
- A first look at 100 Gbps LAN technologies, with an emphasis on future DAQ applications (Adam Otto / CERN)

Session 1: Monitoring

- Monitoring is very important to ensure fully effective and efficient exploitation of resources
- New infrastructure developed at CERN: well-scaling modular system for alerts, displays, archives, streaming, …
- Analytics project at CERN to enable integrated studies using data from different data sources crossing service boundaries [CERN/Wigner]
- Many (new) tools used: Hadoop/MapReduce, ElasticSearch, Kibana, …
  - In particular, looking to technology trends for improvements
- PerfSONAR offers an integrated network monitoring system
  - Becoming more important to detect network degradations/failures

Session 1: Computing platforms

- Experience with Open Compute hardware at CERN
  - Few suppliers, getting specifications right is difficult
  - Significant power saving, potential for even more
  - Price in line with expectations based on volume
  - Might be tendering to buy some OPC CPU and disk – other sites could benefit from this process
- WLCG worker nodes with ARM 64-bit system-on-chip
  - ARM becoming commodity, much like x86 in the late ‘90s
  - Power/performance ratio is interesting
  - Successful proof-of-concept of a Tier 3 with ARM-based servers as workers, batch system head nodes, …
  - CMS tests with CMSSW – still early days for software ecosystem

Session 1: Networking (1)

- Network and transfer metrics
  - PerfSONAR (end-to-end monitoring) fully deployed
  - SDNs offer potential to further optimise workflows
  - Integration project for network and transfer metrics
    - particularly interesting to 'plug' network knowledge from PerfSONAR/FTS3 into experiment *Data Management* frameworks
- 100 Gbps deployment at KIT
  - Tested since 2010, including a transatlantic demo at SC13
  - LCHOPN saturating during Run1
    - LHCOPN/LHCONE and general Internet require higher bandwidths as of now
  - Multi-year project to upgrade to 100 Gpbs
    - Network is not free (in term of efforts and cost)

Session 1: Networking (2)

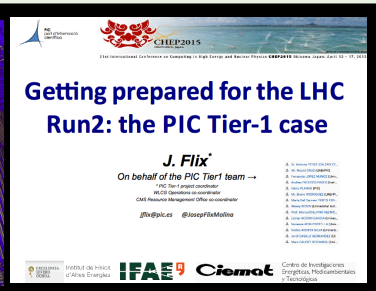
- 100 Gbps LAN for LHCb DAQ – required for upgrade during LS2
  - Possible solutions → INTEL OmniPath, 100 Gbps Ethernet, EDR Infiniband
  - 100 Gbps Infiniband tests: NIC→NIC and NIC→Switch→NIC
  - Measured bandwidth, latency, CPU utilisation, memory bandwidth
    - Several optimizations in the BIOS and in the kernel
  - Obtained 80 Gb/s unidirectional, 150 Gb/s bidirectional
  - Further improvements to be expected
    - pre-production firmware/driver was used

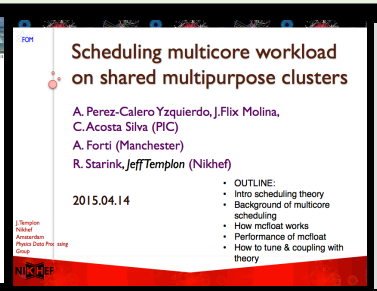
Session 1: Miscellaneous

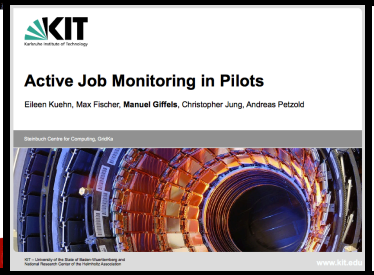
- Job distribution at multi-VO site (DESY)
  - Torque/MAUI: scaling issues beyond 10k jobs, heterogeneous workers not well supported
  - Scheduler study/development, based on the torque C-API
    - Simple re-ordering of job list, preferring multi-core jobs
    - Better use of the farm
  - Pilot jobs change grid model – vacuum model attractive

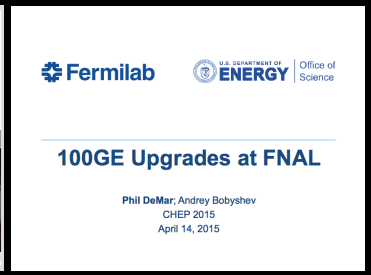
Session 2: Tue 14:00 h – 16:00 h [C209] → Convener: H. Meinhard

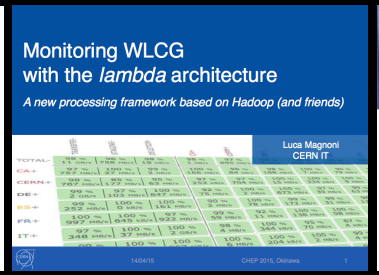
- GridPP – preparing for Run-2 and the wider context (Jeremy Coles / Cambridge)
- Getting prepared for the LHC Run2: the PIC Tier-1 case (Josep Flix / PIC)
- Scheduling multicore workload on shared multipurpose clusters (Jeff Templon / NIKHEF)
- Active job monitoring in pilots (Manuel GIFFELS / KIT)
- Migrating to 100GE WAN Infrastructure at Fermilab (Phil DeMar / FermiLab)
- Monitoring WLCG with lambda-architecture: a new scalable data store and analytics platform for monitoring at petabyte scale (Luca Magnoni / CERN)
- A Model for Forecasting Data Centre Infrastructure Costs (Renaud Vernet / CC-IN2P3)
- High-Speed Mobile Communications in Hostile Environments (Stefano Agosta / CERN)

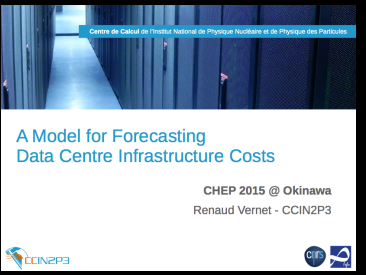


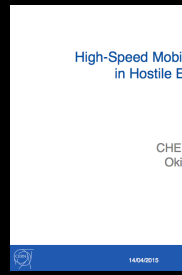












Session 2: Computing Sites (1)

- A few computing sites presented their readiness for LHC Run2
  - Required substantial R&D and tuning: new protocols, data federations, ...
- Costs concerns / **reducing costs**:
  - Detailed studies on estimated efforts for site tasks
  - Detailed studies on hardware costs, electricity costs and trends
  - Infrastructure improvements (free-cooling techniques, UPS upgrades, ...)
  - Detailed plans / costs for network upgrades
  - Deployment of more flexible, cost effective and easy ways to operate services
    - Batch system consolidation (HTcondor, moving from CREAM-CEs,...)
    - Cloud resources R&D / VAC context
  - People operate/run the services, perform R&D, tuning, ask/drive projects
    - **You(we) don't want to lose people!**

Session 2: Computing Sites (2)

- IPv6
  - Many sites addressing migration to IPv6 (dual-stack), within the HEPIX IPv6 WG
- Multicore
  - LHC moving towards multicore – major customer at the sites
  - Efficient scheduling multicore workload on shared multipurpose clusters is needed
    - Controlled draining of WNs is a must
    - Dynamic partitioning works nice for Torque/MAUI (mcfloat) → only 1.5% of resources wasted
    - Implemented at many WLCG sites
- Monitoring pilot jobs
  - Prototype tool presented by KIT which can monitor pilot jobs almost, if not, in real-time
  - Tested with CMS pilots – batch system agnostic tool
  - Capable to know if a job is interacting to local storage or talking to federations

Session 2: Computing Sites (3)

- FNAL presented their plans for 100 GBE migration on their WAN
  - FNAL host the most important Tier-1 center for CMS
    - CMS traffic dominates
  - Science network is isolated / Policy-Based Routing in place
  - Already one 100 GBE is deployed through ChiExpress to StarLight (ESnet)
    - A new 100 GBE, with independent path, to be deployed
    - Can be used as fail-over backup (also, traffic could be load-balanced on the 2x100 GBE)
  - 4x10 Gbps channels in place (Gen. IP and backups)
  - Long term – evaluate SDN-based technologies (OpenFlow) to isolate on-site science data traffic

Session 2: Monitoring

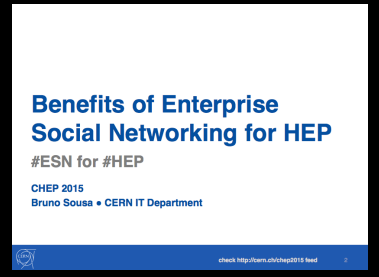
- CERN presented a change of technology for monitoring (***Dashboard***)
  - Current system does not scale
  - Spikes on monitoring information affects UIs
  - Data re-processing is expensive (we have more and more data)
- Monitoring with lambda architectures
  - Hadoop/MapReduce as and appropriate move (+Flume/ElasticSearch/...)
    - Fast data processing: a few minutes to reprocess 2-weeks worth of data!
  - Processing scales with data volume
  - Less code to maintain and operate / easy to upgrade a component
- Some dashboard tools will migrate to the new schema this summer

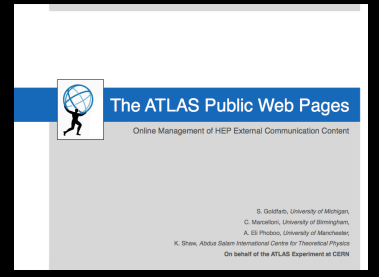
Session 2: Network/Mobile communications

- Have you ever wonder how to transmit voice/data in the LHC tunnel? Stefano's talk!
  - Hostile environment → **radiation**
- Several tools providing data while LHC is running:
  - Train inspection mono-rail
  - Temperature & humidity sensors
- Standard WiFi not convenient (size of the tunnel / radiation)
  - High Speed LTE (4G) technology deployed
  - Antenna cable deployed along the tunnel (resistant to radiation)
  - Repeaters / master units placed outside radiation areas (in the caverns)
- temporary VDSL during LS1 → moved to LTE system
  - High throughputs measured in tunnel + caverns, and ready for use

Session 3: Thu 09:00 h – 10:30 h [B503] → Convener: P. Hristov

- Indico - the road to 2.0 (Pedro Ferreira / CERN)
- Vidyo@CERN: A Service Update (Joao Correia / CERN)
- Status report of the migration of the CERN Document Server to the invenio-next package (Esteban Gabancho / CERN)
- Enterprise Social Networking Systems for HEP community (Bruno Silva De Sousa / CERN)
- ATLAS Public Web Pages: Online Management of HEP External Communication Content (Steven GoldFarb / Michigan)
- Offering Global Collaboration Services beyond CERN and HEP (Joao Correia / CERN)





Session 3: Collaborative tools (1)

- Indico 2.0
  - Planned project to replace old (Python-based) InDiCo code
  - Module by module migration over a 12 month period
    - Challenging project – careful plan is being followed
    - 36% of legacy code has already been replaced
      - New software stack is based on (more/today's) standard technologies
    - Public release (2.0) will be made on the 1st half of 2016
- Vidyo
  - increased use, infrastructure expansion, recording capability:
    - 20k registered users; 1600 simultaneous connections
    - Time in conference 50M minutes/yr (35k hours/yr) → **larger than many commercial providers (!)**
  - Expanded # of routers & gateways (now in Africa – South America next)
  - Majority of users are satisfied. Survey was done
    - areas for improvement identified / new functionalities to be introduced (Lync/Skype gateways, …)

Session 3: Collaborative tools (2)

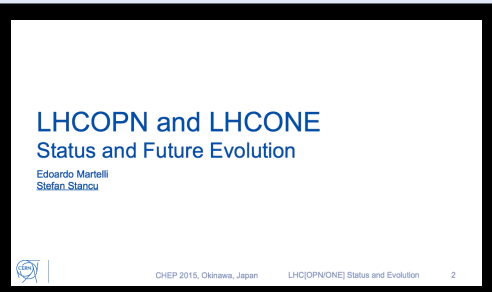
- CERN Document Service (CDS)
  - Usage statistics: 1.4M records; 178k photos; 9k videos; 39k archives (not only LHC)
  - Enhancing the Invenio code that CDS is based on
    - Developed more comprehensive statistics with a new user GUI
    - Author tools: CDS can now read from LDAP or experiment databases (IDs harvesting & names)
    - Enhanced commenting tools (we make ~500 comments/week!)
- Social Networking for HEP
  - Average interaction worker spends 28% of their time on email
    - Mail is: too flexible / locked in knowledge / collaboration is painful
  - Social media makes more sense for collaboration: knowledge base built by community
  - Enterprise Social Networking (ESN) helps in:
    - Identifying expertise & capturing tacit knowledge
    - Flattening communication
  - CERN has a pilot ESN effort (social@CERN)
    - In early stages; participation is voluntary → http://social.cern.ch/ - **DON'T BE SHY BE SOCIAL!**

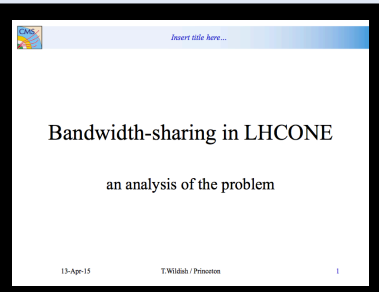
Session 3: Collaborative tools (3)

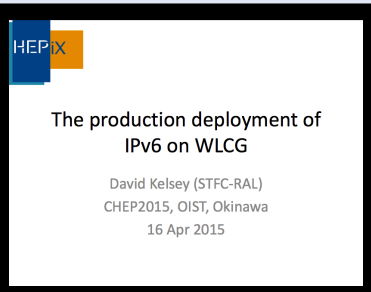
- Atlas Public Web Pages
  - Current ATLAS public web page still work, but needed updating:
    - Limited content workflow & integration with social media. Design is old
    - Desire alignment with CERN Drupal support
  - Three areas of focus: target audience; recommendations for content; performance
    - Following CERN guidelines for content
  - Finishing up with development, complete port by end of year (very professional!)
- Global Communications Services beyond CERN & HEP: aka Vidyo/InDiCo
  - Approached by non-HEP disciplines about using it
  - Both requires tie-in w/CERN for support
    - And requires CERN accounting (guests can't be a meeting owner)
  - Exploring advancing it to a Global Service
    - Discussing partnership w/GÉANT (facilitator – can charge)
    - Full cost recovery model that doesn't disrupt commercial competition
  - Legal & business models proposed with the aim to become a global service by 2016

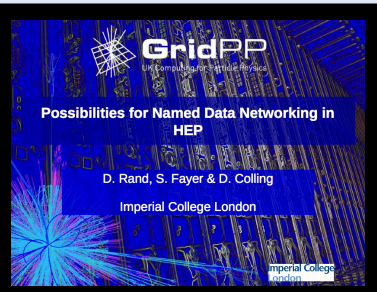
Session 4: Thu 11:00 h – 12:30 h [B503] → Convener: P. DeMar

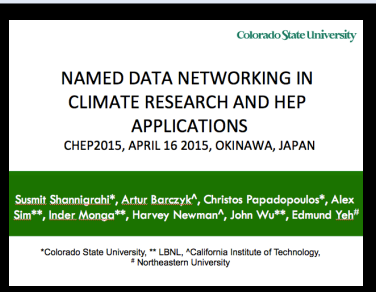
- LHCOPN and LHCONE: Status and Future Evolution (Stefan Nicolae Stancu / CERN)
- The GÉANT network: addressing current and future needs for the High Energy Physics community (Vincenzo Capone / DANTE)
- Bandwidth-sharing in LHCONE, an analysis of the problem (Tony Wildish / Princeton)
- The production deployment of IPv6 on WLCG (Dave Kelsey / SFTC Rutherford)
- Possibilities for Named Data Networking in HEP (Duncan Rand / Imperial College)
- Named Data Networking in Climate Research and HEP Applications (Christos Papadopoulos / Colorado State University)











Session 4: Network (1)

- LHCOPN and LHCONE status and future evolution
  - More and more regions in the LHCONE
    - Level3 VPN – P2P service in development
  - Network upgrades @CERN & Esnet, and elswhere → improving site's connectivity
    - BELLE-II becomes member (KEK, PNNL) – integrated in the LHCONE
  - New Acceptable Use Policy available
  - Benefit from PerfSONAR for network monitoring: ~250 instances
- GÉANT – High-bandwidth, high-performance pan-European communications infrastructure for research and education
  - GÉANT upgrade in 2013: latest transmission and switching technology, improved connectivity (also extra-EU), easy to upgrade
  - All the current and future HEP communities will be part of LHCONE: BELLE-II joined, Pierre Auger Observatory interested, ILC(?)
  - GÉANT: support the emerging trends and needs (SDN, dynamic circuits, CDN, etc.)
  - The GÉANT network is ready **NOW** to satisfy **TOMORROW's** users needs

Session 4: Network (2)

- Bandwidth-sharing in LHCONE: an analysis of the problem
  - LHC Run2 is more demanding →bandwidth-allocation at LHCONE requires a mechanism which is fair, efficient, lightweight, responsive and automatic
  - The **Progressive Second-Price auction** offers this
    - Users negotiate among themselves how much bandwidth they should get on repeat auction (when conditions changed)
    - Network providers get clear statement of what users want at any point in time
  - "Fake money" in HEP with real value for the experiments → similar to the batch quota allocation
- The production deployment of IPv6 on WLCG (HEPIX IPv6 WG)
  - ~10% of sites will run out of IPv4 addresses in 1-2 years
  - IPv6 tests completed in Feb 2015, after two years → more realistic tests to follow (FTS3/dCache): several issues found, per-component view of IPv6 readiness provided
  - Deployment of WLCG dual-stack is starting
    - LHCOPN/LHCONE routing & perfSONAR monitoring
  - Controlled deployment of dual-stack storage and central services

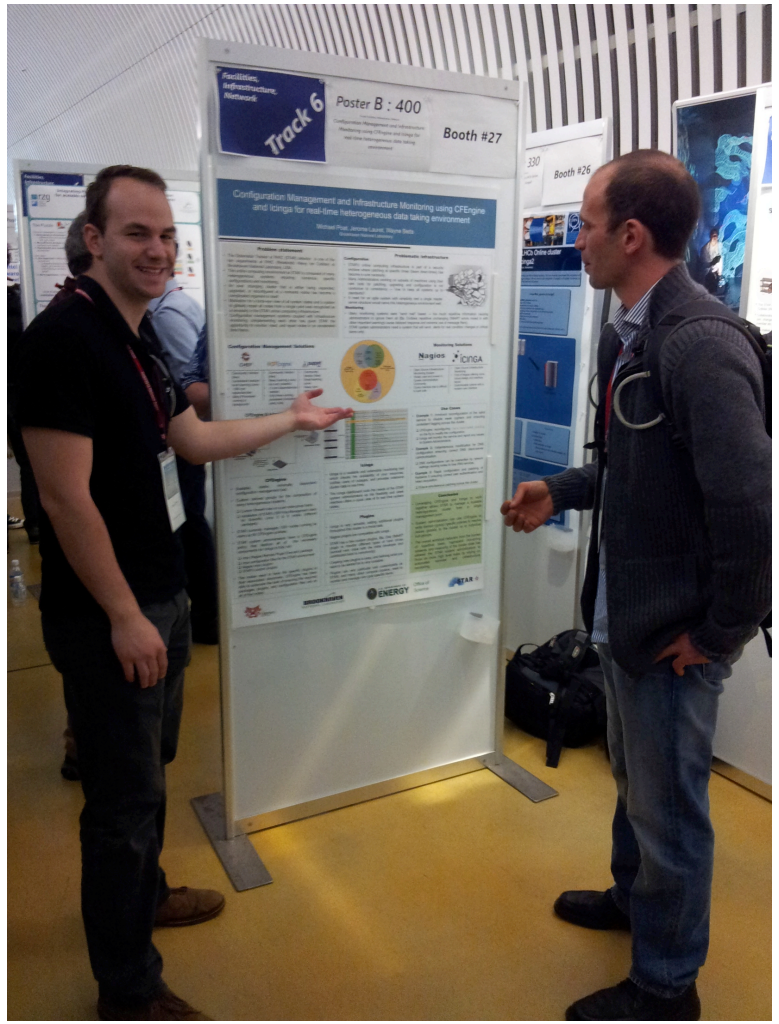
Session 4: Network (3)

- Named Data Networking (http://named-data.net)
  - Part of the NSF Future Internet Architecture initiative (multi-institutional)
    - NDN collaborative Effort to Promote and Sustain the NDN Future Internet Architecture
  - Provides A possible solution for complex data placement/access
    - Need of data, regardless its location → you ask question (to network) – you get answers (data)
    - Named data is signed + encryption
    - Data can easily be cached in routers → fast access to data ("hot" files)
  - HEP – Xrootd vs NDN: no manager, more robust
  - HEP – ROOT NDN plugin → examining the viability of this new networking paradigm in a test-bed (with routers, etc…)
  - Climate research: test-bed in place (central-west US sites)
    - Cisco and others currently building (more) specialized routers – **industry interest!**
- Looks interesting/promising!

- POSTERS -

NEEDED
PART
PLUG-INS
NETWORKS
WORK
JOBS
LHC
WITHOUT
AUTOMATION
ONE
HARDWARE
GLANCE
COMPLEX
RECORD
COLLABORATION
ALLOWS
TEAM
USED
DEVELOPED
FRAMEWORK
ACTIVITIES
DYNAMIC
FRAMES
PROVIDING
MANAGE
PROCESS
LYNC
SOFTWARE
DISTRIBUTED
SERVICE
TRAFFIC
TECHNOLOGY
PUNDIT
FEATURES
EXAMPLE
ACCESS
SERVICES
GRID
NETWORK
USERS
ORDER
HIGH
PHYSICS
AIR
TOOL
TASKS
ENERGY
SDN
RESEARCH
BASIC
ANALYSIS
TOOLS
CORES
ICINGA
RUN
MANY
COMMUNITIES
IHEP
FUNCTIONALITIES
WELLINFORMATION
EXPERIMENT
APPLICATION
INTEGRATION
ALLOWING
SYSTEMS
INCLUDING
ENVIRONMENT
UPGRADES
FLOW
PROBLEMS
COLLABORATION
INFRASTRUCTURE
PAPER
MONITORING
CONFIGURATION
USE
DISCUSSED
CLUSTER
TIER-2
PROJECT
CENTRAL
USER
REQUIREMENTS
RESULTS
MODEL
INTERNATIONAL
PERFORMANCE
BASED
SYSTEM
DEVELOPMENT
TRANSFER
SHIFT
STORAGE
COMPUTING
DESCRIBE
THROUGHPUT
HEP
APPLICATIONS
LAST
WLCG
NUMBER
NEED
STAR
QUALITY
EXPERIMENTS
SOLUTION
RUNNING
PRESENT
PRIVATE
DETECTION
ALICE
MANAGEMENT
DATA
RESOURCES
PROVIDE
COLLIDER
SITES
CORE
CENTER
DEPLOYED
ALSO
ADDITION
STATUS
WEB
ATLAS
MEMBERS
BATCH
KEK
YEAR
NEW
SUPPORT
STATISTICS
TIME
OPERATION
USING
MULTICORE
SET
CERN
PUPPET
PATH
AREAS
JOB
DESIGN
REQUIRES
APPROACH
EFFICIENCY
REPORT
POWER
WAY

Poster Sessions



Poster Sessions

very nice location!

Poster presenters really helpful

Appreciated quality: contents & design


whatis
next?

Trends from the track

- Consolidation to using more flexible/scalable tools, at all levels
- Costs optimizations at the Facilities & Infrastructures
  - Transtition of established services (as of today) to more flexible solutions
    - Services asy-to-operate, and at scale (i.e. changes at many sites)
    - Even, a paradigm change in the way we exploit resources at the sites
- Increased networks around (full deployment of 100 GBE elsewhere)
  - Network is becoming more and more important → we'll have better use/monitors
  - IPv6 for sure... SDNs / Named Data Networks in production?
- Better tools and techniques to monitor "data"
  - subject to revision cycles: what's good today, cannot be used in the future (challenge)
- Better tools for us to better interact, share our results, share knowledge
  - which will positively impact in the quality of the work done

Thanks on behalf of Track6 conveners…

- to speakers and poster presenters for the interesting topics
- we acknowledge the speaker efforts in being on time! (12' talk – not easy)
  - … and the effort spent by the track conveners to keep it on time
  - **(1)+38'', (2)+56'', (3)+14'', (4)+23''** ← Sessions finished within 1' additional time

but successful!

- to the audience of the parallel sessions for their interest and the vital discussions
  - … and the efforts to make short questions! (within 3')
- to the LOC team, and all the scientific secretaries on-site
- **to the conference organizers for a great CHEP 2015!**

ありがとう！